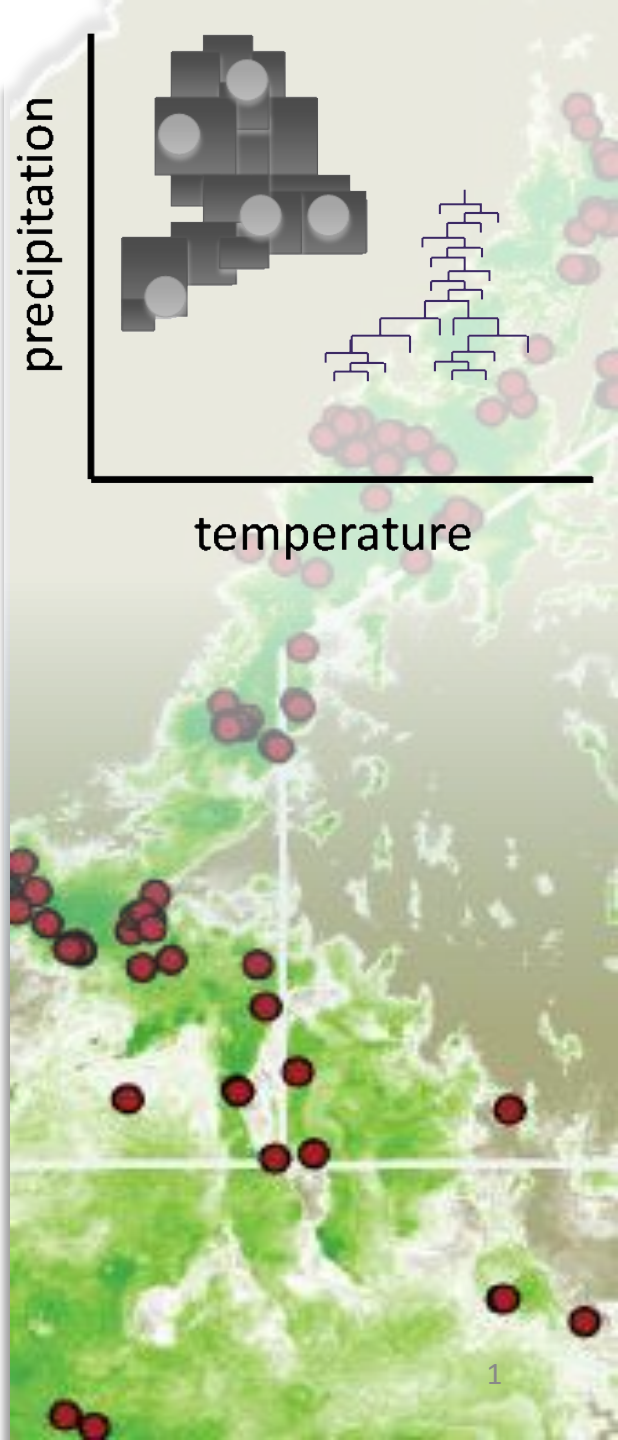
An Introduction to Best Practices in Species Distribution Modeling

Adam B. Smith
Center for Conservation & Sustainable Development
Missouri Botanical Garden
4344 Shaw Boulevard, Saint Louis, MO 63110
adam @ earthskysea org

All course materials can be downloaded at
www.earthskysea.net ("ecology resources").

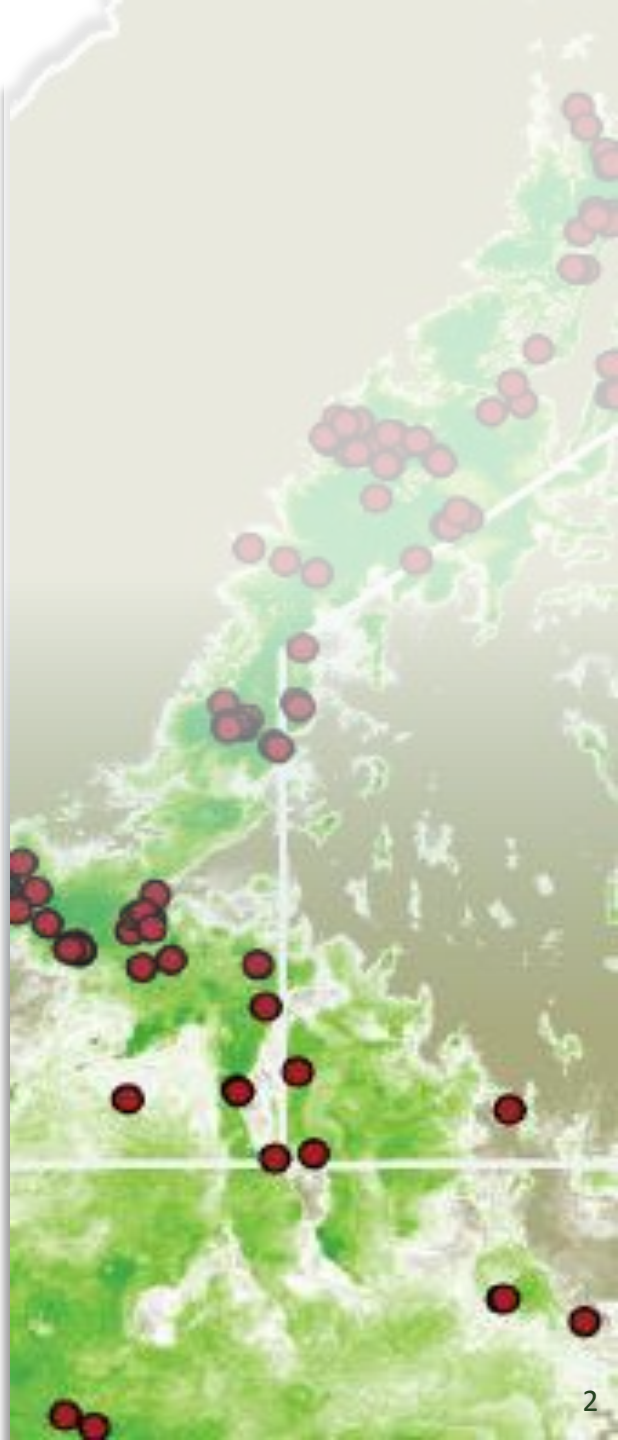
Kansas State University
Friday, October 19, 2012



Why “Best practices”?

- There are **many “cautionary” papers** on how not to model.
- There is **no single, comprehensive guide** on how to model species’ distributions.
- The **field is moving quickly**.

This short course will introduce guidelines for making useful predictions in real-world situations.



A short course on distribution modeling

Outline I

Introduction to distribution modeling
What does a SDM do?
Model algorithms
Input: Species' records
Input: Predictors

Exercise I: Maxent
Input: Species' records, predictor rasters
Output: Maps, thresholds, AUC, variable importance, response functions

Maxent under the hood:
Information entropy maximization
"Feature" functions
Regularization and beta parameter
Assumptions

Exercise II: Maxent with SWD format, k-folds, and projecting to new era/region
Input: SWD format
K-fold data splitting
Projection to new era/region

Best practices: "Truth" vs. "useful bias"
SDM assumptions

Best practices: Training records
Data cleaning
Coordinate uncertainty
Number of records

Best practices: Predictors
Proximate & direct, conditions & resources
Types
Resolution
Accuracy
Correlations
Gradients
Dynamic vs. static vs. dynamic-but-static

A short course on distribution modeling

Outline II

Best practices: Training absences/ pseudoabsences/background

"True" absences

"Pseudoabsences"

"Background"

"Targeted"

Best practices: Study extent vs. range size

Delineation, range size : study extent

Best practices: Model parameterization

Maxent: "Feature" functions, β regularization

Other SDM algorithms

Presence/absence vs. presence-only?

Weighting absences

Ensembling

Best practices: Model evaluation

Maps

Compare to spatial-only model

Autocorrelation in residuals

Response functions

Absences

Performance metrics

Thresholding

Stability/bootstrapping

Techniques for rare species

Using known but extirpated presences or failed reintroductions

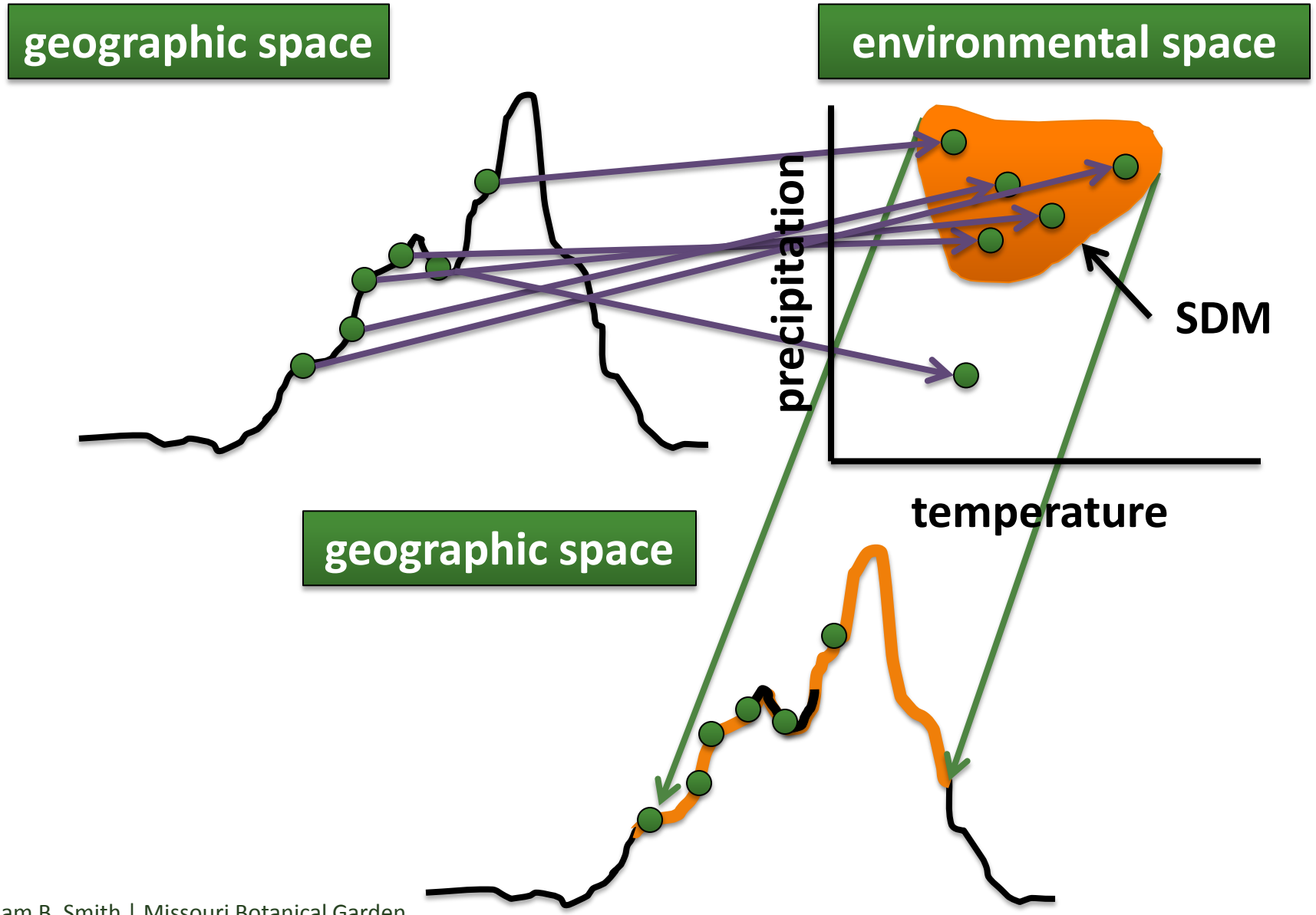
Coarse → fine scale

Uni/bivariate ensembles

Use model to search for more populations

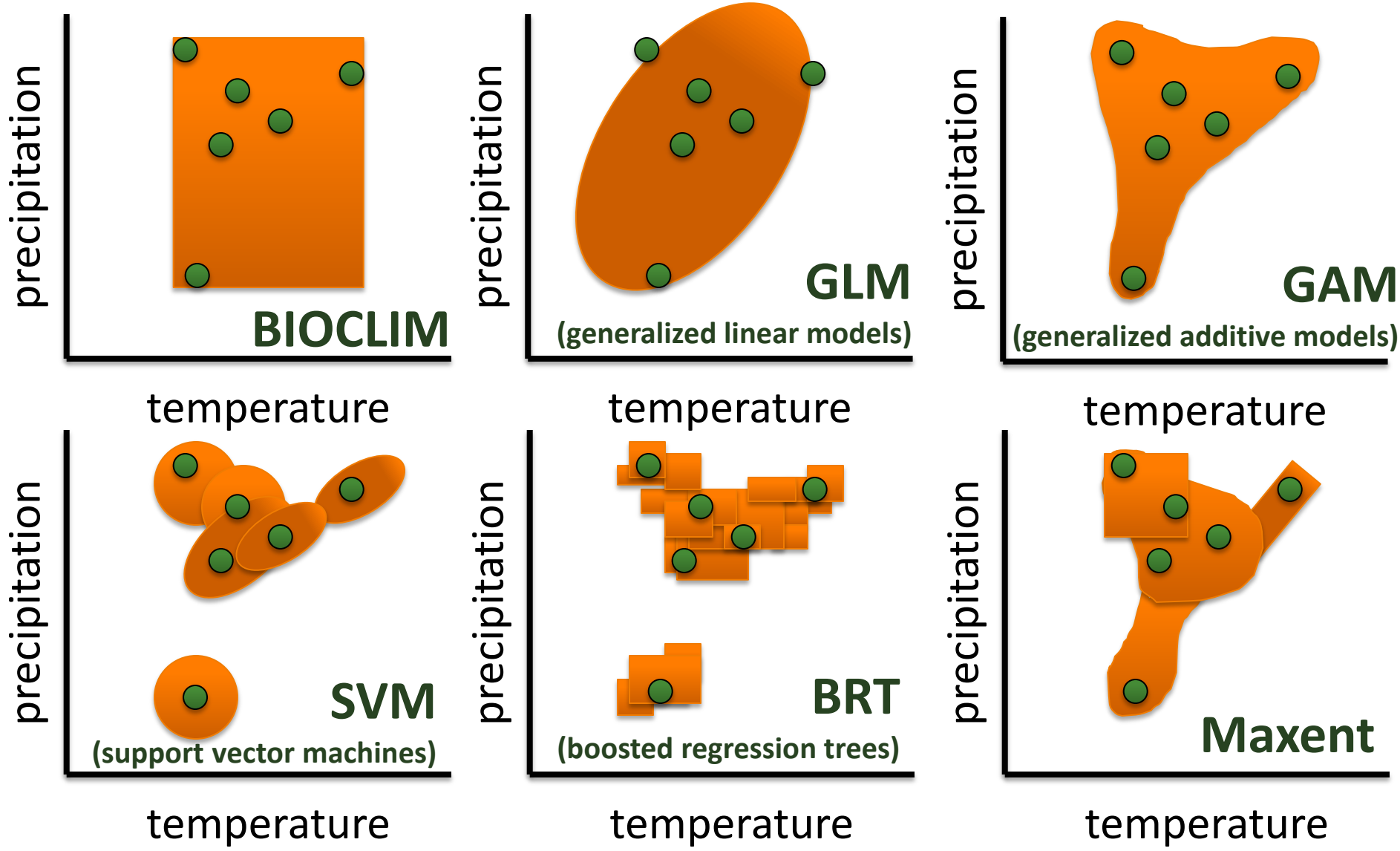
Introduction to distribution modeling

What do SDMs do?



Introduction to distribution modeling

Model algorithms



Introduction to distribution modeling

Model algorithms

SDM	Species data type	Predictor inter-actions	Highly non-linear functions	Categorical predictors OK	Missing data
BIOCLIM	pres-only	no	no	no	discarded
ENFA	presence-only	yes	no	no	discarded
Generalized Linear Models	presence-only* & presence/absence	yes	if user-specified	yes	discarded
Generalized Additive Models	presence-only* & presence/absence	yes	yes	yes	discarded
Boosted Regression Trees	presence-only* & presence/absence	yes	yes	yes	OK
Support Vector Machines	presence-only & presence/absence	yes	yes	yes	discarded
Random Forests	presence-only* & presence/absence	yes	yes	yes	OK
Maxent	presence-only	yes	yes	yes	discarded (default)

* needs pseudoabsences for presence-only

Introduction to distribution modeling

Model algorithms

SDM	Data "hungry"	Software interface*	Deter-ministic	Performs well	"Pre-tuned"**
BIOCLIM	yes	DIVA-GIS, ModEco, "dismo"	yes	no	no
ENFA	no	BioMapper	yes	?	no
Generalized Linear Models	no	ModEco, "BIOMOD", "dismo"	yes	no	no
Generalized Additive Models	yes	"BIOMOD", "dismo"	yes	yes	no
Boosted Regression Trees	yes	"BIOMOD", "dismo"	no	yes	no
Support Vector Machines	no	ModEco, "dismo"	yes	yes (limited testing)	no
Random Forests	yes	"BIOMOD", "dismo"	no	yes	no
Maxent	no	Maxent, ModEco***, "dismo"	yes	yes	yes

* Items in quotes refer to R packages

** Tuning performed with multi-species dataset and presence/absence data for evaluation. (Phillips & Dudík. 2008. Ecography 31:161-175.)

*** not tuned for SDMing

Introduction to distribution modeling

Model algorithms

BIOCLIM

Busby, J. R. 1991. BIOCLIM – a bioclimate analysis and prediction system. In: Margules, C. R. and Austin, M. P. (eds.), Nature conservation: Cost effective biological surveys and data analysis. CSIRO, pp. 64–68.

Generalized linear/additive models (GLMs, GAMs)

Guisan et al. 2002. Generalized linear and additive models in studies of species distributions: Setting the scene. Ecological Modeling 157:89-100.

Boosted regression trees (BRTs)

Elith et al. 2008. A working guide to boosted regression trees. Journal of Animal Ecology 77:802-813.

Random Forests (RF)

Brieman. 2001. Random forests. Machine Learning 45:5-32.

Maxent

Phillips et al. 2006. Maximum entropy modeling of species geographic distributions. Ecological Modelling 190:231-259.

Phillips & Dudík. 2008. Modeling species distributions with Maxent: New extensions and a comprehensive evaluation. Ecography 31:161-175.

Elith et al. 2011. A statistical explanation of Maxent for ecologists. Diversity & Distributions 17:43-57.

Support Vector Machines (SVMs)

Guo et al. 2005. Support vector machines for predicting distribution of Sudden Oak Death in California. Ecological Modelling 182:75-90.

Ecological Niche Factor Analysis (ENFA)

Hirzel et al. 2002. Ecological-niche factor analysis: How to compute habitat- suitability maps without absence data? Ecology 83:2027-2036

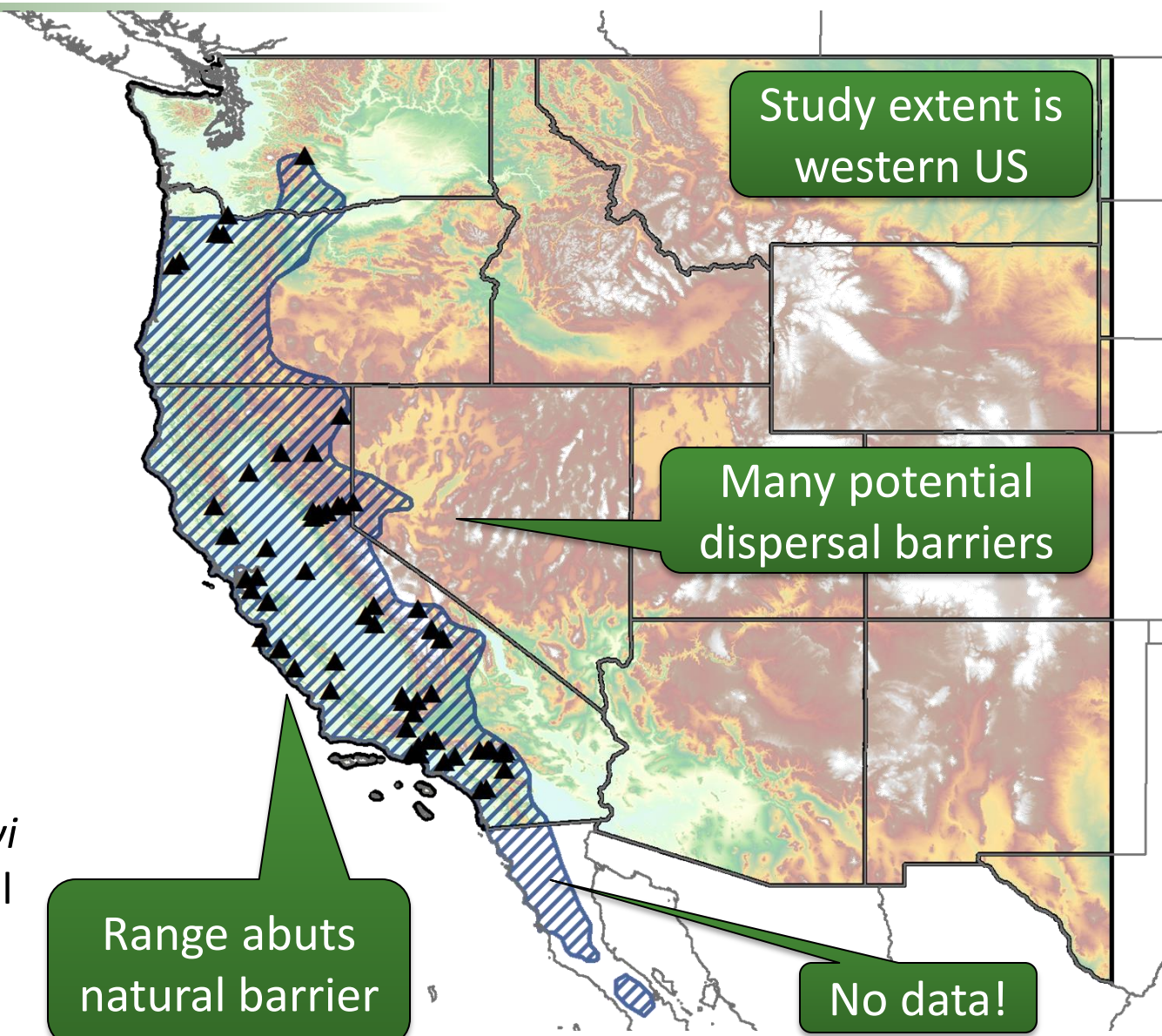
Ensemble models

Araújo, M.B. and M. New. 2007. Ensemble forecasting of species distributions. Trends in Ecology and Evolution 22:42-47.

Introduction to distribution modeling

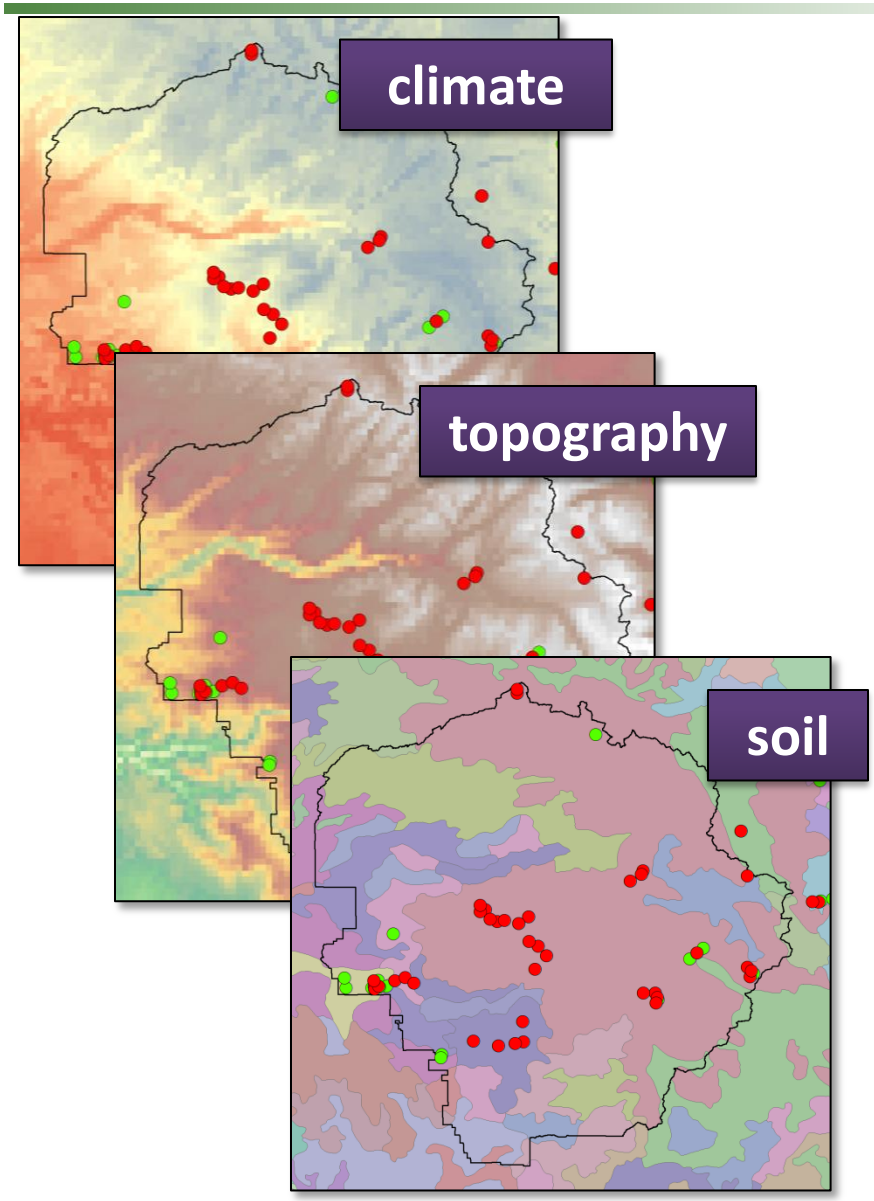
Species' records

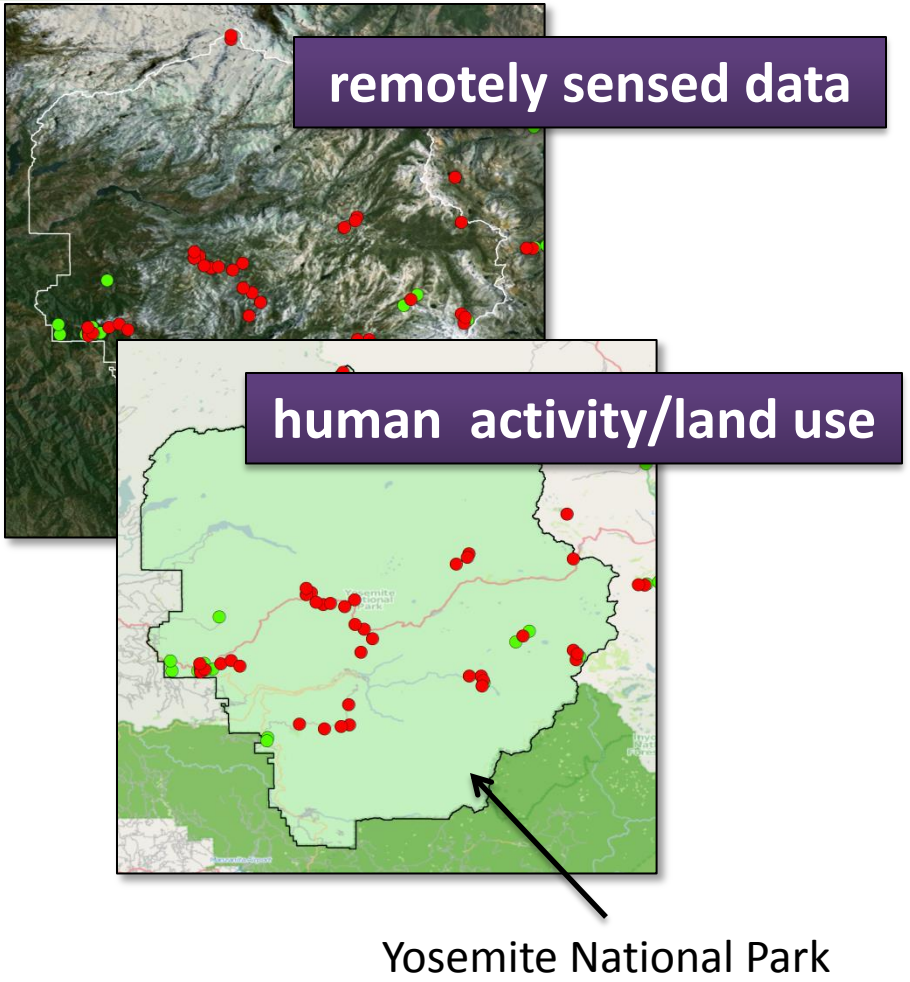
Otospermophilus beecheyi
Beechey's ground squirrel



Introduction to distribution modeling

Predictors





Introduction to distribution modeling

Predictors: "BIOCLIM" climate predictors

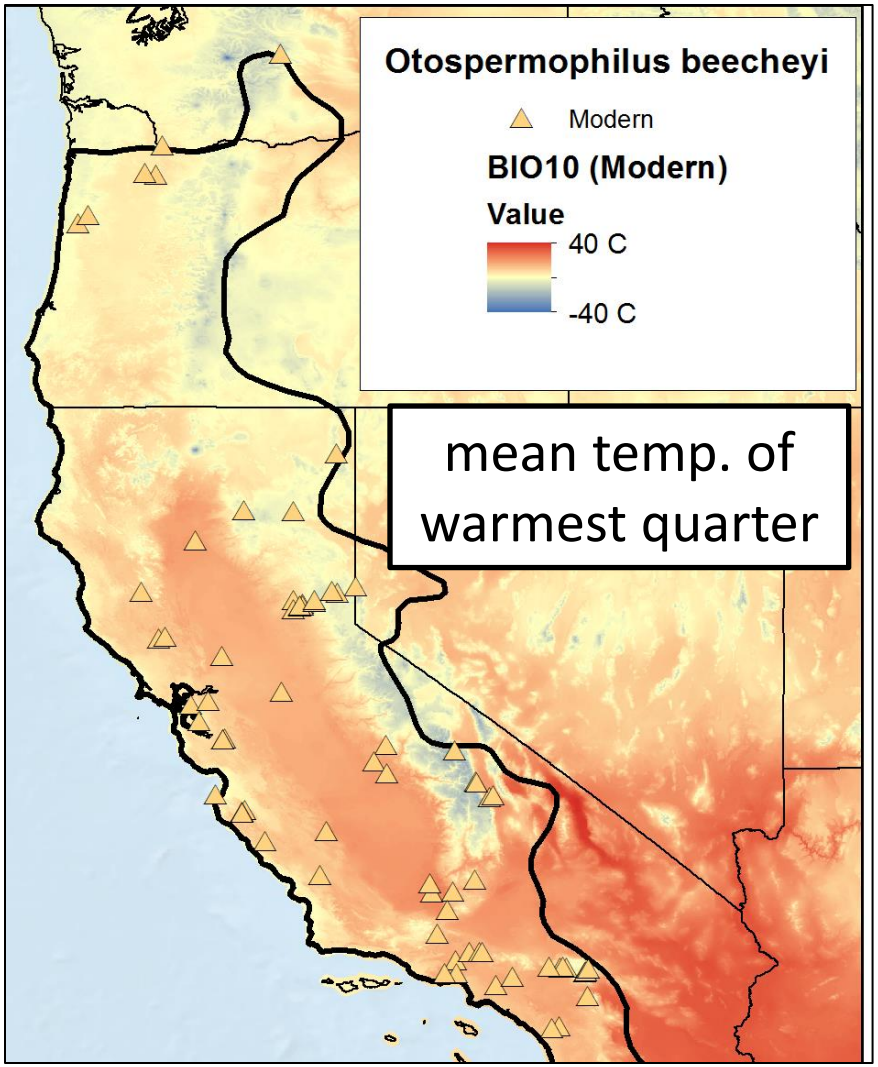
Temperature	
BIO 01	Mean annual temperature
BIO 02	Mean diurnal temperature range
BIO 03	Mean isothermality (BIO02/BIO07)
BIO 04	Mean temperature seasonality (standard deviation across months)
BIO 05	Mean maximum temperature of warmest month
BIO 06	Mean minimum temperature of the coldest month
BIO 07	Mean annual temperature range (BIO05 - BIO06)
BIO 08	Mean temperature of wettest quarter
BIO 09	Mean temperature of coldest quarter
BIO 10	Mean temperature of warmest quarter
BIO 11	Mean temperature of coldest quarter

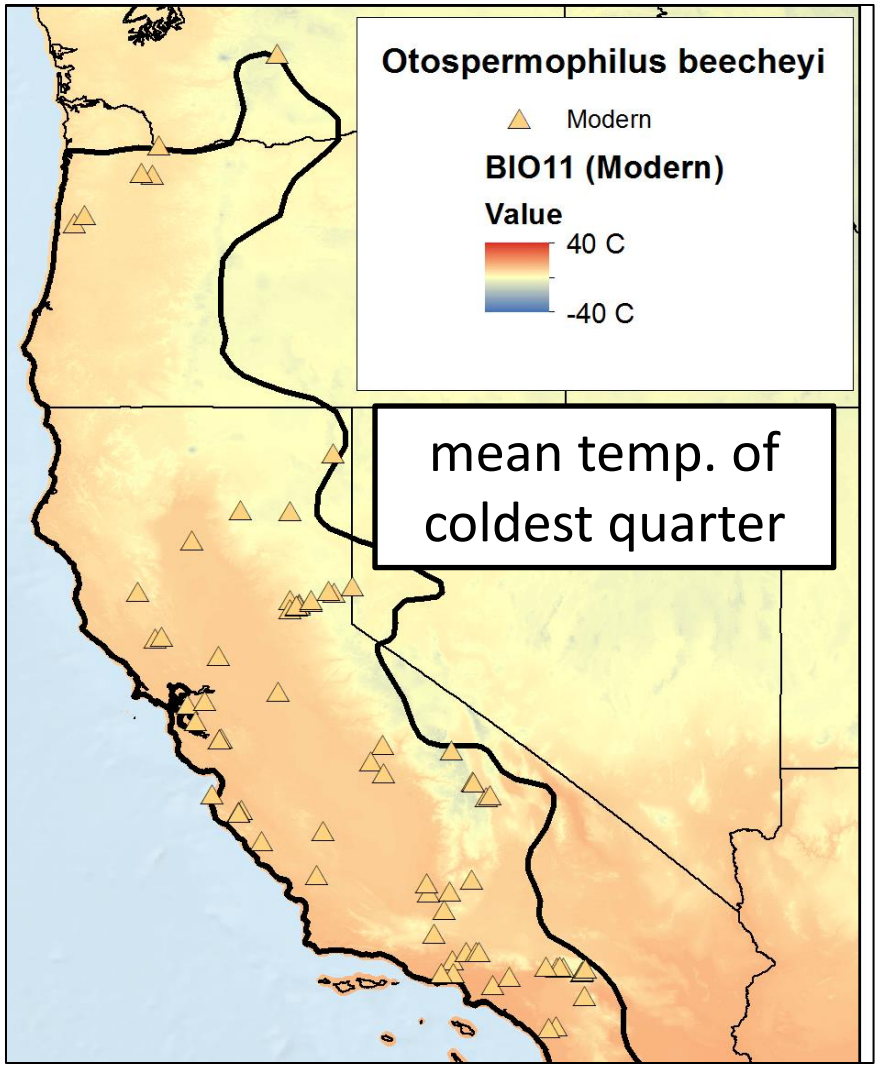
Moisture	
BIO 12	Mean annual precipitation
BIO 13	Mean precipitation of wettest month
BIO 14	Mean precipitation of the driest month
BIO 15	Mean precipitation seasonality (coefficient of variation across months)
BIO 16	Mean precipitation of wettest quarter
BIO 17	Mean precipitation of driest quarter
BIO 18	Mean precipitation of warmest quarter
BIO 19	Mean precipitation of coldest quarter
BIO 20-27	Humidity

Nix, H.A. 1986. A biogeographic analysis of Australian Elapid Snakes. In. Atlas of Elapid Snakes of Australia. (ed.) R. Longmore pp. 4-15. Australian Flora and Fauna Series Number 7. Australian Government Publishing Service: Canberra.

Introduction to distribution modeling

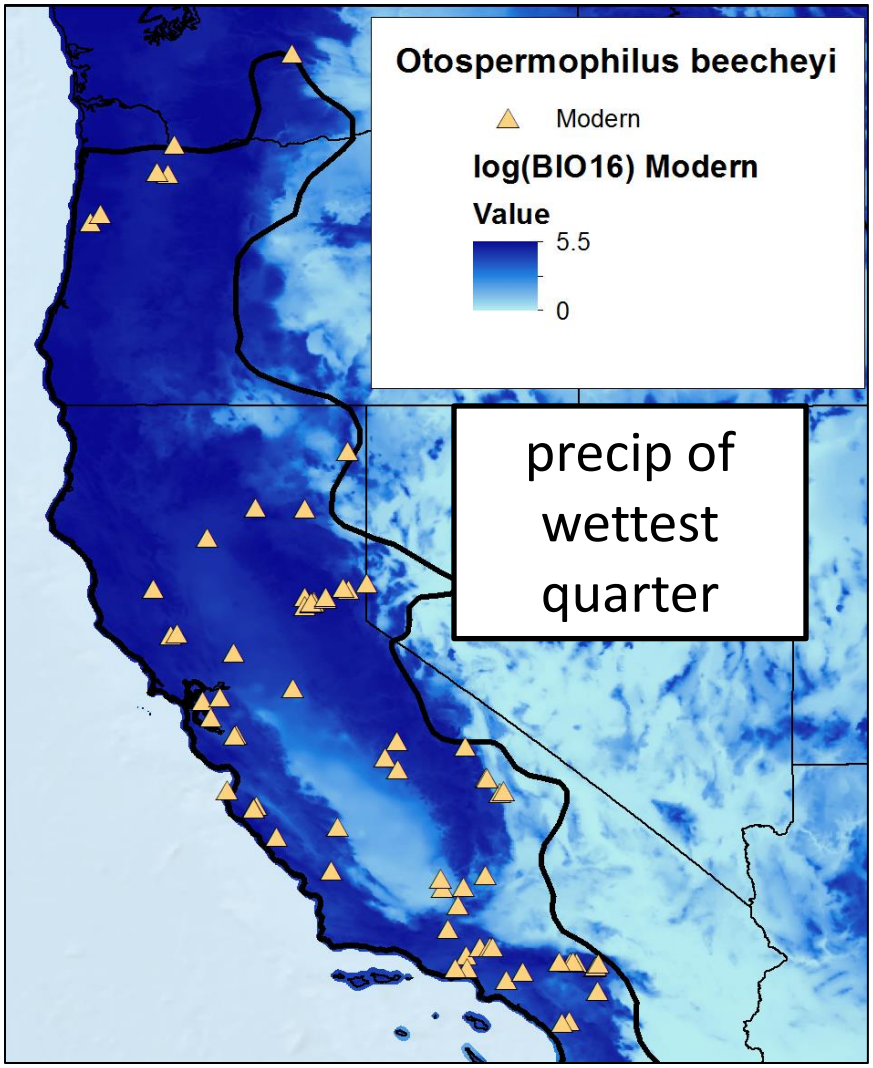
Predictor layers for Exercise I

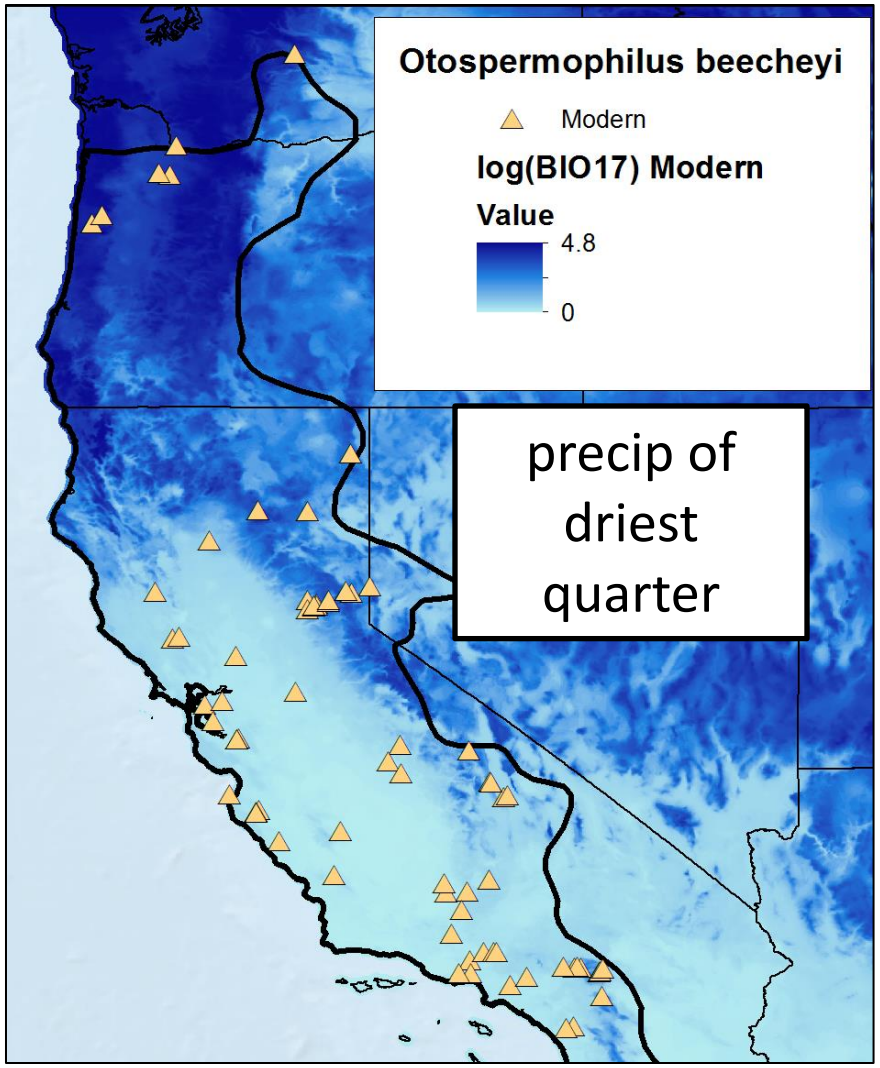




Introduction to distribution modeling

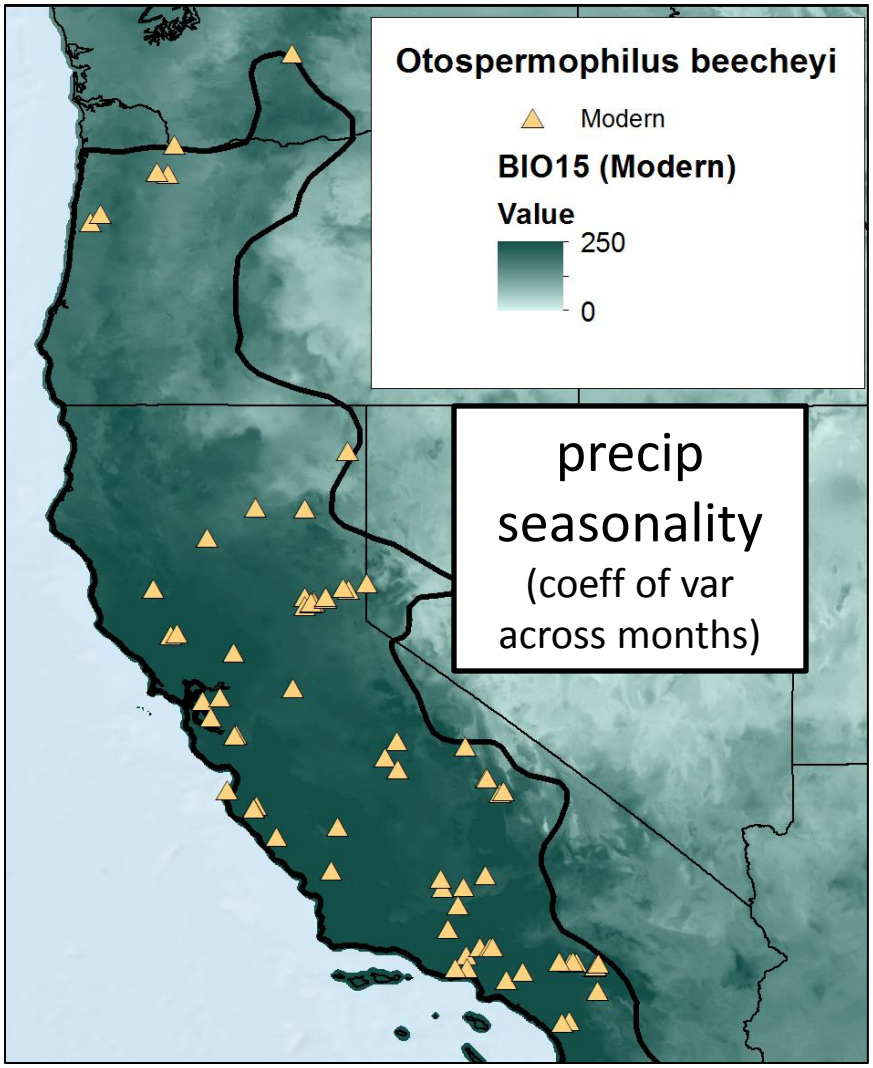
Predictor layers for Exercise I





Introduction to distribution modeling

Predictor layers for Exercise I



A short course on distribution modeling

Outline I

Introduction to distribution modeling
What does a SDM do?
Model algorithms
Input: Species' records
Input: Predictors

Exercise I: Maxent
Input: Species' records, predictor rasters
Output: Maps, thresholds, AUC, variable importance, response functions

Maxent under the hood:
Information entropy maximization
"Feature" functions
Regularization and beta parameter
Assumptions

Exercise II: Maxent with SWD format, k-folds, and projecting to new era/region
Input: SWD format
K-fold data splitting
Projection to new era/region

Best practices: "Truth" vs. "useful bias"
SDM assumptions

Best practices: Training records
Data cleaning
Coordinate uncertainty
Number of records

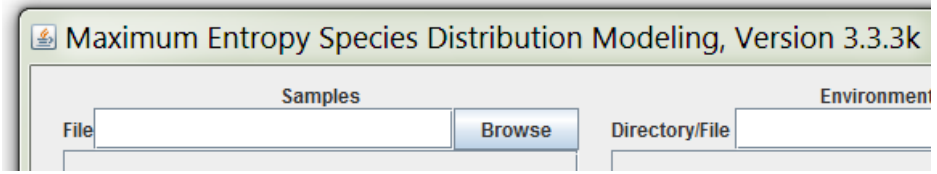
Best practices: Predictors
Proximate & direct, conditions & resources
Types
Resolution
Accuracy
Correlations
Gradients
Dynamic vs. static vs. dynamic-but-static

Exercise I: Maxent

Purpose

- Demonstrate Maxent software and how to run Maxent in R
- Illustrate input data
- Project range of species to new time periods or regions
- Interpret output

Windows (Maxent software)



R using dismo and raster packages

All R code is in "Exercise I - Basic Maxent.r" in the folder "sdmShortCourse_kState."

Exercise I: Maxent

Species' records

1 **species' records**

Each row represents a site where the species is found.

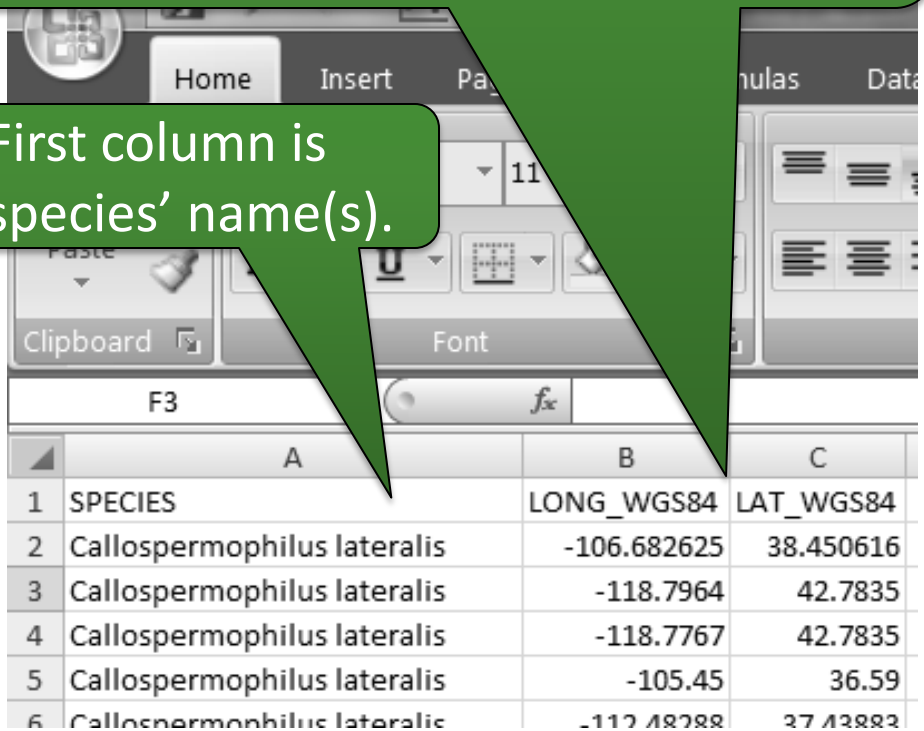
Inside "speciesRecords" folder:

- groundSquirrels_training_historic_swd
- groundSquirrels_training_historic_xy
- groundSquirrels_training_modern_swd
- groundSquirrels_training_modern_xy
- microtusCaliforicus_testData_historic_swd
- microtusCaliforicus_testData_modern_swd
- microtusSpp_trainingData_historic_swd
- microtusSpp_trainingData_historic_swd_10
- microtusSpp_trainingData_historic_swd_40
- microtusSpp_trainingData_historic_swd_80
- microtusSpp_trainingData_historic_swd_120

CSV format (not Excel)

Longitude and latitude columns can have any name but they must be the 2nd and 3rd columns, and longitude must be before latitude!

First column is species' name(s).



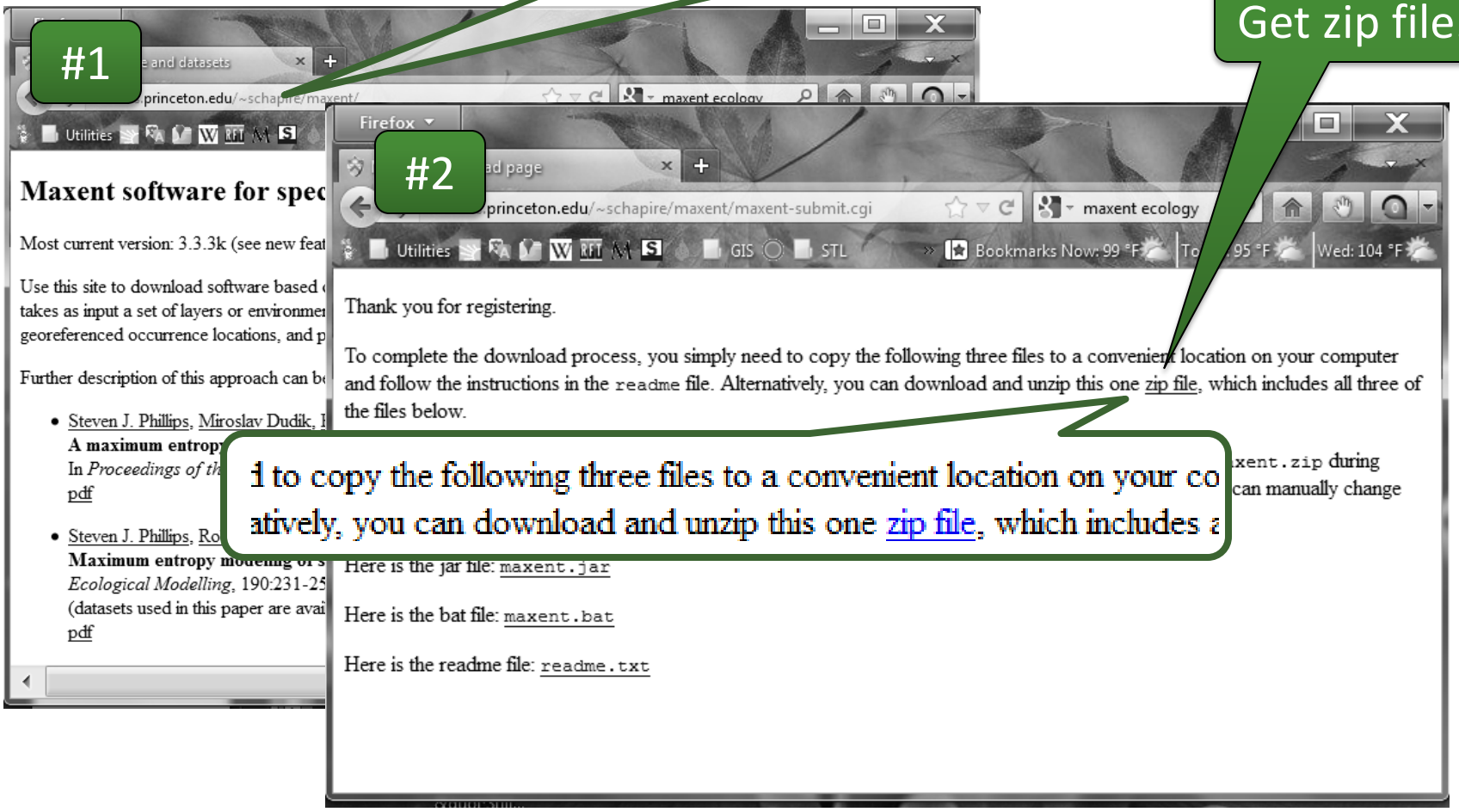
	A	B	C
1	SPECIES	LONG_WGS84	LAT_WGS84
2	Callospermophilus lateralis	-106.682625	38.450616
3	Callospermophilus lateralis	-118.7964	42.7835
4	Callospermophilus lateralis	-118.7767	42.7835
5	Callospermophilus lateralis	-105.45	36.59
6	Callospermophilus lateralis	-112.48288	37.43883

Exercise I: Maxent

Getting Maxent on your computer

2 **download MX**

http://www.cs.princeton.edu/~schapire/Maxent/

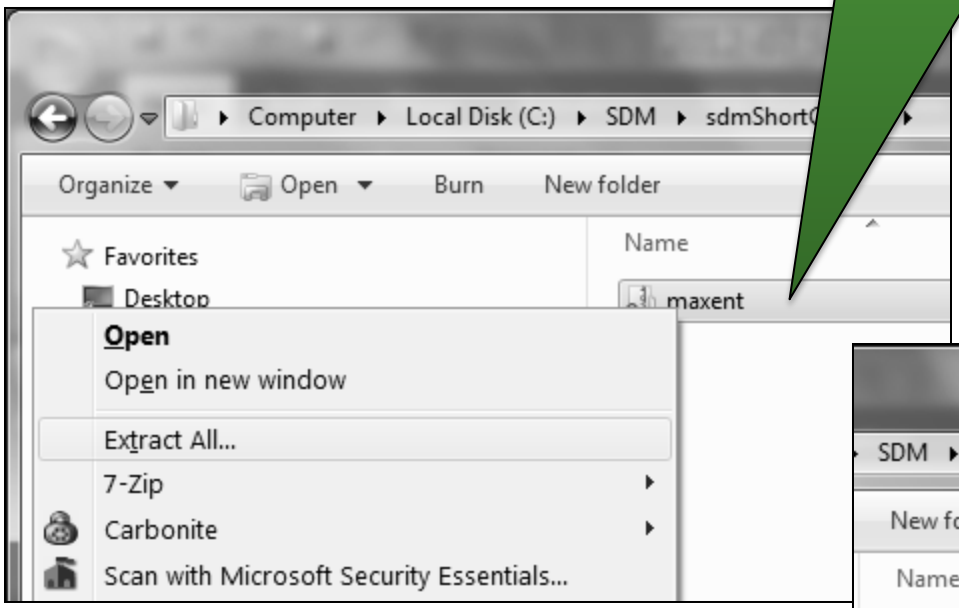


Exercise I: Maxent

Starting Maxent

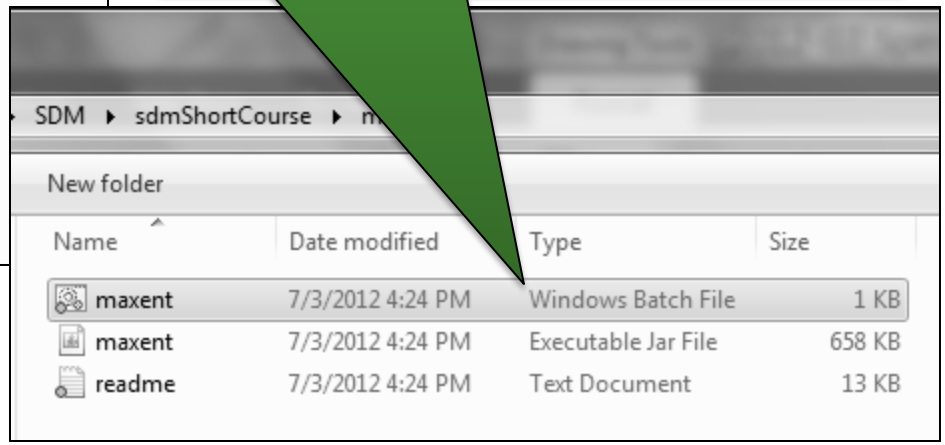
3 **decompress zip file**

Right-click zip file and select “Extract All…”



4 **start Maxent**

Double-click the Maxent “batch” file (not the “Jar” file).



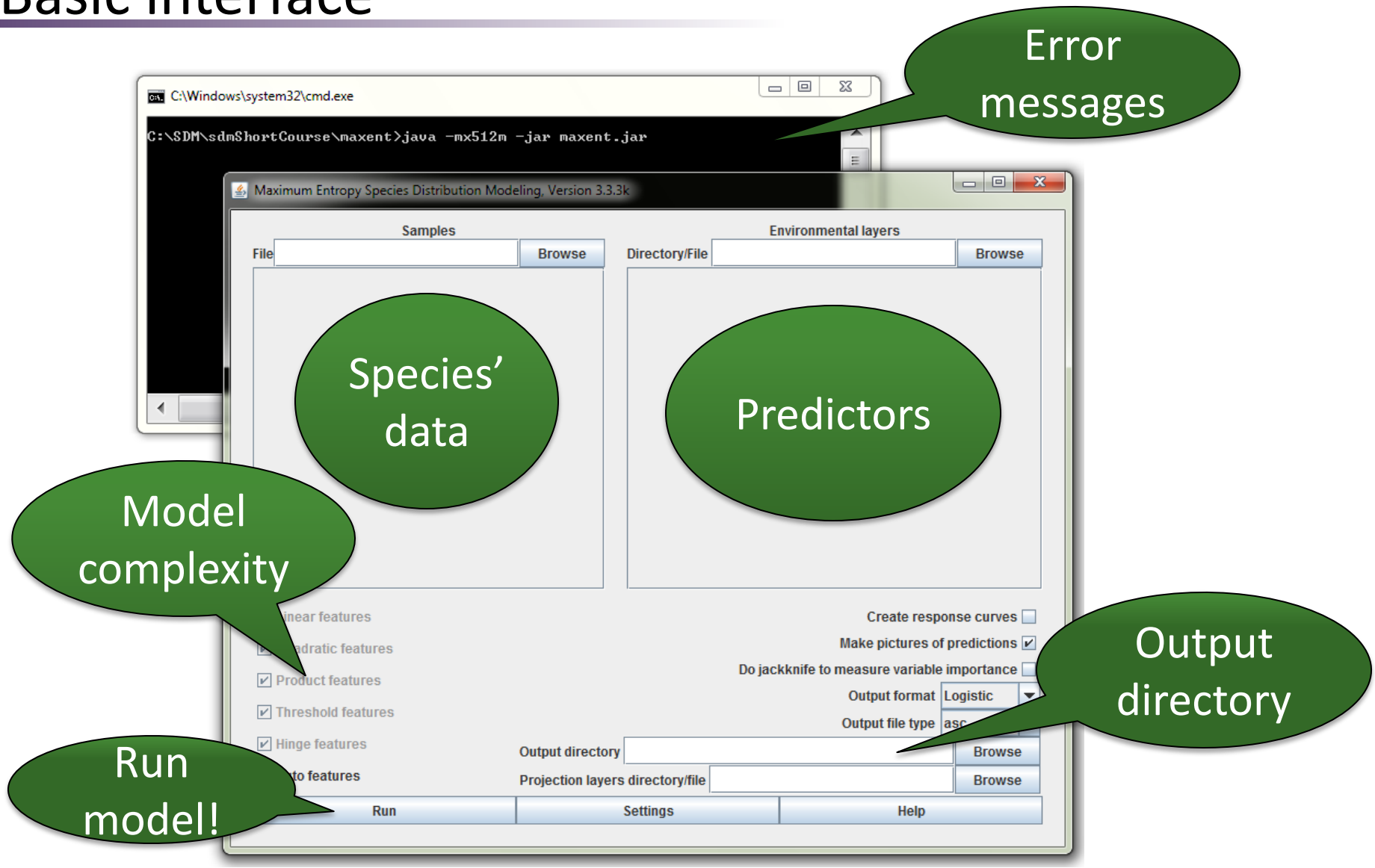
TIP: You can run Maxent with more memory by opening the batch file (with a text editor like Notepad) and changing the value after “-mx” to a larger number (don’t go above ~1300 on Windows machines, though):

java -mx512m -jar Maxent.jar

TIP: Maxent may still work if you double-click the “jar” file but may also crash due to memory problems.

Exercise I: Maxent

Basic interface



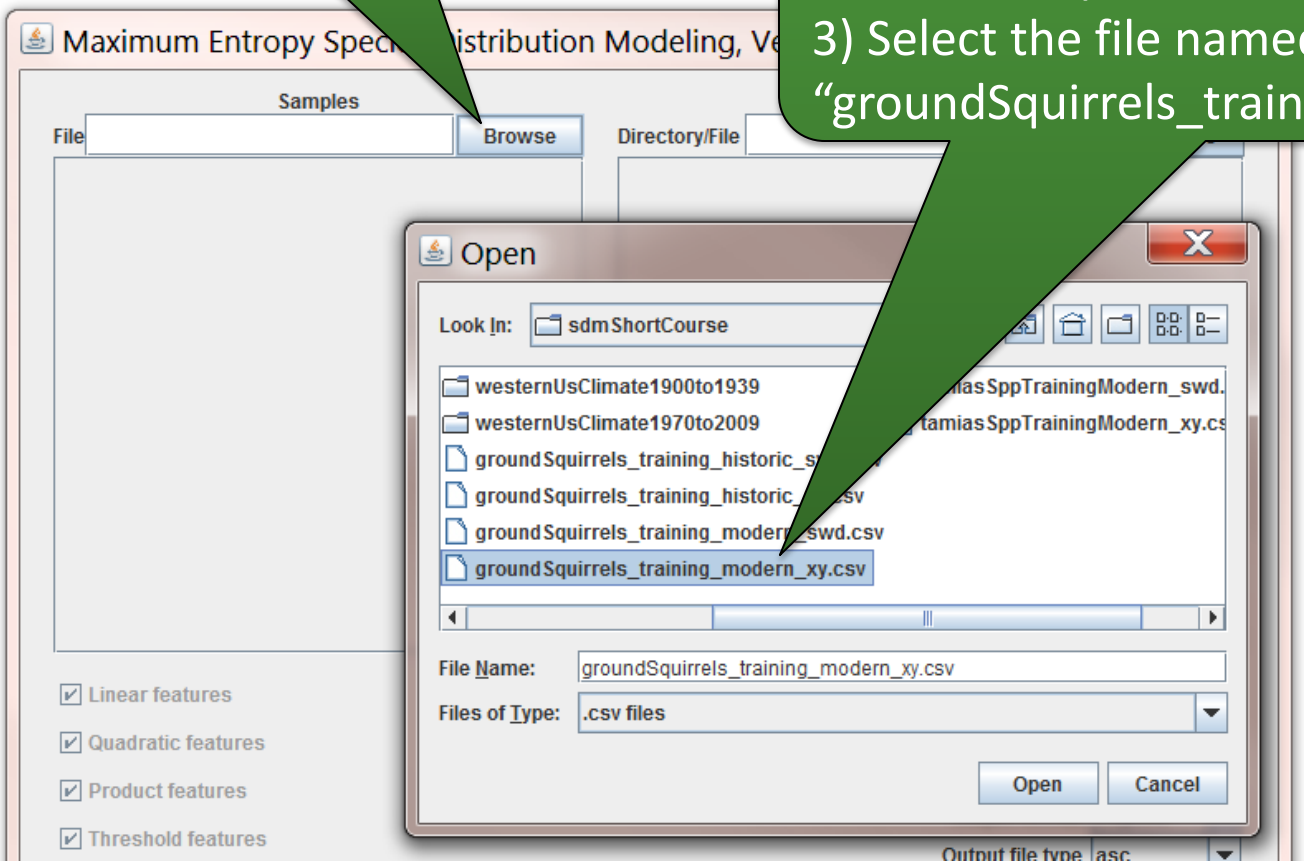
Exercise I: Maxent
Loading species' data

5 **load species' data**

1) Click "Browse"…

2) Navigate to SDM course folder then to the folder "speciesRecords."
3) Select the file named "groundSquirrels_training_modern_**xy**.csv".

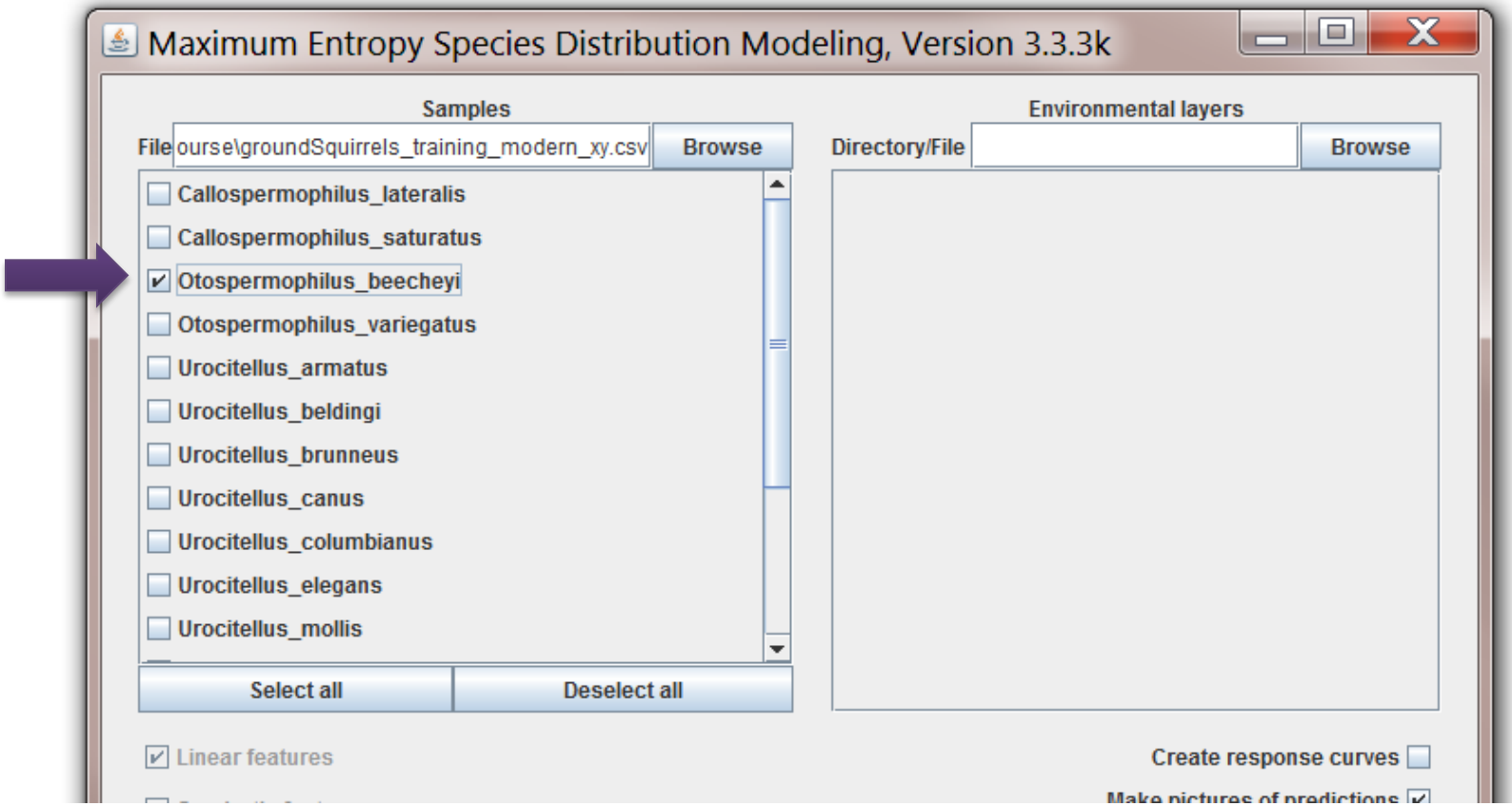
Not "swd"!



Exercise I: Maxent

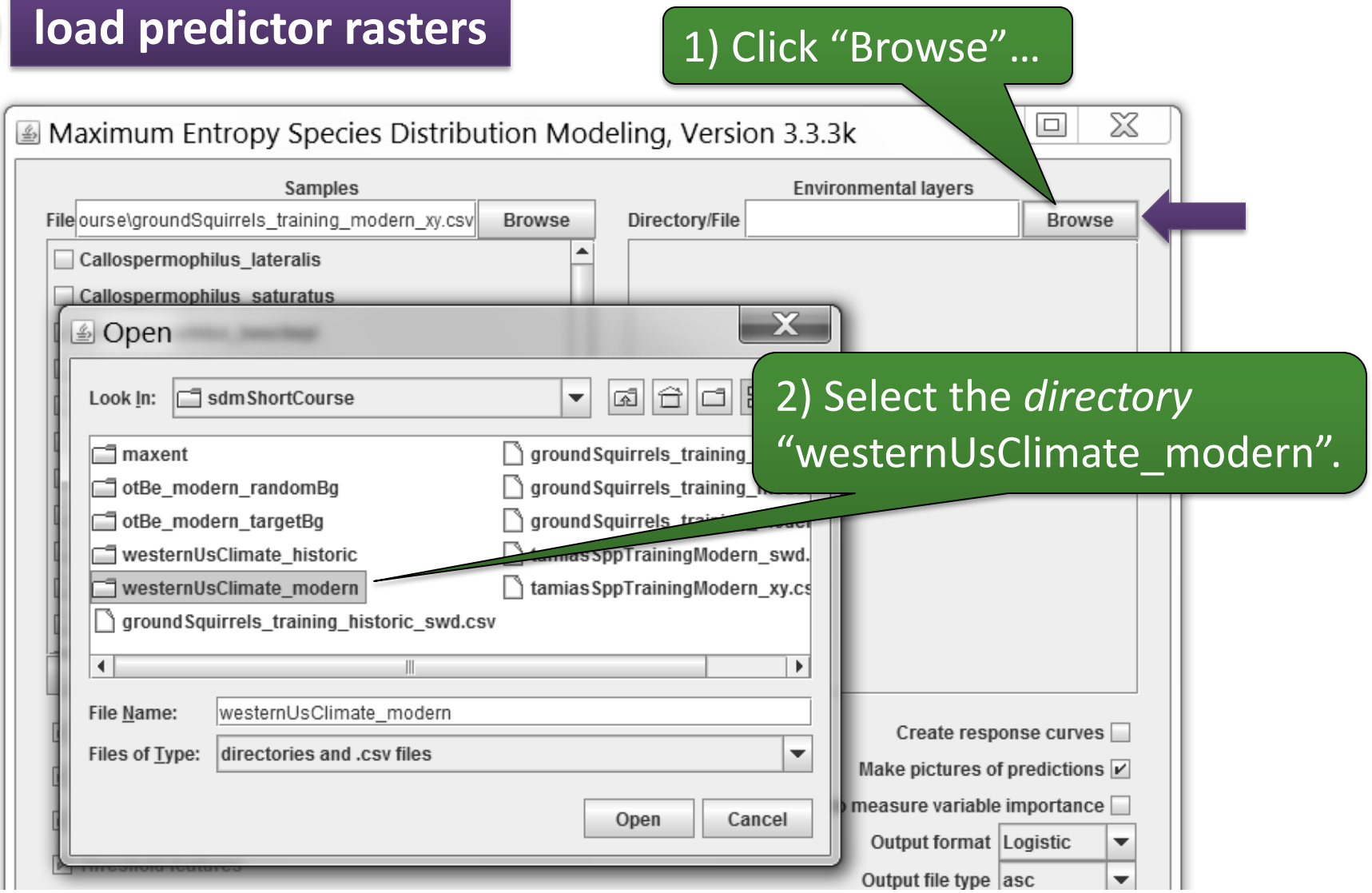
Loading species' data

6 **select only "Otospermophilus_beecheyi"**



Exercise I: Maxent
Loading predictor grids

7 **load predictor rasters**

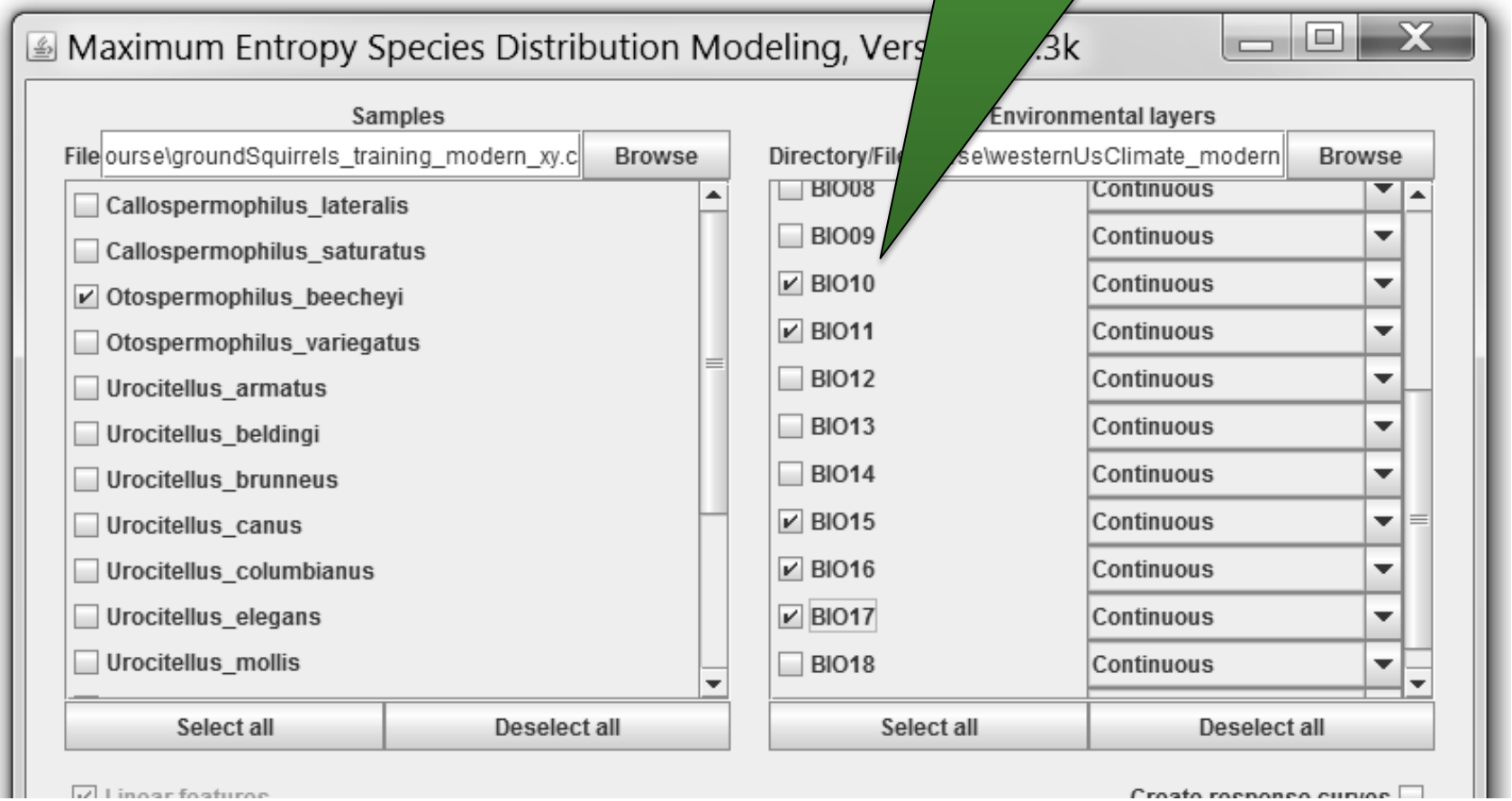


Exercise I: Maxent

Loading predictor grids

8 **select predictor rasters**

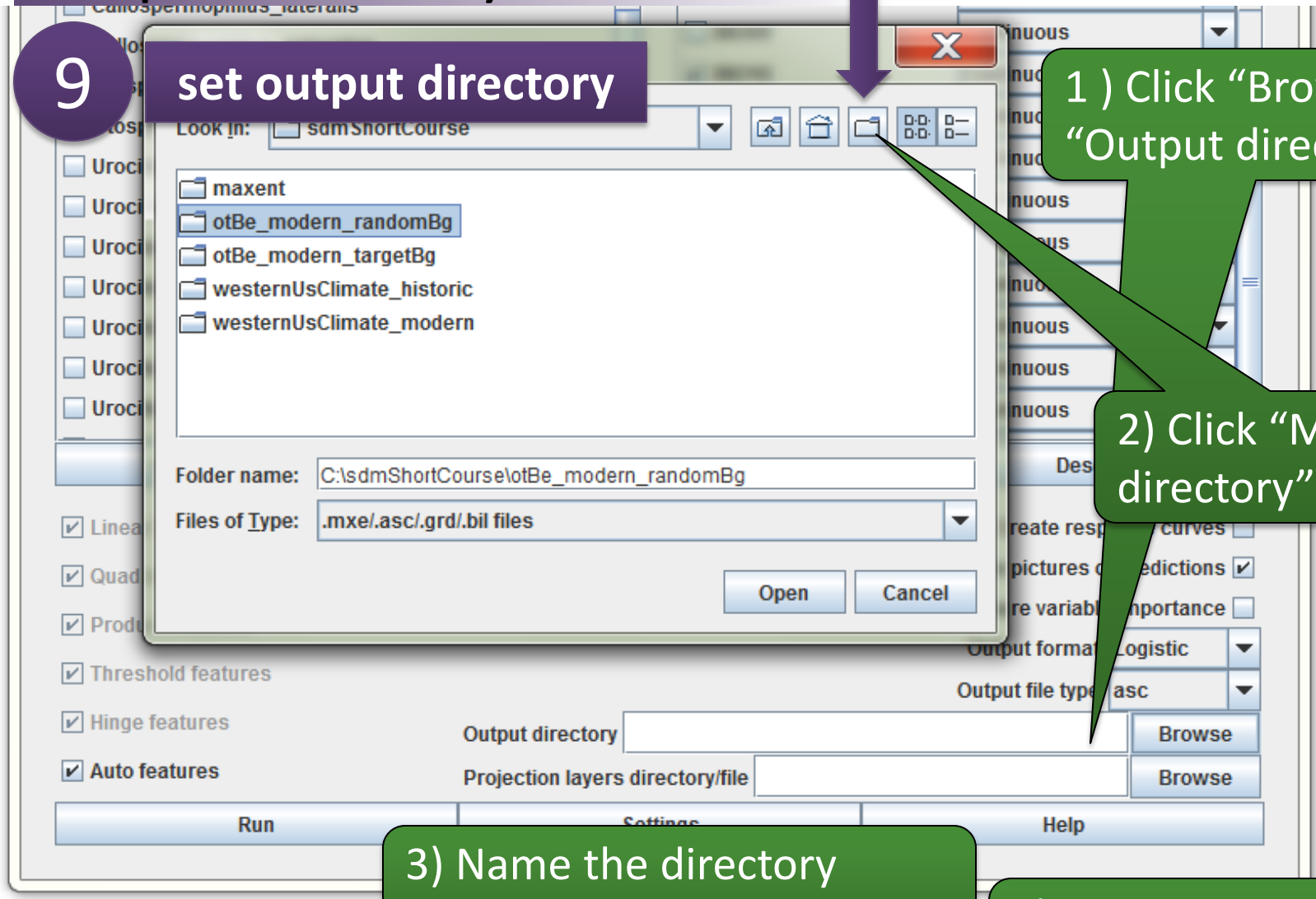
Select BIO10, BIO11, BIO15, BIO16, and BIO17.



Why these predictors? We will cover selection of predictors later in the course.

Exercise I: Maxent

Output directory



Exercise II: Advanced Maxent

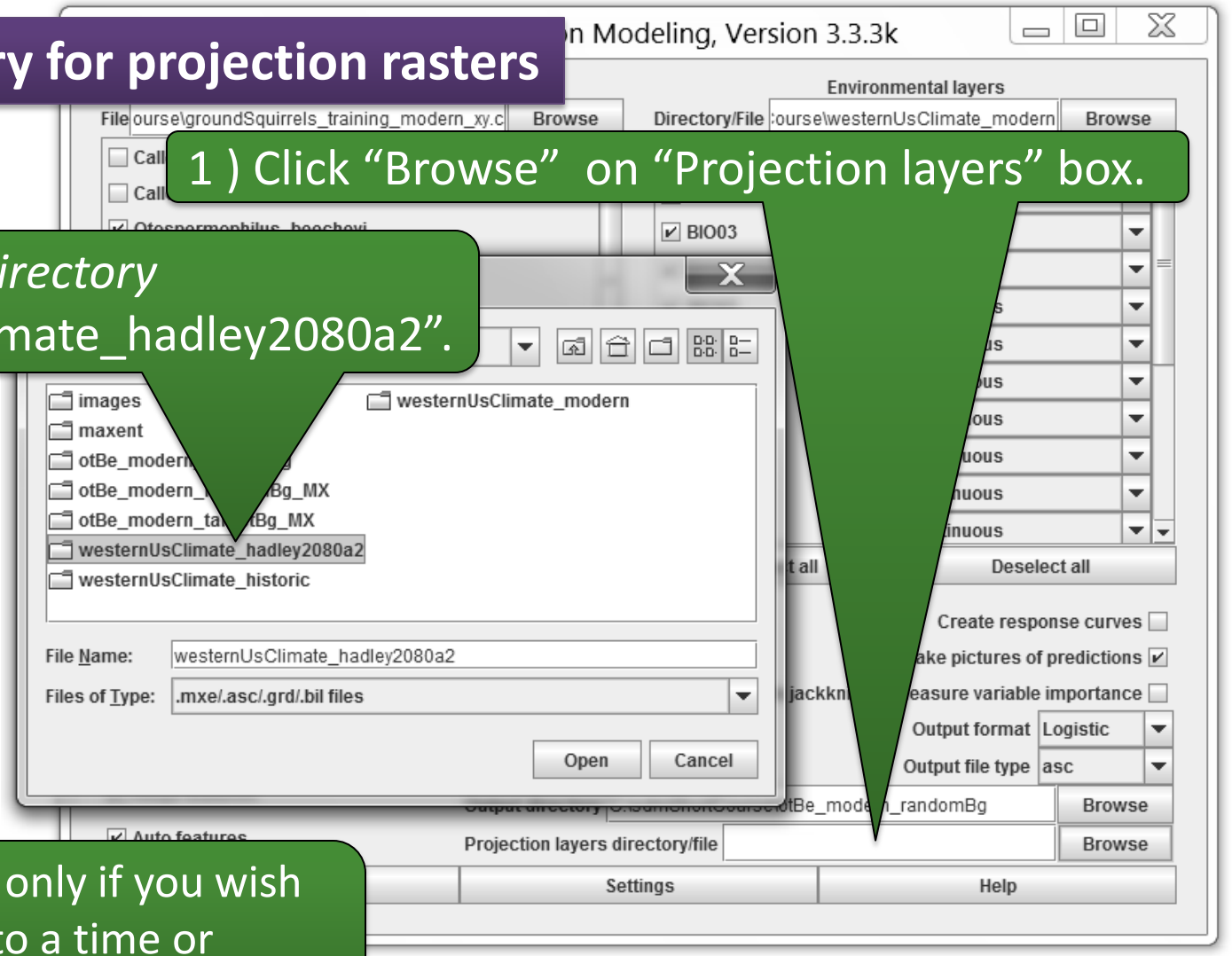
Projecting to a new time period

10 **set directory for projection rasters**

1) Click “Browse” on “Projection layers” box.

2) Select the *directory* “westernUsClimate_hadley2080a2”.

This directory contains rasters for an A2 emissions scenario centered on the year 2080 from the Hadley GCM.



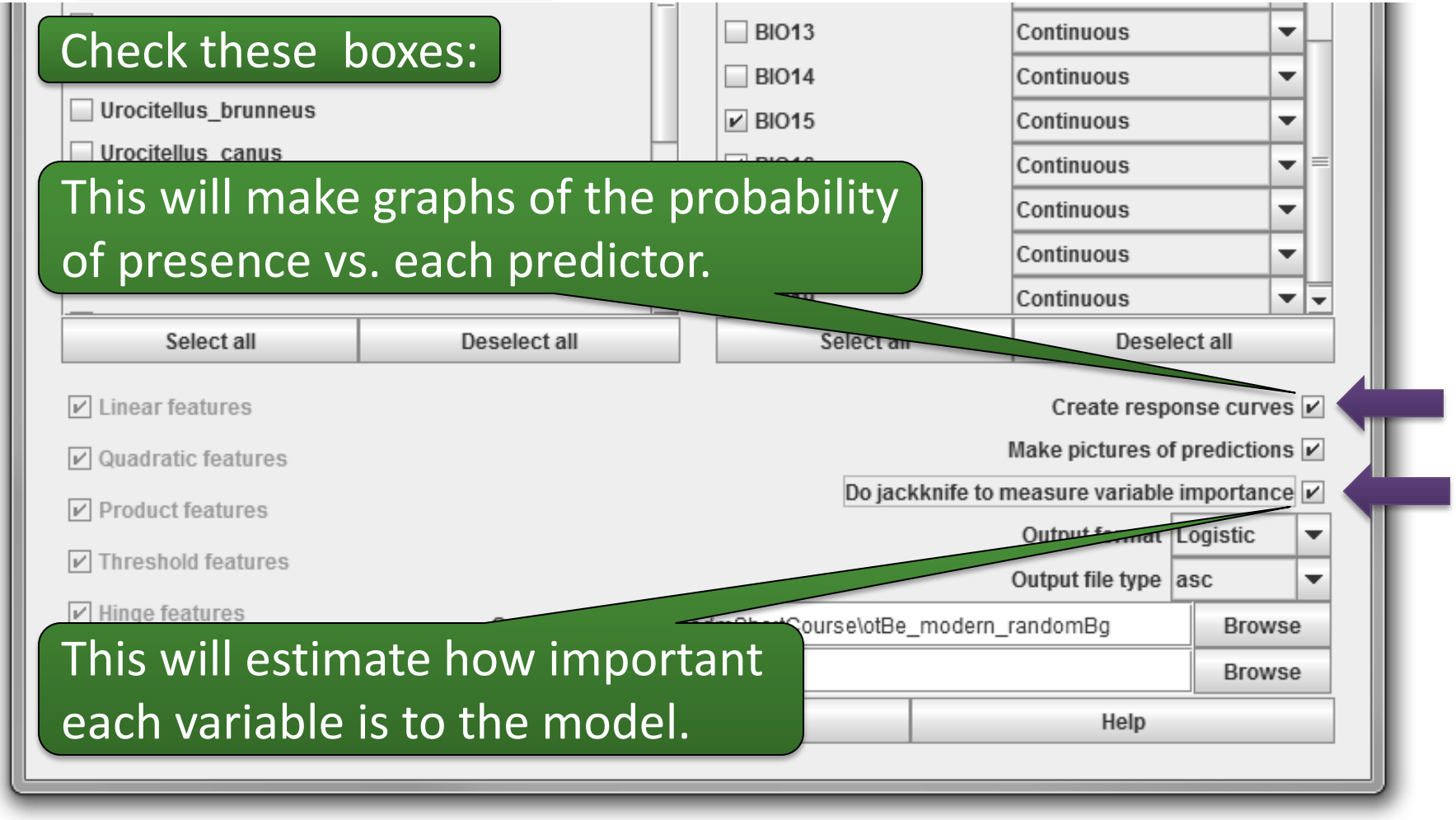
This step is necessary only if you wish to project the model to a time or region different from the training area.

Exercise I: Maxent

Options

11

response curves & variable importance

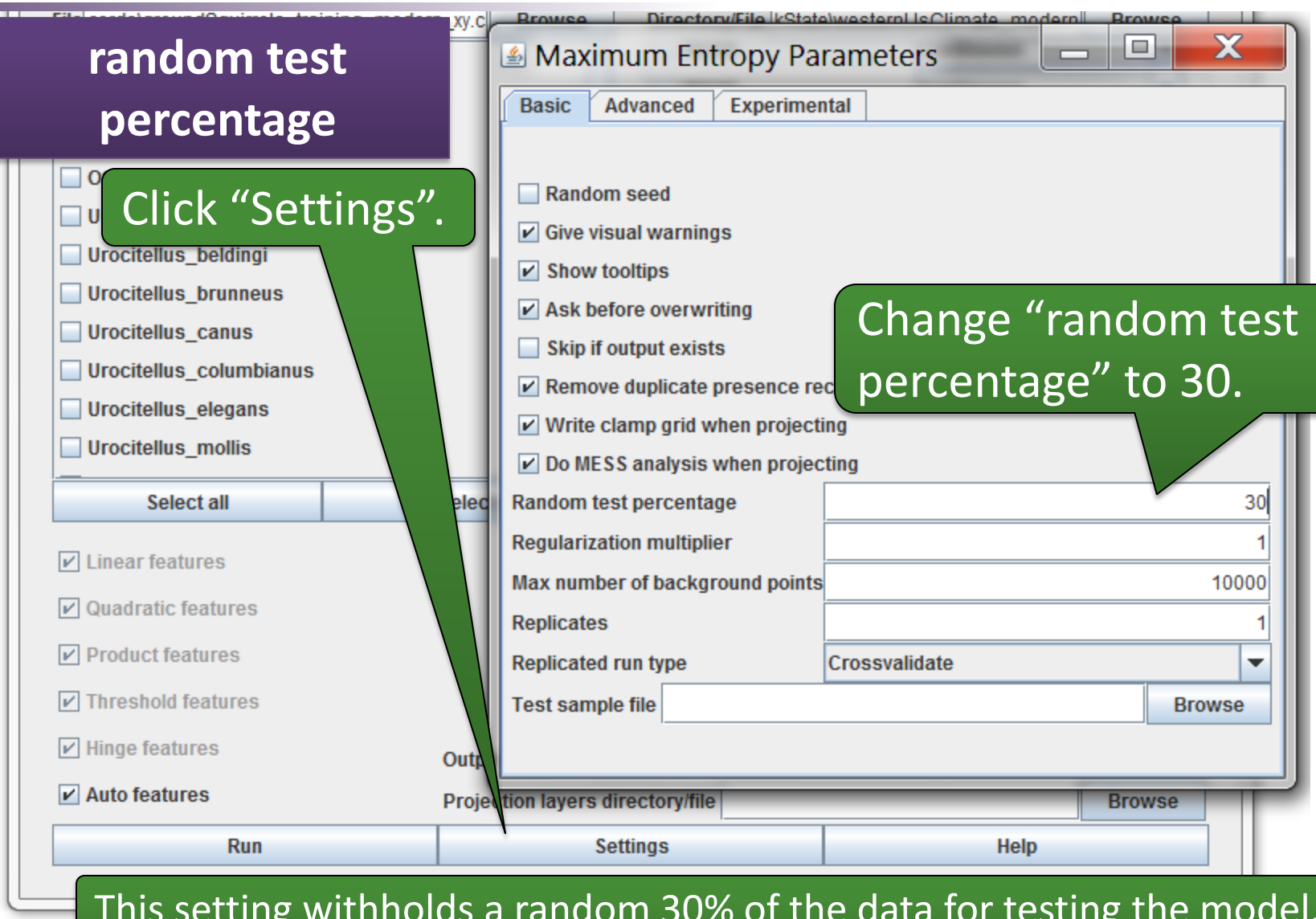


Exercise I: Maxent

Options

12

random test percentage



This setting withholds a random 30% of the data for testing the model. The remaining 70%—a typical value—is used for training the model.

Exercise I: Maxent

Thresholding

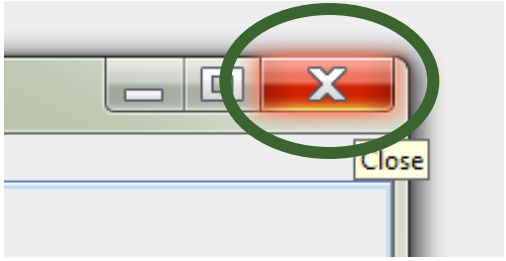
13 **thresholding**

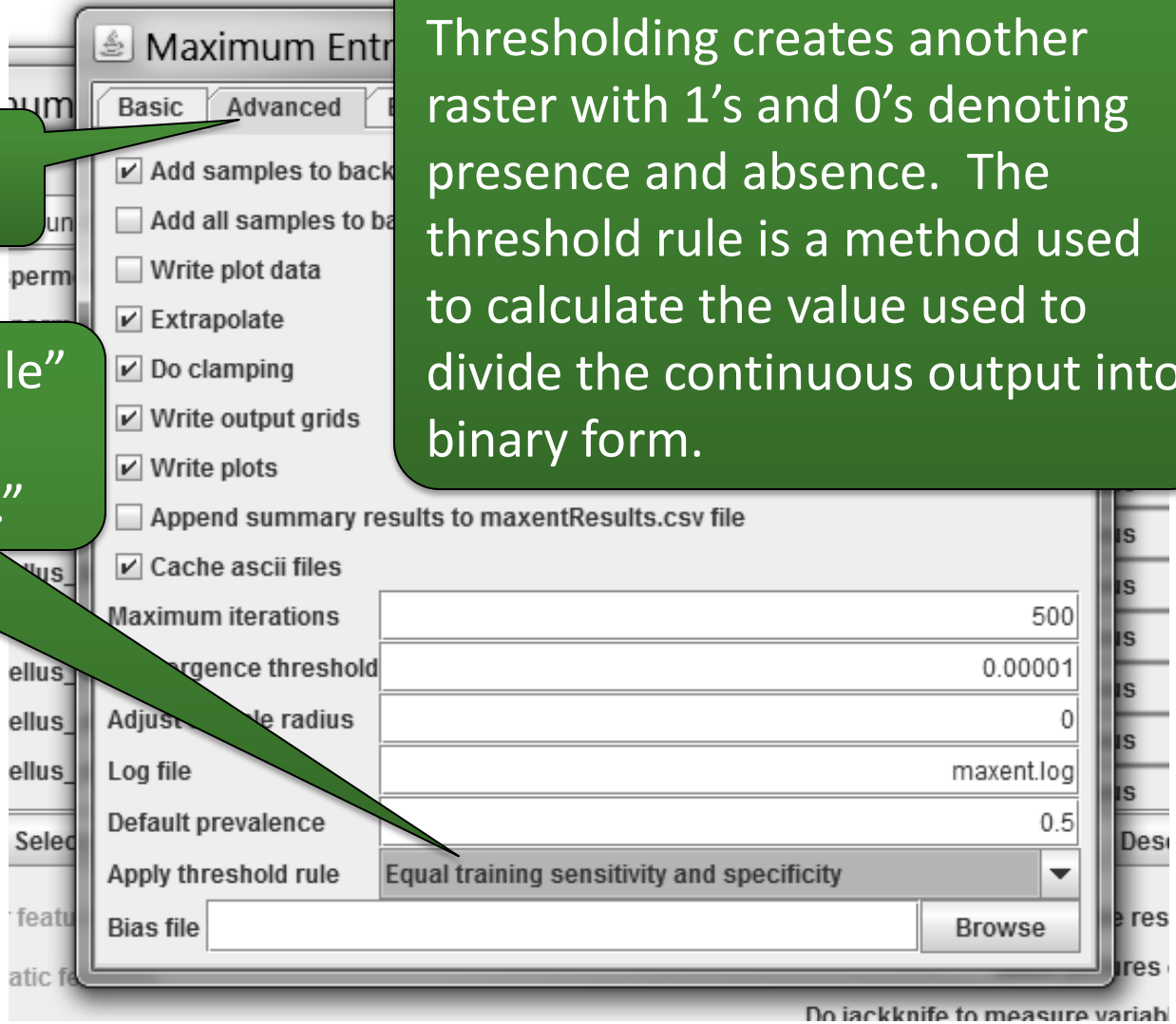
Click the "Advanced" tab to get here.

Set "Apply threshold rule" to "equal training sensitivity & specificity."

Typical model output is a raster with values between 0 and 1. Thresholding creates another raster with 1's and 0's denoting presence and absence. The threshold rule is a method used to calculate the value used to divide the continuous output into binary form.

close settings tab

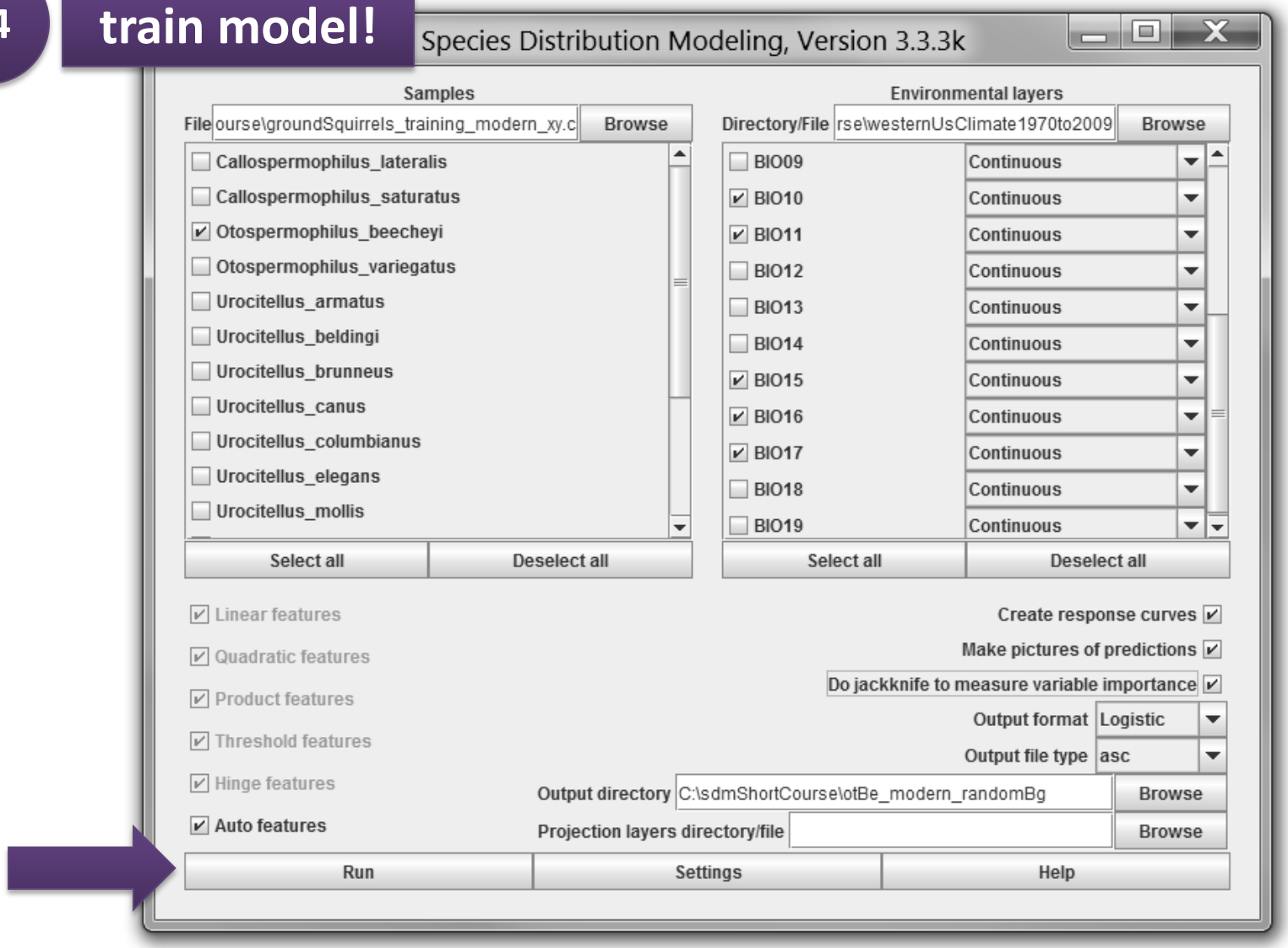




Exercise I: Maxent

Train model!

14 **train model!**



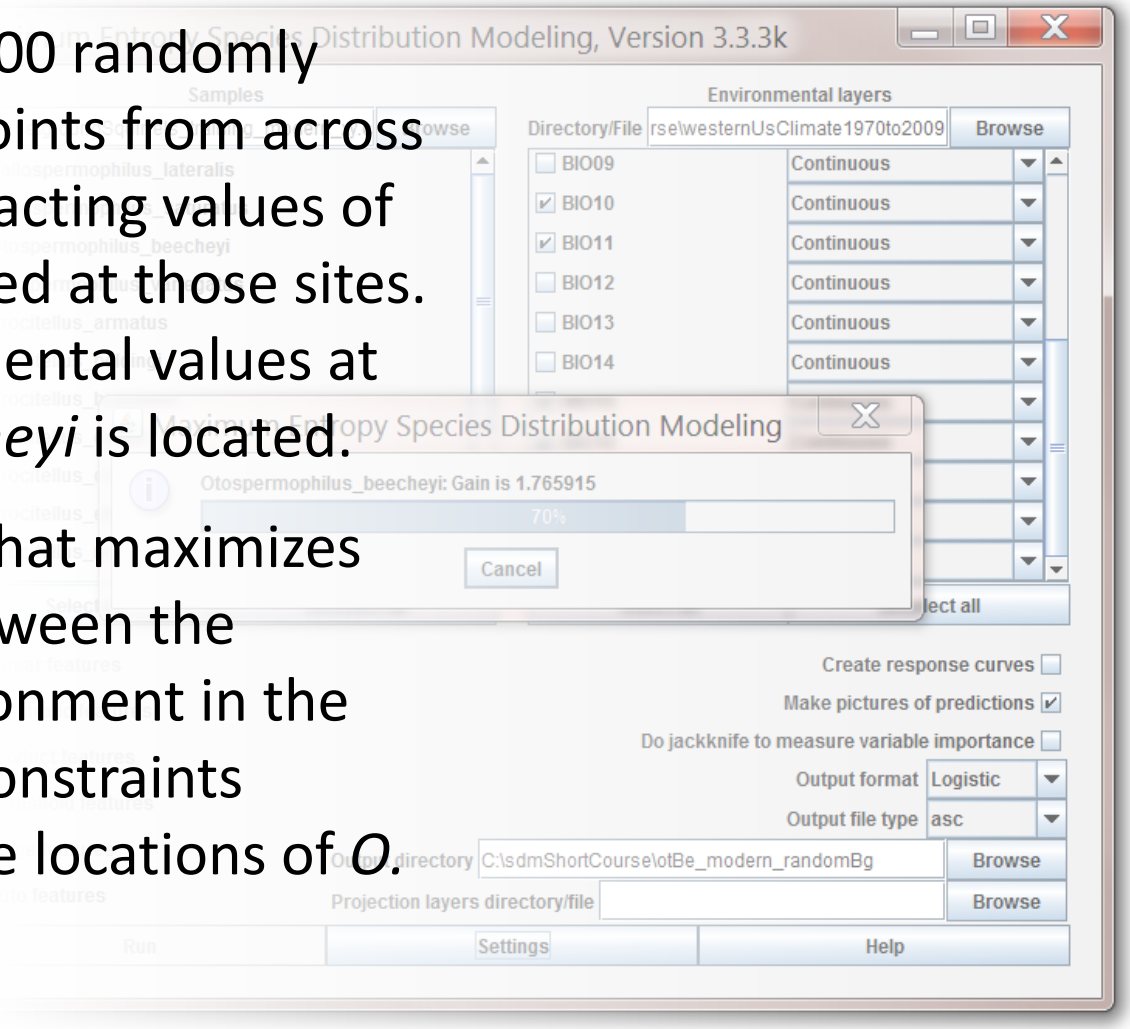
Exercise I: Maxent

What is Maxent doing?

Maxent is sampling 10,000 randomly located "background" points from across the western US and extracting values of the predictors we selected at those sites. It also extracts environmental values at the sites where *O. beecheyi* is located.

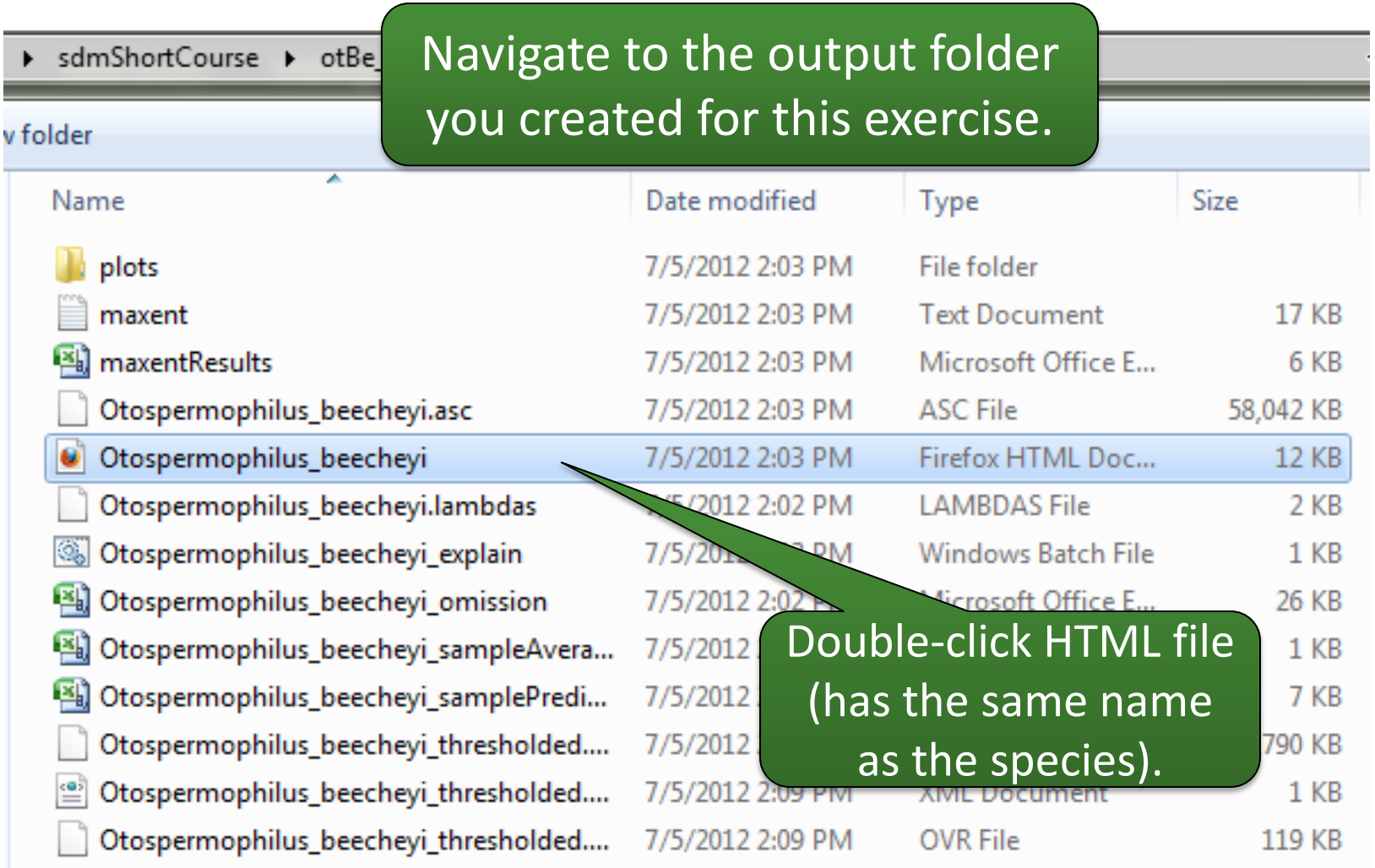
It then finds a function that maximizes information entropy between the distribution of the environment in the background subject to constraints imposed by the presence locations of *O. beecheyi.*

More on this later…



Exercise I: Maxent

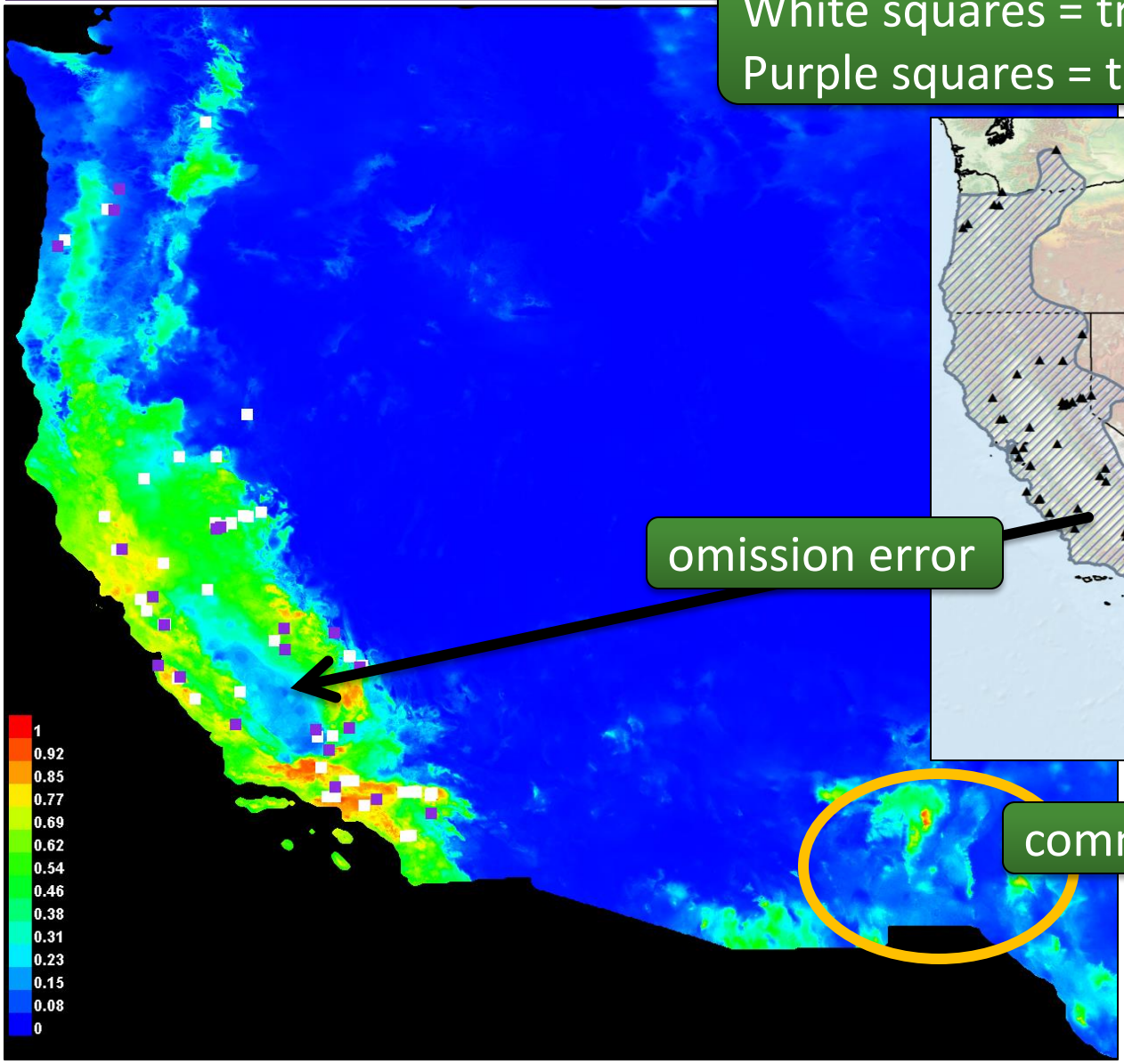
Output

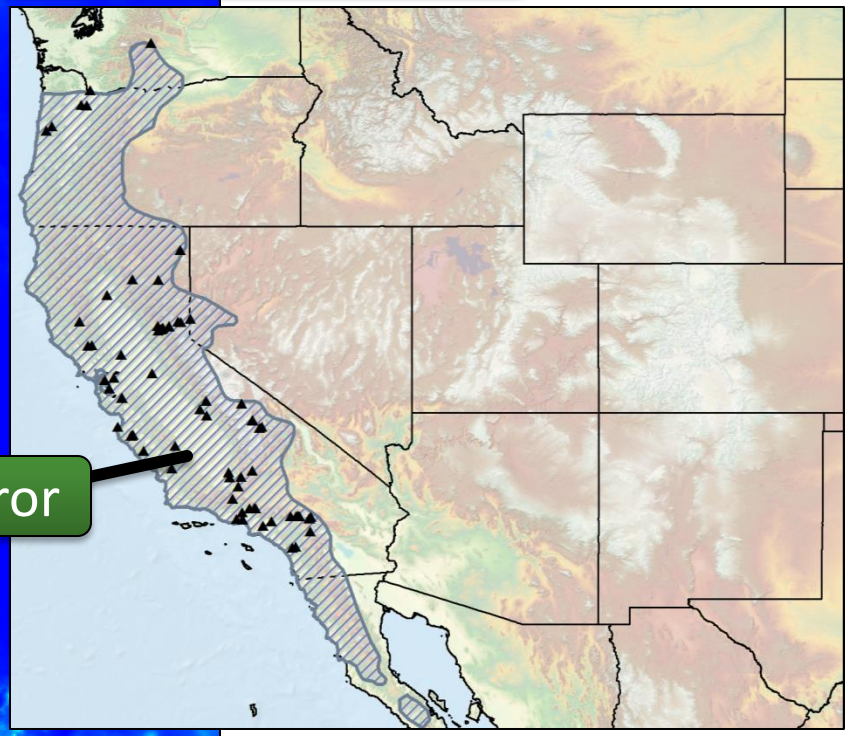


Exercise I: Maxent

Map output: Modern

White squares = training sites
Purple squares = test sites





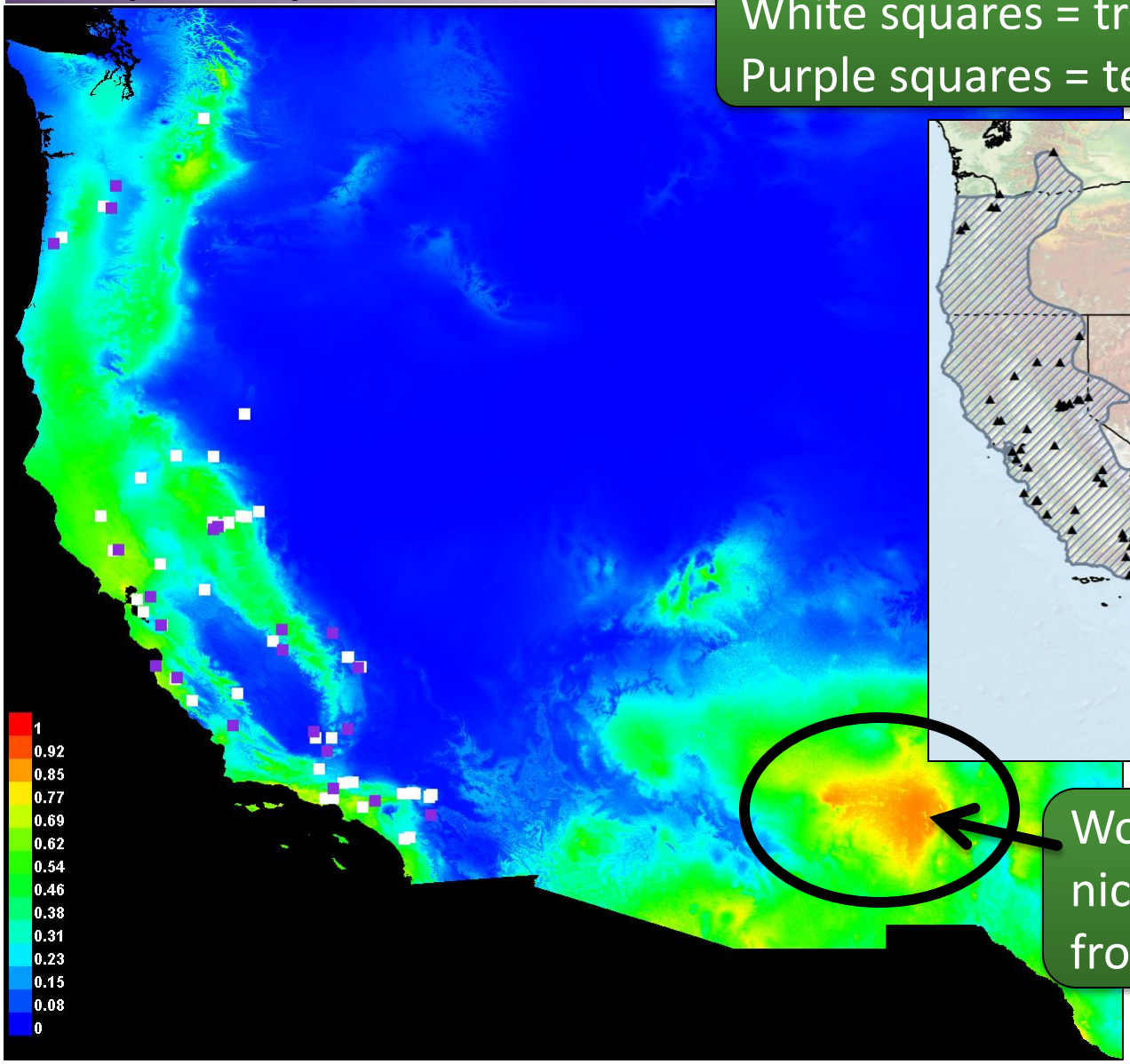
omission error

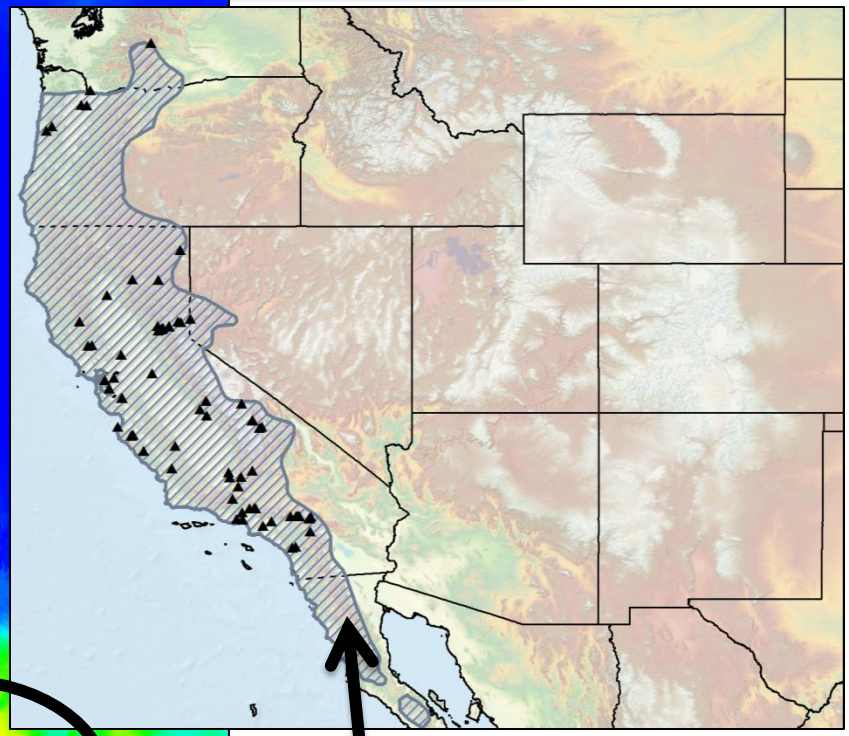
commission error

Exercise I: Maxent

Map output: Future

White squares = training sites
Purple squares = test sites

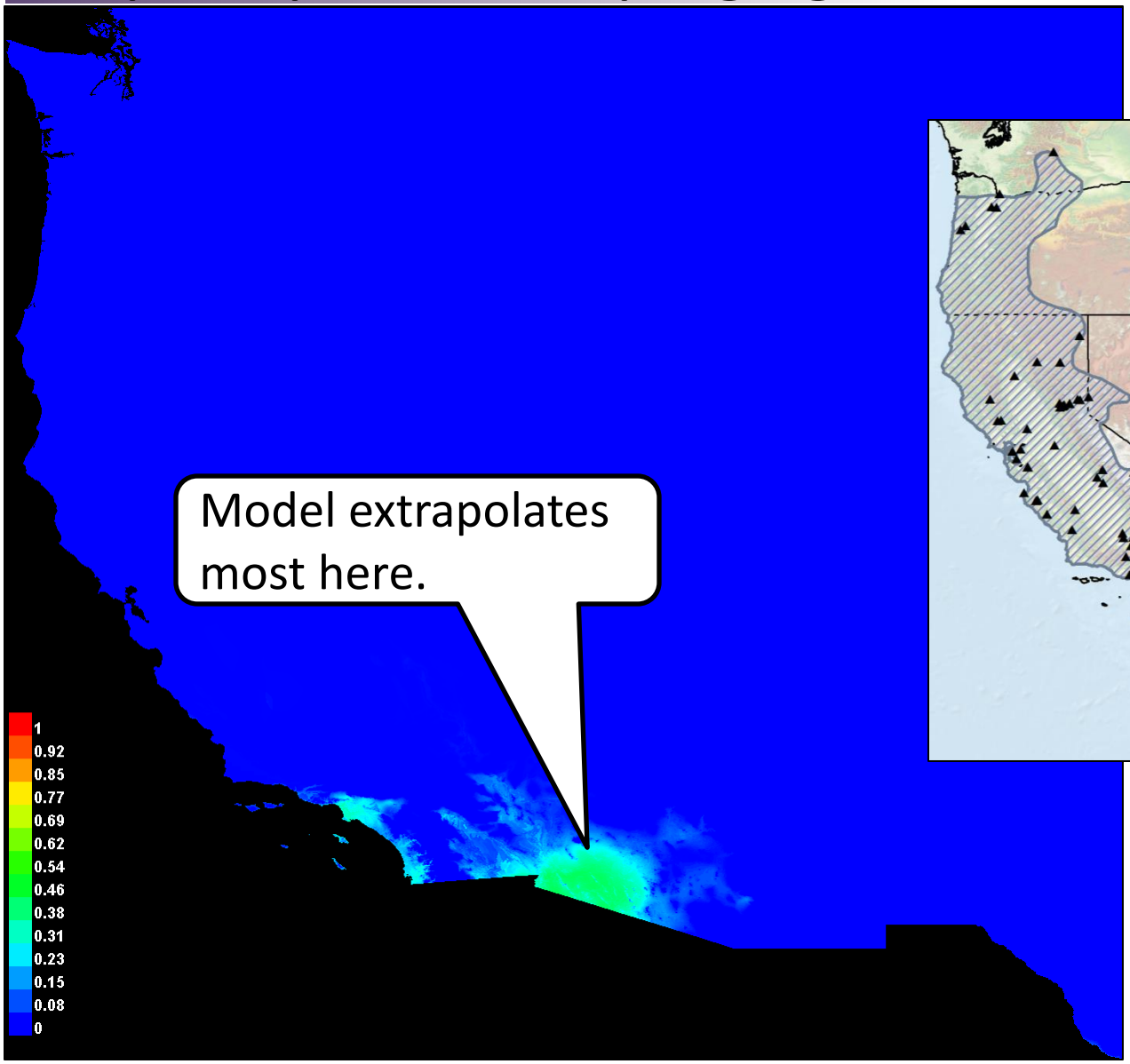


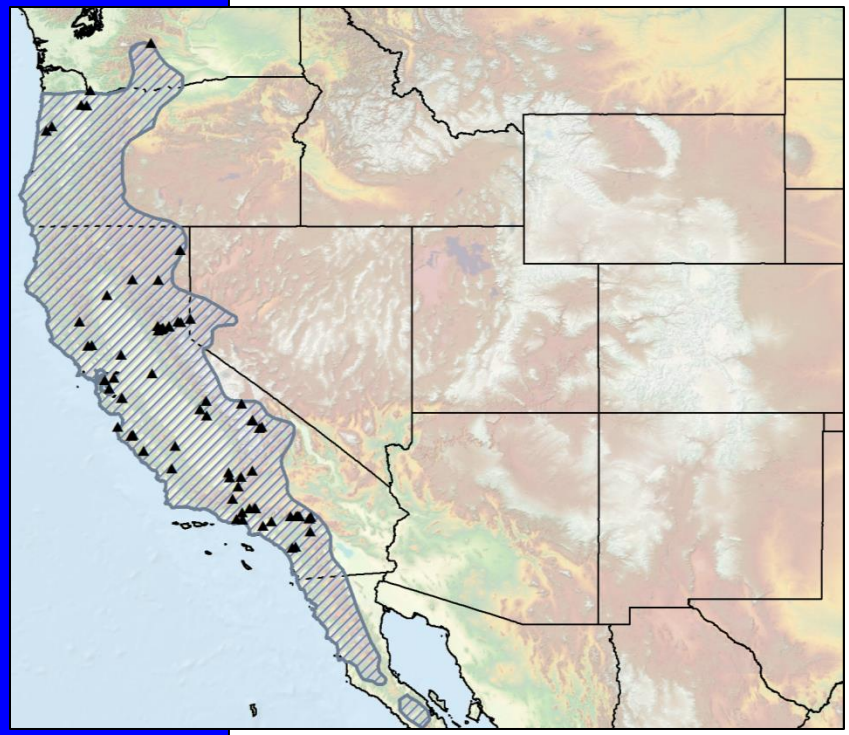


Would have been nice to have data from Baja Mexico…

Exercise I: Maxent

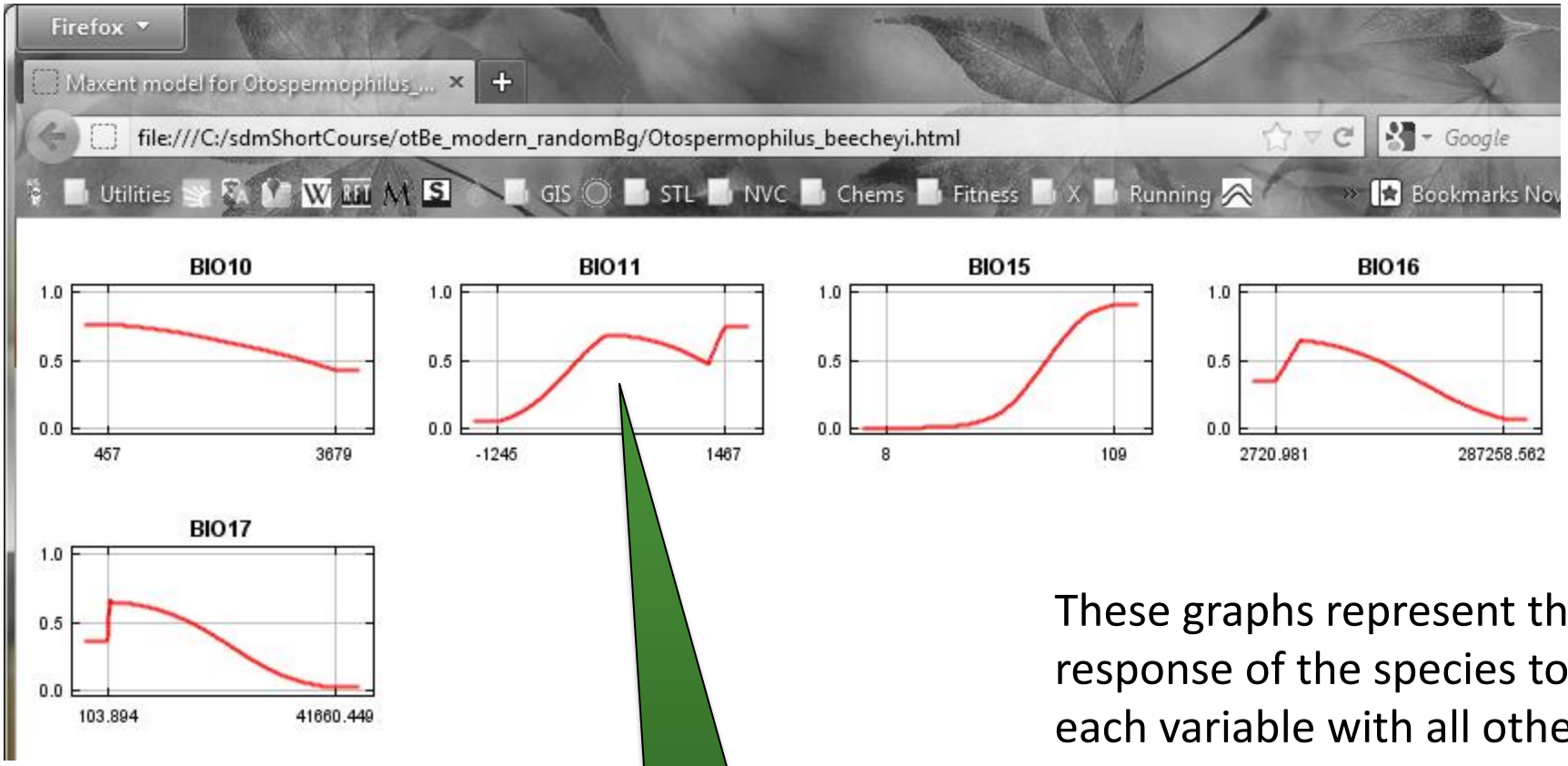
Map output: “Clamping” grid





Exercise I: Maxent

Response function inspection



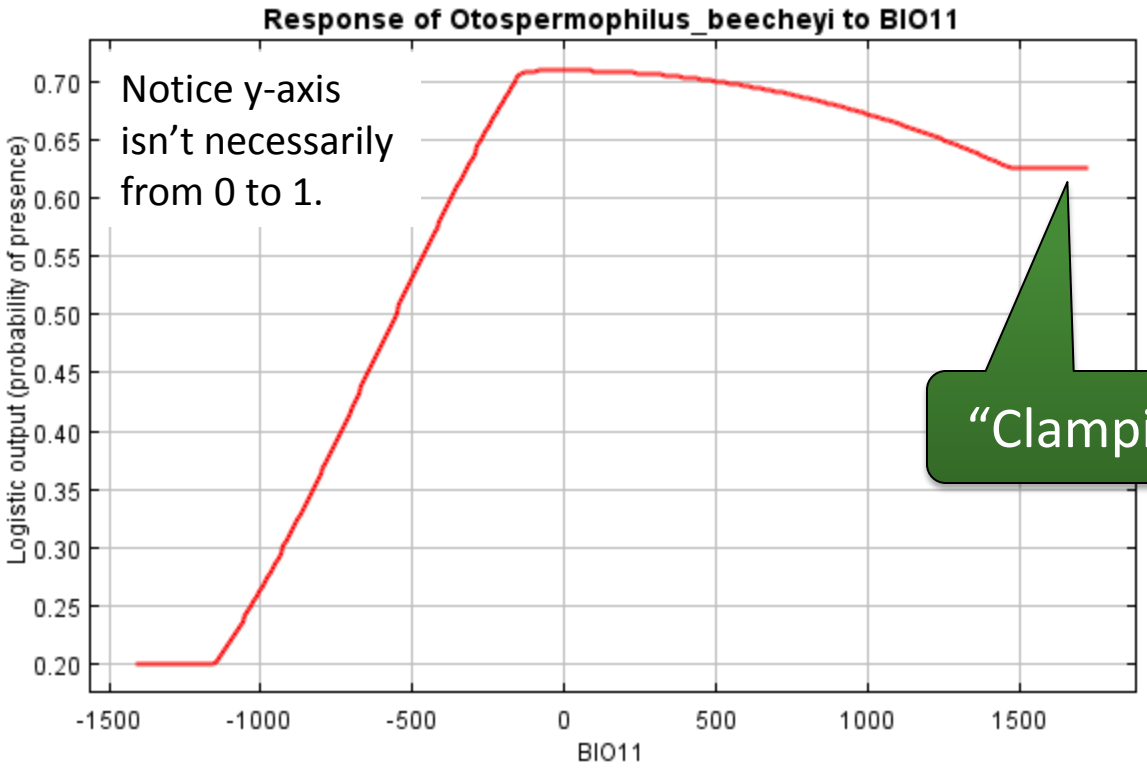
Click “BIO11”.

These graphs represent the response of the species to each variable with all other variables held at their mean value across the species’ training points.

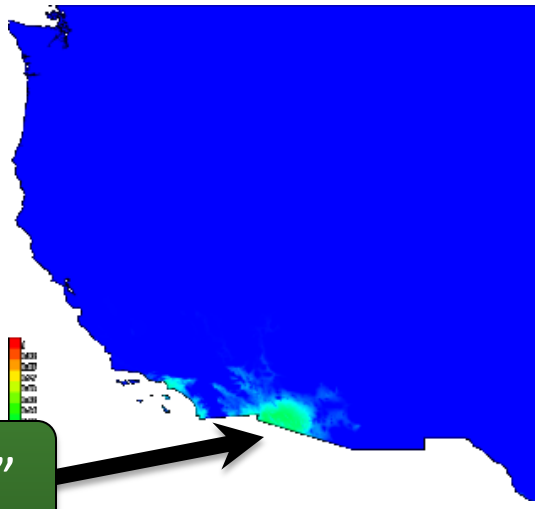
Exercise I: Maxent

Response function inspection

species' response to BIO11
(mean temp of coldest quarter)



clamping map

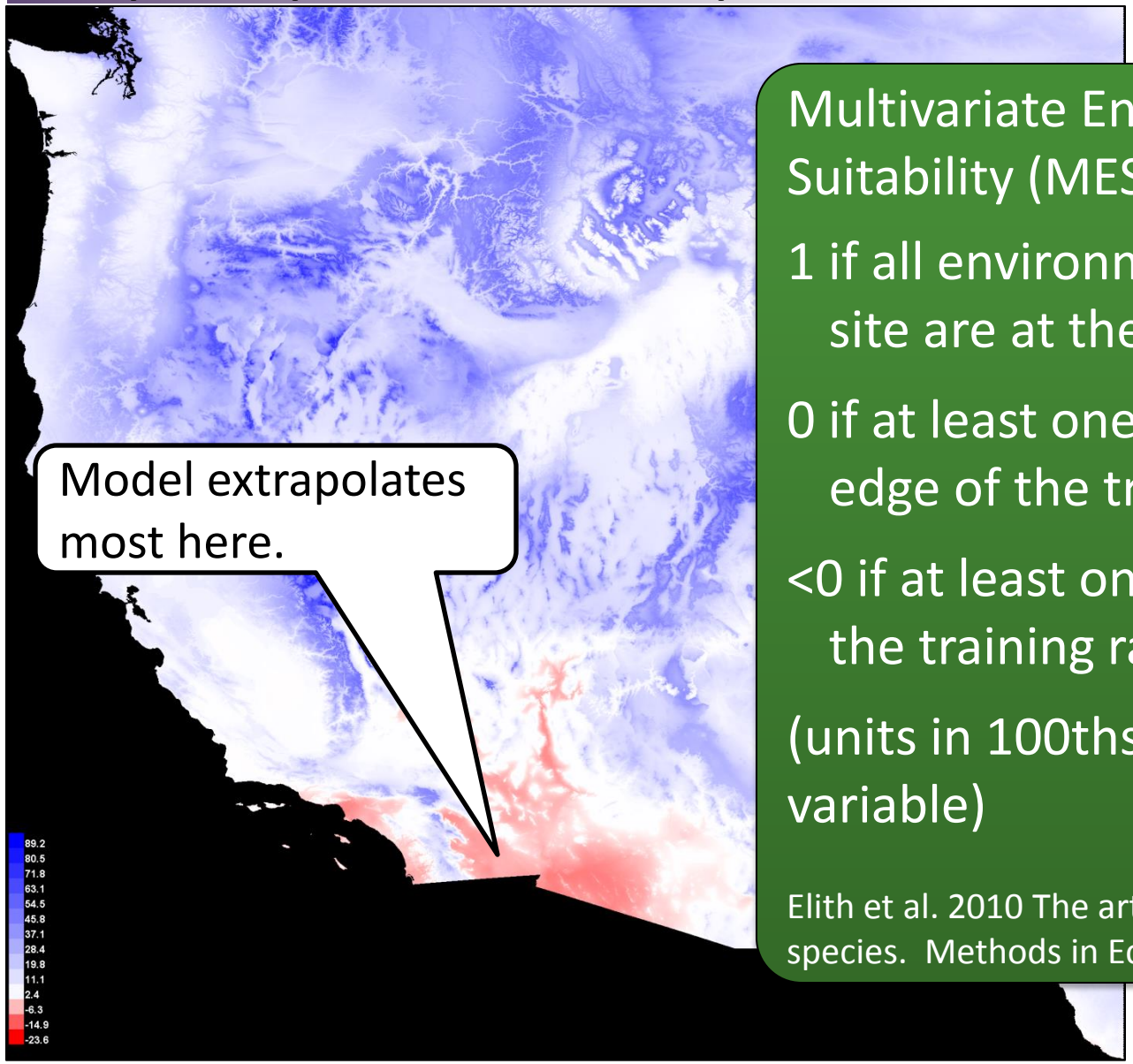


"Clamping"

By default Maxent "clamps" predictions to their last value when predicting beyond the range of training data.

Exercise I: Maxent

Map output: MESS map



Multivariate Environmental Suitability (MES) =

1 if all environmental variables at a site are at their median value

0 if at least one variable is on the edge of the training range

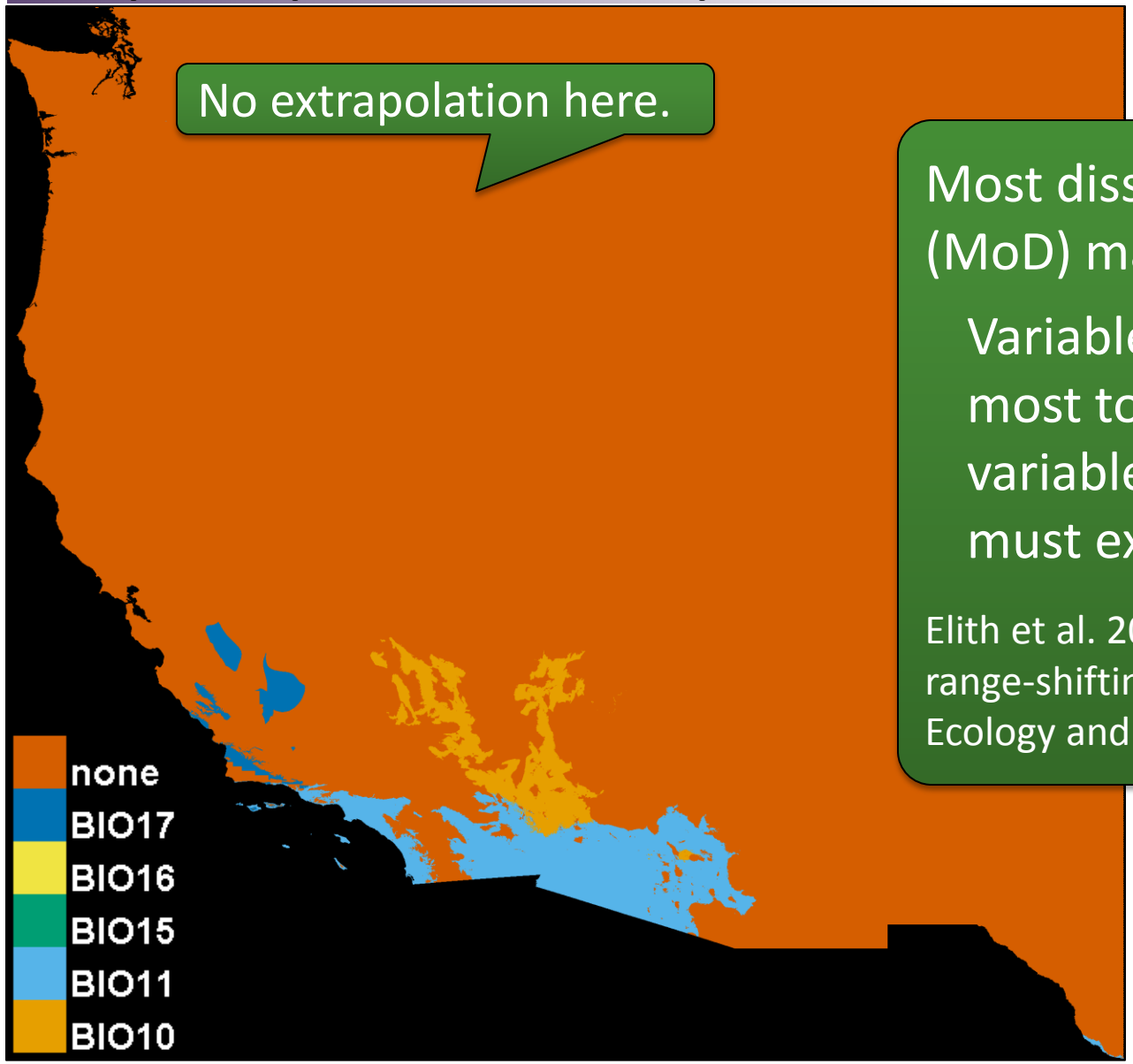
<0 if at least one variable is beyond the training range

(units in 100ths of range of each variable)

Elith et al. 2010 The art of modeling range-shifting species. Methods in Ecology and Evolution 1:330-342.

Exercise I: Maxent

Map output: MoD map



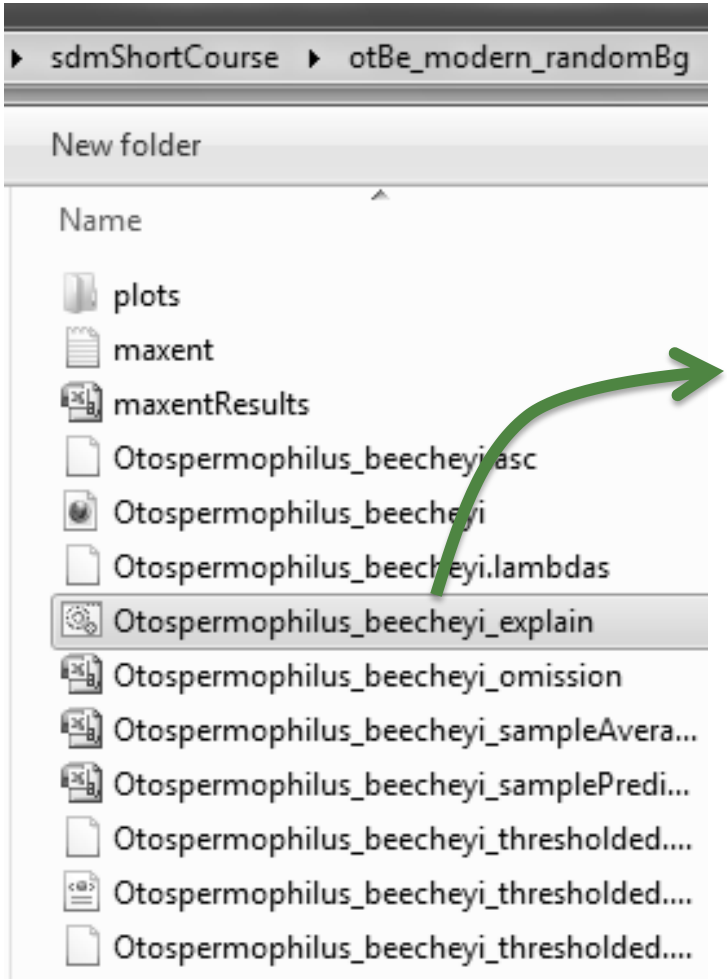
Most dissimilar variable (MoD) map:

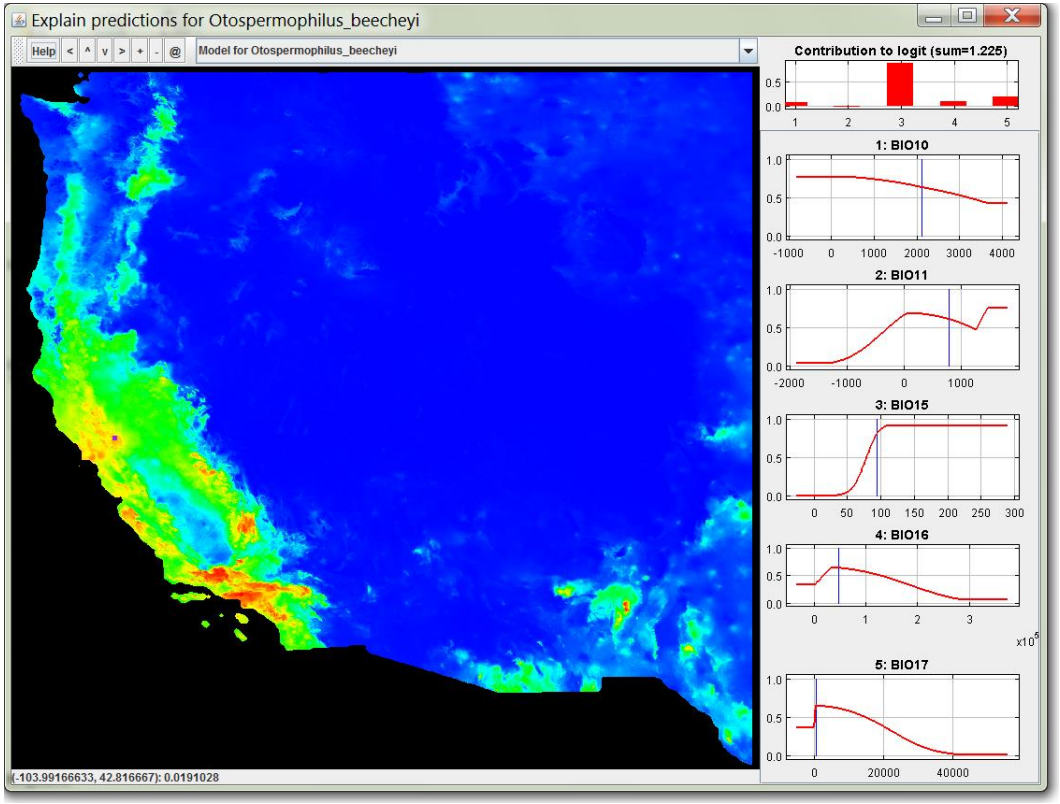
Variable that contributes most to MES score… the variable in which the model must extrapolate the most.

Elith et al. 2010 The art of modeling range-shifting species. Methods in Ecology and Evolution 1:330-342.

Exercise I: Maxent

Response function explorer

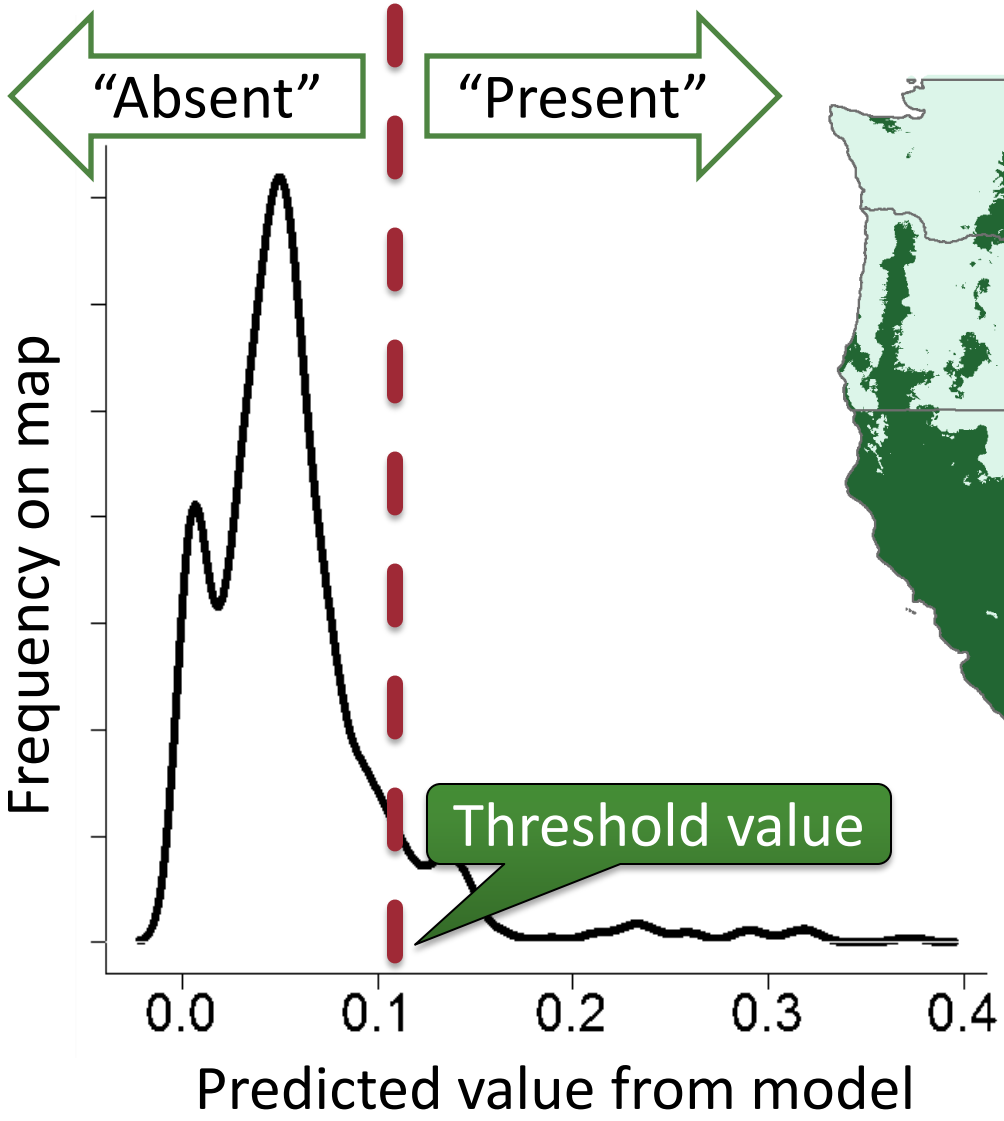


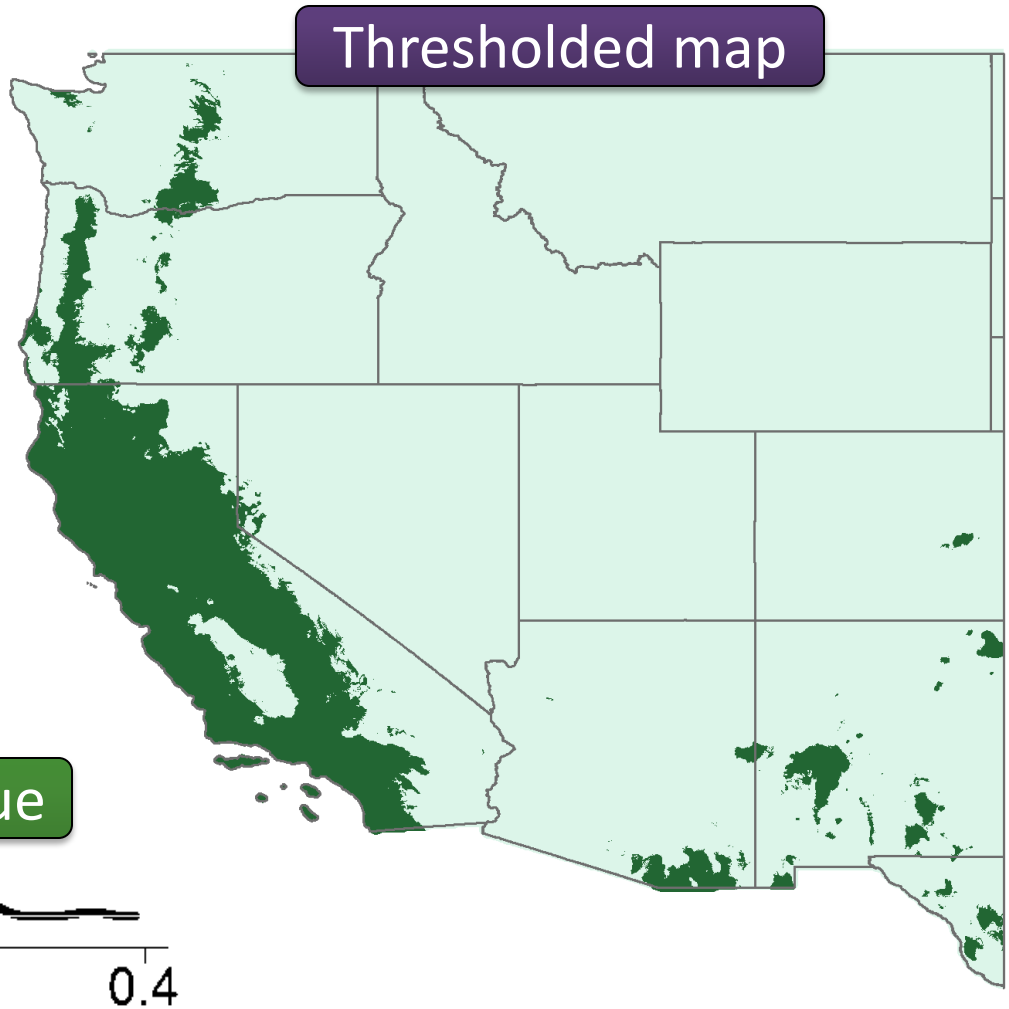


TIP: Can also be used to display each predictor (see drop-down menu at top).

Exercise I: Maxent

Thresholded output





Exercise I: Maxent
Thresholded output

error rates

Sensitivity = proportion of presences correctly predicted

Omission error rate = proportion of presences incorrectly predicted

Specificity = proportion of absences correctly predicted

Commission error rate = proportion of absences incorrectly predicted

See: Liu et al. 2005. Selecting thresholds of occurrence in the prediction of species distributions. Ecography 28:385-393.

common thresholding rules

- Maximizing sensitivity + specificity
- Equal sensitivity and specificity (omission = commission)
- Minimum value at training or test presence
- *n*-th percentile at training or test values (usually 10th)

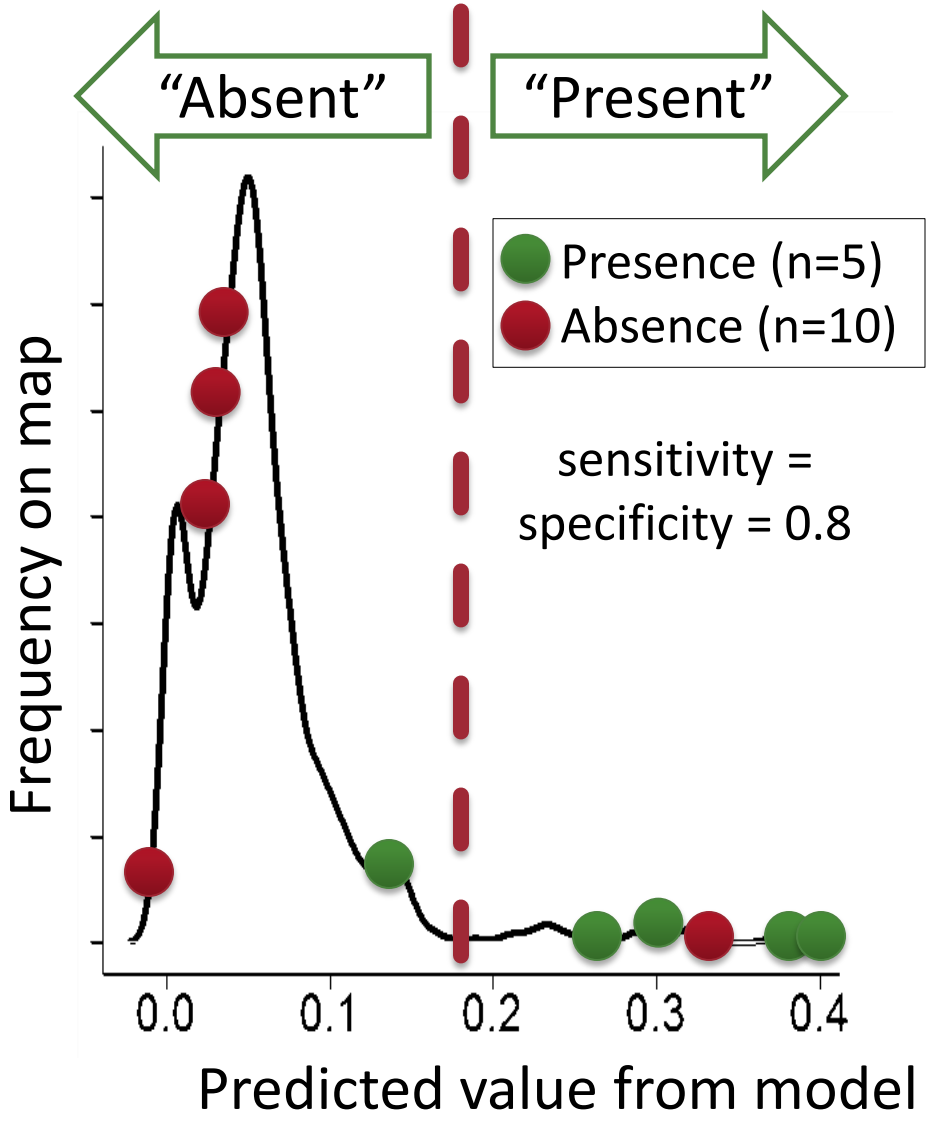
See also:

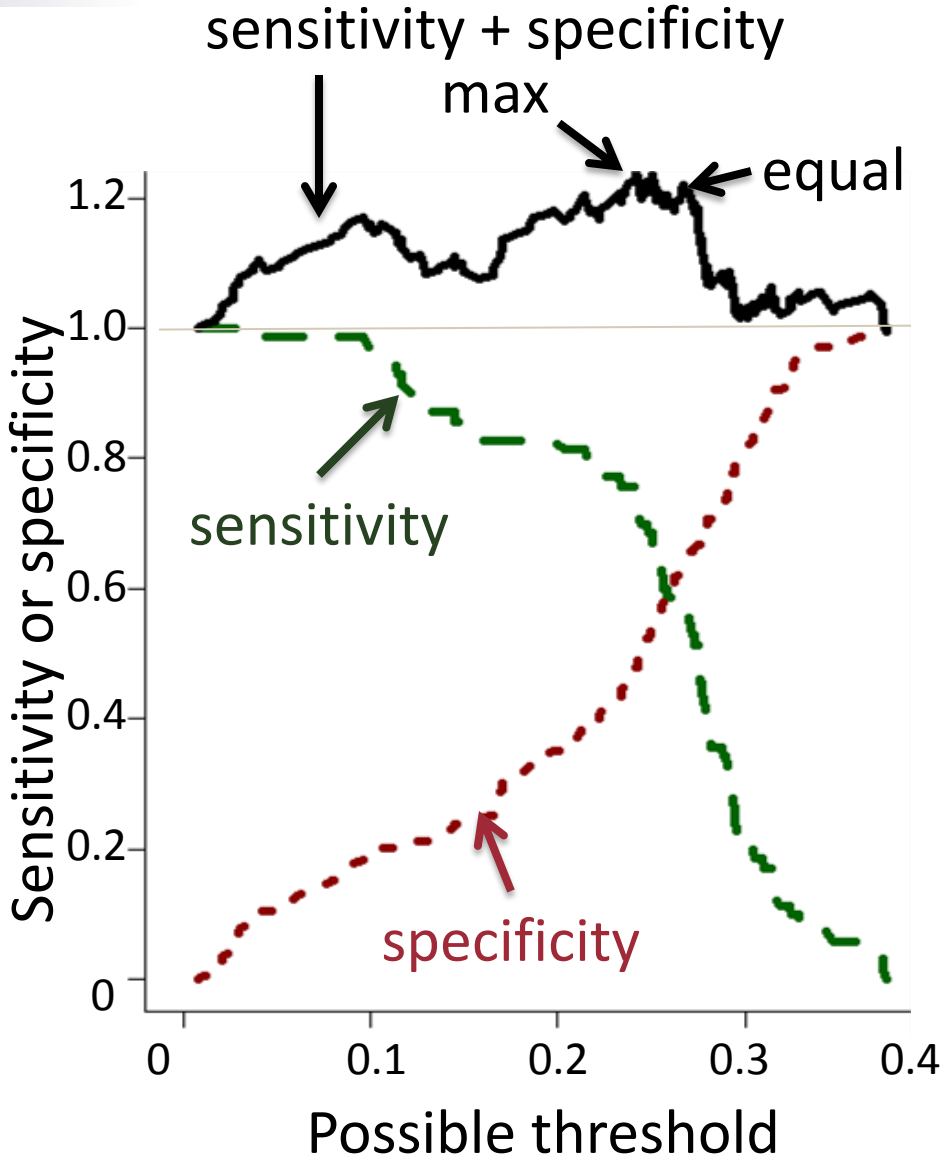
Bean et al. 2012. The effects of small sample size and sample bias on threshold selection and accuracy assessment… Ecography 35:250-258.

Nenzén & Araújo. 2011. Choice of threshold alters projections of species range shifts… Ecological Modeling 222:3346-3354.

Exercise I: Maxent

Thresholded output





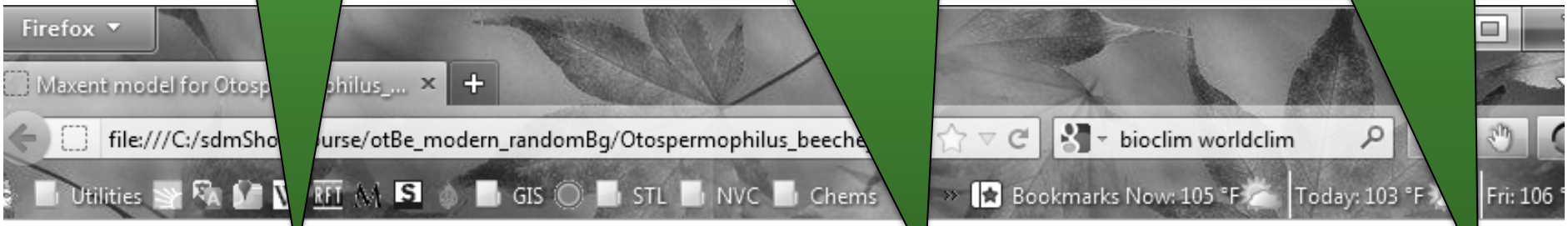
Exercise I: Maxent

Thresholded output

Threshold values

Area of landscape predicted "present" using this rule

Assumes test sites independent but rarely are because of spatial autocorrelation



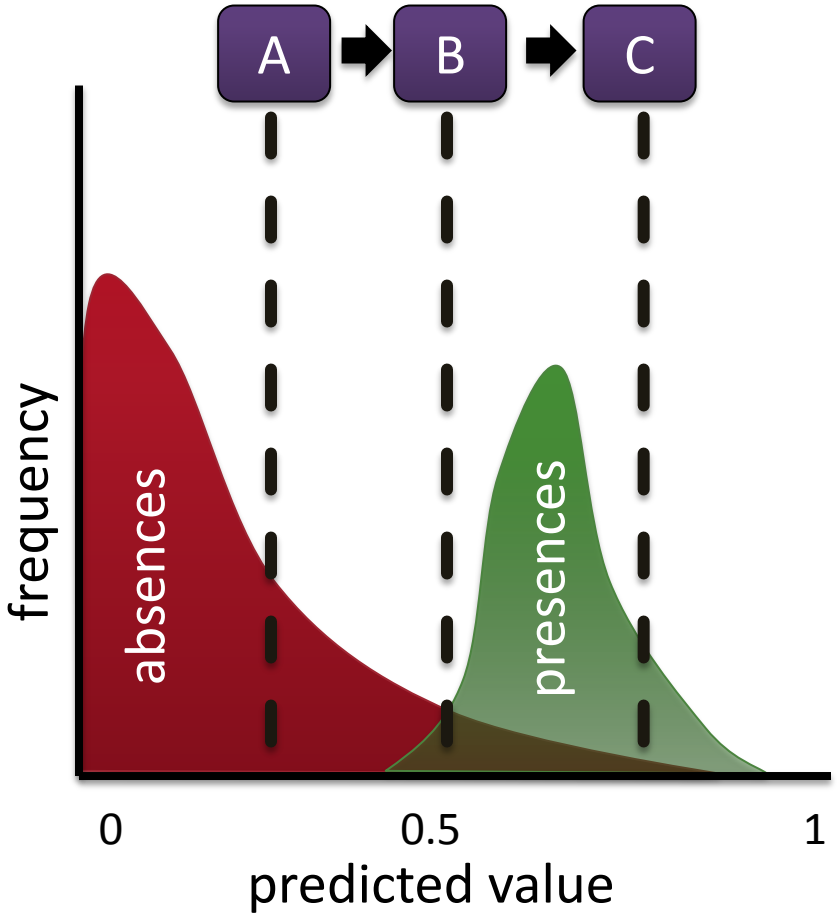
Cumulative threshold	Logistic threshold	Description	Fractional predicted area	Training omission rate	Test omission rate	P-value
1.000	0.012	Fixed cumulative value 1	0.507	0.000	0.000	6.324E-7
5.000	0.055	Fixed cumulative value 5	0.254	0.000	0.000	3.129E-13
10.000	0.139	Fixed cumulative value 10	0.164	0.040	0.095	1.744E-13
5.138	0.057	Minimum training presence	0.250	0.000	0.000	2.227E-13
14.235	0.210	10 percentile training presence	0.129	0.100	0.238	6.09E-11
15.360	0.230	Equal training sensitivity and specificity	0.122	0.120	0.238	2.629E-11
12.163	0.175	Maximum training sensitivity plus specificity	0.144	0.040	0.190	1.566E-11

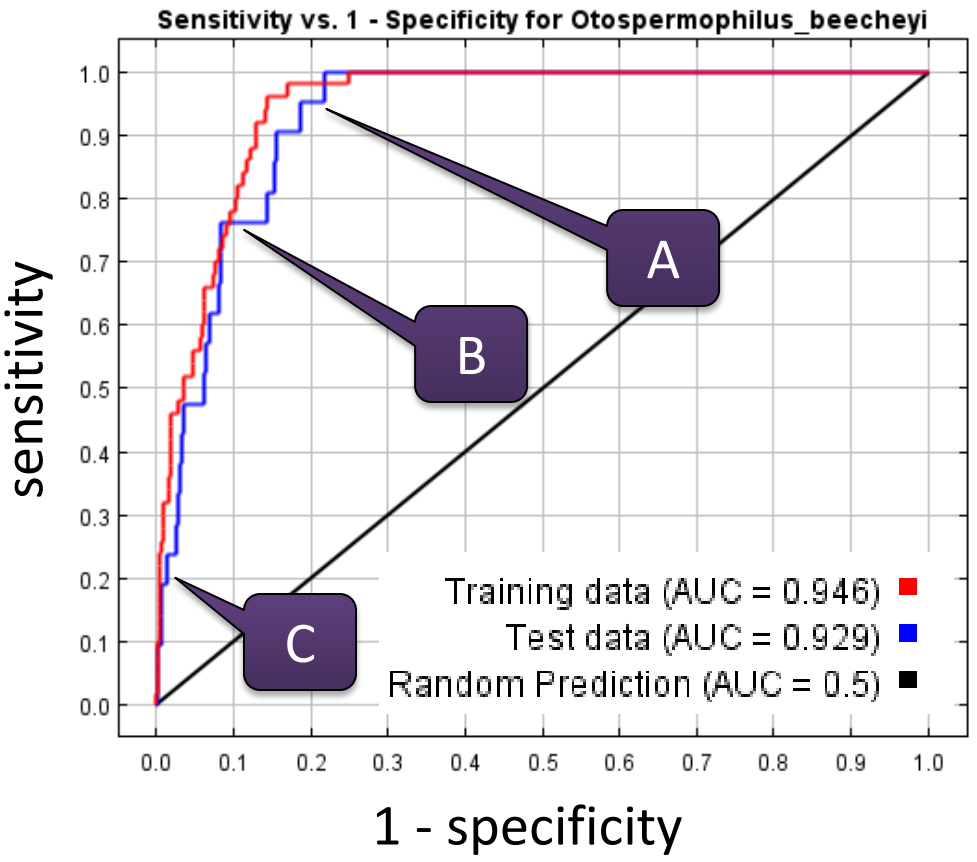
Exercise I: Maxent

Area under the receiver-operator curve (AUC)

Receiver-operator curve (ROC):

Plot of sensitivity vs. 1 - specificity as the threshold is moved from 0 to 1.





AUC is the area under this curve.

Exercise I: Maxent

Area under the receiver-operator curve (AUC)

Receiver-operator curve (ROC):

Plot of sensitivity vs. 1 - specificity as the threshold is moved from 0 to 1.

AUC ranges from 0 to 1

AUC "rules of thumb" (debatable):

1 to 0.9 "excellent"
0.9 to 0.8 "good"
0.8 to 0.7 "fair"
0.5 to 0.7 "poor"
0 to 0.5 "perverse" (less accurate than a random guess!)

Range is >0 to <1 if random "absences" are used instead of real absences, so these rules of thumb won't apply!

No authoritative citation for this, but everyone cites Swets (1988 Measuring the accuracy of diagnostic systems. Science 240:1285-1293), though he doesn't actually pose this breakdown of values.

Interpretation:

With true absences: Probability that a randomly-drawn presence site has *a higher score than a randomly drawn absence site*.

With randomly located sites: Probability that a randomly-drawn presence has a *higher score than a randomly drawn site*.

Range will be from ~0 to ~1 in most cases.

Smith *In review.* A tradeoff between apparent and actual accuracy of species distribution models evaluated with random background sites in place of absences.

A short course on distribution modeling

Outline I

Introduction to distribution modeling
What does a SDM do?
Model algorithms
Input: Species' records
Input: Predictors

Exercise I: Maxent
Input: Species' records, predictor rasters
Output: Maps, thresholds, AUC, variable importance, response functions

Maxent under the hood:
Information entropy maximization
"Feature" functions
Regularization and beta parameter
Assumptions

Exercise II: Maxent with SWD format, k-folds, and projecting to new era/region
Input: SWD format
K-fold data splitting
Projection to new era/region

Best practices: "Truth" vs. "useful bias"
SDM assumptions

Best practices: Training records
Data cleaning
Coordinate uncertainty
Number of records

Best practices: Predictors
Proximate & direct, conditions & resources
Types
Resolution
Accuracy
Correlations
Gradients
Dynamic vs. static vs. dynamic-but-static

Maxent under the hood

Maxent described in Methods sections

to other methods (Elith et al. 2006). Maxent estimates a target probability distribution for each species by finding the probability distribution of maximum entropy (i.e., closest to uniform), subject to a set of constraints (environmental variables) that represents the incomplete information about the target
al. 2006). For each model

MAXENT estimates species' distributions by finding the distribution of maximum entropy (i.e. closest to uniform) subject to the constraint that the expected value of each environmental variable (or its transform and/or interactions) under this estimated distribution
illips et al. 2006). In

Dudík 2008). The Maximum Entropy algorithm estimates the realized niche by finding the probability distribution of species presence that is most spread-out (i.e. closest to uniform), constrained by the data relating presence to the environment (Phillips et al.

'maximum-entropy.' Maxent estimates the likelihood of a species being present by finding the distribution of maximum entropy (i.e. that is closest to uniform) subject to the constraint that the expected value of each environmental variable under this estimated distribution matches its empirical average (Phillips *et al.*, 2006). Maxent uses

Maxent under the hood

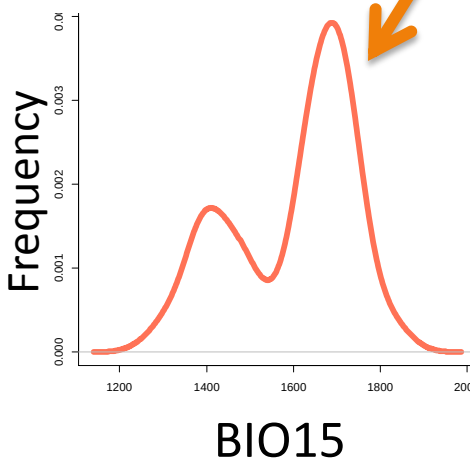
Calculating probability of presence

From Bayes' Rule:

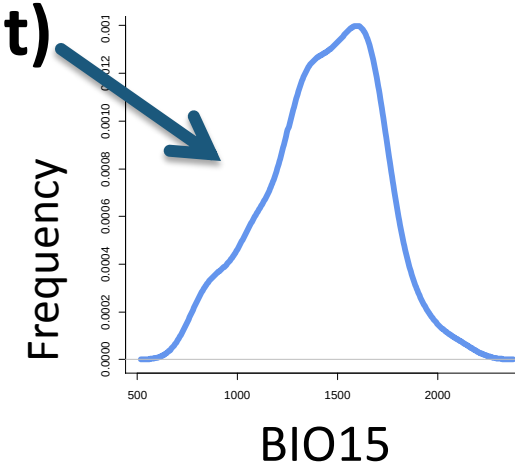
We want to know this.

Unknown constant… MX assumes mean Pr(pres) at training sites = 0.5.

$$\Pr(\text{presence} \mid \text{environment}) = \frac{\Pr(\text{environment} \mid \text{presence}) \times \Pr(\text{presence})}{\Pr(\text{environment})}$$



Pr(env|pres) from species' training sites.



Pr(env) drawn from landscape (often 10,000 randomly located sites)… so extent of landscape directly influences Maxent results!

Maxent under the hood

Information entropy

$$H = -\sum_{i=\min}^{i=\max} p_i \ln p_i$$

information entropy

probability of BIO15 on landscape

Die example:

$$= -\tfrac{1}{6}\ln\tfrac{1}{6} - \tfrac{1}{6}\ln\tfrac{1}{6} - \tfrac{1}{6}\ln\tfrac{1}{6} - \tfrac{1}{6}\ln\tfrac{1}{6} - \tfrac{1}{6}\ln\tfrac{1}{6} - \tfrac{1}{6}\ln\tfrac{1}{6} = 1.79$$

All probabilities are equal therefore this is the "smoothest" distribution.

Maxent under the hood

Maximization of information entropy

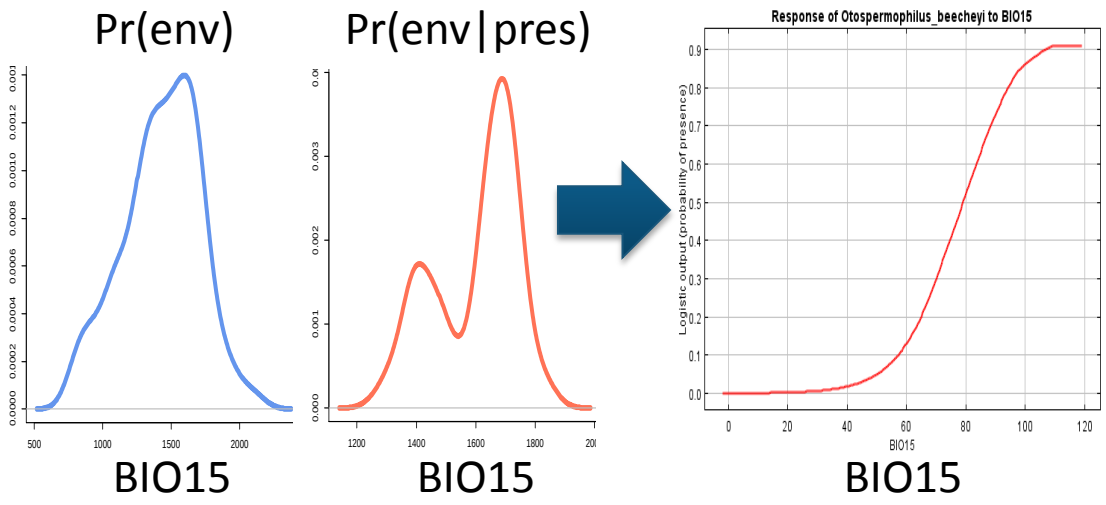
$$\text{maximize}\left(-\sum_{i=\min}^{i=\max} p_i \ln p_i\right)$$

subject to "constraints" (same mean, variance, etc.) as environment at species' presences ± "regularization" factor

Pr(presence | env) =

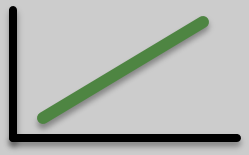
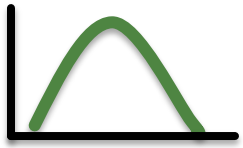
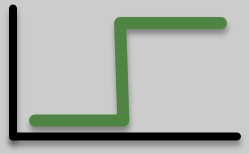
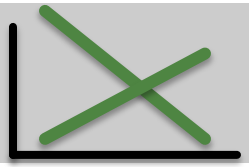
$$\frac{\textbf{Pr(env | presence)} \times \textbf{Pr(presence)}}{\textbf{Pr(environment)}}$$

MX finds a function of BIO15 that is literally the "smoothest" with the same mean, variance, etc. as the distribution of BIO15 at the species' presence sites.



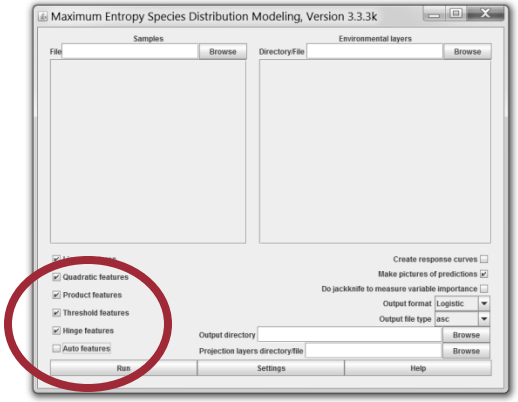
Maxent under the hood

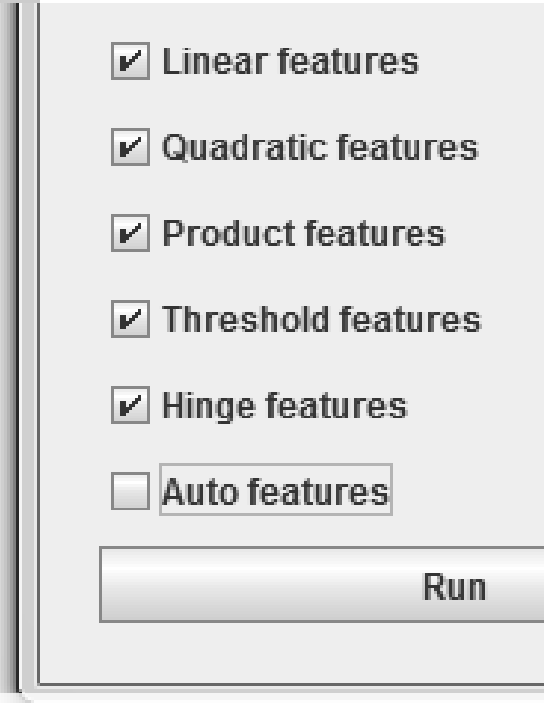
Constraints

"Feature"	Constraint	Name	min num of presences
	mean	linear	>0
	variance	quadratic	≥10
	proportion above/below threshold	step	≥15
	(as above) & mean	hinge	≥80
	covariance	product (2 variables)	≥80
	proportion in each category	categorical	>0

Pr(env | presence)

env





Use only hinge features to avoid overly-complex fits. Elith et al. 2010. The art of modeling range-shifting species. Methods Ecology & Evol 1:330-342.

Maxent under the hood

Regularization

$$\text{maximize}\left(-\sum_{i=\min}^{i=\max} p_i \ln p_i\right)$$ subject to "constraints" (same mean, variance, etc.) as environment at species' presences ± "regularization" factor

"regularization" factor

= β × SE of constraint value (mean, variance, etc.)
(smaller the more training presences there are)

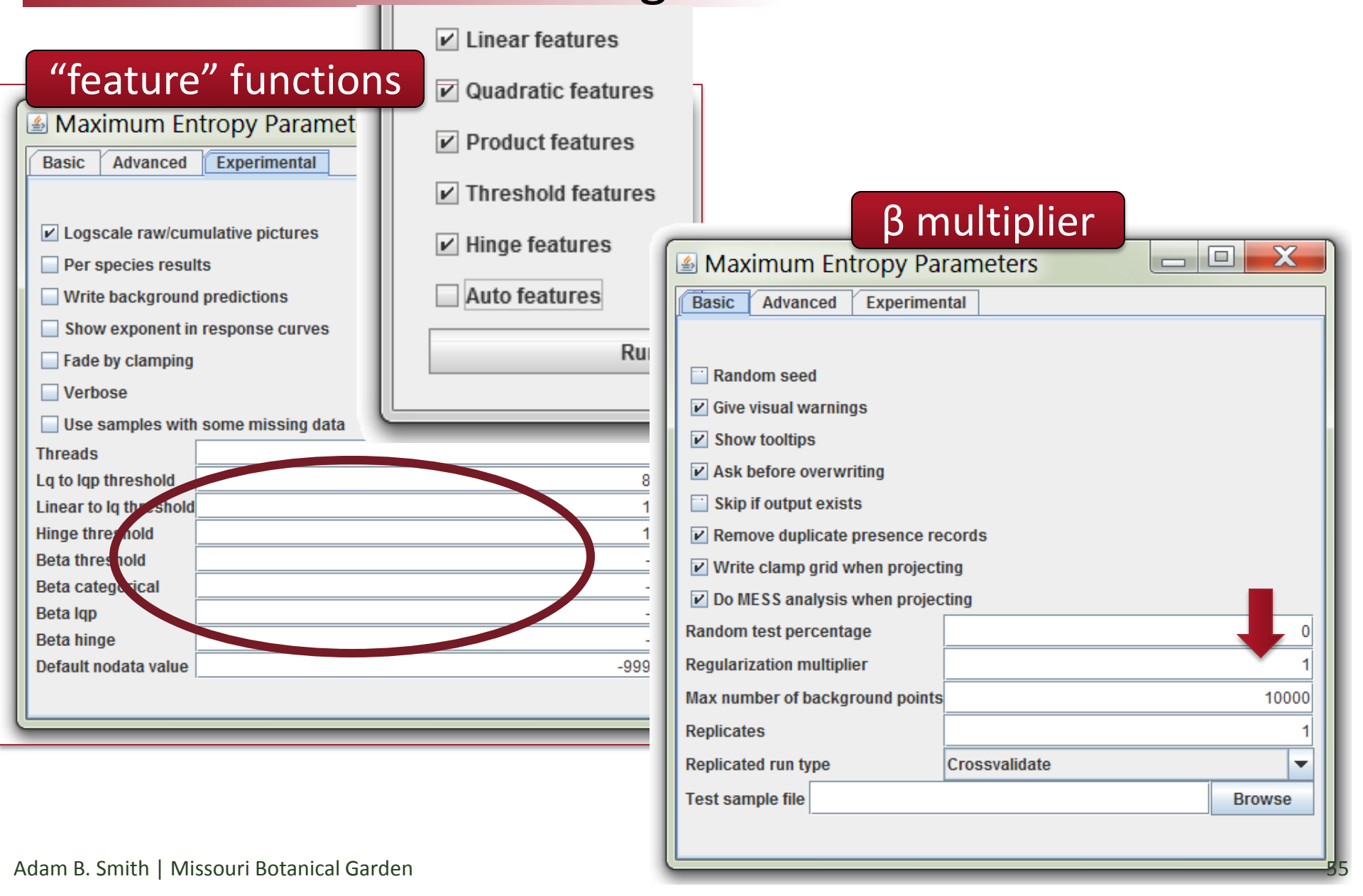
- allows "wiggle" room in having same mean, variance, etc.
- β set by "tuning" to large dataset… default is "1"… Settings → Basic → "Regularization multiplier"

Phillips & Dudík. 2008. Modeling species distributions with Maxent: New extensions and a comprehensive evaluation. Ecography 31:161-175.

- Larger β means smoother curves (good for rare/invading species)
- can use AIC to set β Warren & Siefert 2011. Ecological niche modeling in Maxent: The importance of model complexity and the performance of model selection criteria. Ecological Applications 21:335-342.

Maxent under the hood

Feature functions and regularization



A short course on distribution modeling

Outline I

Introduction to distribution modeling
What does a SDM do?
Model algorithms
Input: Species' records
Input: Predictors

Exercise I: Maxent
Input: Species' records, predictor rasters
Output: Maps, thresholds, AUC, variable importance, response functions

Maxent under the hood:
Information entropy maximization
"Feature" functions
Regularization and beta parameter
Assumptions

Exercise II: Maxent with SWD format, k-folds, and projecting to new era/region
Input: SWD format
K-fold data splitting
Projection to new era/region

Best practices: "Truth" vs. "useful bias"
SDM assumptions

Best practices: Training records
Data cleaning
Coordinate uncertainty
Number of records

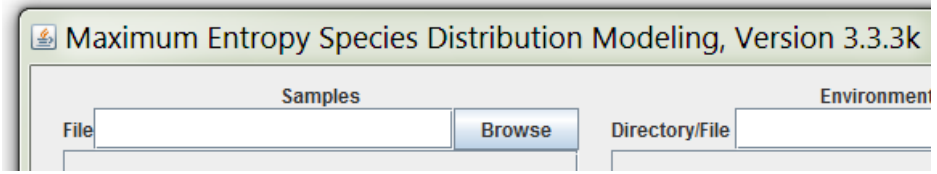
Best practices: Predictors
Proximate & direct, conditions & resources
Types
Resolution
Accuracy
Correlations
Gradients
Dynamic vs. static vs. dynamic-but-static

Exercise II: Advanced Maxent

Purpose

- Demonstrate species-with-data (**SWD**) format… useful for controlling for bias in presence sites, projecting to non-raster environments (e.g., streams), etc.
- Demonstrate use of "**target background**" sites to control for sampling bias of training presences
- Demonstrate measurement of model performance using "**k-fold**" data partitioning

Windows (Maxent software)



R using dismo and raster packages

All R code is in "Exercise II - Advanced Maxent.r" in the folder "sdmShortCourse_kState."

Exercise II: Advanced Maxent

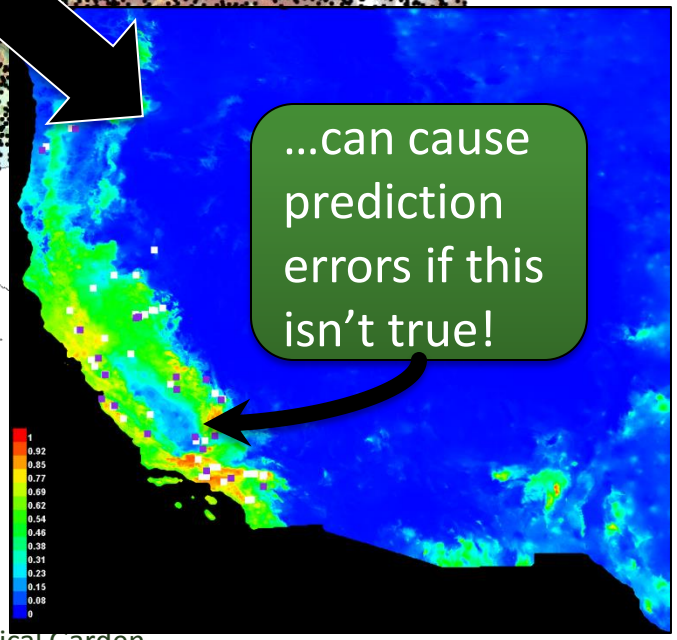
Targeted absences

Random background sites (Maxent default)

Target background sites (control for bias in training presences)

Sometimes this bias can be controlled for using "target background" sites.

Use of randomly located background sites assumes species' records are also randomly sampled from among those in its range.



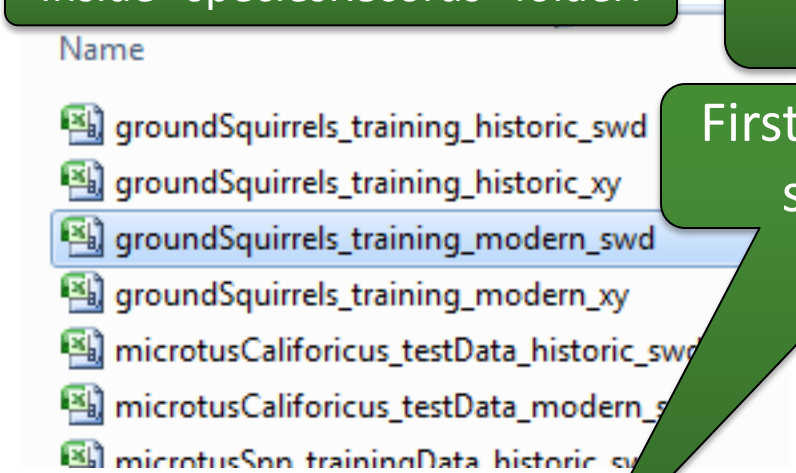
…can cause prediction errors if this isn't true!

Presences of other species expected to have same sampling bias as target species' presences (should include target species' records!)

Exercise II: Advanced Maxent

SWD format

Inside “speciesRecords” folder:



Each row represents a species’ presence.

First column is always species’ name.

Subsequent columns are environmental variables (you don’t actually need latitude or longitude).

	A	B	C	D	E	F	G	H	
1	SPECIES	LONG_WGS84	LAT_WGS84	YEAR	BIO01	BIO02	BIO03	BIO04	BI
2	Callospermophilus lateralis	-106.682625	38.450616	1970	335	1886	41	89283	2
3	Callospermophilus lateralis	-118.7964	42.7835	1970	791	1359	39	72592	2
4	Callospermophilus lateralis	-118.7767	42.7835	1970	729	1335	39	70605	
5	Callospermophilus lateralis	-105.45	36.59	1970	381	1441	41	71005	
6	Callospermophilus lateralis	-112.48288	37.43883	1971	766	1647	43	75458	
7	Callospermophilus lateralis	-122.841982	39.734693	1971	879	931	34	62316	2
8	Callospermophilus lateralis	-118.850219	37.600635	1971	638	1632	46	66548	2
9	Callospermophilus lateralis	-122.17604	41.37069	1971	356	960	35	60248	1

Exercise II: Advanced Maxent

Loading species' data in SWD format

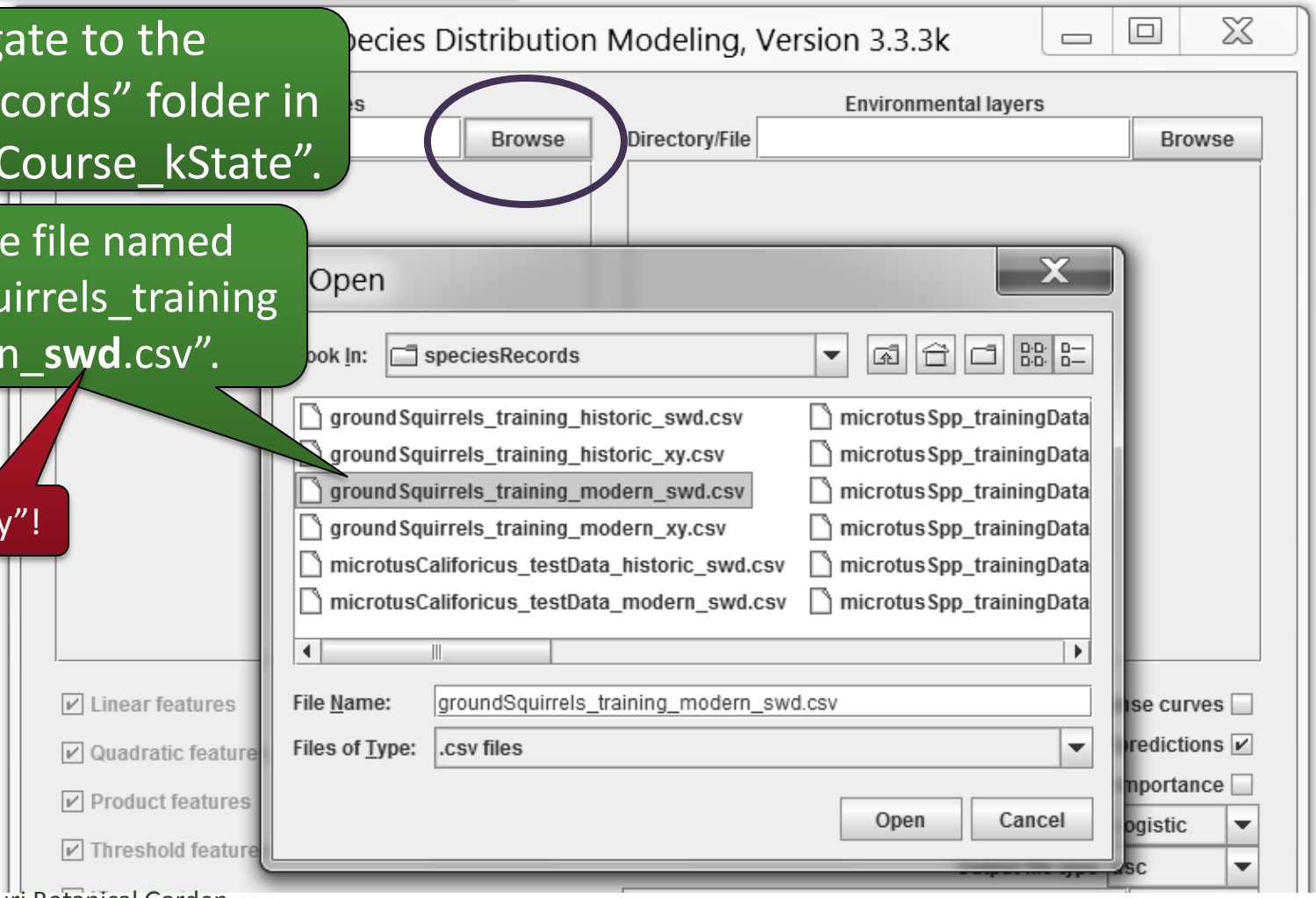
1 **stop then restart Maxent to restore default values**

2 **load species' data**

Navigate to the "speciesRecords" folder in "sdmShortCourse_kState".

Select the file named "groundSquirrels_training_modern_**swd**.csv".

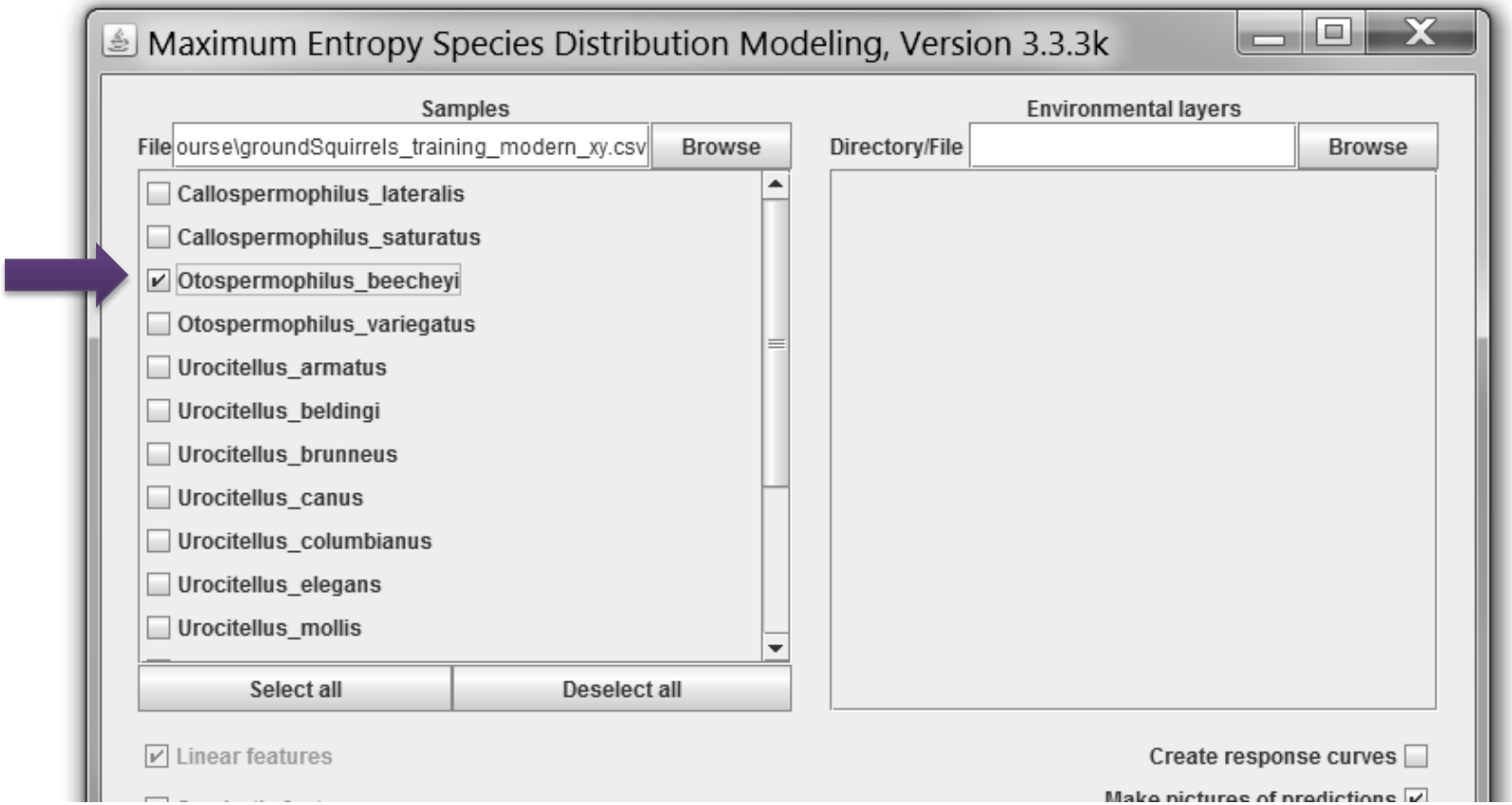
Not "xy"!



Exercise I: Maxent

Loading species' data

3 **select only "Otospermophilus_beecheyi"**



Exercise II: Advanced Maxent

Loading species' data in SWD format

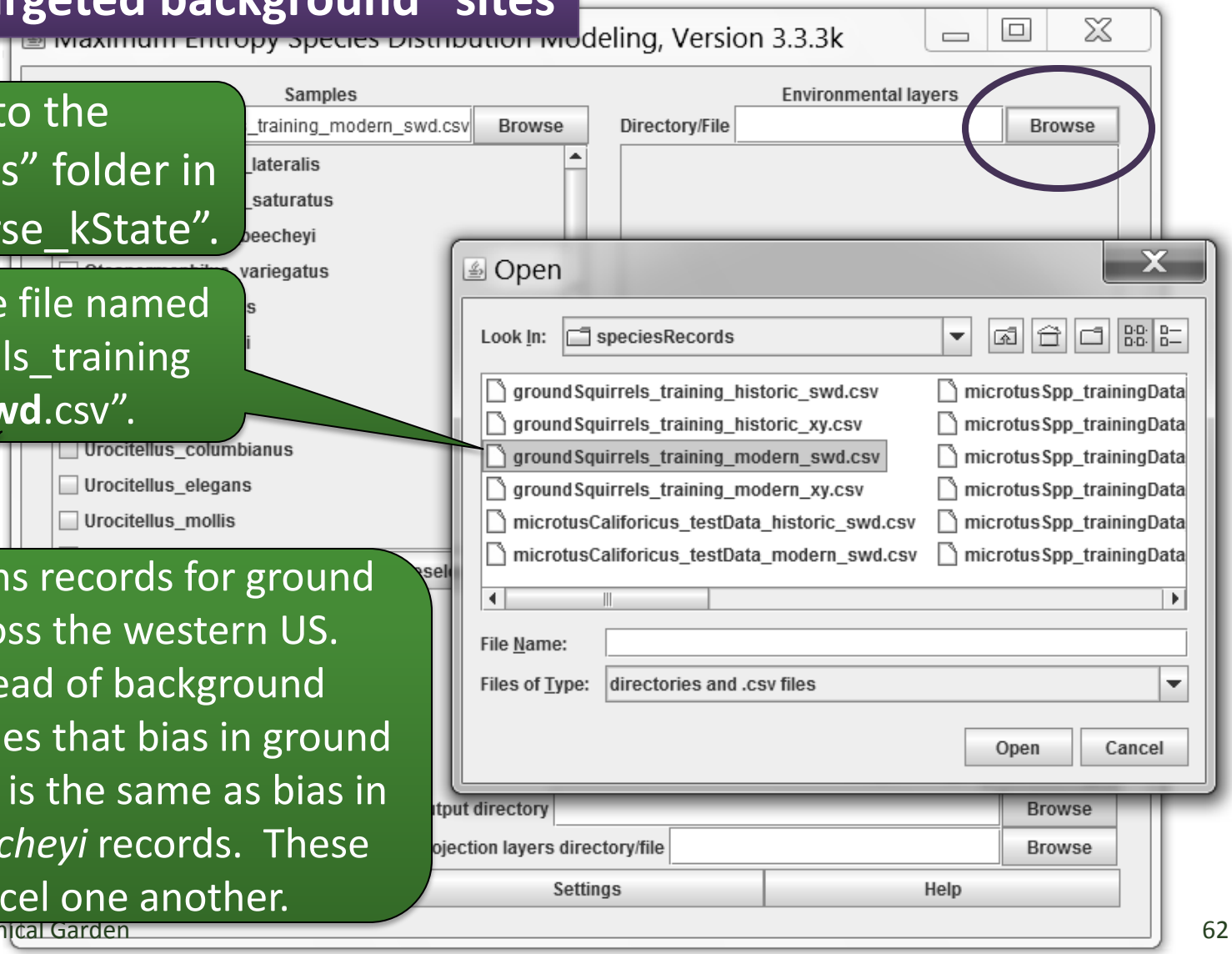
4 **load "targeted background" sites**

Navigate to the "speciesRecords" folder in "sdmShortCourse_kState".

Again, select the file named "groundSquirrels_training_modern_**swd**.csv".

Not "xy"!

This file contains records for ground squirrels across the western US. Using it instead of background absences assumes that bias in ground squirrel records is the same as bias in just the *O. beecheyi* records. These biases cancel one another.

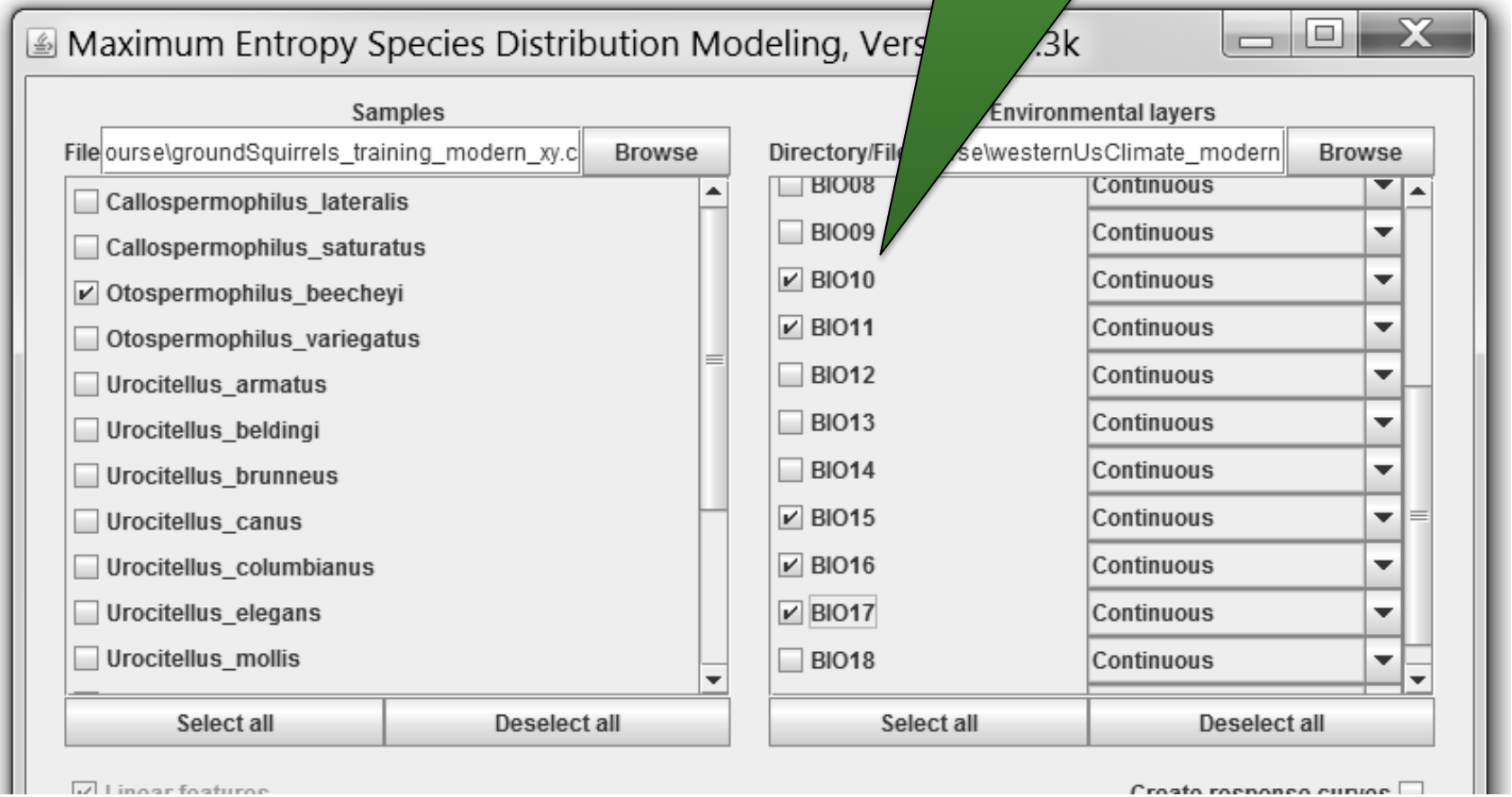


Exercise I: Maxent

Selecting predictors

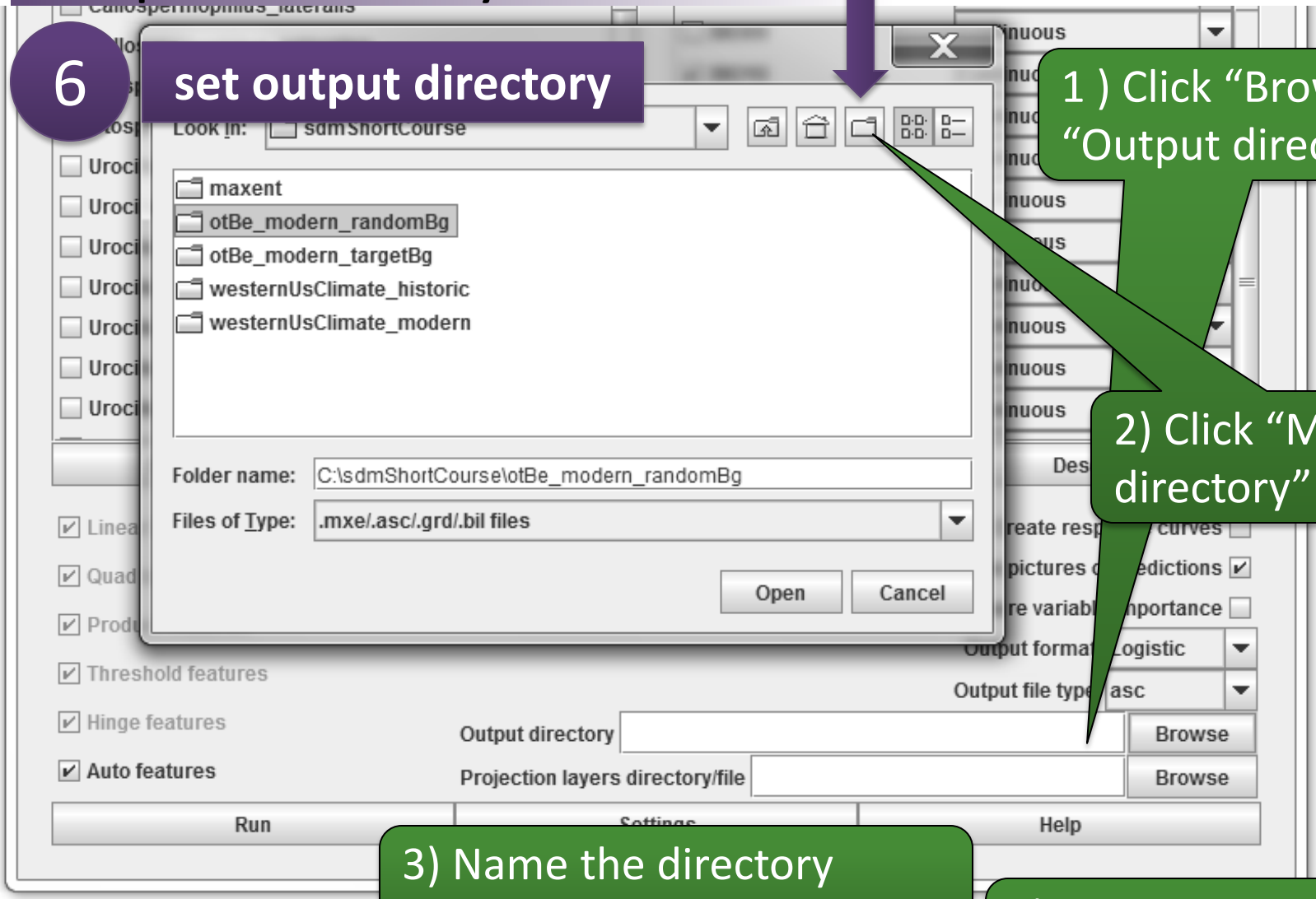
5 **select predictors**

Select BIO10, BIO11, BIO15, BIO16, and BIO17.



Exercise I: Maxent

Output directory

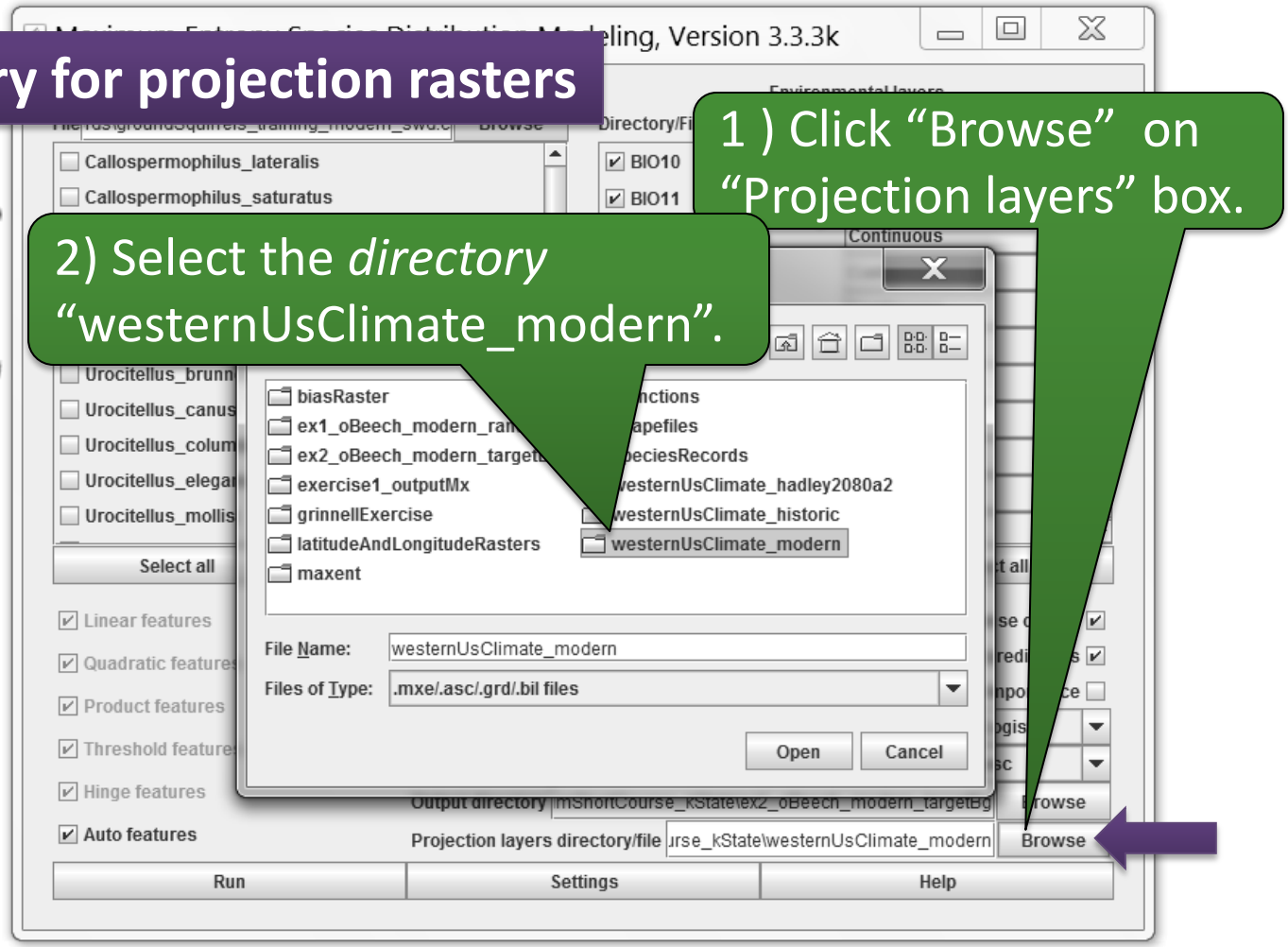


Exercise II: Advanced Maxent

Projecting to a new time period

7 **set directory for projection rasters**

This directory contains climate rasters for the western US in the modern era. We used it last time in the "Environmental layers" box, but this time we loaded a SWD table into that box. Thus, we have to specify where rasters are if we want Maxent to write a prediction raster.



Tip: If you want to project to more than one time/region (e.g., to the present and top the future), list multiple directories in this box separated by a comma.

Exercise II: Advanced Maxent

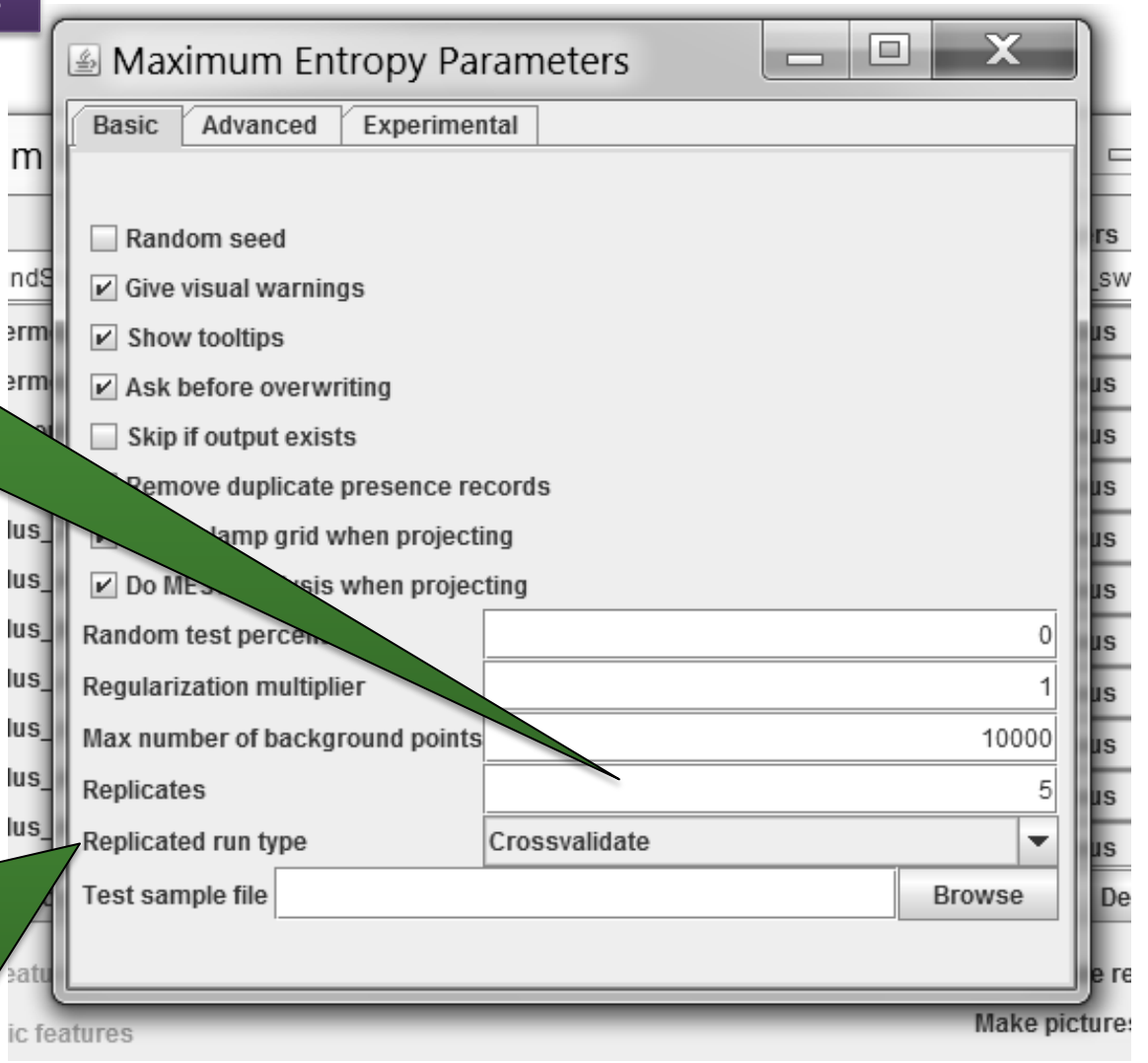
Data splitting for model evaluation

8 **k-fold data splitting**

Click “Settings” to get here.

Set “Replicates” equal to the number of times to split the data and calculate the model. Typically this is set to ≥5 (each split is called a “k-fold”).

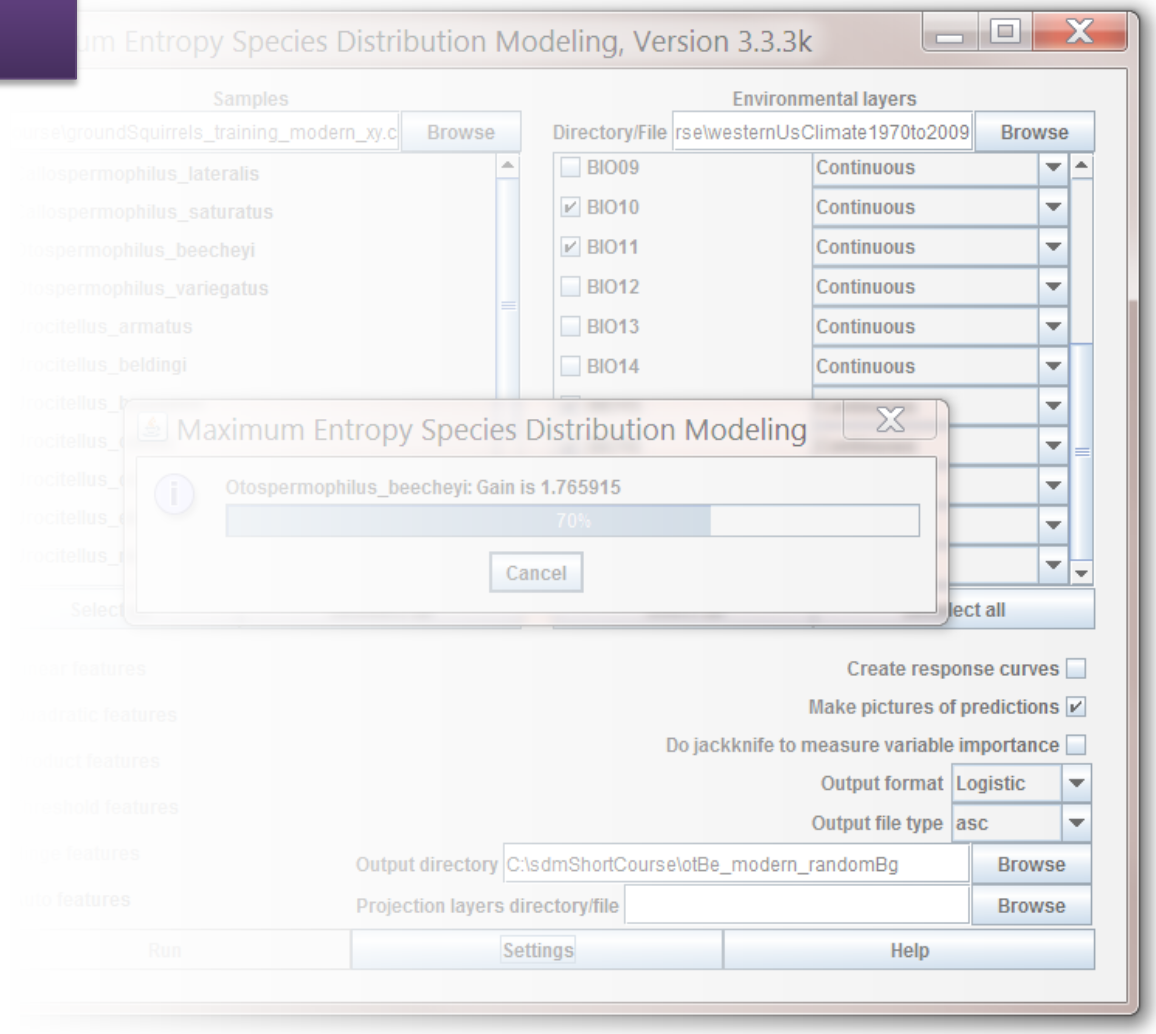
Using “crossvalidate” means to split the data 5 times (20% per partition), train the model 5 times on 80% of the data, and test it each time on the 20% that was withheld. This is more typical than using a single partition like we did in the previous exercise.



Exercise II: Advanced Maxent

Train model

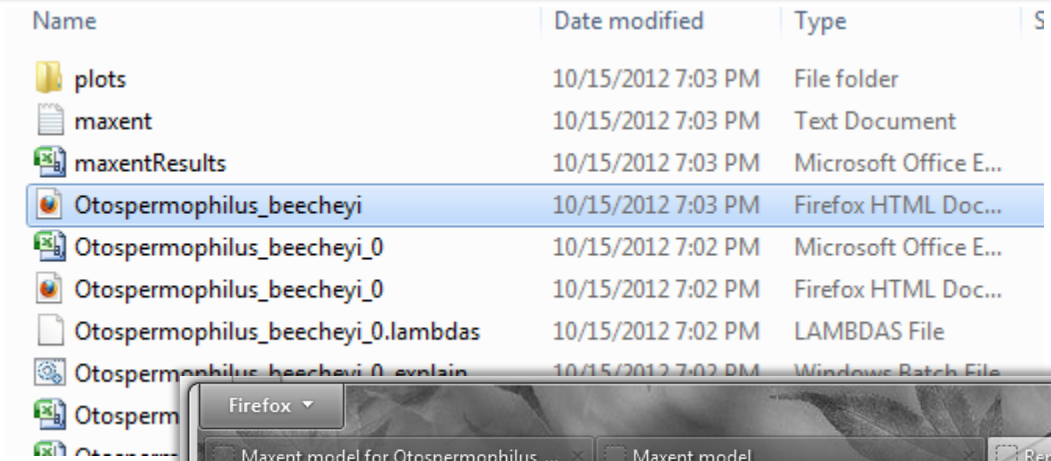
9 **run the model!**



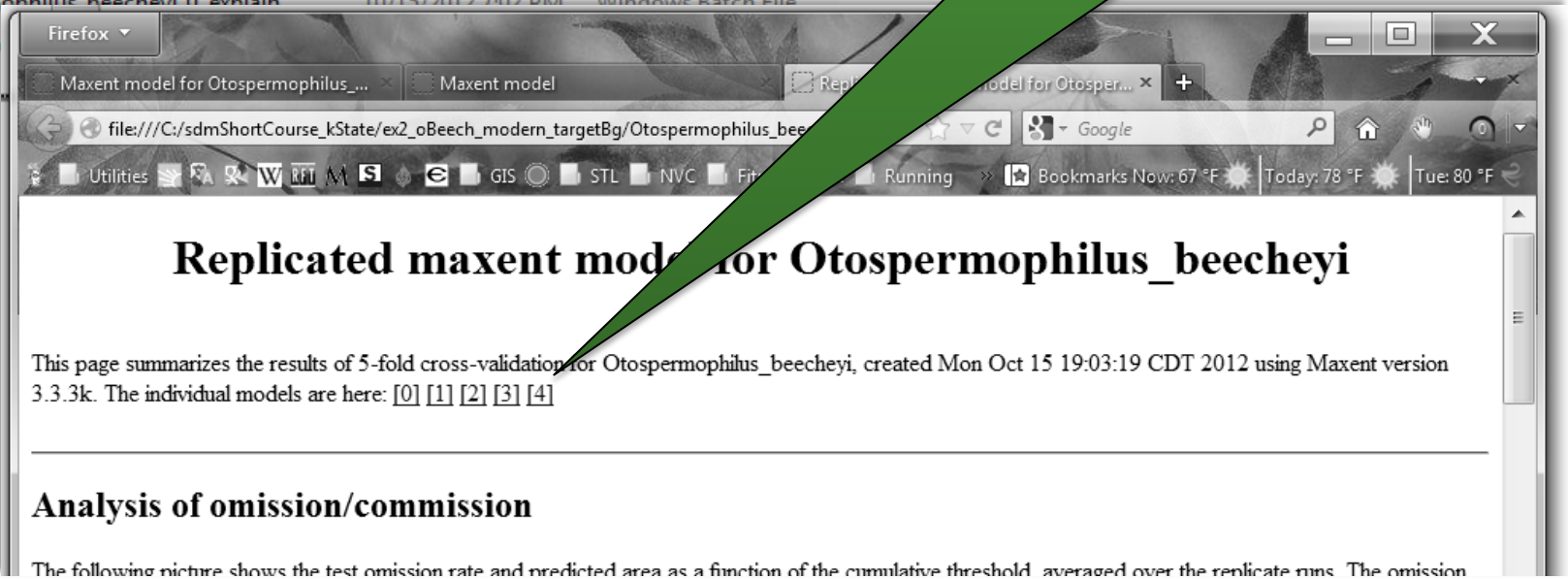
Exercise II: Advanced Maxent

Output

Navigate to the output directory and open the HTML file named “Otospermophilus_beecheyi”.



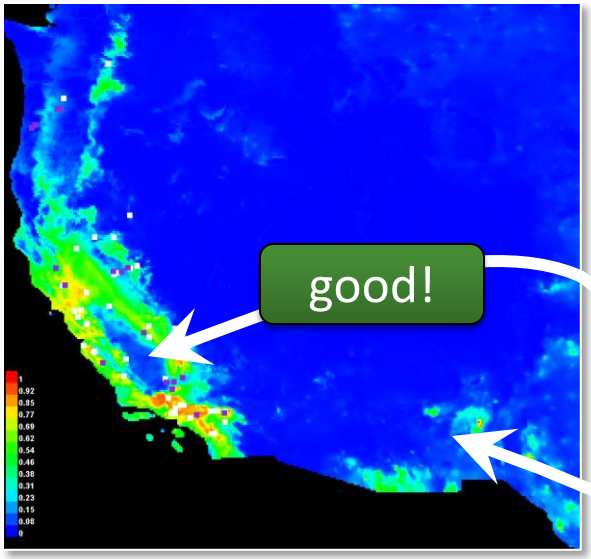
This file is the summary of all 5 k-fold models. You can see the results for each k-fold by clicking these numbers.



Exercise II: Advanced Maxent

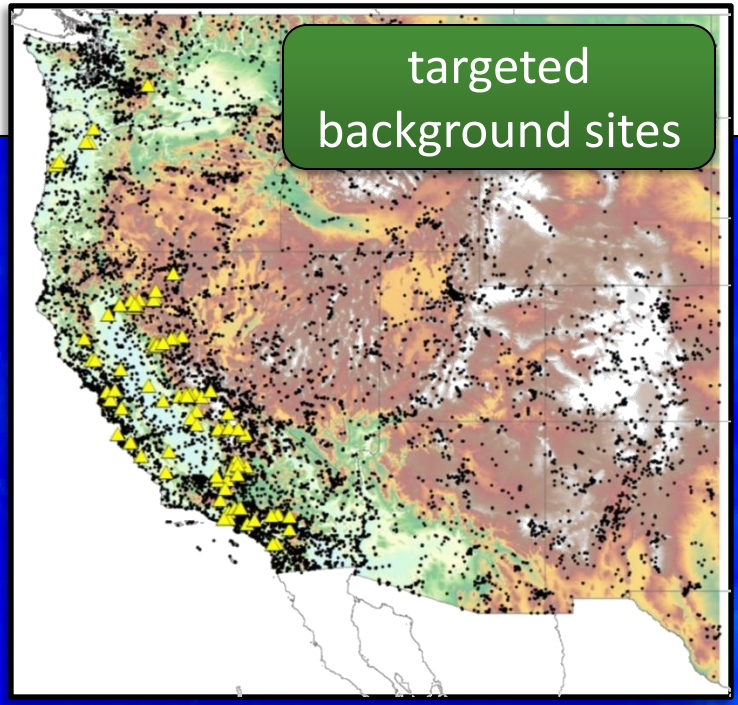
Correction for sampling bias

predictions using random background sites (exercise #1)



good!

predictions using targeted background sites (this exercise)



targeted background sites

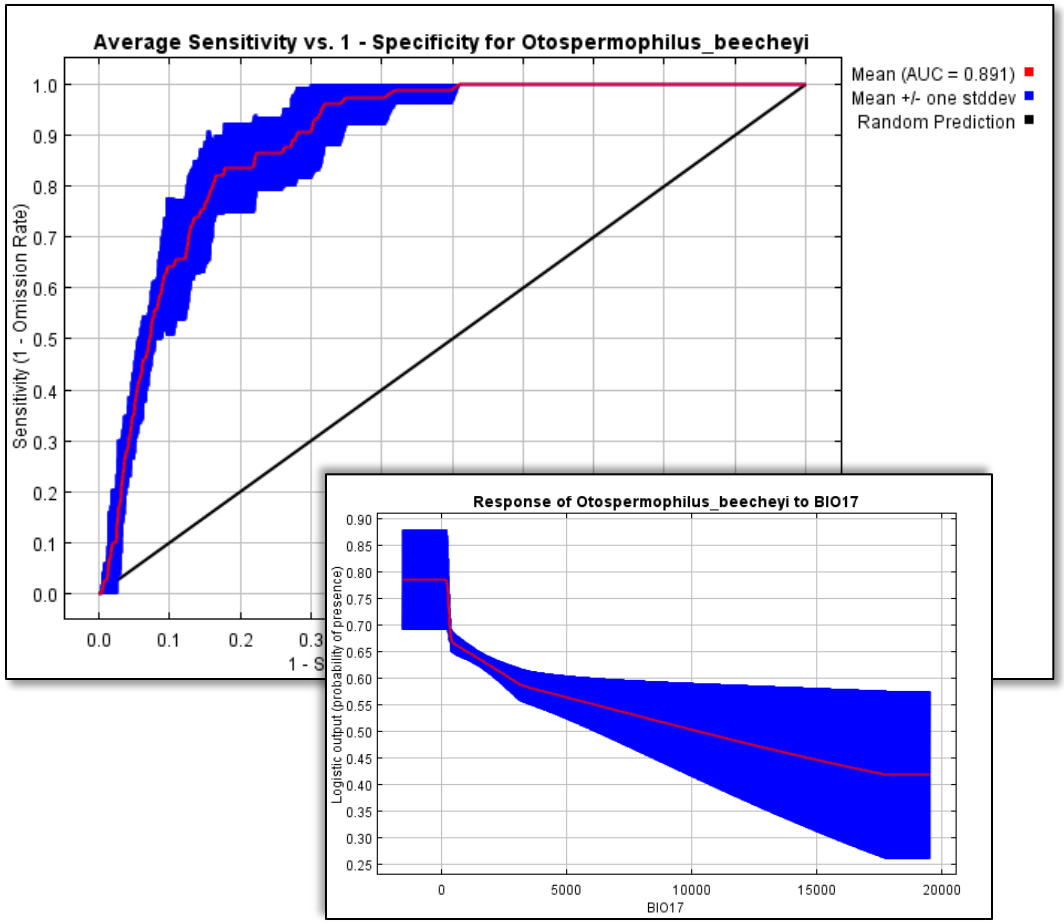
bad…

Use of targeted background sites can correct for bias in some areas while overcorrecting in others.

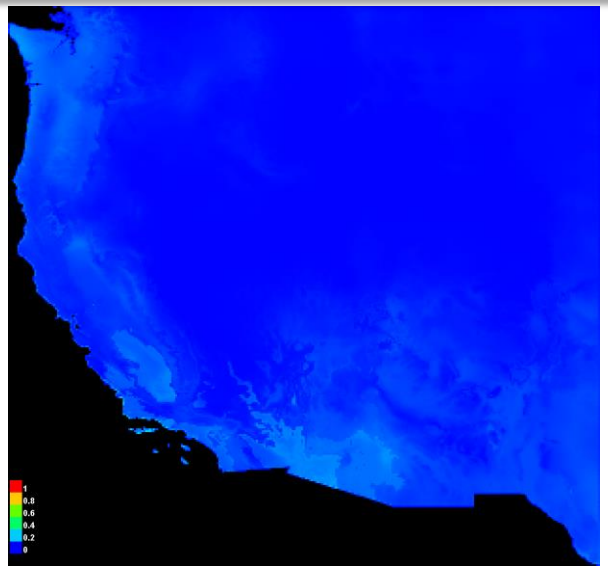
Exercise II: Advanced Maxent

Output

Many of the graphs now show an envelope representing the variation across all 5 k-folds.

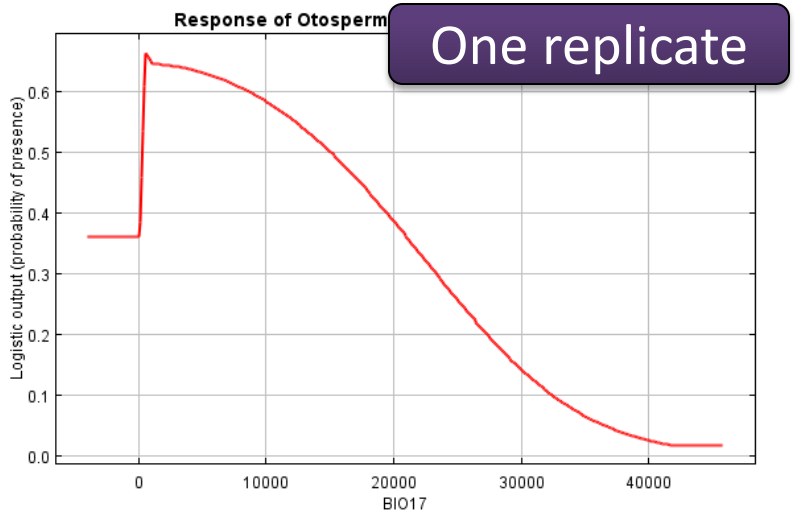


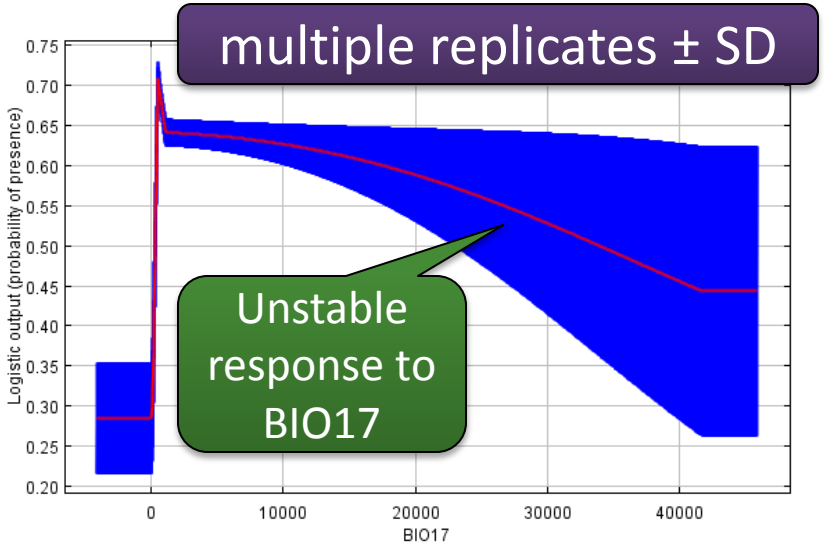
The summary output also includes a map showing the standard deviation of predicted values. Higher values suggest less confidence in those areas.

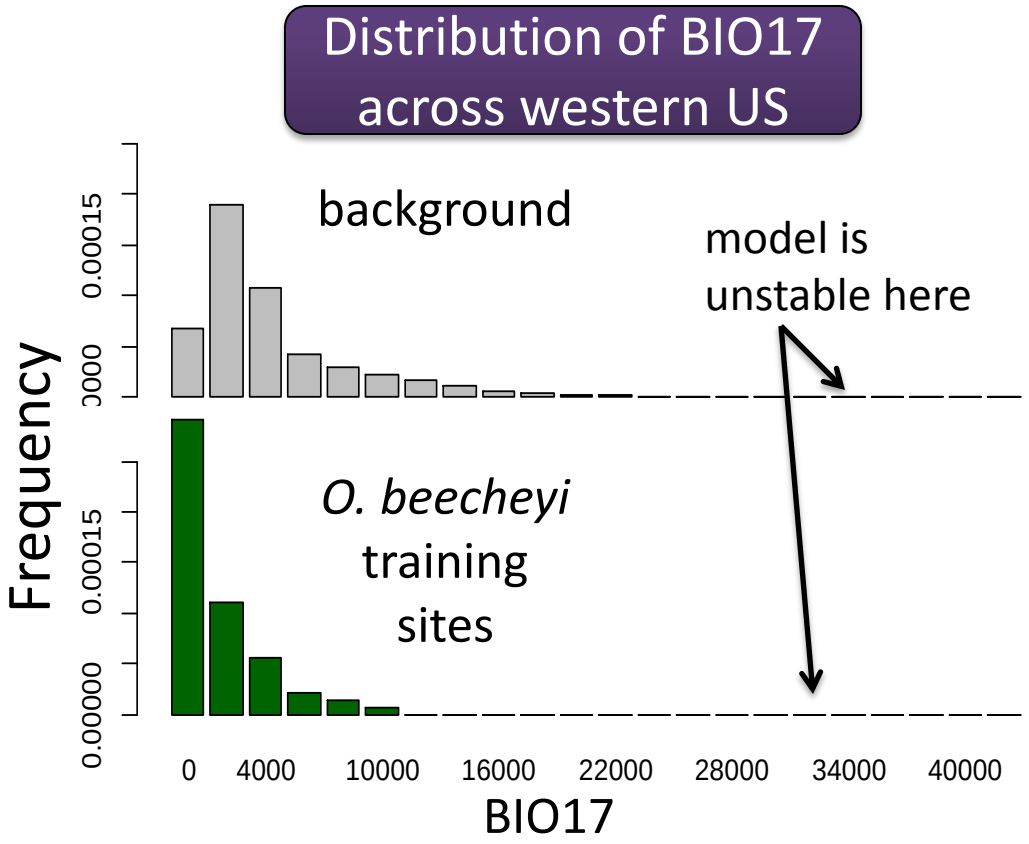


Exercise II: Advanced Maxent

Response function inspection





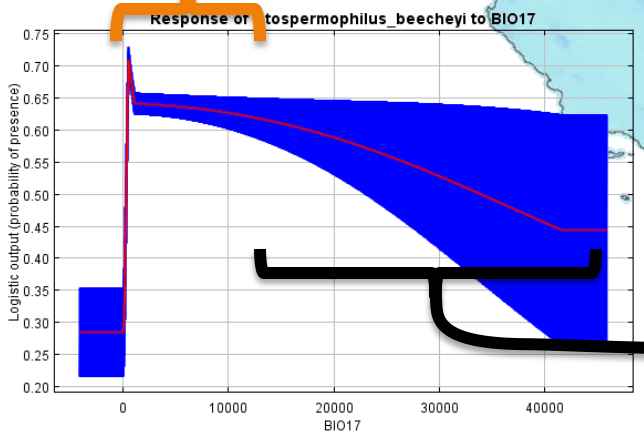


Exercise II: Advanced Maxent

Response function inspection

areas with BIO17 > 10000

Stable for BIO17



Unstable for BIO17

Best practices: "Truth" vs. "useful bias"

Under- and overprediction

The literature on SDMs focuses on "**accuracy**" (=match between model and truth).

However, the "truth" never known (else, why model?). So we **can never know true accuracy** of models.

Nevertheless, we can make decisions likely to **bias** them **in a <u>useful</u> way**:

- **Underprediction**: Modeled range is < true range
- **Overprediction**: Modeled range is > true range

Underprediction useful:

- Identifying "core" areas of range
- Finding favorable reintroduction sites
- Locating previously unknown populations
- "Worst-case" scenario under global change

Overprediction useful:

- Monitoring spread of invasive species
- Identifying areas species could have dispersed to in past (useful for model calibration)

Best practices: "Truth" vs. "useful bias"

SDM assumptions

species in equilibrium with environment

Disrupted by:

- dispersal limitation
- adaptive evolution
- biotic interactions
- population dynamics unrelated to environment
- human persecution/ "encouragement"
- "hold-over" and mass effects (e.g., long-lived adults in environments that currently disfavor reproduction)

Wiens et al. 2009. Niches, models, and climate change: Assessing the assumptions and uncertainties. Proceedings of the National Academy of Sciences USA 106:19729-19736.

populations are "interchangeable"

No local adaptation

Oberle & Schall. 2011. Responses to historical climate change identify contemporary threats to diversity in Dodecatheon. Proceedings of the National Academy of Sciences USA 108:5655-5660.

presence/absence sampled without bias in environmental space

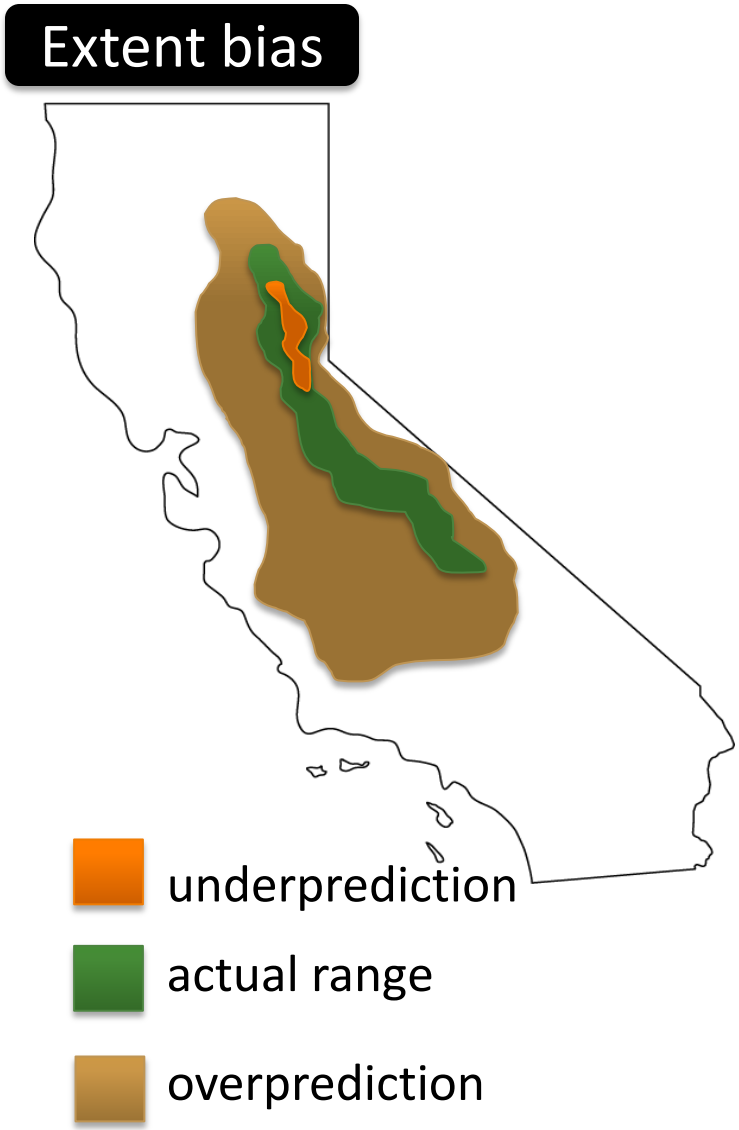
Most records occur near accessible areas (cities, roads)

relevance to conservation SDMing

Rare species and invasive species are more susceptible to different assumptions being broken.

Best practices: “Truth” vs. “useful bias”

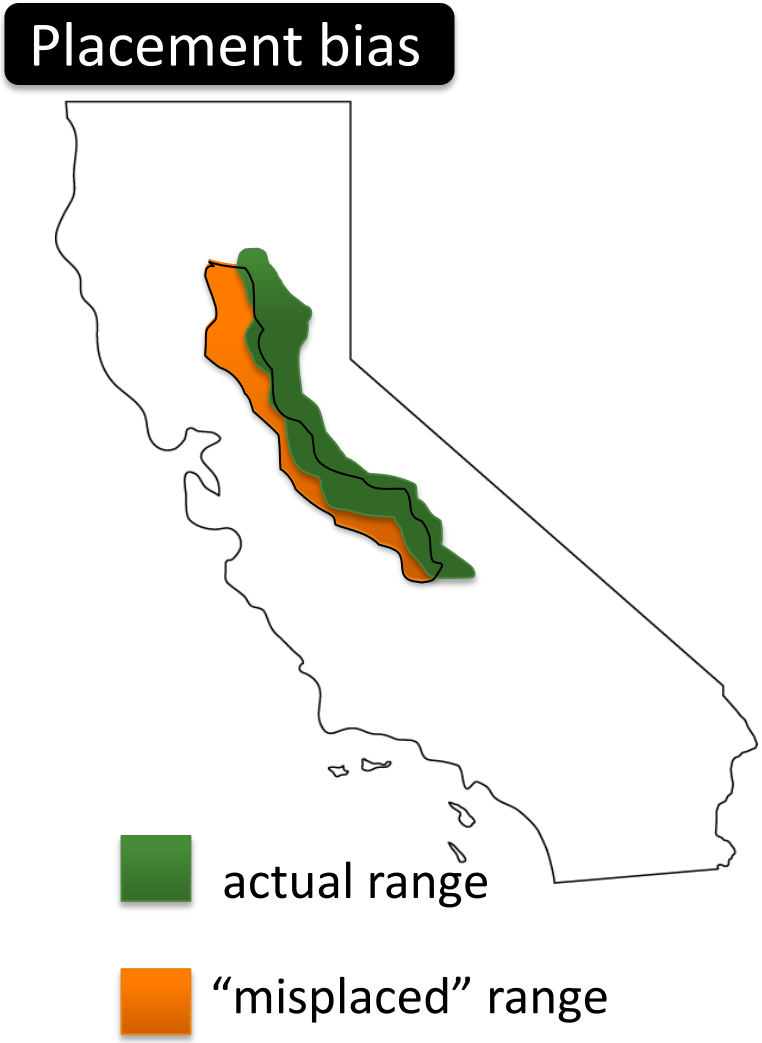
Extent bias



Influenced by:

- Number of training presence sites
- Choice and number of predictors
- Resolution of predictors
- Choice of study extent
- Model algorithm and parameterization
- Threshold used and number of test sites

Best practices: “Truth” vs. “useful bias”
Placement bias

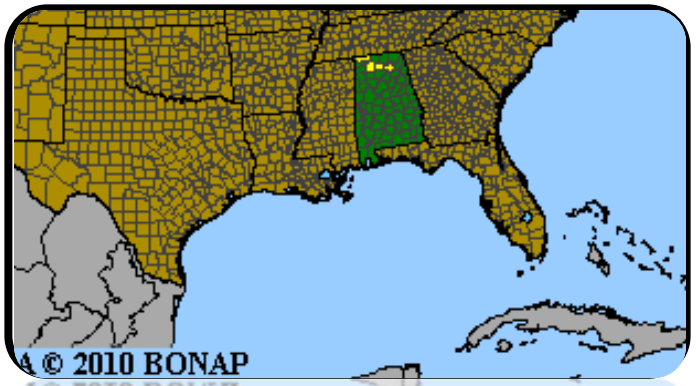


Influenced by:

- Non-random sampling of training presences and/or absences
- Choice of study extent
- Mass effects, long-lived adults with no reproduction
- Disturbance-mediated occurrence
- Dispersal limitation
- Biotic interactions

Best practices: “Truth” vs. “useful bias”

Challenges of rare species

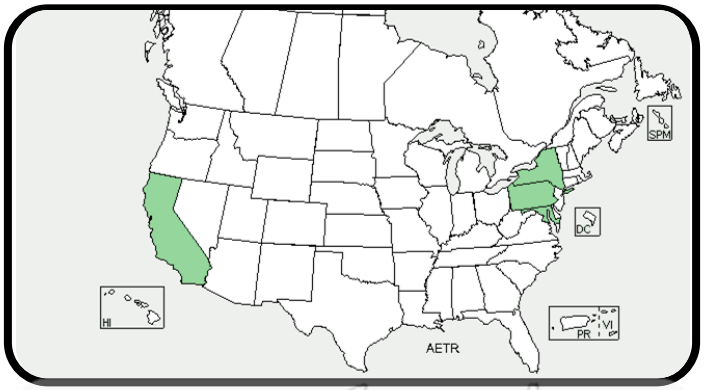


Leavenworthia crassa (n=5)

- Few records for training and testing
- Enhanced problem of overfitting
- May be rare for reasons unrelated to the environment (what does an absence mean?)
- Range undersampled, perhaps biased by a few points… Remaining populations often on the edge of former range Fisher 2011. Trajectories from extinction: Where are missing mammals rediscovered? Global Ecology and Biogeography 20:415-425.

Best practices: “Truth” vs. “useful bias"

Challenges of invasive species

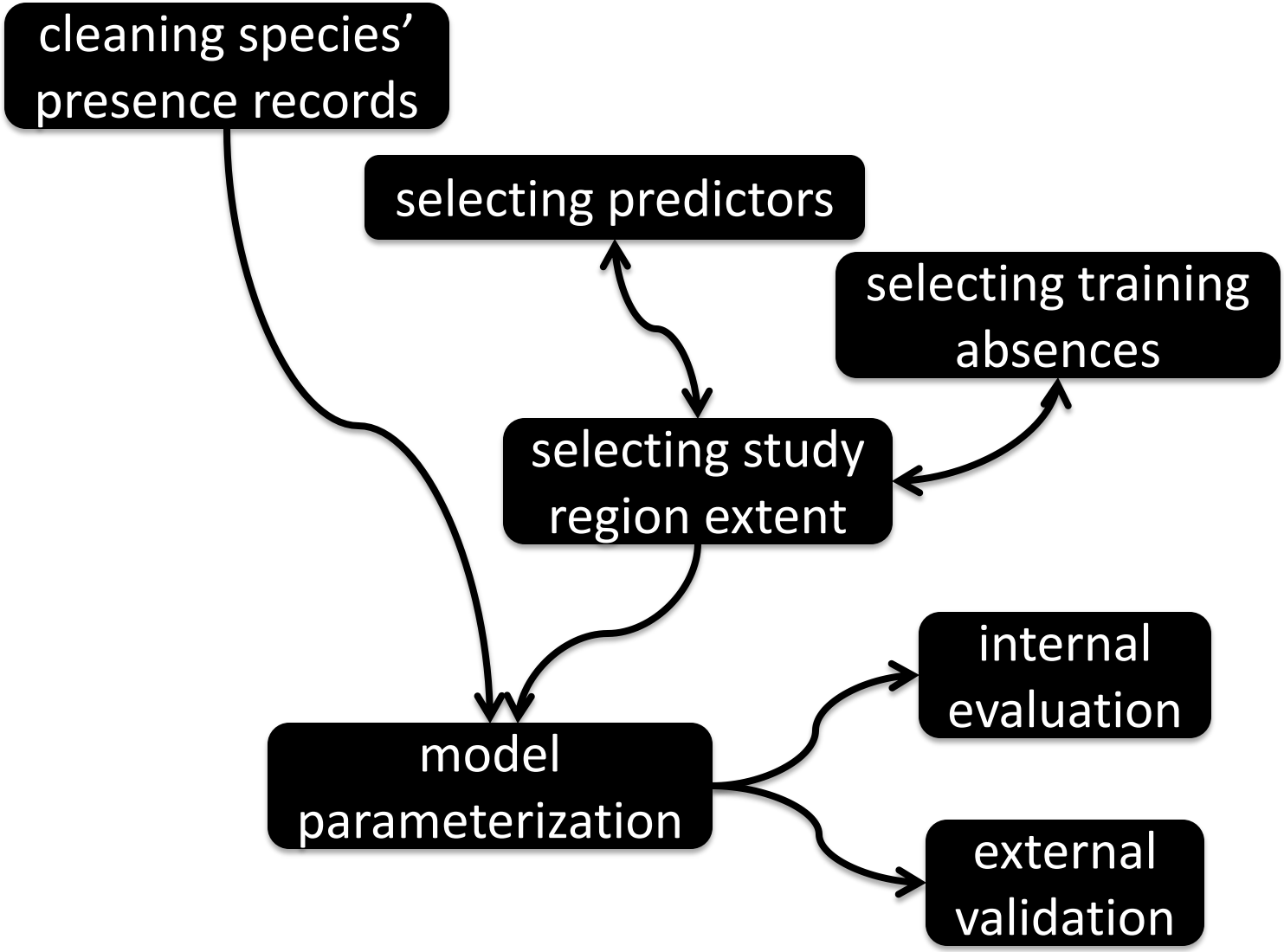


Aegilops triuncialis

Not at equilibrium with environment:

- Current location may not be at optimum though still favorable
- Absences may reflect dispersal limitation more than environmental limitation
- May have shifted niche from native range Broennimann et al. 2007. Evidence of climatic niche shift during biological invasion. Ecology Letters 10:703-709.

“Best practices”
Outline ~ SDM workflow



Best practices: Training records

Online databases

vertebrates

VertNet

- birds, mammals, herptiles
- http://vertnet.org/index.php
- in progress… see MaNIS, ORNIS, HerpNet in the meantime:
  - MaNIS (mammals): http://manisnet.org/
  - ORNIS (birds): http://www.ornisnet.org/
  - HerpNet (herptiles): http://www.herpnet.org/

FishBase

- marine & freshwater
- http://www.fishbase.org

plants

TROPICOS

- worldwide, but especially tropics
- http://www.tropicos.org/

any

Global Biodiversity Information Facility (GBIF)

- http://www.gbif.org/ (records must be carefully vetted!)

Map of Life (in progress)

- http://www.mappinglife.org/

US State Heritage Programs ($)

TIP: Data portals often share data (e.g., FishBase draws from GBIF), so if you use more than one beware of duplicate records.

Best practices: Training records

Data cleaning

- Are species' names correct? (Beware of synonyms, subspecies/variety declensions, etc.).
- Was the species accurately identified? Lozier et al. 2009. Predicting the distribution of Sasquatch in western North America: Anything goes with ecological niche modeling. Journal of Biogeography 36:1623-1627.
  - Is species outside known range? (compare to atlases, expert knowledge)
  - Does lat/long match country/state/province/etc. in which specimen was supposed to have been collected?
  - Does recorded elevation match elevation from GIS? (Note: "0" often used as "no data")

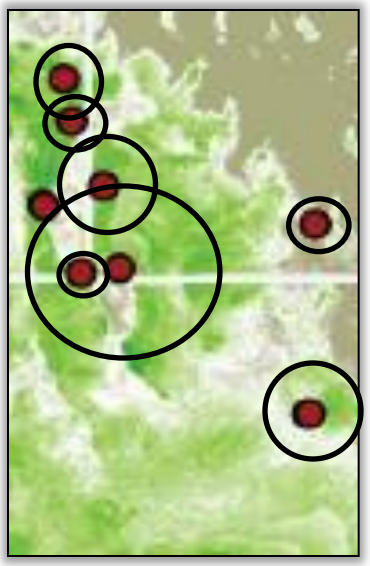
Brown bears mistaken for Sasquatch

Best practices: Training records

Coordinate uncertainty

Coordinate uncertainty: Uncertainty in location of record, often expressed in meters or qualitatively

- Models perform better with less error, but **up to 10 km error** doesn't degrade performance substantially Graham et al. 2008. The influence of spatial errors… J. Applied Ecology 45:239-247… but see Fernandez et al. 2009. Locality uncertainty… Biodiversity Informatics 6:36-52.
- More important for **topographically complex regions** Naimi, et al. 2011. Spatial autocorrelation in predictors reduces the impact of positional uncertainty in occurrence data on species distribution modeling. Journal of Biogeography 38:1497-1509.
- Will generally tend to **cause overprediction** of range (if error is not systematic)
- Beware of **impossibly small errors** (e.g., < few tens of meters before GPS, <~3 m before year 2000)



Specimens are often georeferenced to the middle of a county, state, or even country if locality descriptions are lacking.

Country	Upper	Lower	Elevation	Latitude	Longitude	Date	Qual.	Collectors	Coll No
United States	Alabama	Jefferson		[33°35'00"N]	[086°52'00"W]	19 apr 1931		Ernest J. Palmer	28975
United States	Arkansas	Carroll		[36°20'00"N]	[093°31'00"W]	29 apr 1926		Ernest J. Palmer	29826
United States	Arkansas	Marion		[36°14'00"N]	[092°41'00"W]	19 apr 1920		Ernest J. Palmer	17241
United States	Missouri	Barry		[36°42'18"N]	[093°50'13"W]	30 apr 1926		Ernest J. Palmer	29828

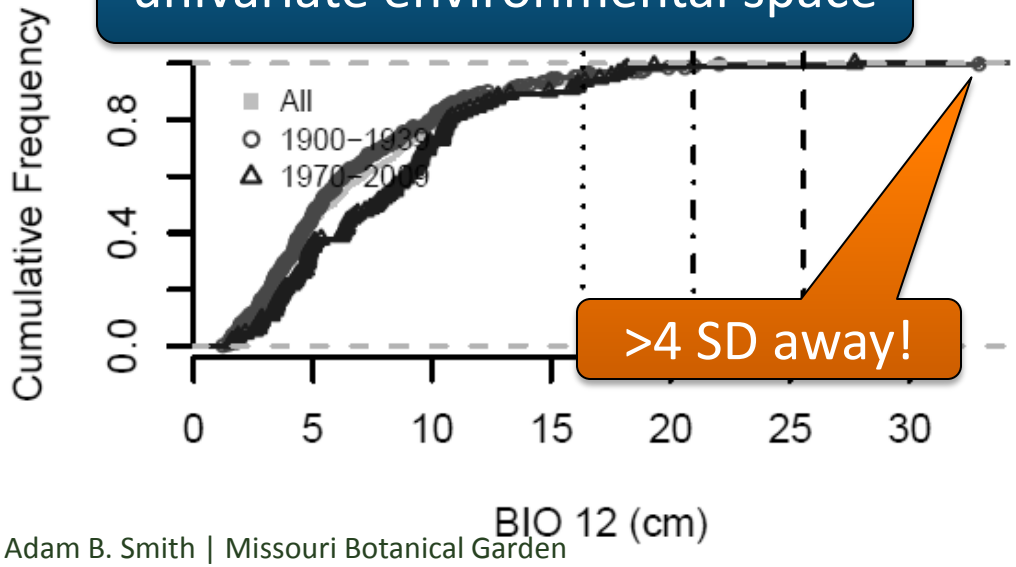
Best practices: Training records

Coordinate uncertainty

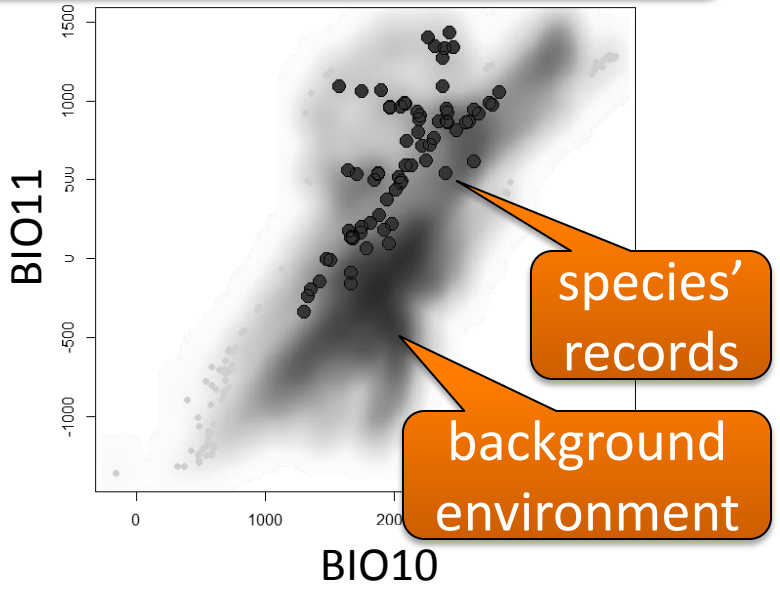
- Beware of records georeferenced to the zoo or garden where it lived (search for “rear”, “raise”, “zoo”, “garden”, “captive”, “captivity”, etc.).
- Can use environmental outliers as surrogate for mistaken georeferencing Chapman 2005. Principles and Methods of Data Cleaning: Primary Species and Species-Occurrence Data, Version 1.0. Report for GBIF, Copenhagen.

See BioGeomancer manual for best practices in geo-referencing (www.gbif.org/orc/?doc_id=1288)

univariate environmental space



bivariate environmental space

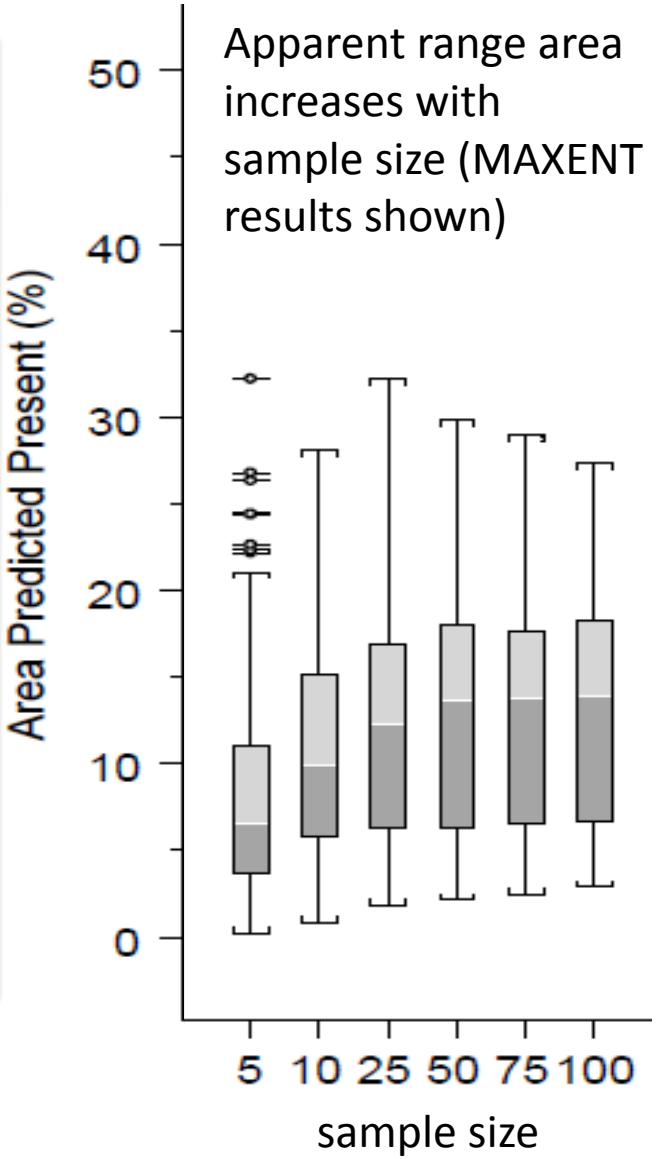


Best practices: Training records

How many records are necessary?

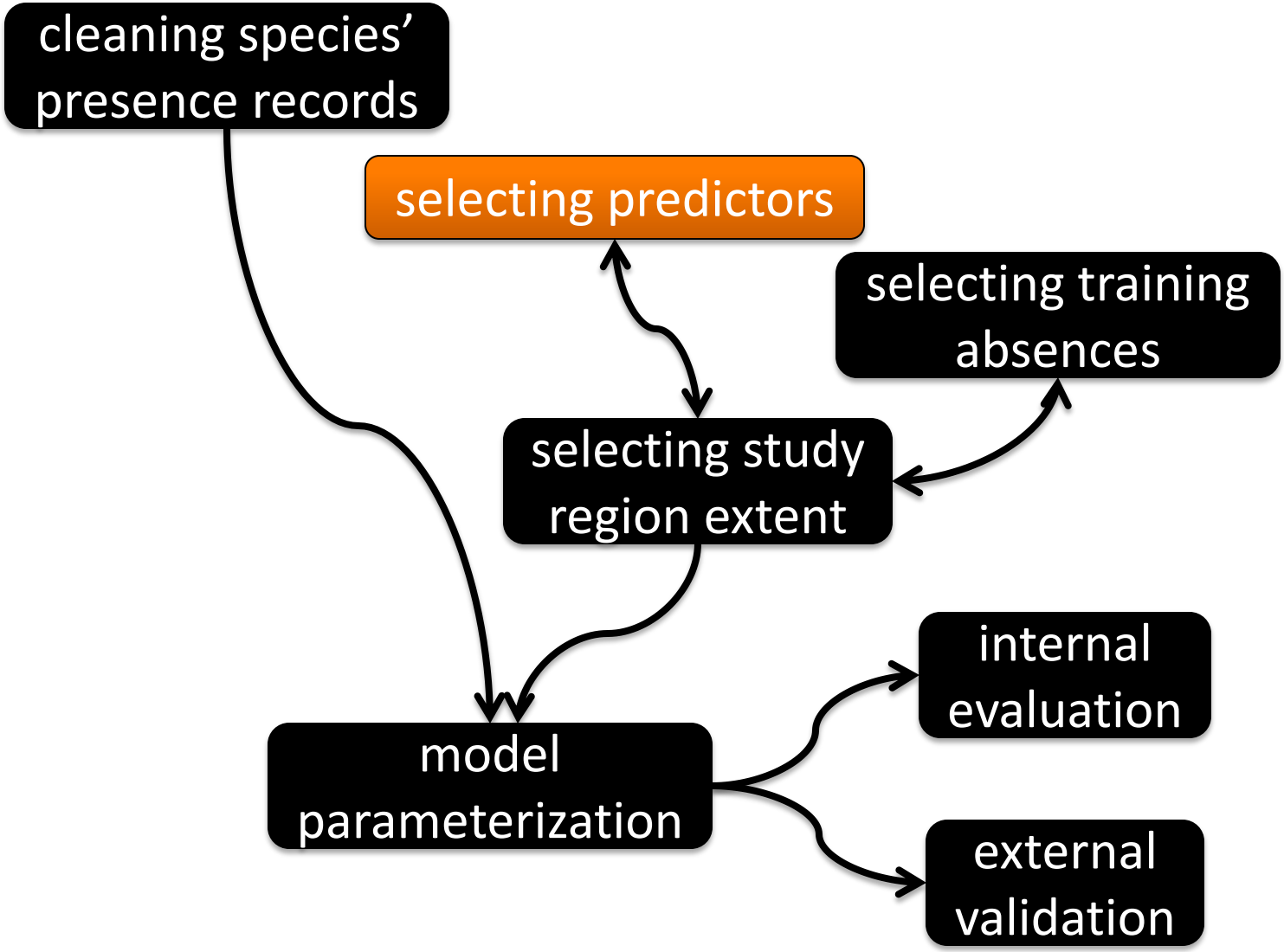
- Greater number of records increases apparent range (small sample size→ ***underprediction* of range**) Hernandez 2006. The effect of sample size… Ecography 29:773-785.
- Generally need **at least 30** training records (not including test records) for stable performance Wisz et al. 2008. Effects of sample size… Diversity and Distributions 14:763-773.
- Presence-only techniques: **MAXENT less affected** by small sample size than other methods Wisz et al. 2008
- Presence-absence techniques: **GAMs need 20-30 *per term*** for stable performance and **BRTs need ≥32** to even work (pers. obs.)

Apparent range area increases with sample size (MAXENT results shown)



Best practices: Predictors

Selecting and preparing predictors



Best practices: Predictors

Predictors: Climate data

WORLDCLIM

- averages over 1950-2000 + IPCC futures + last glacial maximum
- worldwide
- ~1 km resolution
- http://www.worldclim.org/
- Hijmans et al. 2005. Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology 25:1965-1978.

PRISM

- daily/monthly/period from 1900 to 2012
- US + parts of Canada
- 800-m and 4-km resolution
- some parts not free, some free
- http://www.prism.oregonstate.edu/
- Daly et al. 2001. High-quality spatial climate data sets for the United States and beyond. Transactions of the American Society of Agricultural Engineers, 43:1957-1962.

Bias-corrected GCM projections

- 1950 to 2099 (16 GCMs)
- Southern Canada, US, northern Mexico
- ~12-km resolution
- http://gdo-dcp.ucllnl.org/downscaled_cmip_projections/

RNCEP

- Weather station data
- 2.5 ° resolution
- https://sites.google.com/site/michaelukemp/rncep
- Kemp et al. 2012. RNCEP: Global weather and climate data at your fingertips. Methods in Ecology and Evolution 3:65-70. (R package).

Best practices: Predictors

Predictors: Other data

Elevation

- SRTM Digital Elevation Dataset (≥250-m resolution worldwide): http://srtm.csi.cgiar.org/
- USDA's Geospatial Data Gateway (≥30-m resolution for US): http://datagateway.nrcs.usda.gov/

Hydrology

- USDA's Geospatial Data Gateway (above)
- Bias-corrected GCM projections (previous slide)

Soils

SIEMEM CONUS (quantitative soil characteristics for US): http://www.soilinfo.psu.edu/

Harmonized World Soils Database (not so accurate for North America): http://www.iiasa.ac.at/Research/LUC/External-World-soil-database/HTML/index.html

Best practices: Predictors

Selecting predictors

"Expert-driven" selection

proximate vs. distal

Choose predictors that **directly affect the distribution** of the species.

Example: Woodrats (*Neotoma* spp.) are highly sensitive to heat, so choose temperature-related variables (vs. elevation which correlates with temperature).

resource and conditions

Choose predictors you expect to serve as **resources** or establish direct **physiological limits**.

Condition example: Many temperate seeds require cold temperatures to overripen before breaking dormancy, so select variables related to winter temperatures as a predictor.

"Model-driven" selection

Calculate model will all possible predictors and select those with greatest "importance" in model.

Real practice is to use a mix of these.

Best practices: Predictors

Number of predictors

Fewer predictors gives the model less ability to rule out a particular area. ***Fewer predictors → overprediction.***

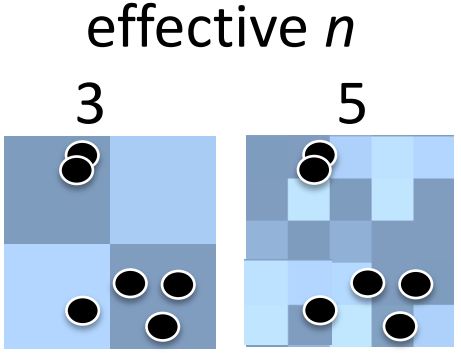
GAM for the highly endemic *Lesquerella perforata*, n=20

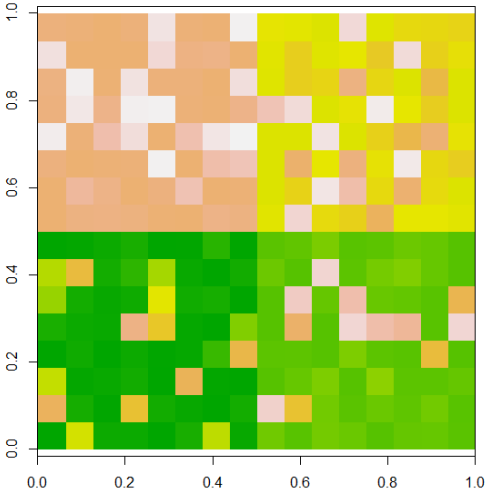
Best practices: Predictors

Resolution

- In practice, you **use what you can get**!
- Ideally, resolution would be **at the scale of range-limiting processes**. Otherwise, you can only expect predictors to be ***correlated*** with limiting factors.
- **Finer resolution** (smaller cells) means you can **use more records** because there are fewer duplicate points in cells.
- Mixing coarse- with fine-grained predictors: Can create "**blockiness**" in predictions.
- **Downscaling** coarse-grained predictors may be OK for those with long characteristic distance of spatial autocorrelation.
- Coarse predictors → ***overprediction***

Menke et al. 2009. Characterizing and predicting species distributions across environments and scales... Global Ecology and Biogeography 18:50-63.





Best practices: Predictors

Accuracy of predictors

- **Rarely known** Parra & Monahan. 2008. Variability in 20th century climate change reconstructions and its consequences for predicting geographic responses of California mammals. Global Change Biology 14:2215-2231.
- If projecting to future climates, should use GCM that **best predicts current climate in region of interest** Beaumont et al. 2008. Why is the choice of future climate scenarios for species distribution modeling important? Ecology Letters 11:1135-1146.

  …can be corrected using **bias-correction** (see CMIP Project at http://gdo-dcp.ucllnl.org/downscaled_cmip_projections/)
- Can perform **sensitivity analyses** on predictors to determine effect of predictor inaccuracy

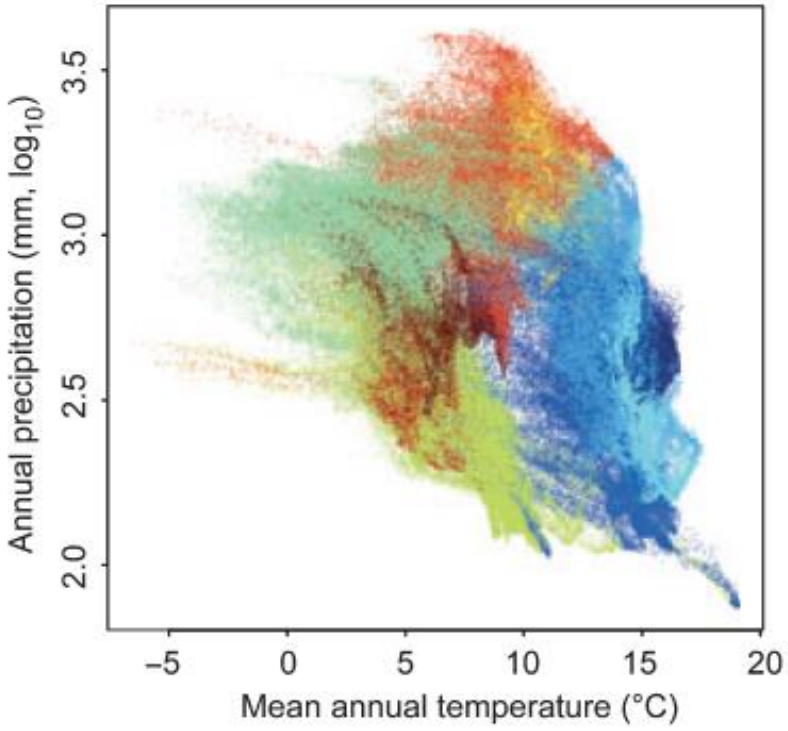
Best practices: Predictors

Problems with highly correlated predictors

- **Regression-based methods** (GLMs, GAMs) unstable
- **Maxent** less affected by correlations between predictors

  Elith et al. 2011. A statistical explanation of Maxent for ecologists. Diversity & Distributions 17:43-57.
- **If correlations change in future** and model incorporates interactions, may have to extrapolate in "covariance space" Smith *In press* The relative influence of temperature, moisture, and their interaction on range limits of mammals over the past century. Global Ecology and Biogeography.

mean annual temperature vs. precipitation for California

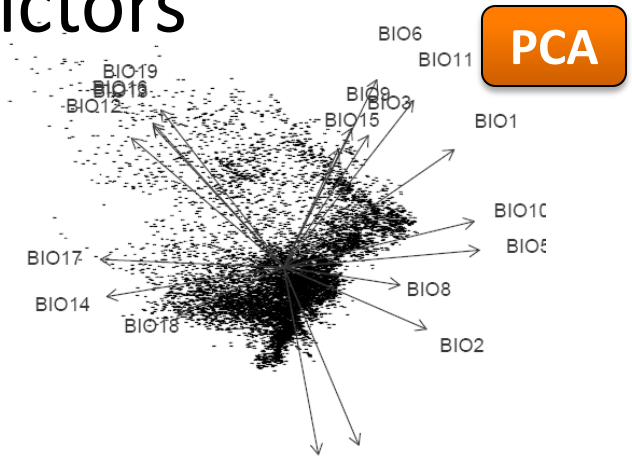


Ackerly et al. 2010. The geography of climate change: Implications for conservation biogeography. Diversity and Distributions 16:476-487.

Best practices: Predictors

Managing highly correlated predictors

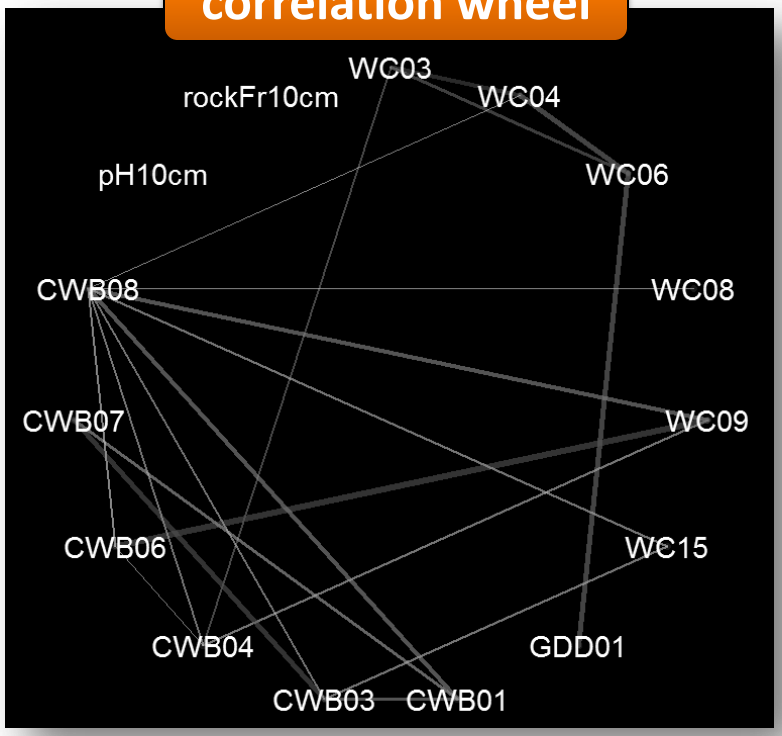
- Use **PCA** axes as predictors (e.g., Loarie et al. 2008. Climate change and the future of California's endemic flora. Public Library of Science ONE 3:e2502.)... **not recommended for Maxent** Elith et al. 2011. (as above)
- Use predictors with pairwise **correlations ≤ some value** (usually 0.6-0.7)



correlations between BIO01 through BIO11 in the western US, |r|≥0.5 highlighted

	BIO01	BIO02	BIO03	BIO04	BIO05	BIO06	BIO07	BIO08	BIO09	BIO10	BIO11
BIO01		0.33	**0.56**	-0.28	**0.84**	**0.85**	-0.15	**0.52**	**0.57**	**0.93**	**0.94**
BIO02	0.33		0.43	0.31	**0.62**	-0.07	**0.63**	0.42	0	0.47	0.17
BIO03	**0.56**	0.43		**-0.68**	0.28	**0.62**	-0.41	0.25	0.41	0.32	**0.71**
BIO04	-0.28	0.31	**-0.68**		0.21	**-0.69**	**0.92**	0.2	**-0.51**	0.1	**-0.59**
BIO05	**0.84**	**0.62**	0.28	0.21		0.49	0.38	**0.51**	0.39	**0.95**	**0.63**
BIO06	**0.85**	-0.07	**0.62**	**-0.69**	0.49		**-0.62**	0.19	**0.71**	**0.62**	**0.96**
BIO07	-0.15	**0.63**	-0.41	**0.92**	0.38	**-0.62**		0.25	-0.41	0.2	-0.46
BIO08	**0.52**	0.42	0.25	0.2	**0.51**	0.19	0.25		-0.22	**0.6**	0.35
BIO09	**0.57**	0	0.41	**-0.51**	0.39	**0.71**	-0.41	-0.22		0.4	**0.67**
BIO10	**0.93**	0.47	0.32	0.1	**0.95**	**0.62**	0.2	**0.6**	0.4		**0.75**
BIO11	**0.94**	0.17	**0.71**	**-0.59**	**0.63**	**0.96**	-0.46	0.35	**0.67**	**0.75**	

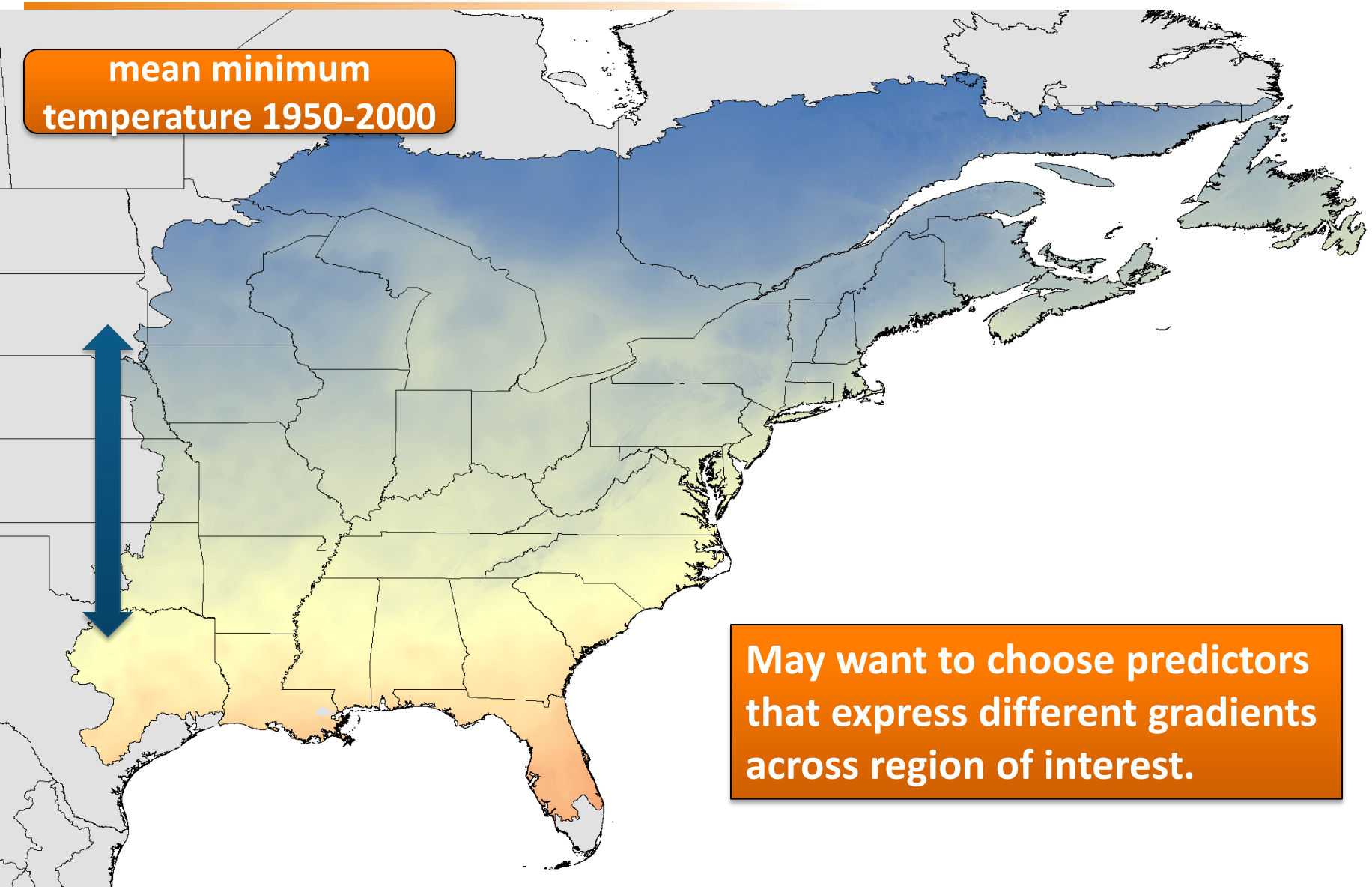
correlation wheel



highly correlated factors connected

Best practices: Predictors

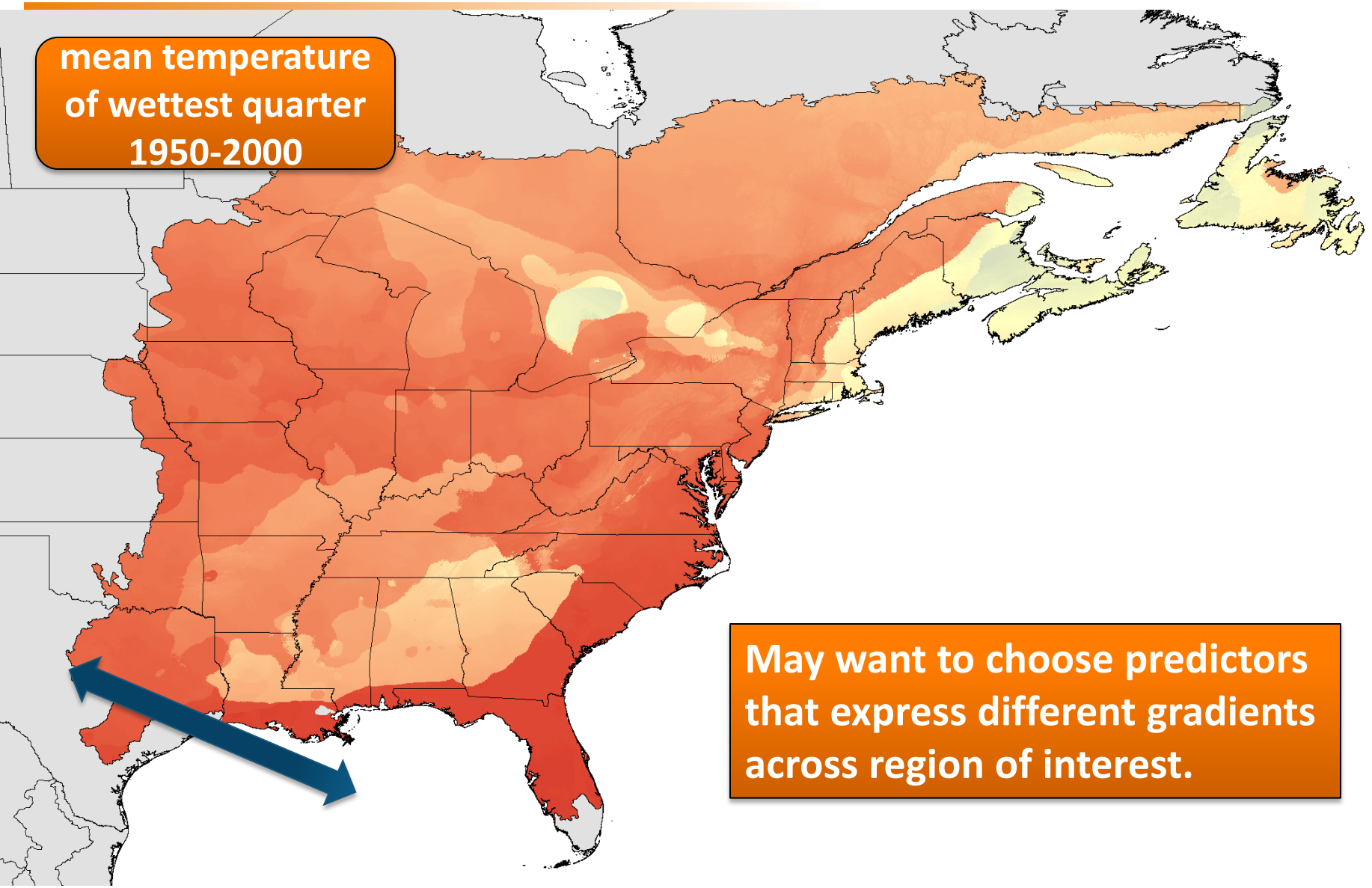
Gradients



May want to choose predictors that express different gradients across region of interest.

Best practices: Predictors

Gradients



May want to choose predictors that express different gradients across region of interest.

Best practices: Predictors

Dynamic vs. static vs. dynamic-but-static

Static predictor: Does not change with time (e.g., elevation)

Dynamic predictor: Changes with time: (e.g., precipitation)

If projecting to new region/time, **avoid static predictors correlated with dynamic predictors** (e.g., elevation and temperature).

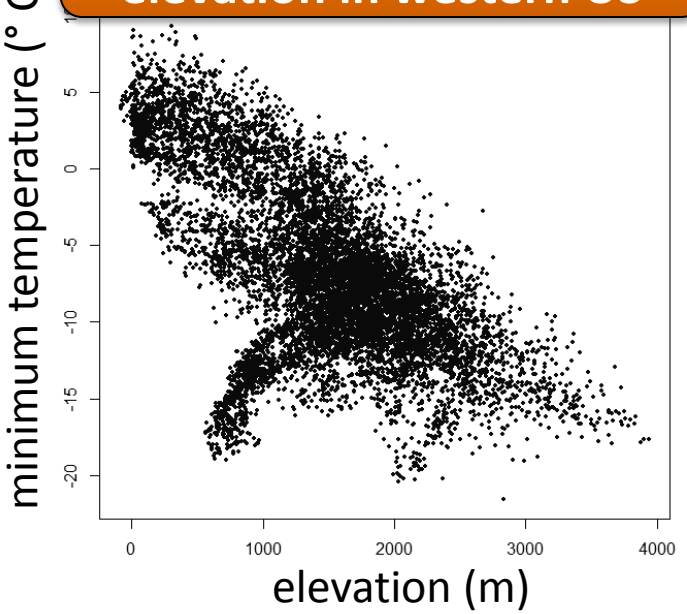
"**Dynamic-but-static:**" Will change with time but don't have data for other time period (e.g., human land use)

Options:

1) Don't use

2) Assume does not change (→ ***underprediction* or *overprediction***)

minimum temperature vs. elevation in western US



Including dynamic-but-static factors may not degrade model performance when projected to new time period. Stanton et al. 2012. Combining static and dynamic variables in species distribution models under climate change. Methods in Ecology and Evolution 3:349-357.

Best practices: Predictors

Time span of predictors

Justification for using long-term averages: Any particular year may have unfavorable weather, but species remains because of storage effects, longevity, etc. because **average year is favorable**.

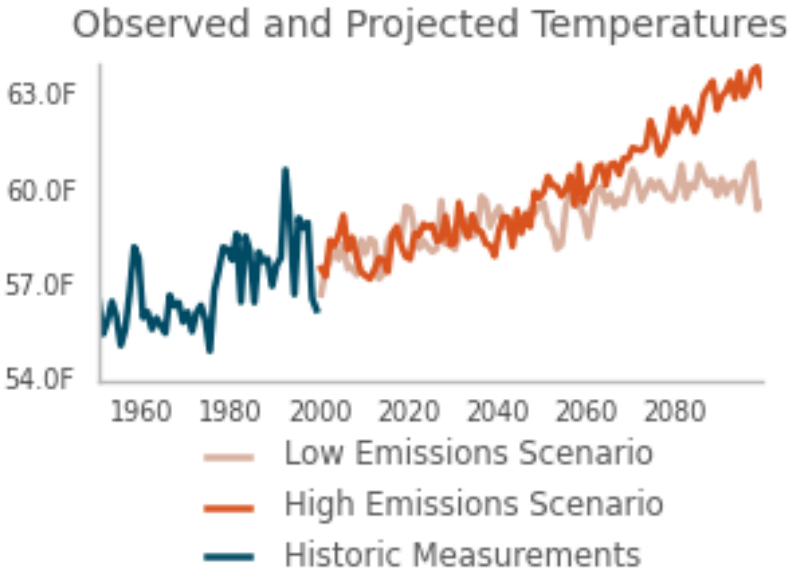
However, yearly variability can determine species' presence…

Long-term averages → ***overprediction*** of range

Reside et al. 2010. Weather, not climate, defines distributions of vagile bird species. PLoS ONE 5:e13569.

Mismatches between presence records and time period of climate can ↑ **error**

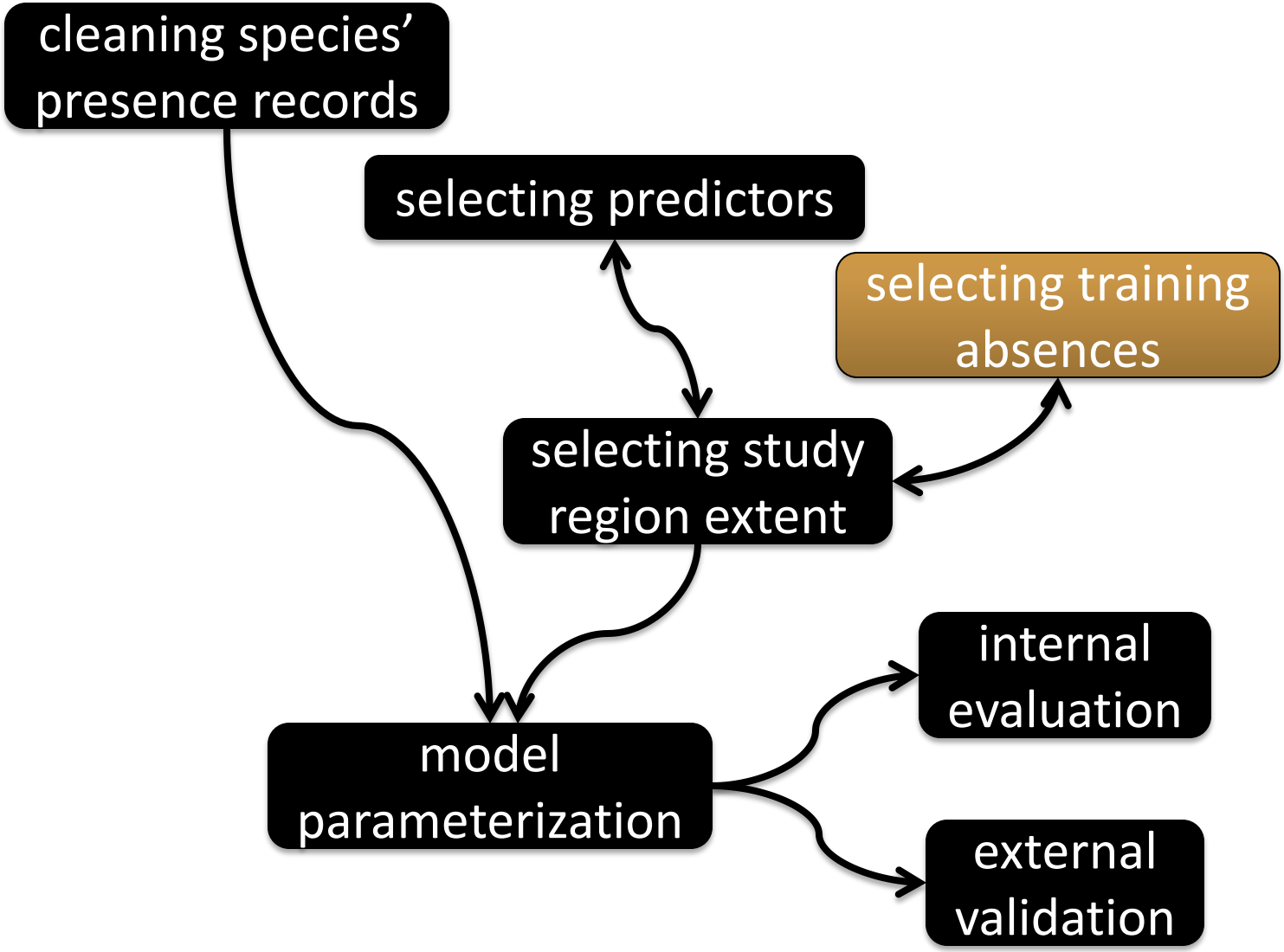
Roubicek et al. 2010. Does the choice of climate baseline matter in ecological niche modeling? Ecological Modeling 221:2280-2286.



Berkeley, CA
(www.cal-adapt.org)

A short course on distribution modeling

"Best practices" ~ SDM workflow



"True" absences

"true" absences available

Absence should be obvious or verified with **occupancy modeling**

MacKenzie et al. 2006 Occupancy Estimation and Modeling. Elsevier, Amsterdam. 324 pp.

For R code see Royale & Dorazio 2008 Hierarchical Modeling and Inference in Ecology. Academic Press.

Should use presence/absence SDMs… (GAMs, BRTs, etc.) should be better than presence-only SDMs

Elith et al. 2011. A statistical explanation of Maxent for ecologists. Diversity & Distributions 17:43-57.

Absences should be **weighted** equally to presences in presence/absence SDMs Maggini et al. 2006. Improving generalized regression analysis… J. Biogeography 33:1729-1749.

Sample bias in presences cancels sample bias in absences if they have the same bias! Phillips et al. 2009. Sample selection bias and presence-only distribution models.. Ecological Applications 19:181-197.

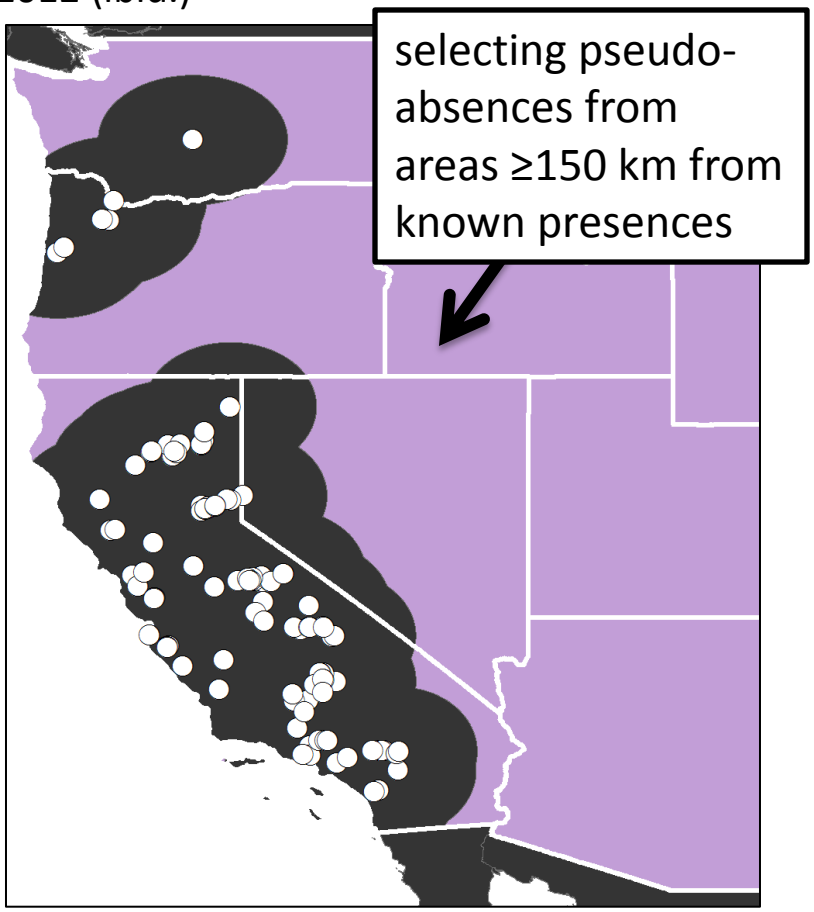
Note: Maxent is not formulated to use true absences! Recall the ratio Pr(env|pres)/Pr(env)… if replace Pr(env) with Pr(env|absent), then won't be true to Bayes' formulation of Pr(pres|env).

Pseudoabsences

"true" absences unavailable

- Can use presence/absence SDMs with "**pseudoabsences**" (sites where species is expected not to exist) Barbet-Massin et al. 2012. Selecting pseudo-absences for species distribution models: How, where and how many? Methods in Ecology and Evolution 3:327-338.
- Pseudoabsences should be **weighted equally to presences** in presence/absence SDMs Barbet-Massin et al. 2012 (ibid.)
- Selecting regions geographically/environmentally far from presences → ***overprediction*** Barbet-Massin et al. 2012 (ibid.)

10:1 ratio of pseudoabsences: presences generally recommended Barbet-Massin et al. 2012 (ibid.)



Background sites

"true" absences unavailable

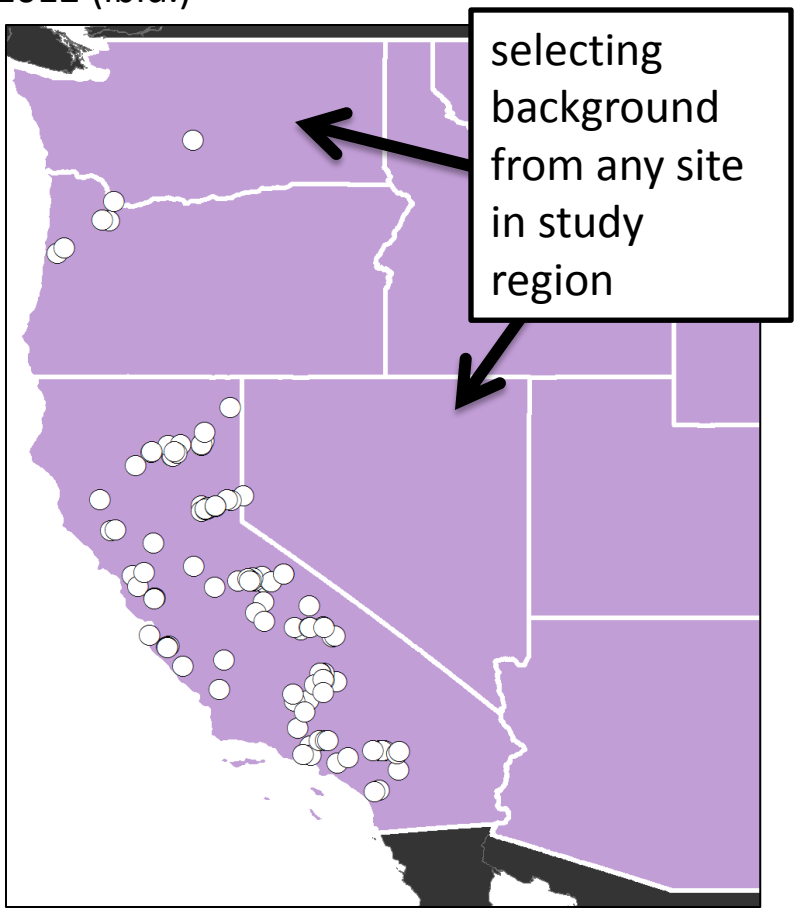
Can use presence/absence SDMs or Maxent with "**background**" sites: Randomly chosen from across landscape

Pseudoabsences should be **weighted equally to presences**

Barbet-Massin et al. 2012. Selecting pseudo-absences for species distribution models: How, where and how many? Methods in Ecology and Evolution 3:327-338.

→ ***underprediction*** Barbet-Massin et al. 2012 (ibid.)

10:1 ratio of pseudoabsences: presences generally recommended Barbet-Massin et al. 2012 (ibid.)



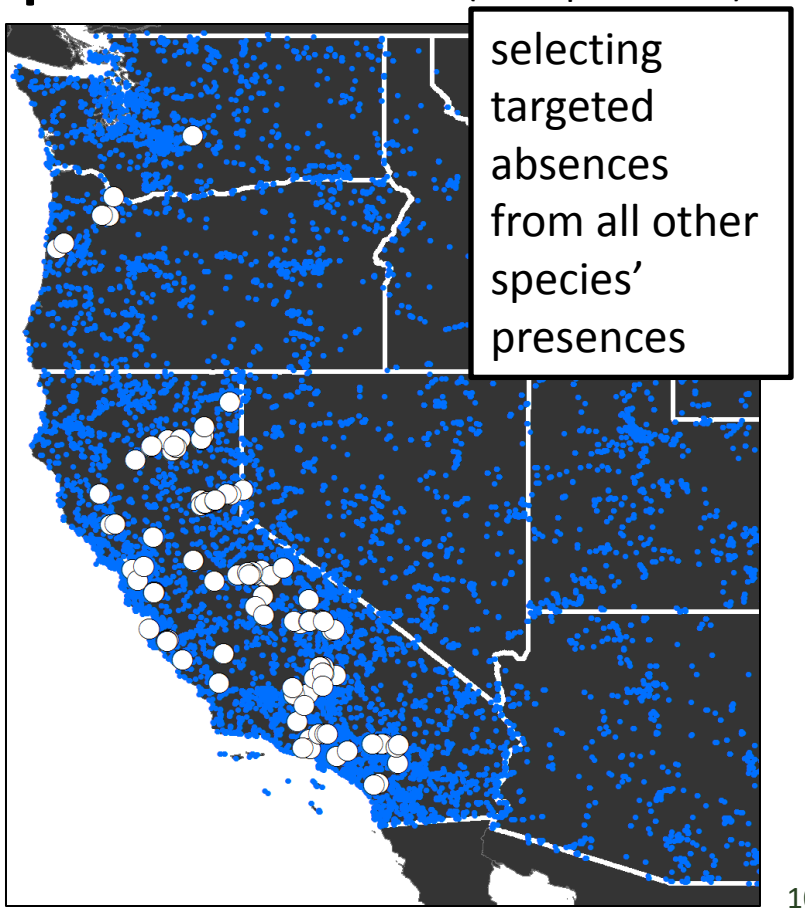
Targeted "absences"

"true" absences unavailable & presences are biased

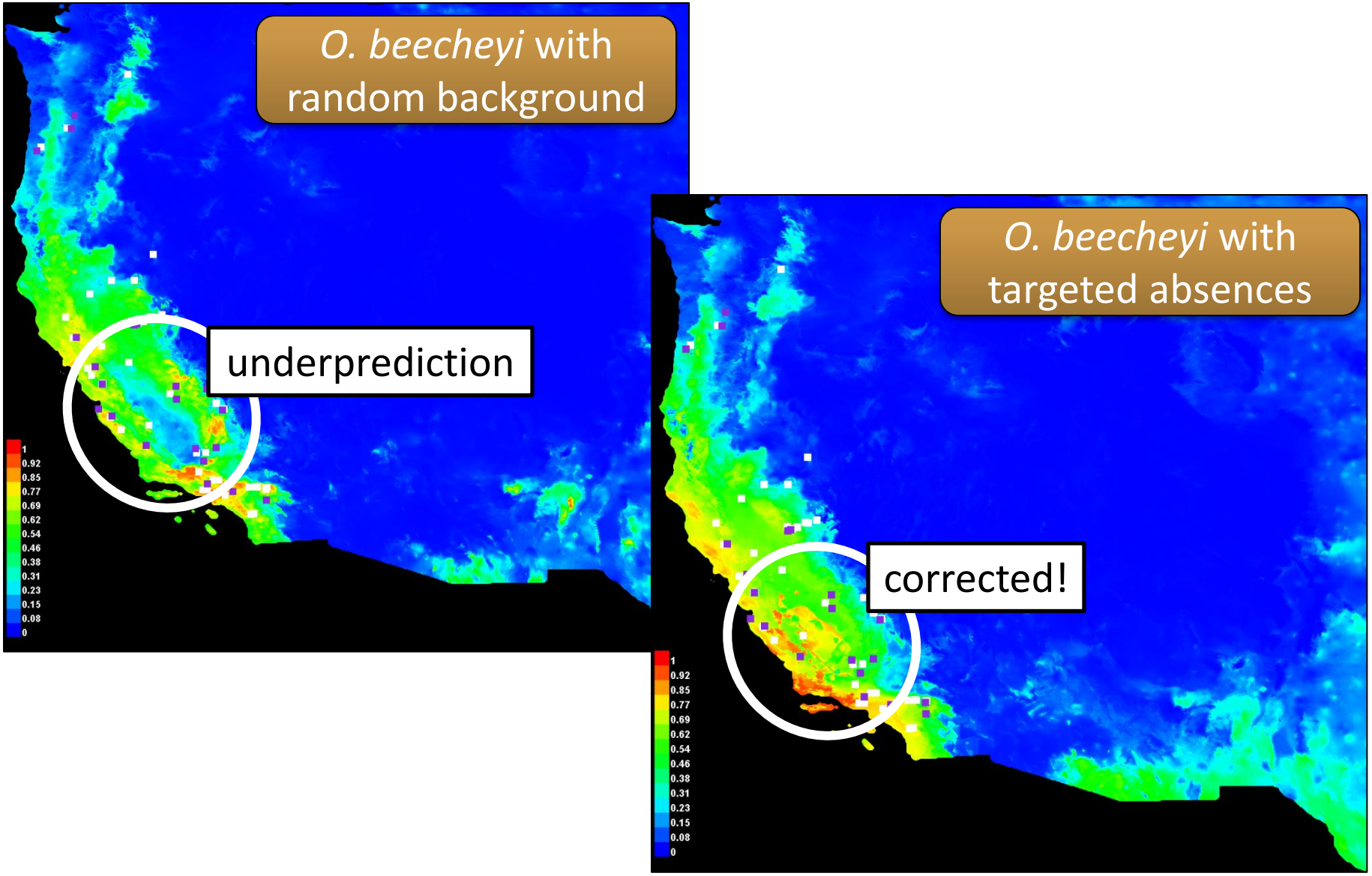
Can use presence/absence SDMs **or Maxent** with "**targeted**" absences: Chosen from sites with same bias as presence sites… e.g., presences of species observed with similar techniques

Targeted absences should be **weighted** equally to presences (not an issue in Maxent) Barbet-Massin et al. 2012. Selecting pseudo-absences for species distribution models: How, where and how many? Methods in Ecology and Evolution 3:327-338.

Targeted absences correct for "extent" bias Phillips et al. 2009. Sample selection bias and presence-only distribution models... Ecological Applications 19:181-197 **but may produce "placement" bias** (ABS, pers. obs.)



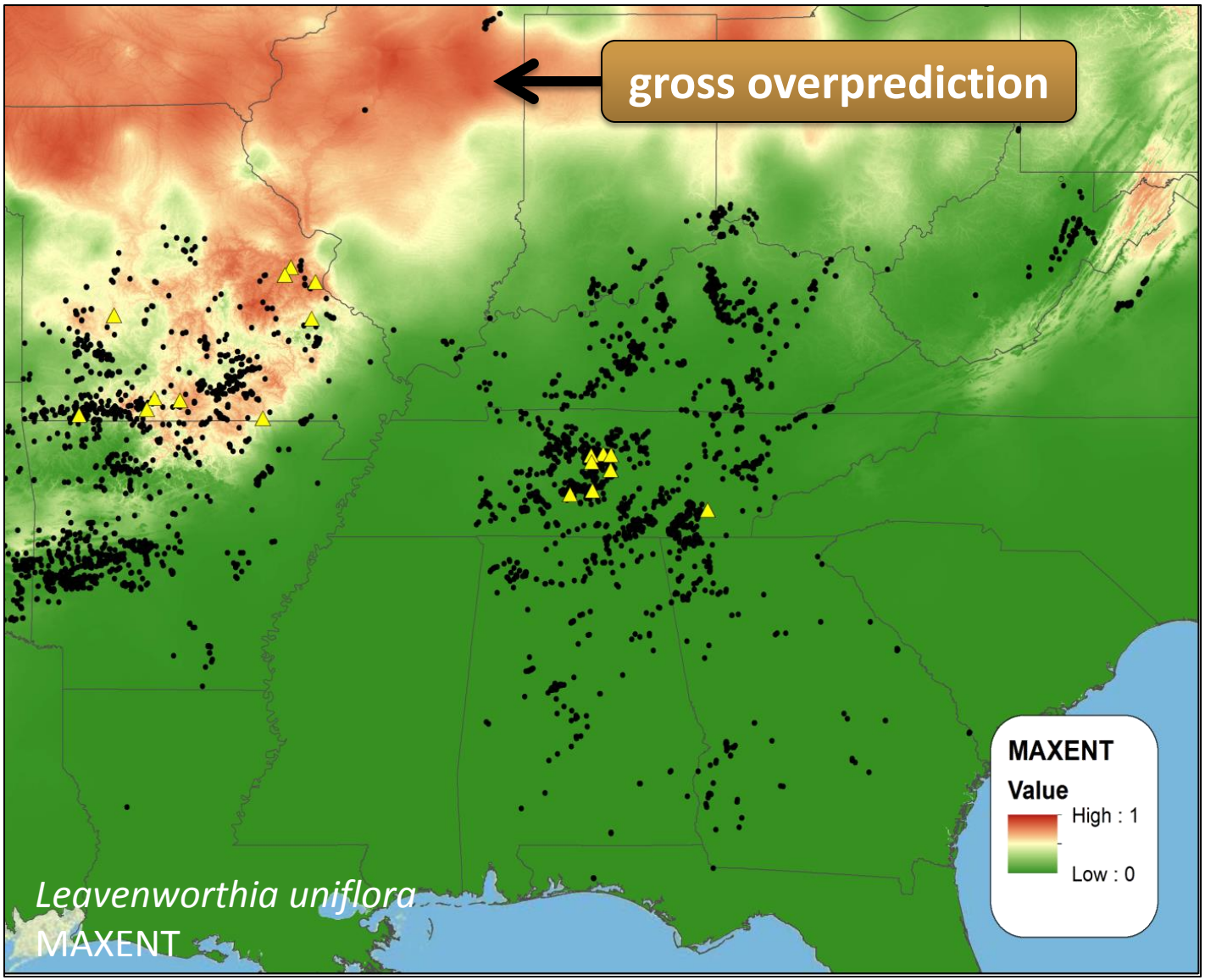
Correcting bias with targeted "absences"



Best practices: Training absences/pseudoabsences/background

Caution with absences/pseuodabsences/background/targeted sites

Geographically restricted absences, pseudoabsences, background, and targeted absences can → dubious overprediction



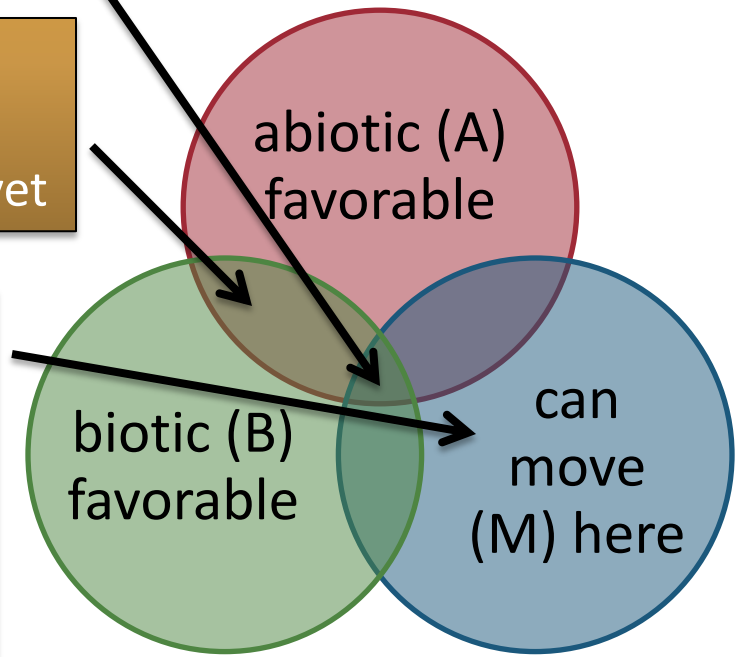
Biotic-Abiotic-Movement (BAM) schema

"Non-presences" should not include areas into which species has had no chance to disperse…

species present here

could exist here but can't/hasn't dispersed here yet

has "sampled" here in past but wasn't favorable… **want "non-presences" from here**



fundamental niche = A (or A ∩ B)
realized niche = A ∩ B ∩ M

how do you estimate M?

- Estimate potential dispersal
- Train model with overprediction bias then sample non-presences from this area
- Project to long ago and work forward iteratively, noting which suitable areas were adjacent to "occupied" areas Barve et al. 2011. The crucial role of the accessible area… Ecological Modeling 222:1810-1819.

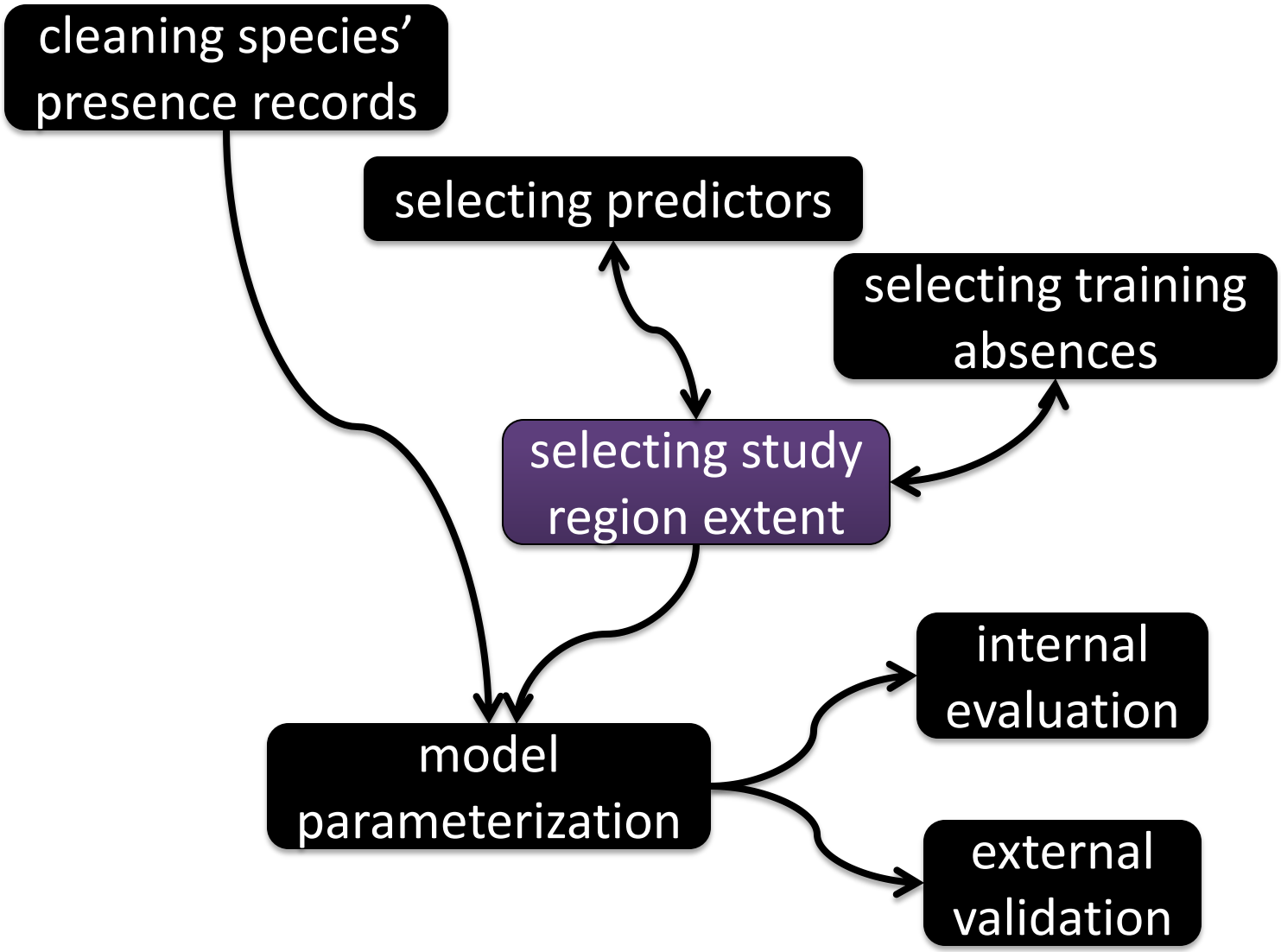
Anderson & Raza. 2010. The effect of the extent of the study region… Journal of Biogeography 37:1378-1393.

BAM diagram: Soberón & Nakamura. 2009. Niches and distributional areas: Concepts, methods, and assumptions. PNAS 106 (Suppl. 2):19644-19650.

See also Peterson et al. 2011. Ecological Niches and Geographic Distributions. Princeton.

Best practices: Study region extent

Effects of scale

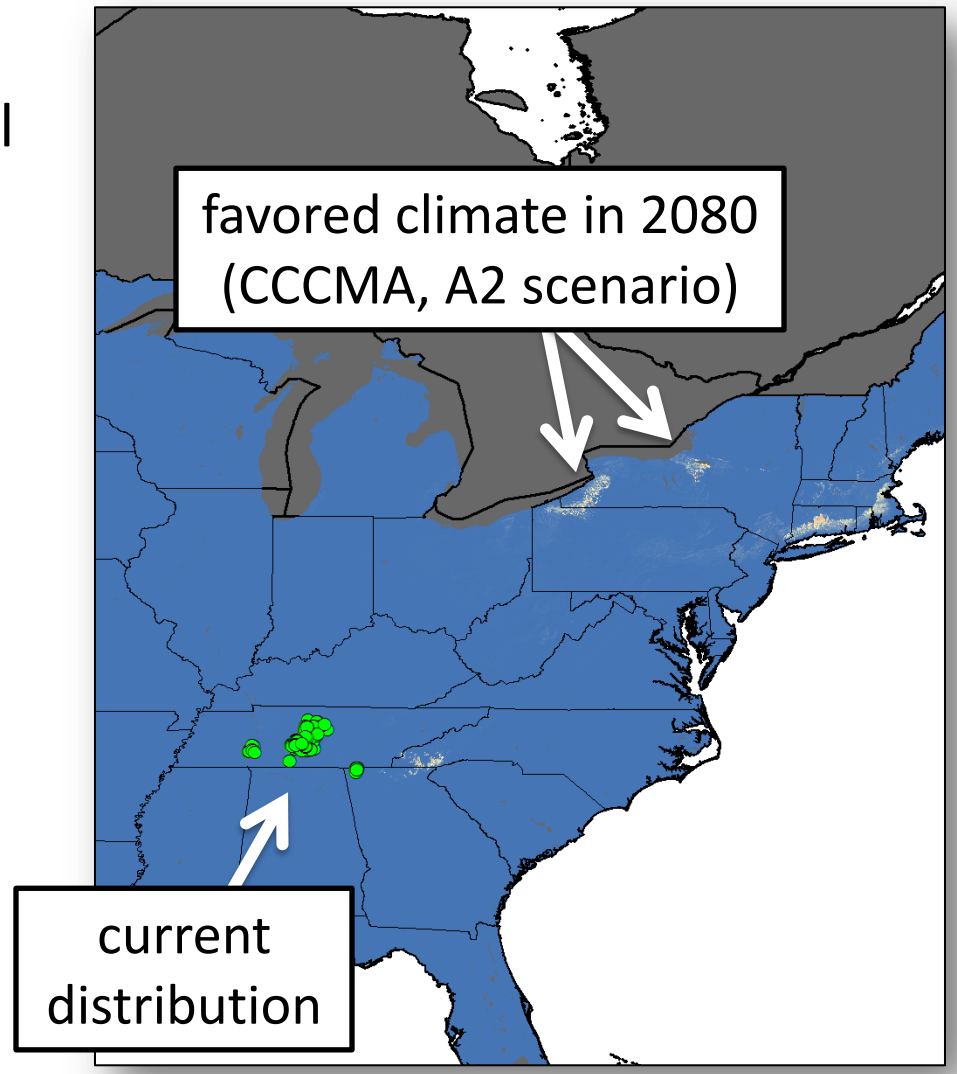


Delineation

- Natural (e.g., islands)
- Biological (e.g., biogeographical provinces)
- Semi-natural (e.g., WWF Ecoregions, EPA Ecoregions, USGS Physiographic Provinces)
- Political (e.g., national borders)
- Arbitrary (e.g., bounding box)

Should include total range of species

Also look at area needed to predict ranges in future



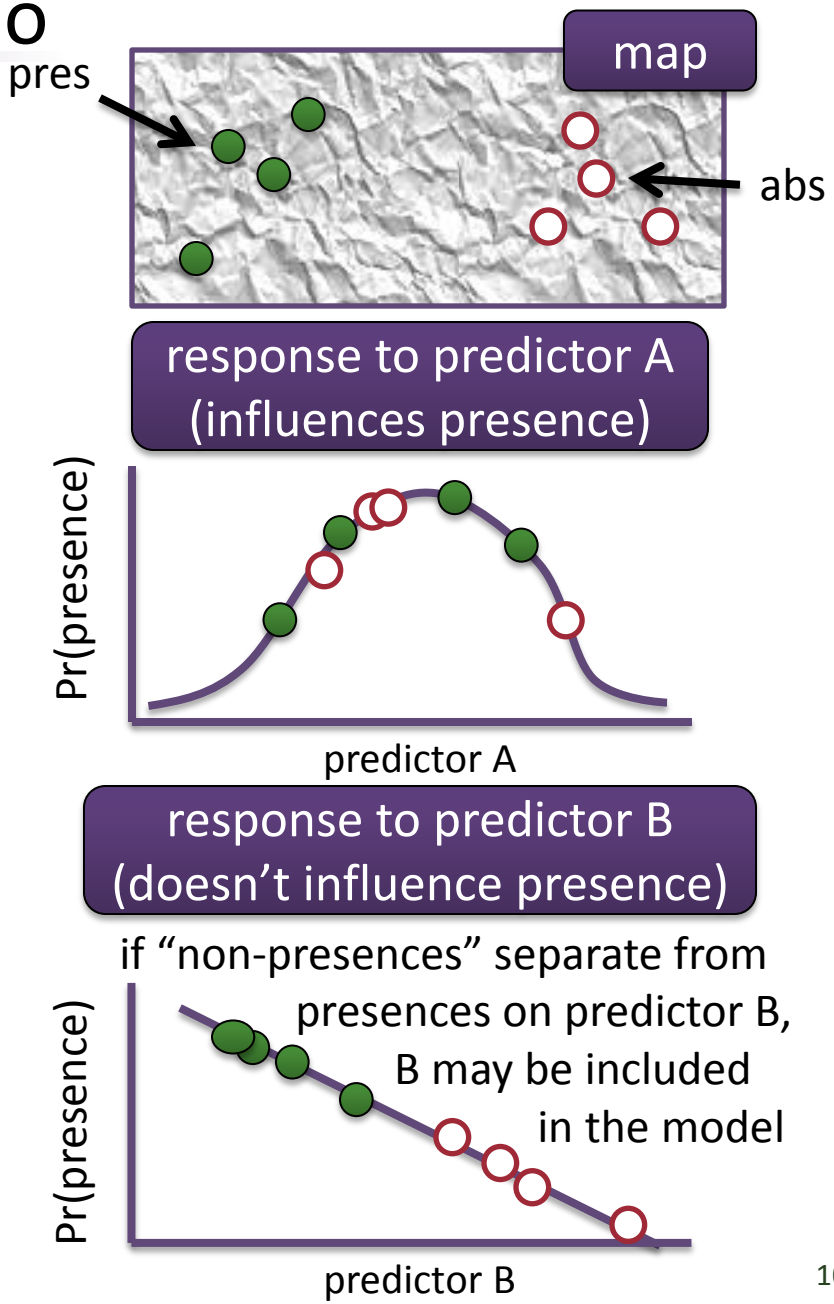
Leavenworthia exigua var. *exigua*
MAXENT

Range size : study extent ratio

Range size : study extent ratio
affects model training

Influences area from which "non-presences" can be chosen for training/evaluation (**larger extent ↑ apparent accuracy**)

↓ range size : study extent can lead to predictor combinations that **spuriously "favor" absence**, → ***underprediction*** Anderson & Raza 2010 The effect of the extent of the study region… J Biogeography 37:1378-1393.

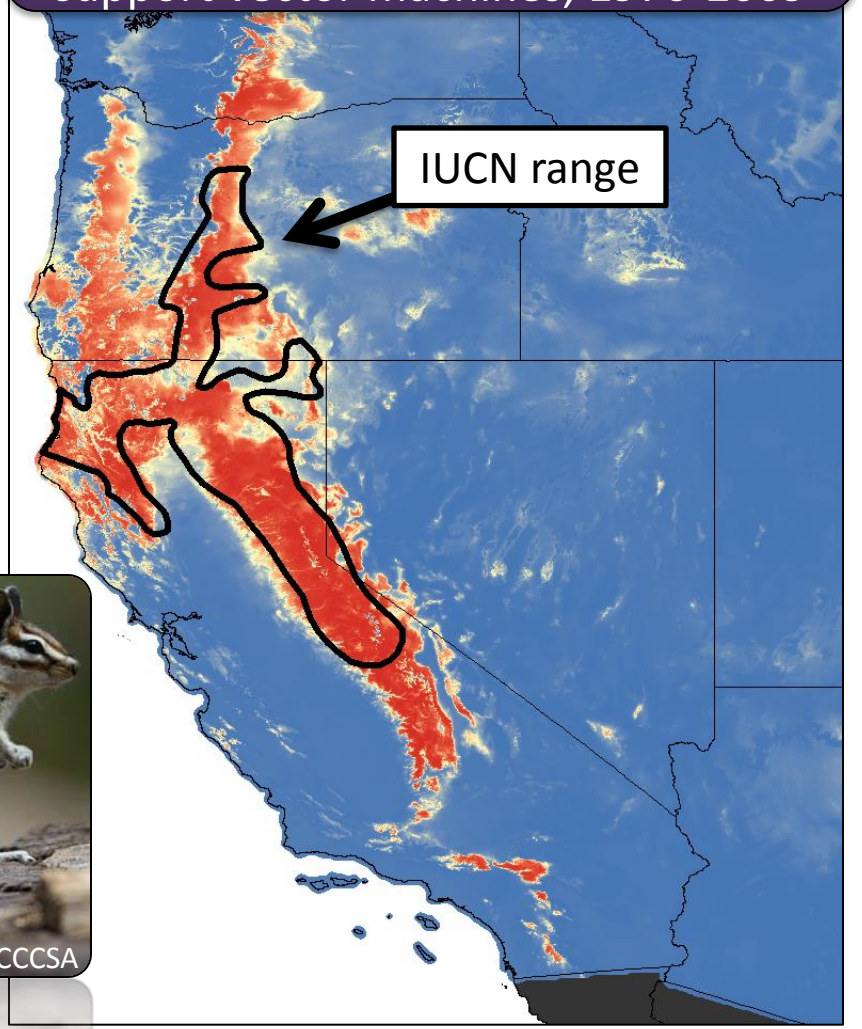


Range size : study extent ratio

↓ range size : study extent may present areas with favorable environments that cannot be dispersed to… → ***overprediction(?)***

Anderson & Raza 2010 The effect of the extent of the study region… J Biogeography 37:1378-1393.

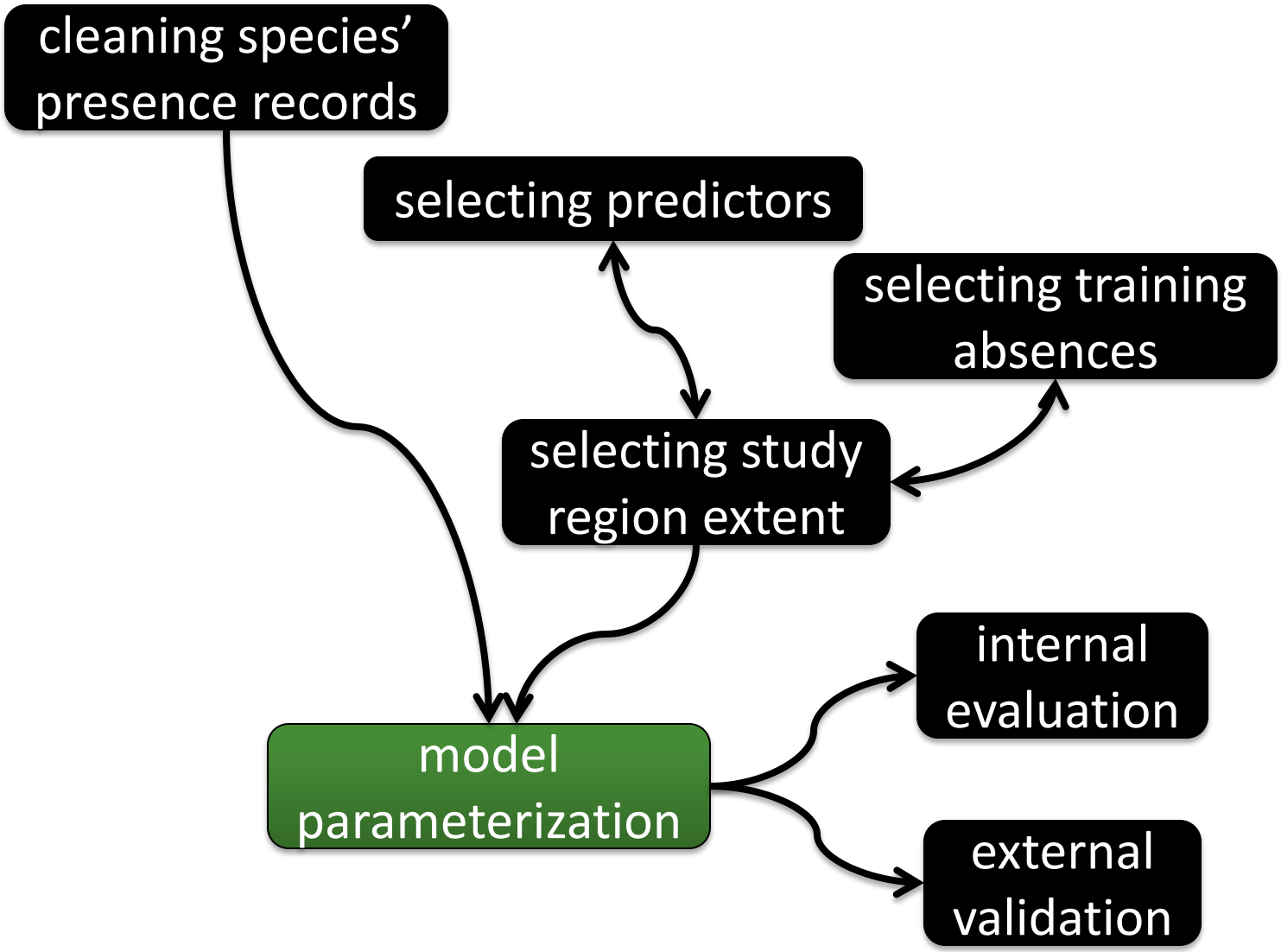
Tamias senex (shadow chipmunk) modeled with support vector machines, 1970-2009




Maholyoak/CCCSA

Best practices: Model parameterization

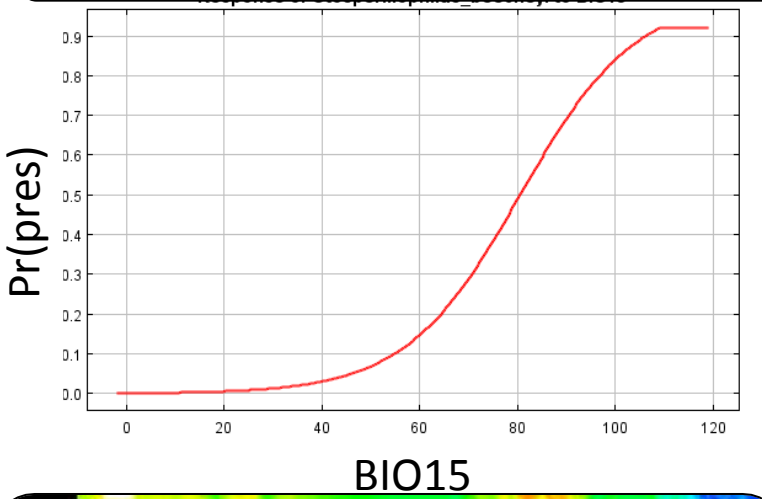
Training the model

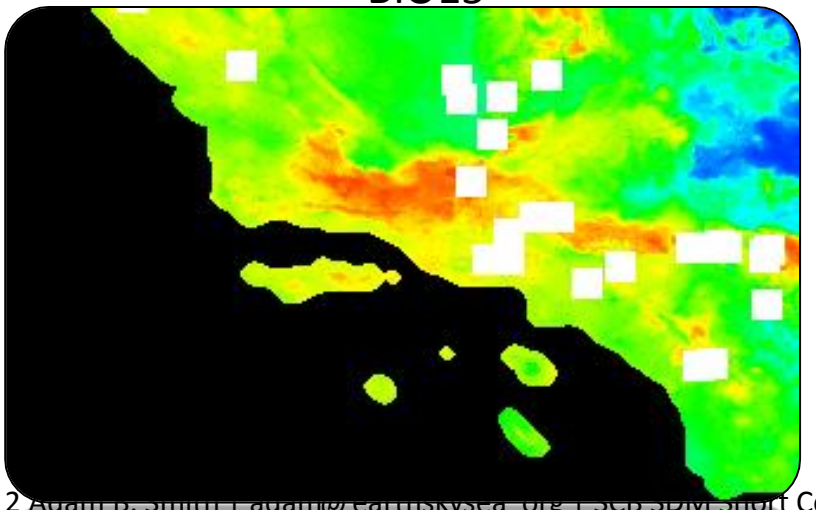


Smooth vs. not smooth

smooth

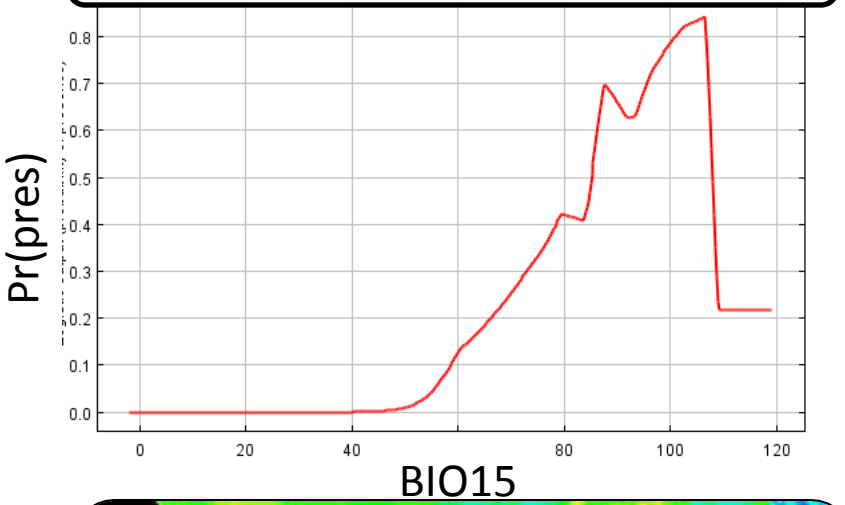
O. beecheyi vs. BIO15 (β=10)

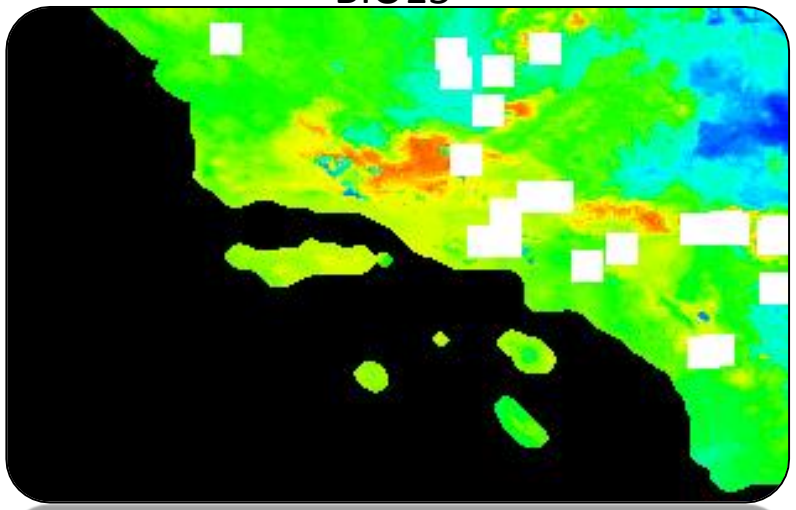




not smooth

O. beecheyi vs. BIO15 (β=0)





Smooth vs. not smooth

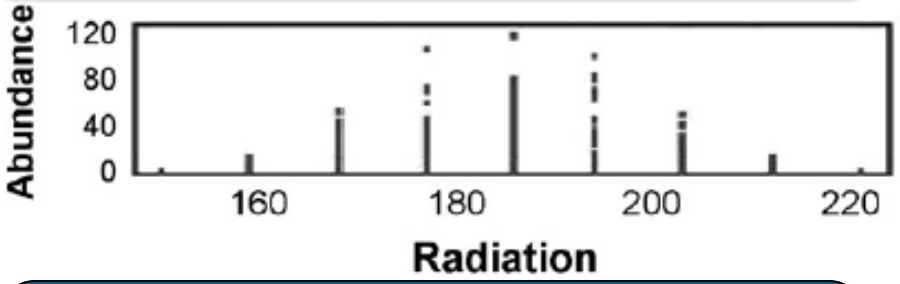
why smooth?

- **avoid overfitting** (gives any one point or set of points less weight)
  Elith et al. 2010. The art of modeling range-shifting species. Methods in Ecology and Evolution 1:330-342.
- **assumes species responds directly** to the predictors (vs. to correlated factors)
- may **"smooth" over local adaptation**

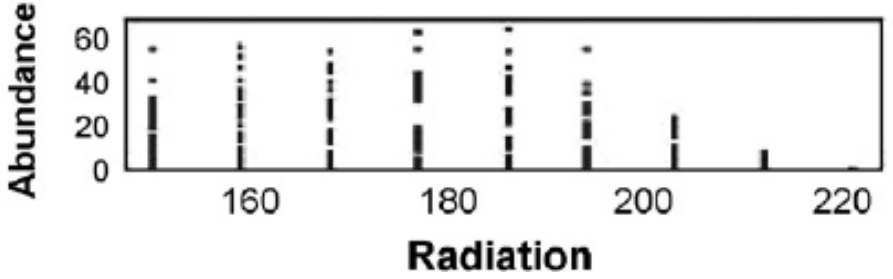
why not smooth?

- response may really be **highly non-linear**
- generally **assumes predictors are only correlated with important factors**
- can **allow for local adaptation**

"true" response of species to light



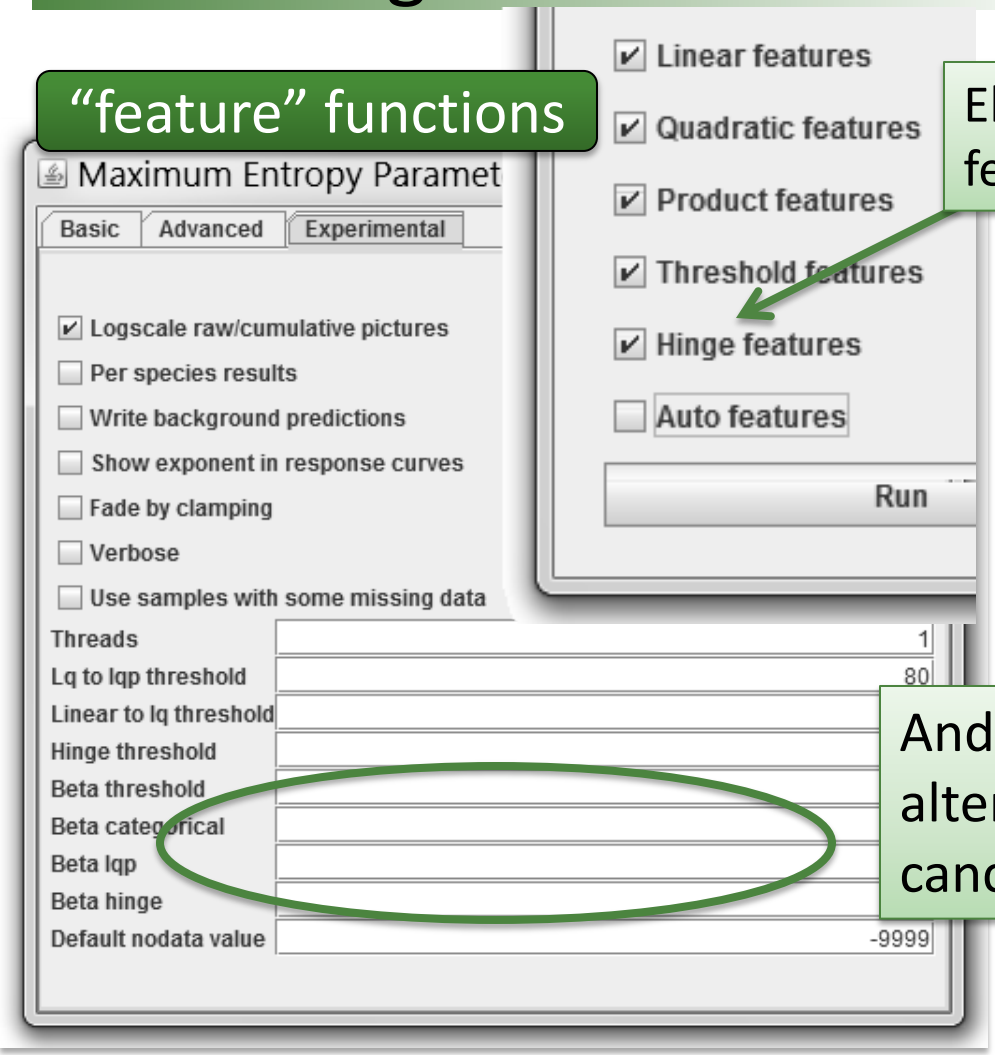
modeled response of species to light when model was trained using factors only correlated with light



Austin et al. 2006. Evaluation of statistical models used for predicting plant species distributions: Role of artificial data and theory. Ecological Modelling 199:197-216.

Best practices: Model parameterization

Smoothing in Maxent



Elith et al. 2010 used only hinge features to model cane toad invasion

Elith et al. 2010 The art of modeling range-shifting species. Methods in Ecology and Evolution 1:330-342.

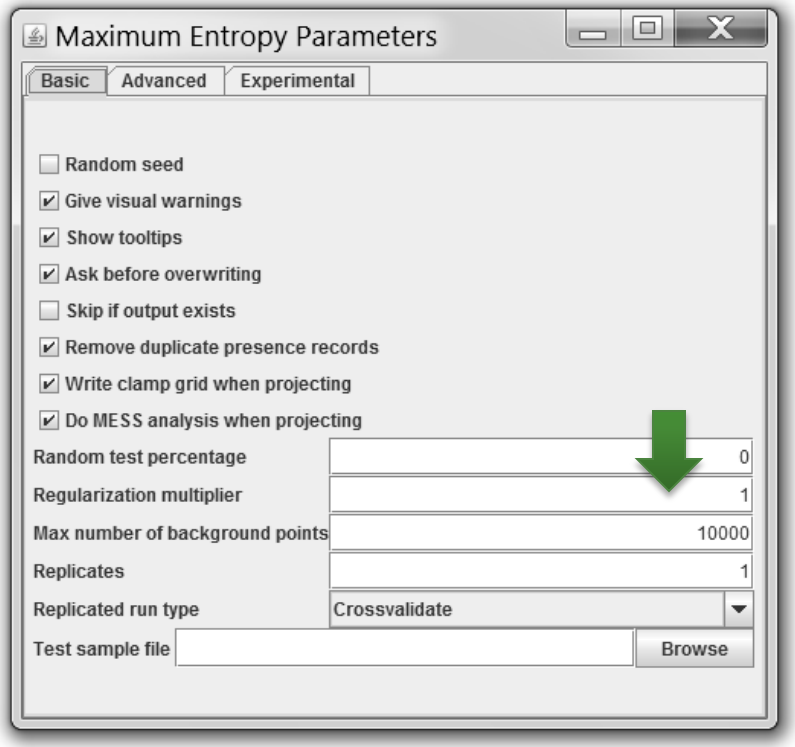
Anderson & Gonzalez (2011) alter "feature" function β's to cancel sampling bias.

Anderson & Gonzalez. 2011. Species-specific tuning increases robustness to sampling bias in models of species distributions… Ecological Modelling 222:2796-2811.

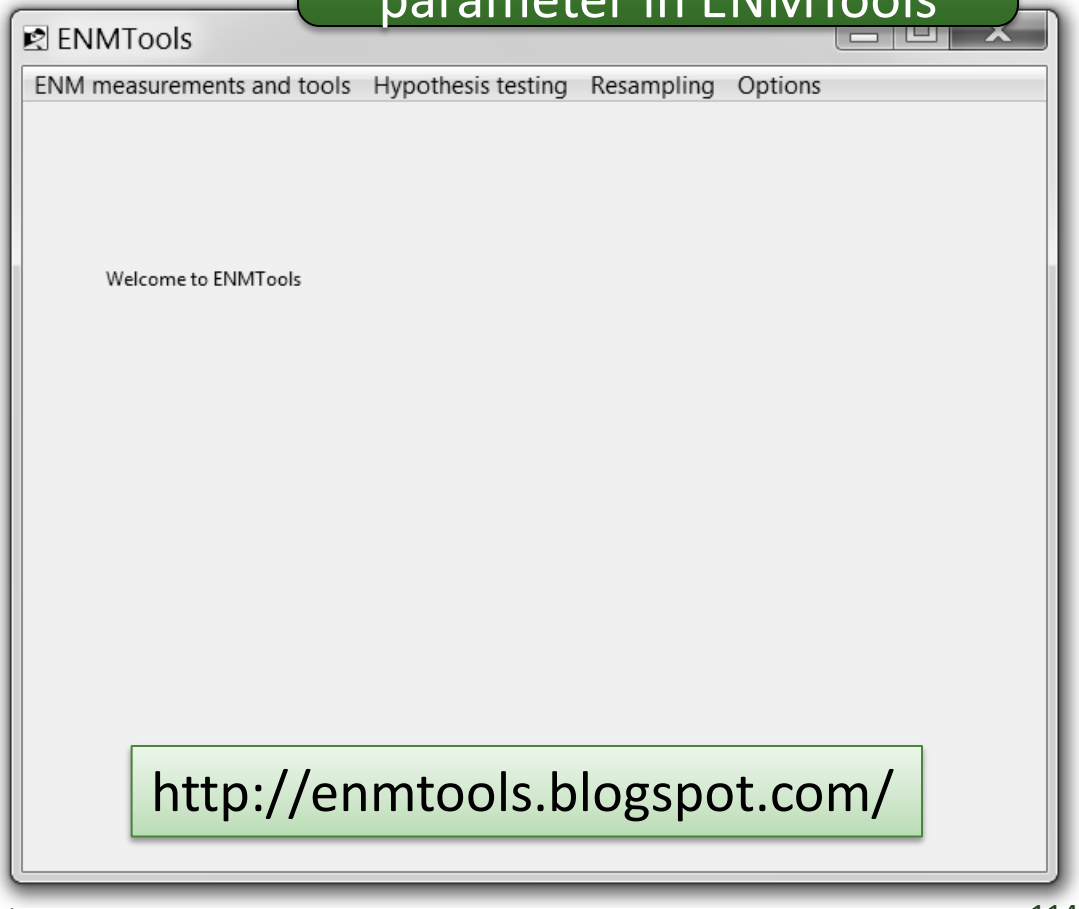
Best practices: Model parameterization

Smoothing in Maxent using AIC

β multiplier setting in Maxent



optimizing β regularization parameter in ENMTools



http://enmtools.blogspot.com/

Warren & Siefert. 2011. Ecological niche modeling in Maxent: The importance of model complexity and the performance of model selection criteria. Eco Apps 21:335-342.

Best practices: Model parameterization

Smoothing in other SDMs

Generalized Linear Models (GLMs)

- Use lower-order terms (linear plus quadratic or just linear)

Generalized Additive Models (GAMs)

- Don't include interaction terms
- Use higher penalty for degrees of freedom (gamma in "mgcv" package… suggest starting with gamma=1.4) Wood 2006. Generalized Additive Models: An Introduction with R. Chapman and Hall

Boosted Regression Trees (BRTs)

- Use smaller number of trees
- Reduce tree complexity

Elith et al. 2010 The art of modeling range-shifting species. Methods in Ecology and Evolution 1:330-342.

Support Vector Machines (SVMs)

- Don't include interactions (one- or two-class SVMs)
- Use linear kernels (one- or two-class SVMs)

Increase ν (one-class SVMs) Tax & Müller. 2004. A consistency-based model selection for one-class classification. Proceedings 17th International Conference on Pattern Recognition (22–26 August 2004, Cambridge UK) (eds J. Kittler, M. Petrou & M. Nixon), pp. 363-366. IEEE Computer Society, Los Alamitos, CA.

Best practices: Model parameterization

Partitioning training/testing data

Common practice is to **split data into k "folds"** (usually 20-30% of data)

If mutually exclusive, then each set is used as test data once and as with other folds as training data k-1 times (typical to use k=5).

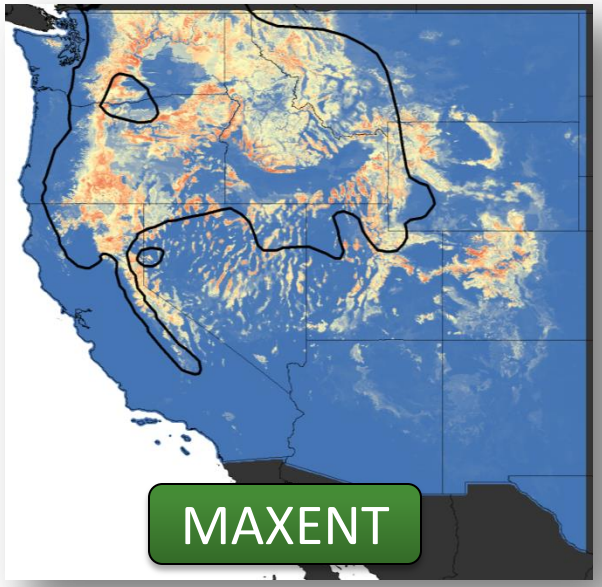
May also be randomly selected each time. Typical to use k=10 to 20…may also randomly choose pseudoabsences/background each time… usually need ~20 to get stable performance Barbet-Massin et al. 2012. Selecting pseudo-absences… Methods in Ecology and Evolution 3:327-338.

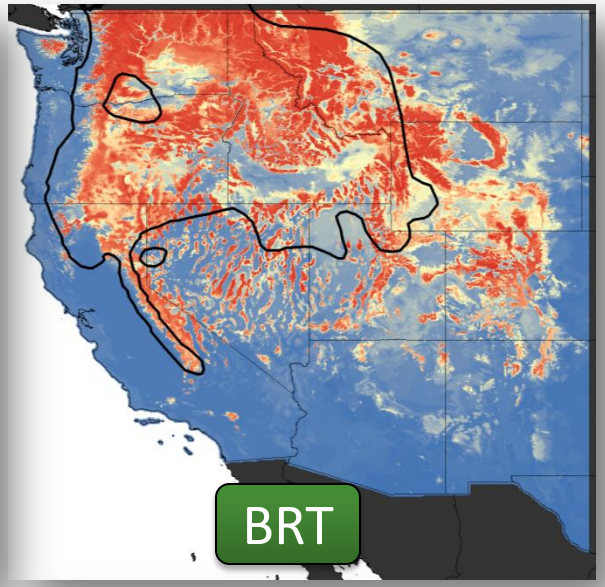
SPECIES	LONG	LAT	KFOLD
Dasypus novemcinctus	-96.57	29.53	2
Dasypus novemcinctus	-106.87	38.81	5
Dasypus novemcinctus	-96.57	29.53	2
Dasypus novemcinctus	-106.88	38.8	3
Dasypus novemcinctus	-100.91	38.92	5
Dasypus novemcinctus	-119.99	38.48	4
Dasypus novemcinctus	-96.57	29.53	1
Dasypus novemcinctus	-111.6	39.11	3
Dasypus novemcinctus	-106.87	38.81	4
Dasypus novemcinctus	-119.99	38.48	3
Dasypus novemcinctus	-111.4	42.22	1
Dasypus novemcinctus	-112.4	39.91	5
Dasypus novemcinctus	-96.57	29.53	2
Dasypus novemcinctus	-111.6	39.11	4
Dasypus novemcinctus	-96.57	29.53	1
Dasypus novemcinctus	-111.6	39.11	1
Dasypus novemcinctus	-112.08	38.28	5
Dasypus novemcinctus	-107.01	40.40	5
Dasypus novemcinctus	-106.87	38.81	2
Dasypus novemcinctus	-111.6	39.11	4
Dasypus novemcinctus	-106.87	38.81	3

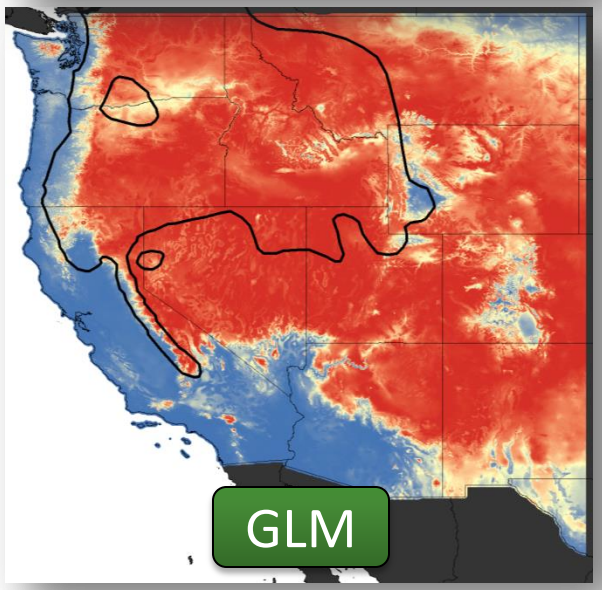
Data split between 5 k folds

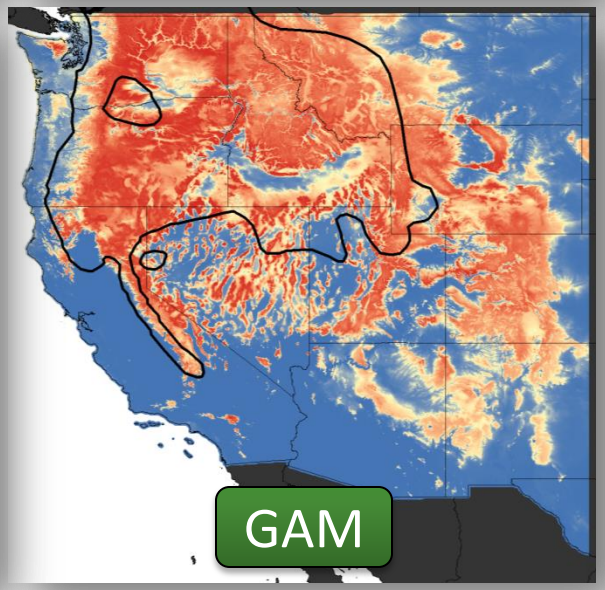
Best practices: Model parameterization

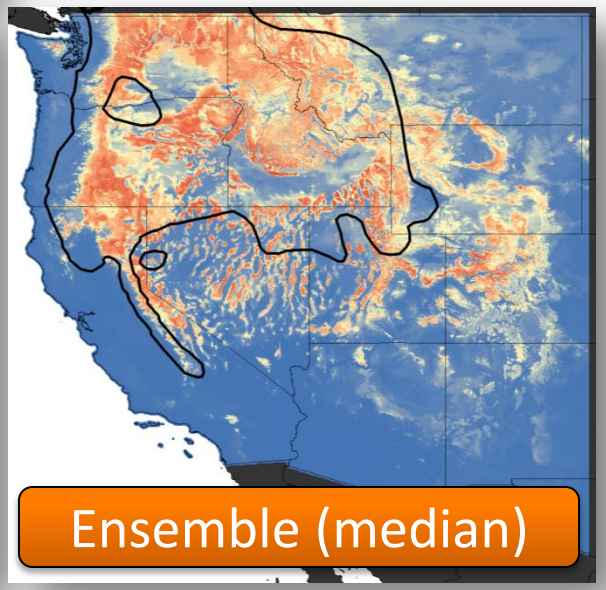
Ensembling









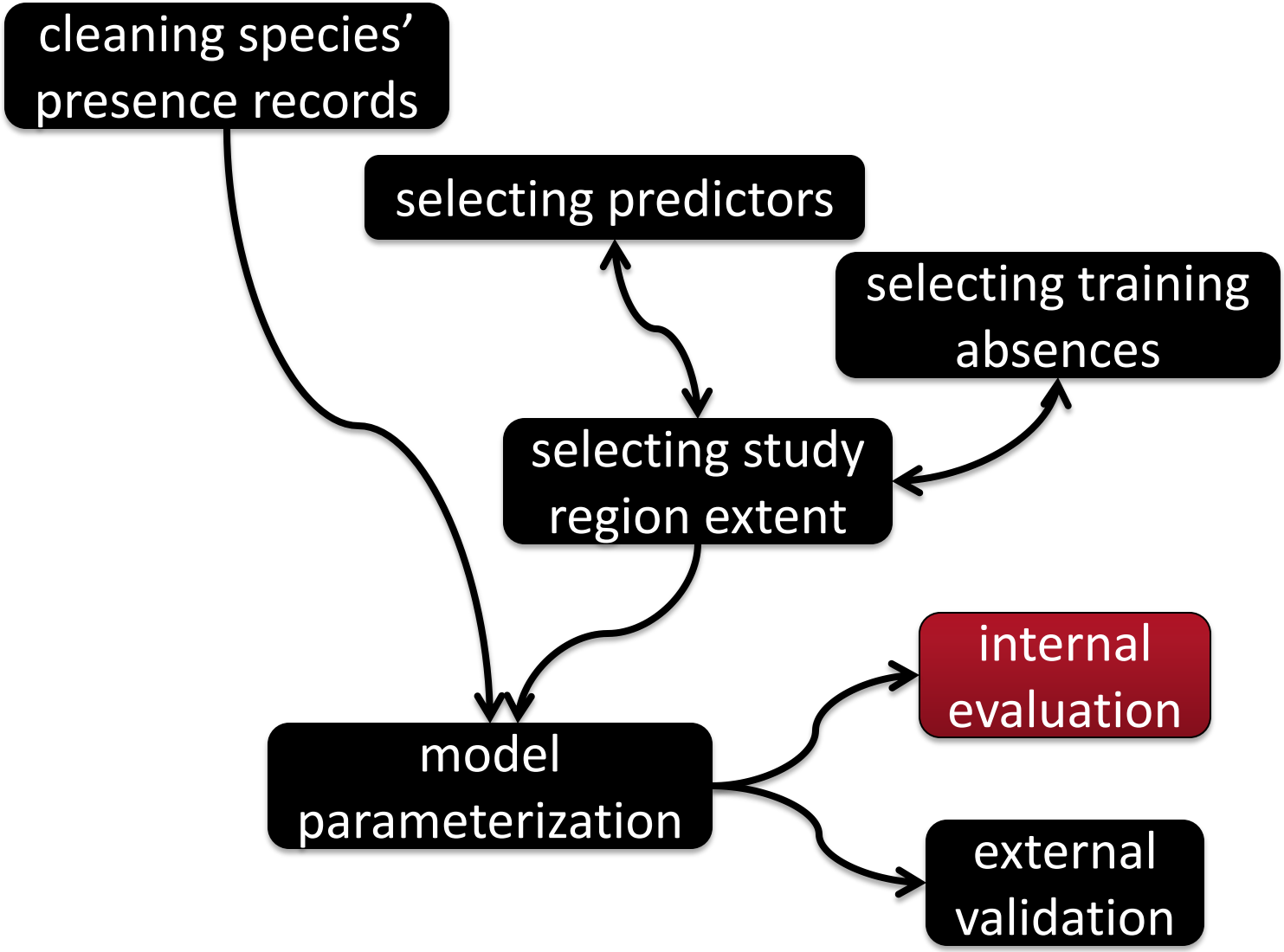


Araújo & New. 2007. Ensemble forecasting of species distributions. Trends in Ecology and Evolution 22:42-47.

Thuiller et al. 2009. BIOMOD – a platform for ensemble forecasting of species distributions. Ecography 32:369-373.

A short course on distribution modeling

"Best practices" ~ SDM workflow



Outline II

Best practices: Training absences/ pseudoabsences/background

"True" absences

"Pseudoabsences"

"Background"

"Targeted"

Best practices: Study extent vs. range size

Delineation, range size : study extent

Best practices: Model parameterization

Maxent: "Feature" functions, β regularization

Other SDM algorithms

Presence/absence vs. presence-only?

Weighting absences

Ensembling

Best practices: Model evaluation

Maps

Compare to spatial-only model

Autocorrelation in residuals

Response functions

Absences

Performance metrics

Thresholding

Stability/bootstrapping

Techniques for rare species

Using known but extirpated presences or failed reintroductions

Coarse → fine scale

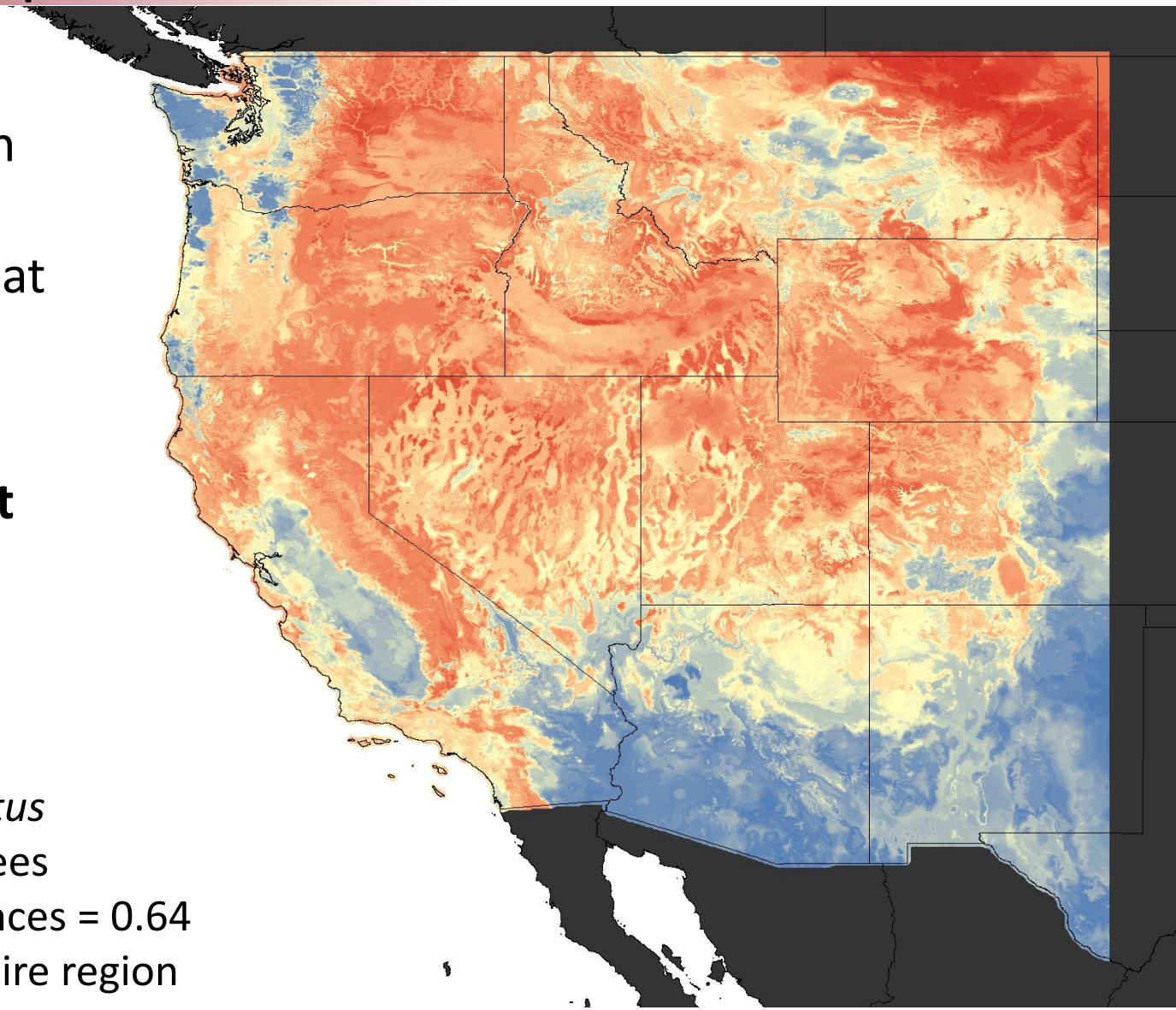
Uni/bivariate ensembles

Use model to search for more populations

Look at the maps

AUC, sensitivity, specificity, etc. can only tell you how good the model is at the test sites!

Relying solely on them **assumes test sites are sampled representatively**.

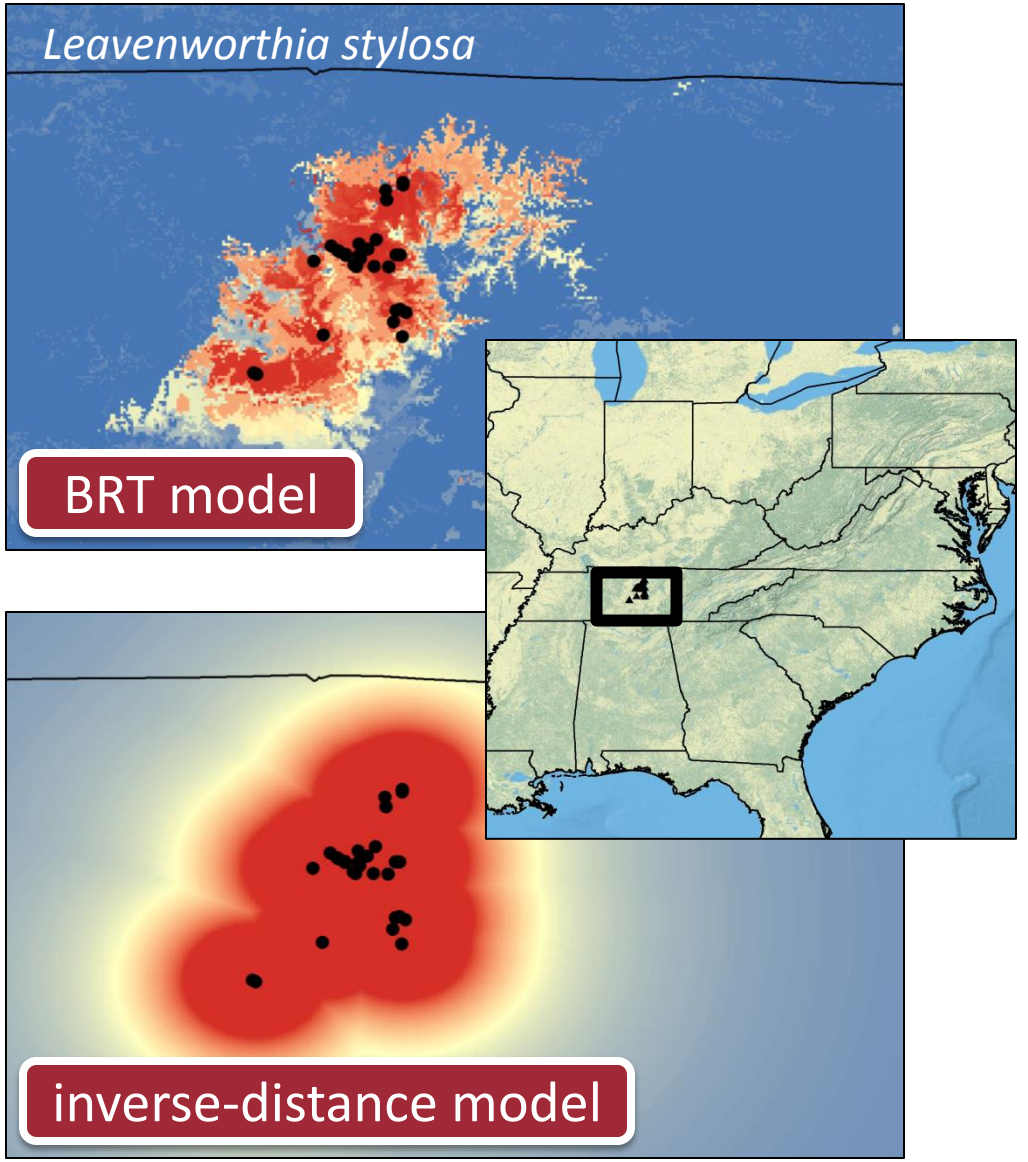


Peromyscus maniculatus
boosted regression trees
AUC vs. random absences = 0.64
IUCN range covers entire region

Compare to model with “space” only as predictors

Does the model trained with environmental predictors **out-perform a model with only space** as a predictor (e.g., lat/long or inverse distance to nearest known presence)?

Test metrics (AUC, sensitivity, specificity) are often **higher** when test points are **closer to training points**.



Segurado et al. 2006. Consequences of spatial autocorrelation for niche-based models. J Applied Ecology 43:433-444.

Veloz 2009. Spatially autocorrelated sampling falsely inflates measures of accuracy for presence-only niche models. J Biogeography 36:2290-2299.

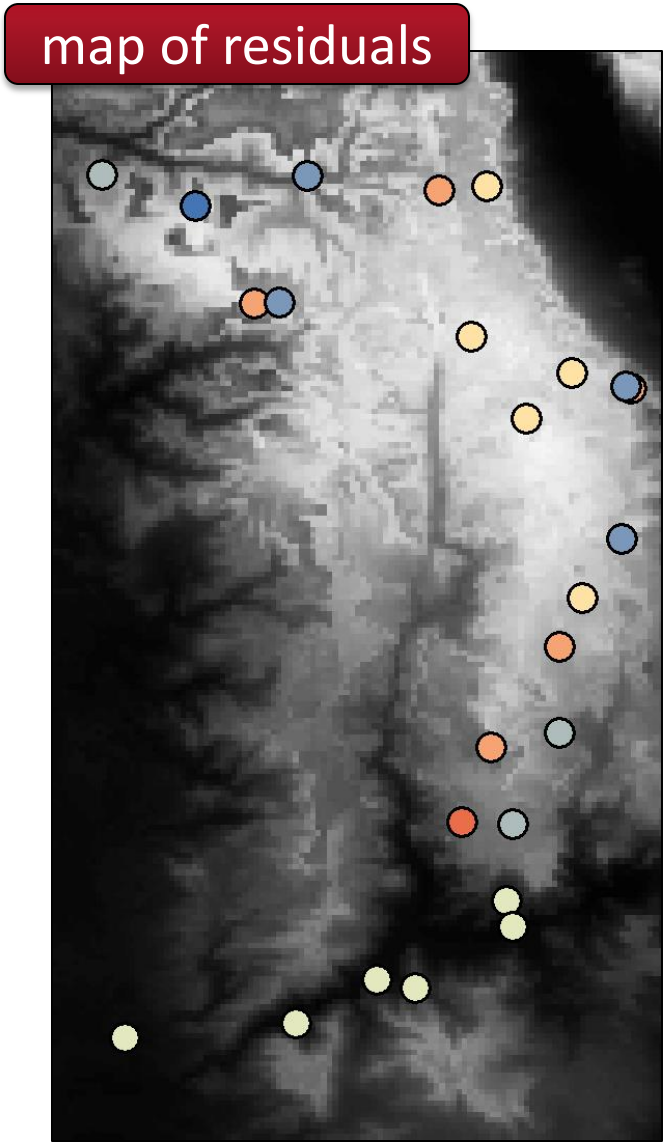
Autocorrelation in residuals

Indicates:

- **Misspecified model** (e.g., lacking a quadratic term)
- **Missing predictors**
- Predictor **resolution too coarse**
- "**Ecology**" (e.g., dispersal limitation, historical effects)

Only problematic if goal is to project range to new region/time (mean modeled response won't change).

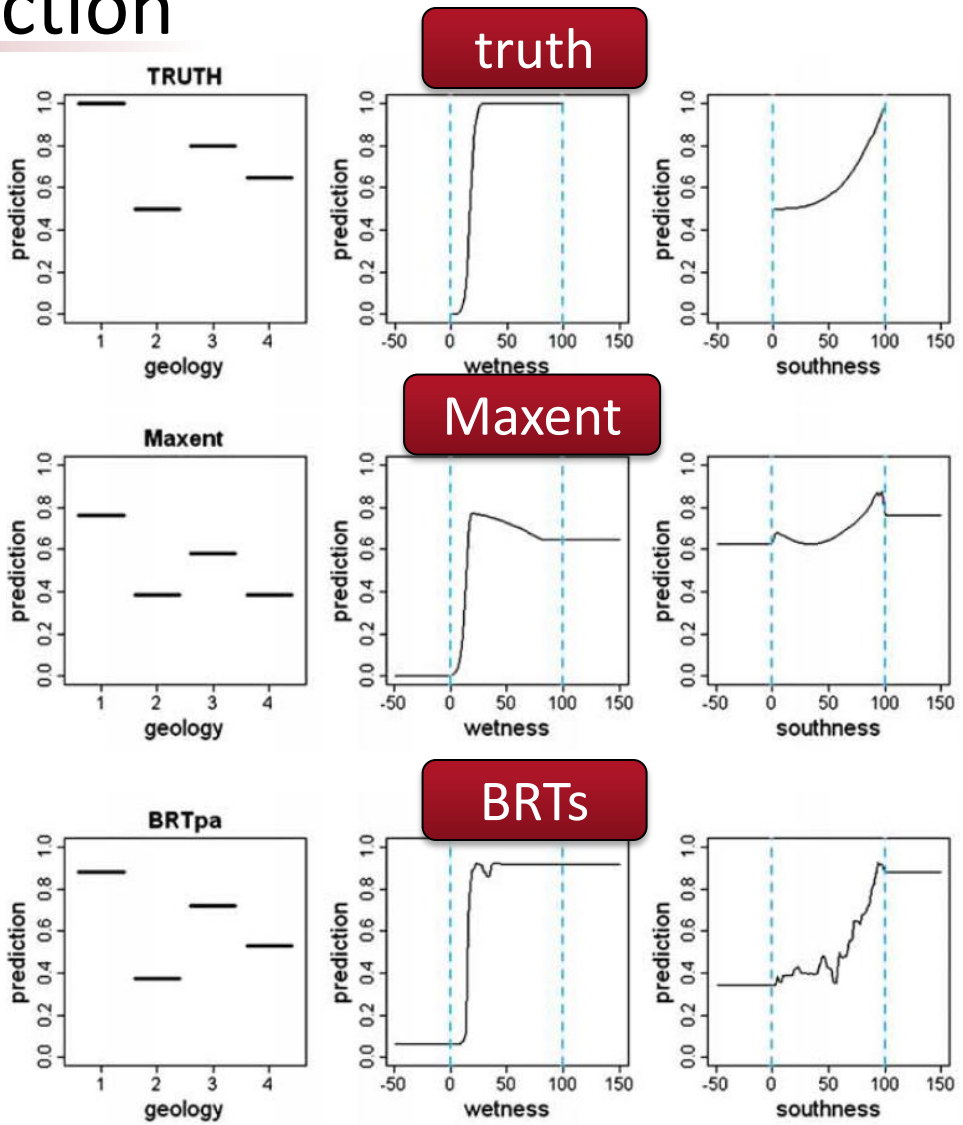
Tends to be more pronounced for regression-based approaches, esp. GLMs.



Response function inspection

Similar to inspection of residuals in regression (i.e., requires user to decide what's acceptable).

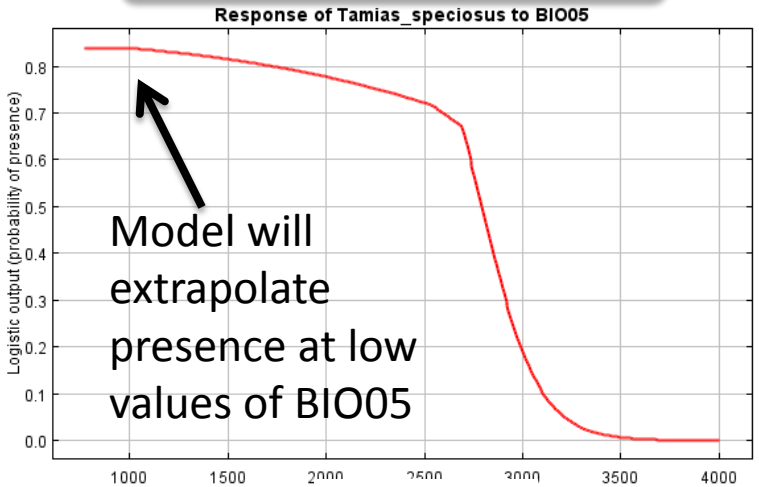
- Is response **reasonable** (e.g., increasing/decreasing when expected)?
- What do **multimodal** responses indicate?
  - **Records are clustered**, over-represent favorable environment
  - **Local adaptation**
  - **Ill-fitting** model



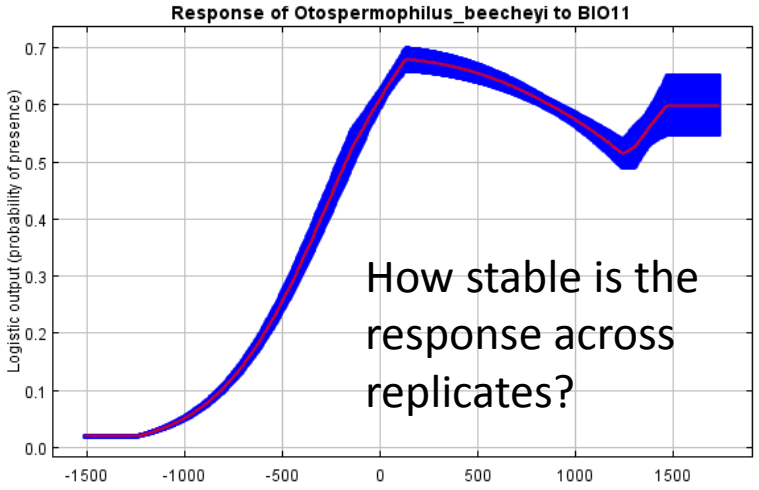
Elith & Graham. 2009. Do they? How do they? WHY do they? On finding reasons for differing performances of species distribution models. Ecography 32:66-77.

Response function inspection

univariate response

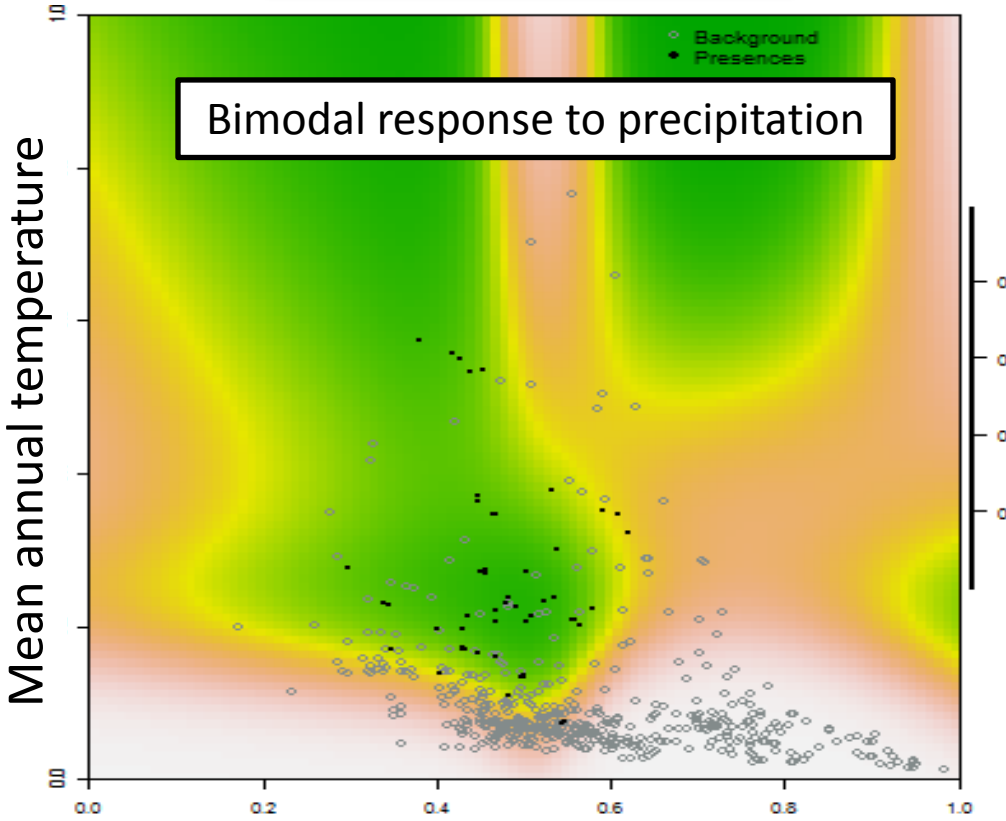


BIO05



BIO11

bivariate response



Mean annual precipitation

From Smith *In press* The relative influence of temperature, moisture, and their interaction on range limits of mammals over the past century. Global Ecology and Biogeography.

Test absences

"real" absences unavailable

Options:

1. Calculate **presence-only** performance **statistics** (e.g., sensitivity, omission rates)
2. Use **randomly located sites** as absences Phillips et al. 2006. Maximum entropy modeling of species geographic distributions. Eco Modeling 190:231-259.
   - **Reflects model's ability to differentiate randomly located sites from presence sites**.
   - Smaller **range size : study region ratio** ↑ apparent performance Lobo et al. 2007. AUC: A misleading measure... Global Ecology and Biogeography 17:145-151.
   - Past an *unknown* threshold, can lead to **biased** predictions Smith *In review* A tradeoff between apparent and actual accuracy of species distribution models evaluated with random background sites in place of absences.

"real" absences available

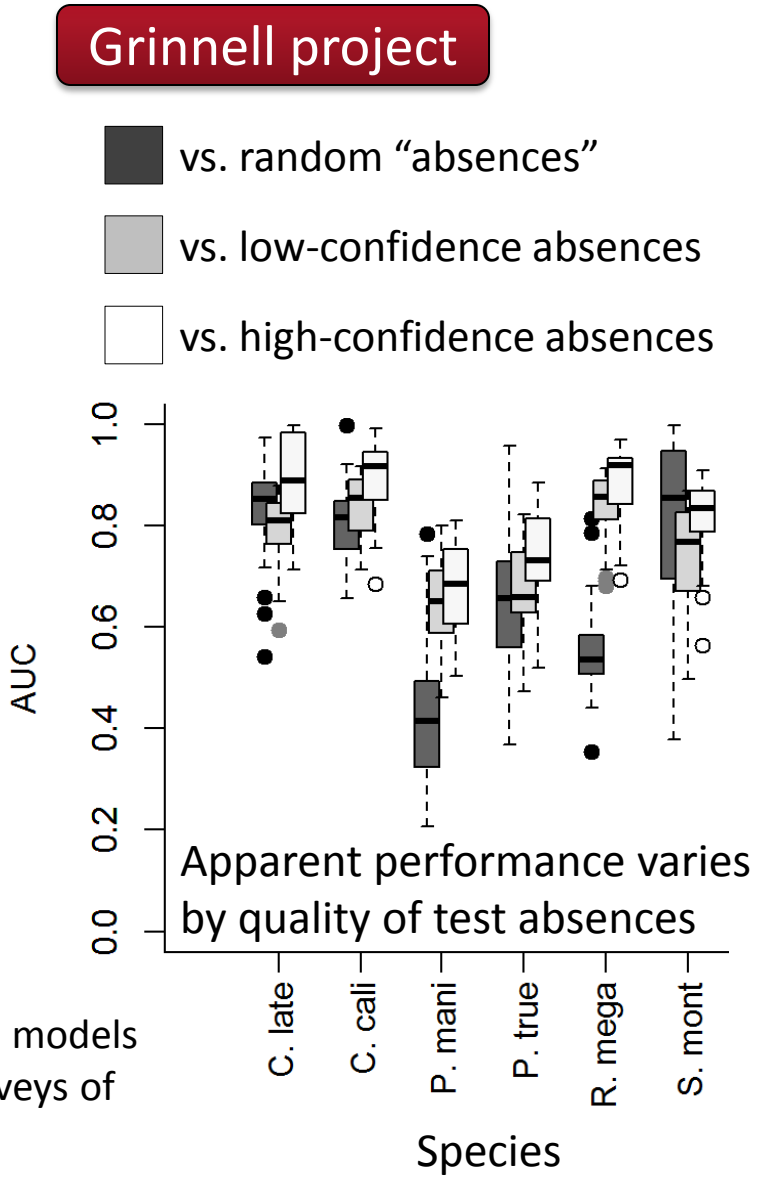
- Should be obvious or verified with occupancy modeling
- Should match scale of inference of "absence" (e.g., absence in small area doesn't connote absence in large area)
- Performance metrics **reflect model's ability to differentiate presences and absences**

Test absences

Performance metrics (AUC, sensitivity, etc.) will only be as good as the **confidence** in presence/absence at the test sites allows.

***Can* produce good models with presence-only data**, but can only know that with high-confidence presence/absence data.

Maybe worthwhile to use presence-only data and test with presence/absence data if latter is scarce.



Smith et al. *Submitted*. Validation of species distribution models projected across a century of climate change with resurveys of sites originally censused by Joseph Grinnell.

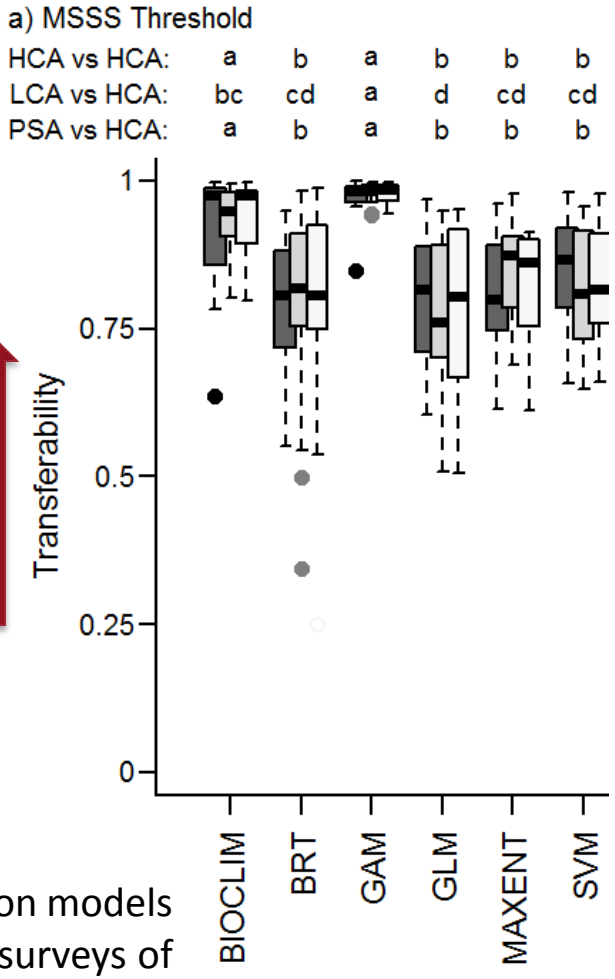
Thresholds

Threshold choice will affect "response" to climate change Nenzén et al. 2011. Choice of threshold alters projections of species range shifts under climate change. Ecological Modeling 222:3346-3354.

Small number of presences/absences will affect threshold selection and apparent accuracy Bean et al. 2012. The effects of small sample size and sample bias on threshold selection and accuracy assessment of species distribution models. Ecography 35:250-258.

"transferability" of thresholds between absence types and across time

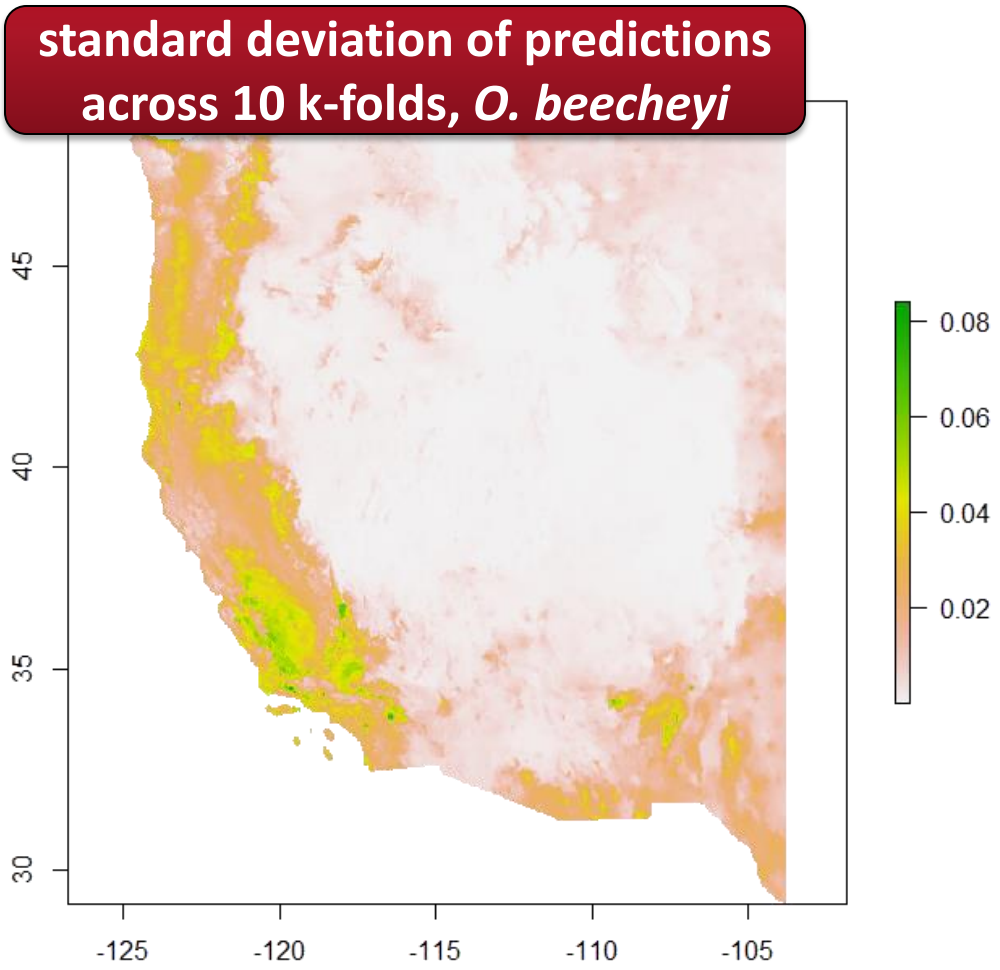
threshold chosen in one era more equal to same threshold chosen with data from another



Smith et al. *Submitted*. Validation of species distribution models projected across a century of climate change with resurveys of sites originally censused by Joseph Grinnell.

Sensitivity

Examine **sensitivity to data permutations** (training records, predictor values, test presences, study region size, etc.) in space and for individual predictors (see response curves)



External validation

Test data in **same time and region** from which training data is drawn is **not independent**! Araújo et al. 2005. Validation of species-climate impact models under climate change. Global Change Biology 11:1504-1513.

Non-independent test data will usually **inflate assessments** of model performance. Segurado et al. 2006. Consequences of spatial autocorrelation for niche-based models. Journal of Applied Ecology 43:433-444.

If test data in a completely different region or time period is unavailable, **try splitting arbitrarily** (e.g., between east and west or study region, or between records collected in different decades) and see if a model trained on one predicts the other. Iván Jimenez, *pers. comm*.

Innovations for modeling rare species

bivariate ensembles

- Train models with two predictors at a time if presence/absence data is available (otherwise use Maxent if just presences available)
- When finished, create ensembles
- Good when you have presence/absence data but not enough to use more than a few predictors at a time
- Lomba et al. 2010. Overcoming the rare species modeling complex: A novel hierarchical framework applied to an Iberian endemic plant. Biological Conservation 143:2647-2657.

coarse → fine-scale modeling

- Model coarse scale first, identify areas with high suitability
- Focusing on one or more of these areas, apply another round of modeling using fine-grained predictors
- Good for identifying core favorable areas
- Good when fine-grained predictors are not available over large region or their demand on computer memory/speed precludes processing all predictors simultaneously

Innovations for modeling rare species

iterative modeling with discovery

- Can use SDM to guide searches for unknown populations
- Must take care also to search favorable areas representatively or addition of new populations to subsequent models may create bias

Le Lay et al. 2010. Prospective sampling based on model ensembles improves the detection of rare species. Ecography 33:1015-1027.

Guisan et al. 2006. Using niche-based models to improve sampling of rare species. Conservation Biology 20:501-511.

Williams et al. 2009. Using species distribution models to predict new occurrences for rare plants. Diversity and Distributions 15:565-576.

leave-one out evaluation

1. Model using all known points but one
2. Threshold output and determine if withheld point was predicted present or absent
3. Calculate probability of correctly predicting that point given total area predicted (assumes no spatial autocorrelation between occurrences)
4. Repeat steps 1-4 for each point

- Good when not enough data exists to train and test model
- Must be wary of non-independence between points!

Pearson et al. 2007. Predicting species distributions from small numbers of occurrence records: A test case for using cryptic geckos in Madagascar. Journal of Biogeography 34:102-117.

SDMing… not just for species!

Distribution models like Maxent only find correlative relationships between predictors and the presence/absence of an "entity".

Usually, "entities" are species, but they can also be:

- crop types
- vegetation types/biomes
- landslides
- fire
- human occupation
- et cetera!

Checklist for SDMing

problem definition

- Define goals of project
- Is under/overprediction error more egregious?

species' records

- Names checked
- Location of records reasonable, verified with external checks
- Coordinate uncertainty ≤~10 km, less in topographically diverse areas
- Enough records for modeling given SDM being used (≥~30 ideal)
- Thin presences so ≤1 per raster cell

predictors

- Justifiable given ecological knowledge
- Resolution matches scale of relevant processes
- No high correlations between variables
- Reflect relevant geographic gradients in region of interest reflected in predictors
- Difference between dynamic/static/dynamic-but-static considered

Checklist

training absences

- Can justify choice of region from which “non-presences” are chosen
- Adequate coverage across study region
- “True” absences: Species is easily detected or absences verified with occupancy modeling
- “Pseudoabsences”: Suggest try several sizes of area from which to select
- Random “background” sites: As per pseudoabsences
- “Targeted” absences: Compare to models using random background/ pseudoabsences

study region

- Appropriate to scale of question
- If too large and overprediction bias is more desirable, training and test absences may have to be restricted to a subset of the region
- Encompasses expected biogeographic responses to climate change

model training and parameterization

- Smooth vs. not smooth… which reflects study goals?
- “Absences” weighted same as presences
- Compare results from different algorithms

Checklist

evaluation

- Are maps reasonable? Where over/underpredict?
- Compare to space-only model
- Is there autocorrelation in residuals?
- Inspect response functions
- What kind of test absences to use? (“real” vs. random sites)
- What effect does study region extent have on performance metrics?
- How does output change with choice of threshold?
- Examine model stability/sensitivity to data permutations
- Can model be externally verified (project to different time/region)?

re-evaluation

- If choices were made to encourage under/overprediction bias, did they work (how sensitive is model output to these choices)?

Advanced modeling

www.earthskysea.net