Using Publicly Available Data to Model Six Species of Lepidoptera in Kruger National Park

Jennifer Sailor
MSSC 6975
Dr. Nate Lemoine

MARQUETTE UNIVERSITY | BE THE DIFFERENCE.

Outline

- Project Goals
- Data Collection - *Shortened Version*
- Machine Learning Method
- Model's Results
- Interpretation
- Future Exploration & What I learned

Goals

- Create Present Day & Future Spatial Distributions for 6 Species of Lepidoptera in Kruger National Park
  - using an ensemble machine learning approach
- Results that show how climate change will impact the distribution of rare and common species

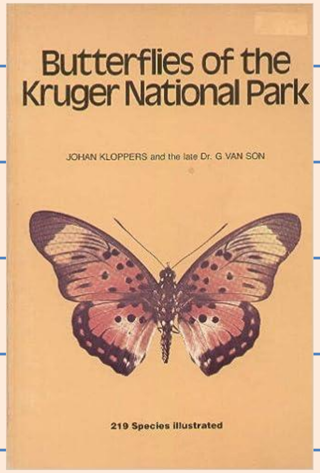
Why Spatial Distributions?

Why Butterflies?

Why Kruger National Park?

Data Collection

Data Set & Sources

	Species	**Latitude & Longitude**	**Climate Data Variables**
0			
1			
⋮		GBIF Global Biodiversity Information Facility	WorldClim
⋮			
⋮			
n			

Data Set

Preprocessing Steps

Removal of Inaccurate Data

1. Null Values
2. Coordinates in Major cities
3. Coordinates in the Ocean
4. Coordinates not in Africa Geographically

Removal of Sample Bias

1. Removing Duplicate Coordinates in Spatial Grid
2. Training Dataset being the true geographical range of species (Africa) not just area of Interest

Generating Pseudo Absent Records

- i.e., uniform distribution

Removal of Variables with strong Correlation

- i.e., multicollinearity testing

Data Set

Training Datasets!

Common

	Species	Climate Data Variables				
0	1 = Present	Temperature Seasonality	Min Temp. Coldest Month	Precipitation Driest Month	Precipitation Warmest Quarter	Precipitation Coldest Quarter
1						
⋮						
n	0 = Absent					

	Species	Climate Data Variables				
0	1 = Present	Temperature Seasonality	Min Temp. Coldest Month	Precipitation Driest Month	Precipitation Warmest Quarter	Precipitation Coldest Quarter
1						
⋮						
n	0 = Absent					

	Species	Climate Data Variables				
0	1 = Present	Temperature Seasonality	Min Temp. Coldest Month	Precipitation Driest Month	Precipitation Warmest Quarter	Precipitation Coldest Quarter
1						
⋮						
n	0 = Absent					

Rare

	Species	Climate Data Variables				
0	1 = Present	Temperature Seasonality	Min Temp. Coldest Month	Precipitation Driest Month	Precipitation Warmest Quarter	Precipitation Coldest Quarter
1						
⋮						
n	0 = Absent					

	Species	Climate Data Variables				
0	1 = Present	Temperature Seasonality	Min Temp. Coldest Month	Precipitation Driest Month	Precipitation Warmest Quarter	Precipitation Coldest Quarter
1						
⋮						
n	0 = Absent					

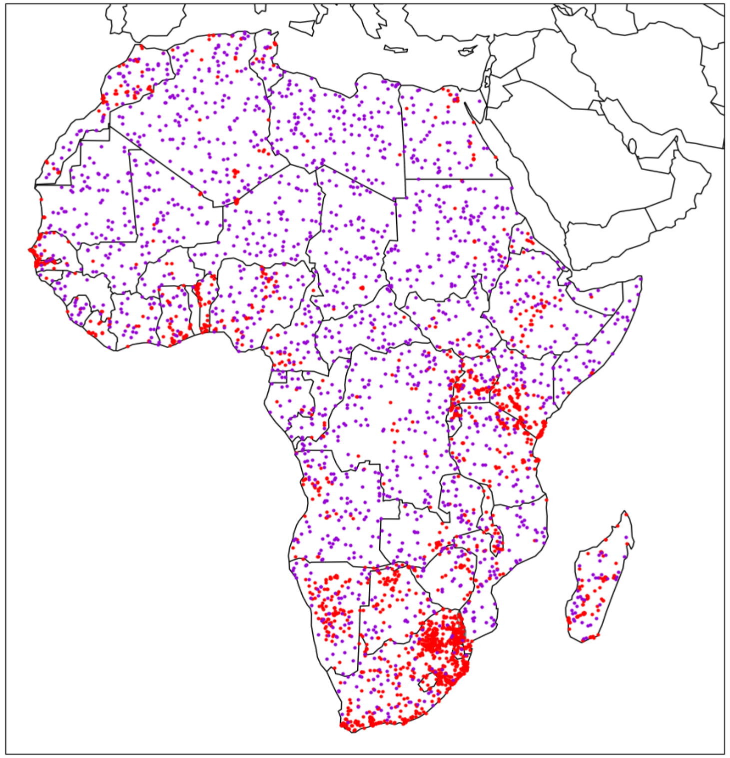
	Species	Climate Data Variables				
0	1 = Present	Temperature Seasonality	Min Temp. Coldest Month	Precipitation Driest Month	Precipitation Warmest Quarter	Precipitation Coldest Quarter
1						
⋮						
n	0 = Absent					

Data Set

Common

	Species	Climate Data Variables				
		Temperature Seasonality	Min Temp. Coldest Month	Precipitation Driest Month	Precipitation Warmest Quarter	Precipitation Coldest Quarter
0	1 = Present					
1						
⋮						
n	0 = Absent					

Danaus Chysippus / Plain Tiger



KEY

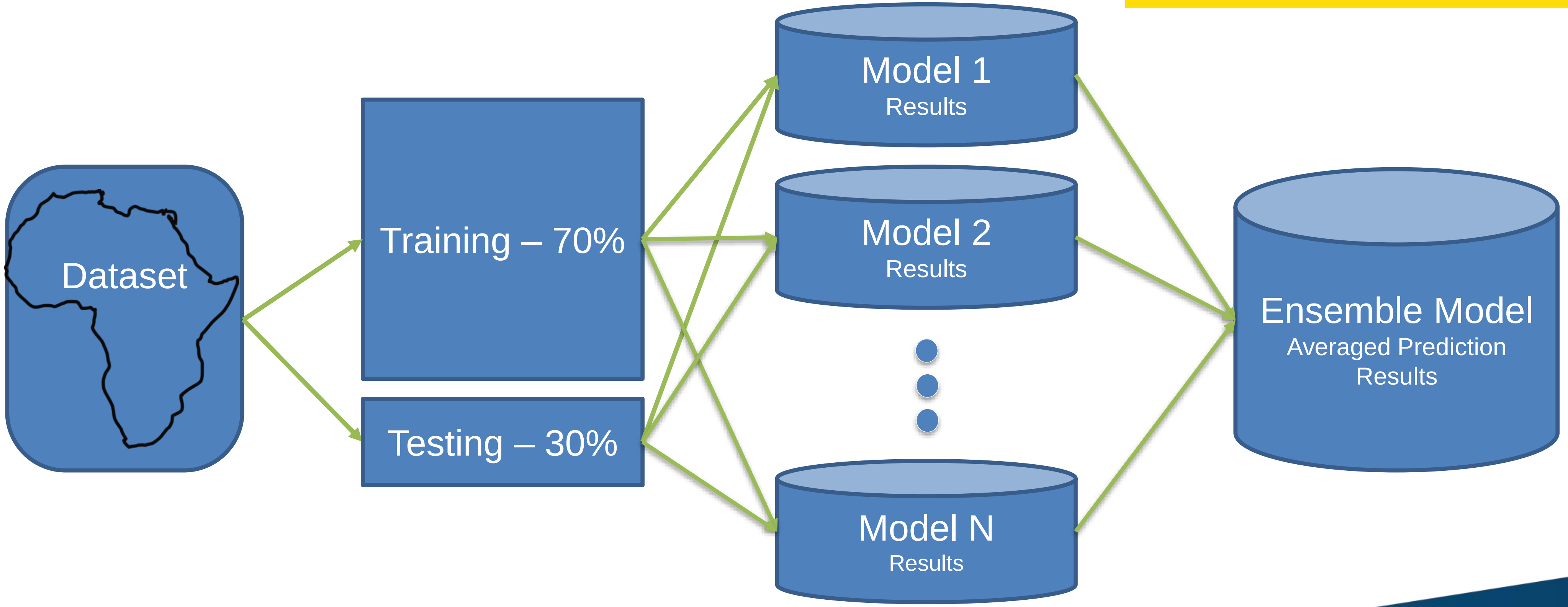
- Present
- Absent

Stacked Ensemble Method

Training & Testing Results

Accounting for Model Uncertainty

ensemble of 9 models



Stacked Ensemble Method

Models of Choice

Model 1
Logistic Regression

Model 4
Decision Tree Clas.

Model 7
Ada Boost Clas.

Model 2
K-Neighbors Clas.

Model 5
Random Forest Clas.

Model 8
Naïve Bayes. Clas.

Model 3
Gaussian Process Clas.

Model 6
Artificial Neural Net.

Model 9
Quadratic Discr. Analysis

Stacked Ensemble Method

Hyperparameter Tuning on AUC-ROC – Non-Default Parameters

Model 1 Logistic Regression	n/a
Model 2 K-Neighbors Clas.	Leaf size = 1, p = 1, n neighbors = 7
Model 3 Gaussian Process Clas.	n/a
Model 4 Decision Tree Clas.	Criterion = ‘entropy’, max depth = 5
Model 5 Random Forest Clas.	max depth = 5
Model 6 Artificial Neural Net.	Max iter = 1000
Model 7 Ada Boost Clas.	Learning rate = 0.1, n estimators = 300
Model 8 Naïve Bayes. Clas.	n/a
Model 9 Quadratic Discr. Analysis	n/a

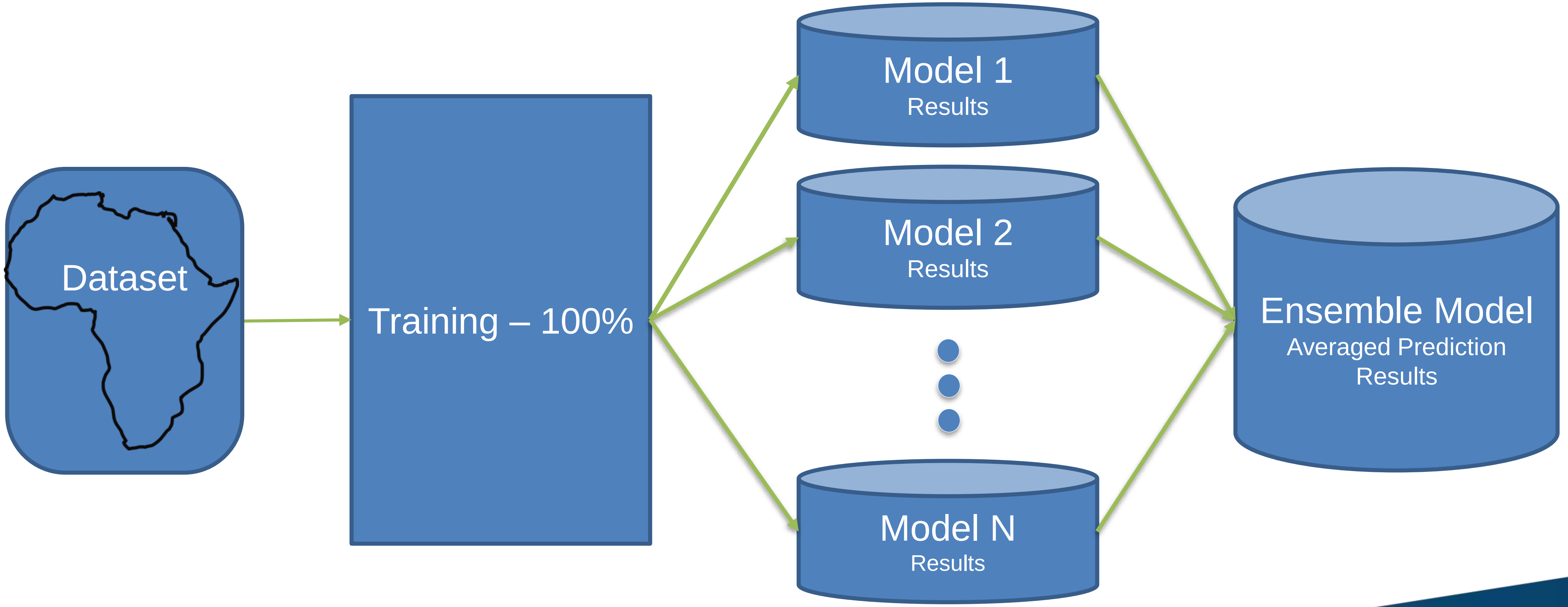
Individual Models

AUC – ROC Estimates						
	Rare			**Common**		
	Mycalesis rhacotis	*Eronia cleodora*	*Cacyreues marshalli*	*Danaus chrysippus*	*Vanessa cardui*	*Belenois aurota*
GLM	0.86	0.85	0.89	0.74	0.83	0.78
KNC	0.94	0.96	0.95	0.86	0.90	0.88
GPC	0.92	0.97	0.95	0.85	0.90	0.86
DTC	0.95	0.90	0.93	0.81	0.88	0.83
RFC	0.94	0.96	0.97	0.88	0.92	0.89
ANN	0.93	0.98	0.95	0.85	0.91	0.86
ABC	0.88	0.97	0.97	0.82	0.89	0.85
NBC	0.88	0.91	0.90	0.77	0.80	0.81
QDA	0.90	0.92	0.92	0.78	0.84	0.81
Ensemble	0.91	0.94	0.94	0.82	0.87	0.84

Abbreviations: GLM, Logistic regression; KNC, K-neighbors Classifier; GPC, Gaussian Process Classifier; DTC, Decision Tree Classifier; RFC, Random Forest Classifier; ANN, Artificial Neural Network; ABC; Ada Boost Classifier; NBC, Naïve Bayesian Classifier; QDA, Quadratic Discriminant Analysis.

Stacked Ensemble Method

100% Training Results



Prediction Datasets (times 6 for each species)

	Species	2023 Climate Data				
		Temperature Seasonality	Min Temp. Coldest Month	Precipitation Driest Month	Precipitation Warmest Quarter	Precipitation Coldest Quarter
0	?					
1						
⋮						
n						

South African
NATIONAL PARKS
2023

	Species	RCP 4.5 Climate Data				
		Temperature Seasonality	Min Temp. Coldest Month	Precipitation Driest Month	Precipitation Warmest Quarter	Precipitation Coldest Quarter
0	?					
1						
⋮						
n						

	Species	RCP 8.5 Climate Data				
		Temperature Seasonality	Min Temp. Coldest Month	Precipitation Driest Month	Precipitation Warmest Quarter	Precipitation Coldest Quarter
0	?					
1						
⋮						
n						

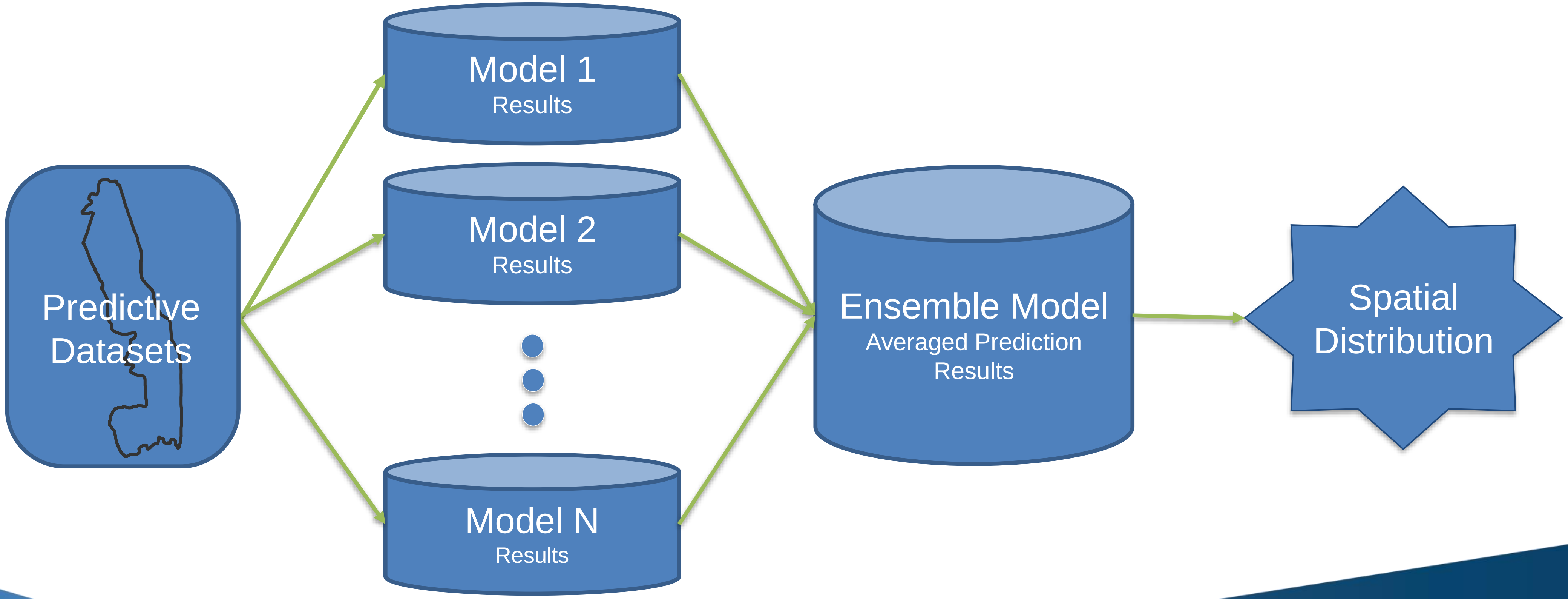
South African
NATIONAL PARKS
2050

Accounting for Climate Uncertainty

ensemble of 4 climate scenarios for each RCP level

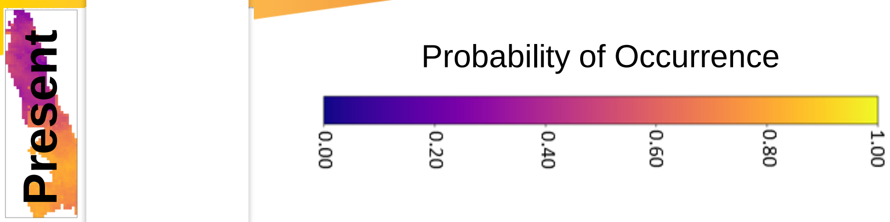
Stacked Ensemble Method

Predictive Results



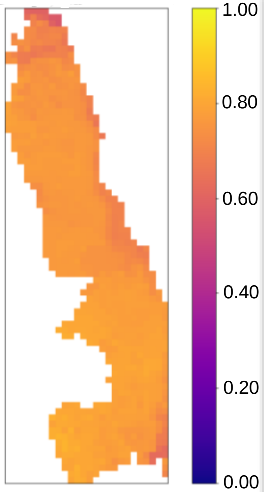
Results

Key

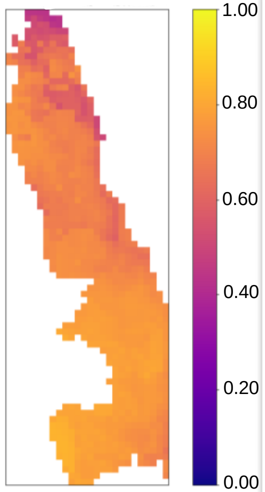


Common

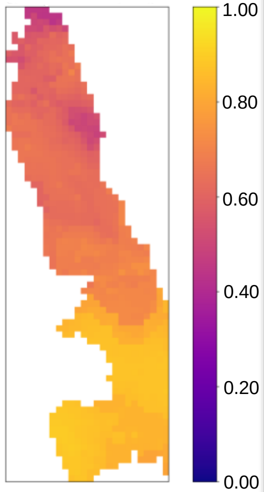
Danaus chrysippus



Vanessa cardui

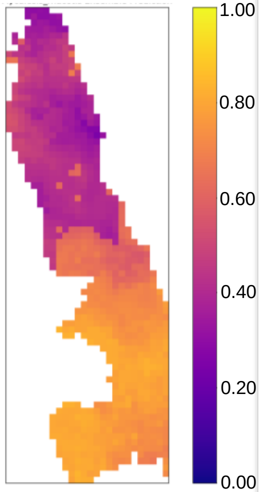


Belenois aurota

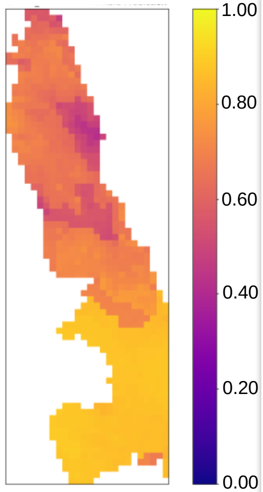


Rare

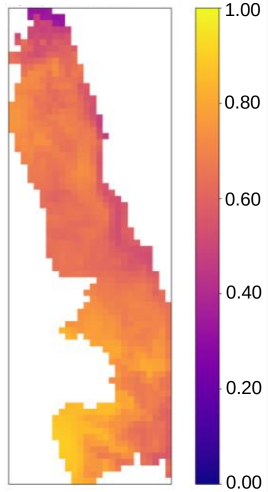
Mycalesis rhacotis



Eronia cleodora

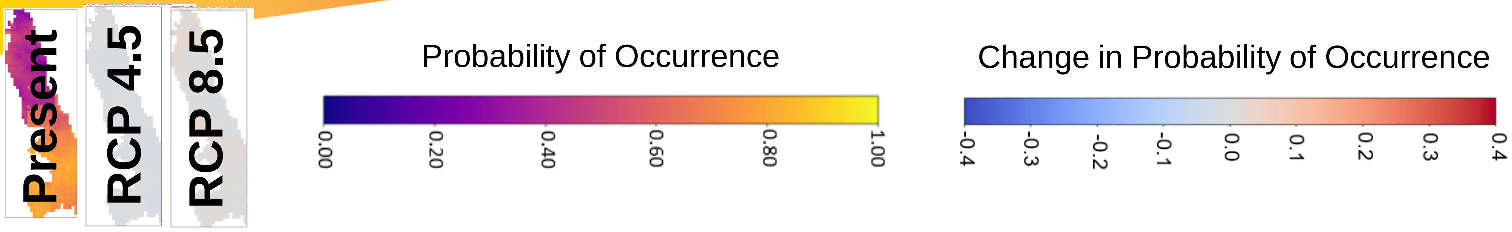


Cacyreus marshalli



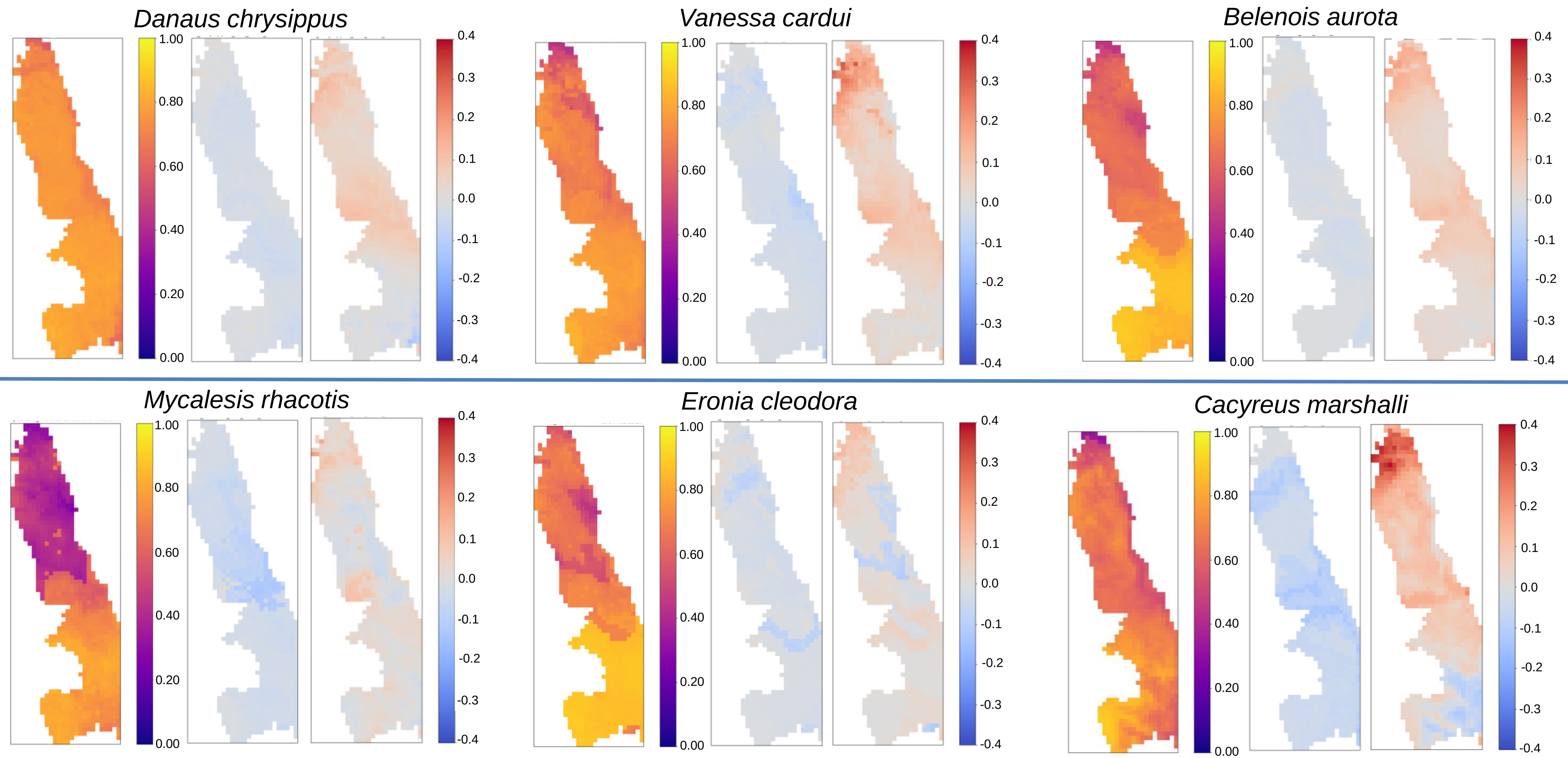
Results

Key

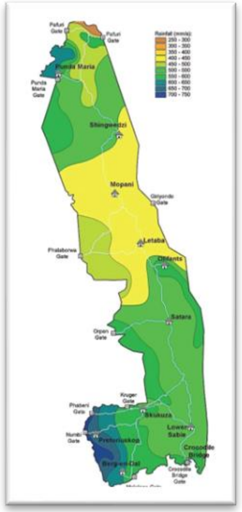


Common

Rare



Factors we're Considering



- Precipitation
- Temperature

Factors we're not Considering

- Plant & Species Distributions!
- Plant & Species Interactions!
- …

Future Exploration

- Expanding to more species to better identify trends
- Adding Additional Variables?

What I Learned

- Expansion/Improvement on my Python coding skills
- GIS!
- Expansion of knowledge in Ecology
- Stacked Ensemble ML Approach

Questions

Using Publicly Available Data to
Model Six Species of
Lepidoptera in Kruger National Park

Thank You to Dr. Nate Lemoine for working with me on this project
&
Thank You to MSSC allowing me to work across disciplines!