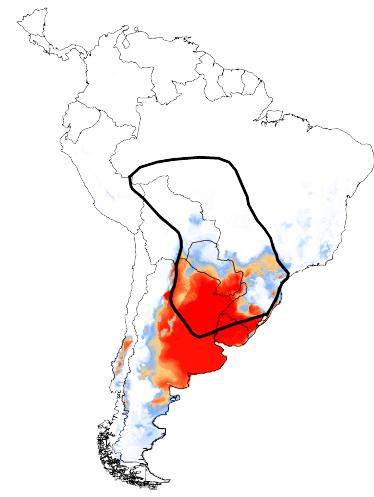
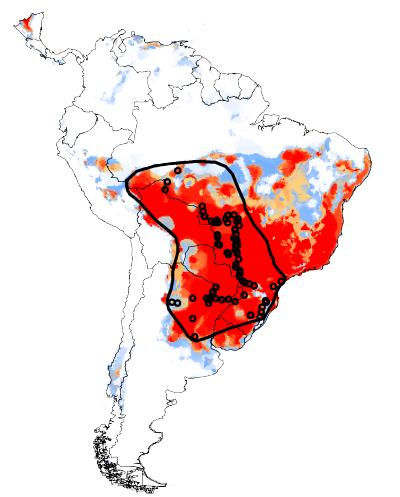


# Evaluating SDMs

# Model evaluation

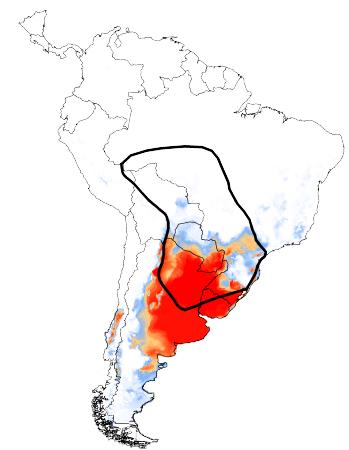
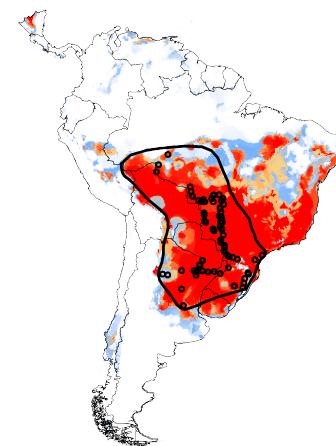
- ▶ Not the same as model selection
  - ▶ Based on information criterion (e.g., AIC, BIC)
- ▶ Most evaluations focus on predictive accuracy
- ▶ Other considerations
  - ▶ Realism
  - ▶ Spatial pattern of error
    - ▶ Ecologically interesting?
    - ▶ Over- vs. under-prediction



# Model evaluation

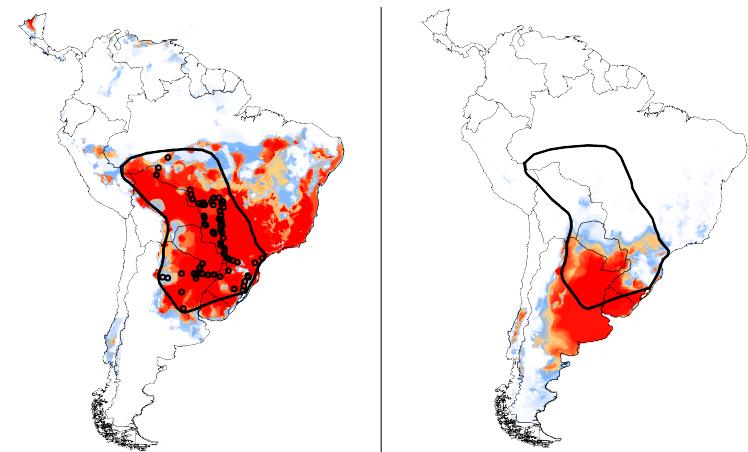
---

- ▶ Should check both:
  - ▶ **Discrimination**
  - ▶ **Calibration**



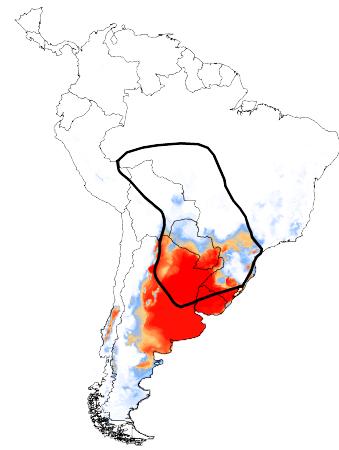
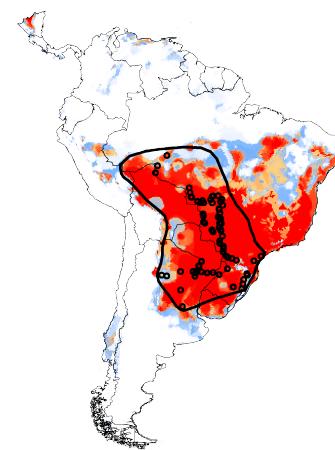
# Model evaluation

- ▶ Should check both:
  - ▶ **Discrimination**
    - ▶ the ability of the model to correctly distinguish between occupied and unoccupied sites
    - ▶ Most studies measure and report only discrimination ability
    - ▶ Two kinds
      - Threshold dependent
      - Threshold independent



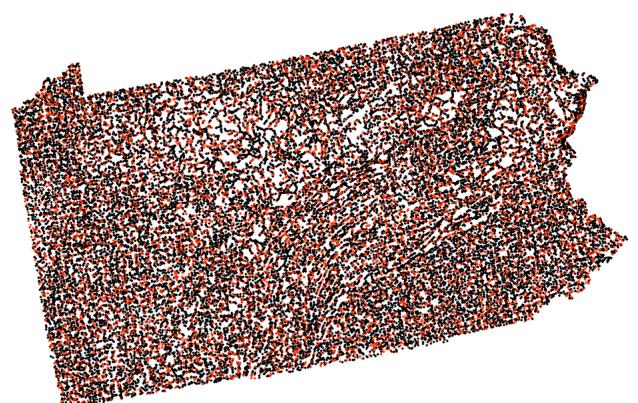
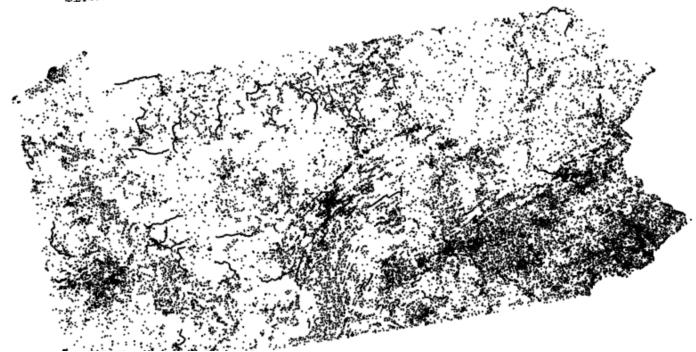
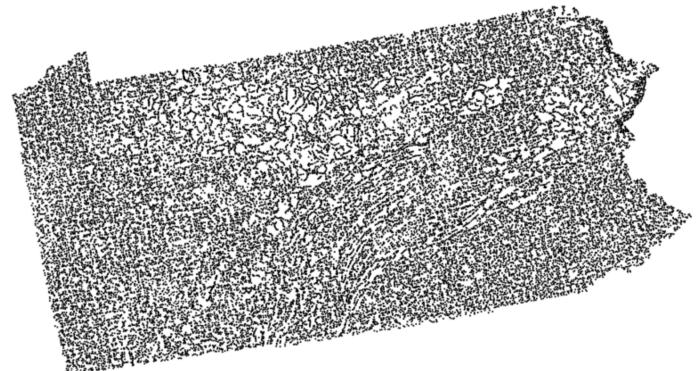
# Model evaluation

- ▶ Should check both:
  - ▶ **Calibration** (Reliability)
  - ▶ Also important
    - ▶ Do predictions of 0.6 have a 60% chance of being occupied and are these twice as likely (or twice as suitable) as predictions of 0.3?
    - ▶ Brier Score
      - ▶ equivalent to mean squared error
      - ▶ applicable when comparing continuous probabilities to presence / absence
      - ▶ Ranges from 0 to 1
        - ▶ 0 = complete agreement between observed and predicted
        - ▶ 1 indicates complete disagreement.



# Data for model evaluation

- ▶ New / independent data are best, but usually not available
  - ▶ Model performance over estimated when using the same data for training & evaluation
- ▶ Partition data into training / testing sets
  - ▶ i.e. 75 training / 25% testing
  - ▶ Testing proportion =  $1/(1+\sqrt{p-1})$  where p is the number of predictors
    - ▶ 50% for 2 predictors
    - ▶ 33% for 5 predictors



# Data for model evaluation

$K = 10$

- ▶ Partition data into training / testing sets
  - ▶ K-fold cross validation
    - ▶ Partition data  $k$  times at random
    - ▶ Fit models to each resulting set
  - ▶ For small sample sizes
    - ▶ Bootstrap sampling with replacement
    - ▶ Jackknife
    - ▶ Use all data to fit final model



# Model evaluation

## ▶ K-fold cross validation using spatial blocks

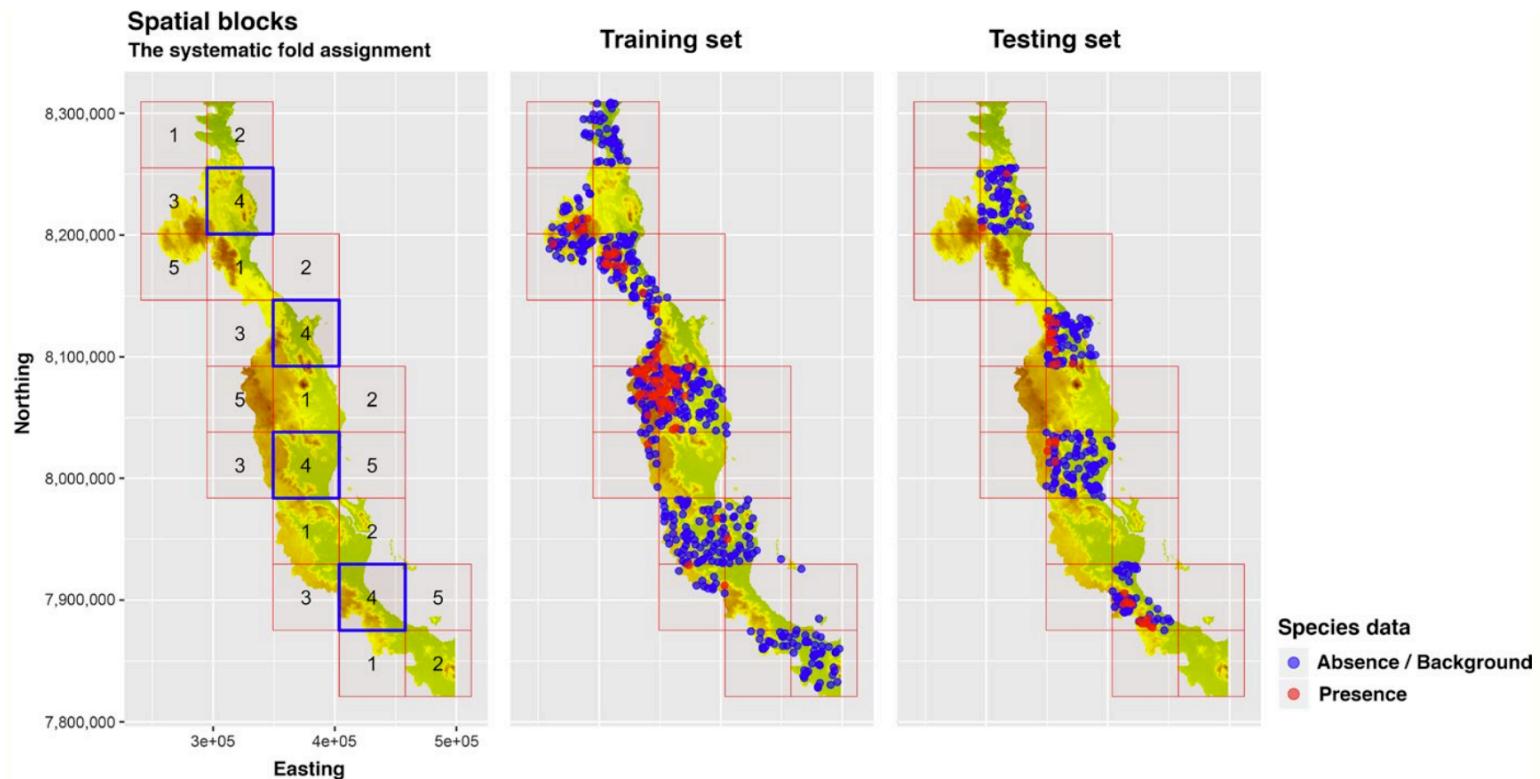
Received: 19 January 2018 | Accepted: 5 October 2018  
DOI: 10.1111/2041-210X.13107

### APPLICATION

Methods in Ecology and Evolution 

BLOCKCV: An R package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models

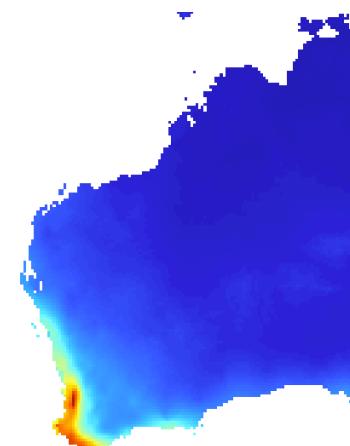
Roozbeh Valavi  | Jane Elith  | José J. Lahoz-Monfort  |  
Gurutzeta Guillera-Arroita 



# Discrimination metrics - threshold dependent

- ▶ Use threshold to convert continuous prediction to presence (1) and absence (0)
- ▶ Construct “confusion matrix”

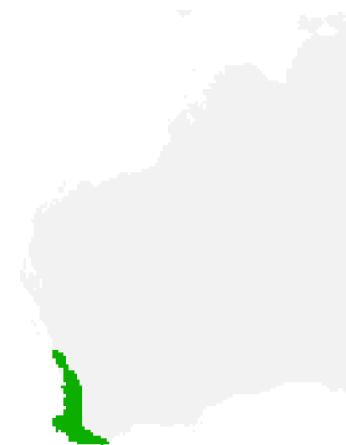
		Observed	
		True	False
Predicted	True	correct	error
	False	error	correct



# Discrimination metrics - threshold dependent

- ▶ Use threshold to convert continuous prediction to presence (1) and absence (0)
- ▶ Construct “confusion matrix”

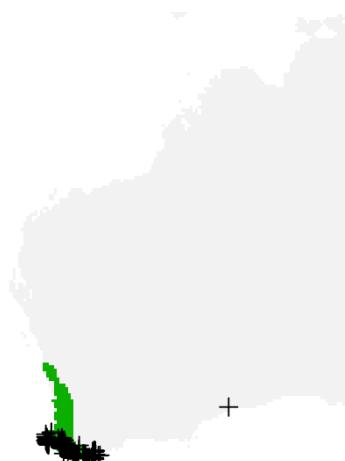
		Observed	
		True	False
Predicted	True		
	False		



# Discrimination metrics - threshold dependent

- ▶ Use threshold to convert continuous prediction to presence (1) and absence (0)
- ▶ Construct “confusion matrix”

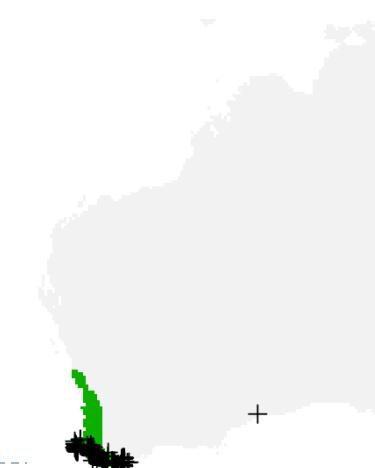
		Observed	
		True	False
Predicted	True		
	False		



# Discrimination metrics - threshold dependent

- ▶ Use threshold to convert continuous prediction to presence (1) and absence (0)
- ▶ Construct “confusion matrix”
- ▶ Compare / assess error rates for presences and absences

		Observed	
		True	False
Predicted	True	639	87
	False	138	391

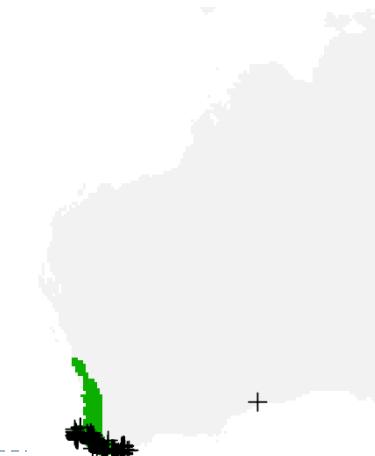


# Discrimination metrics - threshold dependent

## ▶ Sensitivity

- ▶ True positive rate
  - ▶  $TP / (TP+FN)$
- ▶ Proportion of correctly predicted presences
- ▶ Omission error = predicting species absent where it is present
  - ▶ Want to minimize when predicting invasions

		Observed	
		True	False
Predicted	True	639	87
	False	138	391

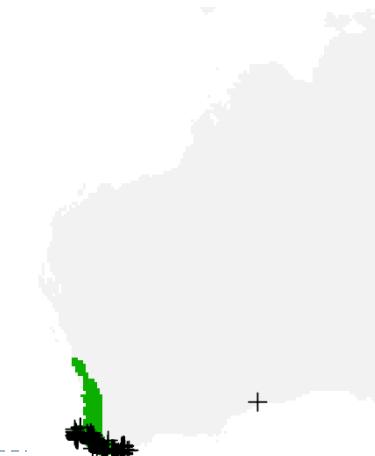


# Discrimination metrics - threshold dependent

## ▶ Specificity

- ▶ True negative rate
  - ▶  $TN/(TN+FP)$
- ▶ Proportion of correctly predicted absences
- ▶ Commission “errors” = predicting species present where it is “absent”
  - ▶ Want to minimize when design reserves / protecting habitat

		Observed	
		True	False
Predicted	True	639	87
	False	138	391



# Discrimination metrics - threshold dependent

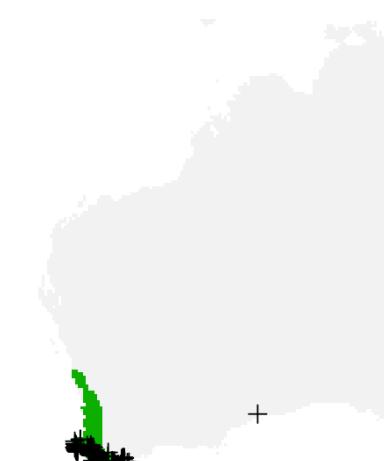
## ▶ Kappa

- ▶ Takes the proportion of correct predictions expected by chance into account
- ▶ Sensitive to prevalence

## ▶ True Skill Statistic (TSS)

- ▶  $1 - \max(\text{Sensitivity} + \text{Specificity})$
- ▶ -1 to +1
  - ▶ +1 = perfect agreement
  - ▶ 0 = no better than random
- ▶ Less sensitive to prevalence
- ▶ Equal to kappa when the number of presences and absences in the validation set are equal (prevalence=0.5)

		Observed	
		True	False
Predicted	True	639	87
	False	138	391

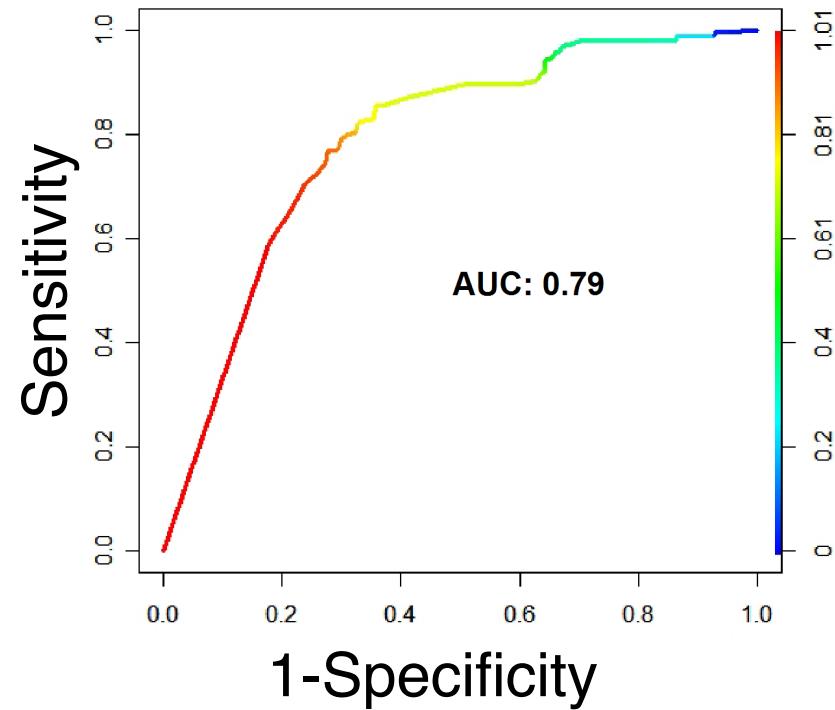


# Discrimination metrics - threshold independent

## AUC-ROC

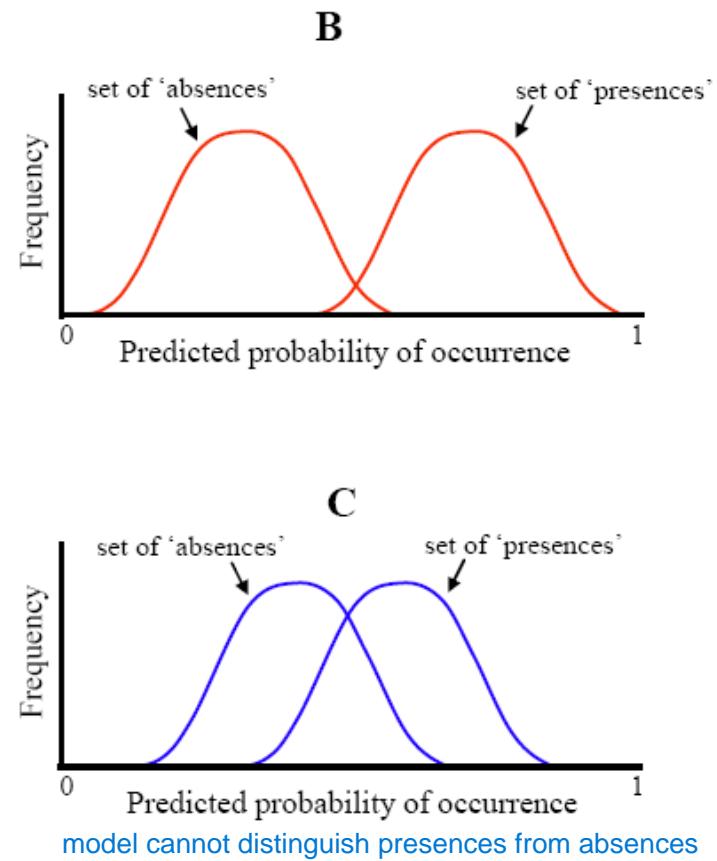
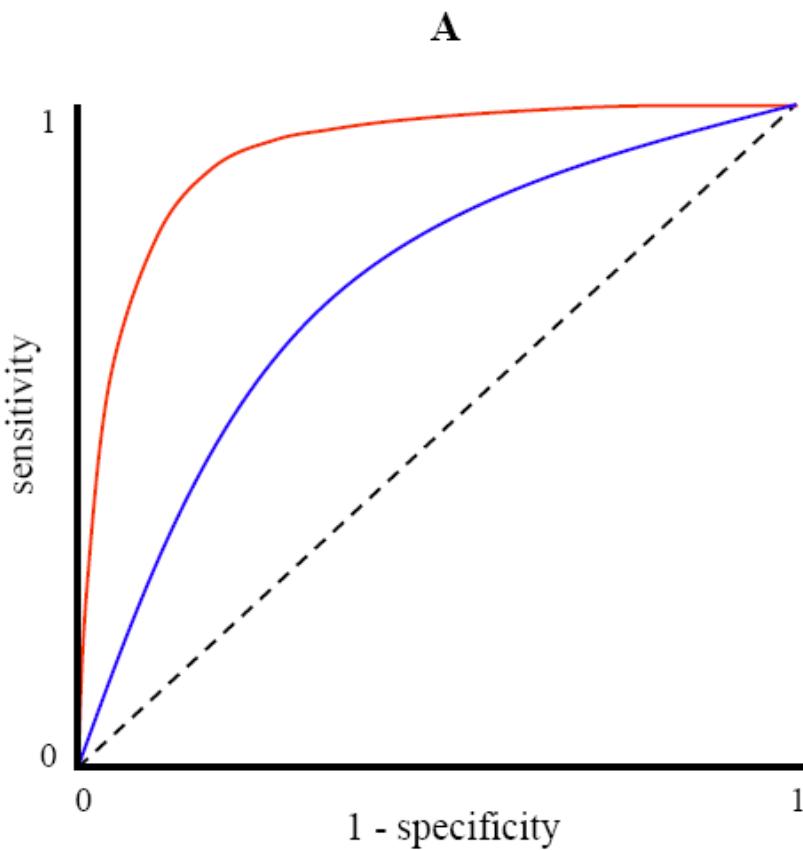
- ▶ Area Under the Curve of the Receiver-Operating characteristic (ROC) plot
- ▶ The proportion of random selections from the positive group that will score higher than random selections from the negative group
- ▶ 0-1
  - ▶ AUC = 0.5 = no better than random
  - ▶ Calculated across all probability values predicted by the model

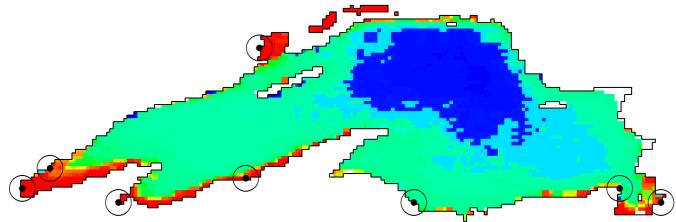
positive is presence and negative is absence



# Discrimination metrics - threshold independent

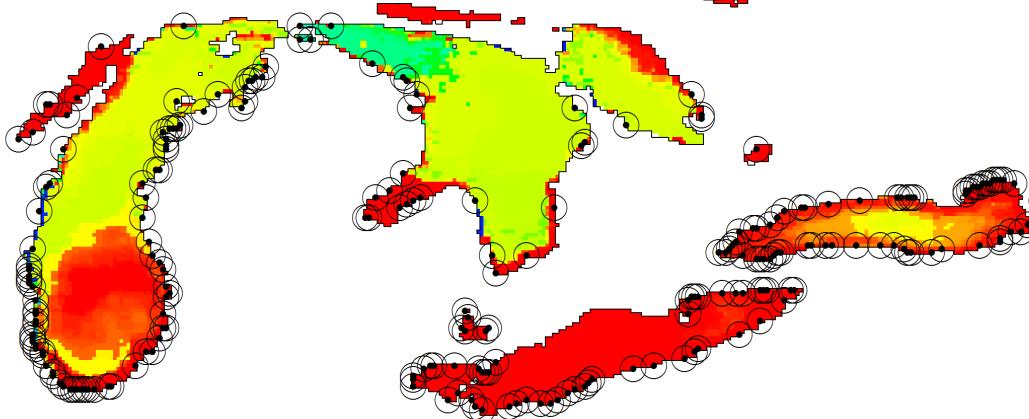
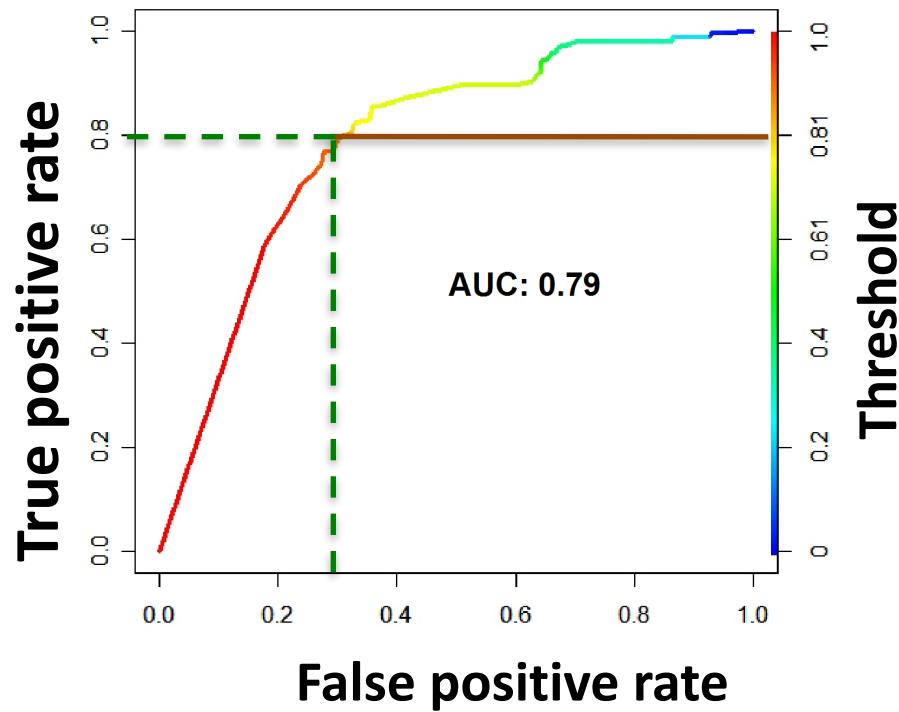
## AUC





**Predicted  
Suitability**

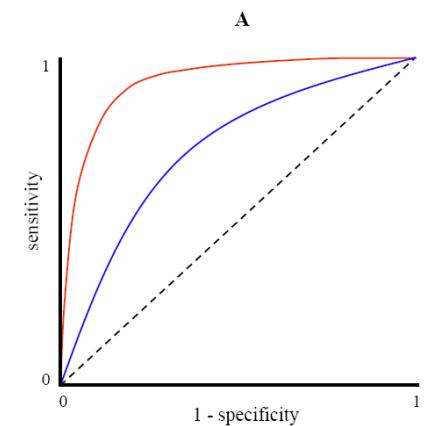
- High (100)
- Mid (50)
- Low (0)



# Discrimination metrics - Presence Only

- ▶ Presence-only data present a problem for model evaluation
- ▶ All discrimination metrics can be applied to presence-background, but interpretation changes
- ▶ AUC=the probability a model scores a random presence site higher than a random background site

		Observed	
		True	False
Predicted	True	639	87
	False	138	391



# Discrimination metrics - AUC-PR

---

- ▶ AUC-PR (Area Under the Precision-Recall Curve)

Received: 22 May 2018

Accepted: 25 November 2018

DOI: 10.1111/2041-210X.13140

**RESEARCH ARTICLE**

Methods in Ecology and Evolution 

## The area under the precision-recall curve as a performance metric for rare binary events

Helen R. Sofaer<sup>1</sup>  | Jennifer A. Hoeting<sup>2</sup>  | Catherine S. Jarnevich<sup>1</sup> 

# Discrimination metrics - AUC-PR

- ▶ Precision = probability that a species is present given a predicted presence
- ▶ Recall (more commonly called sensitivity) = probability the model predicts presence in locations where the species has been observed.

	Observed present	Observed absent	
Predicted present	TP	FP	Precision = TP/(TP+FP)
Predicted absent	FN	TN	
	Recall = Sensitivity = TP/(TP+FN)	Specificity = TN/(FP+TN)	

AUC-ROC is the area under the curve of the plot of sensitivity (recall) versus 1-specificity across thresholds. Sensitivity is based on all observed presences (left column), while specificity is based on all observed (or inferred) absences (right column). Thus, AUC-ROC incorporates all quadrants of the confusion matrix. AUC-PR is the area under the curve of the plot of precision versus recall (sensitivity) across thresholds. Precision is based on all predicted presences (top row). Thus, AUC-PR does not incorporate the number of true negatives (TN); cells used in calculating AUC-PR are shaded grey and outlined in dashes.

TP, True positive; FP, False positive; FN, False negative; TN, True negative.

# Discrimination metrics - AUC-PR

- ▶ AUC-PR does not incorporate correctly predicted absences and is therefore less prone to exaggerate model performance for unbalanced datasets.

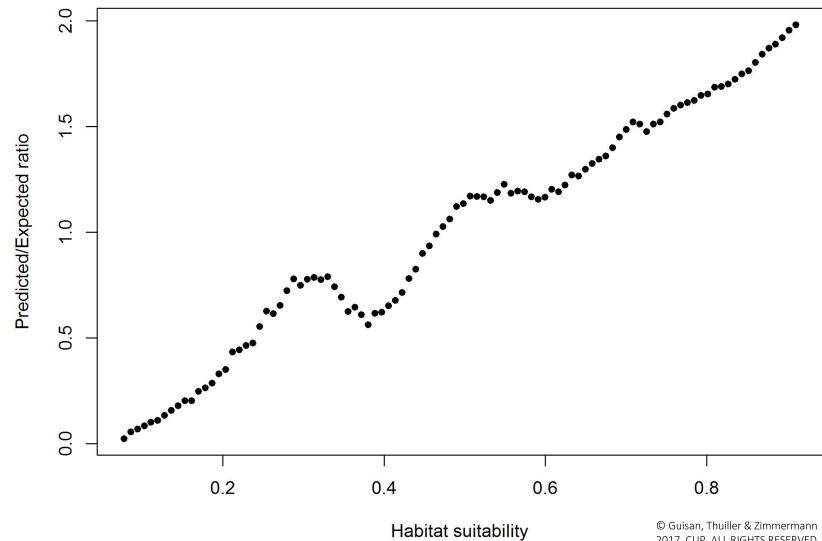
	Observed present	Observed absent	
Predicted present	TP	FP	$\text{Precision} = \text{TP}/(\text{TP}+\text{FP})$
Predicted absent	FN	TN	$\text{Recall} = \text{Sensitivity} = \text{TP}/(\text{TP}+\text{FN})$ $\text{Specificity} = \text{TN}/(\text{FP}+\text{TN})$

AUC-ROC is the area under the curve of the plot of sensitivity (recall) versus 1-specificity across thresholds. Sensitivity is based on all observed presences (left column), while specificity is based on all observed (or inferred) absences (right column). Thus, AUC-ROC incorporates all quadrants of the confusion matrix. AUC-PR is the area under the curve of the plot of precision versus recall (sensitivity) across thresholds. Precision is based on all predicted presences (top row). Thus, AUC-PR does not incorporate the number of true negatives (TN); cells used in calculating AUC-PR are shaded grey and outlined in dashes.

TP, True positive; FP, False positive; FN, False negative; TN, True negative.

# Discrimination metrics - Presence Only

- ▶ Boyce Index
  - ▶ Split model predictions into  $b$  regular bins
  - ▶ Assess the proportion of presences in each bin and compare to the expected proportion if the presences were randomly distributed
  - ▶ **A good model will predict a large proportion of presences in the high value bins**

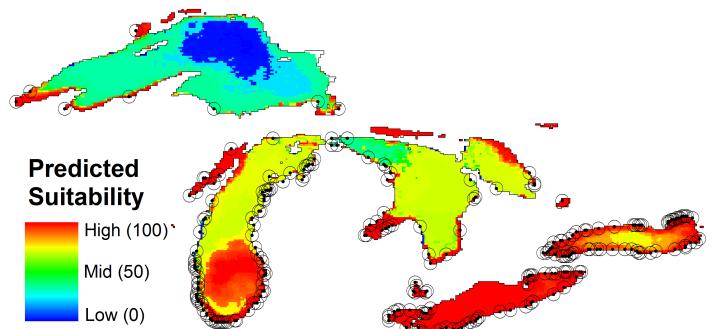


© Guisan, Thuiller & Zimmermann  
2017, CUP. ALL RIGHTS RESERVED

# Discrimination metrics - Presence Only

---

- ▶ Minimum predicted area (MPA)
  - ▶ Proportion of the study area predicted as present using a threshold required to predict as present a user-defined proportion of the test data (e.g., 95%)
  - ▶ Smaller = better



Questions?