Element Distribution Modeling

Overview of Modeling Algorithms

EDM Workshop

December 6, 2004

POTENTIAL MISUSES OF ELEMENT DISTRIBUTION MODELS

Maps produced by EDM are spatial models - not direct representations, as most maps are assumed to be.

Thus there is some degree of uncertainty associated with areas predicted as suitable *and areas predicted as unsuitable* (the latter is often forgotten)

IN GENERAL, MISUSES OF EDM PRODUCTS OCCUR WHEN THE USER FAILS TO RECOGNIZE OR UNDERSTAND THIS UNCERTAINTY

2 points of responsibility:

MODELER must adequately communicate uncertainty

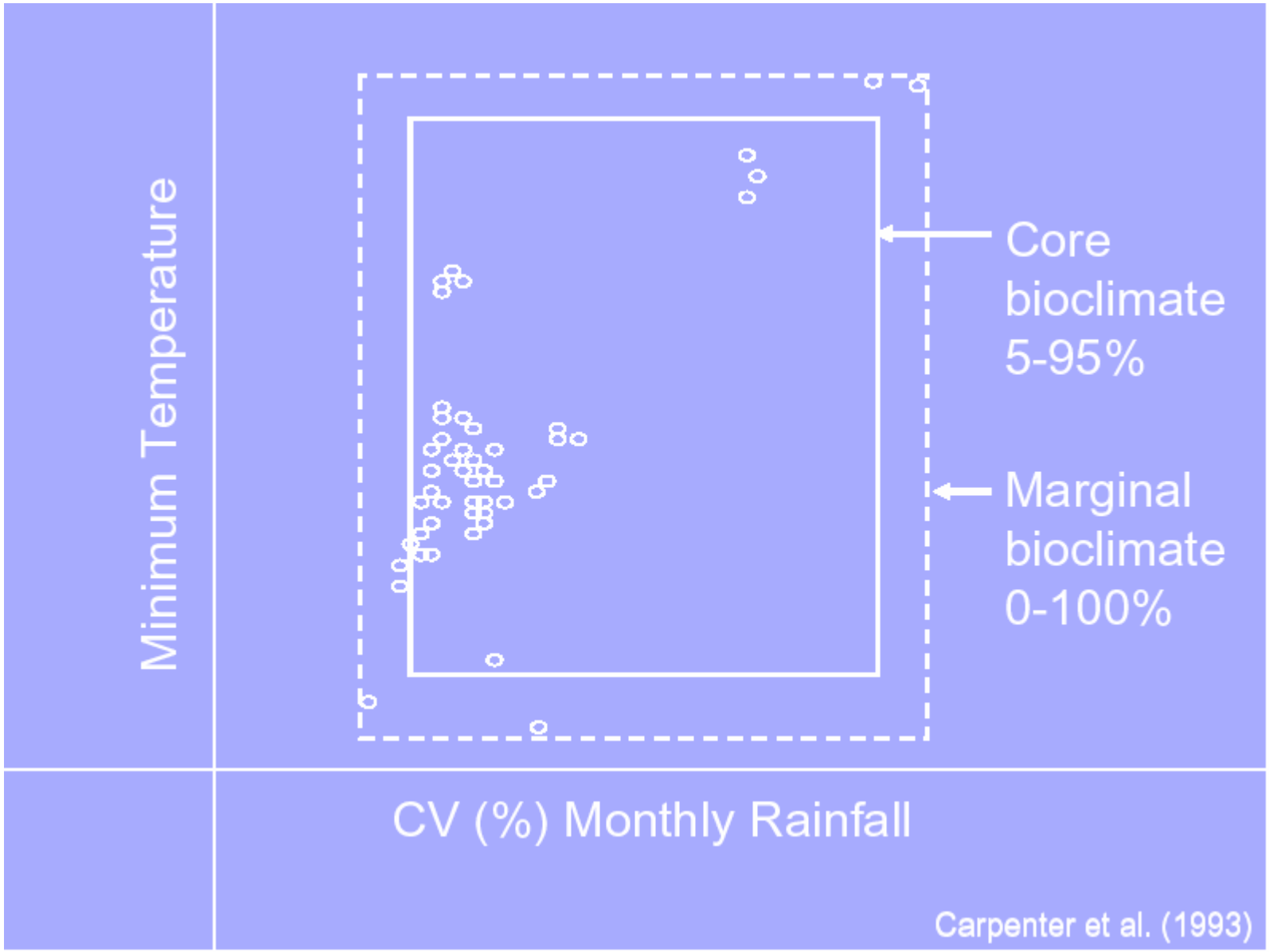
USER must acknowledge uncertainty, take responsibility for application

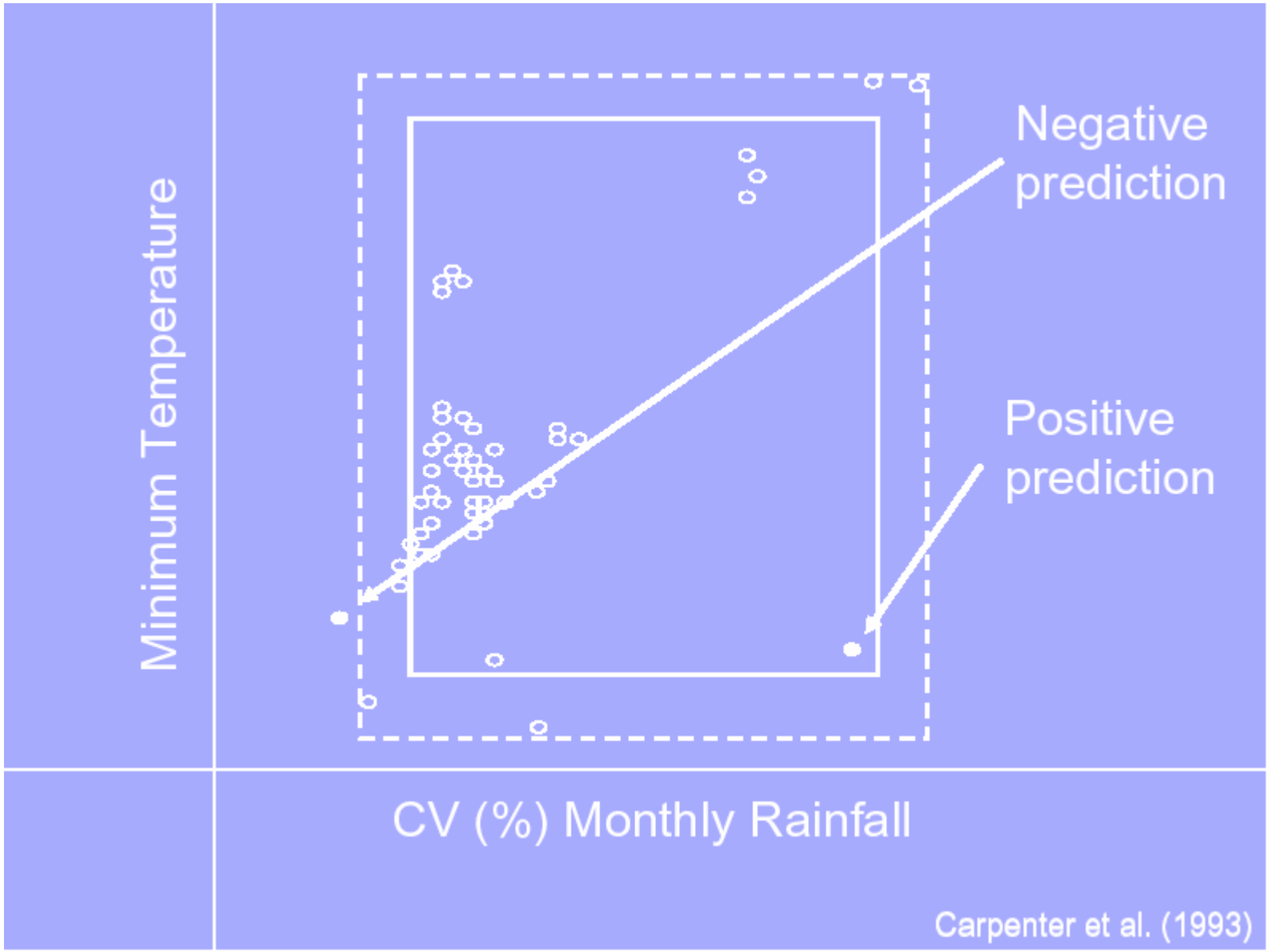
Environmental Envelops
BIOCLIM

- Multidimensional environmental box
- Defines environmental conditions for known occurrences
- Presence-only data

Advantage:

- Easy to explain and implement

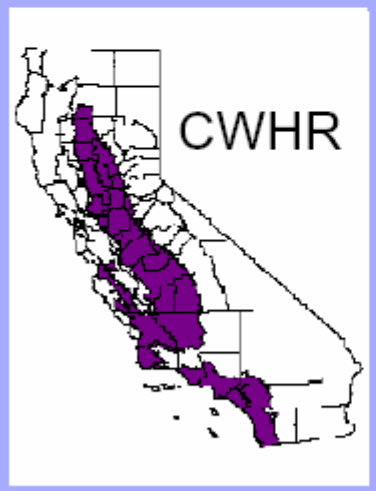

Minimum Temperature
Core
bioclimate
5-95%
Marginal
bioclimate
0-100%
CV (%) Monthly Rainfall
Carpenter et al. (1993)


Minimum Temperature
Negative prediction
Positive prediction
CV (%) Monthly Rainfall
Carpenter et al. (1993)

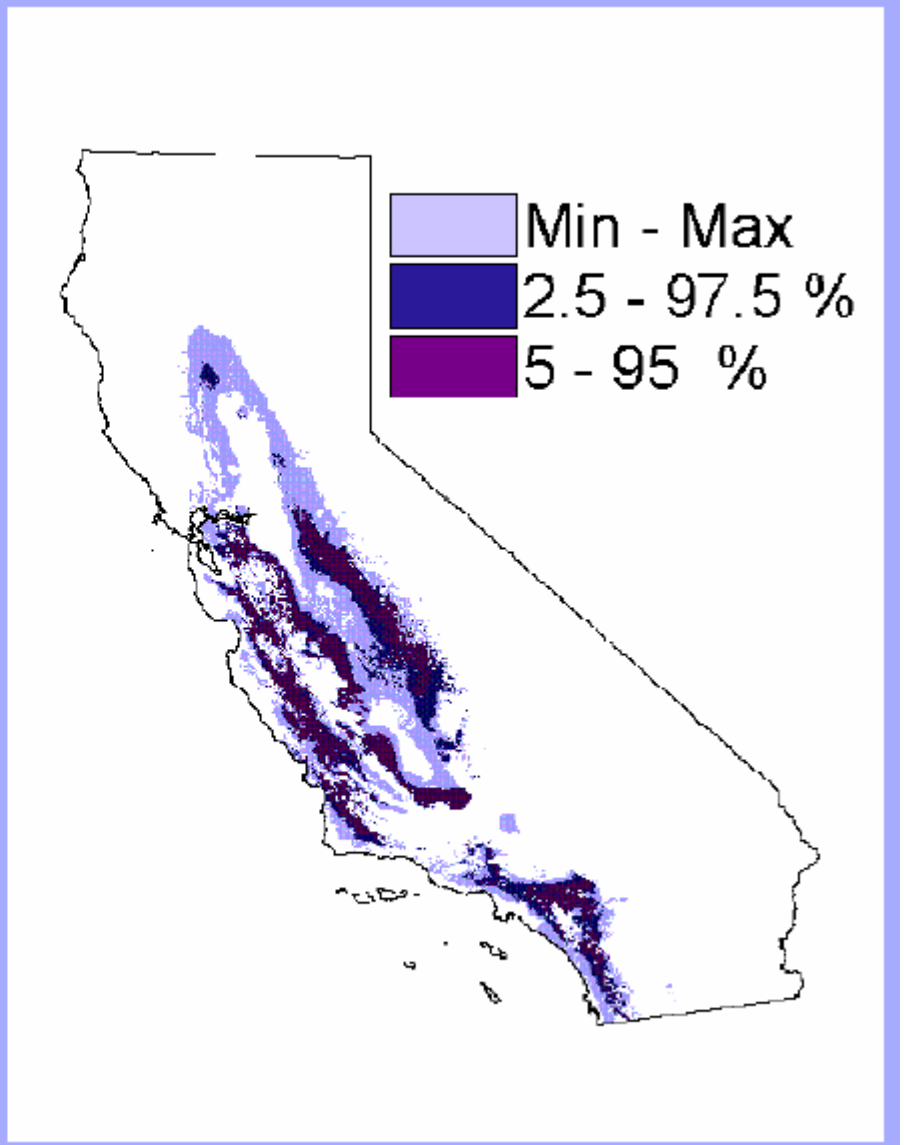
Environmental Envelops
BIOCLIM

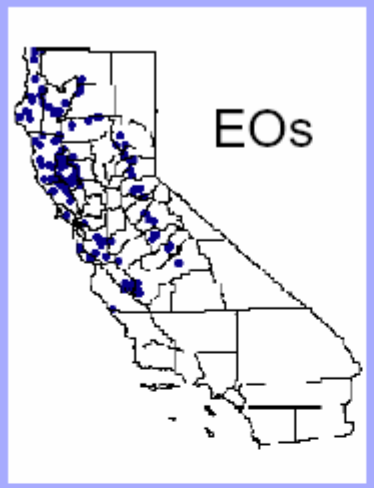
Disadvantages:

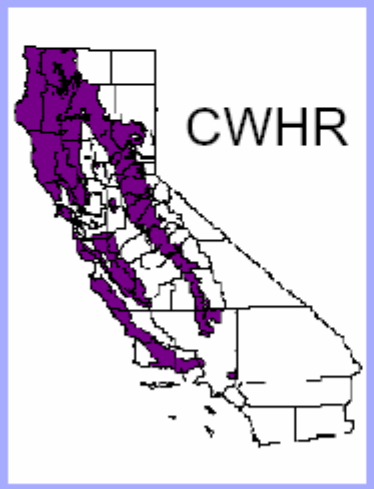
- Unable to consider correlation/interactions
- Gives equal weight to all predictors
- All conditions considered equally suitable
- Sensitive to outliers and sampling bias
- Cannot use categorical data
- No procedure for variable selection



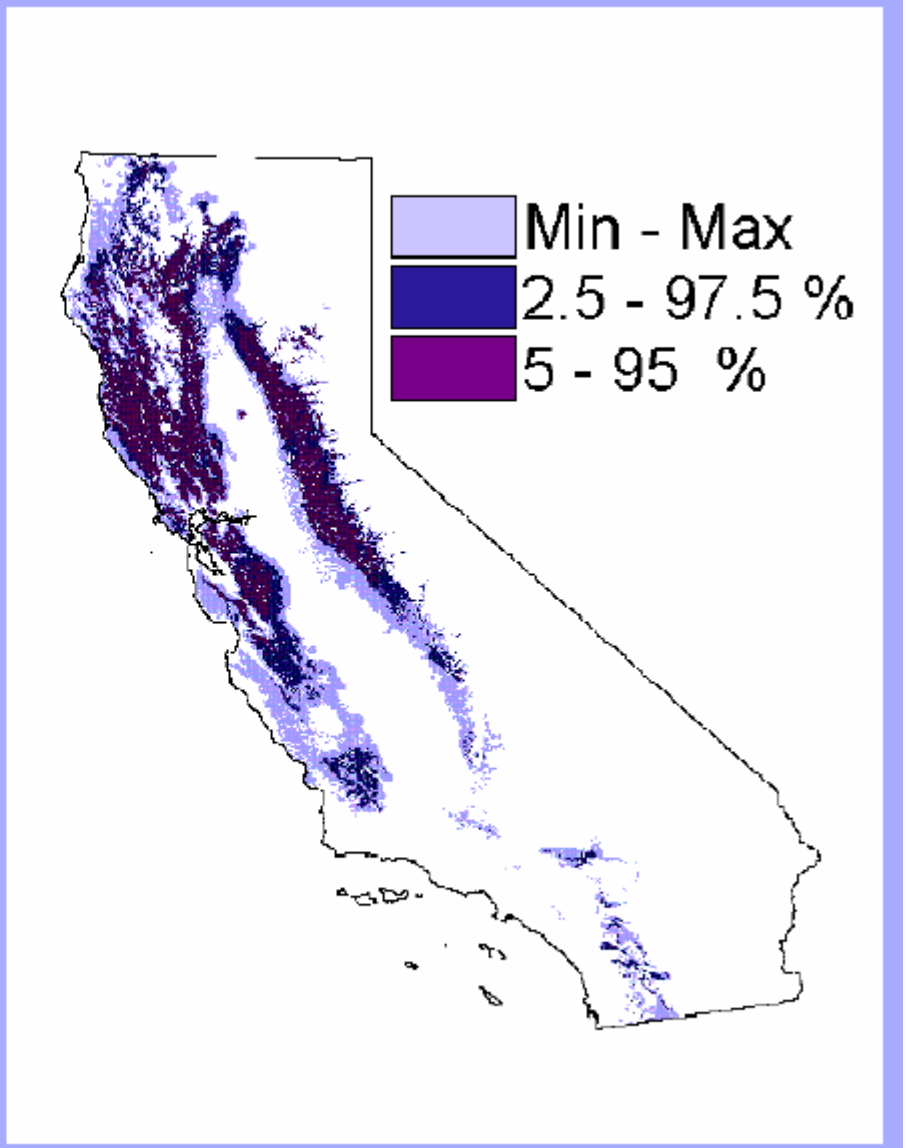
Western Spadefoot Toad
Scaphiopus hammondii

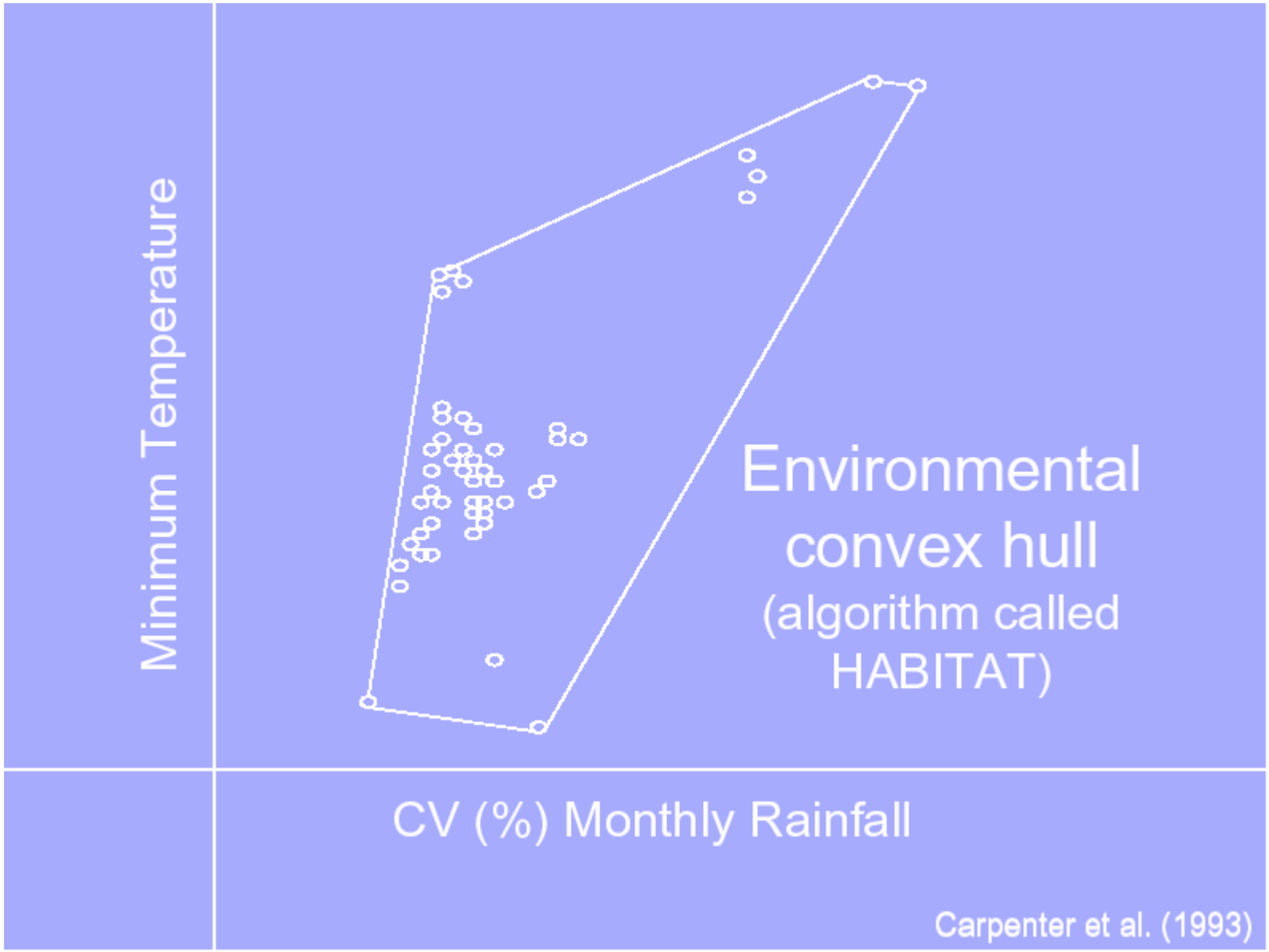


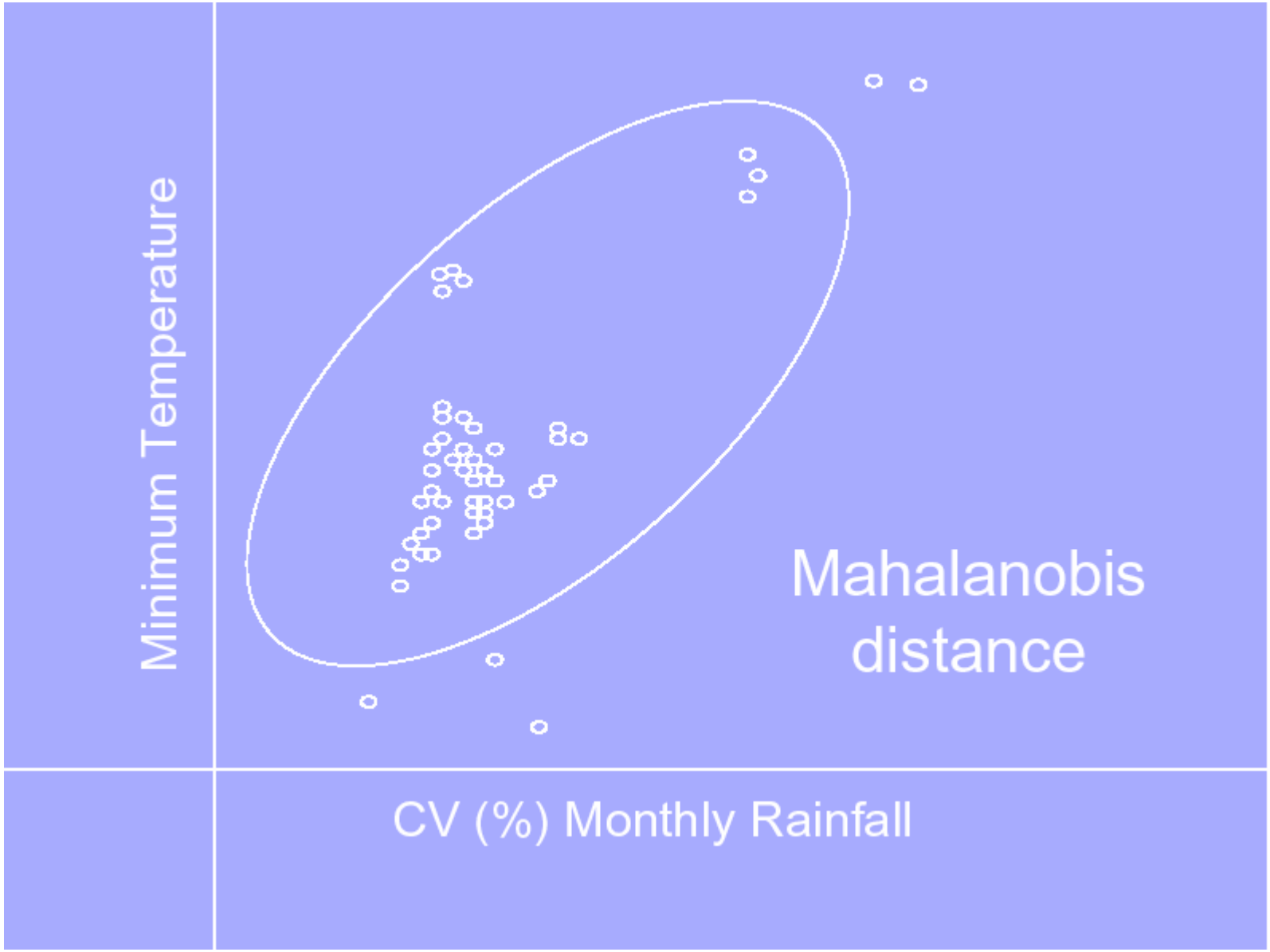




Foothill Yellow-legged Frog
Rana boylii




Minimum Temperature
Environmental convex hull
(algorithm called HABITAT)
CV (%) Monthly Rainfall
Carpenter et al. (1993)

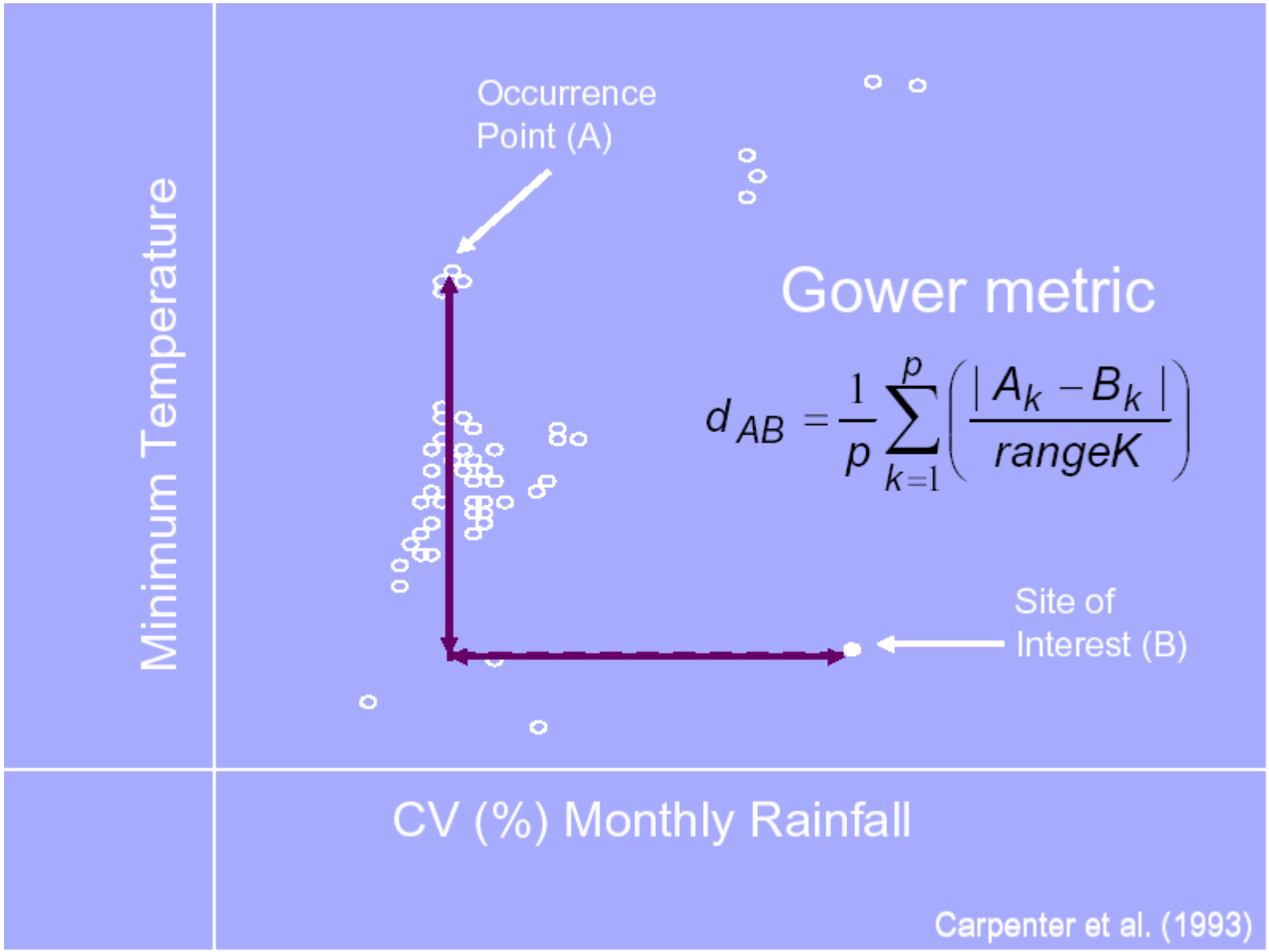

Minimum Temperature
Mahalanobis distance
CV (%) Monthly Rainfall

DOMAIN

- Point-to-point similarity metric (Gower metric)
- Presence-only data

Advantages:

- Easy to implement
- Performs well with small sample sizes


Minimum Temperature
Occurrence Point (A)
Gower metric
$d_{AB} = \frac{1}{p}\sum_{k=1}^{p}\left(\frac{|A_k - B_k|}{rangeK}\right)$
Site of Interest (B)
CV (%) Monthly Rainfall
Carpenter et al. (1993)

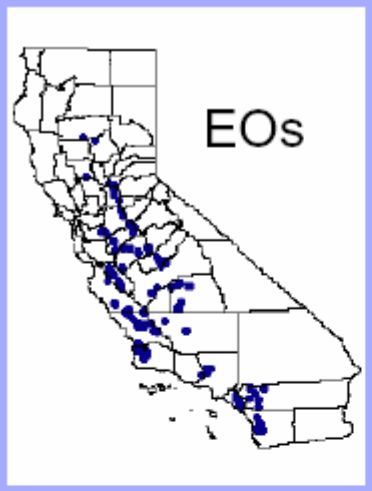
DOMAIN

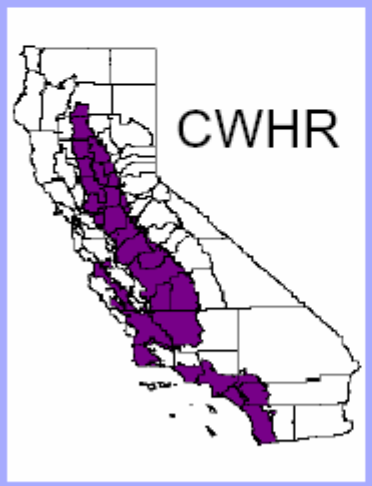
- Converted to a complementary similarity
- Maximum similarity (between 1 or average of # points)
- Predictions are continuous
- Measure of classification confidence

DOMAIN

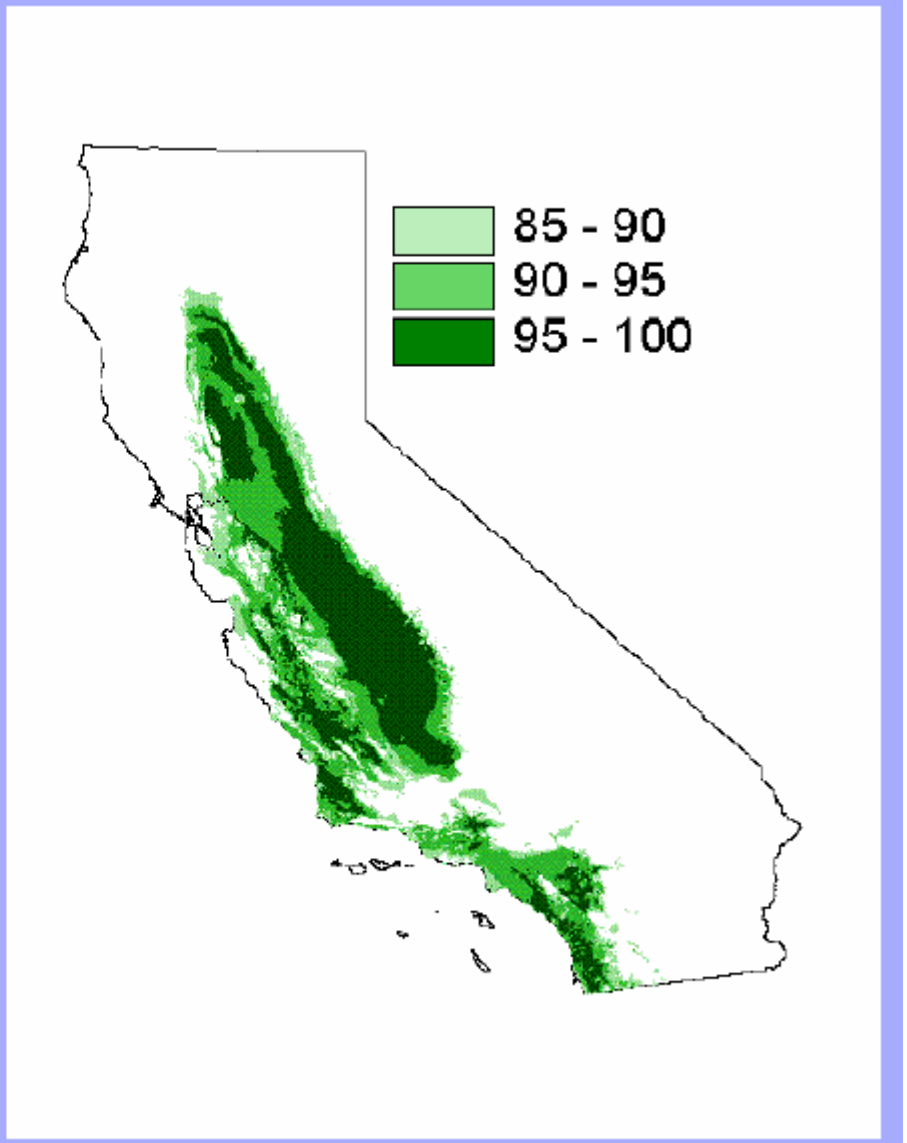
Disadvantages:

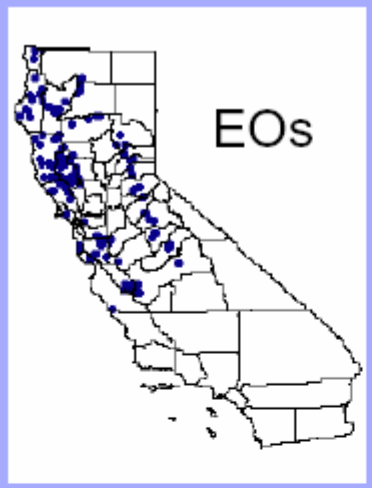
- Unable to consider correlation/interactions
- Gives equal weight to all predictors
- Cannot investigate predictor variable influence
- No procedure for variable selection

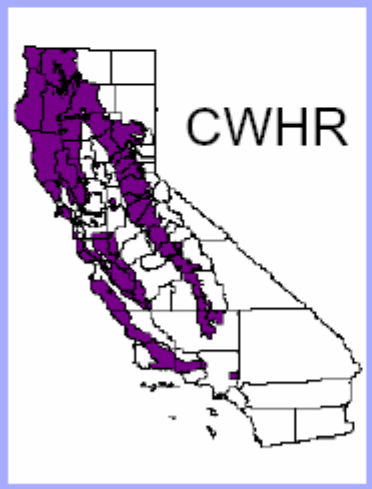




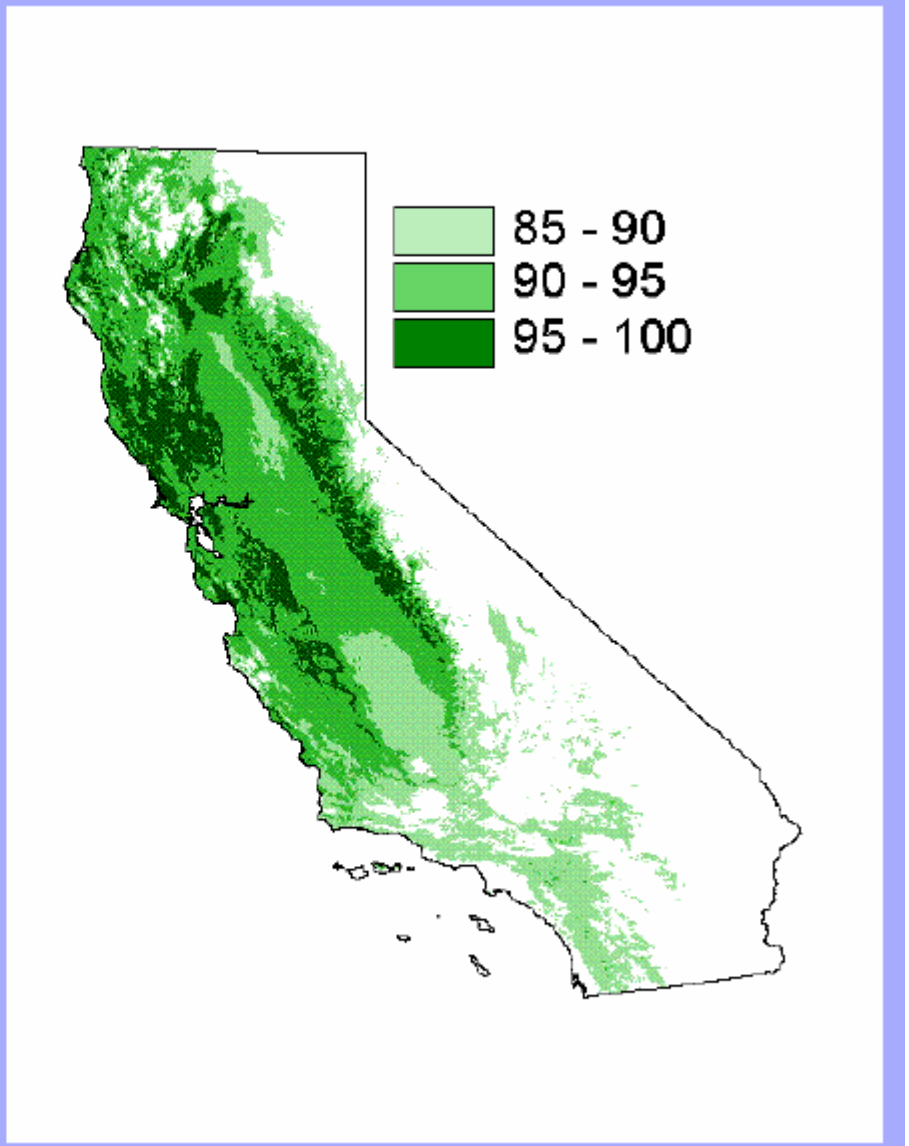
Western Spadefoot Toad
Scaphiopus hammondii







Foothill Yellow-legged Frog
Rana boylii



Logistic Regression

- Formulates a relationship between presence/absence values (response) and environmental predictors
- Generated in Generalized linear models (GLM) framework
- Logit link function and binomial error distribution

Logistic Regression

- Relationship represented as linear function

$$\text{logit}(p) = \log\frac{p}{1-p} = 10.75 + [-0.007 * \text{Elevation}] + [-0.015 * \text{CV of precipitation}] + \ldots$$

- The inverse logistic transformation = probability of occurrence

$$p(x_1, x_2, \ldots, x_k) = \frac{1}{1 + e^{-\text{logit}(p)}}$$

- Predictions are continuous (0-1)

Logistic Regression

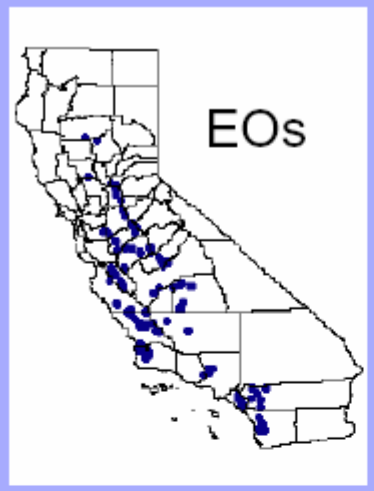
Advantages:

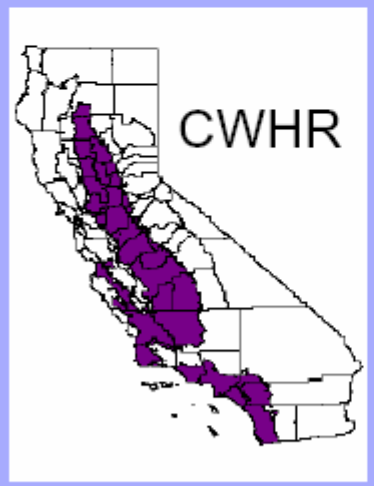
- Easy to implement and interpret
- Can consider interactions and non-linear relationships (polynomials)
- Categorical data accepted
- Variable reduction procedures available
- Possible to investigate variable importance
- Well studied, many refinements available
- Residual analysis for uncertainty investigation

Logistic Regression

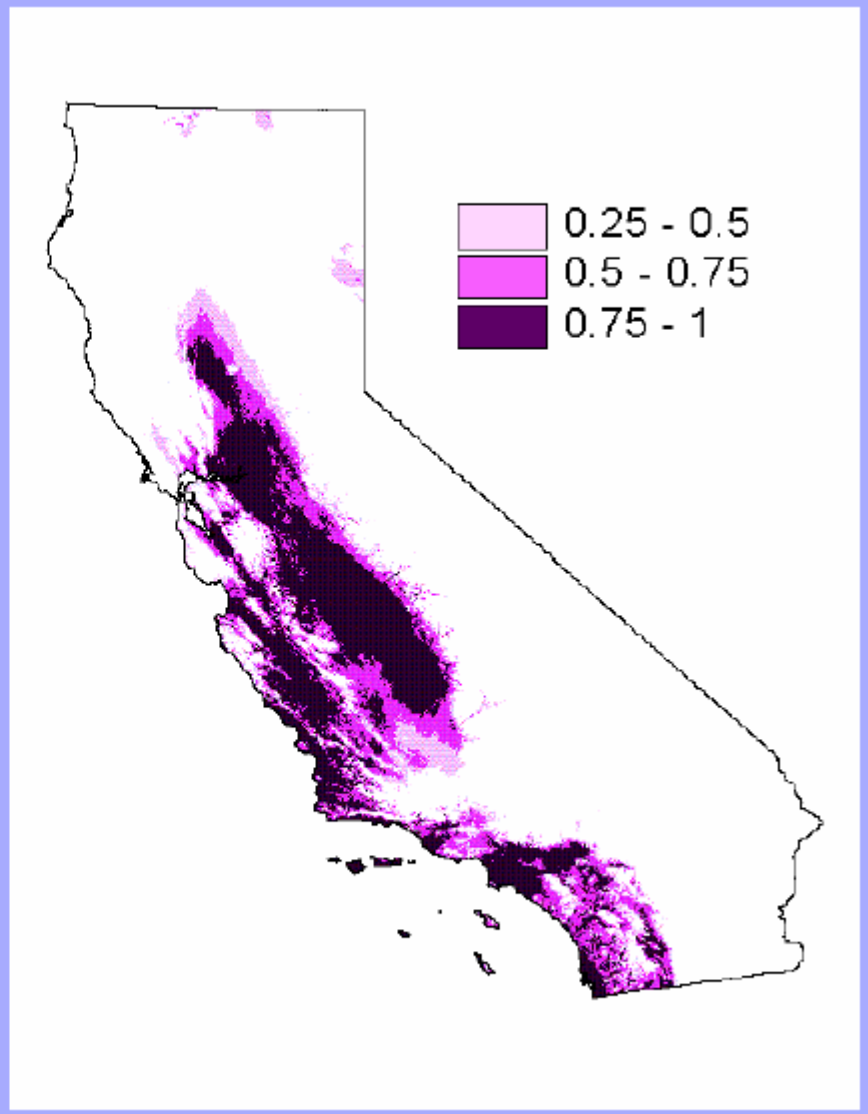
Disadvantages:

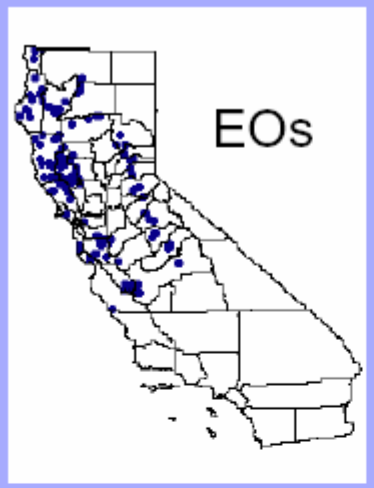
- Sometimes difficult to determine appropriate non-linear relationship
- Variable reduction procedures not perfect
- Extremely sensitive to ratio of presence/absence occurrences
- Requires multiple software

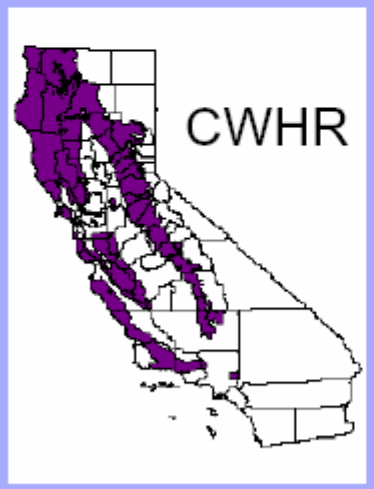




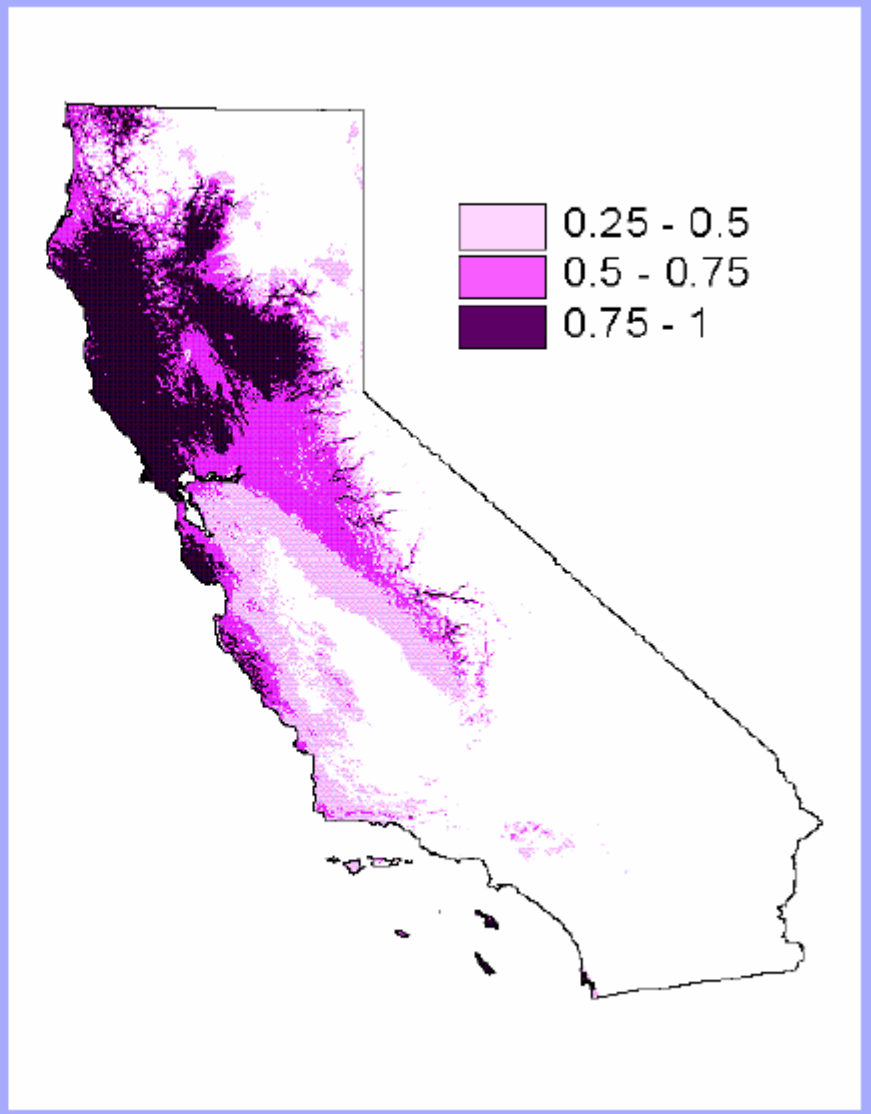
Western Spadefoot Toad
Scaphiopus hammondii





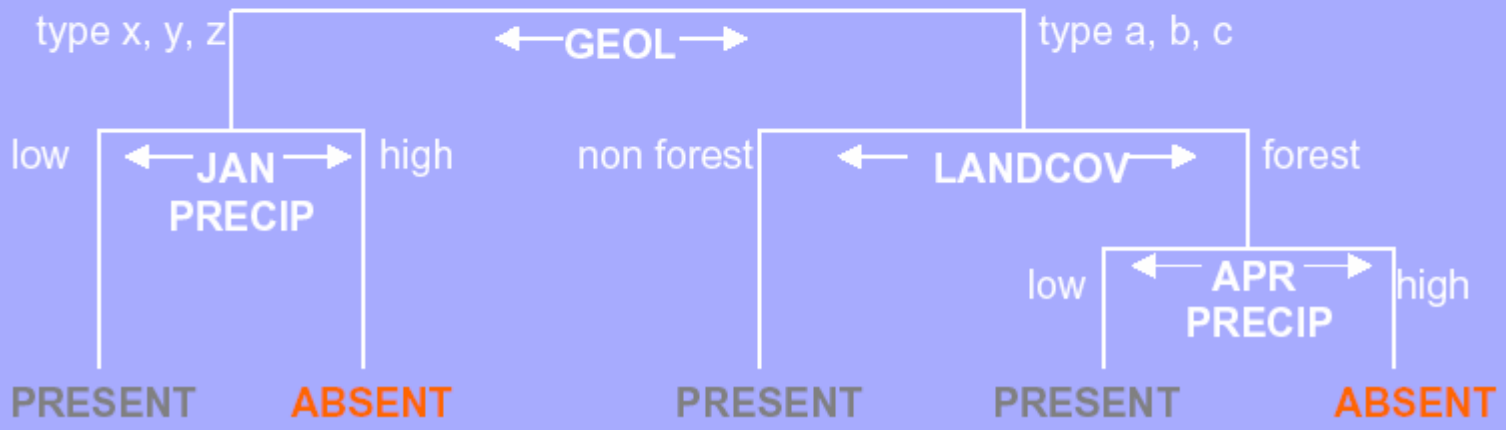


Foothill Yellow-legged Frog
Rana boylii



CART

- Non-parametric, data-driven algorithm
- Identifies threshold values to classify presence/absence occurrences
- Constructs of a dichotomous tree



CART

Advantages:

- Easy to interpret
- No estimation of response shape required
- Useful for non-linear, non-additive and hierarchical relationships
- Possible to investigate variable importance
- Pruning measures available for variable reduction
- Categorical data accepted

CART

Disadvantages:

- Computer intensive
- Less power than parametric methods when response functions simple
- Pruning methods not perfect
- Difficult to generate predicted distribution map
- Requires multiple software

Maximum Entropy

- Statistical mechanics approach
- Estimates the most uniform distribution (maximum entropy) given the constraint that the expected value of each environmental predictor variable matches its empirical mean
- Presence-only data
- Weights each variable by a constant

Maximum Entropy

- Uses a smoothing procedure called regularization
- Predictions are ‘cumulative values’
- Predictions are continuous (0–100)

Maximum Entropy

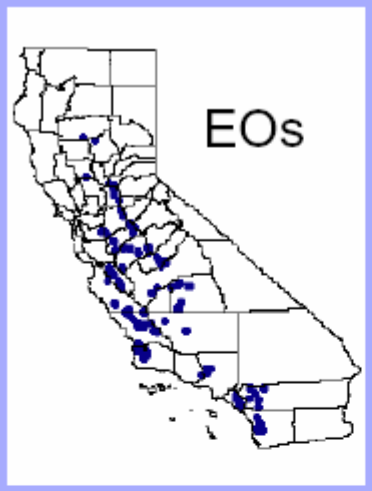
Advantages:

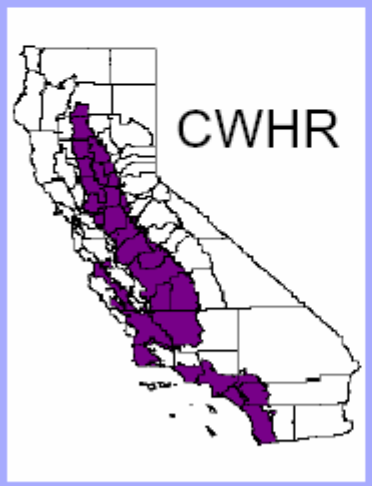
- Easy to implement
- Can consider interactions and non-linear relationships (quadratic)
- Categorical data accepted
- Possible to investigate variable importance
- Performs well with small sample sizes
- Stand alone software (freeware)

Maximum Entropy

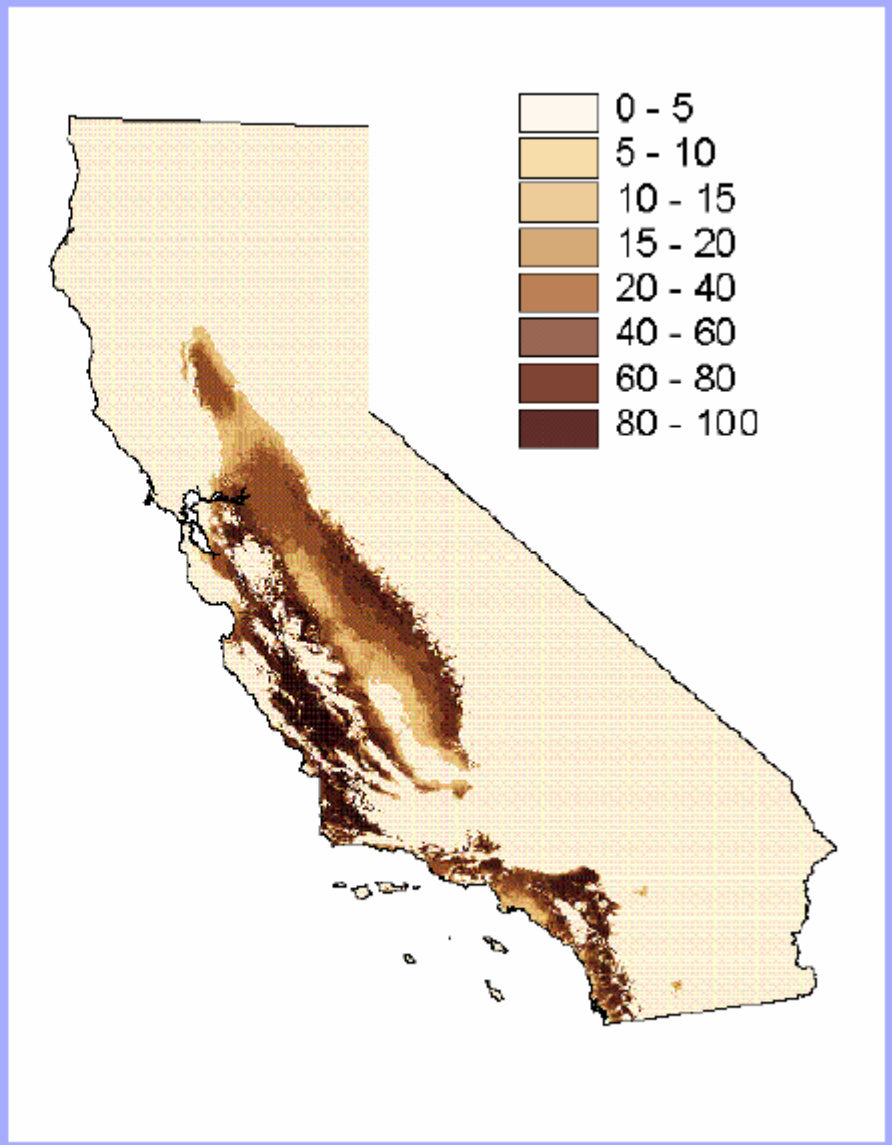
Disadvantages:

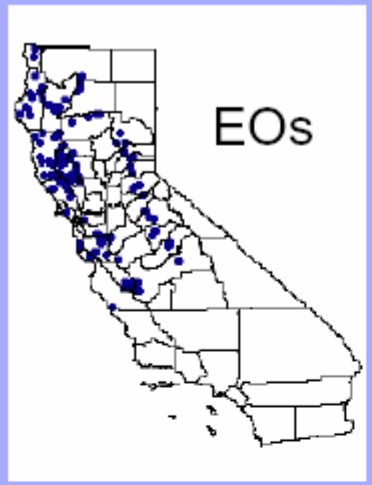
- Extremely computer intensive
- No procedure for variable selection
- New, untested in multiple situations

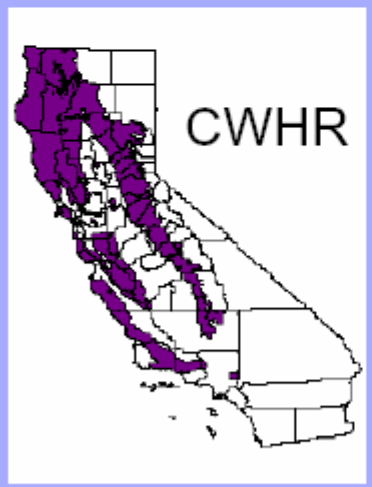




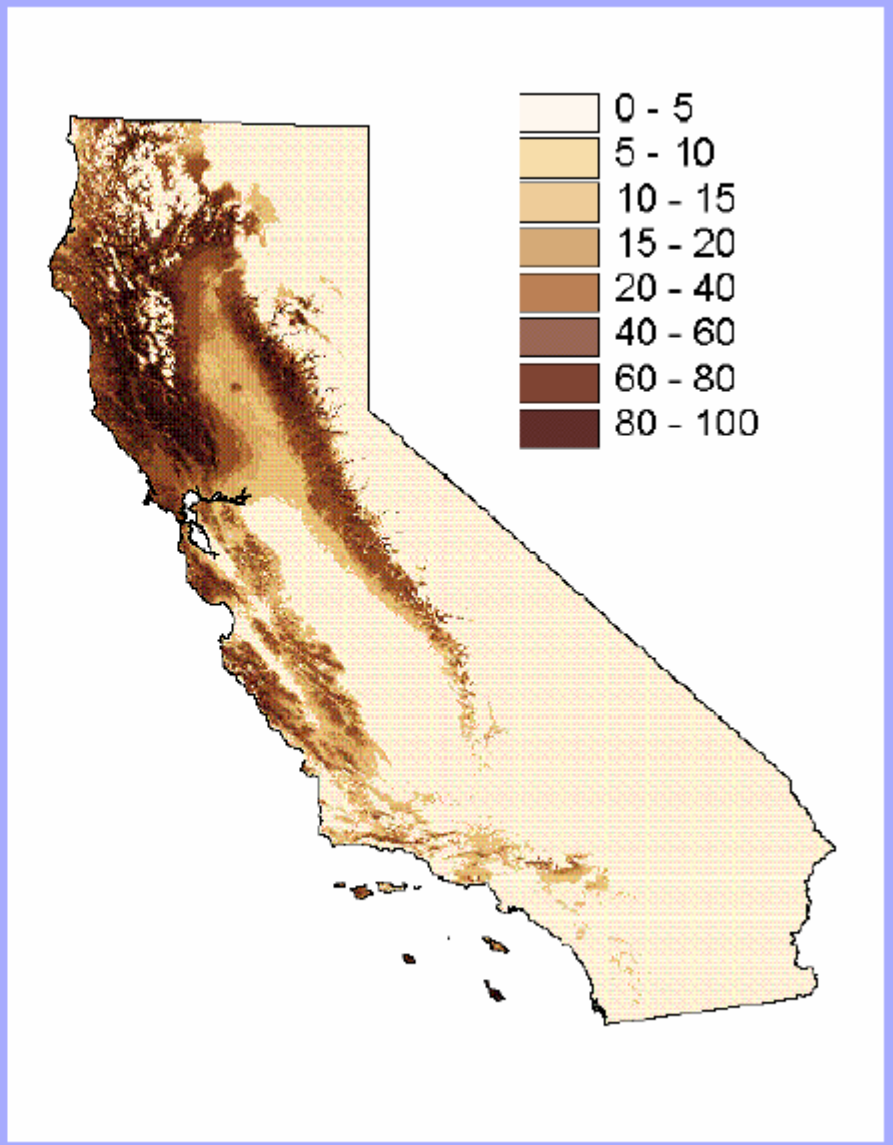
Western Spadefoot Toad
Scaphiopus hammondii







Foothill Yellow-legged Frog
Rana boylii



Genetic algorithms (GARP)

- Genetic Algorithm for Rule-set Prediction (GARP)
- Machine learning algorithm
- Uses several predictive algorithms (e.g. atomic, logistic regression, range rules, and negated range)
- Develops a set of ‘rules’

Genetic algorithms (GARP)

- Presence/pseudo-absence generated by algorithm
- Resamples the data into training and evaluation subsets
- First iteration generates the first rule and evaluates model omission and commission errors

Genetic algorithms (GARP)

- Following iterations - develops more rules
- Rules are included or excluded from the ‘rule set’ based on changes in model accuracy
- Process continues until it cannot create a better model or reached maximum iterations

Genetic algorithms (GARP)

- The final rule set consists of a series of if-then statements
- Predictions are binary values of presence/absence
- Outputs are not deterministic
- Fix - run GARP multiple times, select ‘best subset’ and arithmetically combine

EJEMPLO DE REGLAS DE GARP

The number preceding each rule (e.g., the first line here with numbers in it) gives coded information on:
Rule number, Rule type, Prior, Post-accuracy, Significance, Coverage, Usage (respectively)

5 r 0.50 1.00 29.52 0.22 0.620
IF - clim_01*0.25 - clim_02*0.01 + clim_05*0.17 - clim_20*0.23 + lith*0.26 + clim_06*0.21 +
clim_07*0.36 + clim_08*0.13 + clim_09*0.03 + clim_10*0.34 + clim_11*0.26 - clim_12*0.34 -
clim_13*0.30 + clim_14*0.06 - clim_16*0.04 + clim_17*0.07 - clim_18*0.45 + clim_19*0.08 +
clim_21*0.37 - clim_22*0.46 - clim_24*0.15 - clim_25*0.17 - clim_26*0.38 + clim_27*0.46 +
dem*0.04 + topopos*0.02 + wetind*0.41 - lat*0.01 - long*0.28 - evc*0.10 + slope*0.03 -
geol*0.09 - gully*0.26 + aspeast*0.04 - aspsouth*0.50 THEN species=ABSENT

19 d 0.50 0.85 16.58 0.14 0.023
IF clim_20=[14.9,15.5] AND clim_07=[20.3,25.3] AND clim_11=[4.5,9.2] AND clim_13=[105,178]u
AND clim_17=[184,262] AND clim_19=[334,485] AND clim_21=[24.5,25.4] AND clim_22=[5.9,6.6]u
AND clim_24=[7.5,13.5] AND topopos=[-82,48] AND wetind=[-2.2,23.1] AND lat=[-38.06,-37.10]u
AND long=[145.31,146.31] AND slope=[1.1,17.0] AND geol=[1, 5] THEN species=ABSENT
0 m 0.49 0.95 41.47 0.50 0.013 IF clim_01=[6.8,12.7] AND clim_02=[8.8,10.8] AND clim_05=[20.2,25.1]
AND clim_20=[14.9,15.3] AND lith=[3, 6] AND clim_06=[-2.4,3.4] AND clim_07=[21.2,23.0] AND
AND clim_17=[205,282] AND clim_18=[213,308] AND clim_19=[288,519]u
AND clim_21=[24.5,25.2] AND clim_22=[6.0,6.5] AND clim_24=[9.4,13.5] AND clim_25=[21.2,21.4]u
AND clim_26=[22.9,23.9] AND clim_27=[7.3,7.5] AND dem=[437,1387] AND topopos=[-70,109]u
AND wetind=[1.4,14.7] AND lat=[-37.93,-37.67] AND long=[145.62,146.36] AND evc=[11,30] AND
slope=[-4.4,32.7] AND geol=[3, 5] AND gully=[0, 1] AND aspeast=[-0.95,0.86] AND aspsouth=[-
0.91,0.93] THEN species=PRESENT

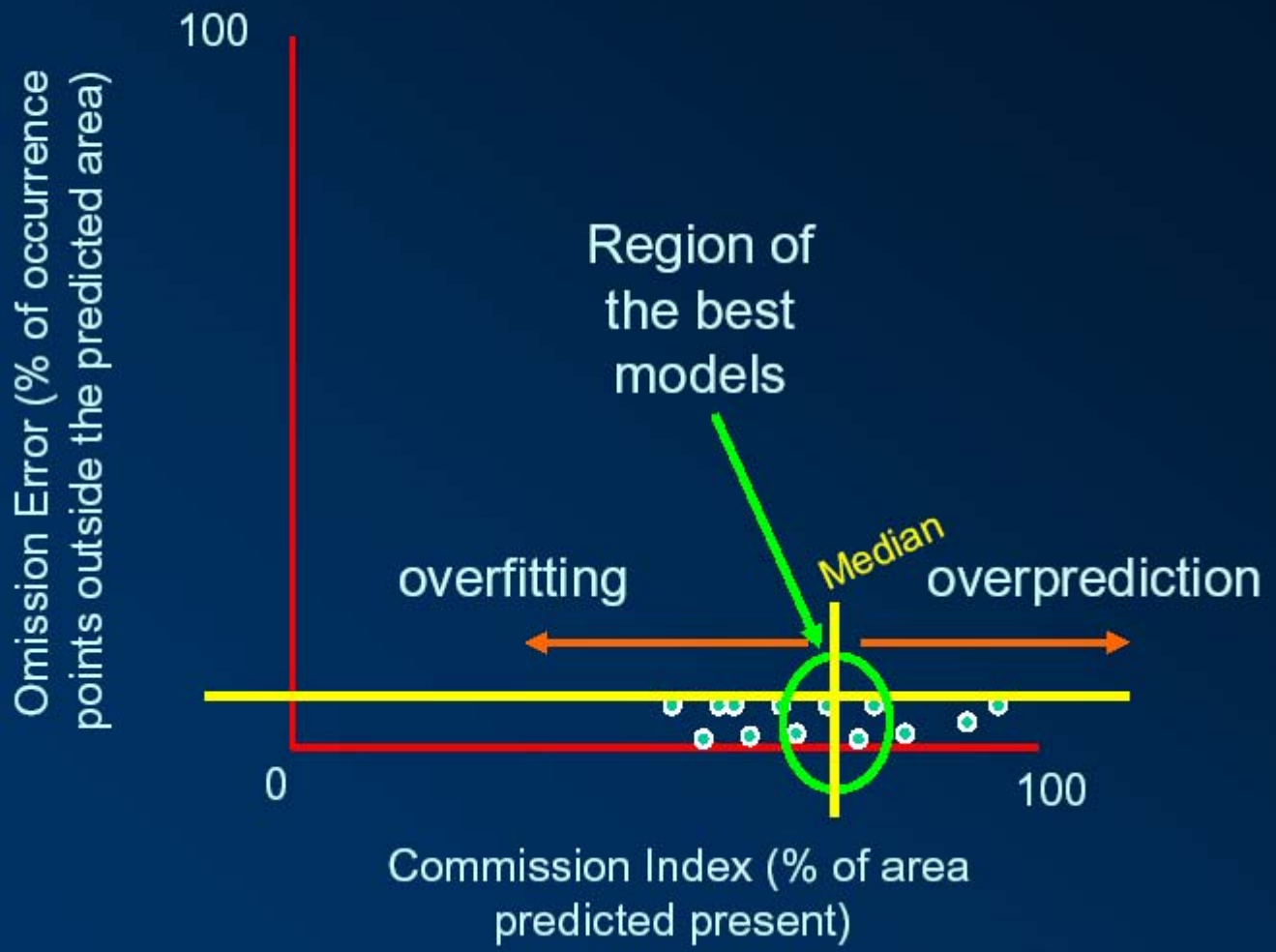
Elith, J. 2002. Predicting the distribution of plants. PhD Thesis. University of Melbourne

Genetic algorithms (GARP)

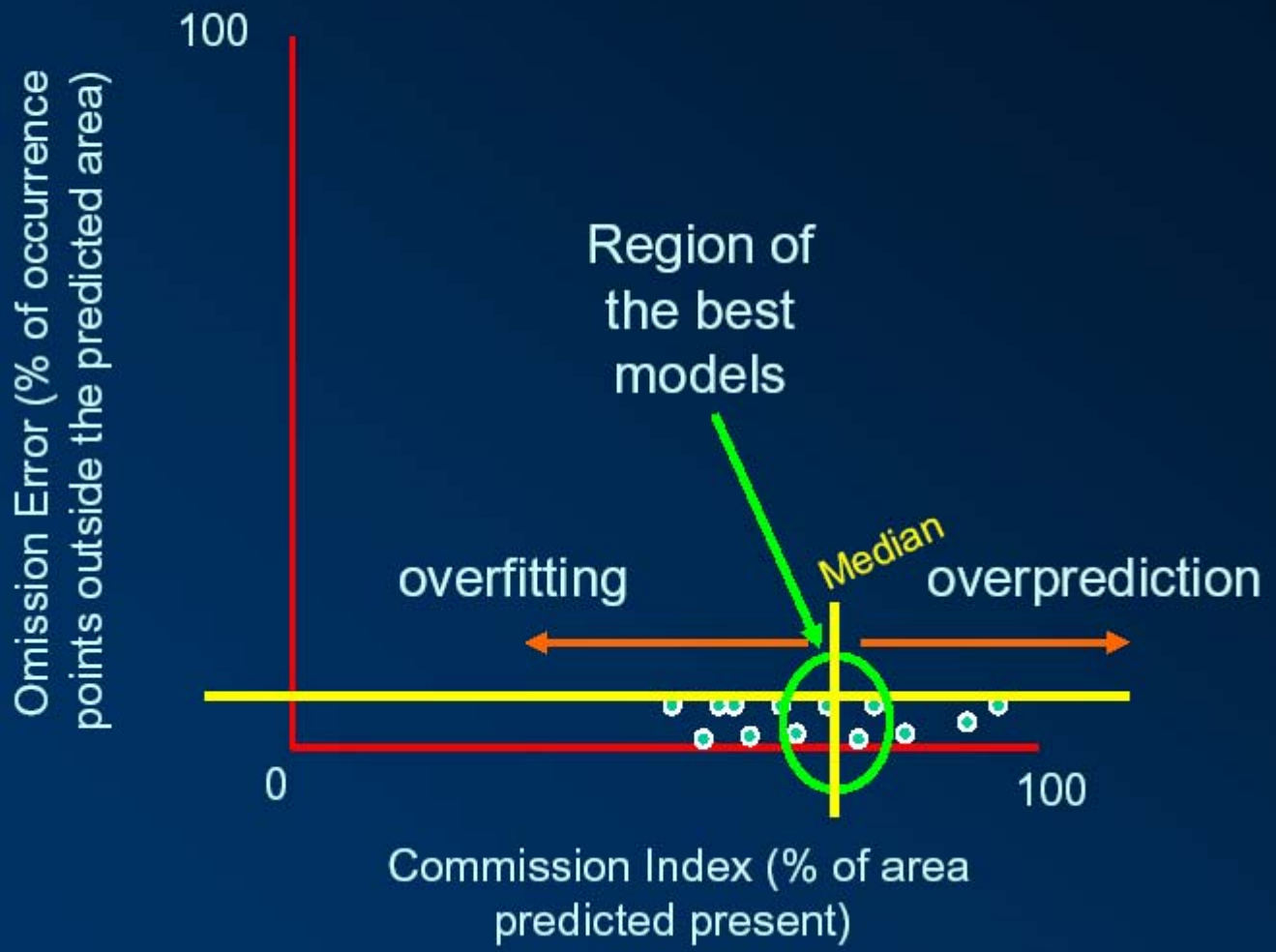
Advantages:

- Easy to implement
- Stand alone software (freeware)

The question now is, which of these models are good and which ones are bad?

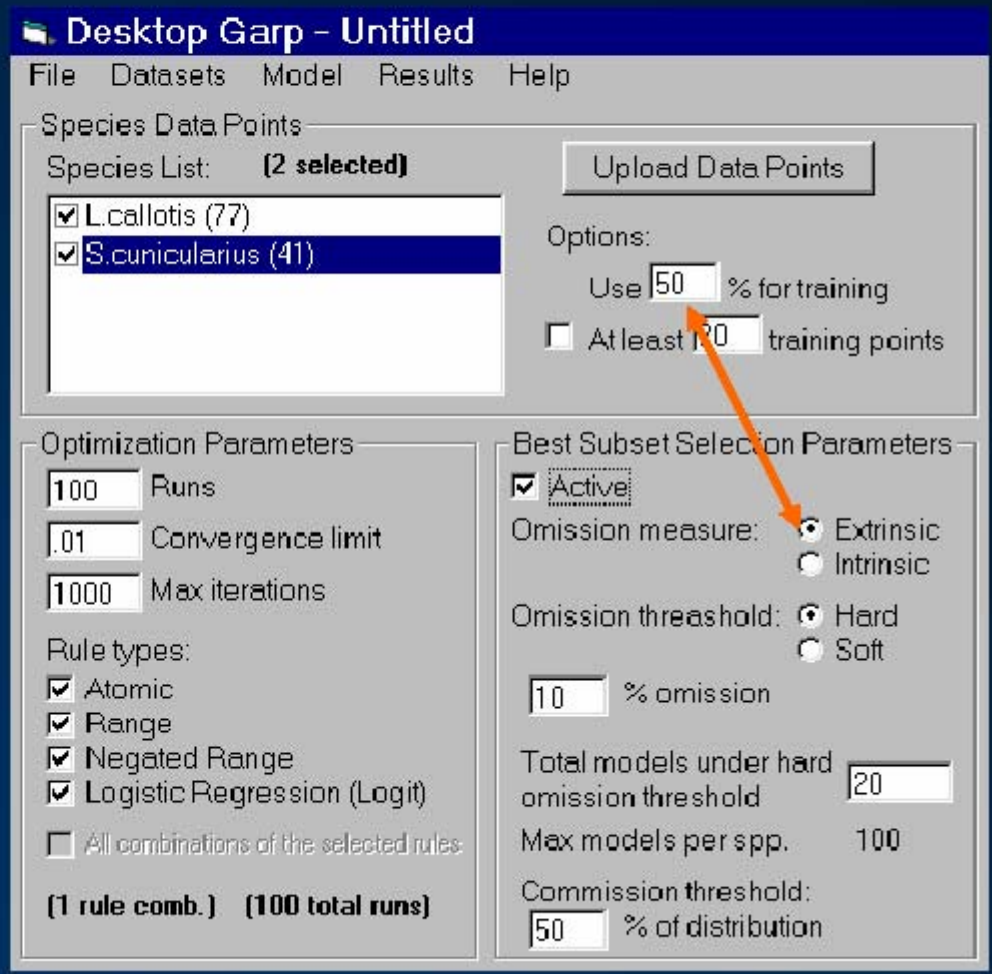


The question now is, which of these models are good and which ones are bad?



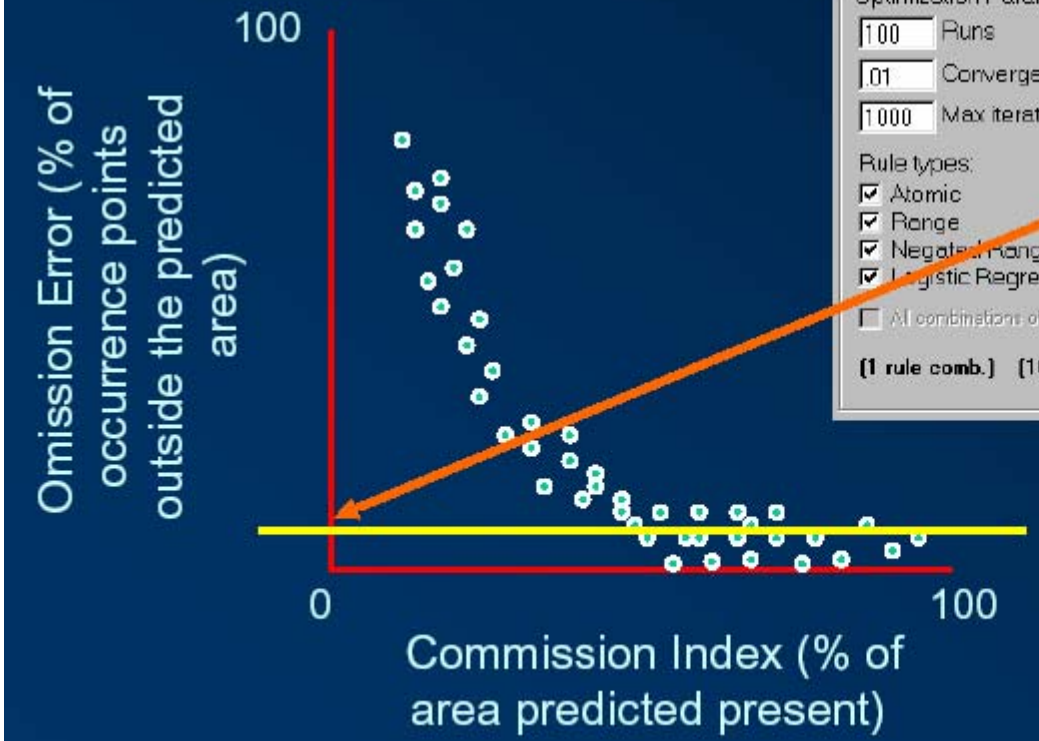
Implementation in Desktop GARP

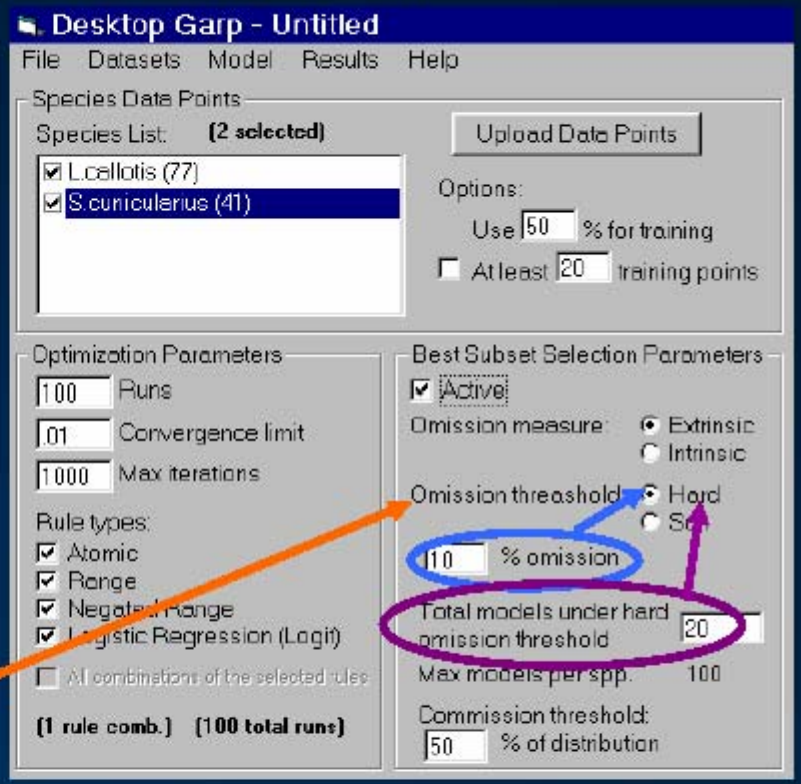
Having enough occurrence data, you can split them into *training* and *testing* datasets. When this is the case, it is convenient to select *Extrinsic* in the *Omission Measure* option. Otherwise, if you have *100% for training*, you have to select *Intrinsic*



Implementation in Desktop GARP

In the *Omission threshold* section, if you select *Hard* means that you will use an absolute value in the omission axis of the plot. You set that value in the *% omission* box

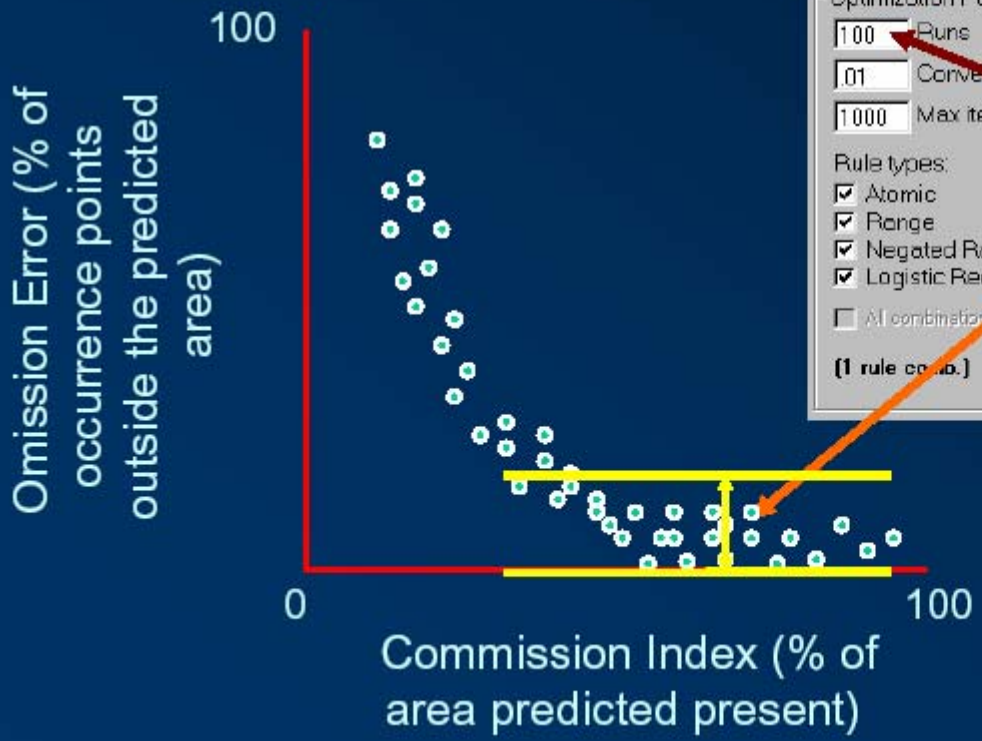


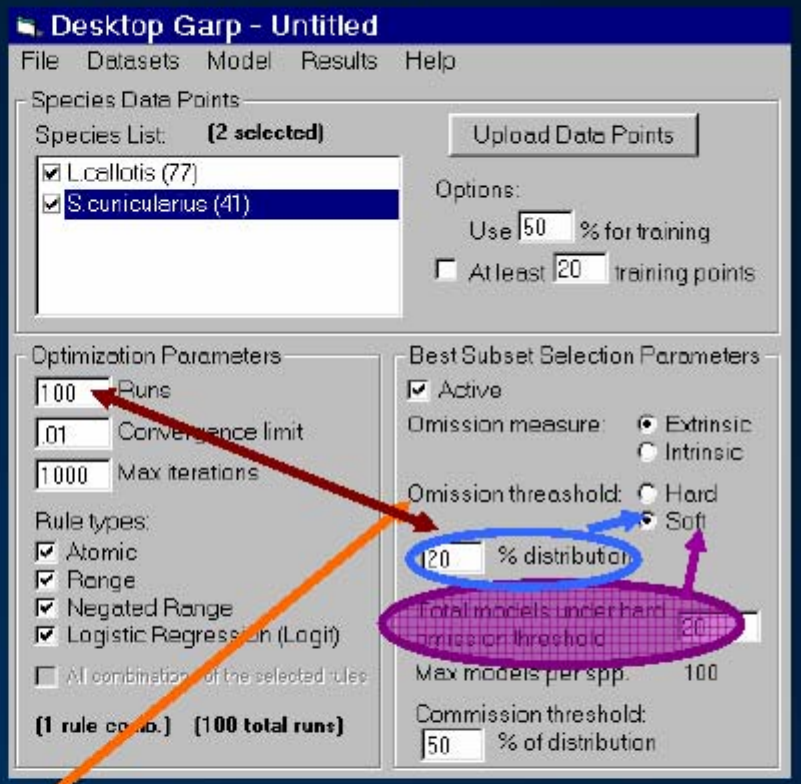


Then you have to select the number of models that you want DG to select under that hard omission threshold

Implementation in Desktop GARP

When you select *Soft* means that you will select certain number of models (in percentage), indicated in the *% distribution* box, with the *least* omission. This is useful when you are running more than one species at a time

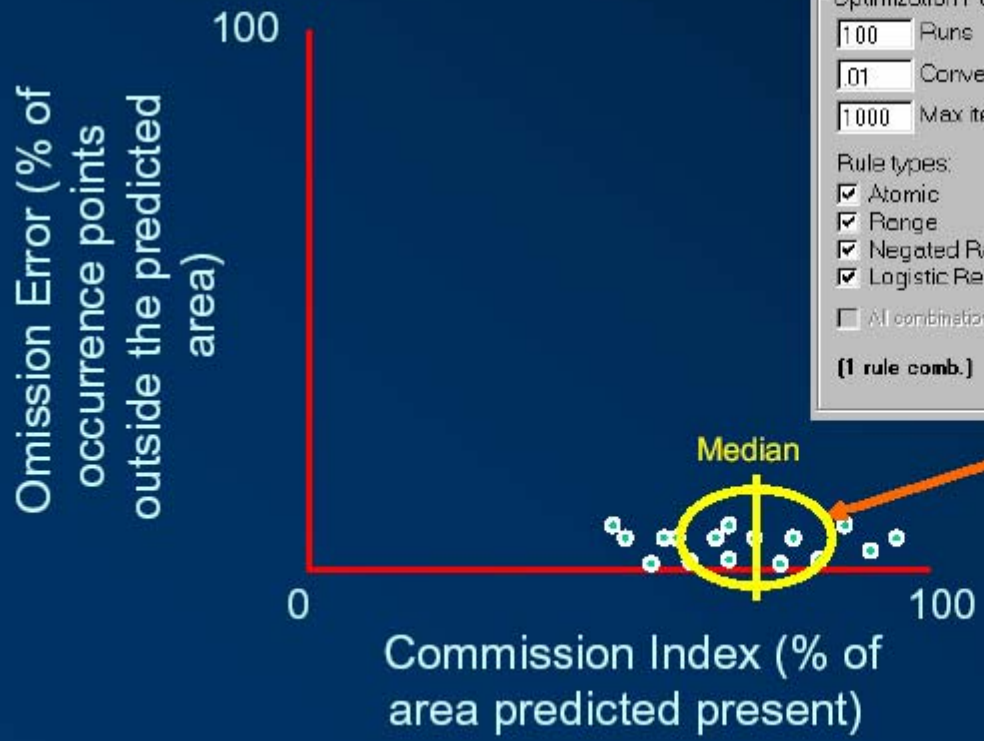


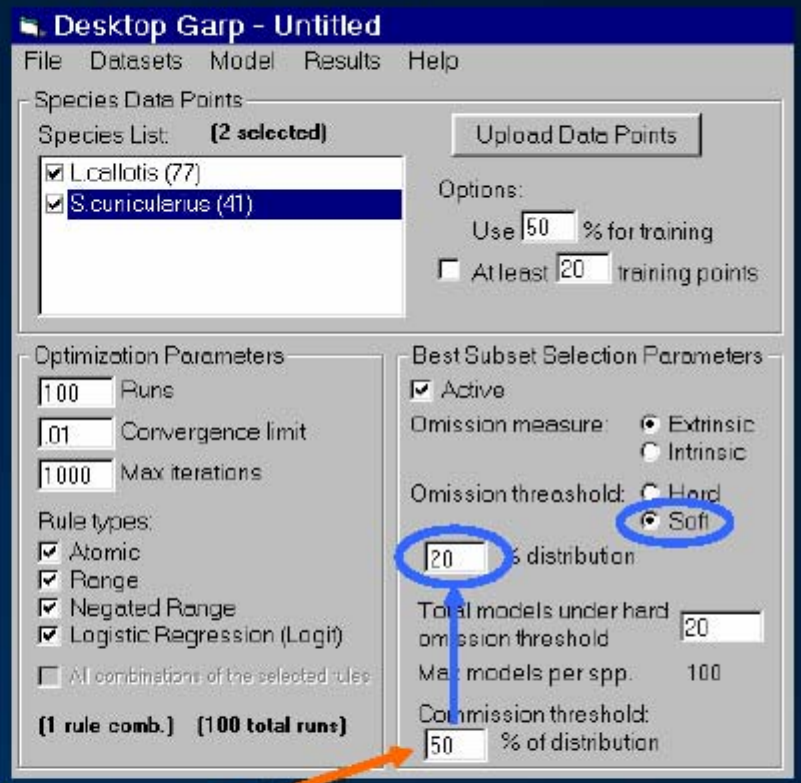


In this case, the *Total models under hard omission threshold* box does not apply

Implementation in Desktop GARP

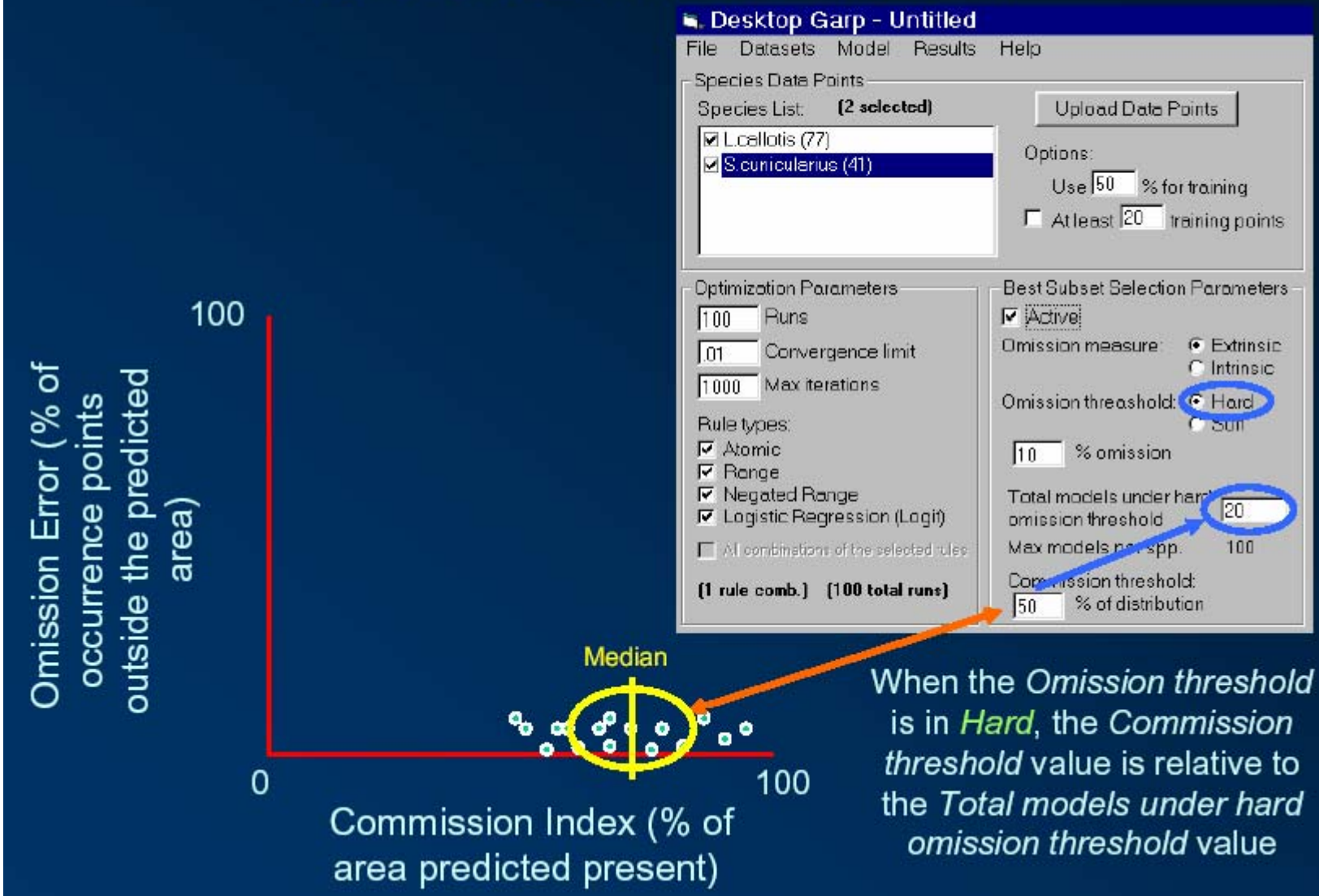
Finally, in the *Commission threshold* box you indicate the number of models (in percentage) closer to the Median in the Commission Index axis that you want to be selected from the remaining models, after filtering with the omission criteria





When the *Omission threshold* is in *Soft*, the *Commission threshold* value is relative to the *% distribution* value

Implementation in Desktop GARP

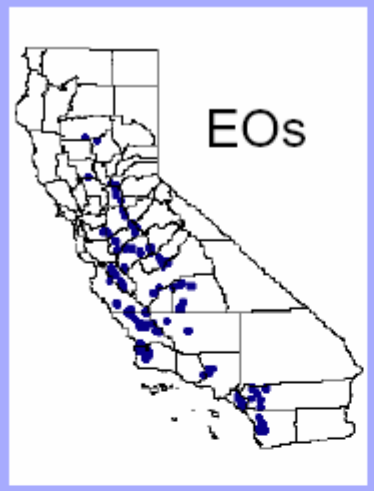


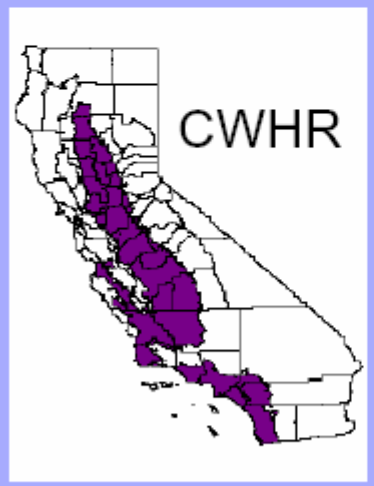
When the *Omission threshold* is in *Hard*, the *Commission threshold* value is relative to the *Total models under hard omission threshold* value

Genetic algorithms (GARP)

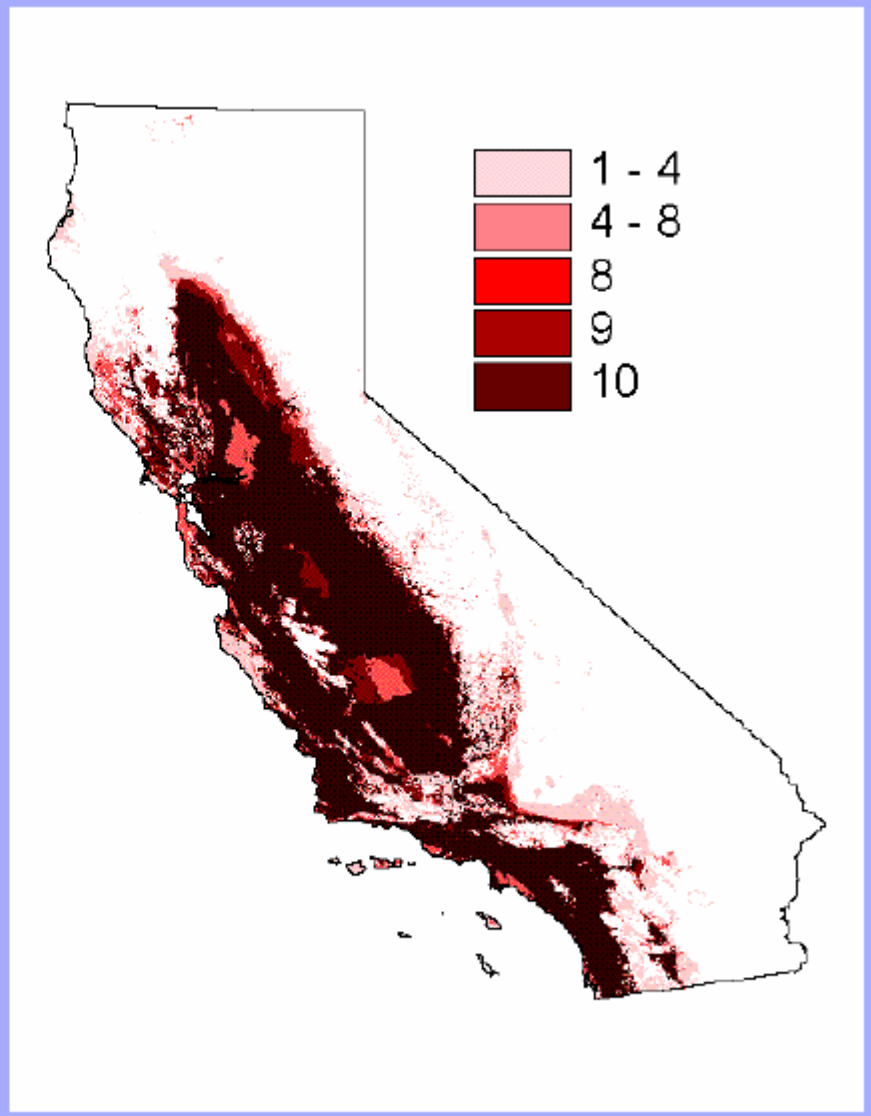
Disadvantages:

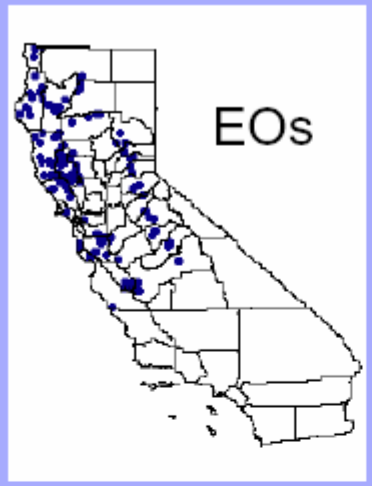
- Black box
- Maps not deterministic
- Internal generation of pseudo-absence
- Tendency for commission errors
- No procedure for variable selection
- Accepts categorical data?

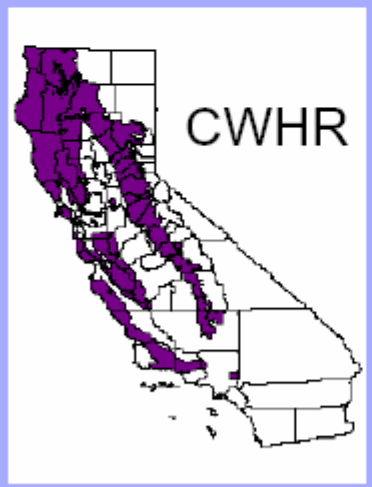




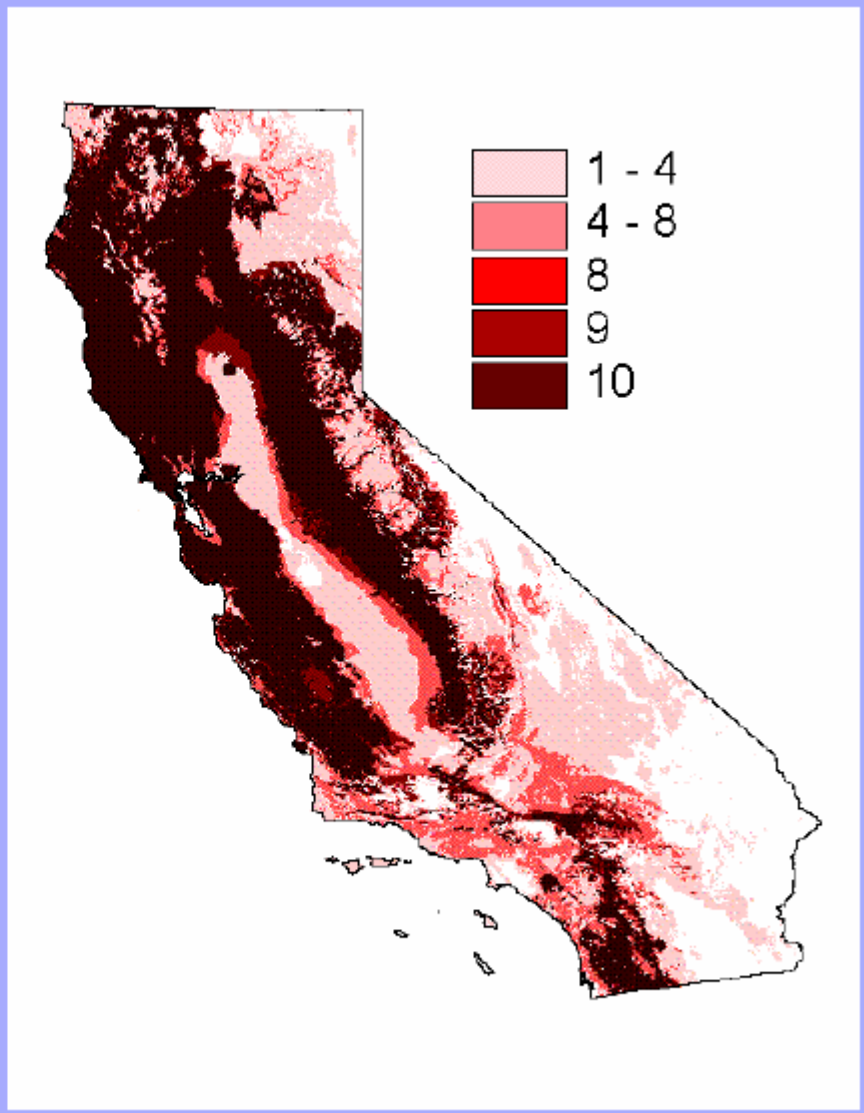
Western Spadefoot Toad
Scaphiopus hammondii



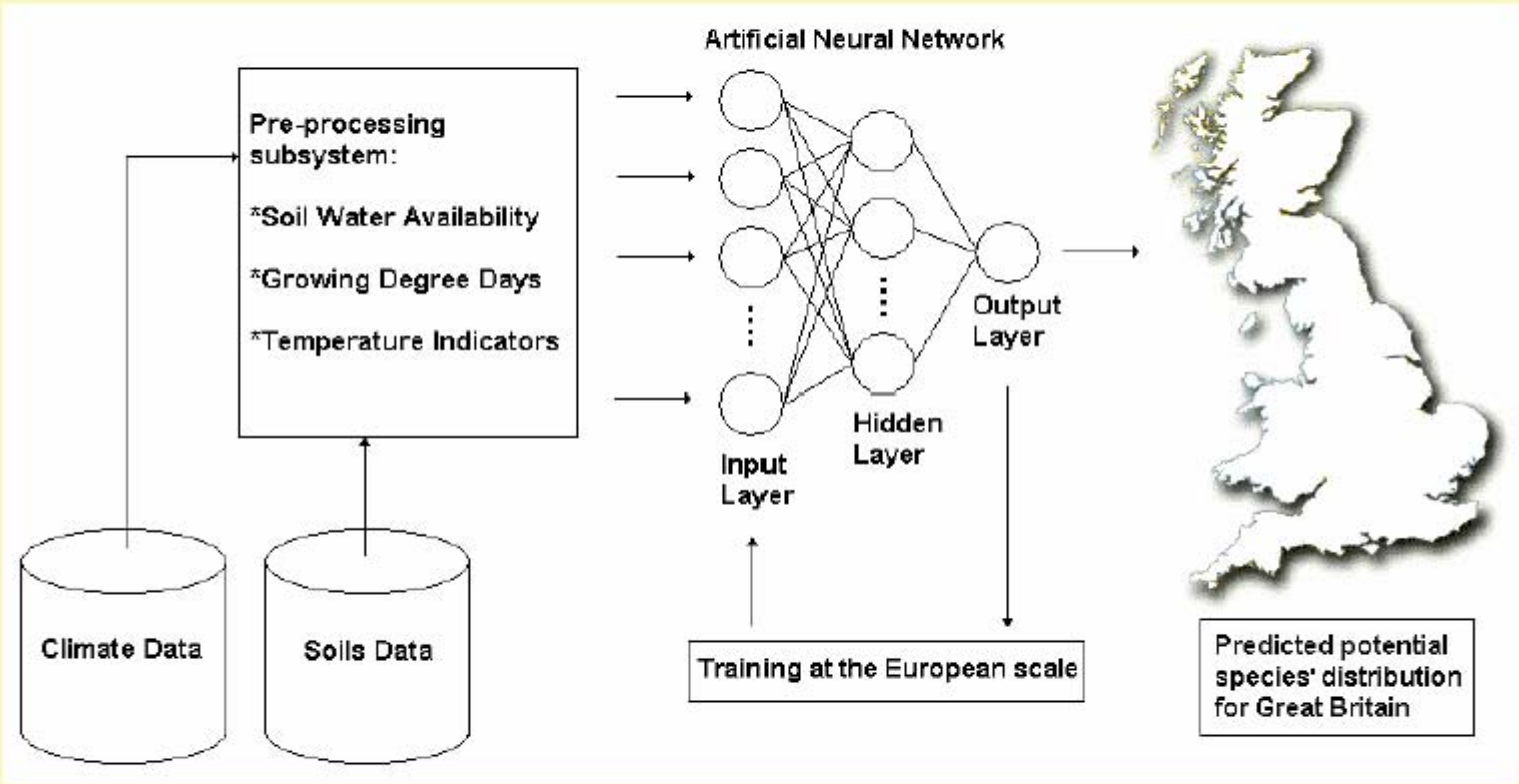




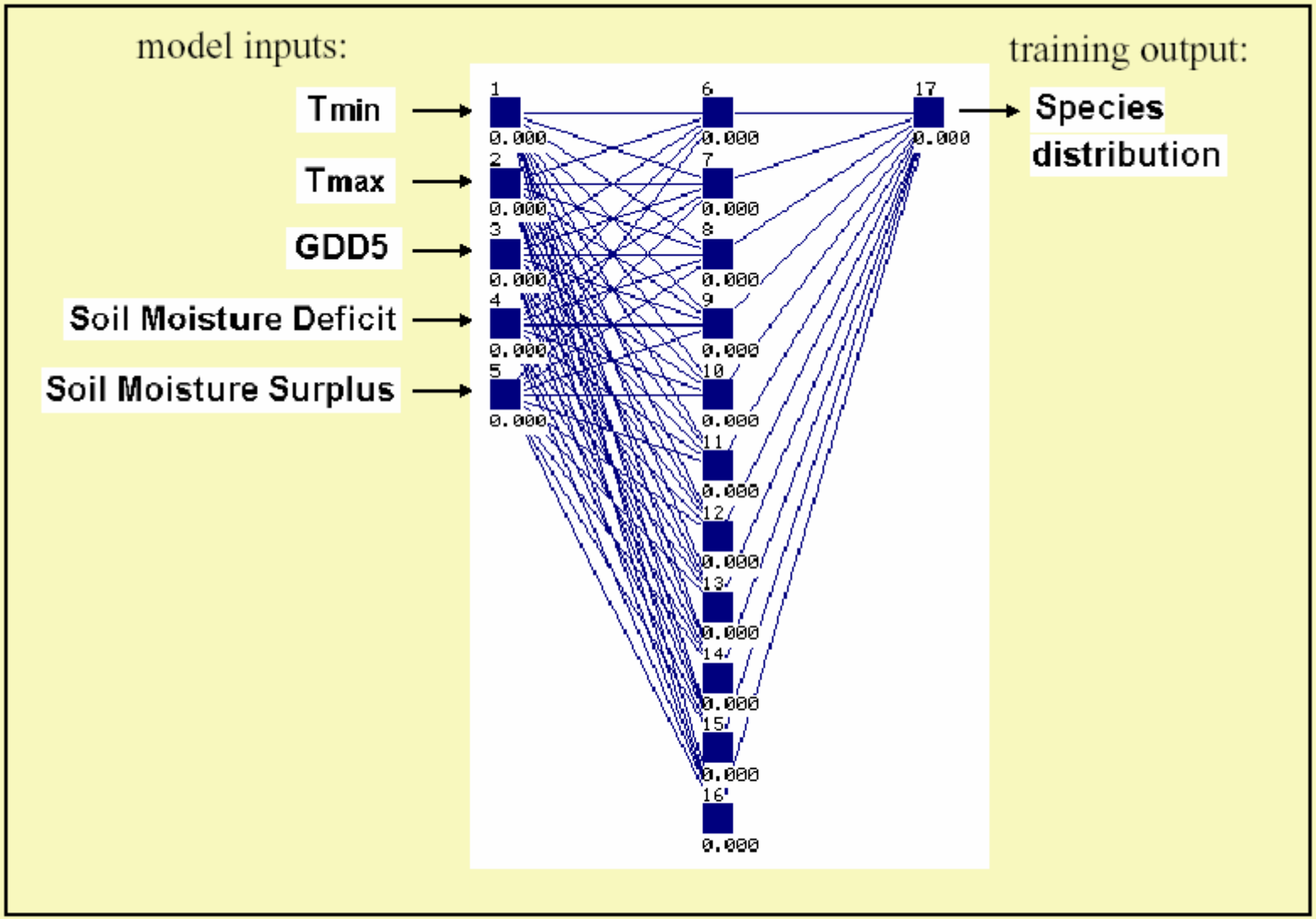
Foothill Yellow-legged Frog
Rana boylii



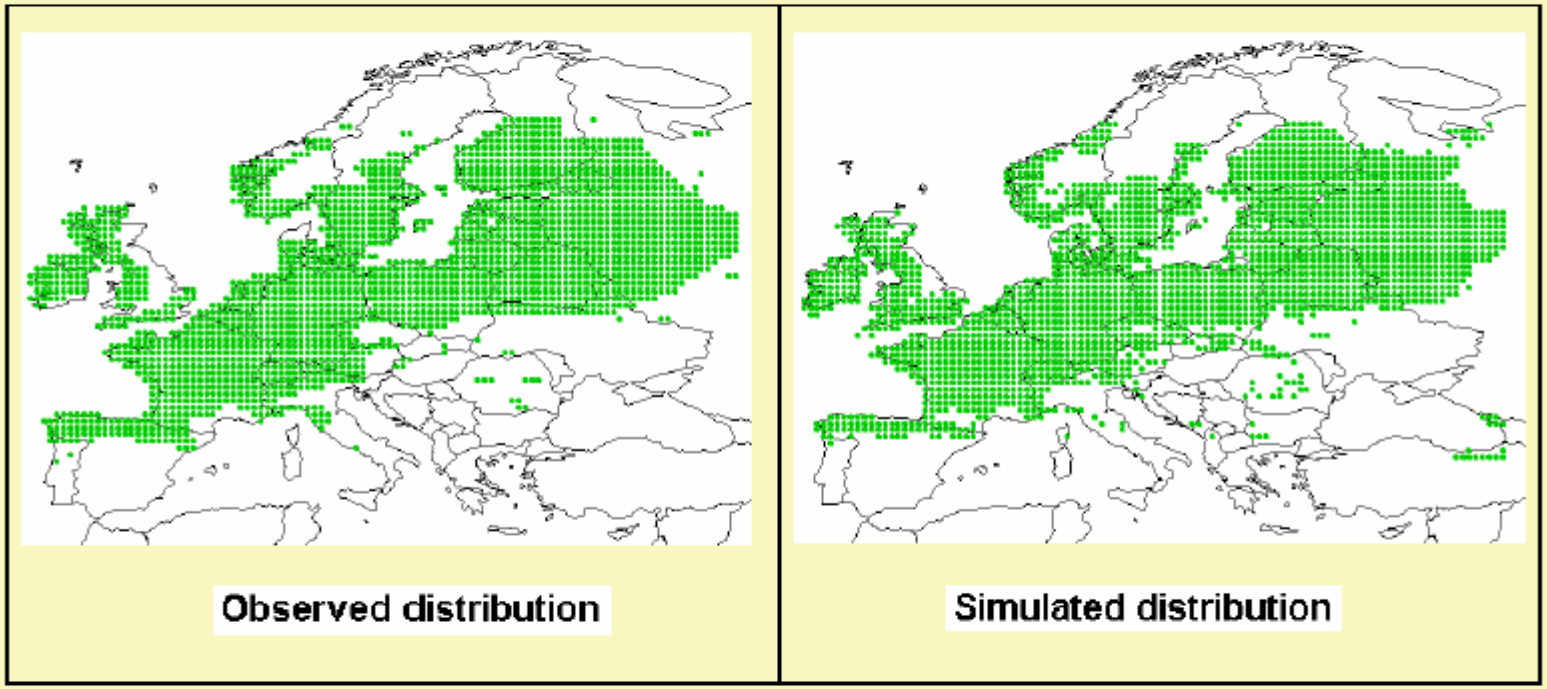
ANNs: The SPECIES model



(Pearson *et al*, 2002, *Ec Mod* 154)


model inputs:
Tmin
Tmax
GDD5
Soil Moisture Deficit
Soil Moisture Surplus
training output:
Species distribution
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
0.000

European climate space simulation for *Rhynchospora alba* (white beaked sedge)



(maximised Kappa = 0.83; mean Kappa for 32 species = 0.77)

Other Algorithms
Ordination Techniques

- Ecological niche factor analysis (Biomapper): presence-only model
- Canonical correspondence analysis (CCA): probably better for communities (Guisan et al., 1999)
- Discriminant analysis (DA) (Manel et al., 1999)

Other Algorithms

- Artificial Neural Networks (ANN, NNETW)
- Multivariate adaptive regression splines (MARS)
- WhyWhere