Species Distribution Modeling: Crash Course Part 1

Nicholas Galle, Rohr Lab




Overview

What drives a species distribution?

The Niche Concept

Broad SDM workflow

SDM Assumptions & Violations

What questions can we ask?

What are Species Distribution Models?

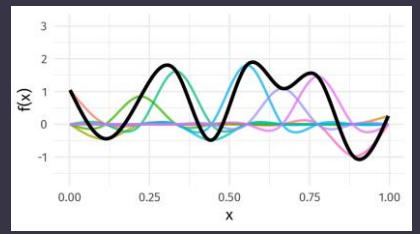
Species Distribution Models (SDMs): "Numerical tools that combine observations of species occurrence or abundance with environmental estimates," in order to project what areas in space are suitable for species persistence.

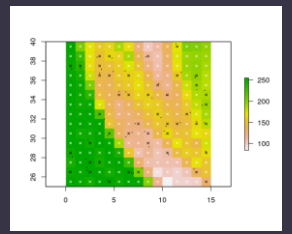
- Also referred to as ecological niche models (ENMs) or habitat suitability models

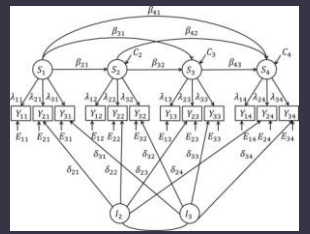
Correlative SDMs: Produce function that encapsulates relationship between environmental variables and species presence

Pro: So many models, minimal knowledge of physiology req.

Con: So many models, use of physiological data minimalized



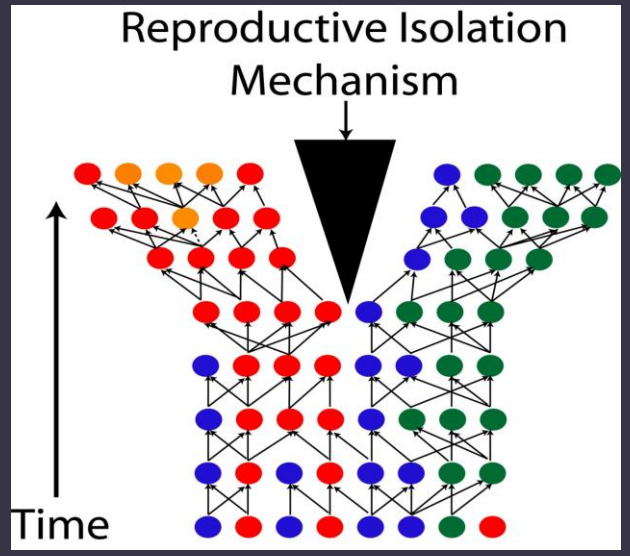


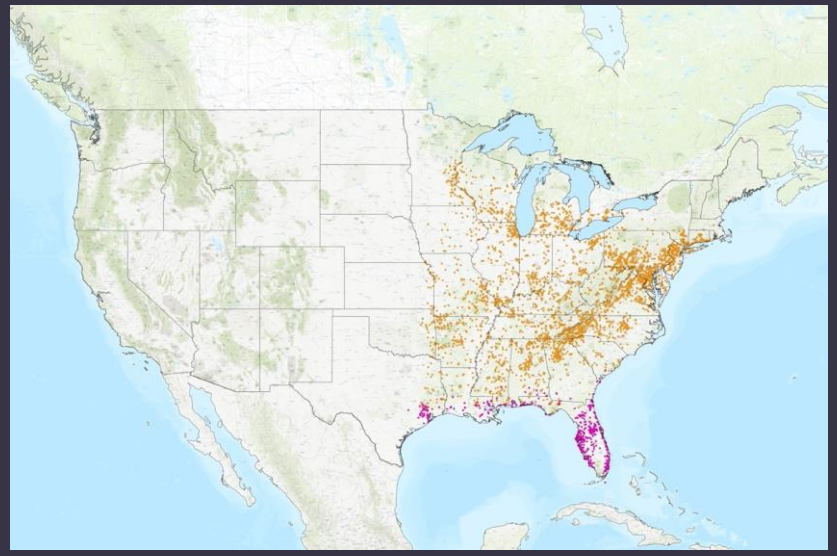


What Drives a Species Distribution?

To persist in a particular place a species must

1. Have *evolved* to survive in that place (or reach a suitable place through dispersal)
2. Tolerate the *biotic* interactions in that place
3. Be able to maintain population growth in the *abiotic* conditions of that place





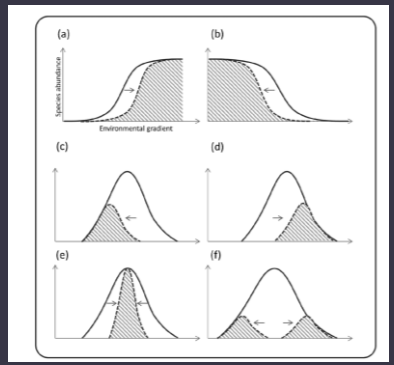
Conditions for living in a given place

Evolved (or dispersed) there

- Dispersal mechanisms available
- Via **allopatry** (due to geographic barrier)
  - River, mountain range, ocean, desert, drainage basin
- Via **sympatry** (other than geographic barrier)
  - Niche partitioning, chromosomal incompatibility, temporal reproductive isolation

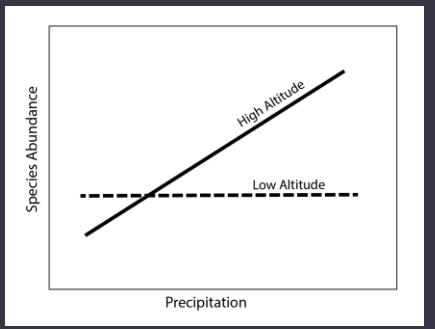
Biotic tolerance

- Trophic interactions
  - Predator / Prey
- Symbiotic interactions
  - Pollinator / Pollinatee
- Competitive interactions
  - Competitive exclusion



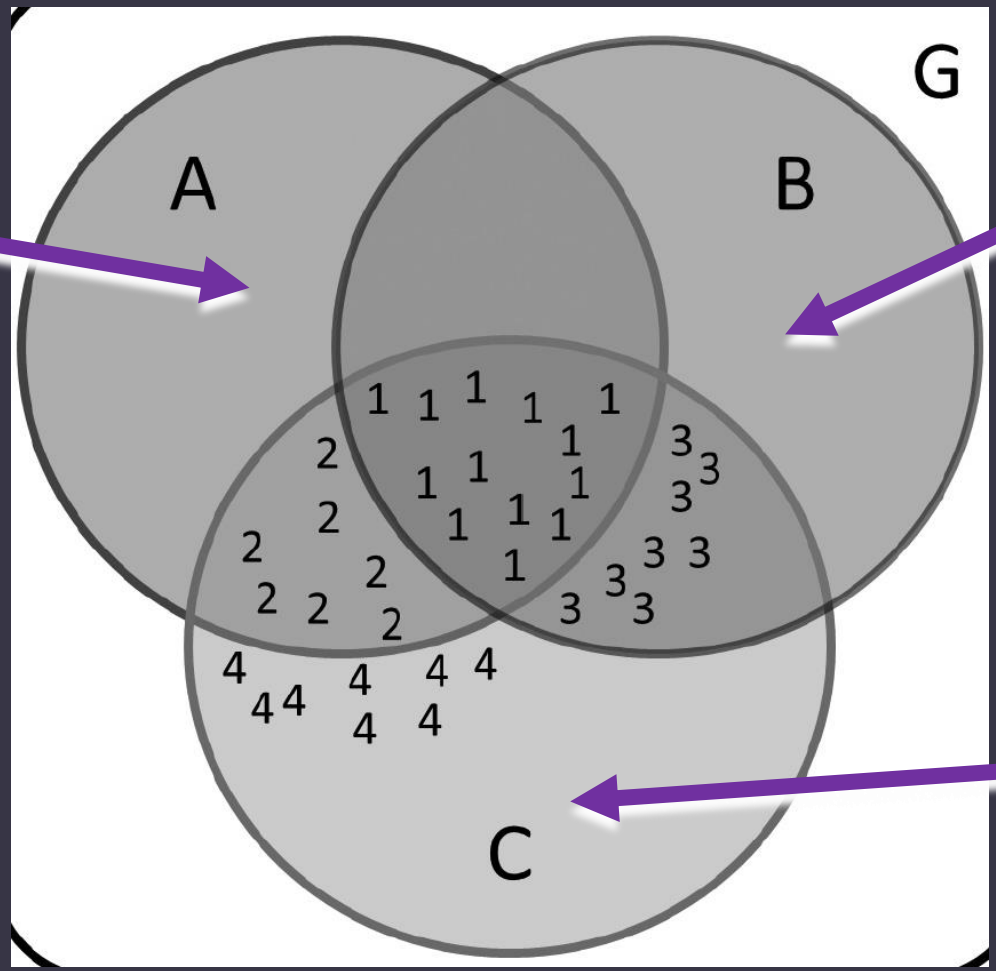
Abiotic tolerance

- Temperature
- Precipitation
- Radiation
- Seasonality
- Soil Minerals
- Etc



Realized Niche: The fundamental niche once it has been constrained by dispersal capability, biotic filtering, and the range of environmental factors that are ACTUALLY available to species.

suitable abiotic conditions

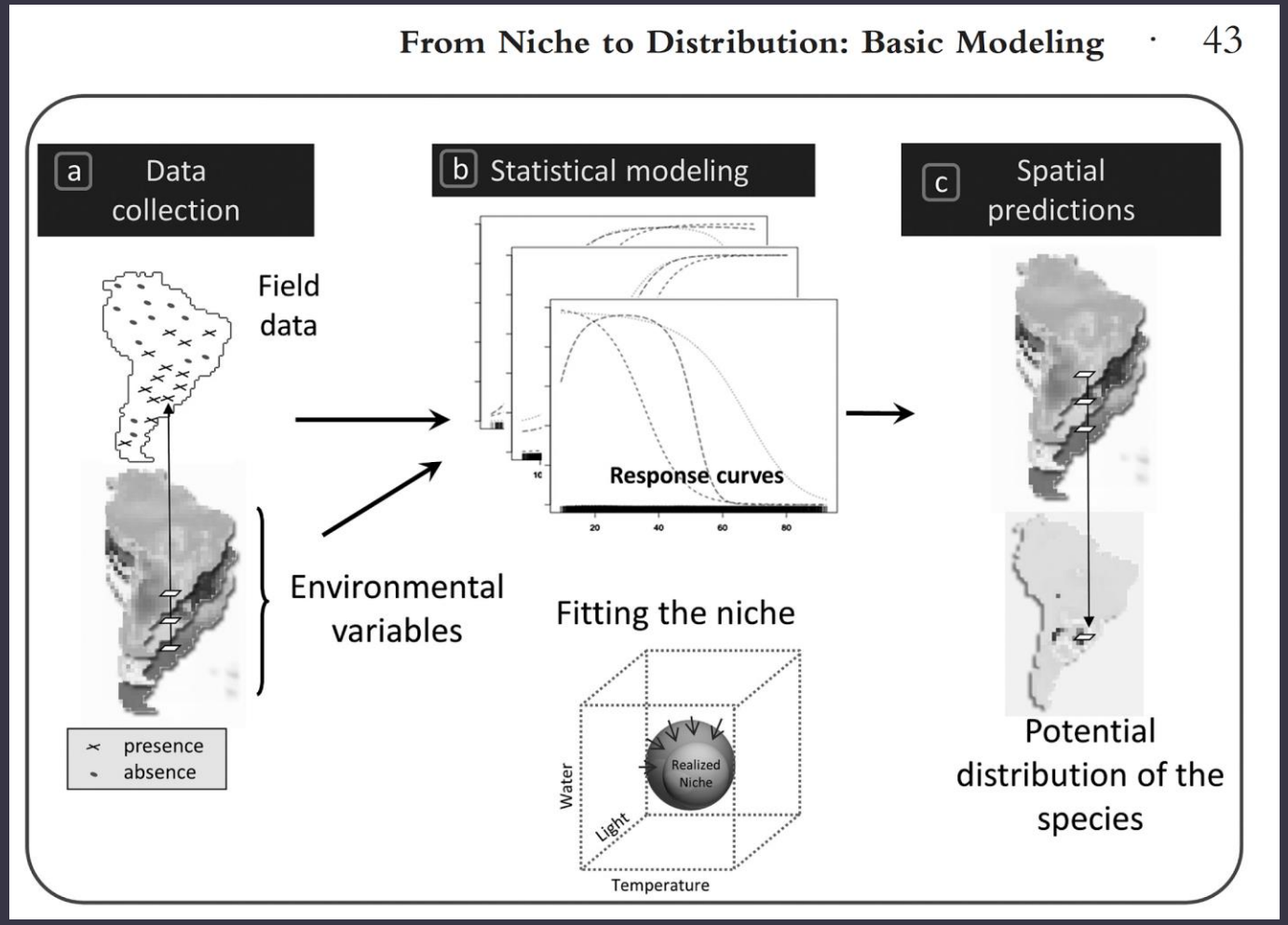


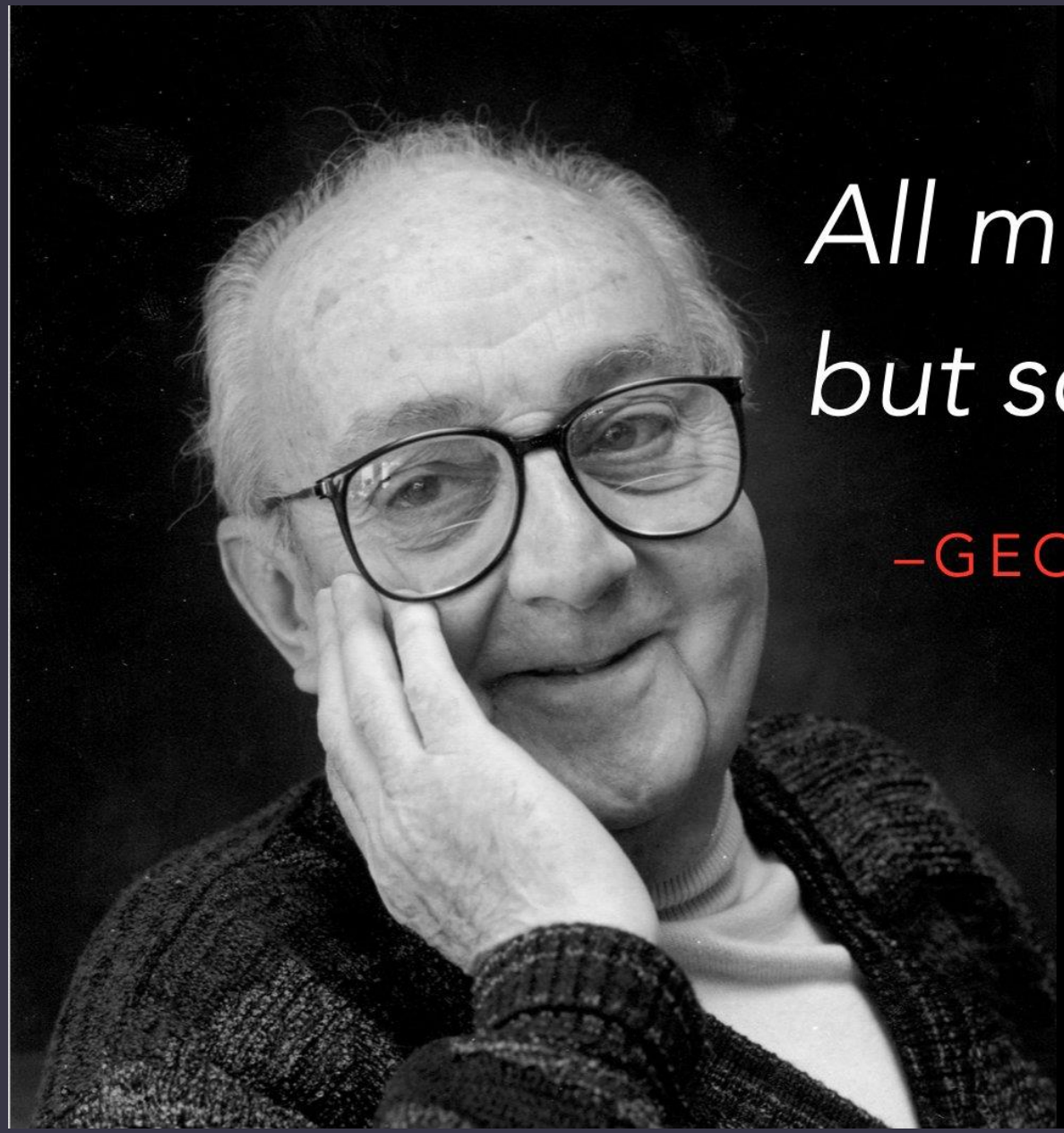
suitable biotic conditions

dispersal range

SDM: A Coarse Overview

1. Collect occurrence and predictor data
2. Fit/tune models
3. Evaluate model fit
4. Project model fit into space





All models are wrong, but some are useful.

–GEORGE BOX, UW-MADISON

SDM: A Finer Overview

Data: Occurrence/Abundance, Background Points, Predictor Variables

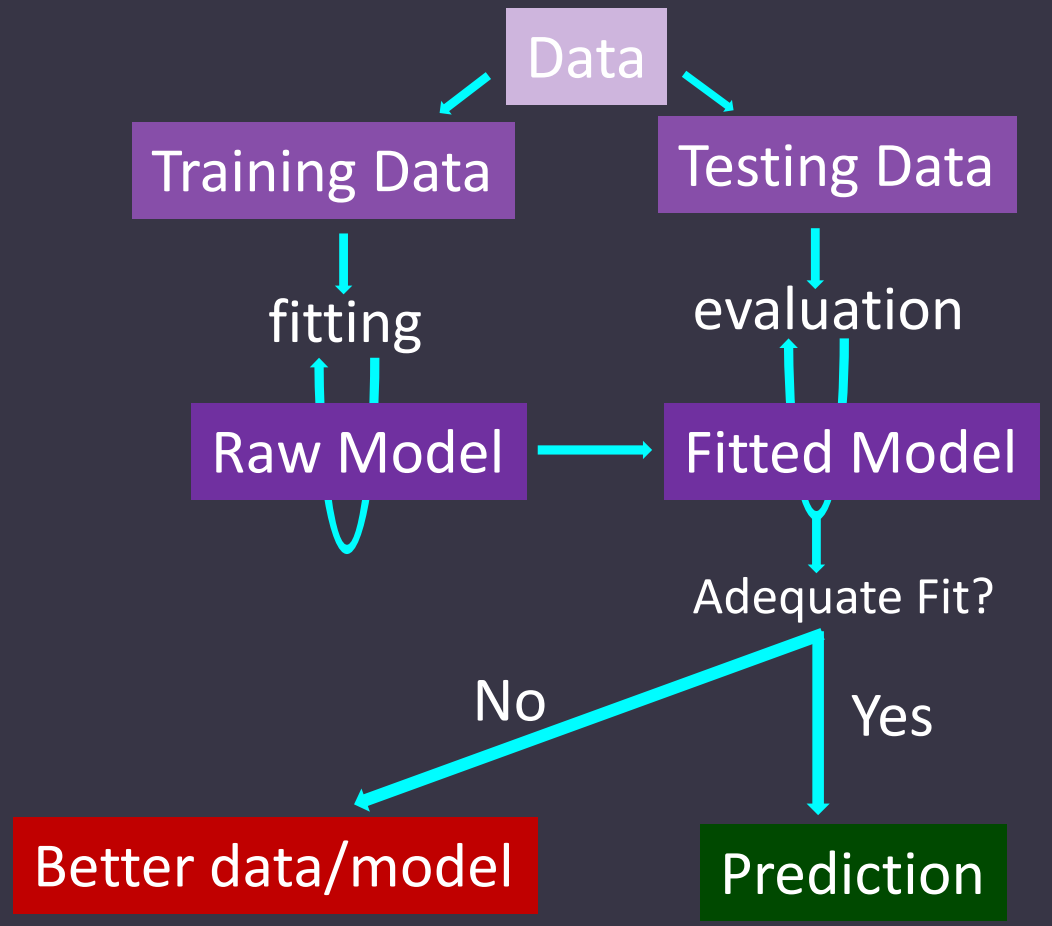
Train/Test Set: Normally 80/20 or 75/25

Fitting: Model tuning (RM, Feature Class, LR, TC)

Evaluation: AUC-ROC, TSS, Max Kappa, Boyce Index

Good Fit?: Proceed to spatial projection

Bad Fit?: Go back to step 1 and re-evaluate



Data: Occurrences

Where is a species known to occur?

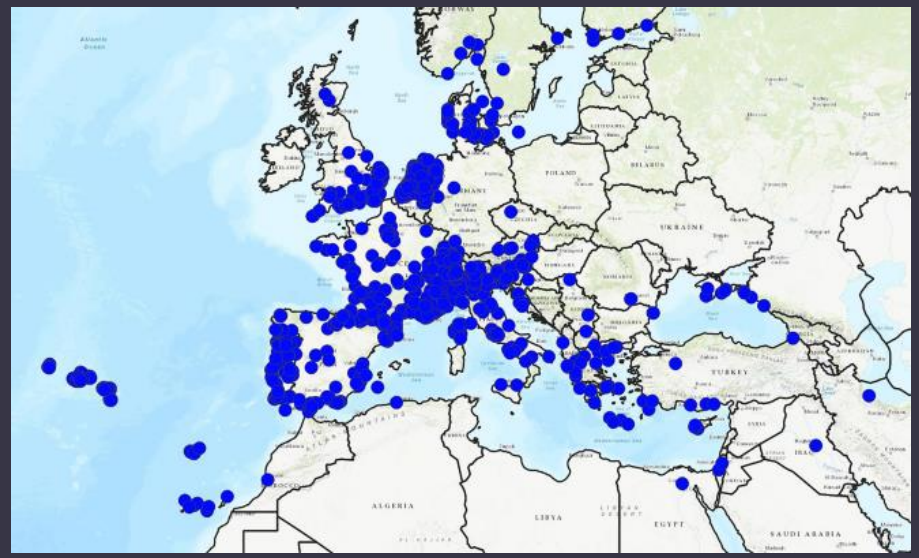
Surveys, collections, databases, literature

Different than abundance data
- # individuals at a site, not commonly available

Records need to be cleaned
- Flip-flop Lat/Lon, terrestrial species in ocean, lat/lon corresponds to museum across the plant from known range, etc.

Some methods require absence data
- More appropriately: pseudo-absence data or background data

Longitude/Latitude coordinates of individuals (*Chrysodeixis chalcites* (Esper))



Data: Background/Pseudoabsence Points

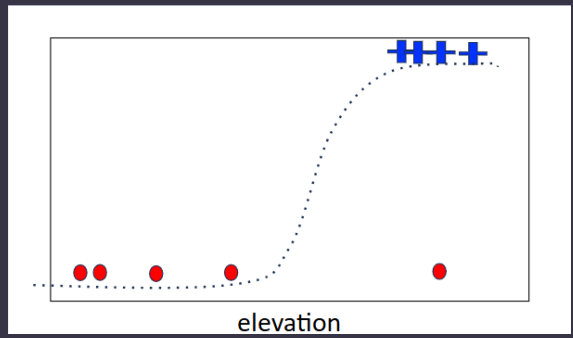
Where is a species not known to occur?

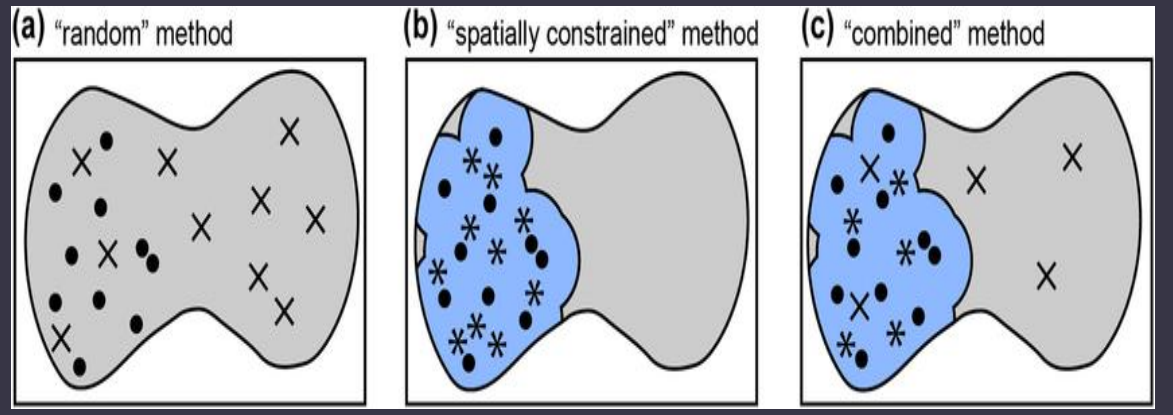
Absence/Pseudoabsence/Background: where a species **is *not* known to occur**

- Most repositories do not include information about where species are not found
- Many researchers don't report it anyways

Why do we need this?

1) Provides contrast to presence/env. correlation (samples the variation of the landscape
2) Allows for model evaluation: how well does the model predict presence data compared to random data?





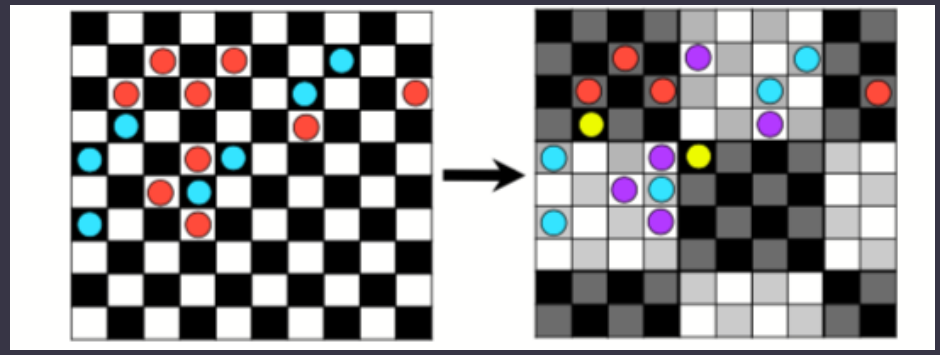
* This topic could use its own presentation

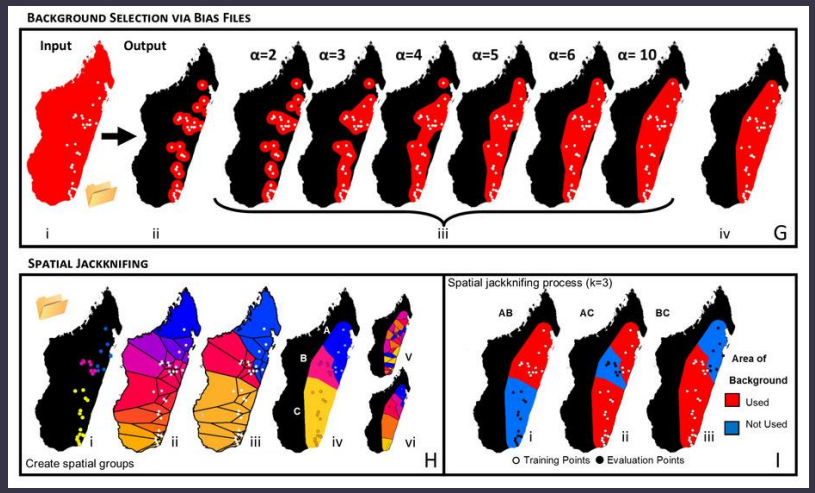
Data: Train / Test Partition

Subset into training and testing

- Random subset (not based on space)
- Checkerboard1/2 sampling (ensures a more even spatial distribution between groups)
- Cluster-based

More in-depth as part of another talk





Data: Environmental

What do we know about where species occur?

Weather and Climate

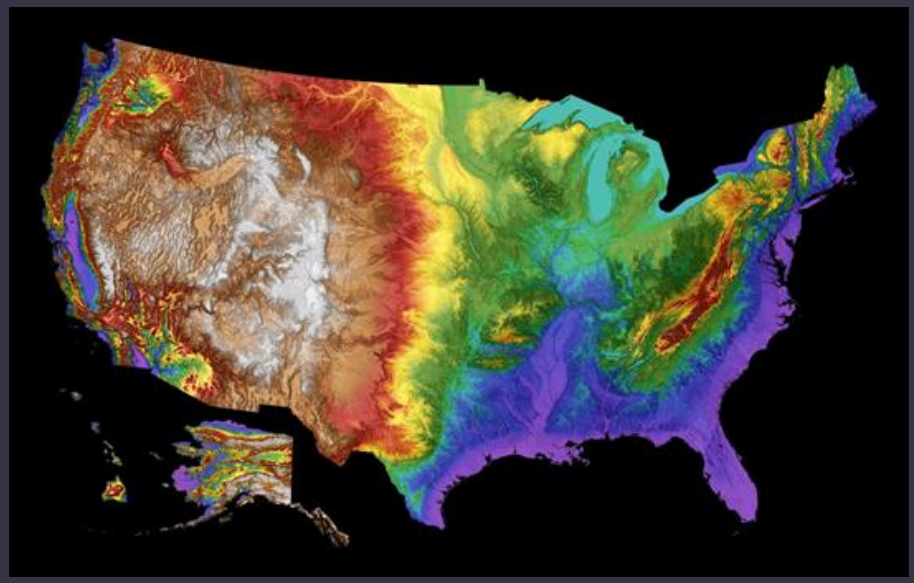
Solar Radiation

Soil Type

Vegetative Cover

Human Disturbance

Elevation/Slope/Aspect



Data: Environmental

Generally stored as raster maps

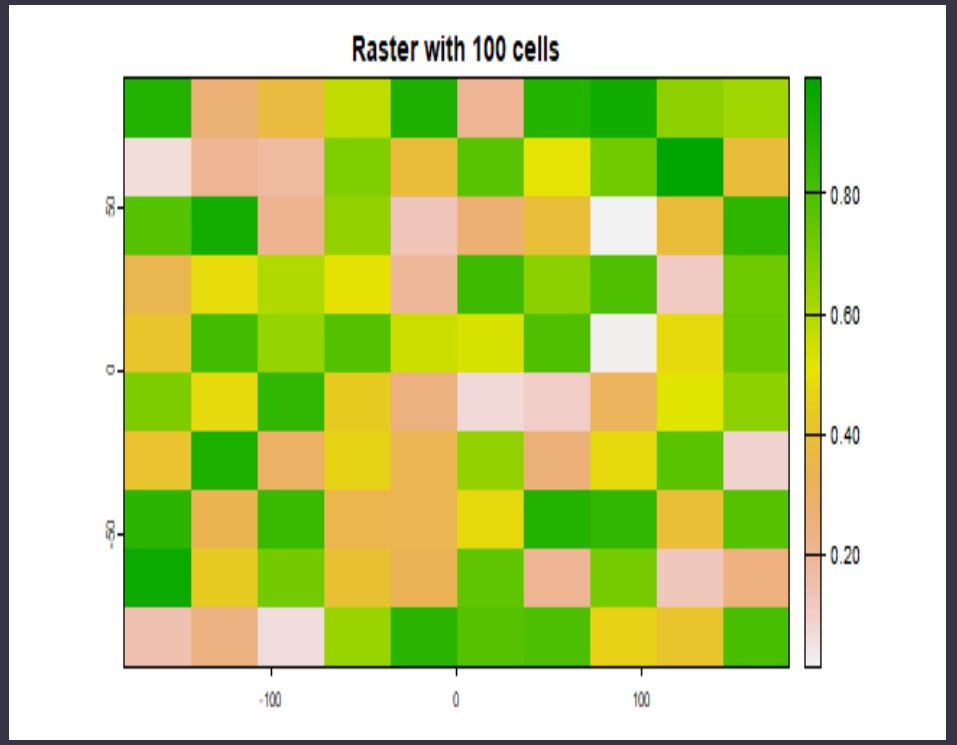
- Raster = grid of cells/pixels, each representing a real space in the world
- Pixels encode value of an environmental condition

Some predictors play a larger role than others for a given species, how can we determine which is which?

- Many approaches out there, many just try a lot of combinations

Consideration of temporal scale important

Consideration of spatial resolution important



Data: Physiological

What do we know about the species tolerances?

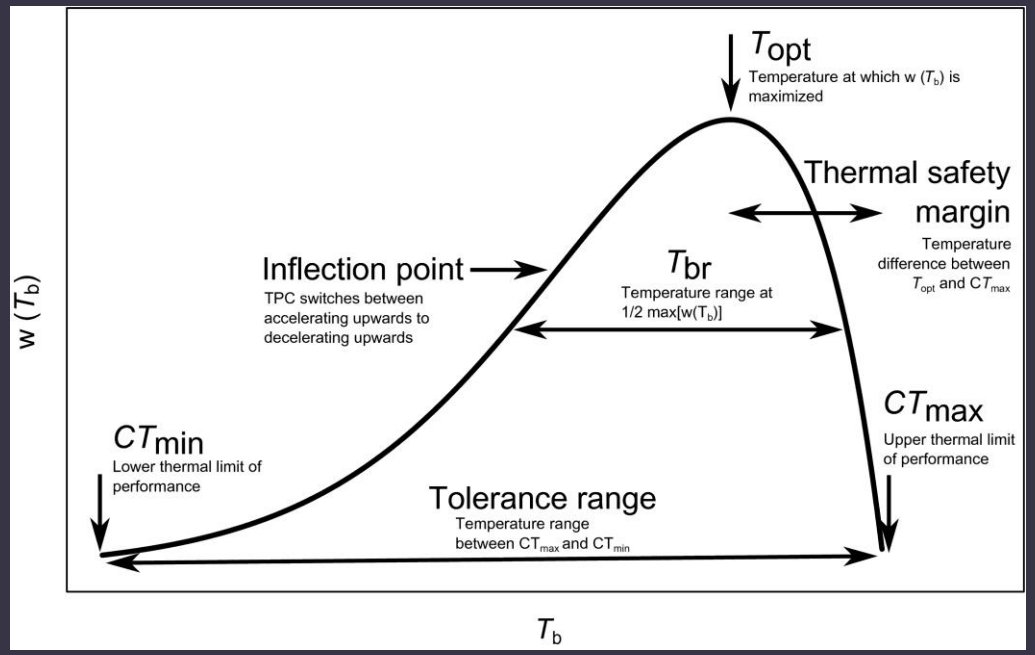
Knowing physiological limits can allow us to:

1) Clamp the extremes

2) Incorporate physiology into correlative frameworks

Temperature-Dependent Performance
- Probably the most abundant, characterized through TPCs

Xyz-Dependent Performance
- Salinity, humidity, etc



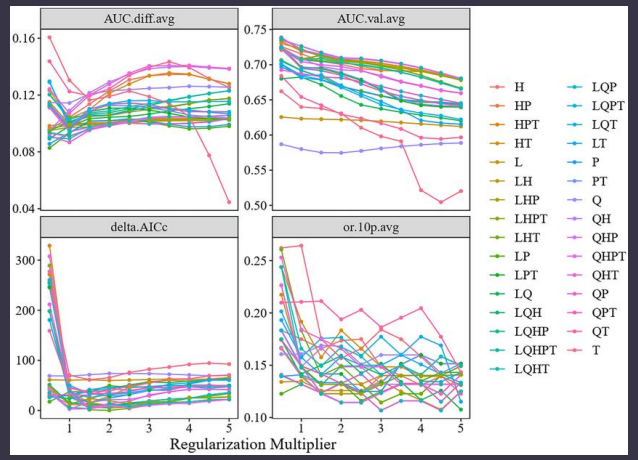
SDM: Fit & Evaluate

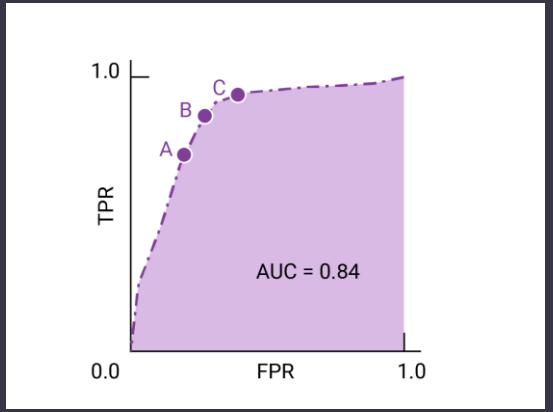
Model Selection: Dependent on modeling type (feature class, regularization multiplier, tree complexity, learning rate, etc.). Usually based on lowest ΔAICc

Fit evaluation: Test the fit against the withheld testing set. Often used include AUC-ROC, TSS, Boyce Index

Spatial Projection: Spatial projections of SDMs are just projections of the residual fit

Ensemble of Models (Optional): Ensemble predictions of multiple models (often performance-weighted averaging)





SDM: Assumptions

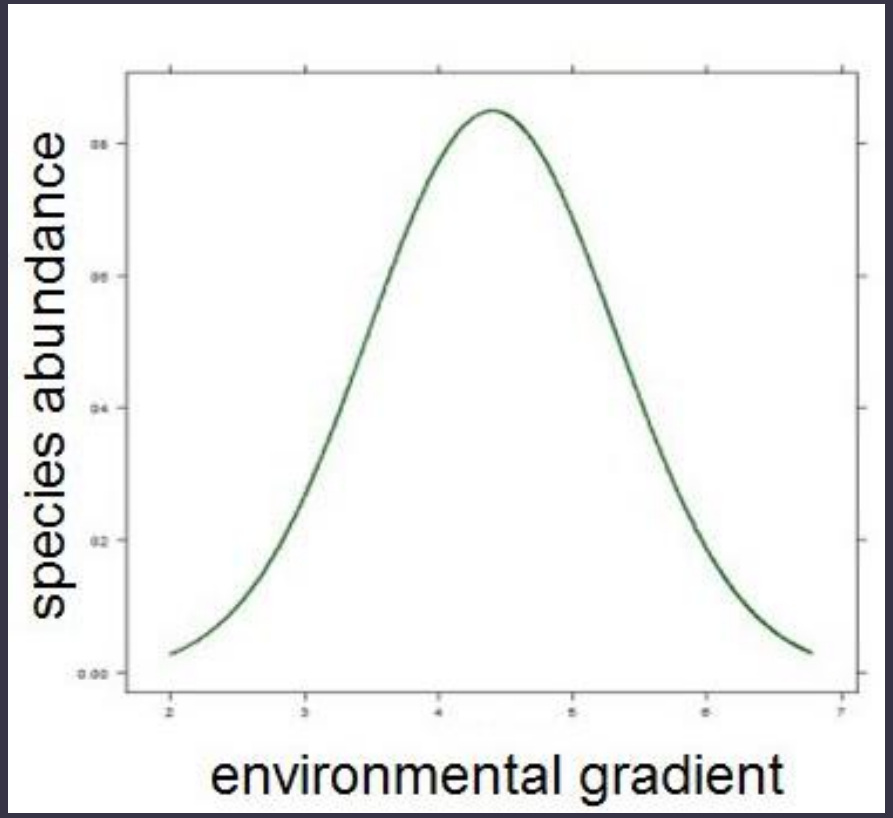
Species distribution modeling follows some general assumptions

1. Species-Environment Equilibrium
2. Availability of Important Predictor Variables
3. Appropriateness of Species Observations
4. Appropriateness of Statistical Methods
5. Occurrence Data is Unbiased
6. Occurrence Data is *not* Autocorrelated

Assumption 1: Species-Environment Equilibrium

Sampled occurrence data captures a sample of a species realized niche

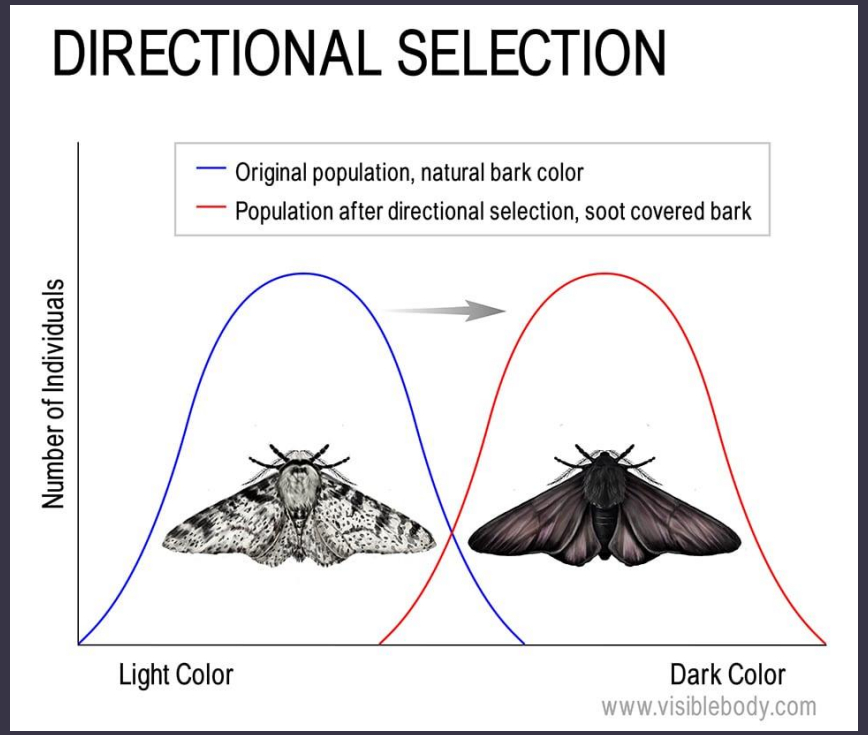
- There is a relationship between the environment & probability of occurrence
- A species fills all of the niche space available to it via dispersal
- Selection ***maintains a stable niche***



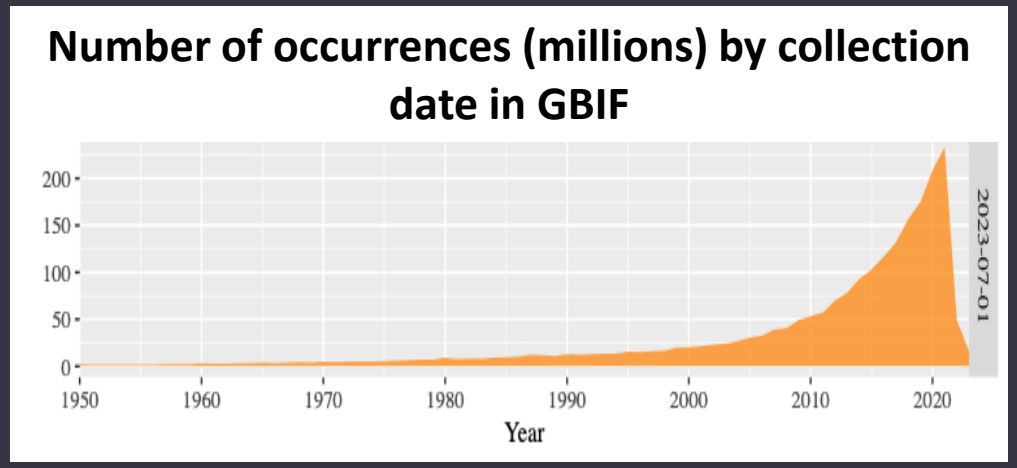
This relationship remains static

Assumption 1: Species-Environment Equilibrium

Violation: Species realized niche is changing due to selection pressures



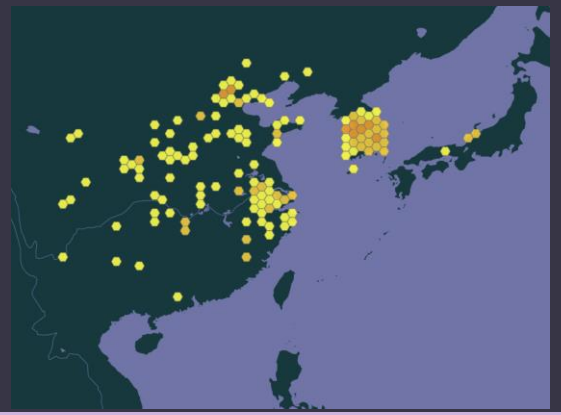
Caveat: This change may be negligible within the time period that occurrence data has been collected



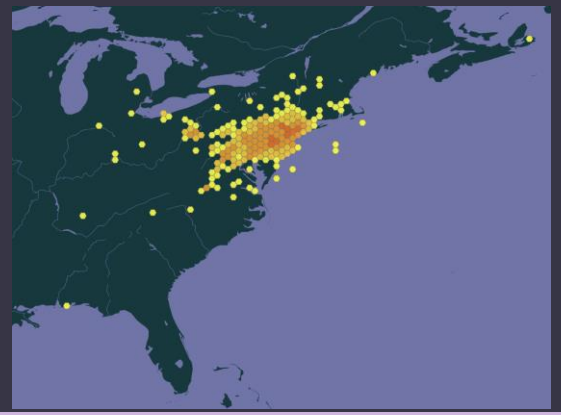
Assumption 1: Species-Environment Equilibrium

Violation: Species realized niche is changing due to selection pressures

Invasive species: once dispersed to a new environment biotic release often allows expansion beyond native realized niche. This means they often do better in their new range due to lack of pathogens/parasites/competitors/predators



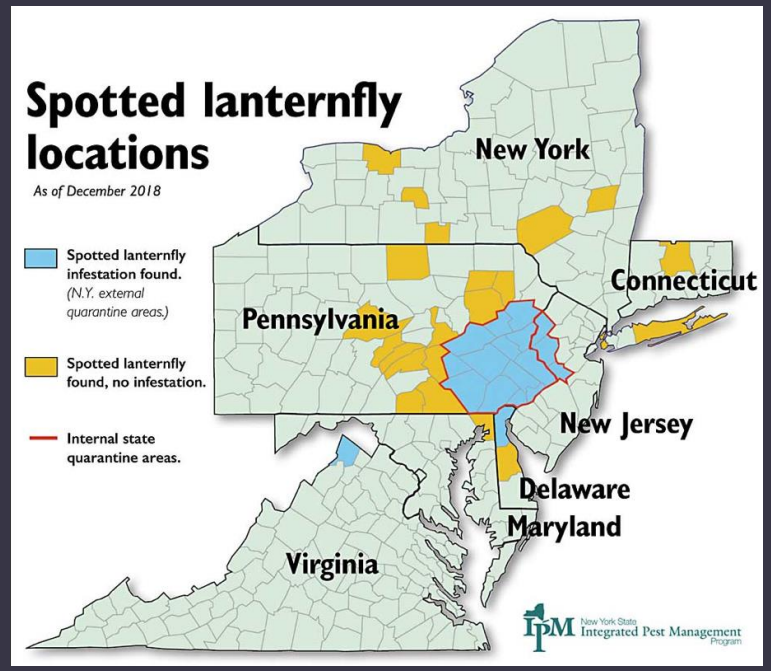
Lycorma delicatula: Native range

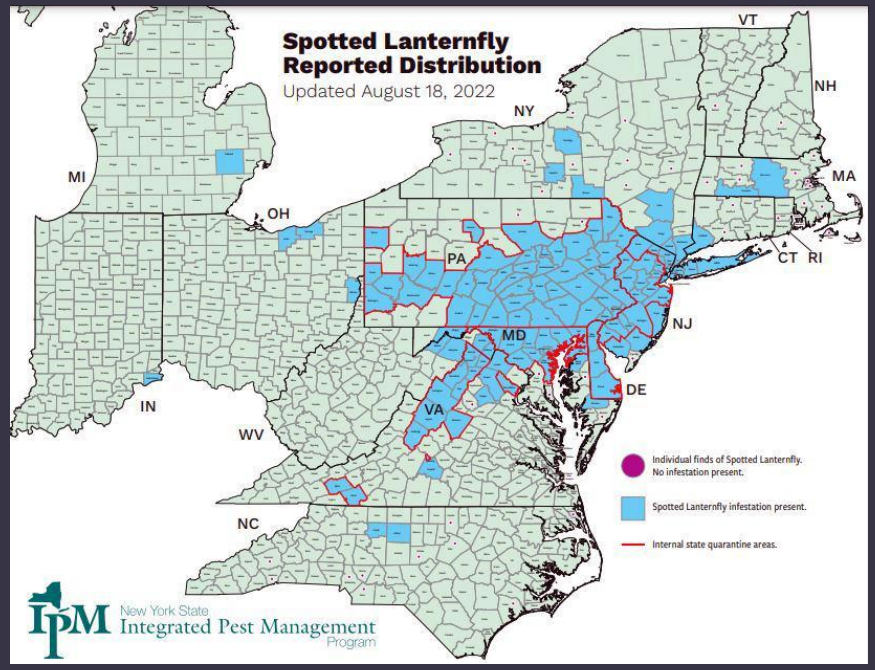


Lycorma delicatula: Invaded range

Assumption 1: Species-Environment Equilibrium

Violation: Species realized niche is changing due to selection pressures





Some invasive species may not yet be in equilibrium with their new habitat.

If that is the case, train your models using the home range only (or both)

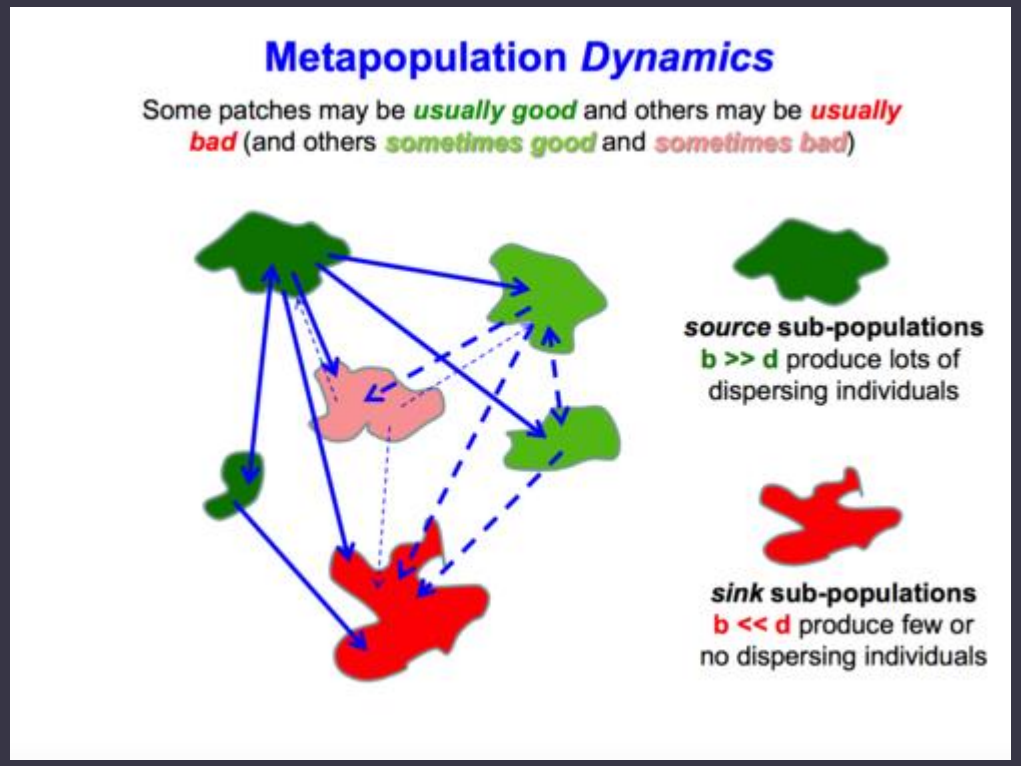
Assumption 2: Important Environmental Predictors Are Available

We do not always data for environmental variables that are important for a given species survival

Ex: very few relevant predictor variables (and their associated data) are available for deep water species

Assumption 3: Appropriateness of Species Observations

Is the species able to experience positive population growth where it was found?



Assumption 4: Appropriateness of Statistical Methods

Is a predictor amenable to being modeled?

Numerical Data

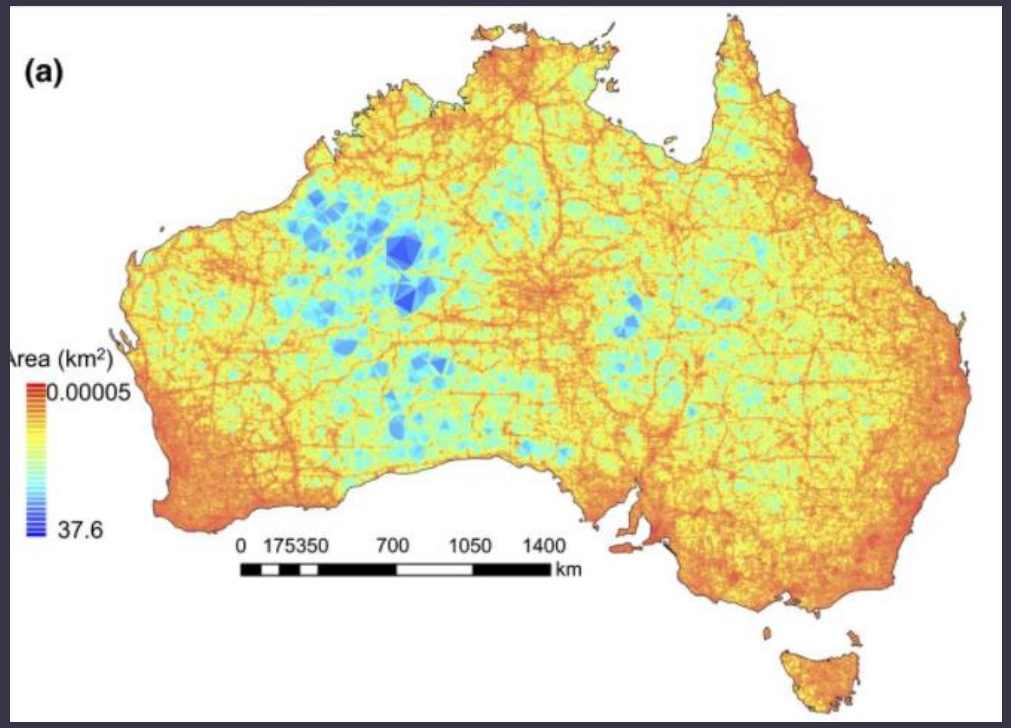
- % Clay
- mm precipitation in March
- Diurnal temperature range
- Units of photosynthetically active radiation per week

Categorical Data

1. **Clay**
2. **Sand**
3. **Silt**
4. **Loam**

You can code soil categories as integers, but these are symbolic and do not have any mathematical relationship

Assumption 5: Occurrence Data is Unbiased



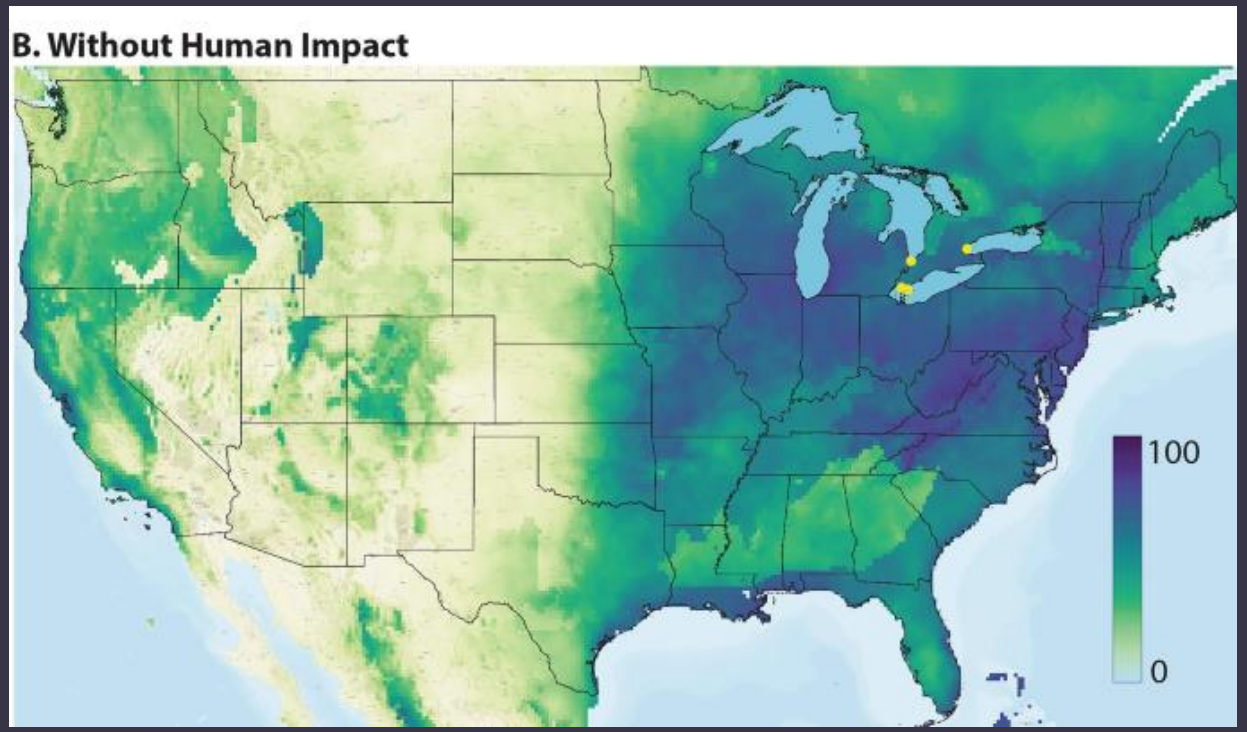
Avg. area / Collection

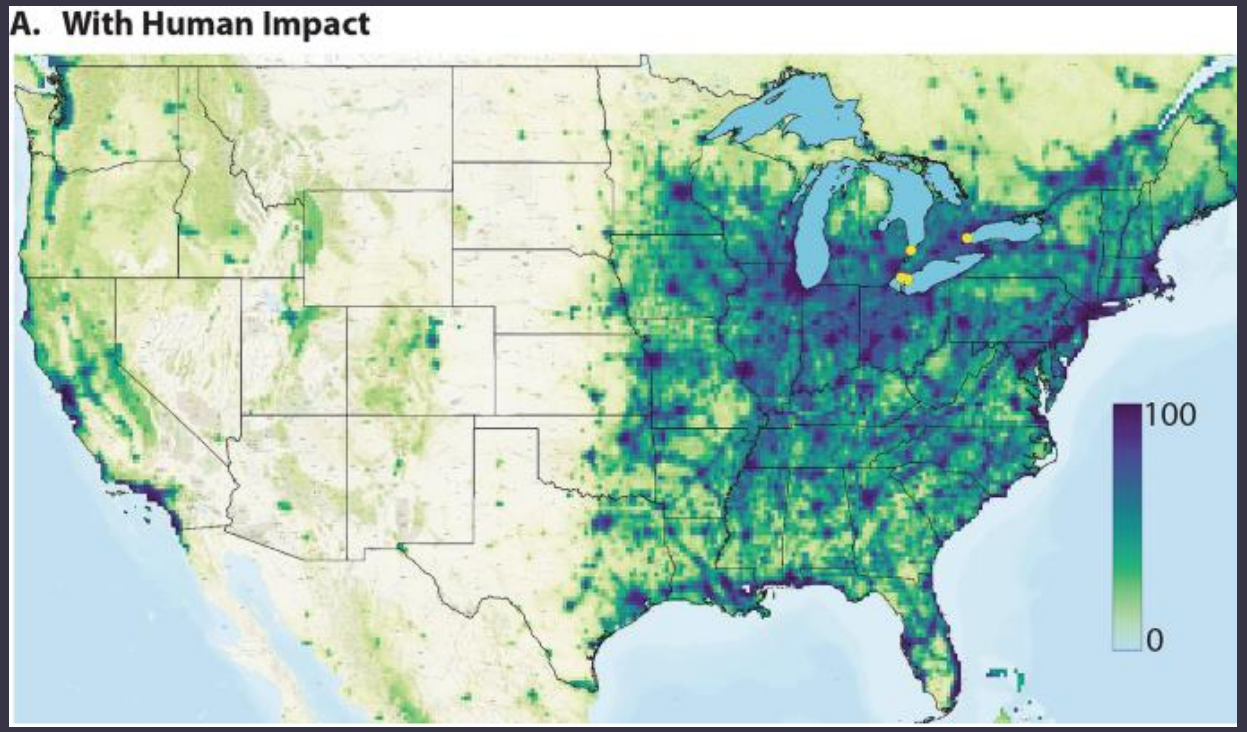
People don't go out and randomly collect occurrence records

- Near roads, habitation/collection facilities
- Megafauna and flora favored in sampling record
- Fungi, micro-organisms under-represented

There are ways to combat this bias

Example of Bias in Spatial Projections





Assumption 6: Occurrence Data is not Autocorrelated

Spatial autocorrelation: the probability of a species occupying area XYZ is dependent on the probability of a species occupying a space adjacent to XYZ

Organisms are likely to live close to where they were born/hatched/budded

Species are also likely to be found where they have been found before, even under changes in condition

Species distributions are driven by previous distributions as well as niche

SDM: Application-Based Uses

Application-based:

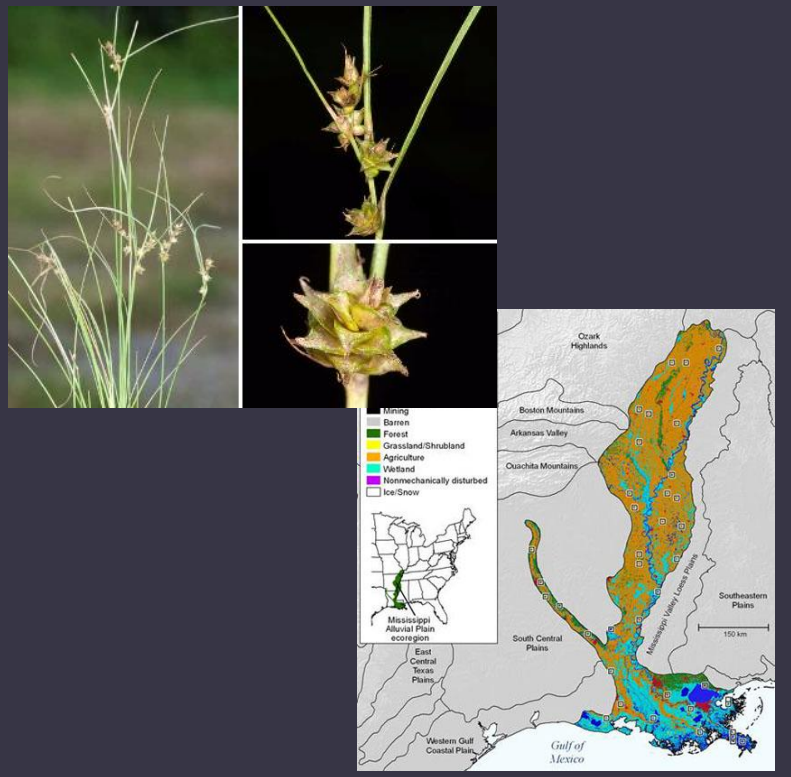
Predict changes in habitat suitability under change scenario/time

Map invasion potential

Help inform what species need more (or less) focus

Predict past distributions and shifts

Identify areas that *may* contain new populations of rare species



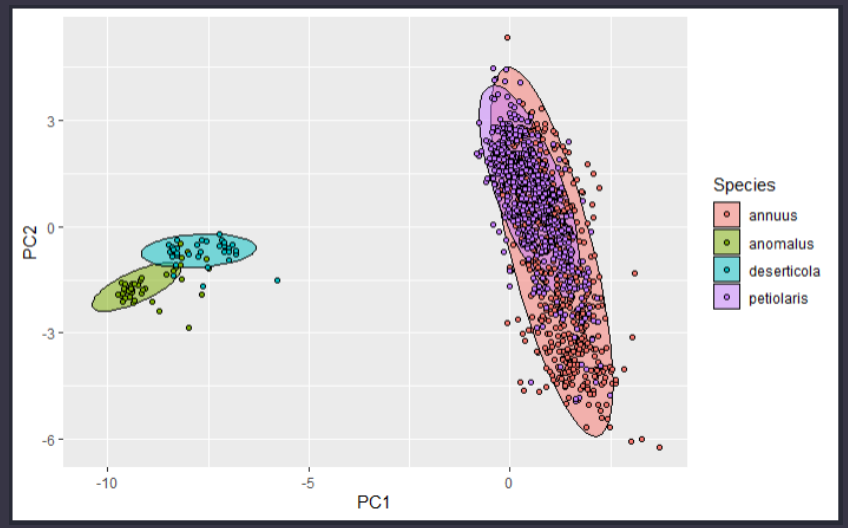
SDM: Theory-Based Uses

Theory-based:

Quantify the niche space of an organism/population

Identify factors that are highly associated with a given distribution

Add support for hypotheses regarding speciation



Future Meeting Ideas

1. Occurrence points, background points, spatial partitioning
2. Predictor data
3. Model selection
4. Evaluating SDMs
5. General workshops: I have code and exercises already written up