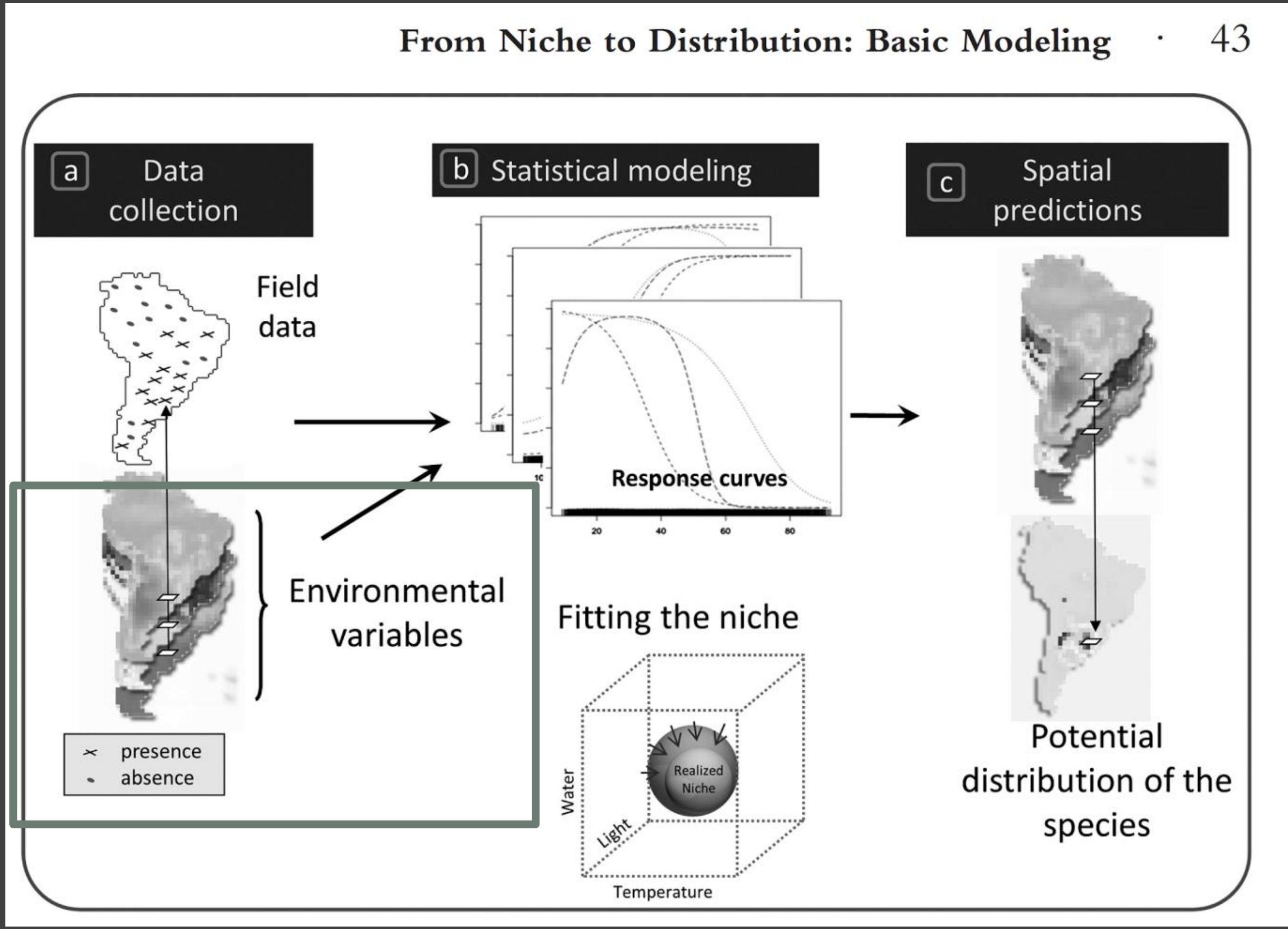
SDM Crash Course #3: Predictor Variables & How to Pick Them




Nicholas Galle, University of Notre Dame – Department of Biological Sciences

What are we doing today?



Predictor Variables:

How are they stored?

How is it gathered?

What Data is available?

Considerations

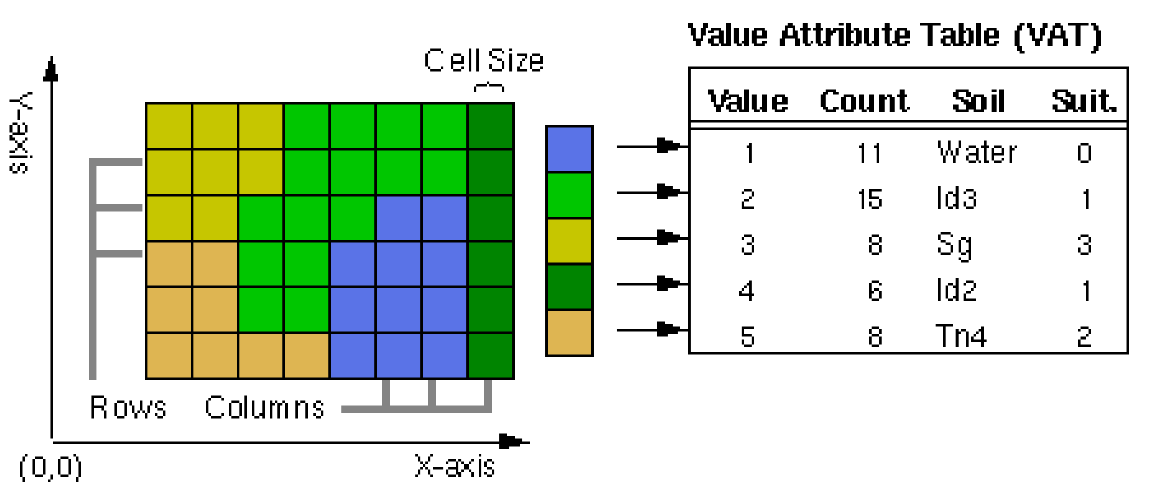
How is Predictor Data Stored?

Models generally use a grid-system known as rasters to store predictor variables

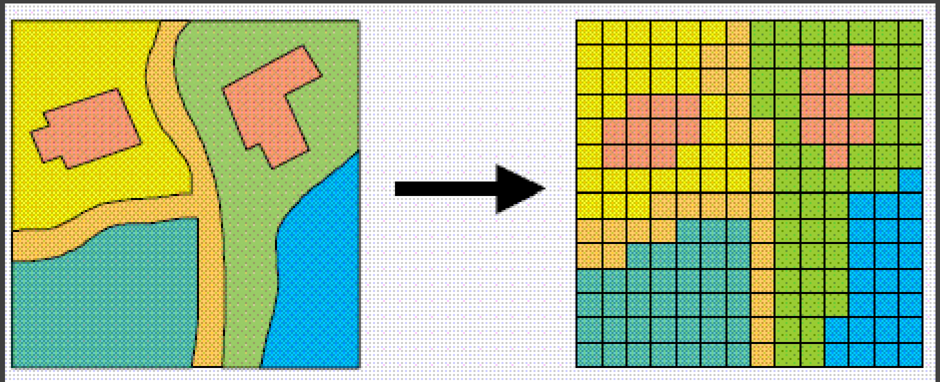
- Raster: a grid of pixels used to store information, where each pixel has **a** value representing a data point

Can be stored in a variety of formats, often .TIFF or .GeoTIFF

Note: Rasters do not capture intra-cell variability, each cell has a single value



Value	Count	Soil	Suit.
1	11	Water	0
2	15	Id3	1
3	8	Sg	3
4	6	Id2	1
5	8	Tn4	2

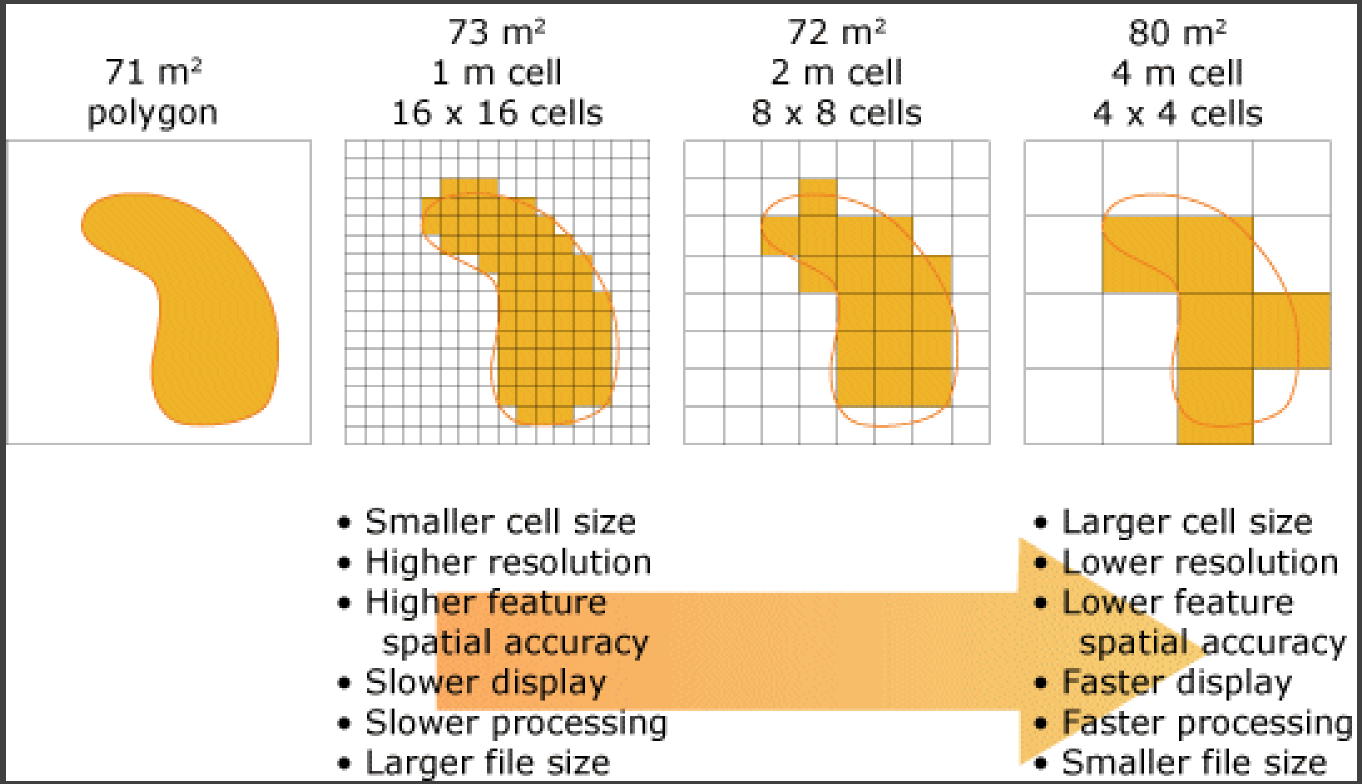


Predictor Variable Attributes: Resolution

Arc
- Decimal degrees
- Arc seconds / Arc Minutes

Size
- Hectares
- Km^2

Map Ratio
- 1:125,000

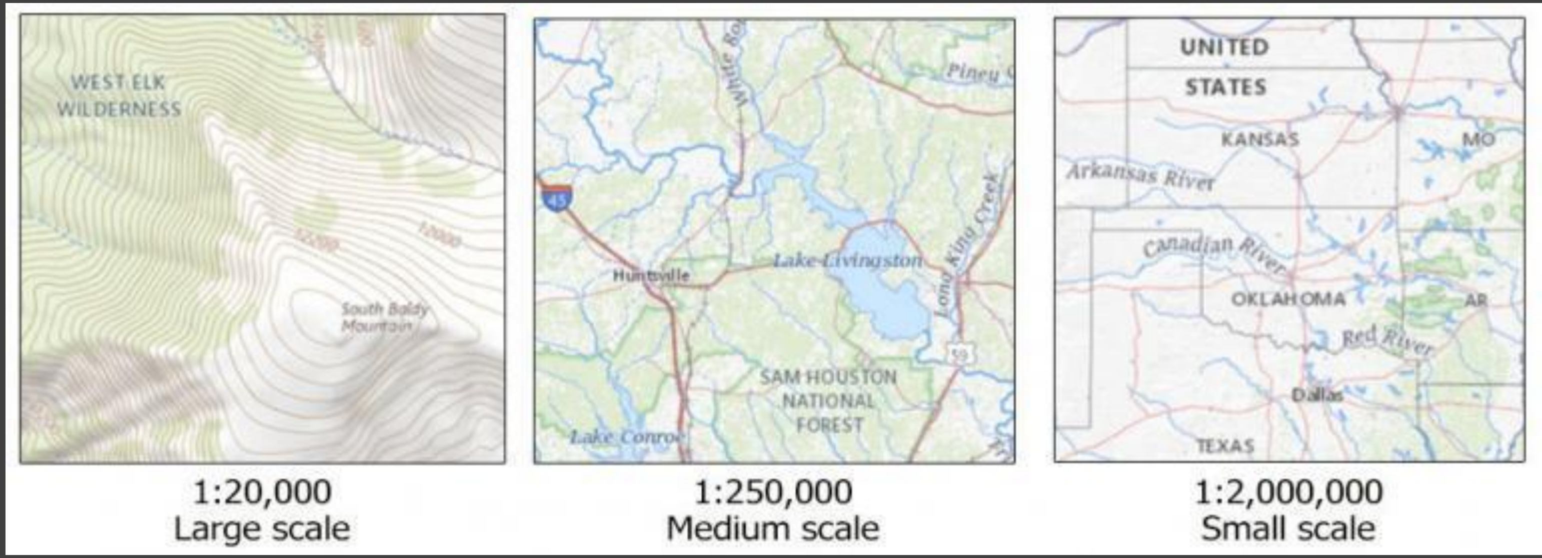


Predictor Variable Attributes: Geographic Scale

<u>Extent:</u> The area over which environmental correlates have been measured

- Will cover in a future talk, study extent should vary depending on 1) the particular species being modeled, & 2) the question at hand

A larger extent requires either more cells or a lower resolution

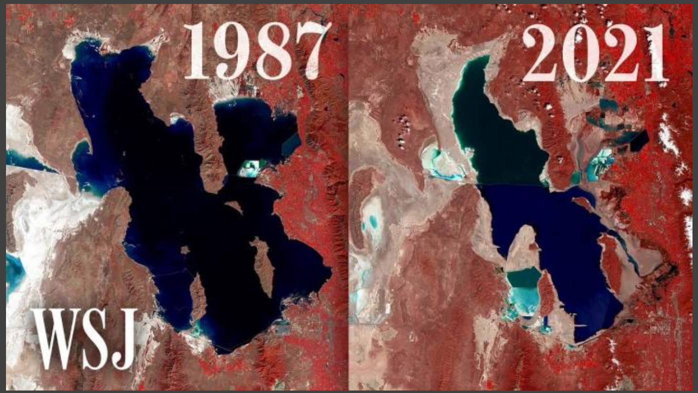


Predictor Variable Attributes: Temporal Scale

Should consider the temporal scale

- Ex: If I am investigating shifts in the last two decades, would average temperature from 1970-2000 really be appropriate?

Should consider seasonality



Predictor Variable Attributes: Temporal Scale

Should consider the temporal scale

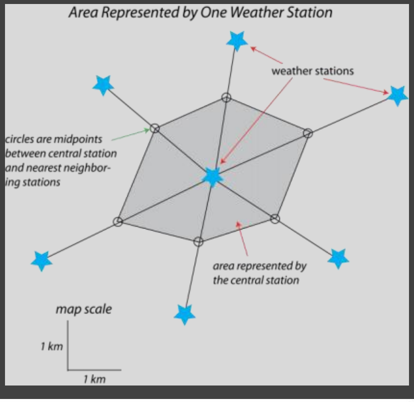
Should consider seasonality

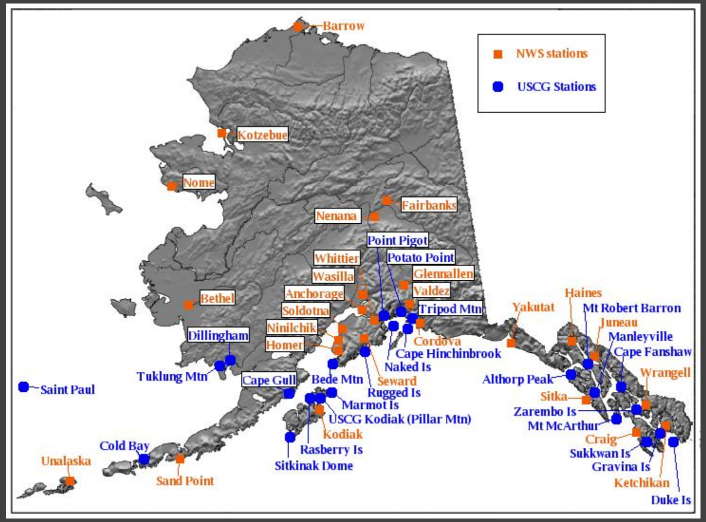
- Cases where a species undergoes diapause or seasonal migration rely on seasonality

How is this data gathered?

Readily available environmental data is not raw

- Interpolated between weather stations
  - Angular Distance Weighting, forms of Kriging, Inverse Distance Weighting, Nearest Neighbor Interpolation, etc, etc
- Remote sensing

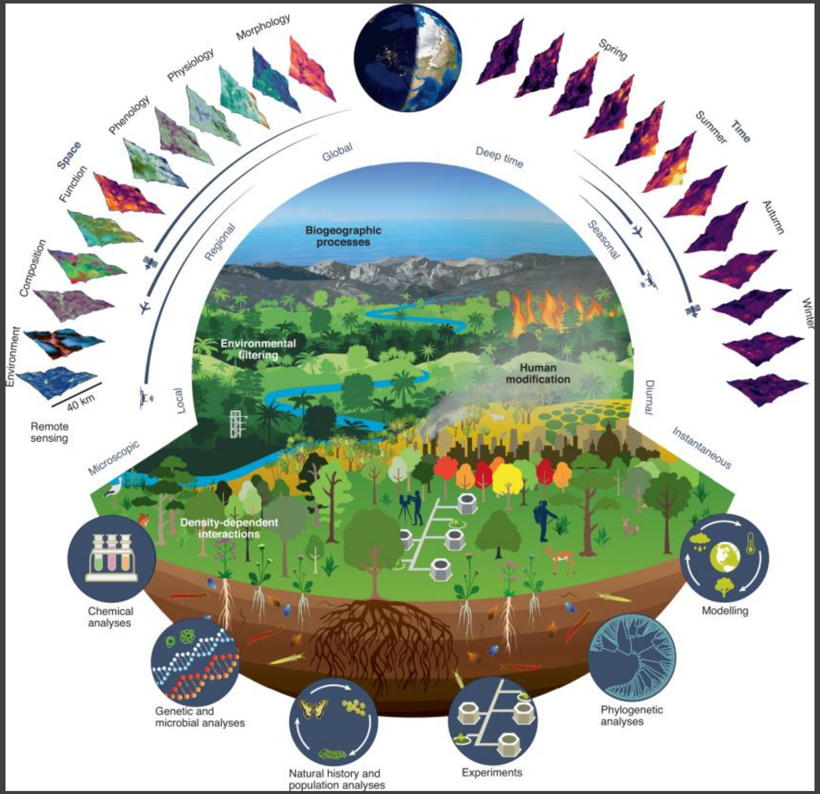




How is this data gathered?

Readily available environmental data is not raw

- Interpolated between weather stations
- Remote sensing
  - Varies in availability at the temporal and geographic scale

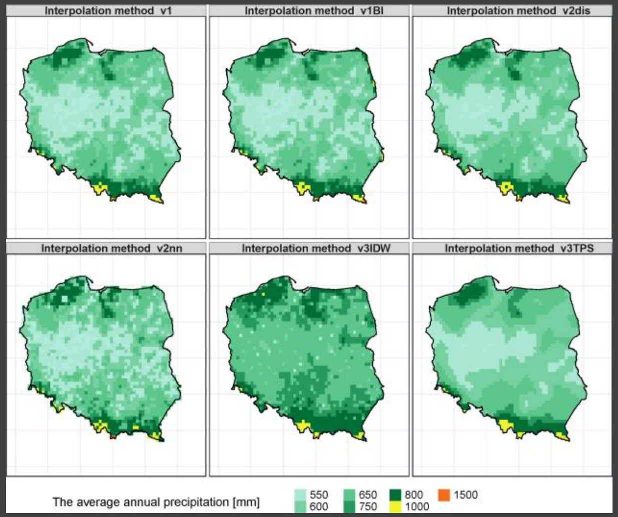


Issues with These Approaches

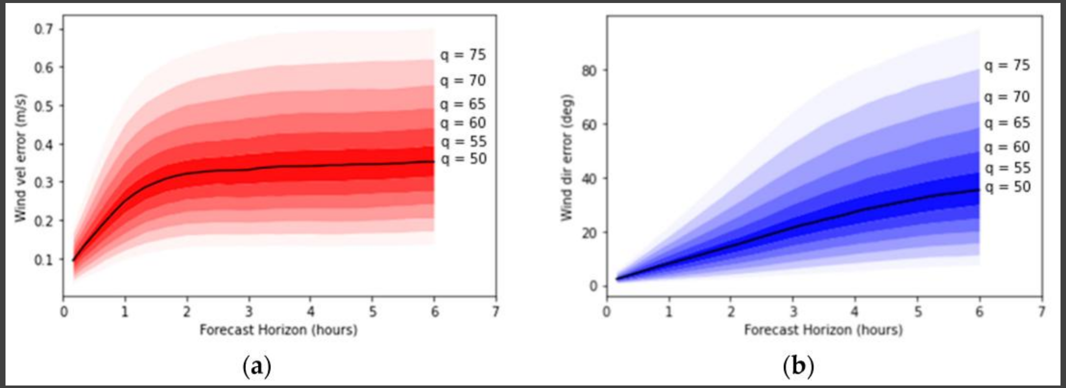
Each predictor will have some degree of error

- Instrumentation imprecision, interpolation differences, etc.

This error is rarely quantified or available



Issues?

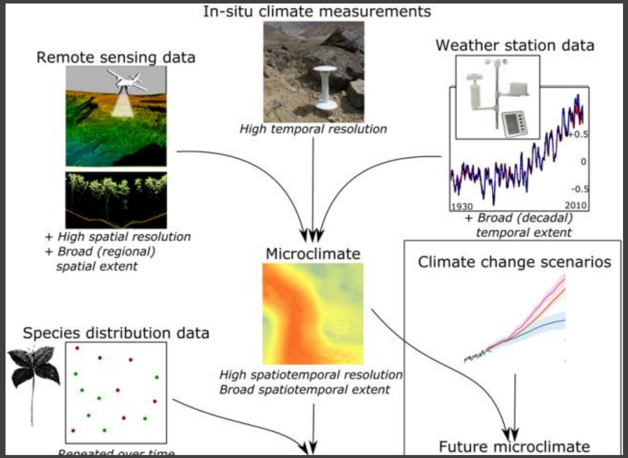


An aside about microclimatic data

Very high-resolution datasets (spatially or temporally)

Used to be predominantly collected via in-situ sensors, remote sensed *seems* to be increasingly common

2018 review on the use of microclimate in SDMs



Review and synthesis | Free Access

Incorporating microclimate into species distribution models

Jonas J. Lembrechts, Ivan Nijs, Jonathan Lenoir

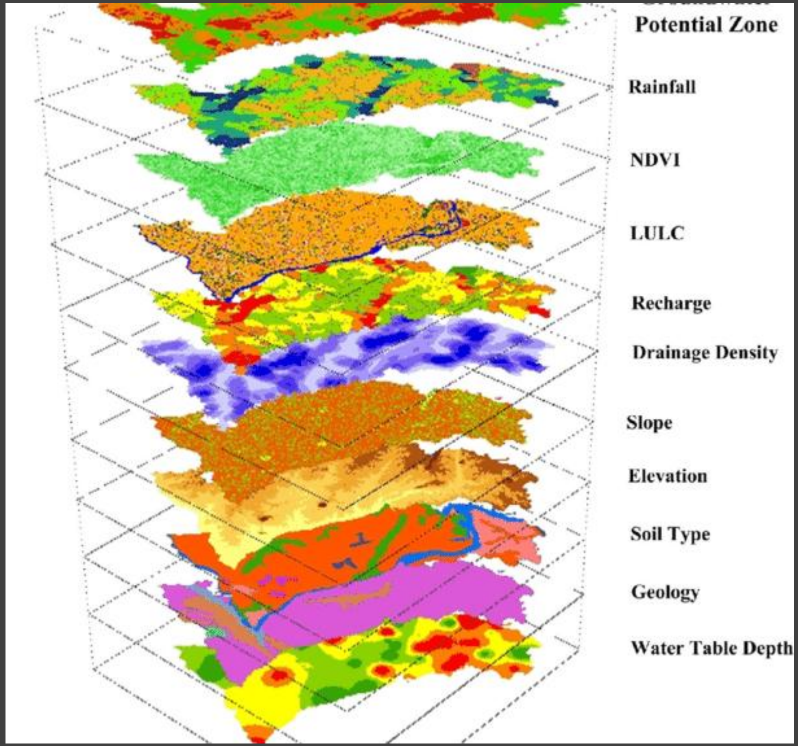
First published: 28 September 2018 | https://doi.org/10.1111/ecog.03947 | Citations: 221

Types of Data Available: Environmental

Environmental

- Bioclimatic variables
  - Permutations of temperature/precipitation/radiation
- Geographic Variables
  - Elevation and derivatives, soil type, etc.
- Anthropogenic Variables
  - Land-use type, human impact, population density

Physiological

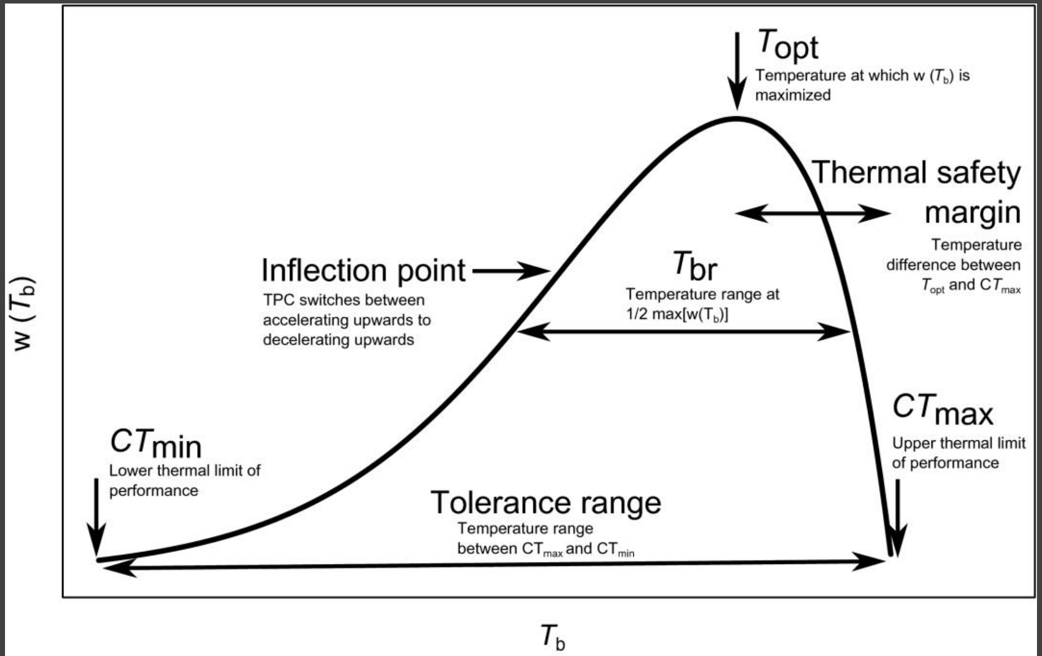


Types of Data Available: Physiological

Environmental

Physiological

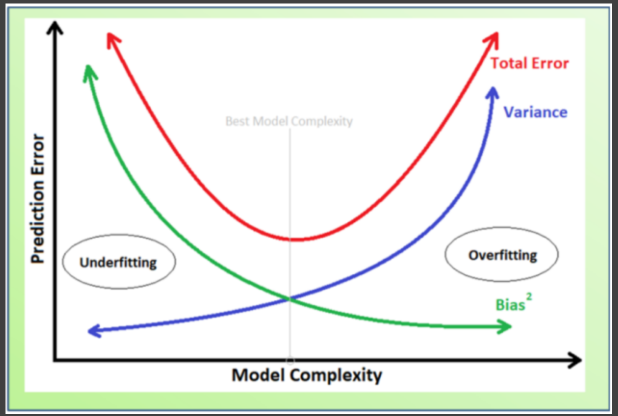
- Temperature-dependent trait performance
  - Most commonly used physiology metric
- XYZ-dependent trait performance
  - Salinity, etc.
- Often generated in part using the prior environmental raster layers

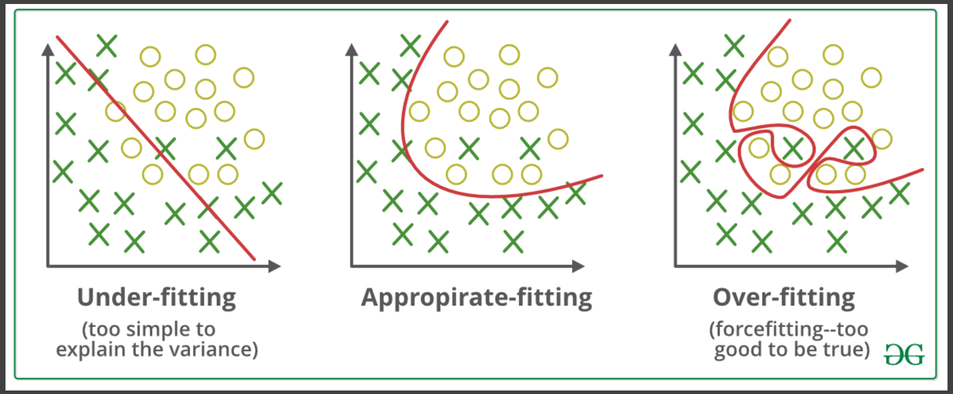


How to Choose?

Why not use every predictor?

- Sufficiency
  - Rough rule of thumb: a robust model requires >10-20 occurrence points for each predictor to avoid overfitting
- Redundancy
- Collinearity

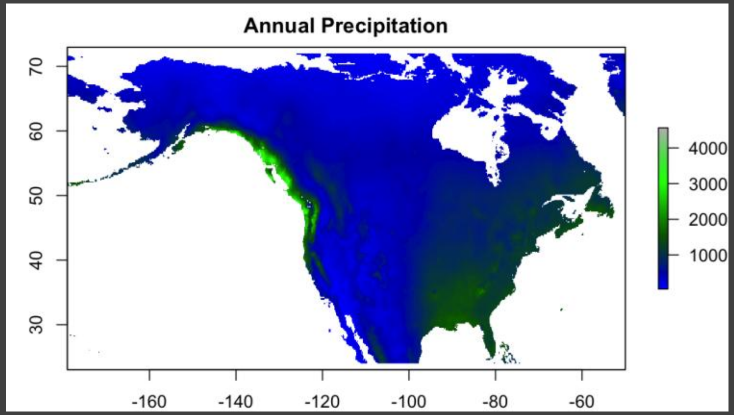


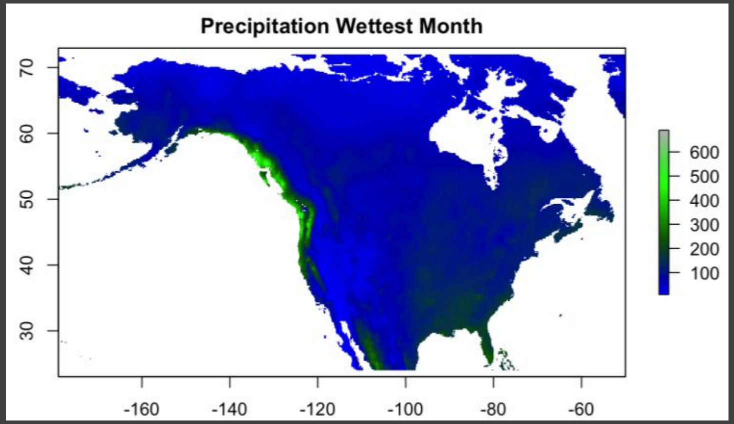


How to Choose?

Why not use every predictor?

- Sufficiency
- Redundancy
  - Some variables measure very similar things, adding little new information to model training other than error
- Collinearity

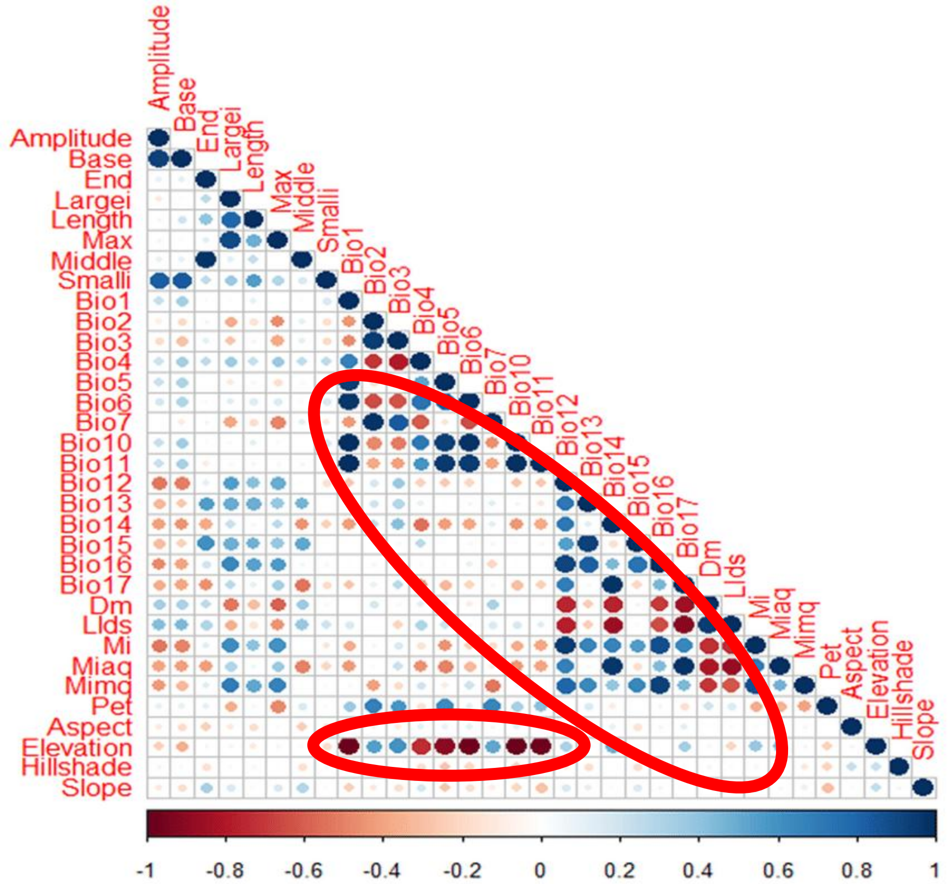




How to Choose?

Why not use every predictor?

- Sufficiency
- Redundancy
- Collinearity
  - Predictors might be collinear with each other, including these together can produce biased models



How to Choose?

Best-practices in a perfect world:

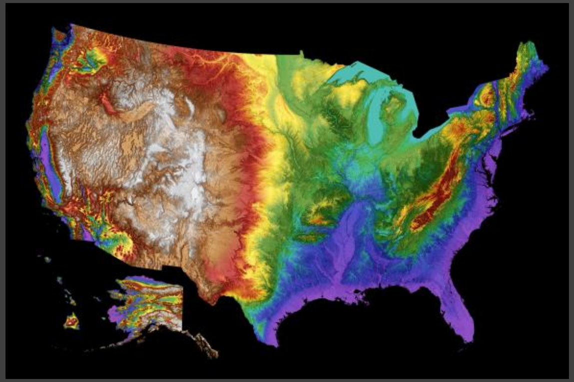
- Figure out what is known about the physiology and ecology of the species, select important variables accordingly

Meaningful

- An underground organism is unlikely to be heavily impacted by cloud cover

Direct vs. Indirect Variables

- Organism might be distributed as a function of a variable, but it could be indirect

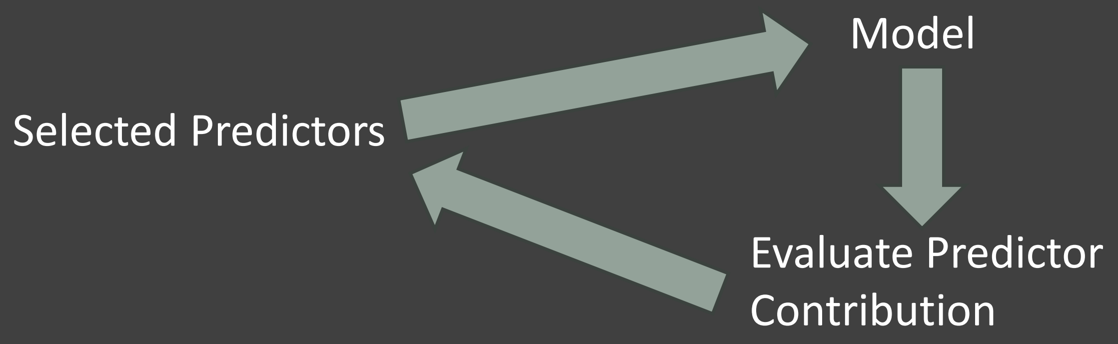


How to Choose?

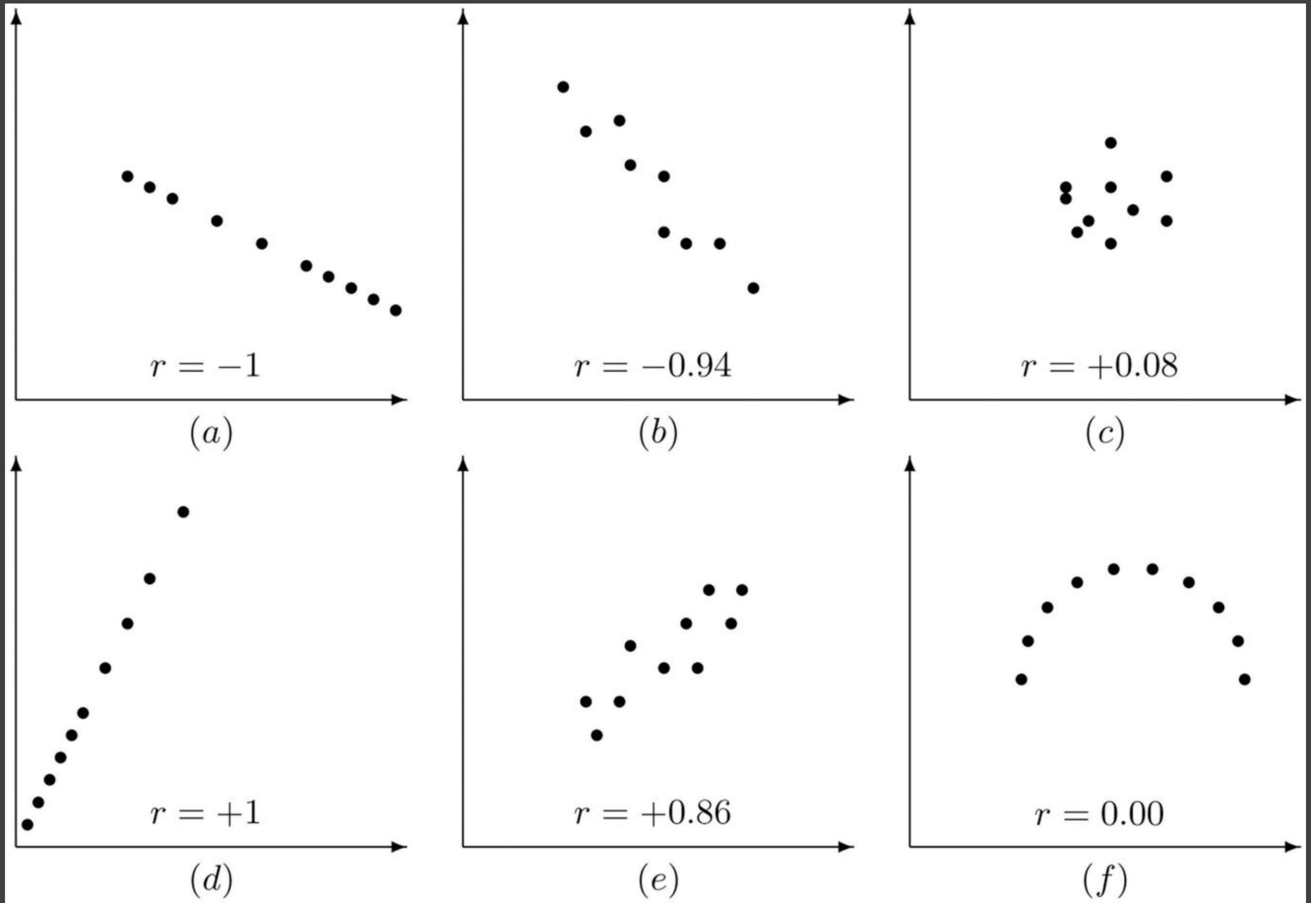
<u>What people actually do:</u>

Statistical approach

- Narrow down a large set of predictors by identifying which are the most informative and show the least collinearity
  - Recursive tuning process: start with many, measure importance and collinearity, remove some, repeat



Dealing with Collinearity



$-1 \leq r \leq 1$

- Sign is direction
- Absolute value is magnitude

r^2

- Conservative estimate of magnitude

$r \geq 0.7$

- Strong correlation

Variance Inflation Factor

Variance Inflation Factor: a statistical measure used to identify how each predictor contributes to multicollinearity

$$VIF = \frac{1}{1 - r_i^2}$$

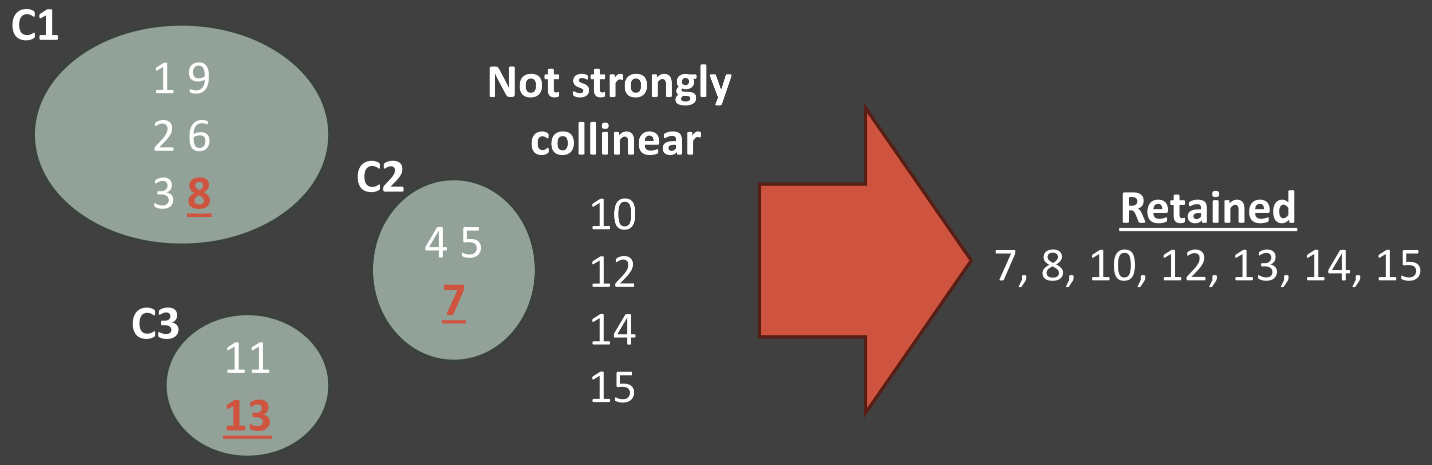
Where r_i^2 is the correlation coefficient of the ith predictor regressed on the other coefficients

VIF > 5: Strong contributor to multicollinearity

1. Estimate VIF for all predictors (If all < 'threshold', stop)
2. If not, eliminate predictor with highest VIF, go back to step 1
   - VIFstep r function does this in a forward, step-wise manner

Cluster-Based Selection

1. Create clusters of highly collinear variables (via some metric)
2. Select and retain the most informative/biologically informed variable per highly collinear cluster
3. Re-run to ensure variables are not highly collinear



Appendix 7. Variable selection using peNMIG spike-and-slab priors#.

Variables	*Schistosoma haematobium*	*Schistosoma mansoni*
Group1†		
Annual mean temperature	-	-
Max temperature of warmest month	-	-
Min temperature of coldest month	-	-
Mean temperature of wettest quarter	-	-
Mean temperature of driest quarter	-	Selected
Mean temperature of warmest quarter	-	-
Mean temperature of coldest quarter	-	-
LST at night	Selected	-
Group 2†		
Mean diurnal temperature range	Selected	-
Isothermality	-	-
Temperature seasonality	-	-
Temperature annual range	-	Selected
Group 3†		
Annual precipitation	Selected	-
Precipitation of wettest month	-	-
Precipitation of wettest quarter	-	-
Group 4†		
Precipitation of driest month	-	Selected
Precipitation of driest quarter	Selected	-
Variables moderately correlated	-	-
Precipitation seasonality	-	Selected
Precipitation of warmest quarter	Selected	Selected
Precipitation of coldest quarter	Selected	Selected
LST in the day time	Selected	Selected
NDVI	Selected	-
Land cover	Selected	Selected
Elevation	-	Selected
Water distance	Selected	Selected
Climatic zones	Selected	-
pH measured in water	Selected	Selected
Soil moisture	Selected	Selected
Human influence index (HII)	-	-
Urban extents	-	-
Gross domestic product (GDP)	Selected	-
Infant mortality rates (IMR)	Selected	Selected
Proportion of improved sanitation	Selected	-
Proportion of improved drinking-water sources	-	Selected
Survey type	Selected	Selected

#Results are based on surveys in mainland sub-Saharan Africa; †a maxiumum of one variable can be selected in each highly correlated group.

Of All Models: Garbage In, Garbage Out