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

Jennifer Sailor **August 22, 2023** **MSSC 6975** **Yu / Lemoine**

Abstract

Summer 2023 Practicum design was to use publicly available data to model three rare and three common species of Lepidoptera in Kruger National Park. This goal was completed with the use of two online data sources that lead to the creation of the finalized dataset. Where I was then able to use a stacked ensemble model approach to produce the current and predicted ecological niche models. The results showed unexpected trends and lead to further research development and questions. While the overall summer practicum led to an increase in knowledge and experience.

Introduction

Climate change has and will continue to affect ecosystems and biodiversity. As our planet begins to change it is important to model how species environmental niches will shape future habitat availability. The ecological niche models will aim to assist conservation efforts of rare and common Lepidoptera species. With this in mind, the goal of the Summer Practicum was to use publicly available datasets to model six (three common and three rare) species of Lepidoptera (butterflies). I focused on Kruger National Park (KNP), a 7,500 square mile national park located in eastern South Africa.

The project is a part of the MSSC 6975 Practicum in Statistics and Data Science course that ran from May 22nd to August 19th, 2023. Now that the course has been completed this paper serves as the required summary report that includes an outline of the work completed during the 12 weeks, a description of milestones, and an explanation of how the work contributed. Later in the Fall of 2023, I will give a 15–20-minute presentation over the same criteria to students and faculty of the Mathematical and Statistical Sciences (MSSC) Department at Marquette University. Over the 12 weeks, I was mentored by Dr. Nathan Lemoine an Assistant Professor in the Biological Sciences department. With a focus on Trophic Ecology and Quantitative Ecology, he has published articles on Lepidoptera and machine learning techniques for ecological sciences. He currently contributes to KNP research. I have attended lab meetings for over a semester, started working in his lab in April, and have completed the work required for the practicum virtually from Kansas City. We met at a minimum of once a week via Microsoft Teams for mentorship and project updates and then kept in contact via email for questions and concerns outside of our weekly meetings.

Methods

Data Collection

The dataset used in this project was developed through a combination of three distinct data collection sources. The primary source encompassed environmental data, sourced from the World Climate website. Specifically, the WorldClim2 environmental dataset is an interpolated climate dataset covering the years 1970 to 2000 on a global scale. For this analysis, the focus was on the yearly data with a 2.5 arc-minute resolution. This is a comprehensive set of 19 environmental variables: annual mean temperature, mean diurnal range, isothermality, temperature seasonality, maximum temperature of warmest month, minimum temperature of coldest month, temperature annual range, mean temperature of wettest quarter, mean temperature of driest quarter, mean temperature of warmest quarter, mean temperature of coldest quarter, annual precipitation, precipitation of wettest month, precipitation of driest month, precipitation seasonality / coefficient of variance, precipitation of wettest quarter, precipitation of driest quarter, and precipitation of coldest quarter. To prepare for Machine learning multicollinearity was evaluated. This revealed significant correlations among several variables. Consequently, an iterative process was developed to address this issue. Variables with a correlation threshold greater than five were systematically removed until all variables were below the established threshold. As a result, only 5 of the 19 variables were used for all further analysis. The 5 pivotable variables are temperature seasonality, minimum temperature of coldest month, precipitation of driest month, precipitation of the warmest quarter, and precipitation of coldest quarter.

For the compilation of species occurrence data, a combined approach utilizing both book references and online data repositories was employed. The process initiated by referencing Johan Kloppers 1978 book, “Butterflies of the Kruger National Park,” which served as a key reference to establish the rarity or commonness of each species within KNP. To select the target species of interest, I cross validated Johan Kloppers 1978 book assessment with GBIF (Global Biodiversity Information Facility) occurrence reports for the continent of Africa and KNP. The GBIF data was obtained using an API (Application Programming Interface), which included occurrence records of the six species of interest. The top three highest occurrences in KNP were chosen for the three common species and the first three rare species with a high number of KNP occurrences and high difference in African occurrence in KNP occurrences were chosen for the three rare species. The rare species were chosen in this manor for the goal of seeing if there is a particular biological reason for the rarity in the park and what conversational efforts can be implemented to help increase their abundance. Therefore, the target species of interest are *Mycalesis rhacotis (Bicyclus safitza)*, *Eronia cleodora*, and *Cacyreues marshalli* as rare species and *Danaus chrysippus*, *Vanessa cardui (Cynthia cardui)*, and *Belenois aurota* which are all classified as common. The next phase involved the comprehensive download (which includes species, latitude, and longitude) of all available occurrence records within the African continent using GBIF’s resources. This step was carried out in May of 2023, yielding a dataset encompassing a total of 23,410 georeferenced locations. The individual species counts within this dataset were as follows: *Mycalesis rhacotis*: 2150, *Eronia cleodora*: 1,377, and *Cacyreues marshalli*: 1,586 as rare species and *Danaus chrysippus*: 6,833, *Vanessa cardui*: 5,673, and *Belenois aurota*: 5,791. This meticulous data compilation process is what laid the foundation for the subsequent analyses and insights into the distribution and occurrence patterns of these butterfly species.

Data Cleaning & Filtering

The refinement of GBIF records followed a systematic sequence of steps, aimed to ensure the quality and reliability of the dataset. I cleaned GBIF records in the following order. Initially, records with null values for latitude and longitude were dropped. Subsequently, records with coordinates falling within a narrow range of 0.01 decimal degrees to capital cities and cities with a population exceeding 2 million were excluded. This decision was motivated by the likelihood such entries weren't true occurrences, but instead preserved specimens housed in a museum. The next filtering stage involved dropping records whose coordinates were situated within oceanic areas or not in Africa geographically. This step involved the use of the GeoPandas Python Package Natural Earth Africa shapefile. Lastly, removed any duplicating geographical coordinates resulting in only one coordinate per 2.5 arc-minute spatial grid. This step avoided any redundances ensuring that the finalized dataset essentially has one point attributing to specific climate features. These steps refined the number of present records to a total of 5,794. The species-specific counts within this refined dataset were distributed as follows: *Mycalesis rhacotis*: 485, *Eronia cleodora*: 169, and *Cacyreues marshalli*: 300 as rare species and *Danaus chrysippus*: 2,034, *Vanessa cardui*: 1,504, and *Belenois aurota*: 1,302. Through these detailed steps, the dataset's reliability and relevance was enhanced.

Dataset

The inherent challenge arose from the realization that while the locations of species' presence were documented, a notable absence of information existed regarding where these species had not been observed. In response, the strategy of introducing pseudoabsence data was devised and implemented. To generate pseudoabsence data latitude and longitude points within Africa were iteratively and randomly generated, from a uniform distribution. The aim was to create a matching number of pseudoabsence points for each species, thus forming a 1:1 ratio between present and absent occurrences. The visualization of the distribution of both present and absent locations for each species can be observed in Figure 1, offering a comprehensive view of the spatial dynamics. In this final stage, the datasets were merged into one dataset. The finalized dataset for each species consisted of 6 variables. The rows within these datasets corresponded to instances of species presence or absence, while the accompanying 5 environmental variables were linked to the geographical coordinates of each data point.

Figure 1

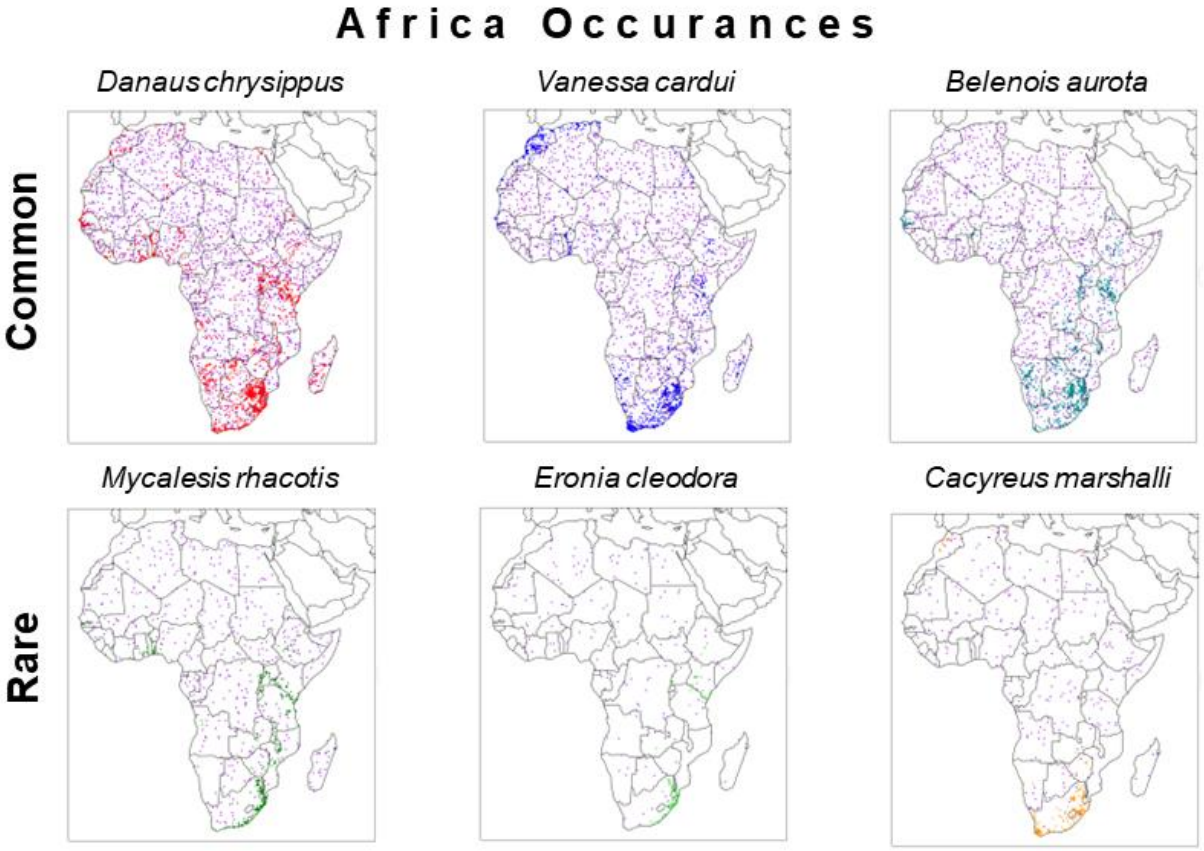

Figure 1: The provided figure showcases the geographical distribution of both present and absent (represented in purple) geolocations across all six unique species. The foundational shapefile outlining the contours of Africa utilized in this visualization was sourced from Cartopy Python package, a widely used tool for geospatial data visualizations and analysis. This presentation encapsulates the spatial patterns and variations across species.

Ecological Niche Models

Ecological Niche Models (ENM) are designed to show the correlation between species and environment. While a straightforward approach might involve the use a singular machine learning model and focus time on increasing the model's accuracy, the inherent uncertainty produced from presence-only sampling and the difficultly of accuracy in climate models compels an exploration of a different approach. To address the limitations and potential biases of individual models, a strategic response was industrialized: the implementation of stacked ensemble approach using 9 diverse machine learning models to construct ENM's. The 9 models are Logistic regression, K-neighbors Classifier, Gaussian Process Classifier, Decision Tree Classifier, Random Forest Classifier, Artificial Neural Network, Ada Boost Classifier, Naïve Bayesian Classifier, and Quadratic Discriminant Analysis. To elevate predictive performance, I implemented hyperparameter tuning on all 9 models based on the best Area Under the Receiver Operating Characteristic Curve (AUC-ROC), thus ensuring the models' capacity to discern between presence and absent points. Hyperparameter tuning across all species and their models showed a consistent trend of parameters and based off those results the most common parameters were chosen. The set of non-default parameters that emerged from this tuning process was applied to each of the 9 models. These specific non-default parameter selections were as follows:

- Logistic regression (n/a)
- K-neighbors Classifier (leaf_size = 1, p = 1, n_neighbors = 7)
- Gaussian Process Classifier (n/a)
- Decision Tree Classifier (criterion = 'entropy', max_depth = 5)
- Random Forest Classifier (max_depth = 15)
- Artificial Neural Network (max_iter = 1000)
- Ada Boost Classifier (learning_rate = 0.1, n_estimators = 300)
- Naïve Bayesian Classifier (n/a)
- Quadratic Discriminant Analysis (n/a).

By combining the predictive strengths of these diverse models through a stacked ensemble approach, the resulting ENMs exhibited enhanced robustness and a capacity to mitigate uncertainties.

Climate Change Projections

There is much uncertainty in climate projections and a lot of it is based on whether humans continue their current habits or if change is implemented. To account for the uncertainty in climate projections I followed Dr. Lemoine's 2021 paper's two approaches.

> First, by predicting ecological niches for unrestricted and moderate representative concentration paths (RCPs) to the year 2050. The intermediate scenario, RCP 4.5, which predicts that CO2 emissions peak in 2040, CH4 emissions stop increasing by 2050, and SO2 concentrations continuously decrease from the present day, results in an increase in world average temperatures of 5 degrees Celsius by 2050. The "worst-case" baseline scenario, RCP 8.5, ignores biological or political feedback as a means of reducing emissions under the current standard. The general circulation model (GCM) forecasts themselves are the final source of uncertainty. The forcings and parameters of each GCM differ, which causes variances in the outputs. As a result, I used four different GMCs to project ENMs into future climates: BCC-CCMS-1-1, CCSM4, IPSL-CM5A-LR, and MIROC5.

For each species 2 meta-ensembles were created for the climate change projections. This was done by finding the one ensemble model (a model averaging the 9 machine learning results) for each GCM thus making a total of 8 (4 for RCP 4.5, 4 for RCP 8.5) then averaging the 4 results for RCP 4.5 to create one meta-ensemble and averaging the 4 results for RCP 8.5 to create one meta-ensemble.

Results

When constructing ENM's the modeling algorithms varied in performance across species but consistently within species (Figure 2 & Figure 3). The AUC-ROC estimates (Figure 2) and accuracy, precision, and recall scores (Figure 3), collectively portray the projected outcomes of the models across all six species, including both individual models and the ensemble model. A noteworthy observation is not one machine learning model consistently outperformed or underperformed across all six species. This is an example of the intricacy of the relationships between models and species characteristics. Overall, these results gave a high degree of confidence in the performance of the ensemble algorithm in generating the ENMs.

Figure 2

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

Figure 2: The provided Figure 2 illustrates the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) outcomes for all 9 models, along with the inclusion of the stacked ensemble model. These visualizations vividly represent the performance of each model in terms of their AUC-ROC scores. The results show diversity with the recorded values spanning from a minimum of 0.82 to a maximum of 0.94 with the ensemble model. The standard deviation for the all-ensemble models was less than 0.05.

Abbreviations: GLB, 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.

Figure 3

Accuracy, Percision, & Recall Estimates																		
	Rare									**Common**								
	Mycalesis rhacotis			*Eronia cleodora*			*Cacyreues marshalli*			*Danaus chrysippus*			*Vanessa cardui*			*Belenois aurota*		
	Acc.	Per.	Rec.	Acc.	Per.	Rec.	Acc.	Per.	Rec.	Acc.	Per.	Rec.	Acc.	Per.	Rec.	Acc.	Per.	Rec.
GLM	0.78	0.76	0.81	0.78	0.73	0.88	0.83	0.80	0.88	0.67	0.67	0.70	0.77	0.76	0.80	0.72	0.71	0.74
KNC	0.87	0.85	0.88	0.86	0.79	0.98	0.89	0.85	0.97	0.79	0.78	0.81	0.84	0.83	0.85	0.80	0.79	0.82
GPC	0.86	0.85	0.86	0.89	0.84	0.96	0.88	0.85	0.95	0.77	0.79	0.75	0.83	0.84	0.81	0.77	0.76	0.82
DTC	0.88	0.88	0.86	0.88	0.84	0.94	0.92	0.89	0.96	0.73	0.70	0.82	0.84	0.86	0.81	0.76	0.73	0.84
RFC	0.89	0.89	0.88	0.87	0.83	0.92	0.92	0.88	0.97	0.80	0.78	0.84	0.86	0.86	0.87	0.81	0.80	0.85
ANN	0.86	0.85	0.86	0.91	0.90	0.92	0.88	0.84	0.96	0.77	0.78	0.76	0.85	0.85	0.84	0.77	0.76	0.80
ABC	0.85	0.85	0.83	0.90	0.85	0.96	0.88	0.85	0.94	0.73	0.72	0.75	0.83	0.83	0.84	0.76	0.75	0.80
NBC	0.81	0.76	0.88	0.77	0.72	0.88	0.85	0.81	0.94	0.71	0.68	0.81	0.72	0.68	0.83	0.74	0.70	0.83
QDA	0.82	0.77	0.88	0.78	0.71	0.92	0.87	0.83	0.94	0.73	0.71	0.78	0.77	0.74	0.82	0.73	0.68	0.86
Ensemble	0.85	0.93	0.86	0.85	0.80	0.92	0.88	0.94	0.95	0.74	0.73	0.78	0.81	0.81	0.83	0.76	0.74	0.82

Figure 2: The depicted Figure 2 presents overview of the performance metrics encompassing accuracy, precision, and recall estimates for all comprehensive ensemble of nine distinct machine learning models, in addition to the stacked ensemble model. Through these graphical representations, a comprehensive assessment of the model's efficacy is apparent.

Abbreviations: Acc., Accuracy; Per., Percision; Rec., Recall; GLB, 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.

The ensemble generated ENMs were able to establish the habitability patterns and anticipated behaviors in the current climate conditions (Figure 4). Withing the "current" plots of Figure 4, a trend emerges: common species tend to exhibit a greater value with higher probability of presence compared to the rare species, aligning with the categorization detailed in Klopper's book.

Interestingly, in the context of the RCP 4.5 climate scenario, a contrasting pattern to our hypothesis unfolds. It was expected that the change in probability of occurrence would have higher values with climate scenario RCP 4.5 when compared to climate scenario RCP 8.5. However, the RCP 4.5 climate scenario plots indicate a large quantity of instances where a decrease in the probability of occurrence is indicated (depicted in blue). Conversely, under the RCP 8.5 scenario, the plots display either a mix of both increased or decreased instances or a larger quantity of instances where an increase in probability of occurrences is indicated (depicted in red). These unexpected outcomes lead to the question of what the percent change in probability from its current to future climate scenarios which Figure 5 answers. The results of Figure 4 highlight the complexity between species distributions and climate scenarios.

Figure 5 addresses the unexpected patterns observed in the previous analyses, raising relevant questions regarding the change in probability as species transition from present/current to future climate scenarios. The results show a pattern of negative percent changes, indicating habitat loss, is notable within the context of the RCP 4.5 climate scenario. In contrast, the RCP 8.5 climate scenario mainly showcases positive percent change, implying an expansion or increase in habitat suitability. The insights require future research on potential reasonings behind these results.

Figure 4

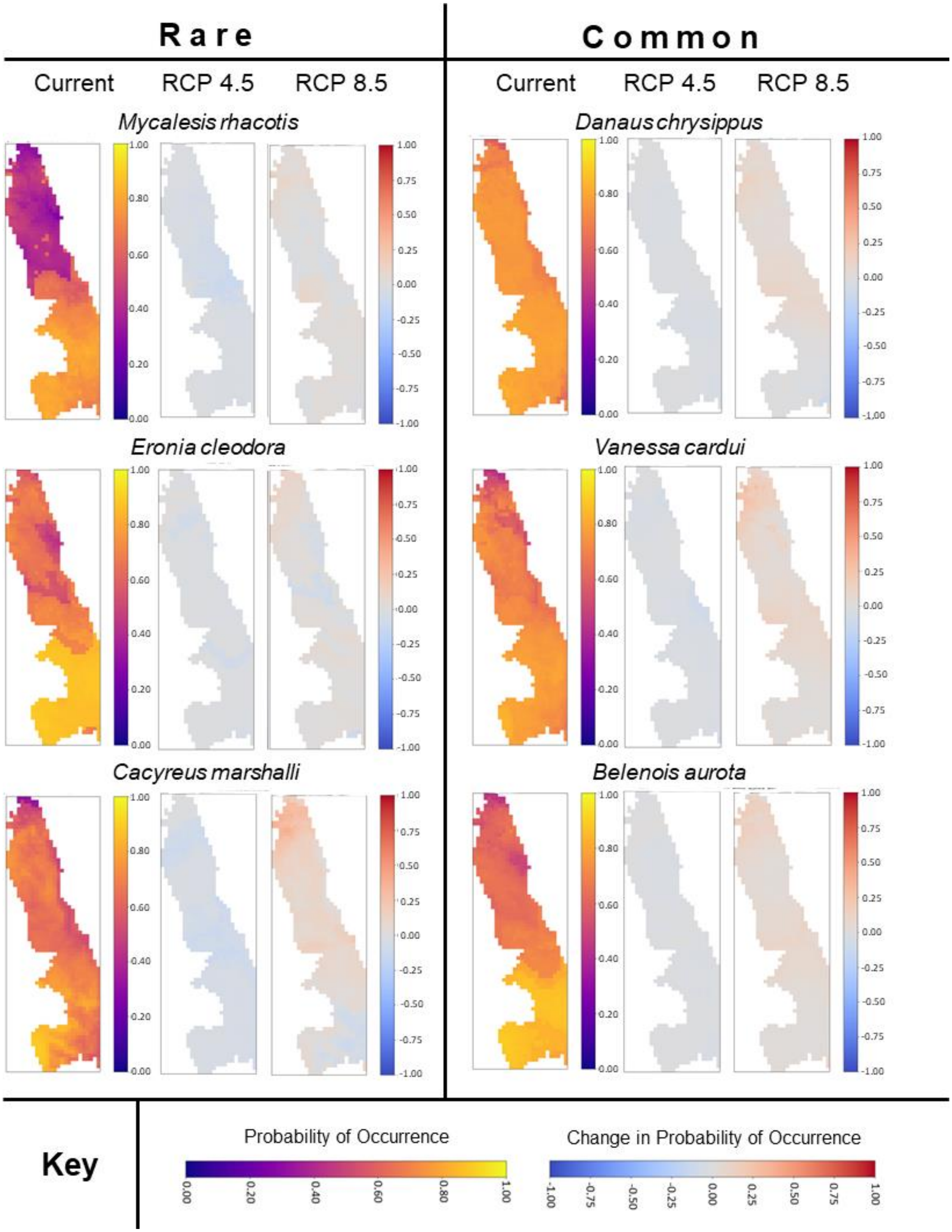
Rare
Common
Current
RCP 4.5
RCP 8.5
Current
RCP 4.5
RCP 8.5
Mycalesis rhacotis
Danaus chrysippus
Eronia cleodora
Vanessa cardui
Cacyreus marshalli
Belenois aurota
1.00
0.80
0.60
0.40
0.20
0.00
1.00
0.75
0.50
0.25
0.00
-0.25
-0.50
-0.75
-1.00
Key
Probability of Occurrence
0.00
0.20
0.40
0.60
0.80
1.00
Change in Probability of Occurrence
-1.00
-0.75
-0.50
-0.25
0.00
0.25
0.50
0.75
1.00

Figure 4: The provided Figure 4 illustrates the ENMs' of Kruger National Park (KNP) probability of occurrence in Current day model and the change in probability of occurrence between future (RCP 4.5 or 8.5) and current data.

Figure 5

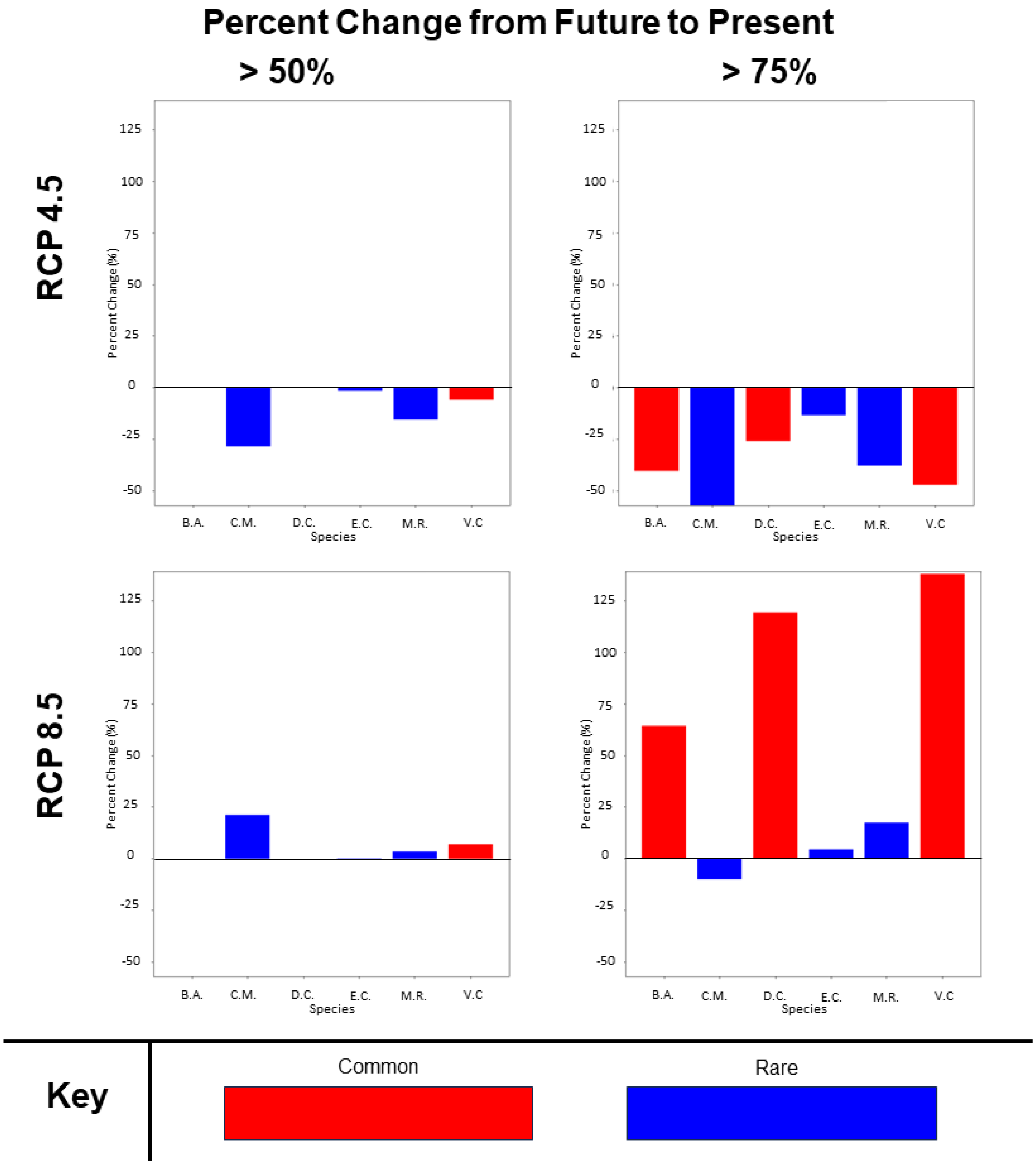

Figure 5: The depicted Figure 5 presents the percent change in probability from Future to Present. This was calculated by finding the number of cells from present and future datasets with probability above the threshold (50%

or 75%) and finding the percent change between these values for each species. Through these graphical representations, an assessment of the change in species' ENM is apparent.

Abbreviations: B.A., Belenois aurota; C.M., Cacyreues marshalli; D.C., Danaus chrysippus; E.C., Eronia cleodora; M.R., Mycalesis rhacotis; V.C., Vanessa cardui.

Discussion

Reflecting on the past 12 weeks, I not only want to reflect on the project itself but also what I learned during this course. During the course I got to practice and expand my python coding skills. A lot of which came from learning how to plot and work with GIS data. I also had to obtain the GBIF data via an API which I was familiar with but had never called data via an API prior. Lastly, using the ensemble model approach was new. I have used most of the Machine Learning models before in other projects or classes but never had heard or used the ensemble approach and it was a great learning experience. And to highlight one last thing it was a great experience / challenge to really dive into documentation on my own and really see how capable I was to be able to sit down, research, and test until the code worked as expected.

In conclusion, the 12-week practicum aimed to develop species distributions and apply models to predict changes in habitat suitability on six species of butterflies. By using statistical approaches, I gained information on how the species will be affected by the changing environmental conditions. Despite the need to do further research to identify trends and reasoning, the overall goals were accomplished. I look forward to continuing working with Dr. Lemoine throughout the semester with a focus of expanding the quantity of species. Additionally, I am excited to present to the MSSC faculty and students an example of how Statistical and Data Science methodologies can be applied to ecological issues.