A gentle introduction to modeling ecological niches and species distributions

Different Definitions of Niche

- Many definitions. Key concepts:
  - **Grinnelian niche** – habitat requirements and characteristics that foster persistence
  - **Eltonian niche** – focused on communities and their trophic levels and interactions. Typically considered at local scales.
  - **Hutichinsonian niche** – Hutchinson considered niche as a n-dimensional hypervolume defining both environmental conditions and resources. Distinguished bet. fundamental and realized niche

- The full range of environmental conditions (biological and physical) under which an organism can exist describes its **fundamental niche**.

- Because of biotic interactions (competition, predation), the **realized niche** is almost always smaller.

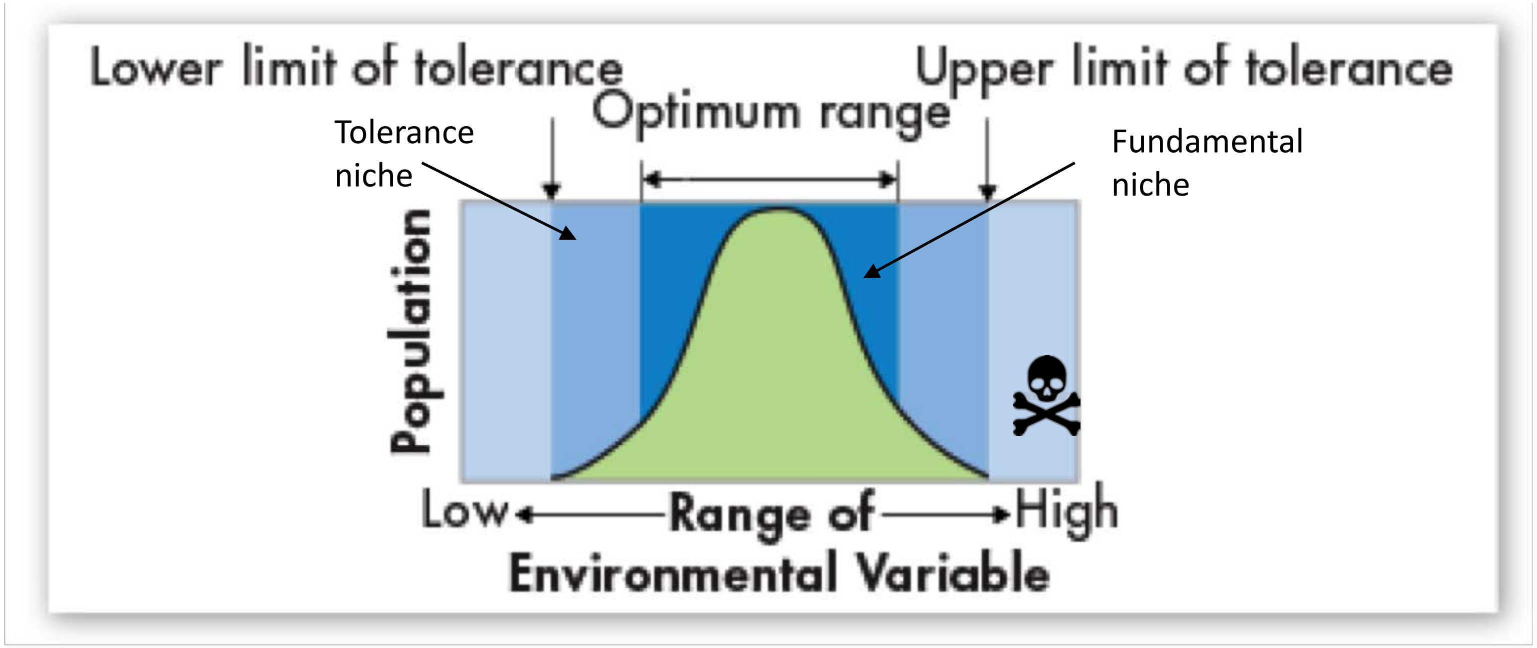
Abiotic Requirements

- Grinnelian niche
- Non-consumable resources
  - Climate
  - Geophysical characteristics
  - Substrate
  - Nutrients

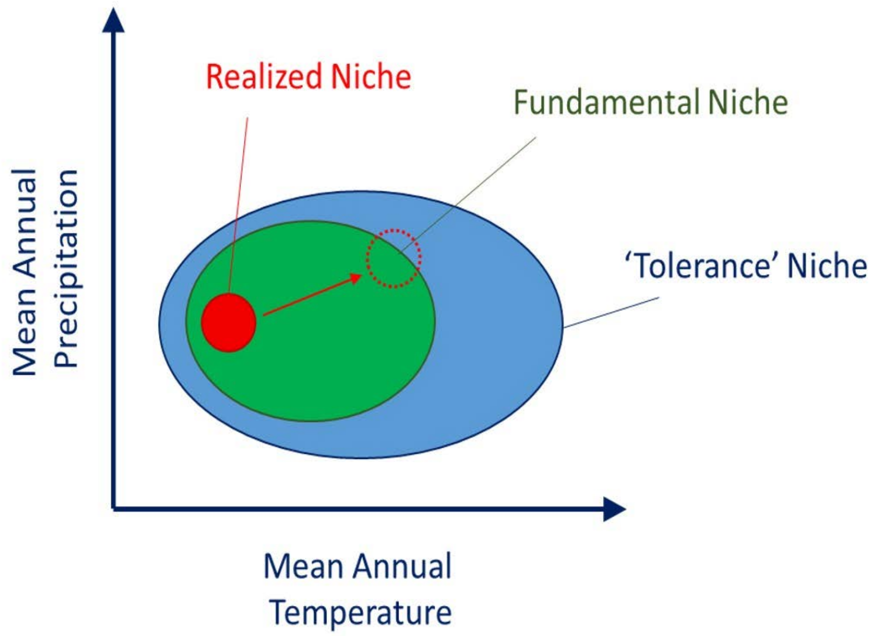
Biotic Requirements

- Eltonian Niche
- Interactions with other species
  - Competition
  - Exploitation
  - Commensalism
  - Mutualism

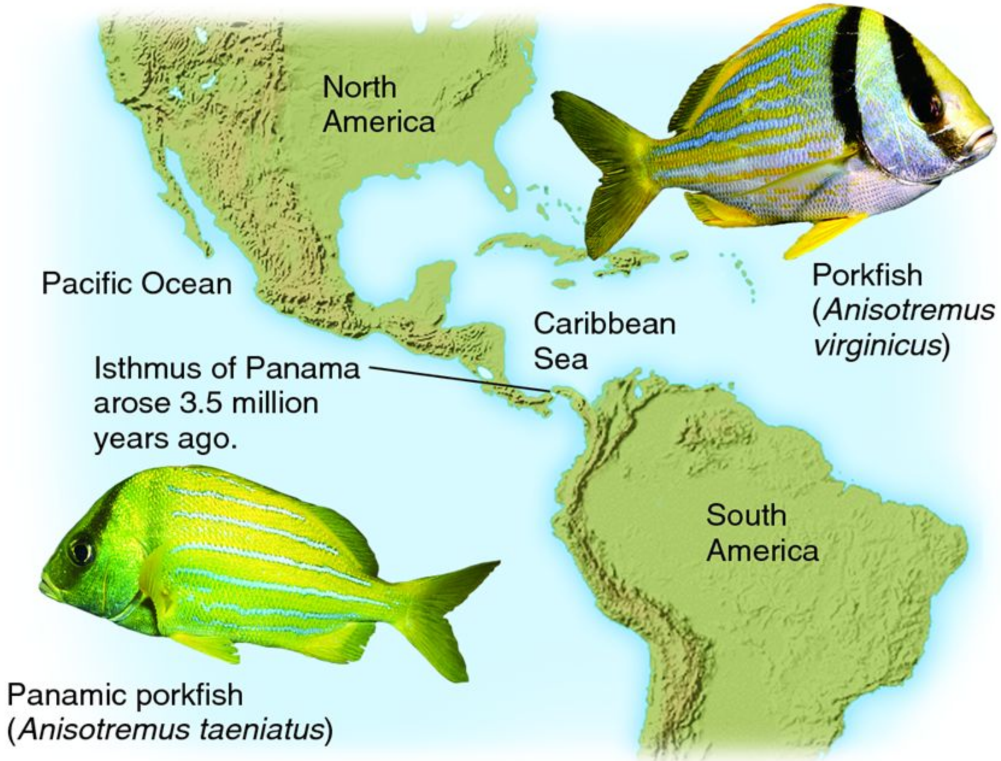
Fundamental niche intimately tied to population/species physiological tolerances



The realized niche is a subset of the fundamental niche - it is the “occupied niche” of the species and includes biotic interactions



Movement



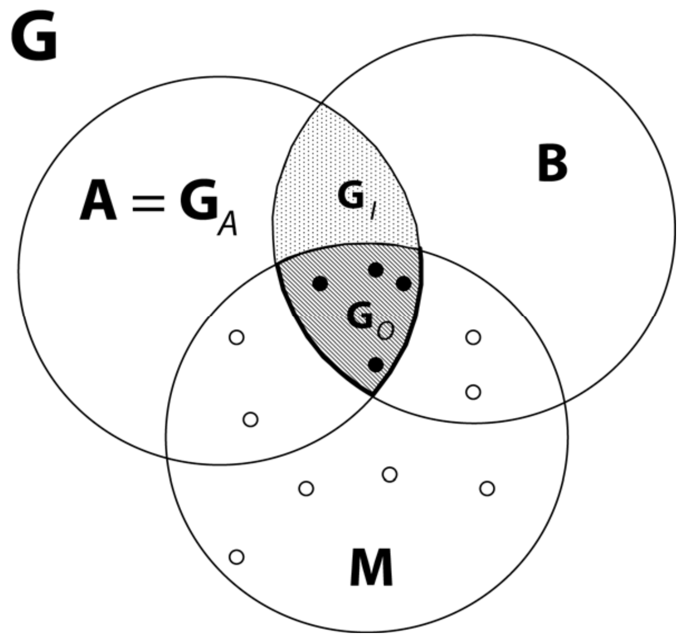
- Where originated
- Dispersal ability

Conceptual background

- Will focus on conceptual basis about what is being modeled in most correlative modeling approaches
- Before beginning: is there a difference between ”ecological niche modeling” and ”species distribution modeling”?
  - To model species distributions, it is critical to also model the niche
  - SDM also incorporates information about movement biology e.g. dispersal/colonization
  - This means we often are working in geographic space and environmental spaces and switching between them

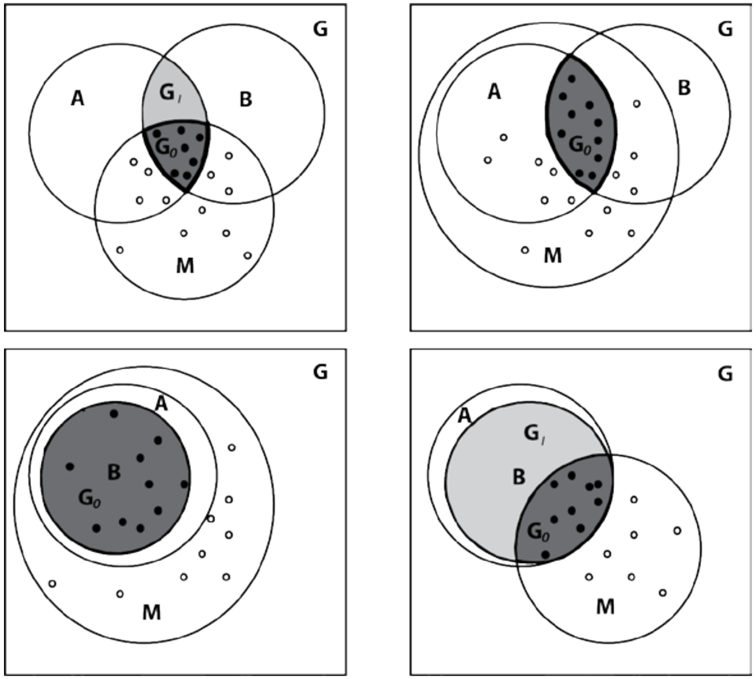
The BAM Diagram

- Why are species found where they're found?
- B – biotic, A– abiotic, M - movement
- Areas of overlap of BAM are critical
  - Potential distributional area
  - Occupied distributional area



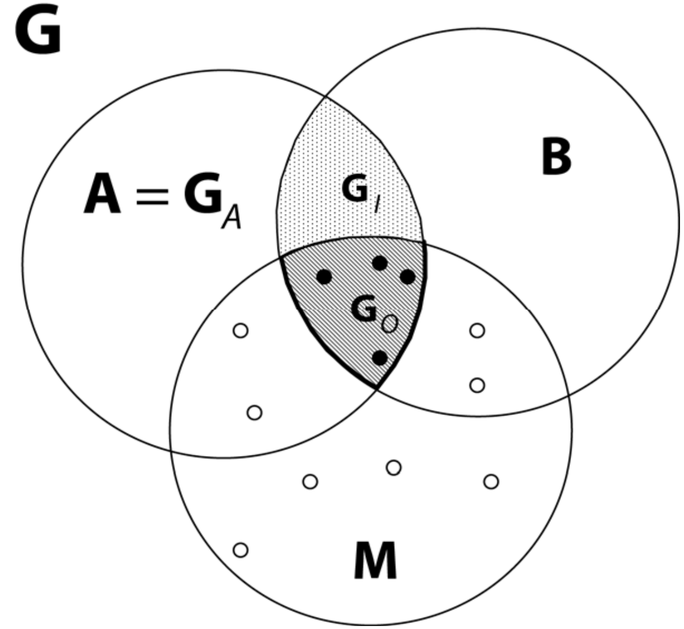
The BAM Diagram: Alternative Scenarios

For each unit of diversity (e.g. species) these overlaps are different (and hard to know apriori)



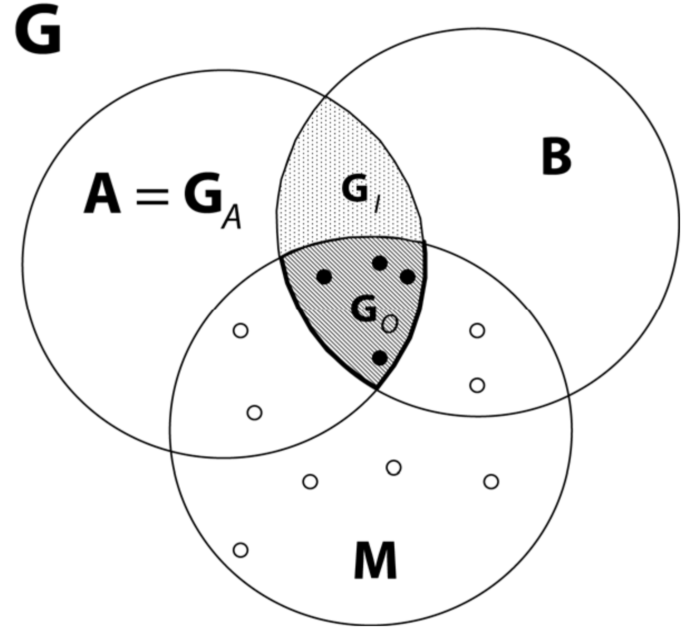
Occupied Distributional Area (G_O)

$A \cap B \cap M$



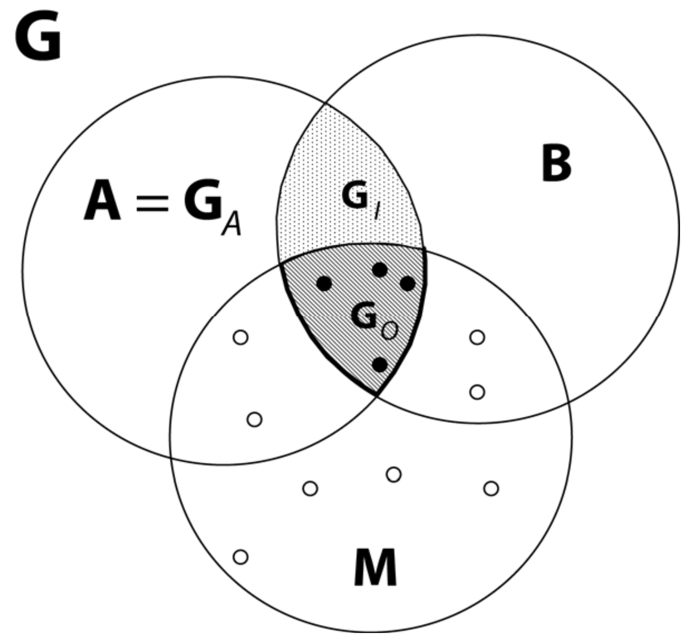
Invadible Distributional Area (G_I)

$A \cap B \cap M^c$



Potential Distributional Area (G_P)

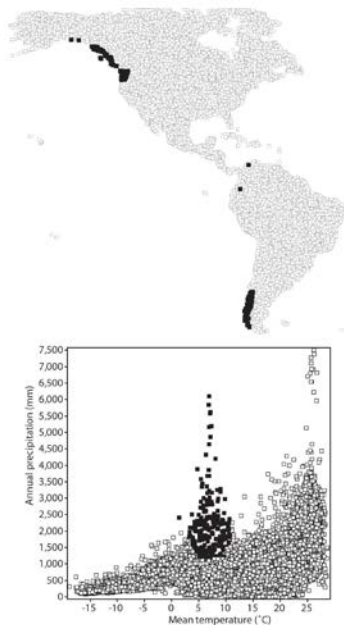
$$G_O \cup G_I = G_A \cap B$$



Switching from G-space (where we have been) to E-space

$\eta(G)$

the multivariate space of environmental dimensions associated with G



$\eta^{-1}(E')$

the geographic location(s) that correspond to a given environmental combination E’

Potential Environmental Niche(E_P)

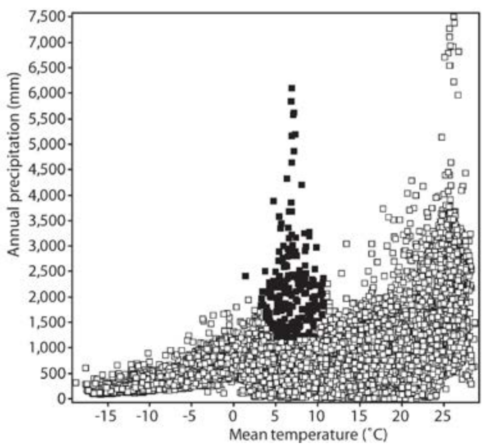
$$E_P=\eta(G_P)=\eta(G_O)\cup\eta(G_I)=E_O\cup E_I$$

The fundamentals of modeling the niche

N^{*}_{F} is the notation for the "existing fundamental niche" and is often what is modeled in correlative approaches

$$N_F \cap \eta(M) = N^{*}_{F}$$

N^{*}_{F} is thus a subset of the fundamental niche based on conditions represented on real-world landscapes



Fundamentals of Niche Modeling

How to model N_F^*

- Multiple methods
- Class of model: Mechanistic
  - Experimentally determine fundamental niche
    - Growth chamber experiments
    - Reciprocal transplants
  - Derive from litertature
  - Union that with environmental conditions for population or species
- Class of model: Correlative
  - The general focus here
  - Use data about species presences and env. covariates to try estimating niche

Mechanistic models: the Gold Standard?

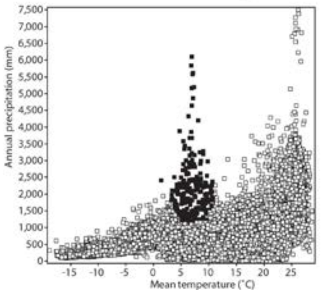
- Experimentally-determined physiological limits
  - Response curves
    - Physiological performance
    - Mortality
    - Reproductive success
  - Know full range of a species' fundamental niche

Mechanistic models: the "but"

- Non-model organisms a pain
  - Access
  - Facilities

Correlative niche modeling

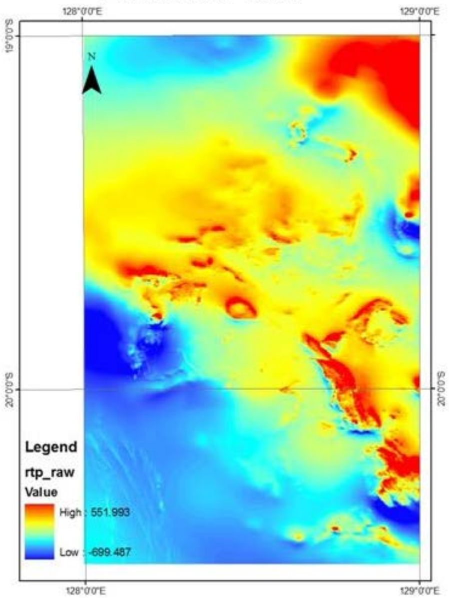
- Relies on species presence data…
- Plus a set of **environmental covariates**
- and the translation from G-space to E-space
- …and then **back** to G-space, which can involve:
  - Interpolation
  - Extrapolation (or transfer)



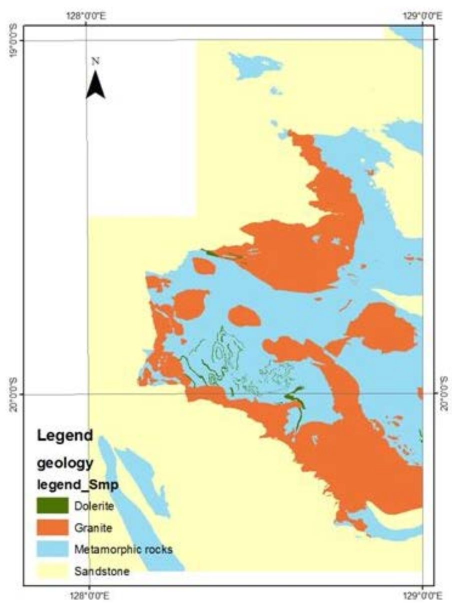
- The goal is typically **predictive** for areas not sampled
- The output is often discussed as “habitat suitability"

About those environmental covariates in niche modeling?

Continuous data



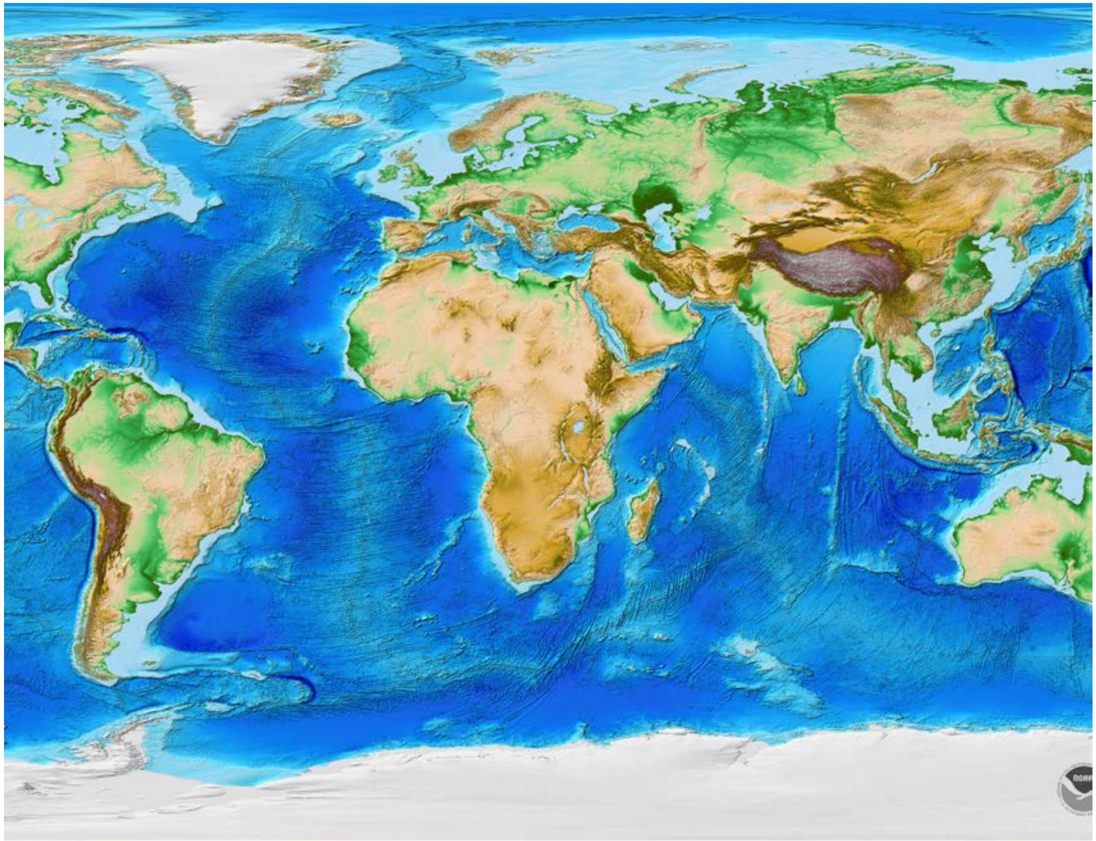
Discreet data



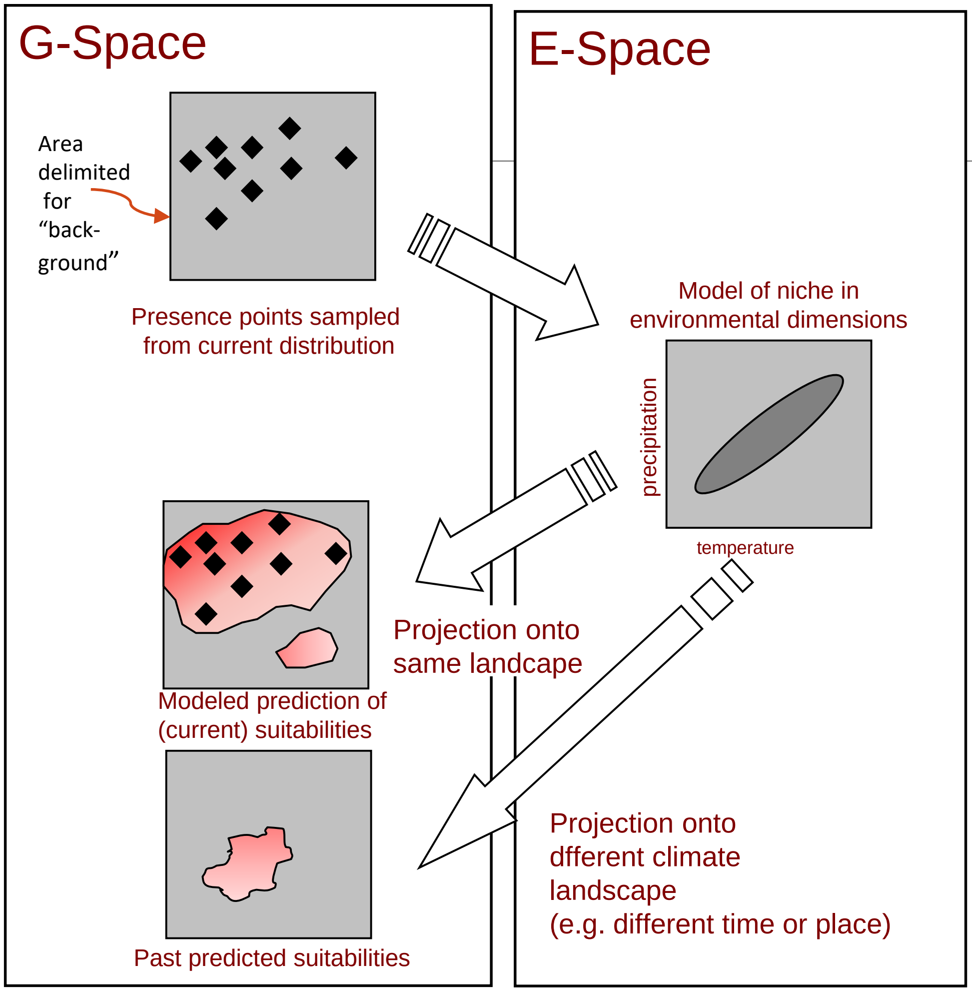
Types of data

- Anything summarized by a raster*
- Continuous
  - e.g. temperature
- Categorical
  - e.g. land cover types

Common data sources



- Altitude/bathymetry
- Slope
- Aspect
- Soil characteristics
- Climate

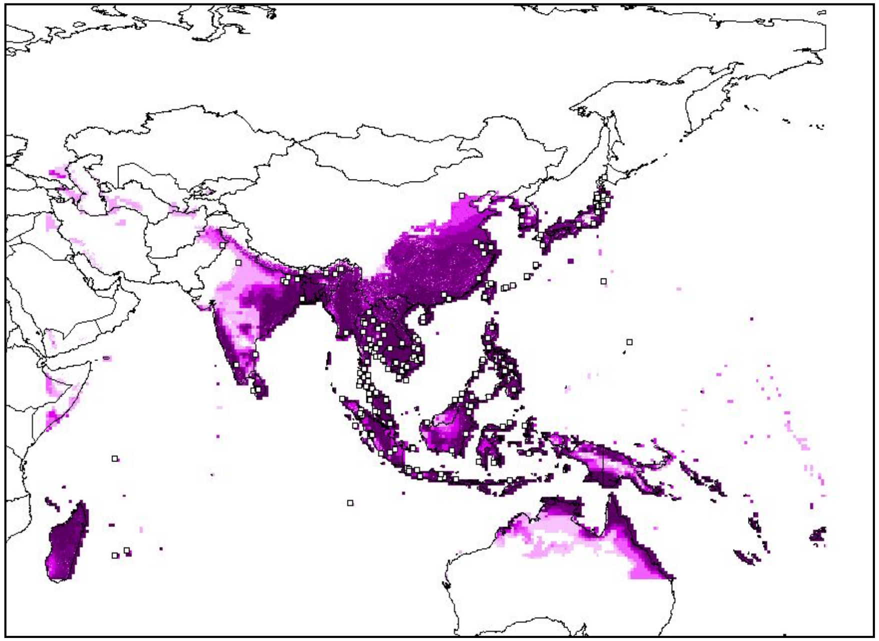


Moving from G-space to E-space and transferring models to new environmental landscapes

Asian tiger mosquito

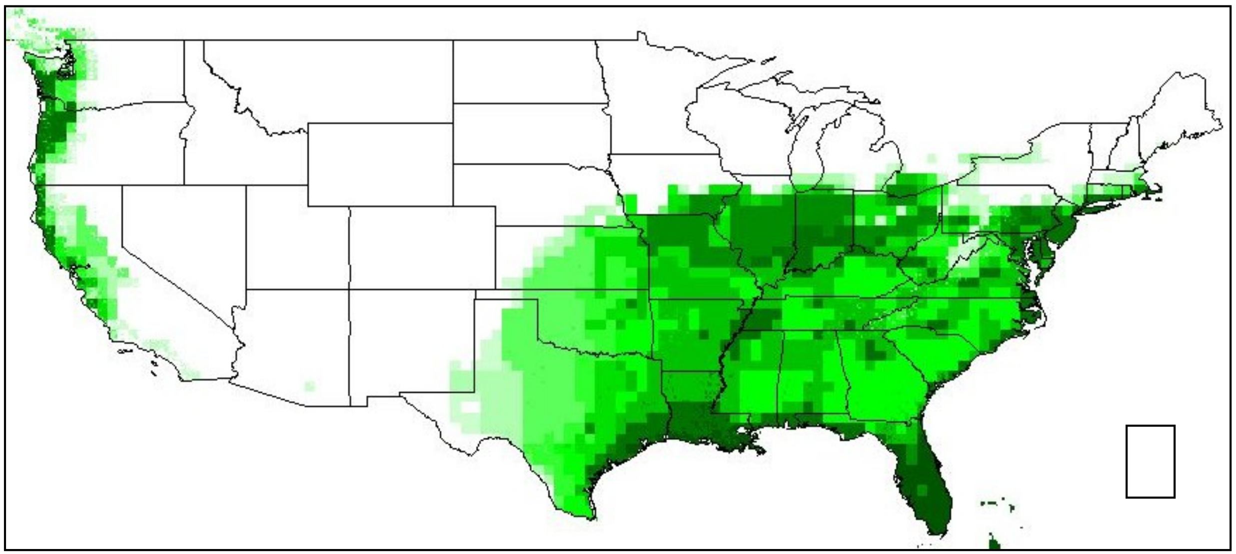
- Native to SE Asia, India
- First found in Houston in 1985
- Has spread throughout SE US and into the NE (to Maine)
- Introduced to Hawai'I before '85

Aedes albopictus



Model of tiger mosquito known to carry dengue in its native range

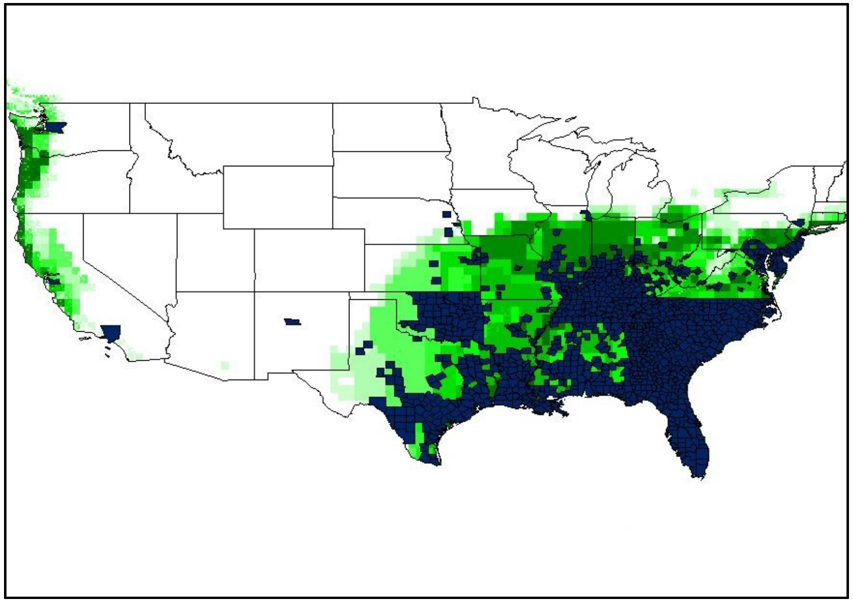
Predicted *Aedes albopictus* in the USA based on niche modeling

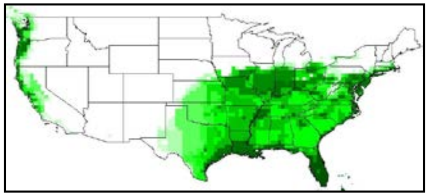


This model can be transferred in space to predict its invadable range in other areas

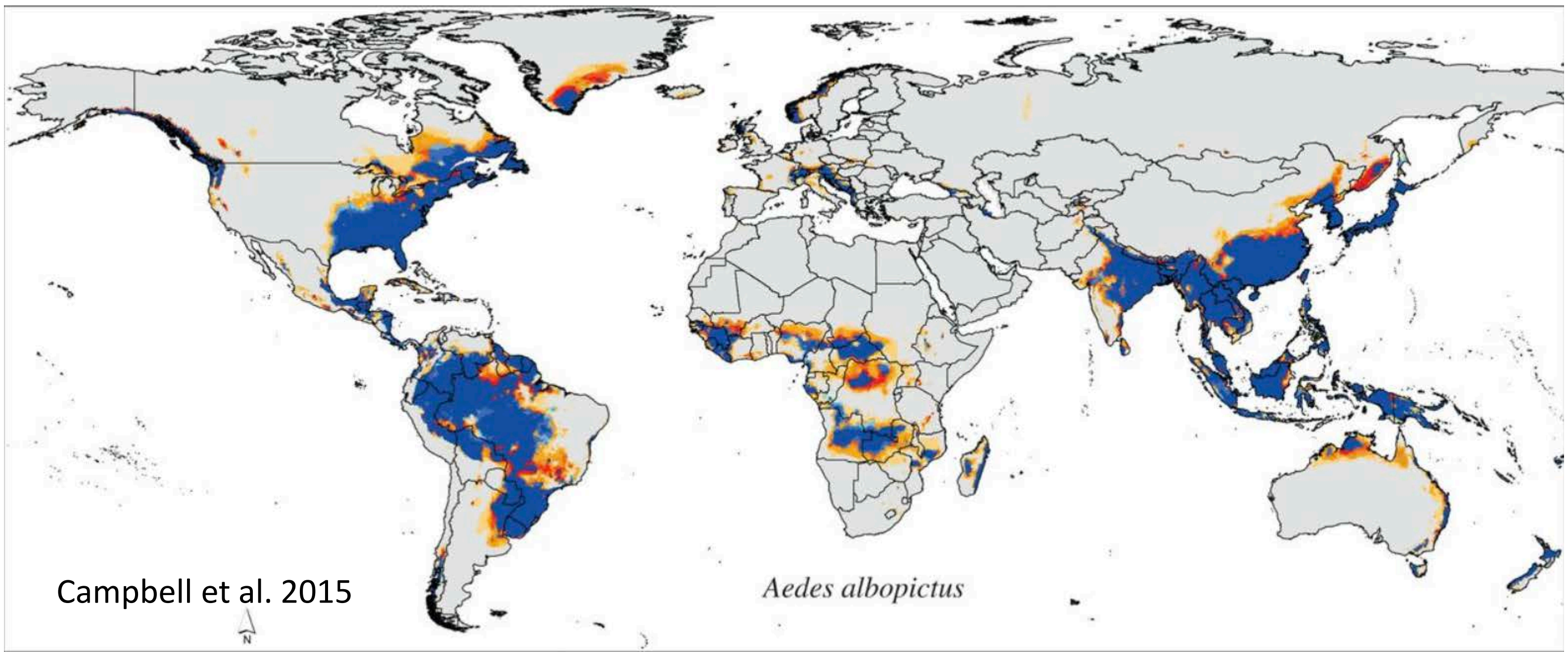
Aedes albopictus US Invasion

Actual spread of
Aedes albopictus
in USA



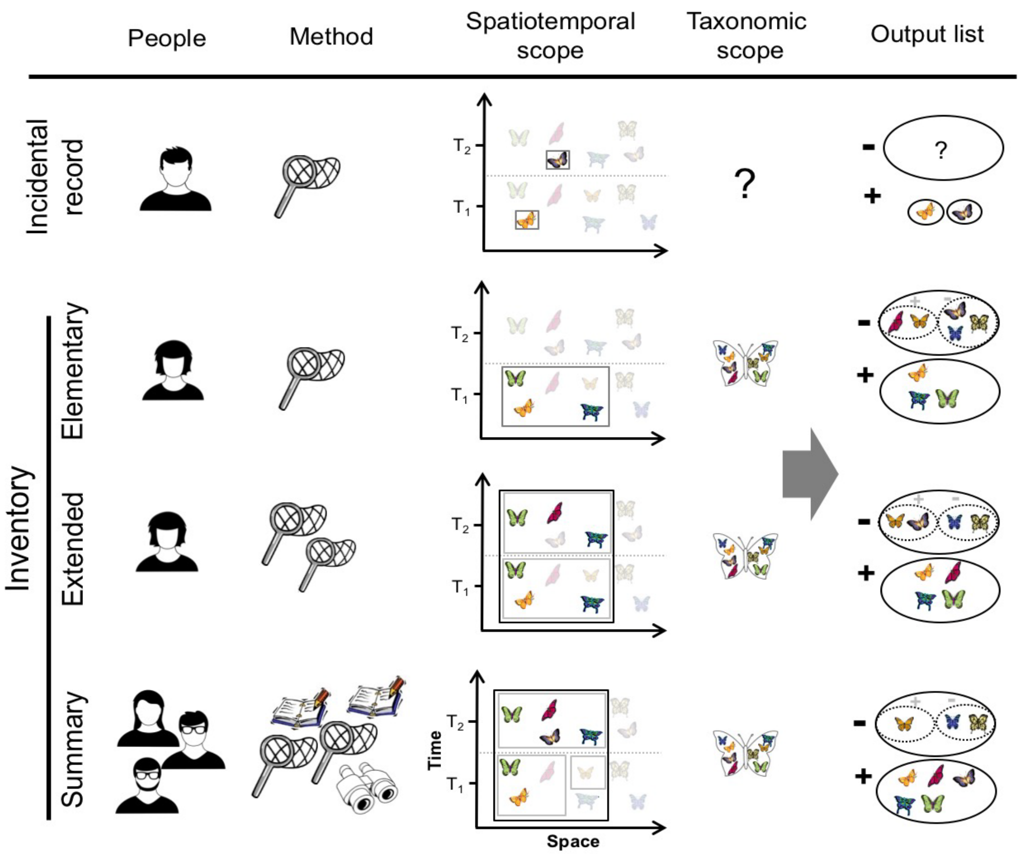


And modeling distributions under future climate conditions (A1B model)



About absences for a moment

- Absences are **probabilistic (unlike presences)**
- Observations that yield no evidence of a taxon as **non-detections**
- **There are whole classes of modeling** that deal with the issues of detection and how to simultaneously assess detection probabilities and site occupancy
- These are **occupancy models**, and they do connect to the larger world of Species Distribution Models (SDMs)



How do we assemble absence data

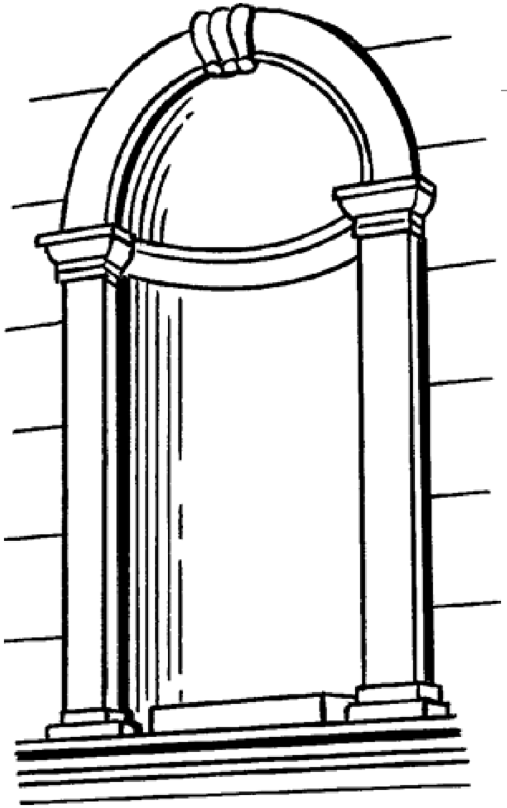
- **Incidence reporting** only generates presences
- IF you do an **inventory** and it is **complete** over an intended spatial and temporal scope, you can document absence
- There are kinds of surveys performed over many scales

Why care about absence?

- If we want to understand **realized distributions** (not potential ones) and **how those change**, absence is critical.

- Absences can be used at multiple points in modeling process.

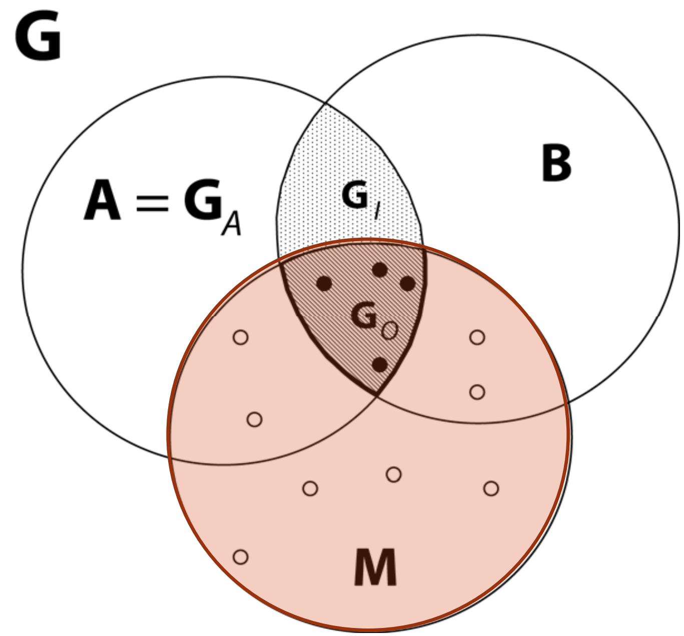
- We will focus for the rest of the time here on presence-only modeling but the key thing is absence data is critical depending on what you intend to model



Presence-Only Modeling

Two strategies

- 1) Pseudoabsence
- 2) Background
- Virtually identical at all but narrowest spatial scales
- The choice of background is *essential* – the model **training region** REALLY matters!



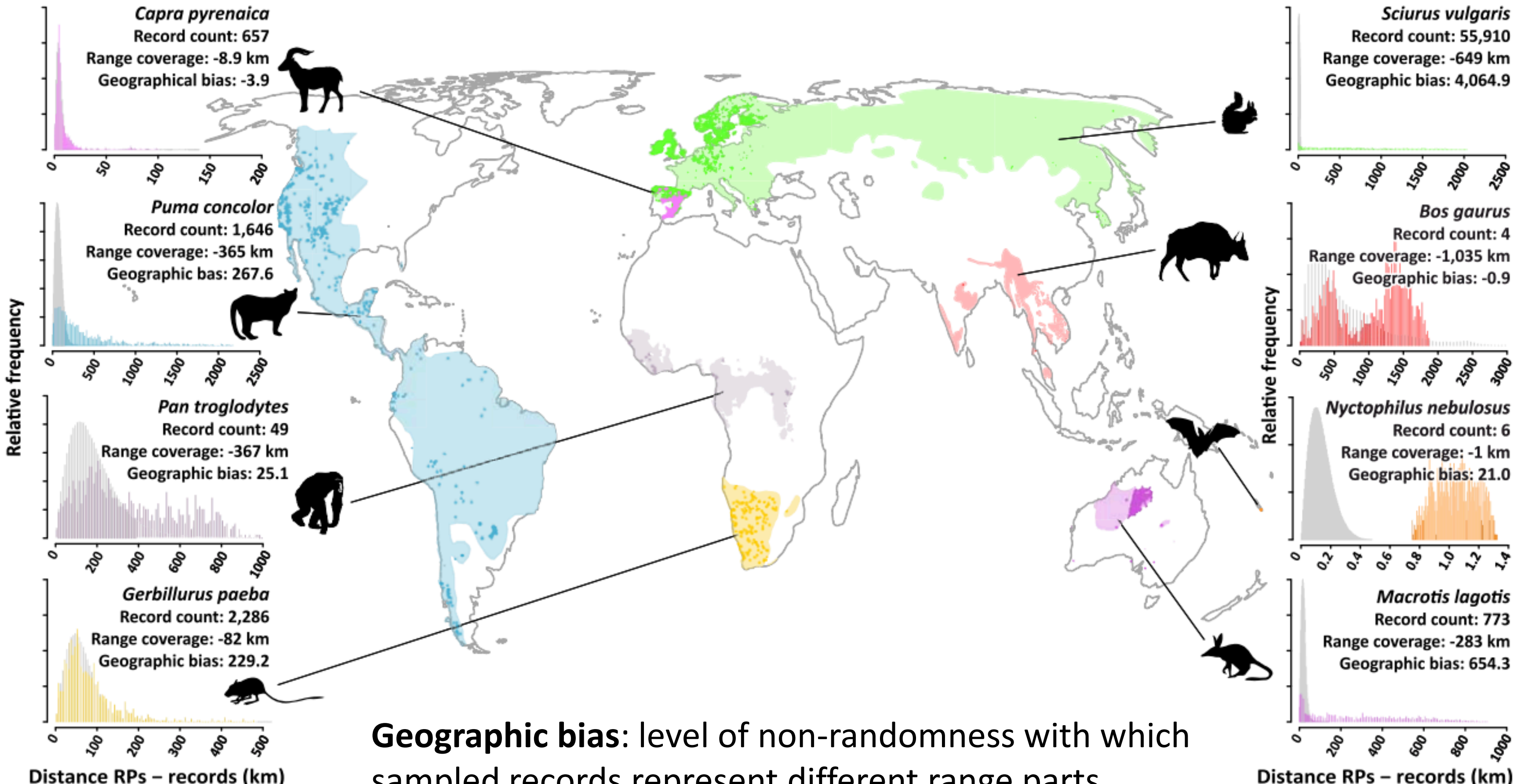
Point presence bias and back to G and E

- Spatial bias in **point presences** is a real problem with ecological niche modeling.

- If you draw a biased sample of the actual G_O, it often (but not always means) you have a biased E_O. This is bad.

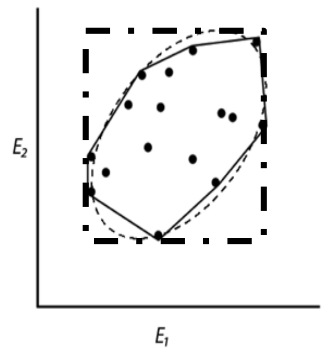
- There are approaches to try to fit sampling bias in niche modeling

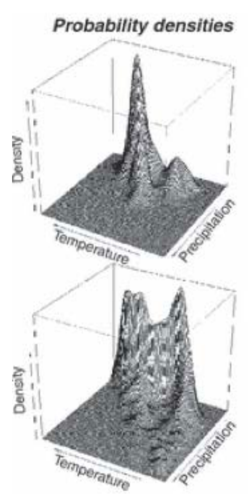
These biases can sometimes be quantified



Geographic bias: level of non-randomness with which sampled records represent different range parts

Meyer et al. 2015

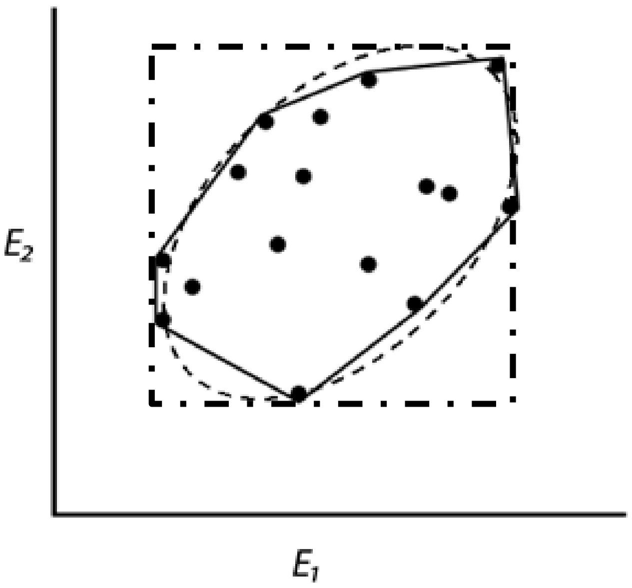




Types of presence-only algorithms

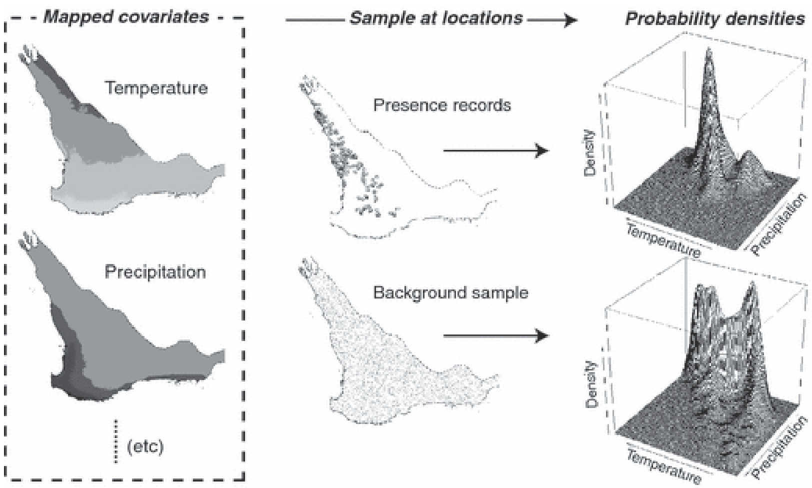
Surface range envelope

- Infer rectilinear envelope
- Or other shape (e.g a hull)
  - Maximum and minimum values
  - Very simple
  - e.g. BioClim
- Everything "inside" the envelope is "suitable" for the taxon in question



Maximum Entropy Approaches

- Maxent – key tool
- Performs relatively well across variety of modeling scenarios
- In geographic space
  - Maximizes dispersedness
- In covariate space
  - Minimizes dispersedness
- Iterative approach



Elith et al., 2010

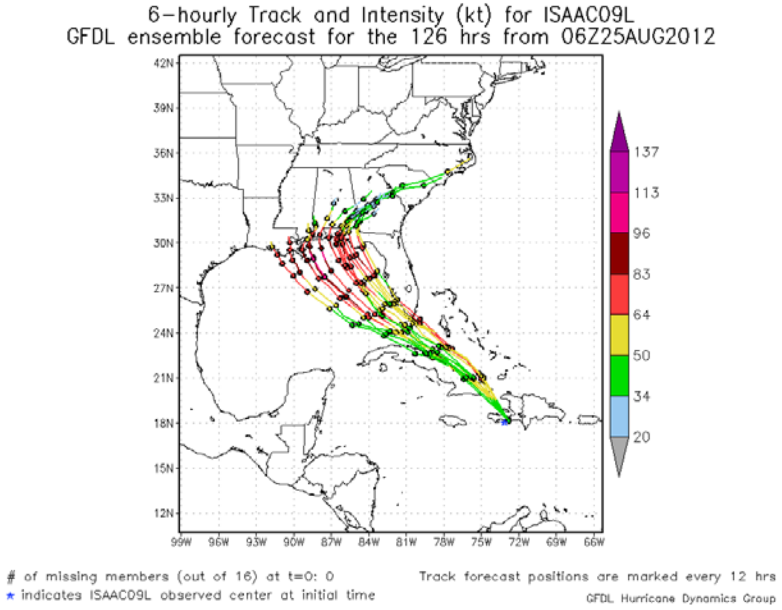
More on Maxent

- Starts with a fully uniform distribution over all grid cells
- Conducts optimization routine to maximize “gain”
  - Likelihood statistic maximizing the probability of the presences
    - Given input data and in relation to the background data
  - Gain will asymptote leading to final probability distribution
- Distribution becomes basis for fitted predictor variable coefficients
  - Coefficients are used to assess probability of presence

Ensemble models

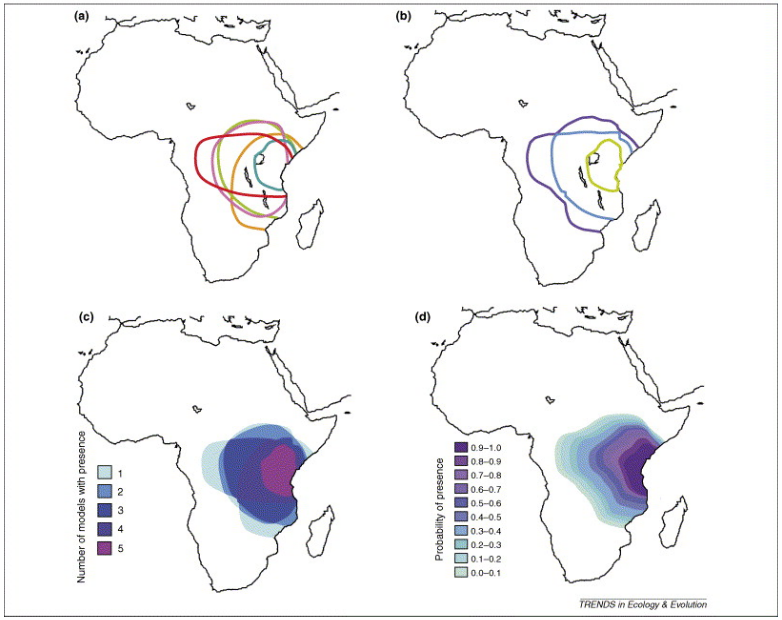
Ensemble approach

- Comes from climatology, weather forecasting
- Biomod2 is a toolkit for this approach in ENMs/SDMs



Biomod2

- Combined model predictions
  - Many ways of combining
  - Get the best of all model types
  - Get the worst of all model types too?



Araujo & New, 2007

TIME CHECK

Evaluating models

Validation data

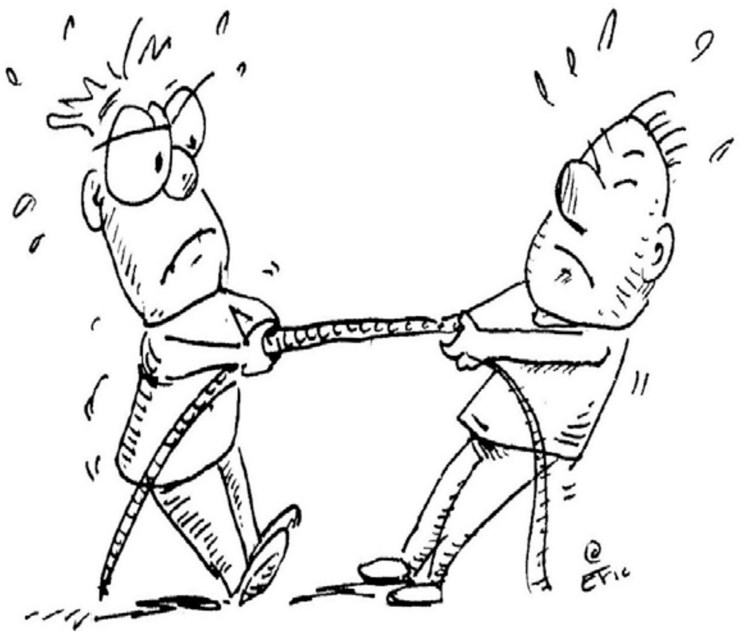
- Independent (or external) validation data (Test file)
- Randomized subsetting (Subsets) (internal)
  - Partition data into two subsets, one for training and one for testing
  - 75-25% training/testing split is common
  - Random subsets are partitioned many times (number of models generated)
- K-fold partitioning (Crossvalidate) (internal)
  - Occurrence data set is partitioned into k subsets
  - Each partition is used for testing once
  - All non-test partitions are combined for model calibration
- Background points
  - Maxent treats them as pseudoabsences for the purposes of post-calibration evaluation

Classification

Sensitivity: True positive rate

Specificity: True negative rate

Would you rather throw out milk that was fine, or drink milk that had spoiled?



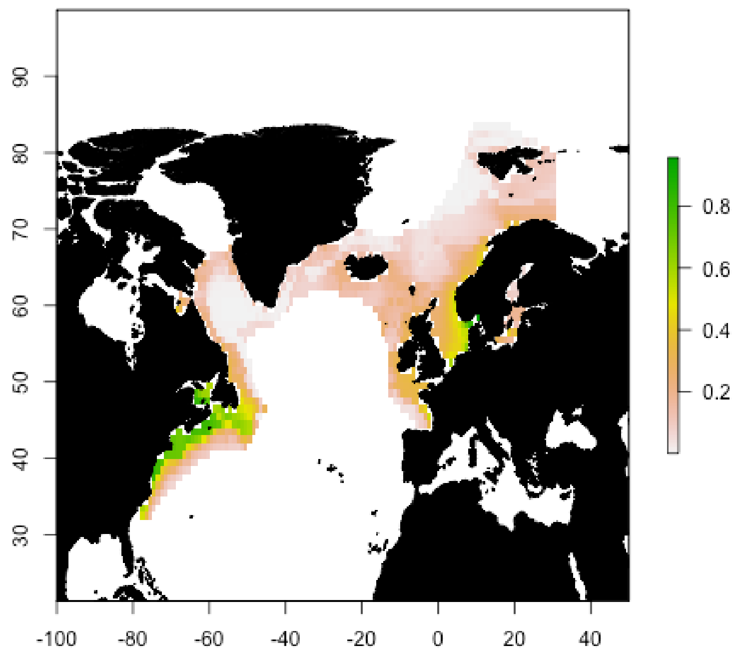
Kinds of metrics

- Threshold-dependent metrics
  - Sensitivity, specificity, Cohen's kappa, and the true skill statistic (TSS)
  - Require binary maps (suitable/unsuitable)
    - User-defined
- Threshold-independent* metrics
  - AUC: Area under the receiver-operator characteristic curve

Thresholding a niche model

Choosing what score determines a presence versus an absence

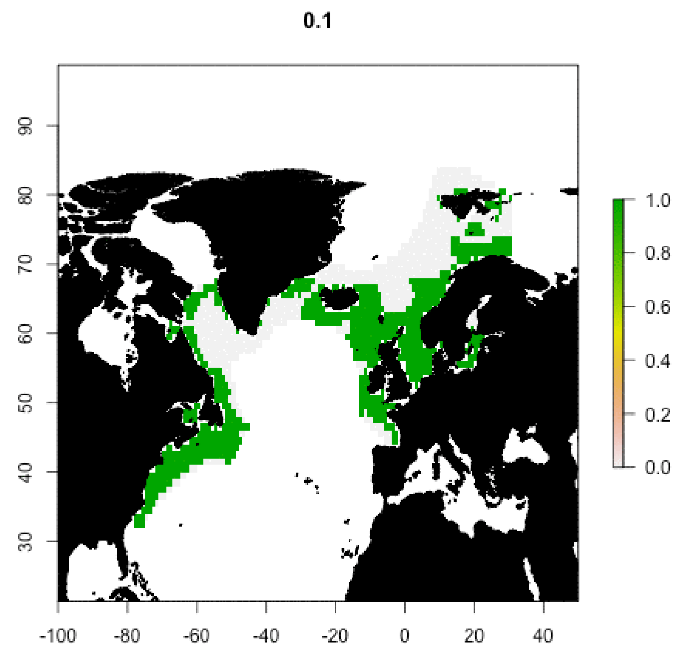
Raw suitability surface



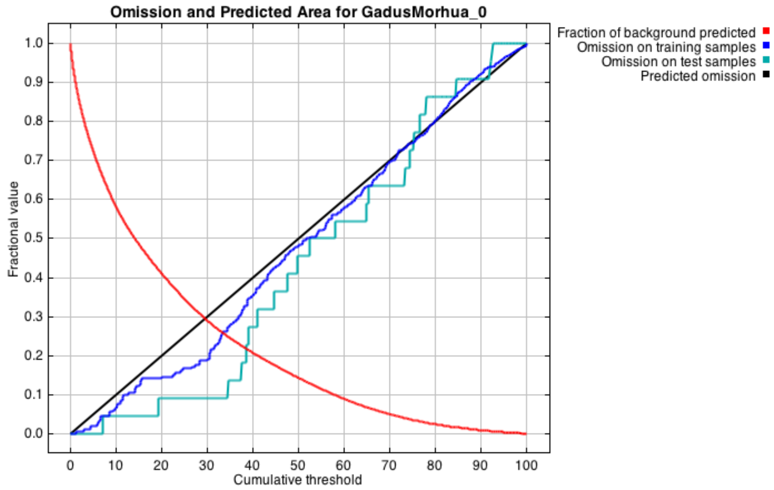
Thresholding a niche model

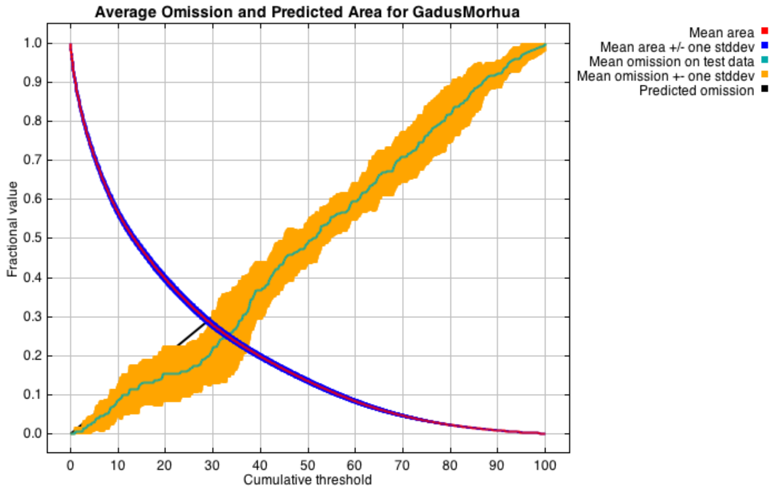
Choosing what score determines a presence versus an absence

Thresholds



Thresholds: a tradeoff





You might notice that as culmulative threshold goes to 100 (all areas are suitable – high commission errors), omission error goes to 0. Same in converse.

At every other threshold, there is a trade-off, but you'll notice the lines **do** cross.

Threshold-dependent evaluations

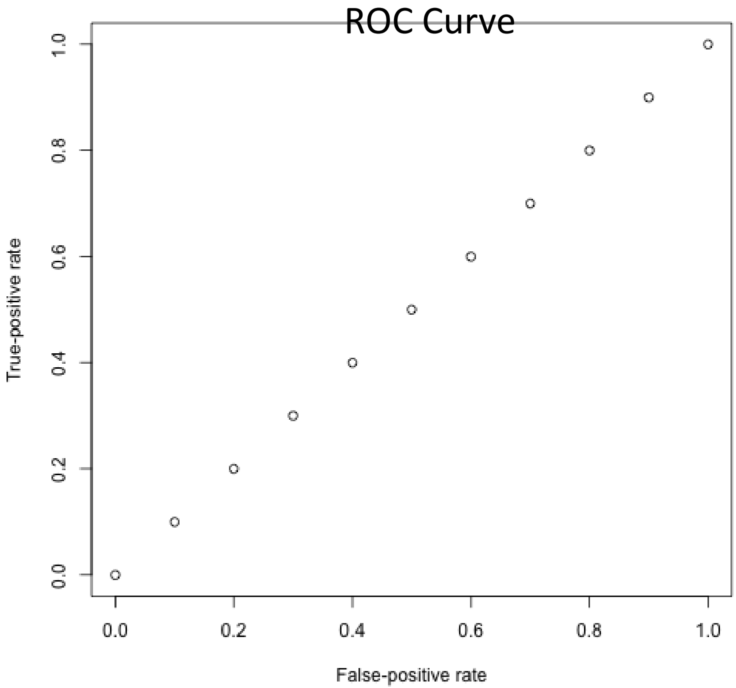
- Cohen’s Kappa
- True Skill Score (TSS)
  - Corrects for prevalence
  - Threshold-dependent
    - A = True presence
    - B = False presence
    - C = False absence
    - D = True absence

$$\mathrm{TSS} = \frac{ad - bc}{(a + c)(b + d)} = \mathrm{Sensitivity} + \mathrm{Specificity} - 1$$

Area Under the Receiver Operating Curve

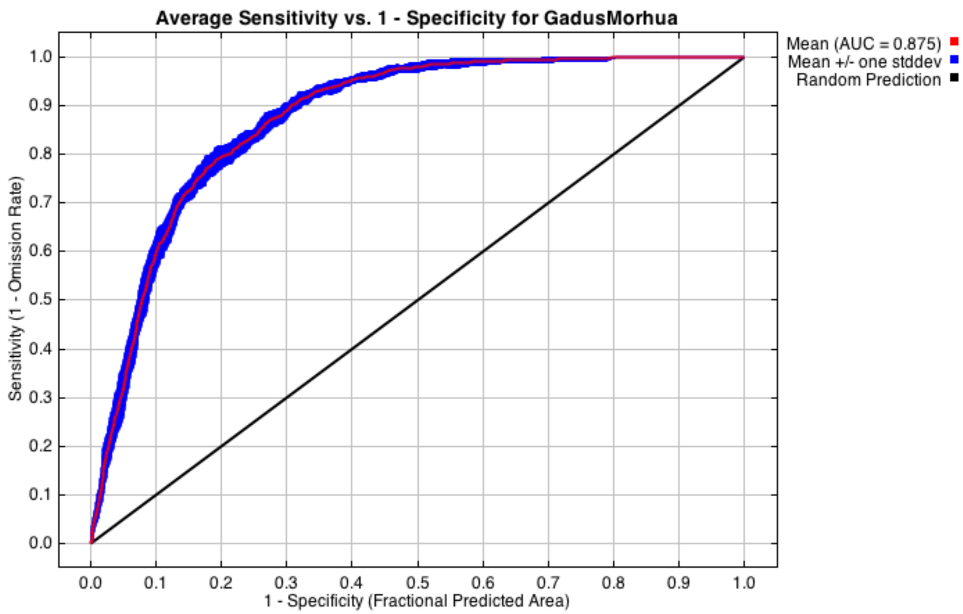
- The first statistic you see reported by Maxent
- AUC: Probability randomly selected presence point has higher predicted suitability than randomly-selected background point
- For **every possible threshold** between 0 and 1:
  - X-axis: false-positive rate
    - (1 - Specificity, aka number of correctly predicted absences*)
  - Y-axis: true-positive rate
    - (Sensitivity, aka number of correctly predicted presences)

* But what are "absences" in a presence-only model?



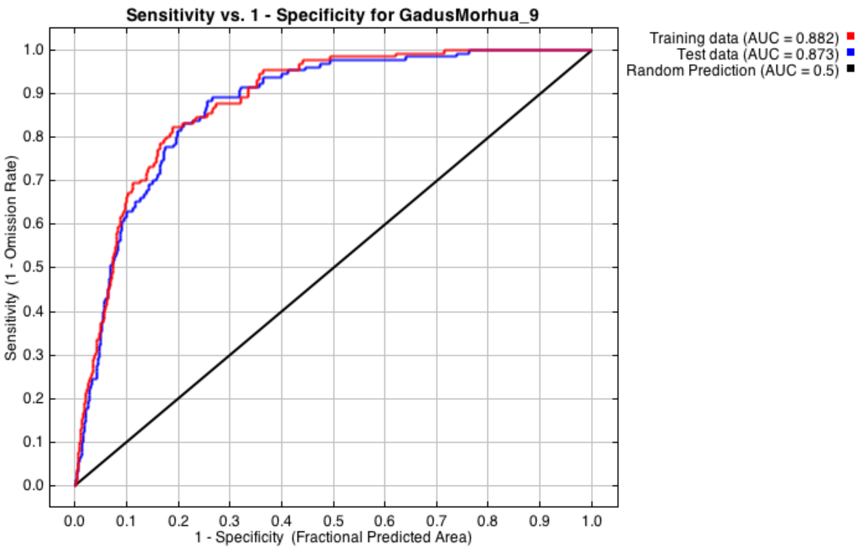
Area Under the Receiver Operating Curve

- Modified for ENMs
  - X-axis based on proportion of area predicted as present
- Models that *fit the data well* have an L shape
  - AUC approaches 1
  - BUT…



Area Under the Reciever Operating Curve

- There's more than one AUC to evaluate
- Found in individual model run reports
  - AUC_{Train} = how well model fits training data
    - Favors overfit models
  - AUC_{Test} = how well model fits testing data
    - Independent of overfitting
    - Favors under-fit models
  - Minimum difference between training and test AUC
    - Minimize overfitting



AUC: Caveat 1

- We don’t have true absences
  - Instead, AUC calculated by Maxent discriminates between presences and background points.

- Background points are treated as pseudo-absences only for evaluation procedure (not for the model fitting)

- This isn’t great if your background samples are a poor representation of absences
  - This is another reason M is important!!

AUC Caveat 2

- AUC scores are often high in ecological niche models. YAY!
- But a lot of time you are fitting spatial autocorrelation and your enthusiasm is misplaced
- Because samples are not really independent
- There are some ways to deal with this
  - partial ROC (Peterson, Papes, Soberon 2008)

Other ways to measure model performance

- AIC, AICc, BIC
  - Balances model complexity with goodness-of-fit
- More fully explore parameter-space
  - e.g. ENMEval