8-2020

The Nexus of Biogeography and GIScience: Utilizing Emerging Big Data Sources and Multiscale Analysis for Species Distribution Models

Adam Alsamadisi
University of Tennessee, aalsamad@vols.utk.edu

To the Graduate Council:

I am submitting herewith a dissertation written by Adam Alsamadisi entitled "The Nexus of Biogeography and GIScience: Utilizing Emerging Big Data Sources and Multiscale Analysis for Species Distribution Models." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Geography.

Liem Tran, Major Professor

We have read this dissertation and recommend its acceptance:

Monica Papes, Sally Horn, Qiusheng Wu

Accepted for the Council:

Dixie L. Thompson

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

The Nexus of Biogeography and GIScience: Using Emerging Big Data Sources and Multiscale Analysis for Species Distribution Models

A Dissertation Presented for the
Doctor of Philosophy
Degree
The University of Tennessee, Knoxville

Adam Guy Alsamadisi
August 2020

DEDICATION

This dissertation is dedicated to my grandparents: Guy D'Urso, who taught me how to use a computer, and his wife Anne D'Urso, who taught me what is important.

ACKNOWLEDGEMENTS

I have so much gratitude to the many people who contributed, in both obvious and subtle ways, through my Ph.D. process. Among the least subtle is my father, Dr. Morsy Alsamadisi, who came from Egypt and built a veterinary hospital that helped foster my love for animals. My dad pushed me to move forward through my education with his words and unsolicited advice, but much more impactful was his own dedication to his craft and lifelong education that taught me the merit of dedicating yourself wholly to a field.

My mother, Maria Alsamadisi, supported me in a variety of important ways: from fostering an academic environment through my childhood full of museums and educational programming to accompanying me on college visits. I appreciate my mom for also passing along her love and skill of teaching, which she admirably fostered in special education classrooms. Teaching has been a fuel for me (both abstractly and financially) to finish this Ph.D., and to that end I am so grateful to have inherited that skill set.

My brother, Noah Alsamadisi, and my best friend, Joey Kowalsky, continue to be rocks I can rely on, and I am especially thankful for their camaraderie as an escape from academics. I further appreciate my friendship with my brothers-in-law, particularly Wesley and James Upham, and am thankful for all of my parents-in-laws as well for their familial support particularly in the years when they were just across the mountains.

I am very appreciative to my advisor, Dr. Liem Tran. In the final phase of my program, circumstances forced me to relocate to New York City. While I am so thankful to Dr. Tran for his scholastic support and deft guidance during this unique period, I would be remiss if I did not consider Dr. Tran's impact on me through my entire graduate education. For all the pushing back on good ideas, trick questions, early morning meetings, sporadically highlighted documents, sketches on printer paper, and (so many) "dad jokes" that it hurt, Thank you. I look forward to a career in which I would be fortunate enough to collaborate with you further.

Professors I learned from during my first semesters of my undergraduate program at Rhodes College, especially Dr. Sarah Boyle and the late Dr. Rosanna Cappellato, helped form the foundation of my academic interests in environmental science and GIS. Specifically, Dr. Boyle's role in developing both the curriculum for the college's GIS program and her role facilitating the student worker program at the GIS lab gave me time and exposure to better understand GIS,

which I undoubtedly credit as core to developing my research skill set. I am also grateful to Dr. Stephen Ceccoli, Dr. Shadrack Nasong'o, and Dr. Elizabeth Thomas for their early academic guidance.

Moving to NYC midway through my program allowed for an opportunity to explore a career I hadn't anticipated: teaching technology classes at a middle school. I felt well-equipped after helping my Little Brother, Brandon Forget, who I thank for providing kinship during my time in Knoxville, and for allowing me to help navigate his own scholastic path. At The Buckley School, I was introduced to colleagues including Dr. Julie King, Greg O'Melia, Willie Dominguez, Edwin Gonzalez, and many others who all played a special role in a new landscape as I finished my dissertation. Finally, a special thanks to the "Buckley Boys" whose contagious energy, creativity, and curiosity helped exercise my own mind (and sometimes patience) every day.

My inclusion in two pottery communities, Mighty Mud in Knoxville and Mud Matters in Manhattan, was central to my success. Seeing the commitment of others to building inspirations, making original pieces, and supporting me as I approached my own projects, taught me invaluable lessons and gave me energy needed to stay dedicated. I'm thankful to all the muddy people I came across, including Barron Hall, Peggy Clarke, Julie Hadley, Diane Waller, Linda Eaton, and David Hollingsworth and for their leadership in these spaces.

Through this journey, many fellow graduate students helped support me in important ways as they completed their own journeys. First and foremost are my academic siblings—who know this ride all too well—including Matthew Miller, Bridgette Fritz, Rachel Craig, and Drs. Sandhya Nepal and Njoroge Gathongo. Those who I met earlier in my graduate career, including Lauren Stachowiak, Neil Connor, and Erik Johanson helped me realize the path and grit required, and were generous in their support for a newcomer to the field. Those who I met later in the program, including Helen Rosko, Dr. Larry Lu, Dr. Jeremy Auerbach, Dr. Pranab Chowdary, Dr. Emma Walcott, Brooke Rose, Kyle Landolt, and Evan Norton, provided a wonderful collegiality and contributed to group learning experiences that helped me grow in important personal and scholastic ways.

Faculty members at the University of Tennessee provided support and instruction that helped scaffold this dissertation. Notably, my original committee members including Dr. Monica Papes, Dr. Sally Horn, and Dr. Yingjie Hu, were incredibly helpful while exploring questions of niche

modeling, biogeography, and volunteered geographic information, and supportive through my comprehensive exams and proposal processes. I appreciate Dr. Qiusheng Wu for stepping in later in the process. I am also thankful for other members at the Department of Geography at UT for their instruction and support including Dr. Nagle, Dr. Kim, Dr. Kalafsky, Dr. Foresta, Dr. Alderman, Dr. Shaw, and Dr. Stewart.

Through my day-to-day, I've been lucky to have the companionship of my two dogs, Simon and Julio, who make me wake up every morning with the mundane purpose of cleaning up after them, which really keeps me grounded. I've been even more fortunate—dare I say the luckiest geographer in the world—to have the partnership of my wife, Cicely Upham, who keeps me on my toes all the time and exercises my creative mind with her wits (which are infinitely stronger than mine). She continues to be a pillar of encouragement who inspires me to do my best, and I thank her for her support and belief in me.

ABSTRACT

Species Distribution Models (SDMs) are important tools for biological conservation and wildlife management as they detail the distributions of biota across landscapes. In this dissertation I explored two emerging big data sources that can be used to enhance SDMs, lidar and Volunteered Geographic Information (VGI). Lidar data can be used in ecological models as explanatory variables that provide information about 3D attributes of space (i.e., structural ecology), and observation data from VGI projects (like eBird) can help inform models about species presence across spatial and temporal scales. In my first research study, I employ a multiscale analysis to address the challenges associated with developing SDMs with high-resolution data from lidar. I present an approach, SBBS, in which the output of SDMs developed with variables that had a spatial resolution of 30-m were used to improve SDMs developed with variables that had a 10-m resolution. This approach produced better models than both a model developed with the default Maxent background sampling area, and a model developed using the conventional approach of resampling environmental data to a common resolution. In my second study I focused on model thresholds to explore the differences between an SDM developed with data from citizen scientists through eBird and one developed with data from wildlife professionals. Results corroborated past research that found SDMs developed with citizen science favor anthropogenic landscapes, but also found factors related to elevation and habitat fragmentation contributed to the mismatch between these models. In my third study, I used inferences from an SDM developed with a scientific presence dataset and the statistical concept of influence to evaluate, categorize, and filter eBird points. Through my methods, I was able to isolate species presence locations from eBird that best matched the environmental characteristics of observation locations from the scientific dataset and analyze attributes of points that differed from that profile. This research contributes to knowledge at the nexus of Biogeography and GIScience, as spatial data methods are used to better understand species distributions, while knowledge about ecological relationships across space serves as a basis to better understand these two emerging spatial data sources.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1 INTRODUCTION

1.1 Scientific and Disciplinary Context of this Research

Generally, scholars recognize two subdomains within the discipline of Geography, which include: (1) Physical Geography, the study of the spatial attributes of natural features and processes across Earth's four physical systems (the biosphere, hydrosphere, geosphere, and atmosphere) and (2) Human Geography, which concerns cultural, socioeconomic, political, and anthropological attributes across space and place (Douglas, 2002; Bonnett, 2008; Warf, 2010; Hanks and Stadler, 2011). However, other research focuses have also been recognized as additional subdomains, including, for example, Nature and Society (Warf, 2010), which explores the intersection of Human and Physical Geography. Another emerging subfield often considered within the discipline is Geographic Information Science and Technology (GIS & T), which involves computational programs (i.e., Google Earth or the popular ArcGIS) and technical tools and data (i.e., lidar, satellite imagery, census data) used to measure and represent geographic features and phenomenon (Warf, 2010; Koeppe, 2012). Because GIS & T also includes the approaches employed to study geographic data in order to mathematically measure and statistically analyze phenomena in Human and Physical Geography, GIS & T is sometimes considered a set of tools within these two subdisciplines, instead of a standalone field (Openshaw, 1991; Wright et al., 1997).

Within the discipline of Geography, I posit my research is a fusion of GIS &T and Biogeography, the subfield of Physical Geography which concerns studies of the biosphere including the spatial distributions of species and ecosystems (Millington et al., 2011; Gavin, 2012). Within GIS & T, this research borrows from two main subject areas that are detailed in the GIS & T Body of Knowledge, a reference document that is the most comprehensive outline of concepts and methods relevant to geospatial science (First edition citation University Consortium for Geographic Information Science 2006, but the work is now updated at https://gistbok.ucgis.org/). The first of these is titled *Citizen Science with GIS & T* which is under the umbrella of *GIS & T and Society*, a subject area that focuses on relationships and engagement between GIScience, spatial technology, public entities, private enterprises, and individual citizens (Rickles et al., 2017). According to the Body of Knowledge, citizen science involves the contributions of non-professionals (i.e., volunteers) in scientific projects and is the basis of

Volunteered Geographic Information (VGI) which, according to Goodchild (2007) is "the harnessing of tools to create, assemble, and disseminate geographic data provided voluntarily by individuals" (Rickles et al., 2017). More specifically, this research fits and contributes to issues related to data quality with VGI, as I investigate the quality and quantify the additional information of VGI from the popular citizen science project eBird relative to a dataset provided by a scientific research organization.

The second area that this research focuses on is titled *Resolution*, which is part of the subject *Foundational Concepts* in the GIS & T Body of Knowledge, presented by Lam (2019). Resolution is most relevant to raster data, pixel-based spatial data, including satellite imagery, used in GIS that is inherently related to the geographic concept of scale. In the context of geographic data, scale refers to: (1) the spatial extent of data coverage (i.e., the Eastern United States or the boundaries of a specific county) and (2) the resolution of spatial data, the size or the spatial coverage of each individual raster cell (Lam, 2019). While data resolution could be easily dismissed as simply a technical attribute of geographic data, resolution is fundamentally related to the ability of spatial data to represent geographic phenomena and features at the scales at which they operate (Lam, 2019). Pixels that comprise raster data are the smallest distinguishable unit of measurement of data (Lam, 2019). While rasters with smaller pixel sizes at local extents are more suitable to analyze some phenomena or features with finer operational scales, like changing successional stages of individual forest stands, raster data at wide extents with larger pixel sizes are more suitable to understand other phenomena at broader operational scales, like shifts in animal migration patterns or seasonal forest dynamics of entire biomes (Hebblewhite, 2008; Goodchild, 2011; Kamal et al., 2015). Grounded in a methodology involving the resolution of environmental data, I sought to explore the concept of scale in developing species distribution models (SDMs), which are computer generated and statistically grounded maps that detail the expected geographic spread of species.

Broadly, Biogeography is the study of spatial distributions of species and ecosystems. However there are a diversity of subfields and specific focuses within the discipline (Millington et al. 2011; Gavin 2012). Millington et al. (2011) provided a comprehensive description of these research focuses. For example, Historical Biogeography is a subfield that emerged from Paleobiology and focuses on changing distributions of organisms in areas over long temporal

spans, and Dispersal Biogeography focuses on the range expansion of species and their barriers, both physical (i.e., roads/rivers) and ecological (i.e., competitors or predator species; Millington et al., 2011).

This dissertation research primarily belongs to two subfields in Biogeography. The first, referred to as both 'Analytical Biogeography' and 'Spatial Biogeography,' focuses on spatial data methods to explore diversity of ecosystems to better understand the relationships between environmental characteristics of place and the distribution of species (Millington et al., 2011). The use of GIS by biogeographers, in conjunction with methods in spatial statistics, has allowed for the analysis of geographic data focused on monitoring and modeling the distributions of species, ecological relationships, and ecosystems across spatial and temporal scales (Elith and Leathwick, 2009; Kreft and Jetz, 2010; Millington et al., 2011). In this research, I employ a number of GIS methods and analyses of spatial data from emerging geographic technologies to understand the distributions of two avian species, and the ways these distributions are affected by particular environmental characteristics.

The second theme of Biogeography that this dissertation research focuses on is Human and Environment Interactions. According to Millington et al. (2011), a growing subset of work in Biogeography is focused on relationships between nature and humans at varying scales (i.e., ranging from individual impacts on nature to the effects of societies on broader ecosystems). This type of work has been employed by academics in disciplines ranging from Political Science to Ecology to History, and often has a focus on conservation. Topics explored in this research focus include the way urban planning might result in habitat and biodiversity loss, or they ways animals and crops have become domesticated across areas through time (Millington et al., 2011). In this dissertation, I explore Human and Environment Interactions in three ways. First, the distribution of the two avian species studied in this research are affected by human threats to their habitats and ecosystems in southern Florida. The red-cockaded woodpecker has specialist niche requirements, and human expansion and urban sprawl has led to a loss and fragmentation of the longleaf pine ecosystem in which the species is endemic (Smart et al., 2012). Crested caracaras in Florida are part of a relict population left after the loss of prairie landscape in the state, and has become adapted to human-dominated, agricultural landscapes (Morrison, 2001).

Second, in one of my studies, I explore differences between two SDMs through a linear regression model, with landscape fragmentation metrics of natural habitat as independent variables. As urban and agricultural development are drivers of landscape and habitat fragmentation in my study area (Murrow et al., 2013), this study additionally borrows from the Human and Environment Interactions theme in Biogeography. Finally, the basis of the second and third chapters involves citizen science data, which touches on noteworthy subfields in the nexus of Social Science and Biogeography. For instance, the contributions and engagement of citizens in conservation efforts have been described as the democratization and decentralization of environmental management (Conrad and Hilchey, 2011). Moreover, citizen science data collection is an important type of engagement between recreation and enjoyment of the natural environment (Sullivan et al., 2009).

In the following research chapters, I employ species distribution models and analytical tools in GIS to explore questions related to both phenomena in Biogeography and characteristics of spatial data. While I use GIS and spatial modeling methods to better understand the utility of fine-resolution environmental raster data and different types of species location data to answer questions about species' distributions, I also used information about biogeographical relationships and ecological theories as a basis to better understand spatial data from two emerging data sources (lidar and Volunteered Geographic Information). Thus, the focuses and methods of these studies are rooted in a fusion of Biogeography and GIS & T, as spatial modeling approaches were used to better understand phenomena in the domain of Biogeography, but also used with biogeographical theory to shed light on underlying properties of spatial data. This dissertation approaches two primary research questions: (1) **How can spatial data with different resolutions be used to improve species distribution models? and** (2) **Are wildlife observation data from citizen scientists valid, and what additional information can they contribute to species distribution models?**

1.2 History of Species Distribution Modeling

Some of the earliest research in Geography focused on the spatial distribution of organisms and species' responses to physical and environmental attributes of place (Elith and Leathwick, 2009; Kreft and Jetz, 2010). In 1793 one of the first recognized biogeographers, Alexander von

Humboldt, explained in his *Flora Fribergensis specimen*: “Plant geography traces the connections and relations by which all plant are bound together among themselves, designates in what lands they are found, in what atmospheric conditions they live, and tells of the destruction of rocks and stones by what primitive forms of the most powerful algae, by what roots of trees, and describes the surface of the earth in which humus is prepared” (Humboldt, 1793; translated in Nicolson, 1987). Based on a qualitative understanding of geographic features, the earliest recognized biogeographers (see, as cited in Kreft and Jetz [2010], the early works of von Humboldt [1806], Buffon [1761], and de Candolle [1855]) drew maps and summarized data from their expeditions (Figure 1.1) to understand areas with similar plant assemblages, and then connected them to the influence of climate and animals (Kreft and Jetz, 2010). von Humboldt in particular was a frontiersman in the tradition of correlating observations of species to environmental features of place; he later developed the isoline technique in cartography that allowed for mapping the distribution of dominant vegetation assemblages relative to environmental characteristics across geography, and is also employed to map climatic attributes like temperature and air pressure (Nicolson, 1904).

Figure 1.1 Alexander von Humboldt's 1807 *Geographie der Pflanzen in den Tropen-Ländern* (Geography of Plants in Tropical Land, public domain reprint), is one of the earliest species distribution models and details changing plant distributions in the Andes across elevation range

More than a century later, an impactful study by Joseph Grinnell (1904) detailed the environmental characteristics of areas inhabited by chestnut-backed chickadee populations (Grinnell, 1904). Though still qualitative in methodology, the study consecrated the term *Grinnellian niche*, which refers to the subset of environmental (in this case, specifically abiotic) conditions, at broad scales, in which the species is present (Soberón and Peterson, 2005). In his 1927 book *Animal Ecology*, Charles Elton introduced a different definition of niche, which has since become known as the *Eltonian niche,* that focused on biotic and interacting components of habitat at the local scale: "the animal's place in its community" (Elton, 1927; Soberón, 2007). Fifty years after these studies helped define the conceptual aspects of species distributions, quantitative methods, fueled by advancements in computational and statistical abilities, began to emerge to measure the relationship between species' distributions and physical geography (Elith and Leathwick, 2009; Kreft and Jetz, 2010). As detailed in Elith and Leathwick (2009), these methods followed the development of statistical models, with environmental envelope models progressing to regression-based models and then evolving to machine learning models like Maxent. While the process of mapping species' distributions now benefits from advances in statistics and computational technologies, the goals are still the same as von Humboldt's and questions of Grinnellian and Eltonian niche remain the conceptual underpinnings for this type of analysis.

Today, SDMs are regularly employed to support the most pressing conservation and environmental challenges faced globally. In their review of standards for SDMs, Araújo et al. (2019) found that published papers employing SDMs in conservation settings did so to: (1) model the suitability of areas after changes in climate or land use, (2) direct surveys for species in areas predicted suitable but without species observations, (3) evaluate the ability of areas to support populations of species, (4) identify and prioritize areas based on their conservation importance and potential, (5) evaluate areas that are susceptible to biological invasions or the transmission of disease, (6) suggest areas suitable for populations being translocated, and (7) predict areas suitable for restoration or predict the effect of restoring specific places. Another subset of papers that use SDMs does not employ them for biodiversity conservation, but rather to further explore their scientific merit through the development of new diagnostics to assess SDM quality or methods to improve SDM predictions, and alongside connections to ecological theory

as a basis for both developing and interpreting SDMs (Araújo et al., 2019). The research presented in this dissertation fits into this category, as I employ SDMs in coordination with spatial data methodologies to answer questions about the role and utility of two emerging spatial sources relevant to ecology, lidar for environmental input variables, and eBird for species presence data.

1.3 Integral Components of Species Distribution Models (SDMs)

While Species Distribution Models are employed for important conservation and management goals and are increasingly relied upon to understand spatial ecology in a changing Anthropocene, researchers have also been engaged in theoretical and semantic discussions about them (Elith and Leathwick, 2009). This dissertation directly addresses some of the challenges and disagreements within the field and integrates them into the study design, including issues related to presence and absence data, model thresholds, and scale (see discussions in Elith and Leathwick, 2009, Li et al., 2011, Liu et al., 2013). However, the most broad and basic of these questions, which is the basis of many of these issues, remains: What are SDMs really modeling? The simplest definition, I suggest, is that SDMs are maps that detail the geographic attributes of species presence. More complex definitions require discussions primarily regarding two factors: (1) the statistical, mathematical, or logical basis to estimate species' distributions and (2) the input variables and data used to develop these models.

The first factor involves the various types of methods and models used to estimate spatial characteristics of species presence. Examples in Drew et al. (2011) show that some SDMs are used to estimate the number of individuals within a species across habitat patches after factoring population dynamics and resource availability, while other SDMs instead represent the environmental similarity of places within a study areas to locations where the species has occurred. Some SDMs are based on uncertain or unclear definitions of the environmental preferences of species, while others, such as agent-based models, have temporal components and are based on the range of behaviors of individuals. Therefore, there are differences between the specific methods involved in SDMs that correspond to the particular utility of models and should be interpreted accordingly. In the following chapters, I rely on two SDMs based on correlative methods. First, the basis of all three studies involve Maxent, a popular machine learning-based

algorithm that calculates the environmental similarity of locations in a study area to the environmental conditions in locations were the species was recorded present. In one of my chapters, I also employ a multivariate logistic regression model, which outputs a map that estimates the probability of suitability by calculating the effect of environmental variables with data that contains both locations where the species is present and where the species is absent–or, at least, predicted absent.

Pivotal aspects of the research questions and methodologies in this dissertation are related to the second factor mentioned: the independent variables (i.e., environmental rasters) and data (i.e., species observation locations) used as inputs to the models. These inputs define models in notable ways, which are present through the dissertation. For example, in the second chapter, one SDM I developed employed broader-scale variables of the red-cockaded woodpecker habitat, including land cover and elevation (Grinnellian niche factors). By comparison, I developed another set of SDMs with environmental attributes at a finer scale with a proxy for ecological relationships (Eltonian niche factors) using 3D characteristics of forests derived from lidar data to represent the structural aspects of forest stands. In the third and fourth chapters, I create one model made with scientific observations of species and another with observations of species from citizen scientists. Therefore, in combination with a variety of methods, the specific data inputs to models allow for a diversity of key characteristics of SDMs, each with a particular utility and interpretation.

In the following dissertation chapters, I explore how models based on specific data inputs can help answer questions about the use of two emerging big data sources for SDMs: observation datasets from citizen science projects and spatial data from statewide lidar sources. Simply put, big data refers to datasets with particularly large volumes which limits conventional data tools and analysis. Given the scale of the current environmental challenges including unprecedented extinction rates and global climate change, big data in the environmental sciences has high value for ecological research (Hampton et al., 2013). Our research objectives were to understand how to leverage the information provided by these emerging big data sources for species distribution models.

1.4 Volunteered Geographic Information as an Emerging Spatial Data Source

Before defining what Volunteered Geographic Information (VGI) is, it is helpful to define what VGI is not. VGI is contrasted with so-called 'authoritative' data sources, and despite the fact that some VGI projects have been shown to be as trustworthy as authoritative data sources, particular aspects of data collection in a VGI framework make the distinctions between the data sources more evident, particularly given their intended use. Coote and Rackham (2008) provided a list of characteristics of conventional data, presented in by Sui, Elwood, and Goodchild (2012), which include:

- Data were specifically created with a set of requirements for use in legal, private, or administrative realms
- The data may be freely available, but usually there is a process for its dissemination (i.e., restrictions of access)
- An organization manages the data for a specific purpose and might have prepared contracts with others who have a stake in the data
- Data are gathered based on "established methods, standards, specifications, and practices"
- Data are collected by a professional and paid staff
- Quality of data is assured during the production of data, and some information on the quality is provided, typically in metadata
- Data are protected by licenses, copyrights, or formal agreements
- Data access is limited for reasons of "security, data protection, or commercial advantage"

By contrast, as presented in Sui, Elwood, and Goodchild (2012), Bruns (2008) discussed four core characteristics of information collected in the Web 2.0 era (i.e., VGI), including:

- Data collection is community-based (using a broad definition of community, from citizens of a particular town to birdwatchers globally) and therefore not limited to a limited number of individuals typically representing a small number of institutions
- Roles of data users and data producers are fluid, alternating between collector, reviewer, and arbitrator
- Data are under continuous review
- Data are typically more freely disseminated and often accessible via APIs

These distinctions are present in the differences between the eBird dataset and the dataset of observations from wildlife professionals, which are a basis for the latter two chapters of this dissertation. eBird is one of the most popular citizen science projects—and by definition, VGI data projects. Through eBird, birdwatchers become data producers who upload data points on locations where they encountered avian species. These data producers are part of a global community of unpaid wildlife enthusiasts with a goal of contributing species presence data for conservation in a broad sense. Data from eBird have been used in a variety of ways other than in SDMs, including to better understand species migration patterns or responses to climate changes. There are important data quality control mechanisms in place through eBird, but generally data contributors are assumed to be untrained and are not expected to collect data with any specific protocol or within any particular time frame. Finally, the data are made freely accessible for anybody to download and analyze. Hence, while there are particular benefits to using eBird data, including data collection over a wide spatial extent and across temporal time scales, there are particular disadvantages that call the use of eBird data into question when developing species distribution models.

Species presence data gathered by trained professionals (authoritative data, or often referred to in this dissertation as scientific data) has many noteworthy benefits. As detailed in Tye et al. (2016), wildlife professionals might be more invested than non-professionals in fieldwork and data collection endeavors with which they are particularly trained for and equipped. Furthermore, professionals may be more familiar with GPS technologies or use other tools used for data collection which may make locational data more trustworthy. Data from professionals might also

be less sensitive to local land cover administration, as professionals might be expected to survey remote areas away from parks and roads. Finally, trained professionals may be more likely to correctly identify species than data contributors to eBird. Hence, despite disadvantages of limited data with particular spatial coverage at specific time scales–attributes that might limit the utility of the data outside of its intended purposes–there are several beneficial characteristics of data collected by wildlife professionals.

The goal of these dissertation studies was not only to shed light on the use of VGI in the domain of Ecology, but also to better understand the utility of VGI broadly as a data source. Hence, with respect to the differences between the two data sources just discussed, the research in this dissertation seeks to contribute a study design and findings regarding VGI's utility as an emerging spatial data source for modeling species' distributions, and broadly within in GIS & T.

1.5 Scale and Spatial Data for SDMs

Issues related to the scale of spatial data (i.e., the data's resolution and extent) in SDMs have provided rich theoretical discussions. For example, Merow (2013) described the Modifiable Areal Unit Problem (MAUP) as it relates to SDMs: Study area boundaries of different shapes and sizes will include different patterns of environmental attributes, and thus a varying diversity of environments for models to discriminate suitable habitat. Therefore, models developed at the statewide extent will provide different estimates from those developed with national boundaries. Guisan et al. (2007) demonstrated that changing resolution of the same set of environmental variables had a range of effects on the quality of SDMs. Instead, I sought to contribute a practical approach based on the accessibility of environmental data derived from fine resolution lidar data. By providing detailed information on 3D attributes of space, across county- and state-wide extents, lidar has been monumental for environmental research. In Chapter 2, I sought to develop a method that approaches the scale-related issues that arise when this novel spatial data source is integrated in SDMs alongside spatial data with a coarser resolution.

As discussed earlier, the extent and resolution of spatial data used for environmental modeling corresponds to the operational scales of ecological processes or characteristics. But further, the operational scales of these phenomena correspond to ecological levels of

organization (Aspinall and Pearson, 1996; Hebblewhite, 2008; Kamal et al., 2015). Species' distributions are scale-dependent phenomena. At coarse scales of analysis, abiotic factors such as geology and climate influence the distributions of all species at the organizational level, but at finer scales of analysis, specific land covers influence species' distributions at the population level; at even finer scales, biotic factors like prey or other symbiotic relationships influence where individuals of the species are found (Soberón and Peterson, 2005; Hebblewhite, 2008). Thus, for each biological level of analysis there are appropriate spatial and temporal scales of explanatory variables that affect species' distributions.

Figure 1.2 highlights two types of input data for SDMs with different resolutions. On the left is raster data with a 30-m resolution, provided by the USGS Gap Analysis program for species distribution research, and on the right is a point cloud of a forested area from a lidar data file provided by the USGS 3D Elevation program. Instead of lidar data having pixel resolution, lidar point clouds provide 3D information at the finest spatial scales available. Scale-related issues in previous research make including both types of variables in SDMs problematic. In a fusion of GIScience and Biogeography, I utilized resolution as a trait of GIS data to explore scale-related issues of environmental variables commonly used in species distribution modeling.

Typically, SDMs are only developed at a single scale (with a single resolution) and it has become common practice to resample variables (manipulating raster data to a different resolution than the native one) when input rasters for SDMs have different resolutions. In the second chapter, I address this convention, which has significant effects on the representation of environmental phenomena. Resampling is particularly problematic with regards to the features derived from lidar, as converting the representation of 3D attributes from lidar data to 2D raster grid cells already results in the loss of information but becomes more problematic with larger raster cells. Thus, I sought to develop a method in which I avoided the pitfalls of including this fine resolution data in species distribution models with previously existing spatial data by utilizing inferences gained by analyzing lidar data at a fine resolution. Instead of resampling, the method introduced in the second chapter involves employing inferences about the relationship between environmental factors and species distributions at a broad scale in order to inform models about the relationship with a fine resolution. Thus, I operationalized the concept of scale,

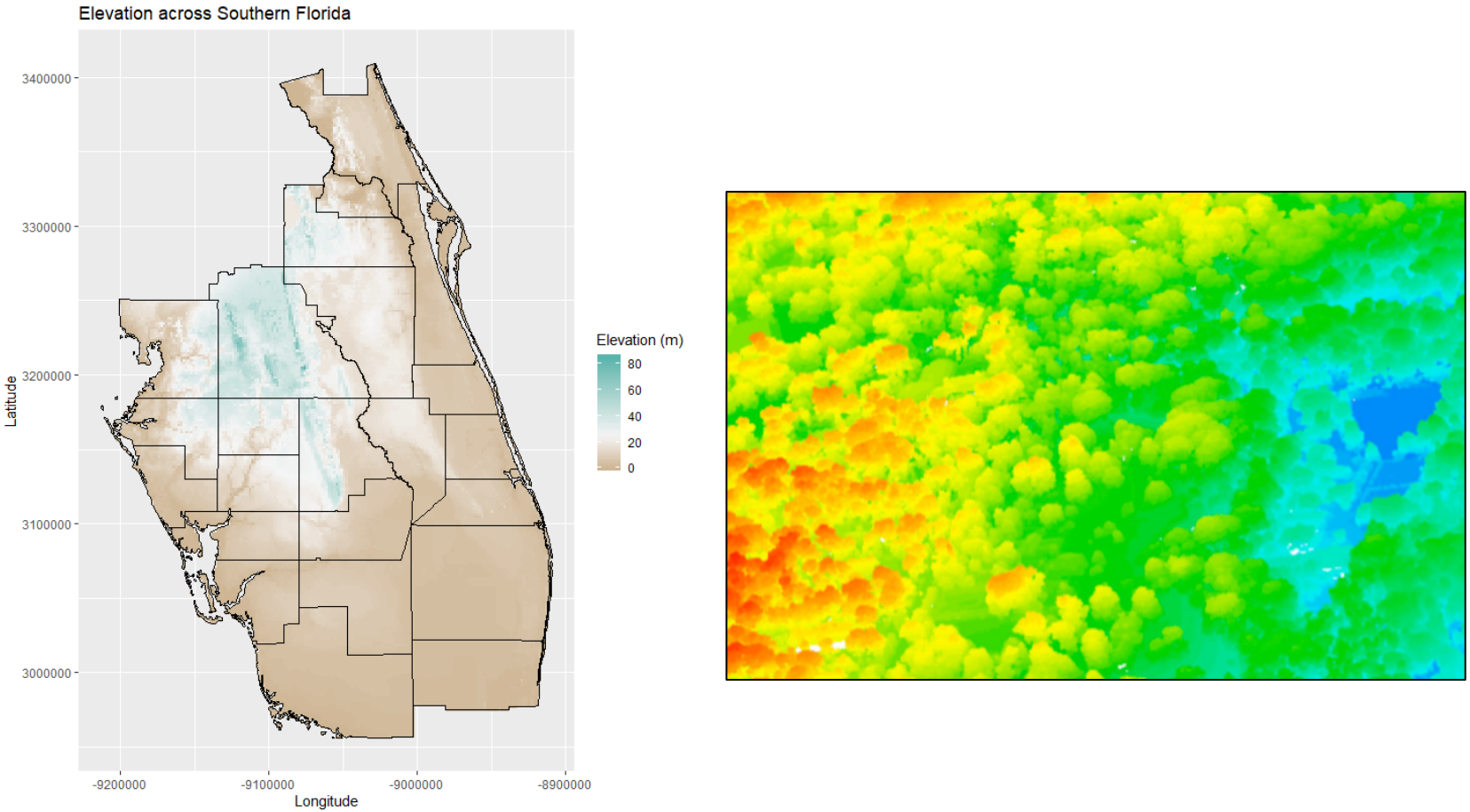


Figure 1.2 Two types of spatial data used as input variables for Species Distribution Models

which has riddled theoretical discussions about species distribution models, in developing a method to incorporate spatial data with different resolutions in SDMs.

1.6 Theoretical Underpinnings of this Research

1.6.1 *Unified Neutral Theory of Biodiversity and Niche Theory*

A fundamental theory foundational to species distribution modeling is the Unified Neutral Theory of Biodiversity (UNTB), presented by Stephen Hubbell (2001). UNTB, like other neutral and null theories in Ecology and the life sciences, suggests that species' distributions and community ecology are results of random walks, and that species' survival, speciation, and dispersal rates are irrelevant to their distributions (Hubbell, 2001). By suggesting the distributions of species are not responsive to abiotic and biotic conditions within different environments, the UNTB stands in opposition and as a null hypothesis to the traditional Niche Theory, which is foundational to the theoretical background of species distribution modeling. In contrast to UNTB, Niche Theory builds on the Grinnellian (i.e., environmental) and Eltonian (i.e., community ecology) factors of species' distributions discussed earlier (Soberón and Peterson, 2005). The Hutchinsonian Niche concept, developed by "Father of Modern Ecology" G.E. Hutchinson, brought together these two perspectives by explaining "niche" as an n-dimensional space defined by two axes: one being scenopoetic and the other being bionomic (Soberón and Peterson, 2005).

Scenopoetic, from the Latin word *setting*, variables are broad-scale by definition and are non-interacting, Grinnellian variables including abiotic attributes of space including terrain and atmospheric conditions, providing a setting for the species over relatively wide time scales like generations of a population (Soberón and Peterson, 2005). By contrast, bionomic variables are those that are discussed in the Eltonian context and include biotic resources that interact with the species, including species sympatry and prey, at fine temporal and local scales (Soberón and Peterson, 2005). In more recent theoretical development of Niche theory, scholars have begun to include variables related to movement and the dispersal of populations as another defining variable of niche. This is because potentially habitable areas require both source populations (Soberón and Peterson, 2005) and the absence of barriers for a species to colonize suitable areas.

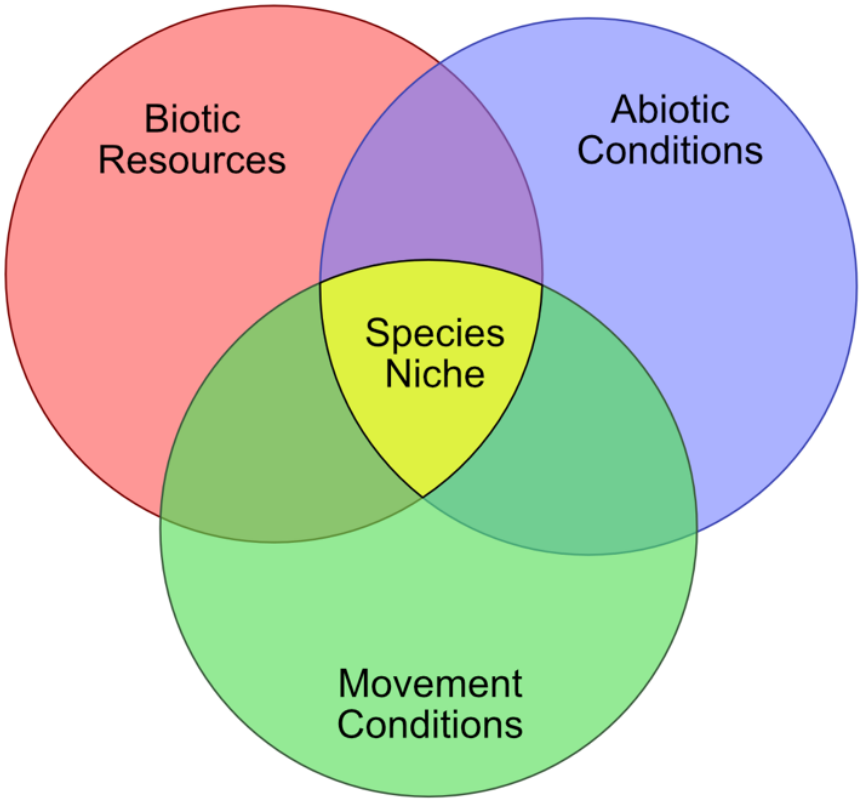


Figure 1.3 The BAM diagram illustrates the overlap of three variables that define a species' niche

These three factors comprise the BAM (Biotic-Abiotic-Movement) model (Figure 1.3), an illustration of the three core components of a species' distribution (introduced by Soberón and Peterson, 2005). Overlap of two of these variables also has ecological meaning, for example the term fundamental niche often refers to abiotic conditions suitable for species survival, and the realized niche for the overlap of biotic resources and abiotic conditions. However, if adequate source populations were not nearby or the niche was already filled by a different species (competitor or ecological equivalence), there would not be a niche available for the species. In the absence of specific biotic resources nearby, a source population might not be able to expand despite hospitable abiotic conditions. Finally, a source population might be limited from expansion despite biotic resources being in their favor because they were sensitive to abiotic conditions. By considering these variables which define the distributions of species, Niche Theory serves as a basis for SDMs, and the research questions addressed in this dissertation.

In the next two subsections, I briefly introduce how Niche Theory engages with two main philosophies of Science: Hierarchy Theory and Positivism. Ecosystems, which are comprised of communities of species each with its specific niche, have been classic examples used to study systems. I discuss one framework recognized in system science, Hierarchy Theory, which is particularly useful given the relationships between biological processes, level of organization, and scale of analysis. In the first study, I used a scale-based attribute of environmental data used in SDMs, and an approach theoretically rooted in Hierarchy Theory, to improve models related to species niche (more broadly, distribution) by informing models with inferences across scales. All of the chapters, from my perspective, borrow from a Positivist epistemology of Science particularly as I explore species' niches based statistical models that are inherently objective. However, the research direction of these studies does delve into questions prodded from a post-Positivist epistemology with questions regarding the data that inform us (and the models) about species niche.

1.6.2 *Hierarchy Theory and Scale*

In the first dissertation study, my research goal was to operationalize scale and the representation of species' distributions at multiple scales. This approach is rooted theoretically in Hierarchy Theory, an important framework for environmental modeling that has been

particularly relevant to spatial suitability modeling/niche distribution models in ecology. Robert O'Neill et al. (1989) provided a seminal text that helped researchers understand ecosystems as hierarchical systems made of nested units that explain the flow of energy among interacting biotic individuals. The organization of ecosystems through Hierarchy Theory has allowed for studies of ecosystems at the system level, while also providing a useful theoretical background for environmental modeling projects at lower levels of analysis. This framework captured and formalized theories on ecosystems (introduced in scientific literature by Forbes, 1887). O'Neill's theory proposal notes that constraints on species distributions occur at many scales of analysis. At the broadest scale, factors including climate and geology constrain distributions of species. At the finest scales, mechanisms in local ecosystems (i.e., symbiotic relationships) determine the distribution of individuals within a species. Thus, the distribution of species is constrained to a certain broad environmental gradient and is further defined by interacting components or mechanisms (i.e., symbiotic relationships) with other species. These are fine-scale attributes that might refer to mutualistic, competitor, or predator species that affect the distribution of the focal species.

To ground the framework of Hierarchy Theory as it applies to Niche Theory, consider the niche of the red-cockaded woodpecker (RCW), the focal species in the first manuscript. At a broad scale, certain conditions ensure the life cycle for the species. For example, RCW populations have been found to be less robust in areas with prolonged rainfall: the species' distribution responds to precipitation regimes which have broad temporal and spatial scales (Neal et al., 1993). At a finer scale, RCW populations are responsive particularly to longleaf pine forests, nested within areas that have appropriate climate conditions for the species (Walters et al., 1988; Walters, 1991; Smart et al., 2012). Furthermore, biotic mechanisms restrict individuals within the species at a fine scale, as individual woodpeckers rely on building their cavities in older longleaf pines that have bark softened by red-heart fungus; thus nested within forested areas nested within areas with appropriate abiotic conditions (Walters et al., 1988; Walters, 1991; Smart et al., 2012). Hierarchy Theory is the basis for the method I introduce in the first paper, scale-based background sampling (SBBS), in which I employ inferences about a species' niche

based on constraints at a broad scale in order to improve models developed based on biotic mechanisms at a fine scale.

From my perspective, the hierarchical concept of ecosystems helps ground the importance of scale, and thus Geography, in Niche Theory. The contributions of Geography to Ecology can be overlooked, but because this research focuses on species' distributions, which are uniquely biogeographic, I find the conversation important. The BAM concept discussed earlier has an inherent geographic basis because biological components of niche (fine-scale components of space) are nested in broad-scale atmospheric, geological, and hydrologic factors, so their diversity and homogeneity are inherently spatial. Furthermore, movement factors, which refers to the accessibility of areas based on historic and current spread of the species and barriers to these including rivers, mountain ranges, and highways, have an explicitly geographic foundation. While Ecology is a basis for several aspects of this dissertation (for instance, the mutualistic relationship between longleaf pine and red-cockaded woodpeckers), it is species' presence along particular environmental gradients across space and time which root the concept of Niche in Biogeography.

1.6.3 *Positivism in GIS*

The third and fourth chapters have a theoretical basis not solely in the domain of Ecology (though both certainly rely on Niche Theory), but more broadly in the Philosophies of Science. The questions I explore regarding the validity and additional information provided by VGI, compared with data from authoritative sources, are inherently epistemological, prompting questions regarding the process of knowledge production, including most broadly: *How do we know what we know?* Given the basis of spatial modeling in this dissertation research, answering this question first demands attention to the epistemologies that underpin science involving GIS.

Since it has emerged as a subfield of Geography, a historic criticism of GIS (and GIScience) has been its roots in Positivism (Lake, 1993). As defined by Robert Lake (1993), science from the positivist framework is inherently objective, strictly adheres to scientific methodologies, does not apply values to results or findings, and promotes the "ontological separation of subjects and objects" (i.e., subjects of studies and the researchers are independent of one another). In part, the

positivist character of GIS research comes from its development through the quantitative revolution in Geography and other fields. GIS was developed in order to apply algorithms and logic to geographic problem-solving, which was criticized as a one-dimensional, reductionist process. Human geographers and those from the field of so-called 'critical GIS,' who embraced a post-Positivist epistemology, felt that the development of GIS betrayed some of the core research priorities and methods within the field of Geography. GIS was lamented as opening the floodgates for reductionist scientific discussions and naïve empiricism, lacking in its ability to address complex geographic issues and methodological processes necessary to study Human Geography with the same ability as studying Physical Geography (Schuurman, 2000).

Post-Positivists often point to power structures present in knowledge production through GIS and GIScience, including through authoritative or scientific data production (Sieber and Hakla, 2015). Prior to VGI, geographic data production embodied power structures, since the wide dissemination of spatial data (such as land cover or census data) came from a limited number of commercial or civil institutions. Hence the production of data and subsequent analysis was centralized, which was considered exclusionary and undemocratic. Thus, GIS from a Positivist philosophy facilitated knowledge production (*How do we know what we know?*) based on data that were produced by a narrow group of people. While data from these sources also came with the confidence these datasets underwent appropriate quality control measures and were effectively representing reality, there were limitations to this type of data production. These limitations were embraced by the post-Positivist approach to GIS.

Data production in a VGI framework is at odds with positivist practices and assumptions, as decentralizing data production increases uncertainty regarding the accuracy and quality of data (i.e., do they represent reality?). Through Web 2.0, databases have been built by decentralized flows of information provided by the public using smart phones. These datasets can completely transform GIS by providing fine-scale information about an endless range of useful information, including regarding the movement of people, shared community experiences, and even abstract spatial phenomena like neighborhood character or regional accents. As technology has flattened inequalities across the Earth, VGI has emerged as a way to give voices to those whose political or economic conditions may otherwise be drowned out or silenced. Against the Positivist framework, VGI can be employed by geographers in a less automated way that meaningfully

promotes public engagement and education. In addition to VGI data collection benefitting scientific endeavors through a novel big data source, crowd sourced geographic information can be the foundation of useful platforms, which range from the promotion of community gardens to traffic avoidance to natural disaster management.

This dissertation research does not radically upend Positivism in the way those critical of the epistemology in Geography might appreciate. In fact, in the third chapter, I employ a Positivist framework as a basis of validating SDMs developed with VGI using data from an authoritative data source (i.e., the scientific model is a benchmark). However, research in both the second and third chapters, while noting differences between VGI and conventional authoritative datasets, objectively finds that citizen science data does provide additional and unique information compared with authoritative data on the representation of niche. As I seek to both operationalize scale through the epistemology of Hierarchy theory, and approach Positivist and post-Positivist schisms in GIScience, I believe the research questions and methods in this dissertation directly contribute to salient and relevant theoretical issues in the field of Geography and, more broadly, Science.

1.7 Research Questions

My intention in this dissertation is to highlight methodologies in which GIScience is used to answer important questions about emerging data sources in the broader field of Geography, while also providing contributions to Biogeography. Thus, as I suggested earlier, the focuses of these studies are not solely rooted in the domain of Physical Geography, but in the use of spatial data to shed light on underlying interrelationships of geographic data and information. Answers to our research questions thus have a broader impact than only improving models of species distributions, though these findings are valuable in that domain. Moreover, they provide insight as to how lidar data (3D representations of space) can be used alongside conventional geographic methods that have traditionally relied on raster imagery, and help assess the utility of emerging data repositories that collect data from the public (VGI) relative to authoritative, scientific data sources.

Through the first research question, I propose an approach that investigates the utilization of spatial imagery data at multiple resolutions to understand niche variables at different levels of

analysis. The second and third research questions involve the use and accuracy of volunteered wildlife observation data at the observation and aggregated levels of analysis, which recognizes the attributes of species distribution models at two biological levels of organization (individual and population). The research questions explored in the following chapters within this dissertation include:

Chapter 2

a. *How can spatial data with different resolutions be utilized in a complementary approach to improve the performance of SDMs?*
b. *How can inferences gained at a coarser scale of analysis be used to improve models developed at local scales?*

Chapter 3

a. *How do SDMs developed with citizen science presence data (i.e., VGI-based SDMs) perform compared with those developed with presence data provided by scientists and professionals?*
b. *In general, what factors or conditions influence the concordance/discordance of VGI-based SDMs with respect to scientific models? (i.e., in what conditions do VGI-based SDMs perform reasonably well?).*

Chapter 4

a. *Relative to observation points provided by scientists and wildlife professionals, what additional information can wildlife observation points from citizen scientists contribute for species distribution modeling?*
b. *What can the differences in the environmental characteristics of eBird points tell us about the species' distribution and species presence data from citizen scientists?*

1.8 Dissertation Organization

The research chapters are presented in Chapters 2, 3, and 4. In Chapter 2 I present our study ***Employing inferences across scales: Integrating spatial data with different resolutions to enhance Maxent models***. In this chapter, I explore scale-based background sampling (SBBS), a method that facilitates the utility of fine-scale lidar data for species distribution models.

In Chapter 3, I present our study ***Assessing the differences between species distribution models developed with scientific data and those developed with data from citizen science projects***. In this study, I developed an approach to calibrate and validate an SDM developed with citizen science data based on an SDM developed using a dataset from wildlife professionals. Finally, in Chapter 4 I present our study ***Evaluating and filtering volunteered geographic information using scientific datasets for species distribution models.*** In this study, I used the statistical measure of influence, along with inferences from a model developed with scientific data, to quantify the additional information provided by individual eBird species observation points.

Finally, in the Conclusions section (Chapter 5), I highlight our results and methodological findings, as I link them to the central research questions and explain the broader merits of the work.

Works Cited

Araújo, M. B., R. P. Anderson, A. M. Barbosa, C. M. Beale, C. F. Dormann, R. Early, R. A. Garcia, A. Guisan, L. Maiorano, B. Naimi, R. B. O'Hara, N. E. Zimmermann, and C. Rahbek. (2019). Standards for distribution models in biodiversity assessments. *Science Advances* 5 (1), 1 – 12.

Aspinall, R.J. and Pearson D.M. "Data Quality and Spatial Analysis: Analytical Use of GIS for Ecological Modeling" GIS and Environmental Modeling: Progress and Research Issues. Goodchild, M. F., Steyaert, L. T., Parks, B. O., Johnston, C., Maidment, D., Crane, M., & Glendinning, S. (Eds.). John Wiley & Sons (1996): 35-39.

Bonnett, A. (2008). *What is geography?* Sage.

Bruns, A. (2008). *Blogs, Wikipedia, Second Life, and beyond: From production to produsage* (Vol. 45). Peter Lang.

Buffon, C. D. (1761). Histoire naturelle, geáneárale et particulieáre, Vol. 9. *Paris, Imprimerie Royale*, *4*.

Conrad, C. C., & Hilchey, K. G. (2011). A review of citizen science and community-based environmental monitoring: issues and opportunities. *Environmental monitoring and assessment*, *176*(1-4), 273-291.

Coote, A., & Rackham, L. (2008). Neogeographic data quality- is it an issue? AGI Geocommunity Conference. ConsultingWhere Ltd.

de Candolle, A. (1855). *Géographie botanique raisonnée ou exposition des faits principaux et des lois concernant la distribution géographique des plantes de l'époque actuelle* (Vol. 2). V. Masson.

Drew, C. A., Wiersma, Y. F., & Huettmann, F. (Eds.). (2010). *Predictive species and habitat modeling in landscape ecology: concepts and applications*. Springer Science & Business Media.

Douglas, I., Huggett, R. J., Robinson, M., & Robinson, M. E. (Eds.). (2002). *Companion encyclopedia of geography: the environment and humankind*. Taylor & Francis.

Elith, J., & Leathwick, J. R. (2009). Species distribution models: ecological explanation and prediction across space and time. *Annual review of ecology, evolution, and systematics*. *40*, 677-697.

Elton, C. (1927). Animal Ecology. The University of Chicago Press, London.

Gavin, D.G. “Biogeography.” *21st Century Geography: A Reference Handbook (Vol. 1)*. Ed. Joseph P. Stoltman. Thousand Oaks: SAGE (*2012)*. 745-752

Goodchild, M. F. (2007). Citizens as sensors: the world of volunteered geography. *GeoJournal*, *69* (4), 211-221.

Goodchild, M. F. (2011). Scale in GIS: An overview. *Geomorphology*, *130* (1-2), 5-9.

Grinnell, J. (1904). The origin and distribution of the chest-nut-backed chickadee. *The Auk*, *21* (3), 364-382.

Guisan, A., Graham, C. H., Elith, J., Huettmann, F., & NCEAS Species Distribution Modeling Group. (2007). Sensitivity of predictive species distribution models to change in grain size. *Diversity and distributions*, *13* (3), 332-340.

Hampton, S. E., Strasser, C. A., Tewksbury, J. J., Gram, W. K., Budden, A. E., Batcheller, A. L., ... & Porter, J. H. (2013). Big data and the future of ecology. *Frontiers in Ecology and the Environment*, *11* (3), 156-162.

Hanks, R., & Stadler, S. J. (2011). Encyclopedia of Geography Terms. *Themes, and Concepts,* Santa Barbara, California: ABC-CLIO.

Hebblewhite, M. (2008). Scientific Review for the Identification of Critical Habitat for Woodland Caribou (Rangifer tarandus caribou), Boreal Population, in Canada. Environment Canada, Ottowa.

Humboldt, A.V. (1793). Florae fribergensis specimen, plantas cryptogamicas praesertim subterraneas exhibens. H.A. Rottmann Berolini. Accessible at: https://www.biodiversitylibrary.org/item/267829

Humboldt, A. V., & Bonpland, A. (1807). Ideen zu einer Geographie der Pflanzen nebst einem Naturgemälde der Tropenländer. *Tübingen, Germany*.

Hubbell, S. P. (2001). *The Unified Neutral Theory of Biodiversity and Biogeography (MPB-32)*. Princeton University Press.

Kamal, M., Phinn, S., & Johansen, K. (2015). Object-based approach for multi-scale mangrove composition mapping using multi-resolution image datasets. *Remote Sensing*, *7* (4), 4753-4783.

Koeppe, M. T. “Environmental Planning and Management.” *21st Century Geography: A Reference Handbook*. Ed. Joseph P. Stoltman. Thousand Oaks: SAGE (*2012)*. 745-752

Kreft, H., & Jetz, W. (2010). A framework for delineating biogeographical regions based on species distributions. *Journal of Biogeography*, *37* (11), 2029-2053.

Lam, N. S.-N. (2019). Resolution. Pages 1 – 8 *in* J. P. Wilson, editor. The Geographic Information Science & Technology Body of Knowledge. 2019 Editi.

Lake, R. W. (1993). Planning and applied geography: Positivism, ethics, and geographic information systems. *Progress in human geography*, *17* (3), 404-413.

Li, W., Guo, Q., & Elkan, C. (2011). Can we model the probability of presence of species without absence data? *Ecography*, *34* (6), 1096-1105.

Liu, C., White, M., & Newell, G. (2013). Selecting thresholds for the prediction of species occurrence with presence-only data. *Journal of biogeography*, *40* (4), 778-789.

Merow, C., Smith, M. J., & Silander Jr, J. A. (2013). A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. *Ecography*, *36* (10), 1058-1069.

Millington, A. C., M. A. Blumler, and U. Schickhoff. (2011). Situating Contemporary Biogeography. *The SAGE Handbook of Biogeography*. Sage.

Morrison, J. L. (2001). *Recommended management practices and survey protocols for Audubon's Crested Caracara (Caracara cheriway audubonii) in Florida* (No. 18). Technical Report.

Murrow, J. L., Thatcher, C. A., Van Manen, F. T., & Clark, J. D. (2013). A data-based conservation planning tool for Florida Panthers. *Environmental Modeling & Assessment*, *18* (2), 159-170.

Neal, J. C., James, D. A., Montague, W. G., & Johnson, J. E. (1993). Effects of weather and helpers on survival of nestling red-cockaded woodpeckers. *The Wilson Bulletin*, 666-673.

Nicolson, M. (1987). Alexander von Humboldt, Humboldtian science and the origins of the study of vegetation. *History of science*, *25* (2), 167-194.

O'Neill, R. V., Deangelis, D. L., Waide, J. B., Allen, T. F., & Allen, G. E. (1986). *A Hierarchical Concept of Ecosystems* (No. 23). Princeton University Press.

Openshaw, S. (1991). A view on the GIS crisis in geography, or, using GIS to put Humpty-Dumpty back together again. *Environment and Planning A: Economy and Space*, *23* (5), 621-628.

Rickles, P., M. Haklay, C. Ellul, and A. Skarlatidou. (2017). Citizen Science with GIS & T (GS-24). J. P. WIlson, editor. The Geographic Information Science & Technology Body of Knowledge.

Schuurman, N. (2000). Trouble in the heartland: GIS and its critics in the 1990s. *Progress in Human Geography*, *24* (4), 569-590.

Sieber, R. E., & Haklay, M. (2015). The epistemology(s) of volunteered geographic information: a critique. *Geo: Geography and Environment*, *2* (2), 122-136.

Sui, D., Elwood, S., & Goodchild, M. (Eds.). (2012). *Crowdsourcing Geographic Knowledge: volunteered geographic information (VGI) in theory and practice*. Springer Science & Business Media.

Smart, L. S., Swenson, J. J., Christensen, N. L., & Sexton, J. O. (2012). Three-dimensional characterization of pine forest type and red-cockaded woodpecker habitat by small-footprint, discrete-return lidar. *Forest Ecology and Management*, *281*, 100-110.

Soberón, J. (2007). Grinnellian and Eltonian niches and geographic distributions of species. *Ecology letters*, *10* (12), 1115-1123.

Soberón, J., & Peterson, A. T. (2005). Interpretation of models of fundamental ecological niches and species' distributional areas.

Sullivan, B. L., Wood, C. L., Iliff, M. J., Bonney, R. E., Fink, D., & Kelling, S. (2009). eBird: A citizen-based bird observation network in the biological sciences. *Biological Conservation*, *142* (10), 2282-2292.

Tye, C. A., McCleery, R. A., Fletcher Jr, R. J., Greene, D. U., & Butryn, R. S. (2017). Evaluating citizen vs. professional data for modeling distributions of a rare squirrel. *Journal of Applied Ecology*, *54* (2), 628-637.

University Consortium for Geographic Information Science. (2006). Geographic Information Science & Technology Body of Knowledge. Page (D. DiBlias, M. DeMers, A. Johnson, K. Kemp, A. T. Luck, B. Plewe, and E. Wentz, Eds.). The Association of American Geographers, Washington, D.C.

Warf, B. (Ed.). (2010). *Encyclopedia of Geography.* Sage.

Walters, J. R. (1991). Application of ecological principles to the management of endangered species: the case of the red-cockaded woodpecker. *Annual Review of Ecology and Systematics*, *22* (1), 505-523.

Walters, J. R., Hansen, S. K., Carter III, J. H., Manor, P. D., & Blue, R. J. (1988). Long-distance dispersal of an adult red-cockaded woodpecker. *The Wilson Bulletin*, 494-496.

Wright, D. J., Goodchild, M. F., & Proctor, J. D. (1997). GIS: tool or science? Demystifying the persistent ambiguity of GIS as "Tool" versus "Science". *Annals of the Association of American Geographers*, 346-362.

CHAPTER 2 EMPLOYING INFERENCES ACROSS SCALES: INTEGRATING SPATIAL DATA WITH DIFFERENT RESOLUTIONS TO ENHANCE MAXENT MODELS

A version of this chapter was published by Adam G. Alsamadisi, Liem T. Tran, and Monica Papeş:
Alsamadisi, A. G., Tran, L. T., & Papeş, M. (2020). Employing inferences across scales: Integrating spatial data with different resolutions to enhance Maxent models. *Ecological Modelling*, *415*, 108857.

My use of "we" in this chapter includes my co-authors, Liem T. Tran and Monica Papeş. As first author, I led experimental design, data analyses, and writing the manuscript.

Abstract

Challenges associated with developing species distribution models (SDMs) with high-resolution data (including from lidar) prompted our investigation into a complementary approach to enhance the performance of SDMs using spatial data with different resolutions. In our experiment we developed a model with Maxent (a presence-background SDM) with variables that had a 30-m resolution, and then used the output of the model to restrict the background sampling area for models developed with variables that had a 10-m resolution. According to common measures of model quality, this approach produced better models than both a model developed with the default Maxent background sampling area and a model developed using the conventional approach of resampling environmental data to a common spatial resolution. We then reviewed the ecological meaning of this approach and observed how model mechanics were impacted as restricting the background sampling areas led to background points that had a greater contrast with the presence points, and therefore different environmental characteristics than background points sampled from the default background sampling area.

2.1 Introduction

Species distribution models (SDMs) are important tools for environmental scientists and are frequently employed for diverse purposes, including as supporting evidence for conservation decisions and understanding environmental threats like climate change, habitat loss, and invasive species. Lidar data provide information about 3-dimensional attributes of habitat, which relate to forest structure and/or successional stages and thus can be used in SDMs as a proxy for the structural and biotic components of the species' niche (Smart et al., 2012; Vierling et al., 2008). Despite the benefits of incorporating lidar data in species distribution research, processing these data is typically computationally expensive and thus often driven by specific needs (Vierling et

al., 2008). As such, models with fine spatial resolutions may be best suited for limited spatial extents compared to those developed using variables with coarser resolutions. In this study, we introduce an approach to improve models developed with fine resolution variables by employing inferences gained from a model developed with coarse resolution variables, at a broader spatial extent.

We introduce our approach with Maxent, a commonly used algorithm for SDMs in the presence-background context (Phillips et al., 1997). Maxent is often noted for having predictive power that is competitive with models that use absence data, despite lacking information on where the species is not present (Elith et al., 2011). Maxent begins with a uniform prediction model which states that the species is equally likely to be present through the study area and uses a machine learning technique that contrasts the environmental characteristics at locations where the species is present with the environmental characteristics at background points (which are by default sampled from the entire study area). As summarized in Elith, et al. (2010), the resulting model minimizes relative entropy between the distributions of independent variables at presence locations and the distributions of independent variables at background points.

One drawback of introducing fine resolution data to SDMs is that the variables used in the model may need to be resampled to a common resolution. Smart et al. (2012) modeled red-cockaded woodpecker (*Picoides borealis*) habitat using lidar at a US Marine Corps Base in Jacksonville, North Carolina, and found that integrating lidar variables improved models; however, the variables had to be resampled to a coarser resolution in order to be included in the models with other variables. Alvarez-Martinez, et al. (2018) used a range of environmental variables to develop Maxent models for a number of vegetation types along the northern coast of Spain. However in order to include climate data in these models, Alvarez-Martinez, et al. (2018) needed to resample rasters to a finer resolution to be analyzed along Landsat and lidar derived data. Although it has become conventional to resample data with different native resolutions to a common resolution, this has important drawbacks, which include introducing uncertainty in the model and sacrificing detail about the independent variables (Dixon and Earls, 2009).

Prompted by the challenges associated with incorporating fine resolution data in SDMs, our research questions were: (1) How can spatial data with different resolutions be utilized in a complementary approach to improve the performance of SDMs? and (2) How can inferences

gained at a coarser scale of analysis be used to improve models developed at local scales? One way that researchers have improved the performance of Maxent models is by restricting the area used to sample background points (needed as inputs to the model) to places where environmental characteristics are unsuitable to the species (Anderson and Raza, 2010; Barve et al., 2011). Published examples of this method, as detailed in Jarnevich & Young (2015), include restricting background-sampling areas with a minimum convex polygon around presence locations (Jarnevich and Young, 2015), limiting the background sampling area based on prevailing environmental conditions like Köppen-Geiger zones or ecoregions (Blanchard et al., 2015; Tererai and Wood, 2014; Webber et al., 2011), eliminating areas that are inaccessible for the focal species from the background sampling area (Elith et al., 2010), and by selecting background points only from areas within a set proximity from presence points (VanDerWal et al., 2009). We hypothesized that by removing areas predicted suitable by models at a different scale from the background sampling area during Maxent model development, the environmental characteristics of the background points and the presence points would be contrasted to a greater degree, which would result in improved model performance.

Employing inferences across scales when developing spatial models is an important focus in spatial ecology which bridges theories of ecological processes with multiscale ecological modeling practice (Miller et al., 2004). Researchers may need to aggregate local observations to improve their understanding of broad scale patterns, or to extrapolate broad scale knowledge to better model local and individual systems (Bombi and D'Amen, 2012; Fernandes et al., 2014; Miller et al., 2004). One scale-based characteristic of environmental variables is the spatial resolution, which is defined as the smallest distinguishable unit of the data (i.e., cell size for raster files). Data with a fine spatial resolution are used to model phenomena like geometric attributes of forest patches or the density of understory across forest stands, whereas data with a coarse spatial resolution correspond to broad-scale phenomena with larger spatial extents like ecoregions or climate regions (Clark et al., 2011; Holt, 2009). In our experiment, we sought to apply knowledge gained by analyzing niche models with a set of environmental variables with a coarse resolution to improve a model developed using variables that had a finer resolution.

2.2 Materials and Methods

2.2.1 *Focal Species: Red-cockaded woodpecker*

The Red-cockaded Woodpecker (henceforth RCW), is a non-migratory woodpecker species found along the coasts of the Southeastern U.S. (Costa, 2002). In 1970, the species was listed as federally endangered facing three major threats: forest management, fire suppression, and urbanization (Hanula and Engstrom, 2000). As an obligate specialist species, RCW distributions are spatially bound by distributions of mature longleaf pines. Individuals excavate cavities in older longleaf pine trees (typically 80 – 120 years old, with the heartwood of the tree softened by red-heart fungus) in coastal pine forests of the Southeastern United States (Smart et al., 2012; Walters, 1991; Walters et al., 1988). Research suggests RCW prefer to construct cavities in taller trees with open understories (maintained by fire regimes), which are hypothesized to make cavities less accessible for snakes and other predators (Costa, 2002; Smart et al., 2012; Walters, 1991). Conservation efforts and research on the species are important because the species plays a keystone role in its ecosystems: an estimated 27 other species will reside in abandoned or usurped cavities the RCW excavates (Costa, 2002; Jusino et al., 2015).

2.2.2 *Species Presence Data and Environmental Data*

Presence data for the RCW were recorded from 2000 – 2011 and provided by the Florida Natural Areas Inventory, a non-profit research organization at Florida State University that maintains state biodiversity inventories (FNAI, 2013). 1,858 RCW presence points within Florida's statewide extent were classified in the dataset as 'Active' and used to train and test the 30-m model. Because of the smaller extent of the fine-scale model, only 222 of these points were used to train and test the 10-m models. Six of the independent variables used to develop the models in our experiment were provided by the US-Gap Analysis Program (elevation, slope, aspect, Human Impact Avoidance, distance from forest edge, and land cover) and had a native resolution of 30-m (Table 2.1). The Florida Forest Service provided a raster of forest age that also had a native resolution of 30-m.

Table 2.1 Variables used in Maxent models developed in our experiment

Variable	Variable Description	Data Source	Resolution
Forest Age	Categorical raster corresponding to biomass stand age classified in 10-year intervals, derived from Landsat imagery	Florida Forest Service (2013)	30-m
Land Cover	GAP National Terrestrial Ecosystems 2011 includes detailed vegetation and land cover information for the entirety of the United States, representing 590 land cover classifications nationally and 102 land cover classifications within Florida	USGS Gap Analysis Program	30-m
Elevation	Continuous raster, digital elevation model derived from the National Elevation Dataset	USGS Gap Analysis Program, Gesch et al., (2002)	30-m
Distance from Forest Edge	Raster layer which corresponds to the distance to the forested edge when in nonforested land cover or from the forest edge when within the forest interior	USGS Gap Analysis Program	30-m
Human Impact Avoidance	Rank based variable in which lower values indicate landscape with less human disturbance	USGS Gap Analysis Program	30-m
Slope	Dataset derived from the elevation dataset which reflects the slope (i.e., rise over run) of the area, reported in degrees	Derived from Elevation raster, Gesch et al., (2002)	30-m
Aspect	Categorical raster derived from elevation dataset which corresponds to the cardinal direction the area within each cell is facing (i.e., North-facing)	Derived from Elevation raster, Gesch et al., (2002)	30-m
Vertical Complexity Index	The VCI is based on the evenness of vegetation heights grouped in bins. Lower values of VCI indicate forest stands in early development stages while larger values indicate the uneven vertical distribution of vegetation in old growth forest stands.	USGS 3D Elevation Program	Processed 10-m Raster
Standard Deviation	Standard deviations of (first return) lidar points, hence representing the standard deviation of tree heights within a 10-m grid. Higher standard deviation indicates forest stands with a more uneven height distribution of vegetation.	USGS 3D Elevation Program	Processed to 10-m Raster
Percentage of Ground Points	Percentage of lidar points (from all lidar returns) classified as a ground point within a 10-m grid. Higher percentages would indicate a less dense vegetation understory.	USGS 3D Elevation Program	Processed to 10-m Raster
Mean Height	Mean height of first returns, hence representing the mean tree height within a 10-m grid. Large values indicate average vegetation height that is taller within each cell.	USGS 3D Elevation Program	Processed to 10-m Raster
Entropy	Entropy or unevenness of first returns, hence representing the evenness of vegetation heights. Cells with greater entropy have less consistent tree heights than those with smaller entropy	USGS 3D Elevation Program	Processed to 10-m Raster
Maximum Height	Maximum height of first returns, hence representing the maximum tree height within a 10-m grid. Higher lidar maximum heights would indicate forest stands with taller trees.	USGS 3D Elevation Program	Processed to 10-m Raster
Kurtosis	Kurtosis of first returns point heights, hence representing the kurtosis of tree height within a 10-m grid. As kurtosis captures the sharpness of the distribution, cells with larger values indicate forest stands with one tall tree and shorter surrounding vegetation.	USGS 3D Elevation Program	Processed to 10-m Raster
Skew	Skew of first returns heights, hence representing the skew of tree height distribution. Cells with a larger skew have a more uneven distribution of tree heights, with larger trees, than cells with a greater number of trees with equally short vegetation.	USGS 3D Elevation Program	Processed to 10-m Raster

We processed lidar data provided by the USGS 3D Elevation Program (average point spacing of 0.36 points/m^2) to a 10-m raster using common forest metrics with the *lidR* package in R (Roussel et al., 2017). The 10-m resolution was chosen following a study conducted by Zellweger, et al. (2014), which stated the resolution would result in cells containing a sufficient quantity of lidar points to allow for the preservation of structural attributes of vegetation and habitat. Lidar forest metrics correspond to three-dimensional measures of forest stands, some of which are known to affect RCW distribution, including the Vertical Complexity Index (VCI), a measure of vegetation height similarity which can correspond to stand age dynamics (van Ewijk et al., 2013), and the percentage of ground points which can be used to measure understory density (Martinuzzi et al., 2009).

2.2.3 ***Methods***

Maxent models were developed using the *dismo* package in R developed by Hijmans, et al. (2017) using the cross-validation approach (with 10 model runs). The cross-validation method was chosen because it uses all presence data provided, randomly selected by Maxent as either a testing point or training point for each model run during model development (Phillips et al., 2006). We used a commonly employed product of Maxent, the logistic output, a raster (which has the same resolution as the model's input variables) with values that range from 0.0 – 1.0 that correspond to the similarity of environmental characteristics of each location to the locations where the species was recorded present (Elith et al., 2011; Phillips et al., 2006).

We developed one Maxent model with the extent of the state of Florida using the variables with a native resolution of 30-m (Figure 2.1). We then developed a model with variables from lidar data processed to a resolution of 10-m that had a spatial extent of Okaloosa County (located along the Florida Panhandle) which used background points from the default background sampling area, the entire study area. To use the output of the 30-m model to restrict the background sampling area for the 10-m model, we averaged the outputs of model runs from the 30-m model and clipped the averaged raster to the spatial extent of the 10-m model (Okaloosa County).

We then produced three binary rasters (suitable/unsuitable) from the 30-m resolution model using three thresholds provided with the logistic output of Maxent. The first two were the minimum/maximum values of these thresholds: the minimum training presence threshold (0.006)

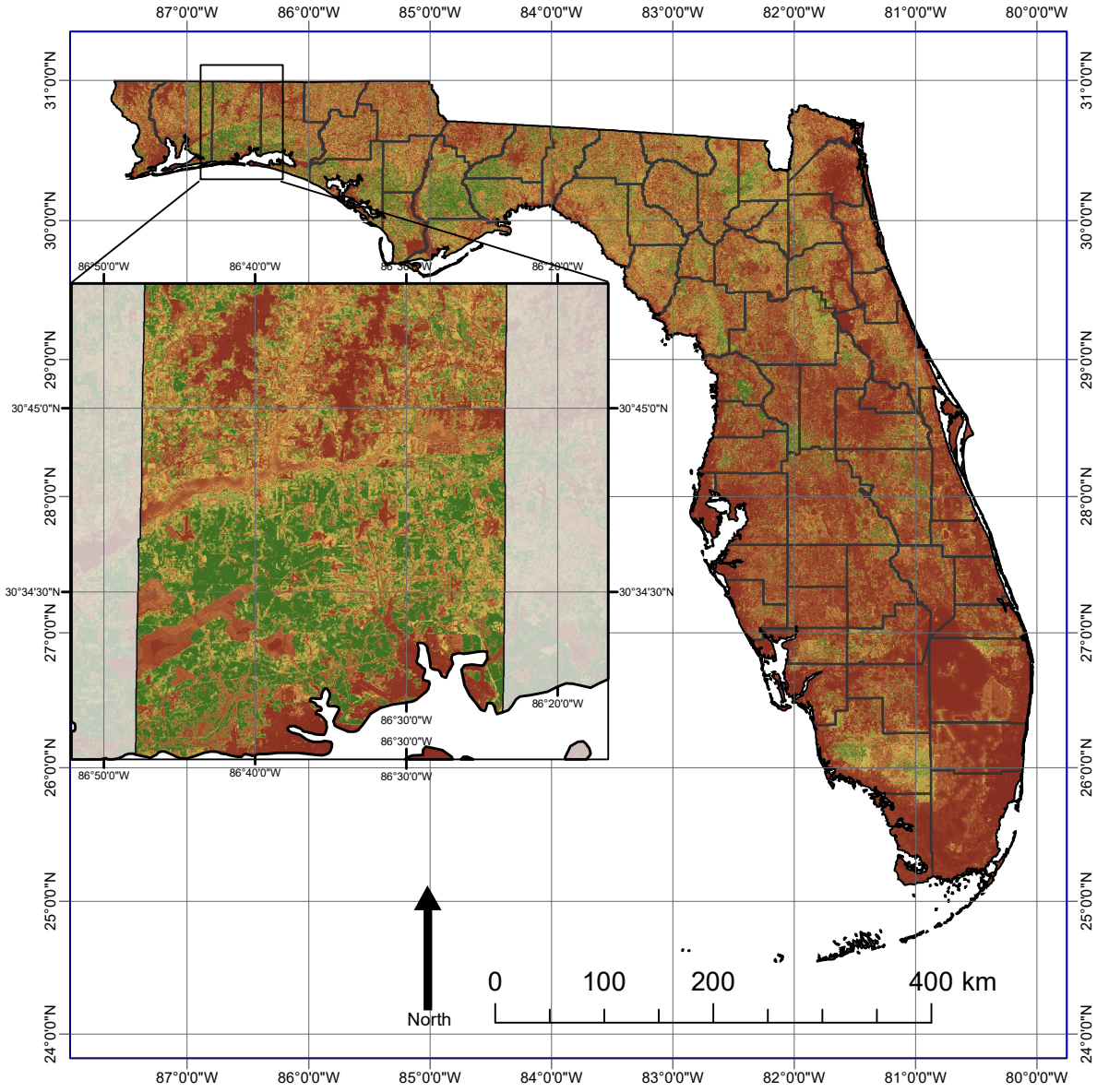


Figure 2.1 Maxent model for red-cockaded woodpecker developed using variables that had a 30-m resolution which had the spatial extent of Florida. Okaloosa County, in the map subset, was used to test our approach.

and the equal training sensitivity and specificity threshold (0.3761). The third was the mean value (0.1919) of all logistic thresholds provided by Maxent. The testing data omission rate is the ratio at which the binary suitable model (e.g. the model after the threshold has been applied) identifies cells with presence locations as unsuitable. Because the rate is sensitive to the threshold, the testing omission rate at the maximum threshold was highest (≈ 21.16%) and was lower at the mean threshold (≈ 8.18%), while at the minimum threshold there was a 0% testing omission rate (i.e., at this threshold all testing points were predicted suitable). We tested our approach with different thresholds based on the rationale that Maxent threshold values correspond to different attributes of model performance, and because the ecological meaning of the thresholds has been debated (Liu et al., 2005). The different background sampling areas are reported in Figure 2.2.

We developed three additional models with the lidar variables and the full presence dataset, however, during model development we restricted background point sampling to areas predicted unsuitable by the averaged 30-m resolution model at the three thresholds. To address the possibility that spatial sampling bias in the presence dataset affected models, we developed three additional models using the same approach after employing the chessboard sampling method on the presence data (which led to 117 presence points used in these models). One benefit of the chessboard sampling approach is that testing and training samples are spatially separated so that evaluation statistics are not affected by their proximity. We presented the diagnostics of models developed with the chessboard sampled presence points alongside those of the models developed with the full set of presence points (however all other figures and tables are based on the models with the full presence dataset). In order to compare this approach, which we will refer to as scale-based background sampling (SBBS), to the common convention of resampling environmental data, we developed a resampled variable model with both the original 10-m variables and the 30-m variables resampled to a 10-m resolution with the nearest neighbor technique (selected to avoid influencing the model with an interpolation of variable values).

We assessed the performance and improvement of model performance using the gain of the models, the Area Under the Receiver Operating Characteristic Curve (AUC), the sample size corrected Akaike Information Criterion (AICc) calculated using the *ENMeval* package in R (Muscarella et al., 2018), and the maximized True Skill Statistic (TSS_{max}) calculated using the

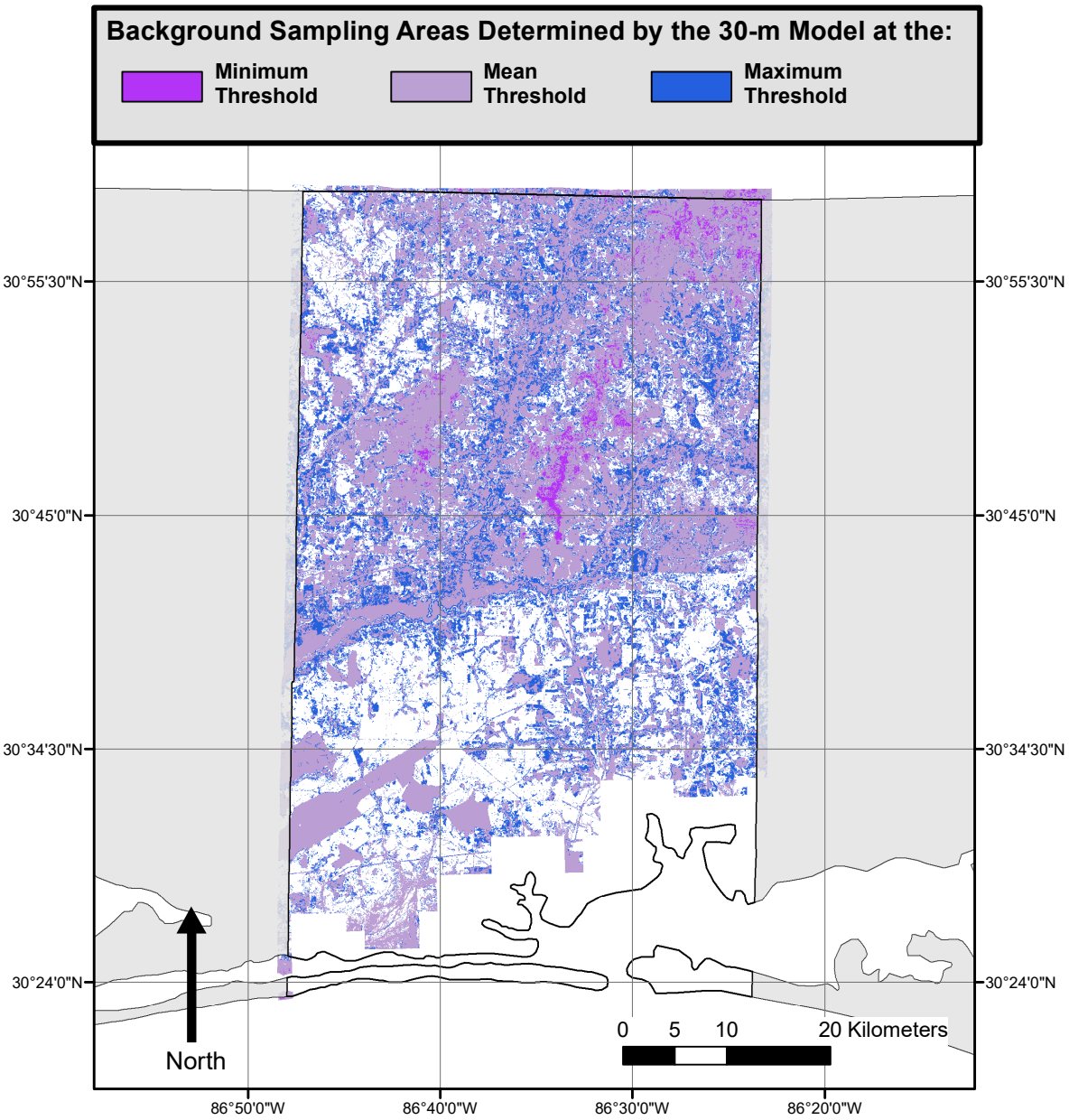


Figure 2.2 Background sampling areas for the model developed with 10-m resolution variables were determined using the output of the model developed with 30-m resolution variables. We compared three background sampling areas which were based on three thresholds.

Ecospat package in R (Broennimann et al., 2018). In Maxent, the gain of the model (detailed in Equation 2.1) is central to the machine learning process used to determine the prediction surface. When the gain value is exponentiated, it is the likelihood ratio of a typical presence point to a typical background point. The iterative model development process performed in Maxent depends on the gain as the model explores and determines the model's solution. When the gain is maximized, the solution of the model best discriminates locations with observation points from background point locations. The gain is calculated by the following (Merow et al., 2013): The second term in Equation 2.1 is based on the predicted values at background locations. Through SBBS we expect there will be fewer background points in suitable areas where there are presence data compared with background points randomly sampled through the entirety of the study area. Thus, we will minimize term 2 in Equation 2.1, resulting in a larger model gain.

$$gain = \frac{1}{m}\sum_{i=1}^{M} z(x_i)\lambda$$ *1. Sum of predicted suitability values at presence locations*

$$- log \sum_{i=1}^{N} Q(x_i)e^{z(x_i)\lambda}$$ *2. Sum of predicted suitability values at background locations*

$$- \sum_{j=1}^{J} |\lambda_j| * \beta * \sqrt{s^2[z_j]/M}$$ *3. Regularization Parameter*

Equation 2.1 Gain in Maxent, adapted from Merow, Smith and Silander 2013.The first term is the sum of the predicted values at presence locations, where M is the number of presence locations, z are the environmental variables at presence locations x_i, and $\boldsymbol{\lambda}$ is the set of regression coefficients. The second term is the sum of predicted values at background locations and reduces the gain as larger values are associated with background locations, N are the background locations and Q(x) are the environmental data values across the background area. The third term is the regularization coefficient, in which β is the regularization coefficient, and $s^2[z_j]$ refers to the variance of presence locations on environmental variable *j*.

The AUC is a frequently employed validation metric for threshold-based models; and in Maxent corresponds to the probability that a presence location chosen at random is ranked by the model as more suitable than a random background location (Phillips and Dudík, 2008). AUC ranges from 0 to 1.0 with larger numbers indicating a better model fit. AICc is an indicator of model fit that penalizes for complexity based on the number of parameters. Further, AICc values are comparable when models are nested versions of one another (smaller AICc values indicating a better model), which is the case for the models developed with 10-m resolution variables (Warren

and Seifert, 2014). However, because AICc is calculated based on the number of presence points, the AICc values between models with the full presence dataset and the chessboard dataset are not comparable. Additionally, we provided the True Skill Statistic (TSS) which is calculated by Sensitivity + Specificity – 1 (Allouche et al., 2006) and thus requires a threshold to calculate: however, the TSS_{max} is found by determining the highest TSS value by testing all thresholds. Higher TSS_{max} values are used as an indicator of better model performance (Jiang et al., 2018). Finally, using the default model as a benchmark we compared each of the models developed with SBBS and the model developed with the resampled variables using the Warren's *I* statistic. Warren's *I* is reported on a scale from 0 (no overlap) to 1 (identical models) and is calculated by summing pairwise differences between the output surfaces at each cell (Warren et al., 2008).

2.3 Results

Generally, compared with the 10-m model developed using the default background sampling area (Figure 2.3), models developed with SBBS (Figure 2.4) found northwestern regions in the county were more suitable whereas the resampled variable model (Figure 2.4) found central areas of the county were less suitable. All models were in agreement that southern areas in the county had the highest suitability. According to Warren's *I*, the least similar model outputs to the model using default background sampling was the SBBS model with the minimum threshold ($I = 0.93$), while the resampled variable model ($I = 0.96$) and the mean ($I = 0.98$) and maximum ($I = 0.99$) threshold models had higher Warren's *I* values. While the default background model found VCI (31.4%), standard deviations of lidar heights (27.5%), the percentage of ground points (16.1%), and the mean height of lidar return heights (9.9%) were variables with largest contributions to the model, restricting the background area resulted in models with which VCI played a smaller contribution, as the other three variables were determined more important in calculating the model's output.

Model diagnostics in our experiment (Table 2.2) indicated that the inclusion of inferences from the broad scale model through SBBS improved the model in comparison with the default background sampling area (chessboard sampling the presence data did not affect these results). AUC (Minimum Threshold: 0.989, Mean Threshold: 0.953, Maximum Threshold: 0.937) and gain (Minimum Threshold: 3.5, Mean Threshold: 2.15, Maximum Threshold: 1.86) for the three SBBS

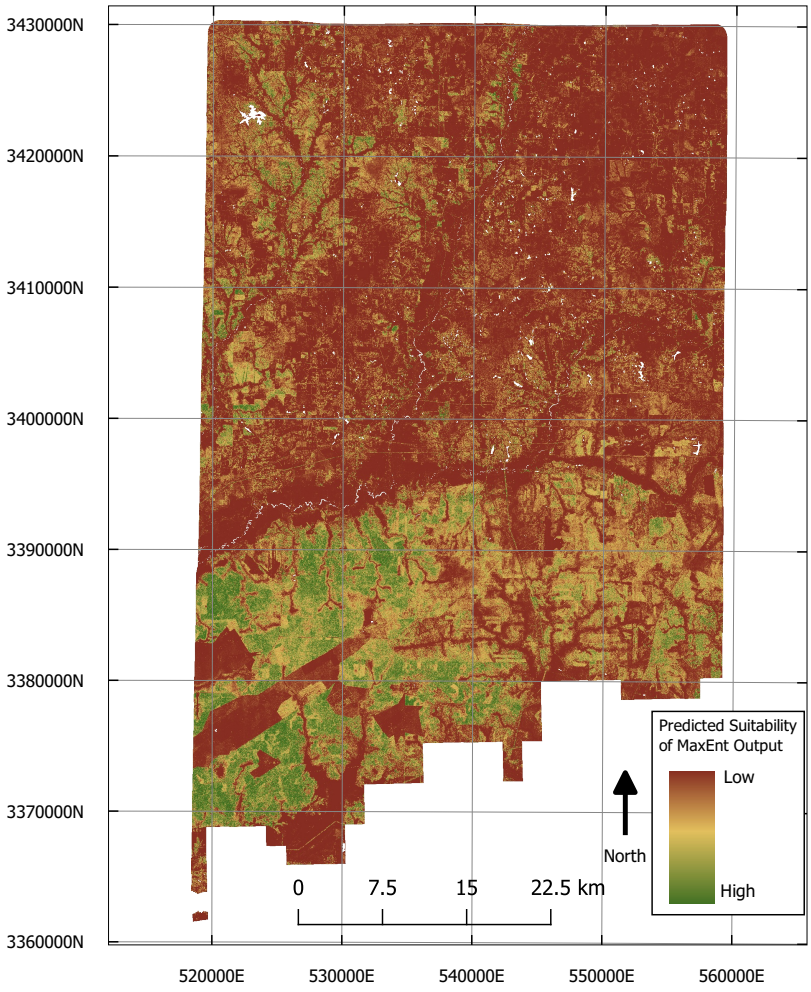


Figure 2.3 Model developed with variables that had a 10-m resolution and sampled background points from the entire study region (i.e., the default background).

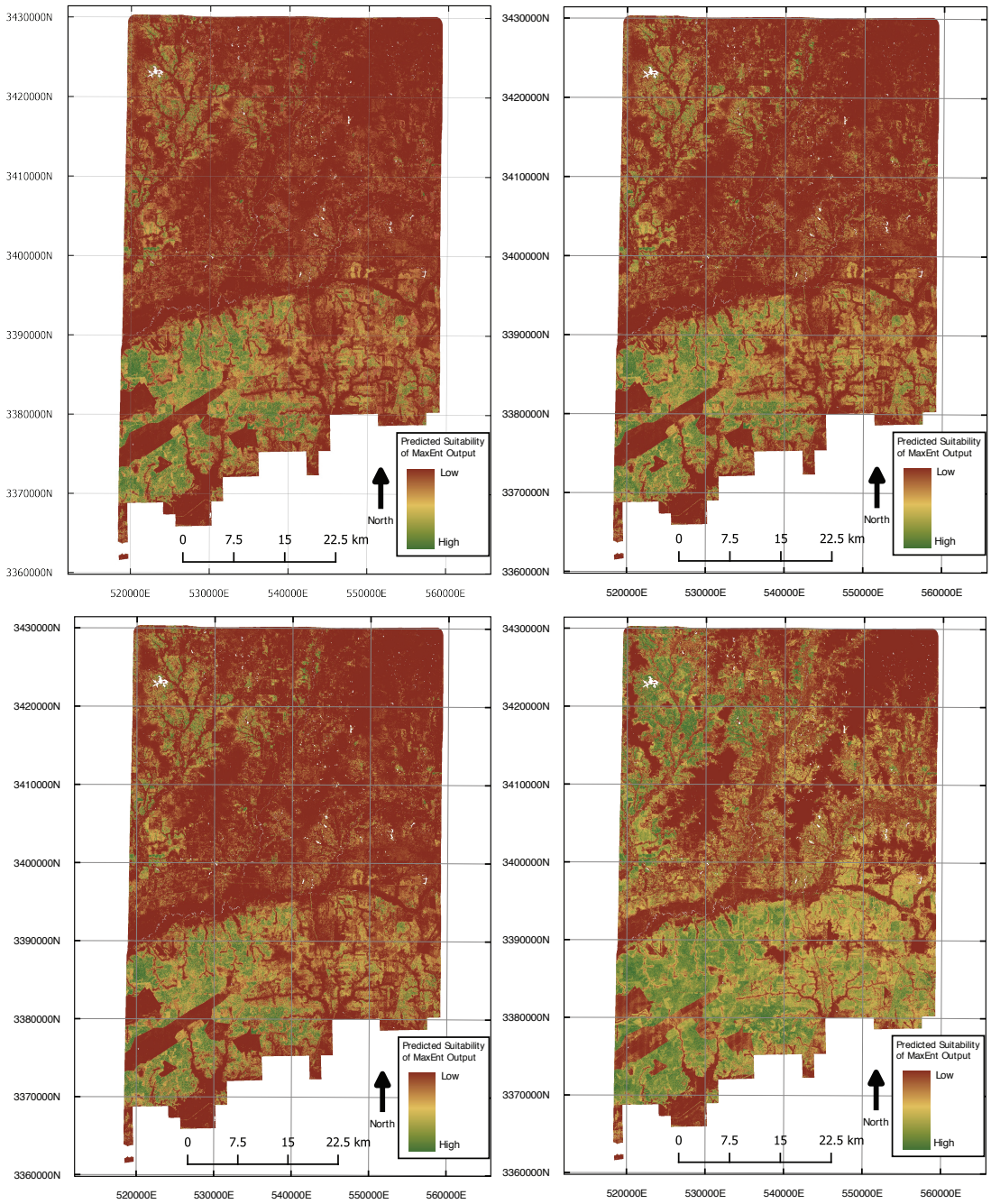


Figure 2.4 Comparison of two approaches of incorporating fine resolution variables in models, scale-based background sampling and resampling environmental variables. Maxent models developed with variables that had a 10-m resolution which also had 30-m variables resampled to 10-m (top left), had a background sampling area determined by maximum threshold (top right), had a background sampling area determined by mean threshold (bottom left), had a background sampling area determined by minimum threshold (bottom right).

Table 2.2 Comparison of model diagnostics: models developed with the scale-based background sampling approach were better than the models developed with the default background sampling areas and the model developed with resampled variables**.**

	AUC	AICc	TSS_{max}	Gain
30-m Model	0.86	67107.57	0.6668	0.91
10-m Model with Default Background	0.883	6939.01	0.72	1.23
Minimum Threshold Model	0.989	7148.73	0.67	3.5
Mean Threshold Model	0.953	6927.7	0.72	2.15
Maximum Threshold Model	0.937	6920.26	0.75	1.86
Resampled Variable Model	0.903	6833.72	0.74	1.49
10-m Model with Default Background and Chessboard Sampling	0.87	3646.32	0.74	1.2
Minimum Threshold Model with Chessboard Sampling	0.993	3727.82	0.66	4
Mean Threshold Model with Chessboard Sampling	0.946	3639.16	0.7	2.07
Maximum Threshold Model with Chessboard Sampling	0.932	3629.78	0.7	1.83
Resampled Variable Model with Chessboard Sampling	0.89	3574.28	0.74	1.58

models were all higher than the AUC (0.883) and gain (1.23) of the model developed using the default background area. AICc values were smaller for the SBBS models developed with the background restricted at the mean (6927.697) and maximum threshold (6920.259) than the AICc values for model developed with the default background area (6939.013), though the AICc was larger for the model developed with the background restricted at the minimum threshold (7148.733). Similarly, TSS_{max} values favored the models calibrated with the background restricted by the mean (0.7225) and maximum threshold (0.7494) to the model developed with the default background area (0.6668). However, the model developed with the minimum threshold had a lower TSS_{max} (0.6683) than the default background model.

The resampled variable model performed better than models developed using the SBBS approach with regards to AICc (6833.718), however the AUC of this model was smaller than the SBBS approach in all cases (0.903). The TSS_{max} for the resampled variable model (0.7424) preferred this model for the minimum and mean threshold model, though the maximum threshold model performed marginally better by this metric (0.7494). Finally, the gain of the resampled variable model (1.49) was lower than the gains of the models developed with SBBS. While comparing the models developed with SBBS, we observed that diagnostic measures of model performance were sensitive to the threshold used to restrict the background sampling area: AUC values favored the minimum threshold; the model gain, AICc, and TSS_{max} favored the models developed with the background restricted by the maximum threshold.

2.4 Discussion

We provide evidence that SBBS can deliver improved results when developing SDMs using spatial data with different resolutions. The 10-m models developed with the SBBS approach had higher diagnostic measures than the 10-m model with the default background sampling area. Further, we observed that models developed using variables at their native resolutions yielded similar and, in some cases, better diagnostic measures compared with a model developed using the conventional approach of resampling variables (Table 2.2). Variations in the results are useful to better understand the SBBS approach and to help determine which threshold yielded the better model. Though widely employed, the AUC is arguably an inflated diagnostic in this experiment particularly because in Maxent the background points (in lieu of absence points) are used to

determine the model's Specificity. Hence, one reason that the model developed with the minimum threshold background sampling area had the largest AUC was that the background points came from the smallest sampling area of the three models, in areas predicted the least suitable by the broader scale model.

Despite being developed with more variables, the resampled variable model had a lower AICc than the models developed with SBBS. Of the models developed using SBBS, the maximum threshold model had the lowest AICc. TSS_{max} was calculated using the presence locations and an equal number of background points (which were sampled through the entire study area) to determine Sensitivity and Specificity at all thresholds. According to TSS_{max}, the maximum threshold model had the highest performance of all the models that we developed. One reason the maximum threshold model may have had larger AICc and TSS_{max} is that background points in this model were sampled from a larger and hence more representative area, while still being restricted by SBBS. Thus, we hypothesize the model had more information about variables (relative to the other two SBBS models) to determine suitability, and consequently the model had better predictive and discretionary ability. Because the minimum threshold model had a higher gain and the largest AUC, and as the maximum threshold model had the smallest AICc and the largest TSS_{max}; there may be tradeoffs regarding the selected threshold, aspects of model performance (e.g. Specificity, Sensitivity), and model diagnostics that should be considered when employing SBBS.

The information that is provided to Maxent in the SBBS approach is fundamentally different than the information provided to Maxent in the resampling context. When a Maxent model is developed using resampled variables (e.g., resampled variables are added to a model), additional relationships are determined between the environmental predictors and a set of presence points that has been partitioned as training and testing data. By comparison, through SBBS the information that is provided to the model is based on relationships between environmental variables and the entire set of presence points, which were calculated by multiple model runs in Maxent. A particular reason SBBS might be preferred to resampling variables is that there is less uncertainty using the variables at their native resolution (Dixon and Earls, 2009), though further work to quantify this uncertainty and compare it to models developed with SBBS might help understand the utility of the approach. Finally, while including resampled variables in models provides further information to explain the presence of a species, SBBS provides the model with

information on species' predicted absence, which is usually lacking when researchers are working with presence-background models.

The SBBS approach is rooted in mechanical aspects of the Maxent model but also carries ecological meaning. By comparing the distribution of environmental characteristics (visualized using the *ggplot* package in R developed by Wickham, 2016) within the background-restricted sampling areas and those within the default background sampling area, we capture the effect that restricting the background sampling area had on background points and model variables. For example, by reviewing the 30-m variables (Figure 2.5), we show the SBBS approach resulted in fewer background points from areas with older forests and longleaf pine land cover, which are known attributes of RCW niche (Smart et al., 2012; Walters et al., 1988). Reviewing the 10-m variables (Figure 2.6) confirms that through the SBBS approach, fewer background points were selected in areas with similar environmental profiles as the presence points (demonstrated using VCI, the most influential variable in the 10-m model). Hence, with SBBS, the distribution of environmental variables at the presence locations and the distribution of these variables at background points were contrasted prior to model development, which resulted in improved model performance. Additionally, based on the niche overlap statistics (Warren's *I*) and differences between model outputs particularly in the northwestern part of the county, we found that limiting the background sampling through SBBS can help identify suitable areas for the species that models with default background sampling might underestimate.

The gain of the models developed with SBBS was larger than the gain of the model with default background sampling and the resampled variable model (Figure 2.7). Another way that researchers have increased the gain in Maxent is by making adjustments to the regularization parameter, which is the penalization term (3) in Equation 2.1. The penalty reduces overfitting in models by ensuring that constraints on the independent variables allow for variability. The penalty is calculated based on the variance of the species' presence on the independent variables, multiplied by the regularization multiplier, which is a parameter that has been demonstrated to improve model performance when tuned (Merow et al., 2013; Radosavljevic and Anderson, 2014). The modification of the regularization penalty, which is based on the environmental characteristics at presence locations, is contrasted with SBBS, which is instead based on

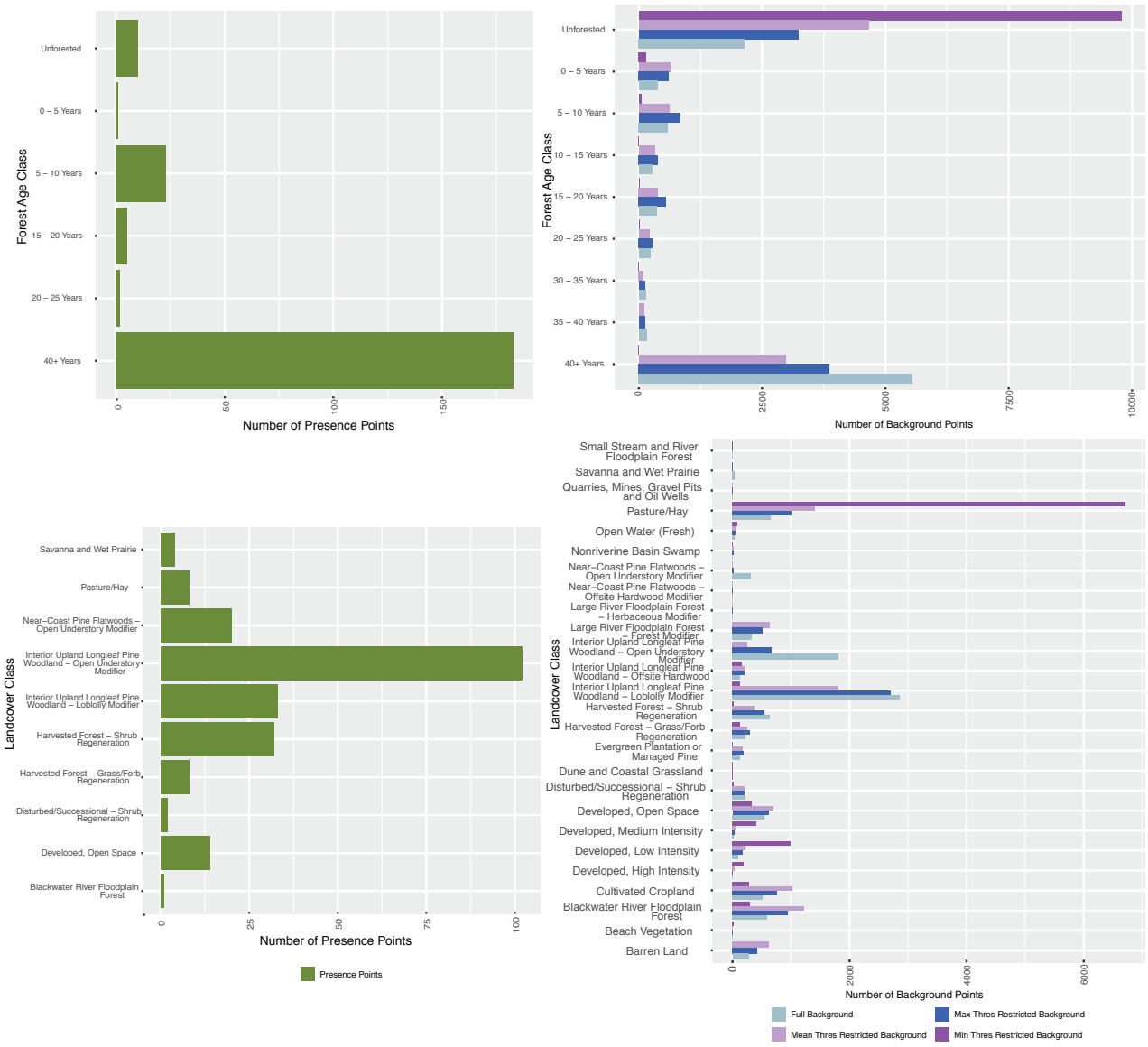


Figure 2.5 Land cover and Forest age classes at presence points (left) and at 10,000 background points randomly sampled within the full background area and the background area partitioned by the three thresholds we tested (right)

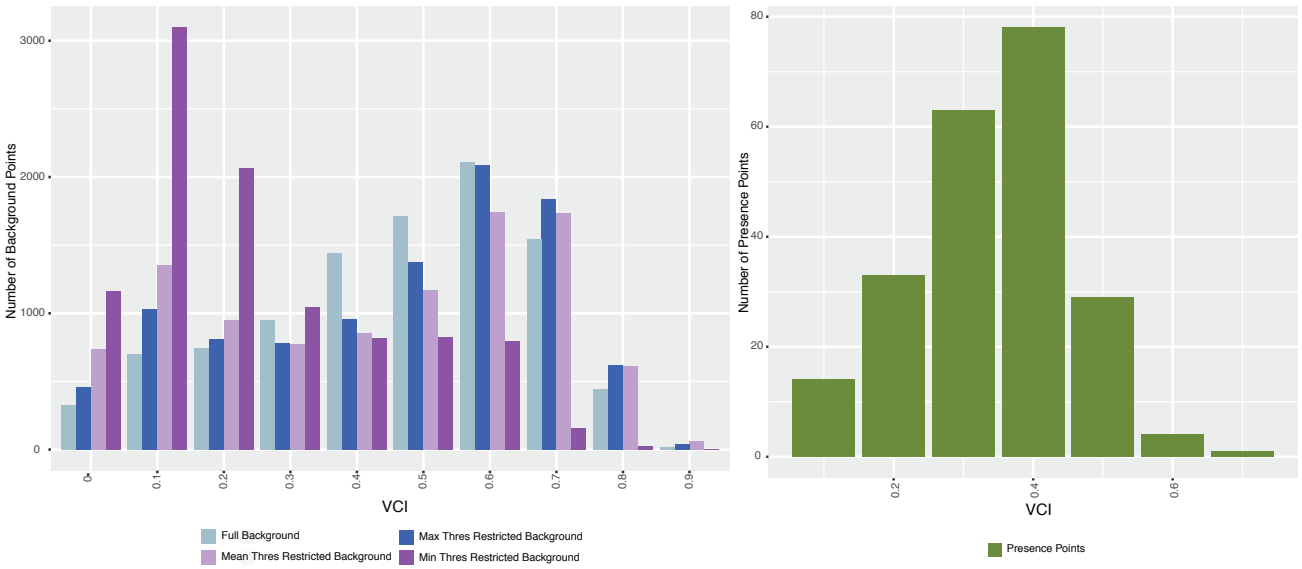


Figure 2.6 Values of Vertical Complexity Index (VCI) at presence points (left) and at the 10,000 background points randomly sampled within the full background area and the background area partitioned by the three thresholds we tested (right)

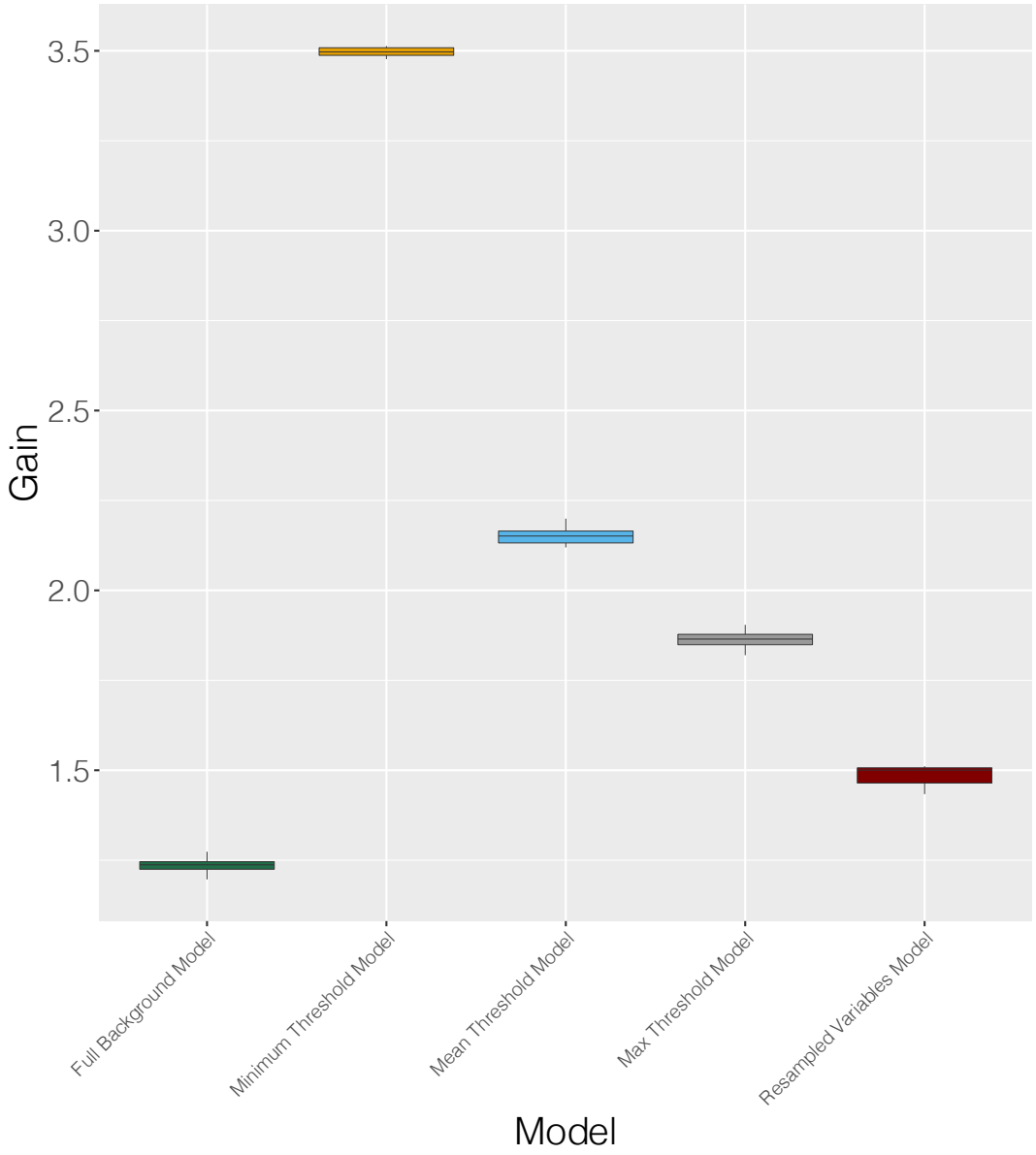


Figure 2.7 Model gain for models developed with 10-m resolution variables. The models developed with the scale-based background sampling approach had higher gain values. Results suggest smaller thresholds produce higher model gains and larger AUC values.

information inferred from the background points. Future work might compare the effects of SBBS with the tuning of the regularization parameter and other methods that have been proposed to optimize Maxent models, as well as explore how these methods may complement one another (Merow et al., 2013; Phillips and Dudík, 2008; Warren and Seifert, 2014).

While SBBS resulted in better models according to conventional measures, future work is needed to confirm the success of the approach. One limitation in our study design was that we only tested this approach with one species, and further research to ensure the approach is successful when modeling species that have key differences in their spatial ecology (i.e., a niche generalist) would help confirm its utility among taxa.

2.5 Conclusions

We introduced a scale-based background sampling approach that can be applied when developing SDMs using spatial data with varying resolutions. We demonstrated the approach improved models from those that sampled background points from the entire study area (the default in Maxent). In our approach, we utilized inferences and knowledge gained by analyzing a species' distribution at a broad scale to improve models developed at a more local scale. By restricting the background sampling area used to develop a Maxent model at a local scale, the environmental characteristics at background points were more distinct from the environmental characteristics at presence points. We showed that this approach resulted in better model diagnostics and evaluated the differences the approach had to the mechanics of the model. Given the growing accessibility to remote sensing, lidar, and other environmental data, this approach might be particularly useful to researchers who are developing SDMs using variables that have different spatial resolutions.

Works Cited

Allouche, O., Tsoar, A., & Kadmon, R. (2006). Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). *Journal of applied ecology*, *43*(6), 1223-1232.

Anderson, R. P., & Raza, A. (2010). The effect of the extent of the study region on GIS models of species geographic distributions and estimates of niche evolution: preliminary tests with montane rodents (genus Nephelomys) in Venezuela. *Journal of Biogeography*, *37* (7), 1378-1393.

Barve, N., Barve, V., Jiménez-Valverde, A., Lira-Noriega, A., Maher, S.P., Peterson, A.T., Soberón, J. and Villalobos, F. (2011). The crucial role of the accessible area in ecological niche modeling and species distribution modeling. *Ecological Modelling*, *222* (11), 1810-1819.

Blanchard, R., O'Farrell, P. J., & Richardson, D. M. (2015). Anticipating potential biodiversity conflicts for future biofuel crops in South Africa: incorporating spatial filters with species distribution models. *GCB Bioenergy*, *7* (2), 273-287.

Bombi, P., & D'Amen, M. (2012). Scaling down distribution maps from atlas data: a test of different approaches with virtual species. *Journal of Biogeography*, *39* (4), 640-651.

Clark, J.S., Bell, D.M., Hersh, M.H., Kwit, M.C., Moran, E., Salk, C., Stine, A., Valle, D. and Zhu, K., (2011). Individual-scale variation, species-scale differences: inference needed to understand diversity. *Ecology Letters*, *14* (12), 1273-1287.

Dixon, B., & Earls, J. (2009). Resample or not?! Effects of resolution of DEMs in watershed modeling. *Hydrological Processes: An International Journal*, *23* (12), 1714-1724.

Elith, J., Kearney, M., & Phillips, S. (2010). The art of modeling range-shifting species. *Methods in ecology and evolution*, *1* (4), 330-342.

Elith, J., Phillips, S. J., Hastie, T., Dudík, M., Chee, Y. E., & Yates, C. J. (2011). A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions*, *17* (1), 43-57.

Fernandes, R. F., Vicente, J. R., Georges, D., Alves, P., Thuiller, W., & Honrado, J. P. (2014). A novel downscaling approach to predict plant invasions and improve local conservation actions. *Biological invasions*, *16* (12), 2577-2590.

Hanula, J. L., & Engstrom, R. T. (2000). Comparison of red-cockaded woodpecker (Picoides borealis) nestling diet in old-growth and old-field longleaf pine (Pinus palustris) habitats. *The American Midland Naturalist*, *144* (2), 370-376.

Holt, R. D. (2009). Bringing the Hutchinsonian niche into the 21st century: ecological and evolutionary perspectives. *Proceedings of the National Academy of Sciences*, *106*, 19659-19665. https://doi.org/10.1073/pnas.0905137106.

Jarnevich, C.S., Young, N. (2015). Using the MAXENT program for species distribution modeling to assess invasion risk., in: Pest Risk Modeling and Mapping for Invasive Alien Species. pp. 65 – 81. https://doi.org/10.1079/9781780643946.0065

Jiang, Y., Wang, T., Wu, Y., Hu, R., Huang, K., & Shao, X. (2018). Past distribution of epiphyllous liverworts in China: The usability of historical data. *Ecology and evolution*, *8* (15), 7436-7450.

Jusino, M. A., Lindner, D. L., Banik, M. T., & Walters, J. R. (2015). Heart rot hotel: fungal communities in red-cockaded woodpecker excavations. *Fungal Ecology*, *14*, 33-43.

Liu, C., Berry, P. M., Dawson, T. P., & Pearson, R. G. (2005). Selecting thresholds of occurrence in the prediction of species distributions. *Ecography*, *28* (3), 385-393.

Martinuzzi, S., Vierling, L. A., Gould, W. A., Falkowski, M. J., Evans, J. S., Hudak, A. T., & Vierling, K. T. (2009). Mapping snags and understory shrubs for a LiDAR-based assessment of wildlife habitat suitability. *Remote Sensing of Environment*, *113* (12), 2533-2546.

Merow, C., Smith, M. J., & Silander Jr, J. A. (2013). A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. *Ecography*, *36* (10), 1058-1069.

Miller, J. R., Turner, M. G., Smithwick, E. A., Dent, C. L., & Stanley, E. H. (2004). Spatial extrapolation: the science of predicting ecological patterns and processes. *BioScience*, *54* (4), 310-320.

Phillips, S. J., Anderson, R. P., & Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, *190* (3-4), 231-259.

Phillips, S. J., & Dudík, M. (2008). Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography*, *31* (2), 161-175.

Radosavljevic, A., & Anderson, R. P. (2014). Making better Maxent models of species distributions: complexity, overfitting and evaluation. *Journal of biogeography*, *41* (4), 629-643.

Smart, L. S., Swenson, J. J., Christensen, N. L., & Sexton, J. O. (2012). Three-dimensional characterization of pine forest type and red-cockaded woodpecker habitat by small-footprint, discrete-return lidar. *Forest Ecology and Management*, *281*, 100-110.

Tererai, F., & Wood, A. R. (2014). On the present and potential distribution of Ageratina adenophora (Asteraceae) in South Africa. *South African Journal of Botany*, *95*, 152-158.

U.S. Fish & Wildlife Service, Costa, R. (2002). Red-cockaded Woodpecker. Fac. Publ. 425.

van Ewijk, K. Y., Treitz, P. M., & Scott, N. A. (2011). Characterizing forest succession in Central Ontario using LiDAR-derived indices. *Photogrammetric Engineering & Remote Sensing*, *77* (3), 261-269.

VanDerWal, J., Shoo, L. P., Graham, C., & Williams, S. E. (2009). Selecting pseudo-absence data for presence-only distribution modeling: how far should you stray from what you know? *Ecological Modelling*, *220* (4), 589-594.

Vierling, K. T., Vierling, L. A., Gould, W. A., Martinuzzi, S., & Clawges, R. M. (2008). Lidar: shedding new light on habitat characterization and modeling. *Frontiers in Ecology and the Environment*, *6* (2), 90-98.

Walters, J. R. (1991). Application of ecological principles to the management of endangered species: the case of the red-cockaded woodpecker. *Annual Review of Ecology and Systematics*, *22* (1), 505-523.

Walters, J. R., Hansen, S. K., Carter III, J. H., Manor, P. D., & Blue, R. J. (1988). Long-distance dispersal of an adult red-cockaded woodpecker. *The Wilson Bulletin*, 494-496.

Warren, D. L., Glor, R. E., & Turelli, M. (2008). Environmental niche equivalency versus conservatism: quantitative approaches to niche evolution. *Evolution: International Journal of Organic Evolution*, *62* (11), 2868-2883.

Warren, D. L., & Seifert, S. N. (2011). Ecological niche modeling in Maxent: the importance of model complexity and the performance of model selection criteria. *Ecological Applications*, *21* (2), 335-342.

Webber, B.L., Yates, C.J., Le Maitre, D.C., Scott, J.K., Kriticos, D.J., Ota, N., McNeill, A., Le Roux, J.J. and Midgley, G.F. (2011). Modeling horses for novel climate courses: insights from projecting potential distributions of native and alien Australian acacias with correlative and mechanistic models. *Diversity and Distributions*, *17* (5), 978-1000.

CHAPTER 3 ASSESSING THE DIFFERENCES BETWEEN SPECIES DISTRIBUTION MODELS DEVELOPED WITH SCIENTIFIC DATA AND THOSE DEVELOPED WITH DATA FROM CITIZEN SCIENTISTS

My use of "we" in this chapter includes my co-authors, Liem T. Tran and Monica Papeş. As first author, I led experimental design, data analyses, and writing the manuscript.

Abstract

Although data gathered by citizen scientists and volunteers can provide information about the distribution of species across wide spatial extents and at fine temporal scales, there are few published examples in which species distribution models (SDMs) developed with citizen science data are contrasted with models developed with data from qualified contributors, such as scientists and wildlife professionals. In our experiment, we compared a model developed with eBird data to a model developed with scientific data, and then analyzed the differences in areas where the two models agreed and disagreed. We developed a regression model to understand how concordance between the eBird and scientific model was affected by different environmental characteristics across the study area, to characterize the type of areas where the eBird model and the scientific model performed similarly. Our results provide insight regarding the ways environmental characteristics of locations in different species presence datasets manifest themselves in differences between model outputs. The average concordance between the two models in all counties within the study area was 80% (and ranged from 65% – 95%); and based on statistical analysis the percent of model concordance between the two models was affected by the proportion of developed land cover per county, the size of the elevation range per county, and the fragmentation of the species' habitat. While there are several benefits to using citizen science datasets for species distribution models, researchers should also be aware of their limitations, biases, and differences compared to scientific data.

3.1 Introduction

3.1.1 *Volunteered Geographic Information for Species Distribution Models*

One important trend in the computational sciences through Web 2.0 has been the increased utility of human sensors as data collectors (Capineri, 2016; Hultquist and Cervone, 2018). When these data have geographic attributes, they constitute Volunteered Geographic Information (VGI), in contrast with more conventional authoritative data. These data types are primarily

distinguished by their collection methods. VGI is collected by volunteers, who are not required to have any training or specific background and often lack a specified protocol, while authoritative data are gathered by professionals with particular credentials who follow established approaches and procedures (Sui, Elwood, & Goodchild, 2012). Moreover, whereas authoritative data are often collected over specific time periods, such as part of research expeditions, VGI production often has an undefined time scale as data repositories are always open for data input (Sui, Elwood, & Goodchild, 2012). Additionally, VGI is typically freely disseminated once available compared with authoritative data, which often have limited accessibility to the public (Sui, Elwood, & Goodchild, 2012). For these reasons, VGI has been proposed to aid data collection for phenomena that have large spatial footprints and large time scales; however, validating data products developed with VGI using authoritative, scientific datasets is particularly important when using VGI data to understand pressing environmental issues (Fonte et al., 2015; Hultquist and Cervone, 2018).

Species presence datasets provided through volunteers and citizen scientists, under the umbrella of VGI, can be valuable resources for conservation science by providing decentralized, comprehensive biodiversity data over broad geographic extents and long periods of time. Launched in 2002, eBird is currently one of the largest citizen science projects (Sullivan et al., 2009). Through the eBird platform, volunteers upload lists of species they encounter with their locations, thereby becoming active participants in biodiversity conservation as providers of commonly used species presence data. The eBird project has been recognized as the reason that avian species are well-represented in large species occurrence record repositories, like the Global Biodiversity Information Facility (Amano et al., 2016). In under-researched areas and data-poor regions where the lack of technical infrastructure has previously led to gaps in knowledge about many species, eBird, and more broadly, citizen science projects, have emerged as possible solutions to these data limitations (Amano et al., 2016).

Although provided by volunteers, species presence datasets from eBird have been employed in published scientific research, including studies to understand the distribution of invasive species (Trautmann et al., 2012), explore aspects of species evolution (McCormack et al., 2010), track distribution changes prompted by climate change (Hurlbert and Liang, 2012), and monitor populations of rare and endangered species (Clark, 2017). While eBird has quality control

mechanisms that are rooted in both scientific knowledge and machine learning (Kelling et al., 2013; Sullivan et al., 2009), there is limited published research that assesses the quality of models developed with citizen science datasets relative to those developed with scientific datasets.

The quality of VGI for environmental research has been explored in the context of monitoring algal blooms (Hultquist and Cervone, 2018), assessing the reliability of volunteered land cover classifications (Comber et al., 2013), measuring air temperature using smart phones (Muller et al., 2015), and recording radiation following the Fukushima disaster (Hultquist and Cervone, 2018). Few studies have specifically focused on the validation and comparison of citizen science data for species distribution models (SDMs), which are statistical models that estimate potential distributions of species across a study area using species presence locations and associated environmental data. Tye et al. (2016) found that SDMs developed with citizen science data and those developed with data submitted by wildlife professionals were comparable in quality; however, they noted models developed with citizen science datasets seemed to favor developed landscapes.

Our research was focused on identifying spatial factors that led to concordance (both models agreed on the suitability and unsuitability of an area) and discordance (models disagreed about the suitability or unsuitability of an area) between the model outputs from eBird and from the authoritative –henceforth scientific– datasets. In our experiment, we investigated the differences in SDMs developed with eBird and scientific presence points by first calibrating the models with a model threshold (i.e., a value used to reclassify the suitability model to a binary surface). We then analyzed the two binary suitability models by reviewing the environmental characteristics that resulted in differences between the two models. Prompted by the growing utility of species presence data provided by volunteered wildlife projects, our research questions were: (i) how do SDMs developed with citizen science presence data (i.e., VGI-based SDMs) perform compared with those developed with presence data provided by scientists and professionals? and (ii) In general, what factors or conditions influence the concordance/discordance of VGI-based SDMs with respect to scientific models? (i.e., in what conditions do VGI-based SDMs perform reasonably well?).

3.1.2 *Crested Caracara*

We modeled distributions of the crested caracara (*Caracara cheriway*), a threatened non-migratory bird of prey that occurs in an isolated population in southcentral Florida (Morrison and Humphrey, 2001). Historically the species was part of the dry prairie ecosystem, which has undergone significant land cover changes (Morrison and Humphrey, 2001). Caracaras are territorial and, in contrast to most avian species, spend significant time on the ground, making them sensitive to local land cover conditions (Morrison, 2001). Hence, though they may not prefer them, forests are hospitable land cover for them. While their preferred natural habitat included marshes, grasslands, and prairies, increased urbanization and agricultural/pasture expansion mean the Florida caracara population is now often found in areas with scattered trees, short/ground vegetation, and with a minimal understory or shrub layer—areas such as pastures, cropland, and cattle ranches. The scavenger species is also found feeding on carrion along roadsides, waste facilities, and near slaughterhouses (Morrison, 2001).

3.2 Materials and Methods

3.2.1 *Species Presence Data and Environmental Data*

The scientific dataset used in our study included 225 presence records for the crested caracara with input dates ranging from 1978 – 2013 provided by the Florida Natural Areas Inventory (FNAI), a non-profit research organization at Florida State University that maintains state biodiversity inventories. The original data were provided in the form of buffer polygons used to represent spatial uncertainty, but we used the centroid of these polygons as presence locations for use in Maxent. The eBird dataset was downloaded from the Global Biodiversity Information Facility. Because of the volume of data available across all years, we only chose data points that were recorded in 2017, resulting in 2,831 presence points (Figure 3.1). To confirm that the spatial footprint of observations was similar to those in previous years, we created heatmaps using presence data for every year from 2013 – 2016 (Appendix 3.1). Table 3.1 outlines the variables used to develop the models in our experiment. Four of the independent variables used in the study were provided by the US-Gap Analysis Program (elevation, slope, aspect, and land cover), and had a native resolution of 30-m. The Florida Department of Agriculture and Consumer Services

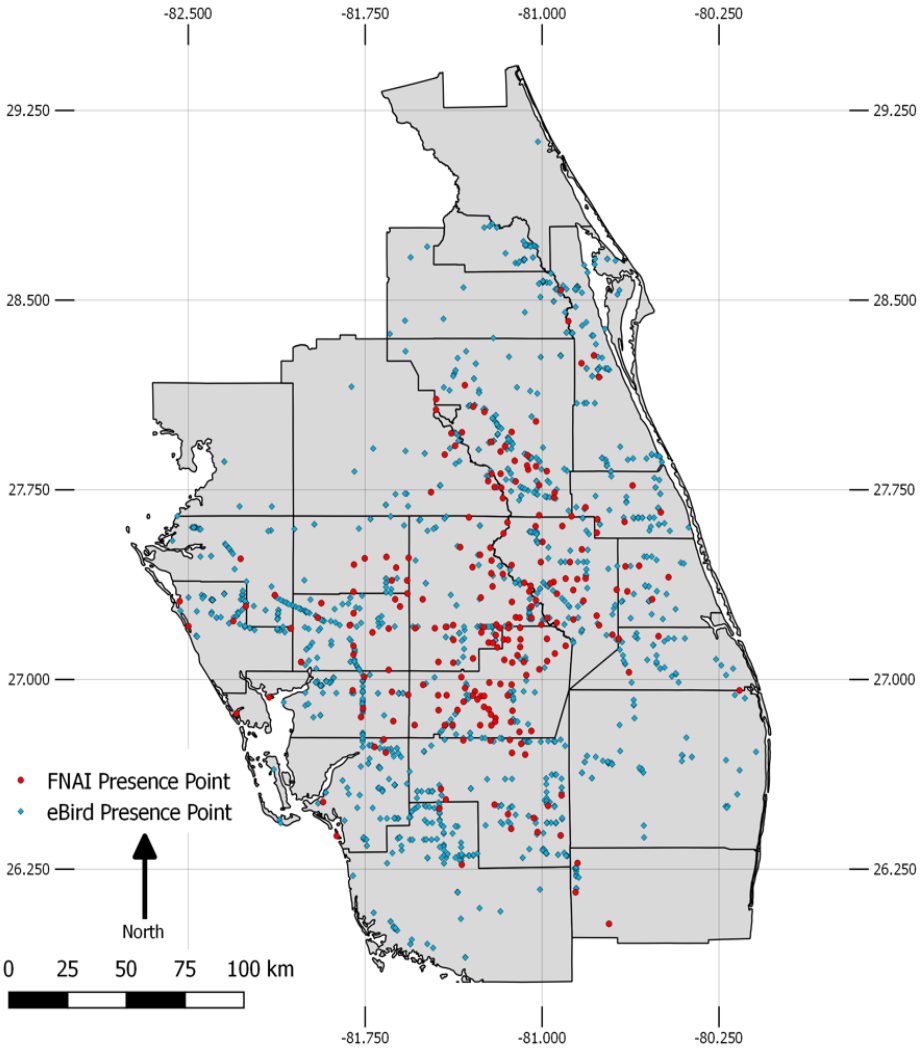


Figure 3.1 Distribution of eBird and scientific presence points

Table 3.1 Variables used in Maxent

Variable	Variable Description	Resolution	Data Source
Forest Age	Categorical raster corresponding to biomass stand age classified in 10-year intervals, derived from Landsat imagery	30-m	Florida Department of Agriculture and Consumer Services
Land Cover	GAP National Terrestrial Ecosystems 2011 includes detailed vegetation and land cover information for the entirety of the United States, representing 590 land cover classifications nationally and 102 land cover classifications within Florida	30-m	USGS Gap Analysis Program
Elevation	Continuous raster, digital elevation model derived from the National Elevation Dataset	30-m	USGS Gap Analysis Program
Slope	Dataset derived from the elevation dataset which reflects the slope (i.e., rise over run) of the area, reported in degrees	30-m	Derived from Elevation raster
Aspect	Categorical raster derived from elevation dataset which corresponds to the cardinal direction the area within each cell is facing (i.e., North-facing)	30-m	Derived from Elevation raster

provided a raster of forest age that also had a native resolution of 30-m.

3.2.2 *Methods*

First, we developed two SDMs using the same set of environmental variables, one model used the eBird presence dataset, while the other used the presence dataset from the FNAI. Models were developed using the Maxent program with the cross-validation approach (with 10 model runs). The cross-validation method was chosen because it uses all presence data provided, randomly selected by Maxent as either a testing point or training point during model development (Phillips, Dudik, & Schapire, 2004). We performed our analysis using a commonly-employed product of Maxent, the logistic output, a raster (which has the same resolution and extent as the model's input variables) with values that range from 0.0 – 1.0. Raster cells in the logistic output correspond to the similarity of environmental characteristics of each location (raster cell) to the environmental characteristics at locations where the species was recorded present (Phillips, Dudik, & Schapire, 2004). Following model development, we compared models and analyzed environmental factors that led to concordance and discordance between corresponding cells of the two models using the *SDMtools*, *raster*, *rgdal*, and *spatialEco* packages in R.

We assessed the performance of the models using the gain of the models, the Area Under the Receiver Operating Characteristic Curve (AUC), the sample size corrected Akaike Information Criterion (AICc), and the maximized True Skill Statistic (TSS_{max}). In Maxent, the gain of the model is central to the machine learning process used to build the prediction surface. When the gain value is exponentiated, it is the likelihood ratio of a typical presence points to a typical background point (thus larger gains indicate models with more discriminatory power). A frequently employed validation metric for threshold-based models, the AUC in Maxent corresponds to the probability that a presence location chosen at random is ranked by the model as more suitable than a random background location (Phillips & Dudik, 2008). AUC ranges from 0.0 – 1.0 with larger numbers indicating a better model fit. AICc is a useful model diagnostic because it is an indicator of model fit that penalizes for complexity based on the number of parameters, with smaller AICc values indicating a better model fit. Additionally, we provided the TSS_{max}. The True Skill Statistic is calculated by Sensitivity – Specificity – 1 (Allouche, Tsoar, & Kadmon, 2006; Ruete & Leynaud, 2015), and thus requires a threshold to calculate. However, the TSS_{max} is

found by determining the highest TSS value by testing all thresholds (Ruete & Leynaud, 2015). Higher TSS_{max} values indicate better model performance (Ruete & Leynaud, 2015).

Following the initial model development outlined above, we created a binary output from the scientific model (converting numeric values of cells in the model from to a suitable/unsuitable classification) to use as a benchmark to compare the eBird model. We binarized the scientific suitability model using a threshold provided by Maxent at which 10% of training presence data were omitted (i.e., 10% of the presence points were in cells that had lower suitability values than the threshold and were thus omitted in cells classified as unsuitable). Then, our analysis was divided into two steps employed in an iteration for every county in the study area: a calibration and a validation step. For each county, we isolated the eBird model output in county *i* and extracted the eBird model within the remaining part of the study area, calibration *i* (all counties in the study area – county *i*). Within calibration *i*, we tested the percentage of concordant cells between the benchmark scientific model and the eBird model when reclassified at every threshold from 0 to the maximum value of the models (by intervals of 0.05) to determine which threshold resulted in a binary eBird model most similar (i.e., highest concordance) to the benchmark scientific model, which we will refer to as the optimal threshold. Thus, we calibrated the eBird model in a majority of the study area to the benchmark scientific model by finding a threshold for the eBird model that led to the highest model concordance with the scientific benchmark model.

In the validation step we applied the optimal threshold to the eBird model output within county *i* and then evaluated its similarity, measuring concordance and discordance, with the benchmark scientific model. Thus, we used the threshold to help conform part of the eBird model left out of the calibration to the scientific model so that we could assess the reliability and consistency of the threshold, and the differences in model concordance, for every county in the study area. For each county *i* we reclassified the eBird output using the optimal threshold (determined in calibration *i)* and reported the percentage of concordant cells (corresponding cells that both models predicted suitable and those that both models predicted unsuitable) between the two reclassified outputs within each county. The concordance between the two models in each county, the percentage of cells with shared suitable/unsuitable values, was the basis for our comparison of the eBird and scientific models to understand the variation of model performance across the study area.

To analyze environmental trends associated with areas where the models agreed and disagreed, we explored the environmental characteristics in Area (A) both the eBird and scientific models determined as suitable (concordance), Area (B) both the eBird model and the scientific model determined as unsuitable (concordance), Area (C) the eBird model determined as suitable but the scientific model disagreed (discordance), and Area (D) the scientific model determined as suitable but the eBird model disagreed (discordance). We reported histograms of environmental characteristics in six counties, the three counties with the highest concordance and three with the lowest concordance between the two models, to understand how these attributes contributed to differences between the two models in these counties.

We performed additional statistical analyses to better understand the environmental conditions that led to the variation in model concordance across the study area and to answer our second research question regarding areas where the eBird models perform well. We first explored the relationship between the percentage of concordant cells between the two binarized outputs per county and the environmental characteristics related to the model variables of the county. Further, we reasoned that factors related to fragmentation of the species' habitat might also influence the concordance between the two models because human and wildlife interactions are more frequent in areas with fragmented habitats, possibly leading to more citizen science presence points (Stewart et al., 2003). Thus, because the crested caracara is often found in cultivated cropland and pastures (based on knowledge about the species and the distribution of scientific presence points), we isolated that land cover in our land cover dataset to derive habitat fragmentation statistics per county (using metrics provided by *SDMtools*).

We then explored correlations between the percentage of concordant cells between the two models per county and the environmental characteristics, including habitat fragmentation measures, per county. Based on those correlations, we compiled a set of variables to develop a regression that incorporated both fragmentation measures and variables used in the Maxent model. With the regression model, we could estimate the variance of the dependent variable (percent of concordant cells between the two models) that is attributed to a set of independent variables. To complement the results from the regression model with insight on the effect of each independent model variable on model concordance per county, we additionally reported partial correlations (Agresti and Finlay, 1997). Partial correlations informed us about the relationship between the

percent of concordant cells per county and each environmental variable in the model after removing the effects from other variables. Further, partial correlations are determined after standardizing the variables, in terms of variance, instead of units the variable was measured in which is how regression coefficients are determined. This allows for direct comparison. We also reported the squared partial correlation, which represents the percent of variance in the model concordance per county attributed to each variable. Results from the correlation and regression models could provide information regarding which general factors and environmental characteristics influence the mismatch between models developed with the two different presence datasets, and to what extent.

3.3 Results

According to conventional measures of Maxent model performance including AUC, AICc, TSS_{max}, and model gain, the model developed with scientific presence data yielded better results compared to the model developed with eBird presence data (Table 3.2). The logistic outputs displayed similar spatial patterns (Figure 3.2), although the scientific model suggested higher suitability in the interior parts of the state, while the eBird model suggested higher suitability along the coast. Another general difference in the logistic outputs is that the eBird model predicted more areas suitable than the scientific model, while the scientific model attributed higher suitability values to a more limited area. This finding is related to the differences between the gain, and thus discriminatory power, of the two models. Maxent provides the contribution of each input variable to the model's solution in a percentage to help assess the relative importance of the environmental predictors. The variables with the largest contribution to the eBird model were land cover (53.7%), elevation (31%), and forest age (12.1%). In the scientific model land cover (72.9%) and elevation (25.7%) contributed the most, while forest age (0.6%) did not substantially contribute to the model's solution.

The calibration phase, during which we determined the optimal threshold for the eBird model, yielded consistent results for every county in the study area: The highest concordance with the benchmark scientific model was found when the eBird model in calibration *i* was thresholded at a suitability value of 0.50 (Figure 3.3). When that threshold was applied in the

Table 3.2 Model Diagnostics

	Scientific Model	Full eBird Model
AUC	0.763	0.718
AICc	8102.354	96159.69
TSS_{max}	0.3408	0.1522
Gain	0.5134	0.3257

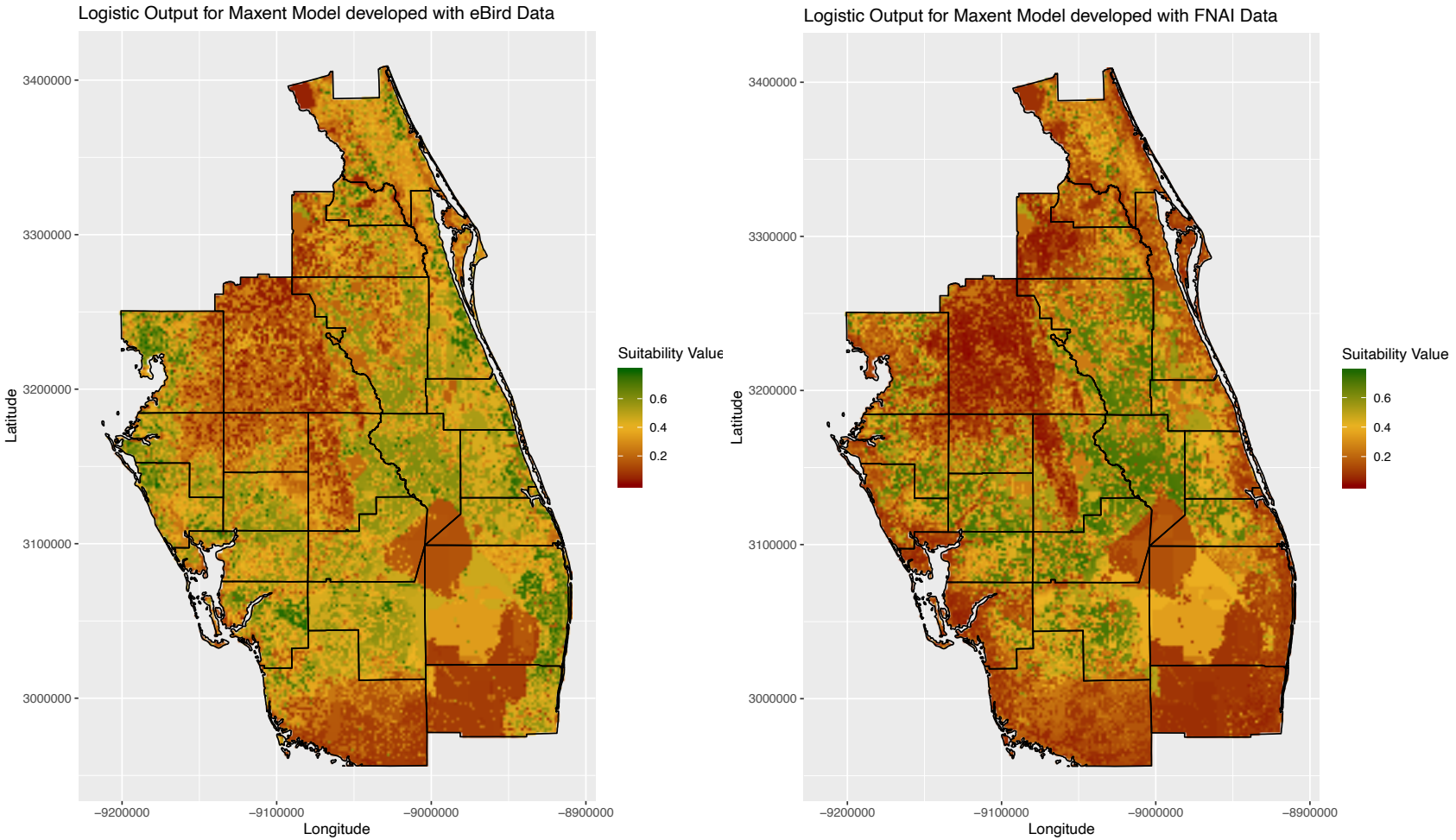


Figure 3.2 Maxent logistic outputs for model developed with eBird presence points (left) and model developed with scientific presence points (right)

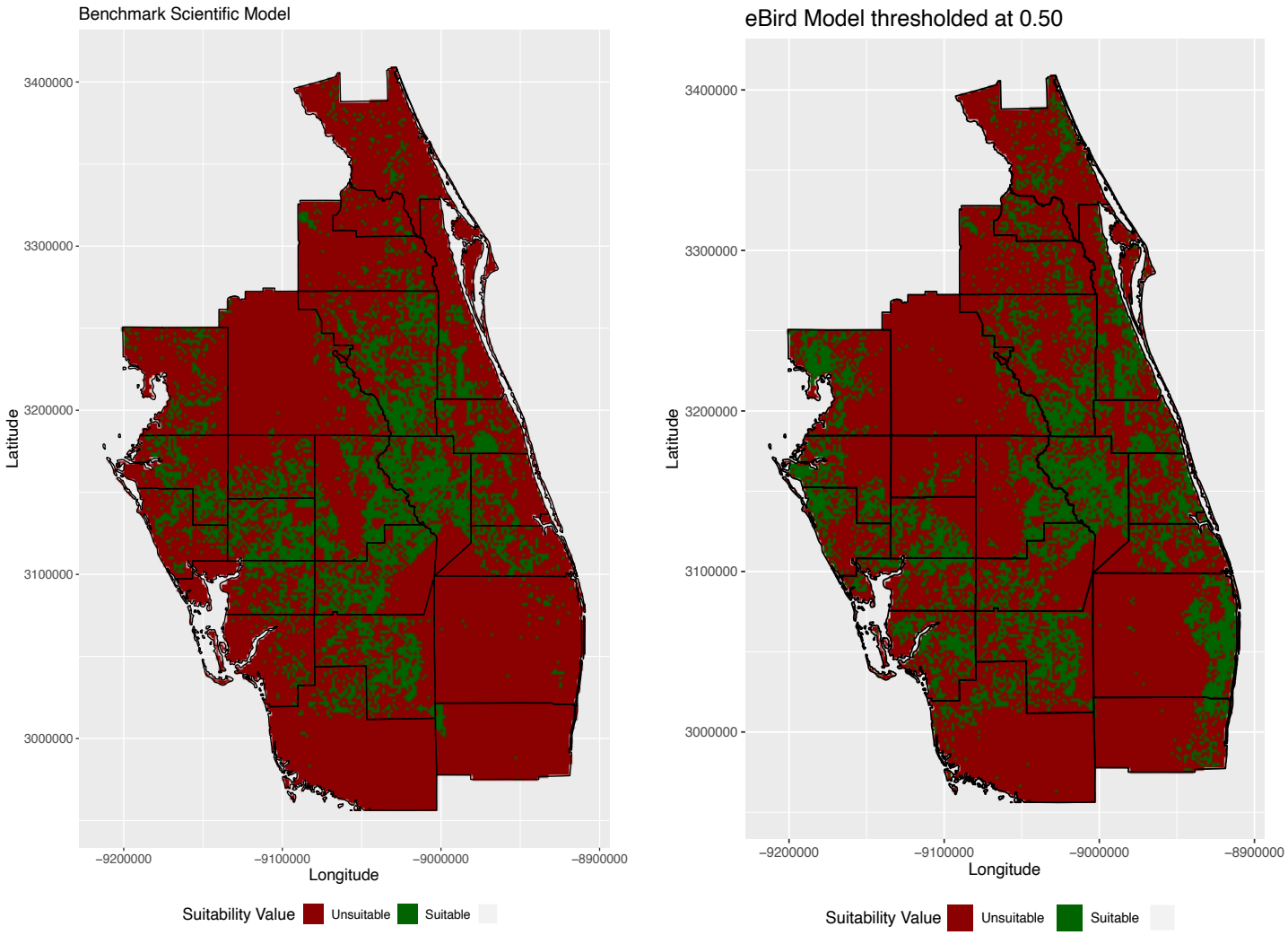


Figure 3.3 Binary suitability models developed with scientific data (left) and eBird data (right)

validation phase, however, the percentage of concordant cells between the two models per county ranged from 65% – 95% concordance, with an average model concordance per county of 80% (Figure 3.4). Geographically, coastal counties along the eastern part of the study area had greater discordance between the two models than counties in the southern and western part. We reported the environmental characteristics within six counties in the study area: the three counties with the highest concordance—Polk, Collier, and Hendry counties (Figure 3.5) —and three counties with the highest discordance—Sarasota, Hillsborough, and Lee counties, which are found along the eastern coast of the study area (Figure 3.5 cont'd).

As land cover was the most influential input variable in modeling crested caracara suitability for both models, it was particularly important to analyze the spread of land covers within Areas A, B, C, and D in the six counties. From these histograms, we observed cells classified as developed land cover (open space-, low-, medium-, and high- intensity development) were generally predicted to be suitable by the eBird model, while the scientific benchmark model disagreed. We additionally observed that land cover classes including pasture/hay, cultivated cropland, hydric hammock, swamps, and prairies were sometimes predicted suitable by the scientific model, but concurrently predicted unsuitable by the eBird model. Particularly given these areas include environmental attributes that the species is known to prefer, we hypothesized that some areas predicted suitable by the scientific model but unsuitable by the eBird model were perhaps so pristine that citizens were less likely to access them to submit eBird observations. Further, we observed that these areas also included older forests stands and places with a relatively high elevation (hence less accessible to volunteers given thick understories and higher terrain). Thus, our results suggest that volunteers to eBird, in addition to being more commonly found in anthropogenic landscapes, might also avoid higher elevations and thick understories of old-growth forests, compared with wildlife professionals who might target specific areas known to be suitable for the species and are also expected to sample across environmental gradients. These differences in data collection manifested themselves in SDMs.

Based on the histograms of environmental characteristics in Areas A, B, C, and D, we looked for Pearson correlations (Figure 3.6) for candidate variables to develop a regression. We corroborated the findings from the histograms and found that the percent of concordant cells between the two models and the proportion of developed land cover by county had a negative

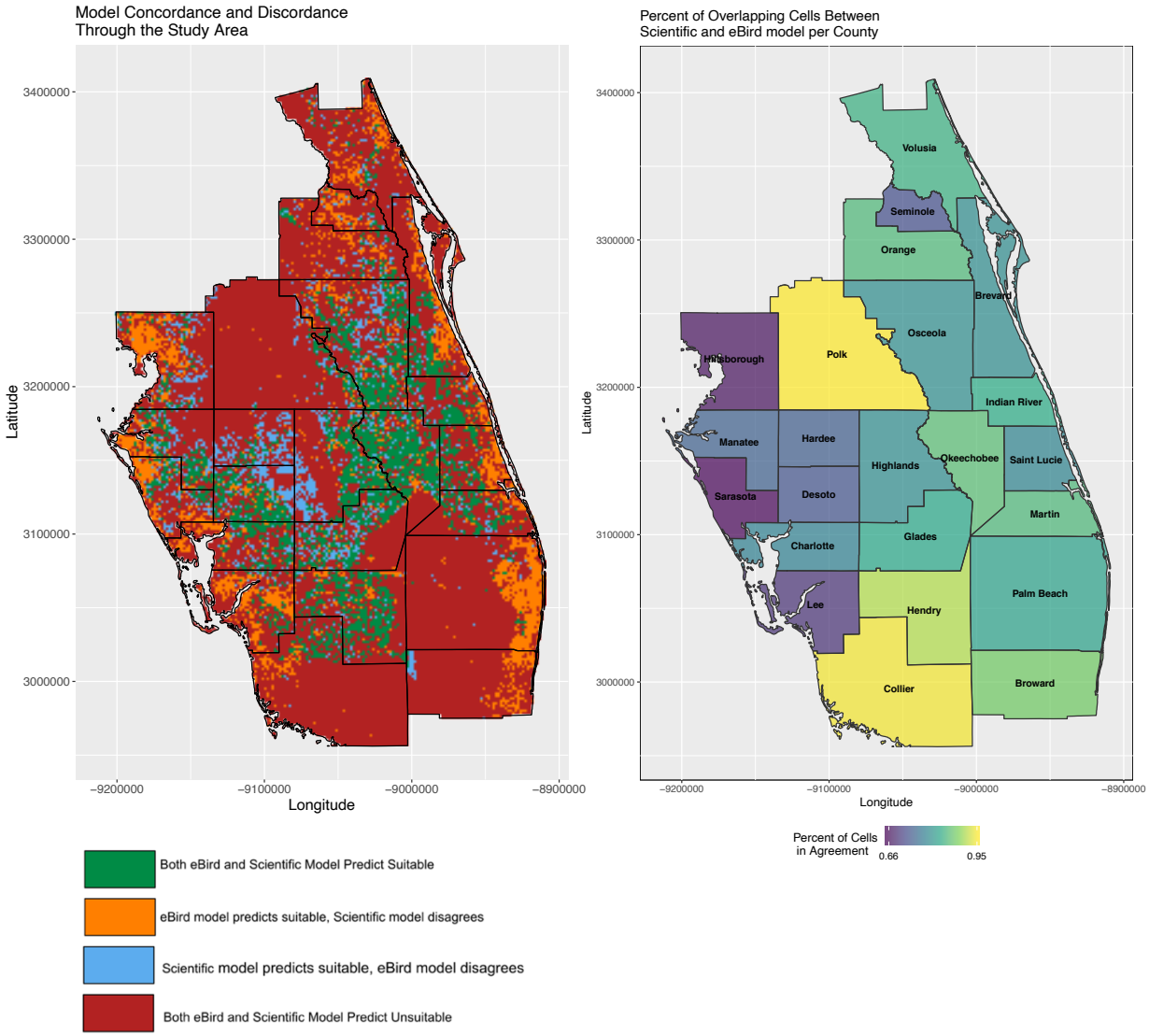


Figure 3.4 Output of overlaid eBird and scientific model (left) and percent agreement between the two models per county (right)

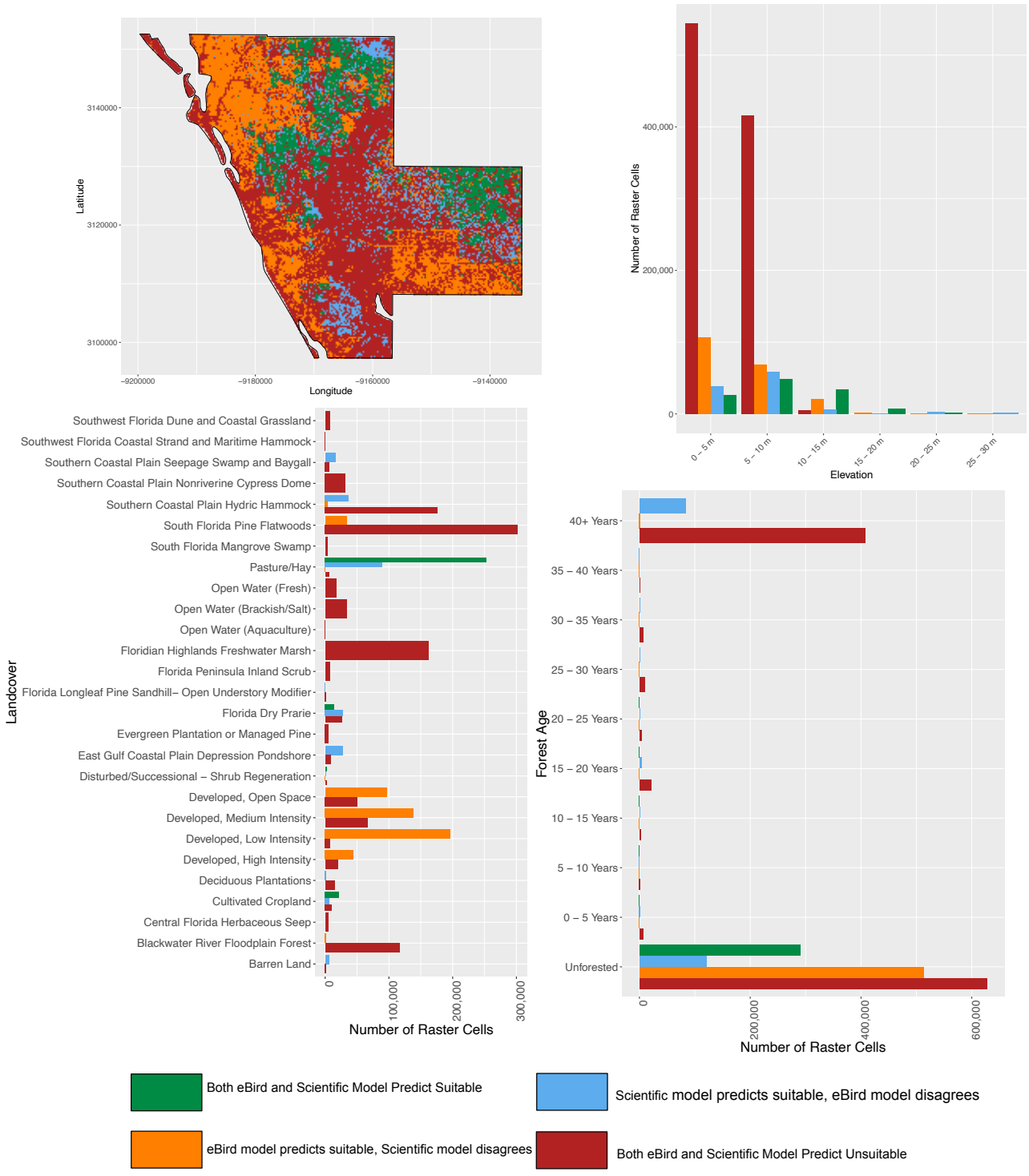


Figure 3.5 Map of model agreement and distribution of environmental characteristics in Areas A, B, C, and D in Sarasota county, where there was high discordance between the scientific and eBird model

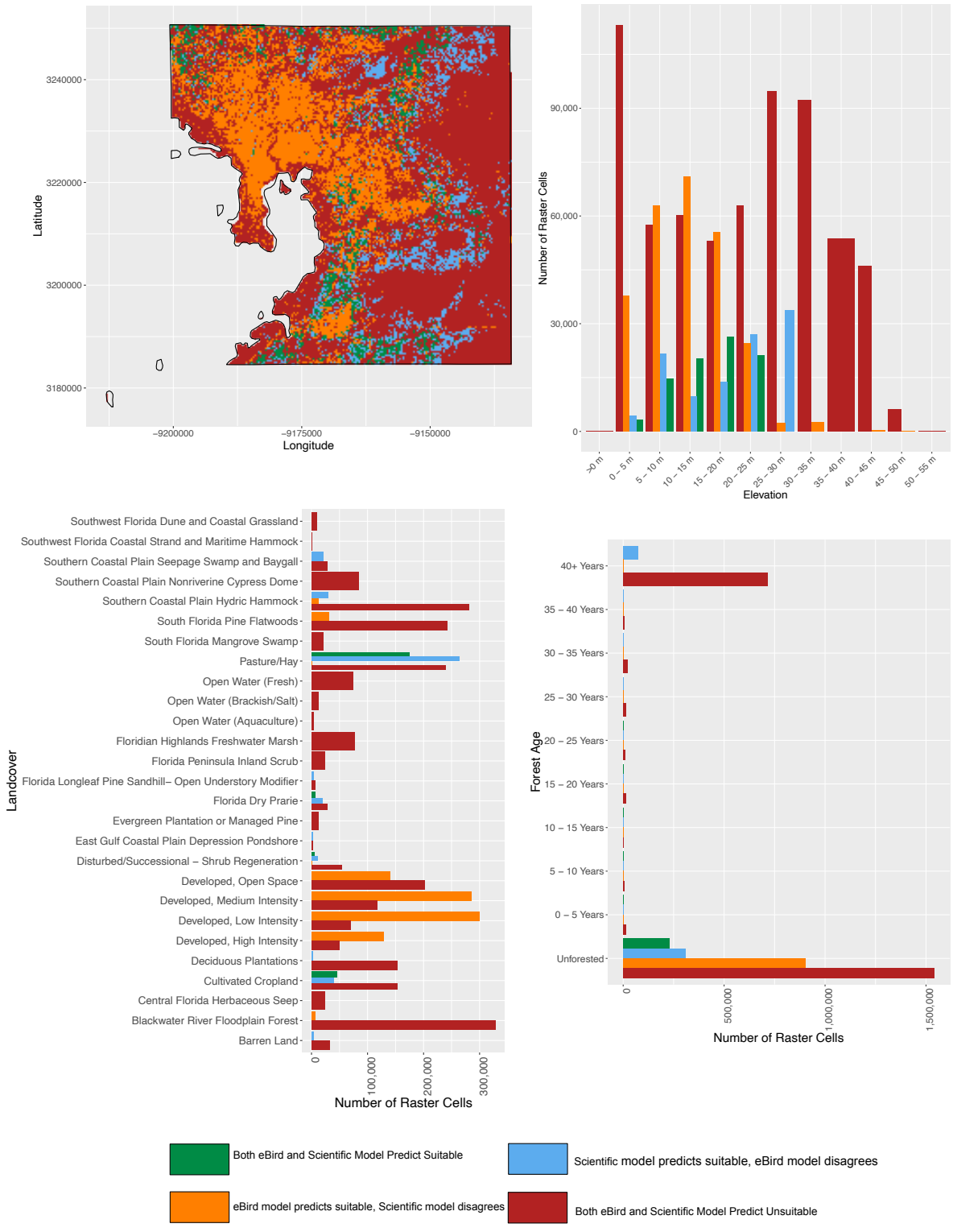


Figure 3.5 (cont.) Map of model agreement and distribution of environmental characteristics in Areas A, B, C, and D in Hillsborough county, where there was high concordance between the scientific and eBird model

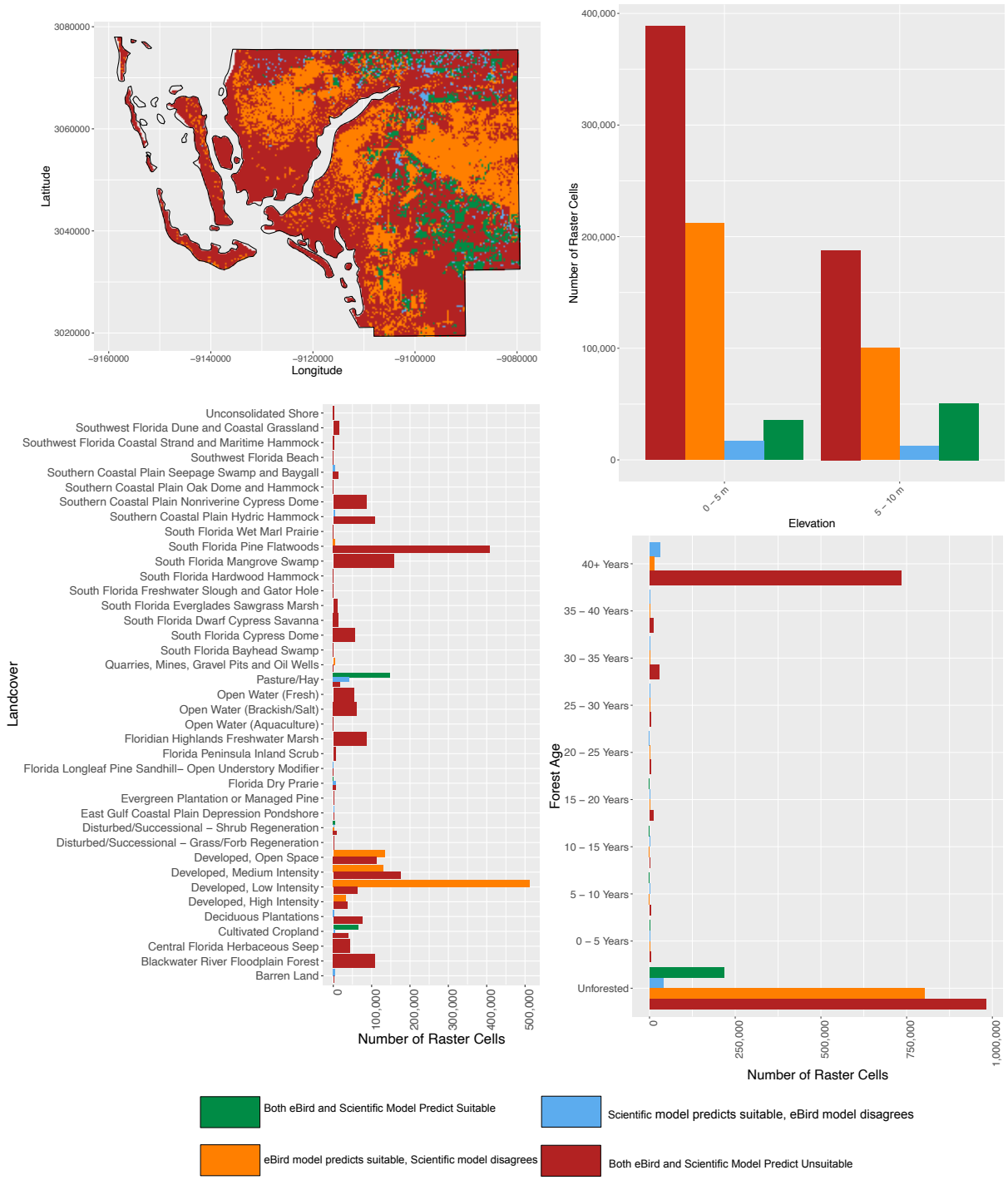


Figure 3.5 (cont.) Map of model agreement and distribution of environmental characteristics in Areas A, B, C, and D in Lee county, where there was high concordance between the scientific and eBird model

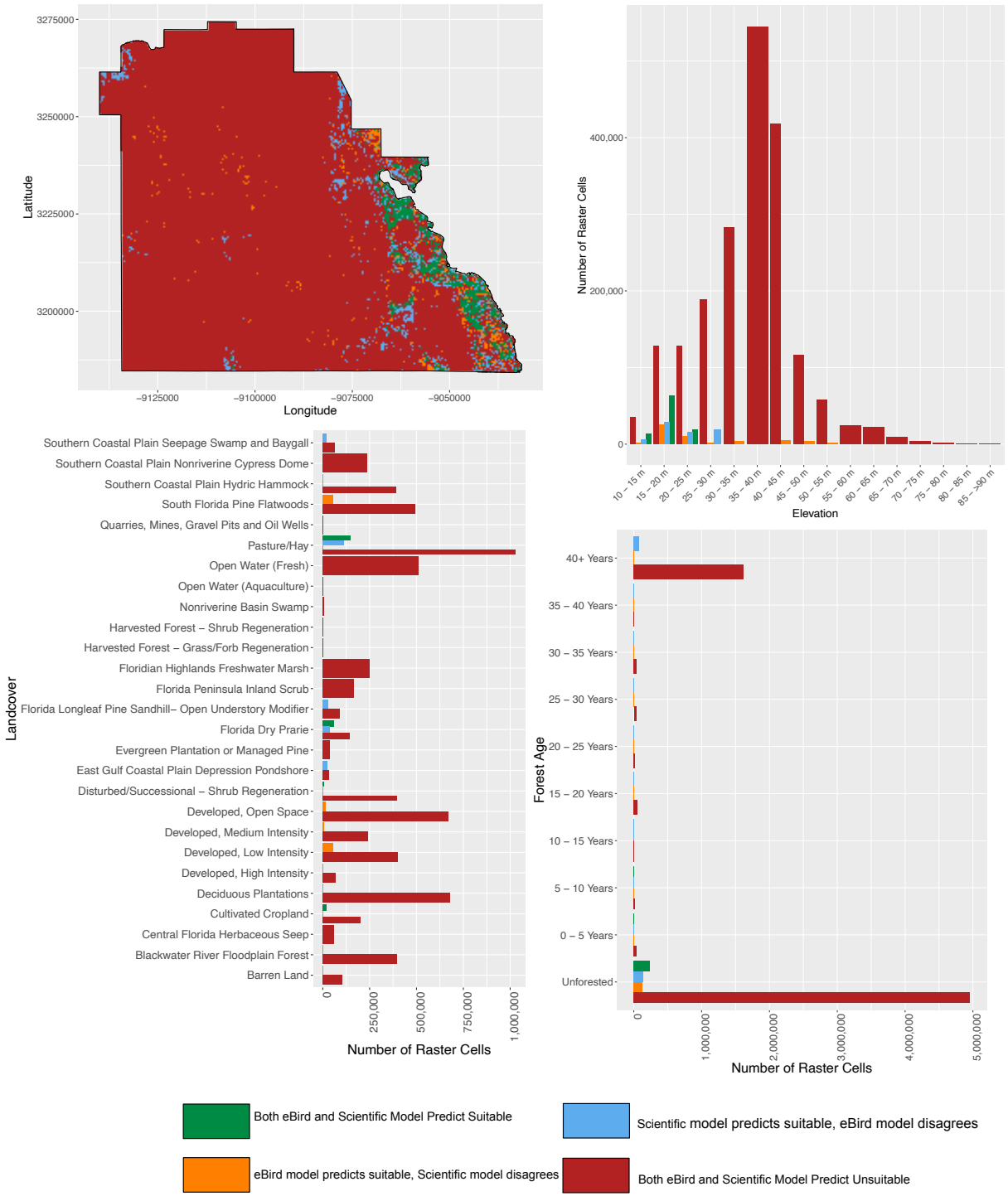


Figure 3.5 (cont.) Map of model agreement and distribution of environmental characteristics in Areas A, B, C, and D in Polk county, where there was high concordance between the scientific and eBird model

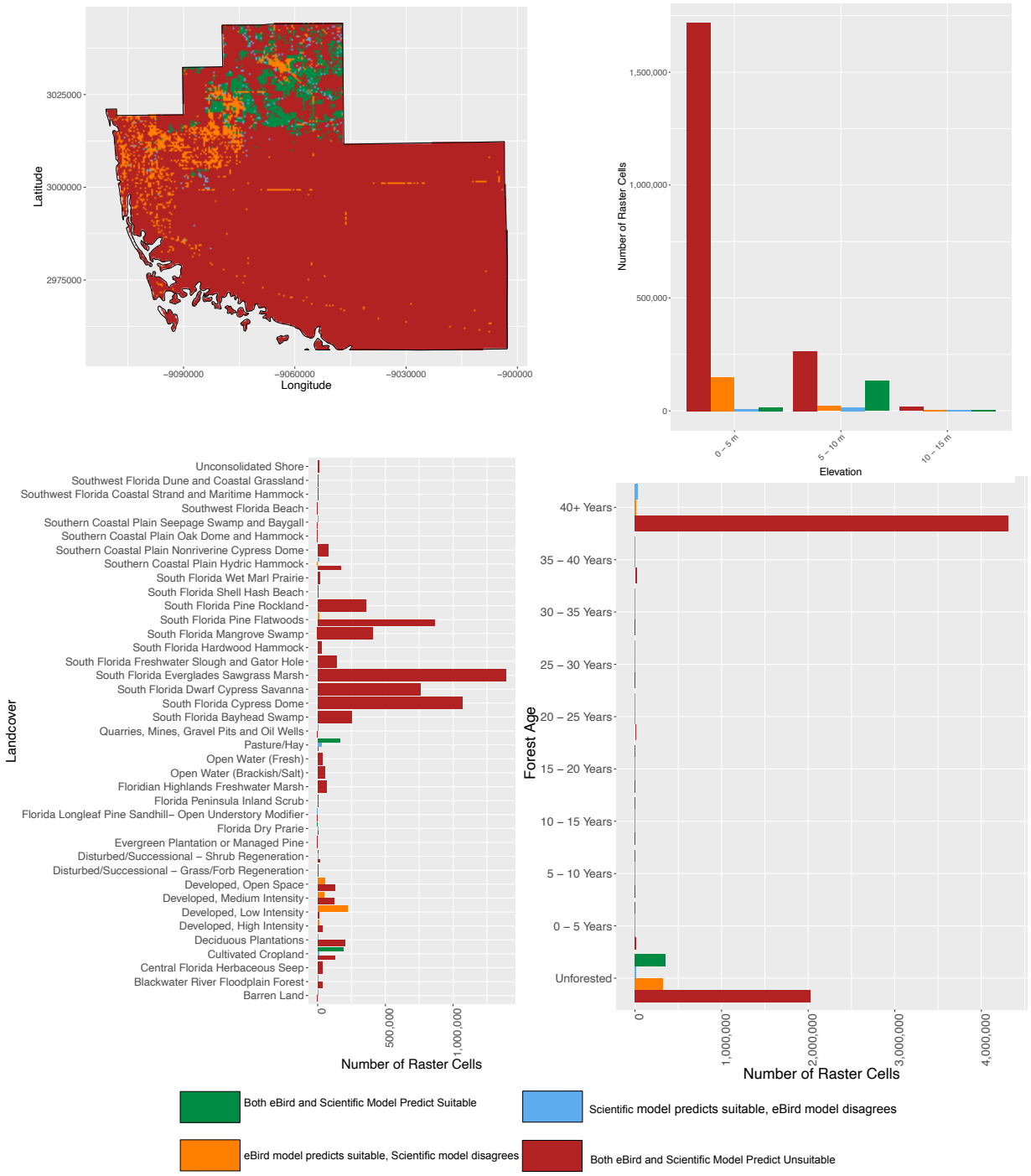


Figure 3.5 (cont.) Map of model agreement and distribution of environmental characteristics in Areas A, B, C, and D in Collier county, where there was high concordance between the scientific and eBird model

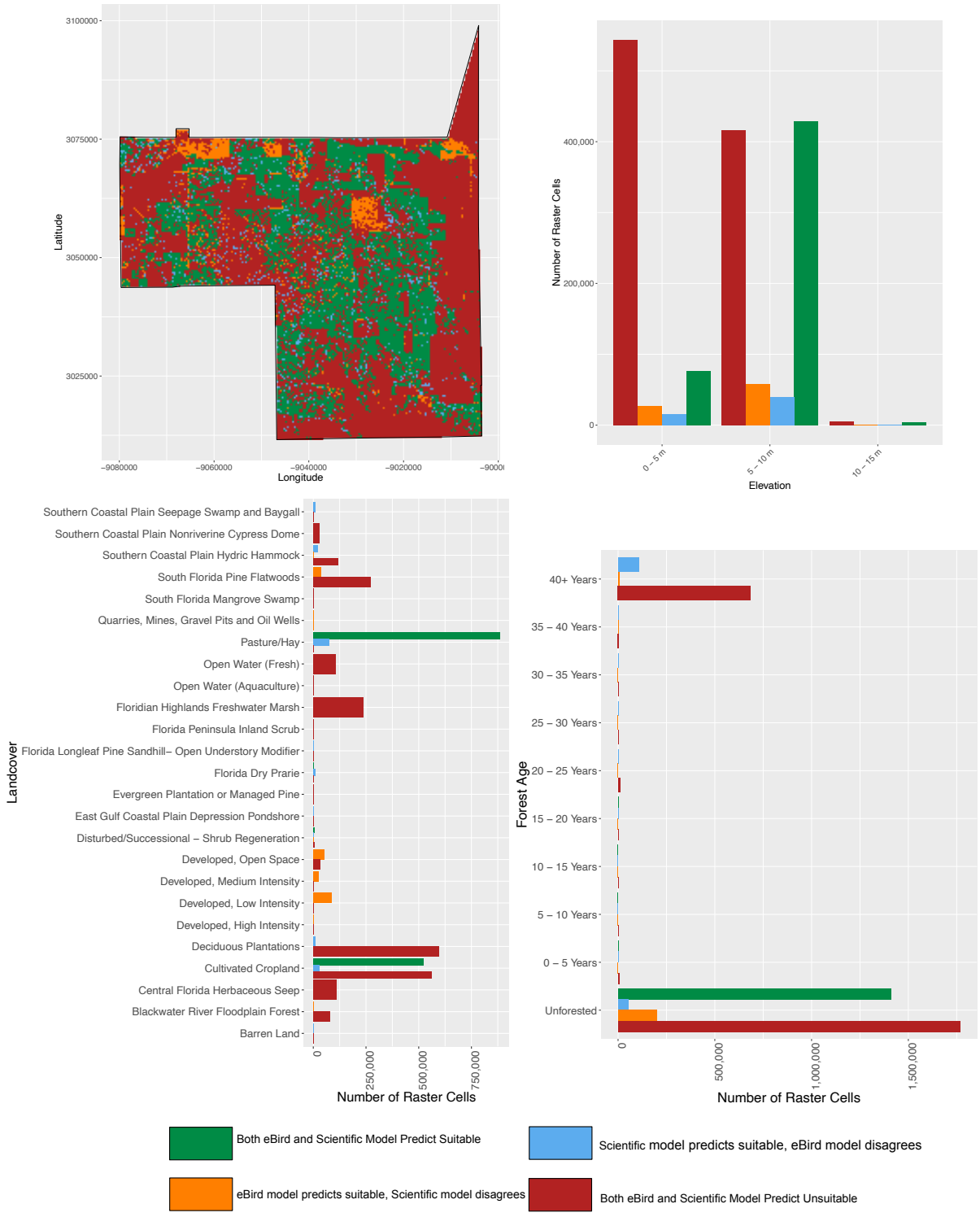


Figure 3.5 (cont.) Map of model agreement and distribution of environmental characteristics in Areas A, B, C, and D in Hendry county, where there was high concordance between the scientific and eBird model

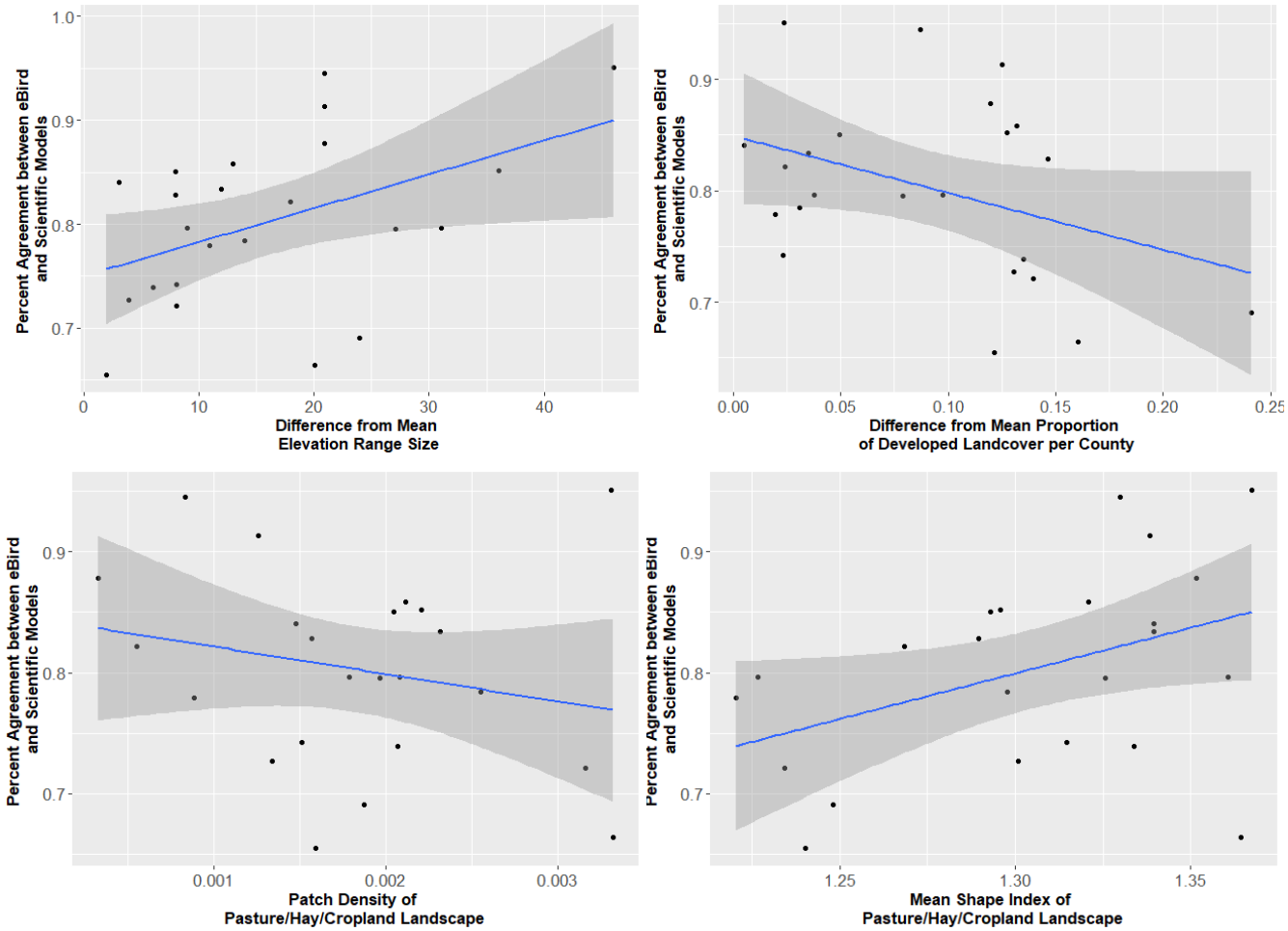


Figure 3.6 Scatterplots used to assess the relationship between model variables/environmental characteristics of counties and the percent of concordant cells between the scientific and eBird model per county (each point represents one county)

correlation (r = -0.378, p = 0.0752). This indicated that counties with higher portions of developed land cover would have the highest discordance between the two models. We additionally tested the correlation between the percent of concordant cells between the two models with the difference between each county's elevation range and the mean elevation range of all counties (as a broad indicator of similarity among all counties), and found the relationship was positive and also significant at $p < 0.05$ (r = 0.449, p = 0.03146). This indicated that counties with a wider or narrower elevation range compared with the average elevation range of all counties had greater discordance between the two models.

We then explored the statistical relationships between the fragmentation variables and the percent of concordant cells between the two models through correlations and regression. We found that the patch density (r = -0.598, p = 0.0025) and mean shape index of the cropland and pasture (r = *0.*423*,* p = 0.0443) both had significant relationships at $p < 0.05$. Finally, we developed a regression model that included all four variables as candidate independent variables. However, because the mean shape index was collinear to the difference in elevation range per county, we removed it from the model. Hence the regression model made to help explain the percent of concordant cells between the two Maxent models per county included three independent variables: (1) the patch density of cropland and pasture/hay land cover per county, (2) the proportion of developed land cover per county, and (3) the difference between each county's elevation range and the mean elevation range of all counties. The model (reported in Table 3.3) had an R^2 of 0.65 and an adjusted-R^2 of 0.59. Hence, approximately 60–65% of the variance in the percent of concordant cells in the scientific and eBird models per county could be attributed to different environmental characteristics (related to land cover, elevation, and fragmentation) within the study area.

The partial correlation analysis (Figure 3.7) suggested that the model variables had generally similar effect sizes. We found that the difference between the county's elevation range and the mean elevation range of all counties had the largest effect on concordance between the two models, a positive relationship (partial correlation = 0.62, p = 0.003) which indicates that this variable explains approximately 38.4% of the variance in model concordance per county and controlling for patch density of agricultural land cover and proportion of developed land cover, counties with elevation range sizes similar to the mean had the highest model discordance

Table 3.3 Coefficients and model summary for linear regression model with the percent of overlapping cells (between the scientific and eBird model) as the dependent variables and environmental factors as the independent variables

	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	St. Error	Beta		
(Constant)	0.88	0.034	---	25.78	3.02e-16
Proportion of Developed Landscape	-0.27	0.11	-0.37	-2.57	0.019
Patch Density of Cropland/Pasture	-46.40	14.30	-0.46	-3.24	0.004
Difference from Mean Elevation Range	0.004	0.001	0.48	3.46	0.003
Residual Standard Error: 0.05 on 19 degrees of freedom F-stat: 11.55 on 3 and 19 DF, p-value 0.0002			Multiple R-squared: 0.65 Adjusted R-squared: 0.59		

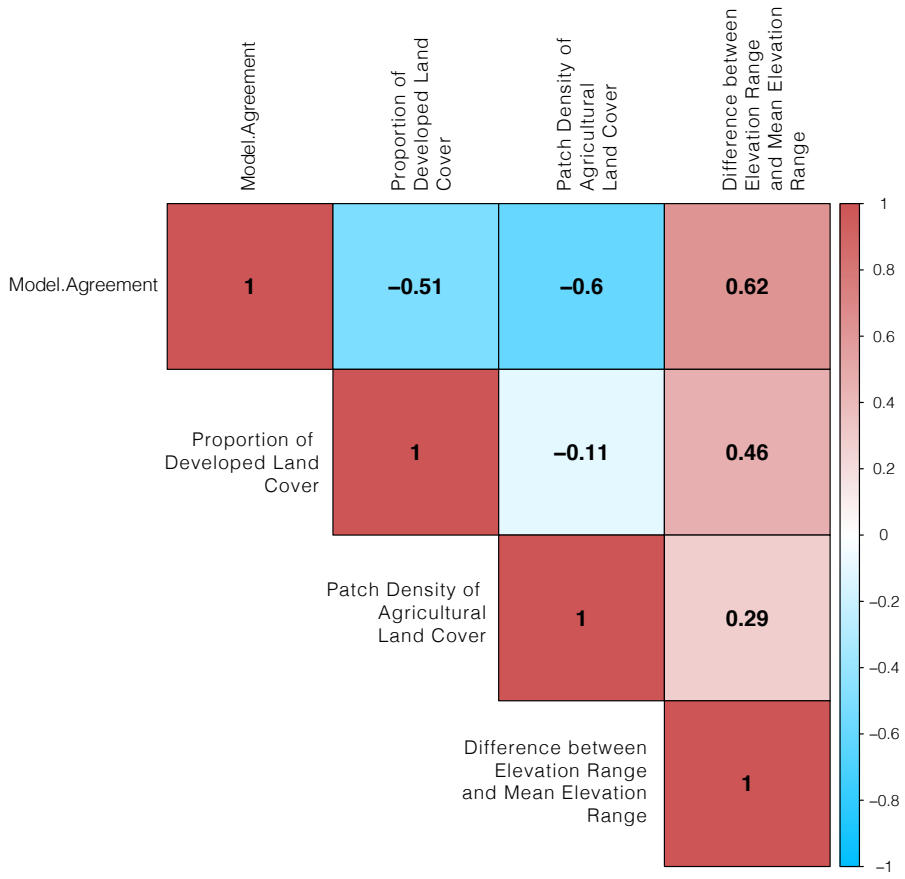


Figure 3.7 Partial correlation results

(concordance increased as the difference increased). The patch density of agricultural land cover had a negative relationship with the model concordance (partial correlation = -0.60, p = 0.004), which indicates patch density explains 36% of the variance in model concordance. In counties with a patchier agricultural land cover controlling for proportion developed landscape and elevation range, the eBird and scientific models had less discordance. Finally, the proportion of developed landscape per county had the smallest, though still comparable effect size, a negative relationship (partial correlation = -0.51, p = 0.02) which indicates that approximately 26% of the variance in model concordance is explained by the portion of developed land cover per county. Controlling for patch density of agricultural land cover and elevation range, counties with more developed land cover had higher discordance between the two models.

3.4 Discussion

Our analysis suggested that although eBird presence data can be used to develop models with comparable diagnostics and similar suitability surfaces as models developed with scientific presence data, there were inconsistencies between model concordance that are related to differences between the two datasets. Further, while the patterns between the two suitability outputs were in many ways similar, when reclassified with a threshold selected to lead to the highest model concordance (i.e., the optimal threshold) there were varying degrees of concordance across the study area.

The histograms of the counties with the highest and lowest model concordance revealed several noteworthy trends. Both models assigned highest suitability values for the species in agricultural landscape, in line with previous knowledge about the distribution of the species, however, there were some cells with this land cover that only the scientific model determined suitable. By contrast, the eBird model assigned high suitability values to many cells classified as developed land cover, in addition to underestimating the suitability of forested landscapes, including old-growth forests that were predicted to be suitable by the scientific model, and areas with relatively high elevations. Interestingly, a small number of cells were in locations classified as prairie, the last remaining areas of the species' native ecosystem. Sometimes the models agreed on the suitability or unsuitability of these cells, but generally only the scientific model

predicted them as suitable (hence occurrences in this land cover were likely not present in the eBird dataset).

Results from the linear regression and partial correlation analysis corroborated trends from the histograms, and findings from previous studies that have concluded that citizen science observation datasets, including those from eBird, have a spatial bias towards anthropogenic landscapes, compared with data collected by scientists or wildlife professionals (Kutner et al., 2005; Li et al., 2019; Sullivan et al., 2009; Tye et al., 2017; Zhang, 2019). Additionally, these results might be particularly useful as a means to understand the sensitivity of the species to different degrees of human development. For example, the eBird model sometimes found less-developed (open-space/low-intensity development) landscape as suitable, and though these areas may lack the necessary characteristics for survival of populations, they still might be hospitable for individuals, particularly for scavenger species given their reliance on carrion and roadkill (Morrison, 2001).

Our statistical analyses additionally suggested that there was higher model discordance in areas where natural habitat for the species was more fragmented. Possible explanations for this include spatial sampling biases from an increased chance for interactions between eBird volunteers and individuals moving along a patchy landscape network, or the potential preference of eBird volunteers to live in areas with more greenspace and thus wildlife. Finally, the difference between the county's elevation range and the mean elevation range of all counties had a positive correlation with the percent of concordant cells per model, indicating that eBird models were more similar to the scientific models in counties with an elevation range that was larger or smaller than average. We hypothesized that the models were more similar in counties with wider ranges of elevation because areas at the highest or lowest elevations were determined unsuitable by both models as they were more infrequent elevation values in the study region, as well as dissimilar from the elevation profile of both eBird and scientific data points. We additionally suggest that counties with a smaller elevation range might be more homogenous (i.e., entirely urban, entirely agricultural) leading to higher concordance between the two models.

We hypothesized that the differences in data collection protocols led to areas of model discordance between the scientific and eBird data based on our exploration of model variables. While an eBird contributor might log the presence of a species during a walk around their

neighborhood or a local greenspace, scientists and wildlife professionals were perhaps more likely to have recorded data in areas they had identified as ideal natural habitats for the species. Because older forests often have denser understories, citizen scientists might have had limited accessibility in these areas, whereas scientists might focus more on these areas as pristine examples of the species' distribution. Scientists or professionals also might be more likely to look for the species in less accessible areas of higher or lower elevation with a goal of finding the species in their known surroundings. This aligns with our findings on the relationship between model differences and habitat fragmentation. While citizen scientists might encounter the species and record species observations in fragmented landscapes as they are areas where human and wildlife interactions are more prevalent, scientists and wildlife professionals may be drawn to more pristine areas with less human development and ergo less fragmentation.

These findings particularly help inform our second research question. Because areas with more anthropogenic landscapes and fragmented natural habitats might highlight the biases and sampling trends in species presence data collected by citizen scientists, they might be areas where VGI based models generally perform poorly. Hence, we posit that SDMs developed with citizen science data are most similar to those developed with scientific data in study regions where biases towards developed landscapes are not exacerbated by a high portion of developed land cover, natural habitats are less fragmented, there are a wide range of different environmental characteristics so that models have greater discriminatory power.

The discrepancies between the scientific and eBird models, which are a reflection of differences in the environmental characteristics at presence locations, intersect a fundamental assumption of SDM development: imperfect versus perfect habitat use (Laurent et al., 2011). An implicit assumption when developing SDMs is that species only occupy suitable habitat, and that the species saturates all the suitable habitat over unsuitable habitat. However, in Maxent, the model's solution is found based on the similarity of areas to the sites of the species presence. Thus, the scientific model, which generally assigned high suitability values explicitly to natural landscapes, predicted similar natural areas as suitable and all areas with developed landscapes as unsuitable. The scientific data might be more appropriate in modeling perfect habitat use. By contrast, the eBird model included more observations that were in a larger range of places compared with the scientific model, as observations were logged in natural habitats as well as in

developed landscapes, which can provide knowledge about a species' imperfect habitat use. Thus, the eBird models might be better suited to understand attributes of non-habitat that influence the species' distribution, including the species' sensitivity to human development.

While the scientific presence data might better model the natural habitat of the species, eBird models may represent changes in the species' spatial ecology as the population adapts to human landscapes. Further, the eBird model might identify areas with the environmental characteristics appropriate for the species' presence that lack the necessary natural resources for survival which might be better represented by the scientific model. However, Florida's crested caracara population has now grown adapted and become reliant on agricultural landscapes. One ecological example that demonstrates that the species has become adapted to human landscapes and giving some credibility to the eBird model is that a part of the population's foraging activities includes scavenging for road kill and finding prey in slaughterhouses, poultry houses, and municipal landfills (Morrison, 2001). Conservation researchers should clearly define the utility of presence datasets in representing the species' distribution in consideration of imperfect and perfect habitat assumptions that define a species' distribution. Further, these findings prompt discussions relating to what is being modeled in species' distribution models, and how different data collection protocols might serve as a means to model species presence along these environmental gradients.

In our study, citizen science data demonstrated broad utility for species distribution models and research. On average, the concordance between the scientific and eBird model at the county level was approximately 80%, indicating that when calibrated, model outputs were generally similar. However, this study also illustrates there are spatial biases that differentiate VGI from scientific data, which prompt opportunities for further research on the causes, implications, and methods to alleviate this effect such as filtering points in residential landscapes prior to SDM development. There also may be scale-based attributes of these biases that may facilitate a species presence like a small forest patch in a residential area near a larger forest. Finally, in a time where natural habitats continue to be altered by human development, VGI perhaps contributes to spatial distributions in human landscapes that may be underrepresented in authoritative scientific data, as the data collection protocol may have focused more on the natural

or pristine areas of the species distribution and underrepresent the species' distribution in human altered landscapes.

While our experiment provided insights into the differences between models developed with citizen science and scientific data, there were limitations in the study design that should be mentioned. Regarding the focal species of our work, it might be helpful to run this experiment using presence data of a species with a different spatial ecology as some behavioral aspects of the crested caracara in Florida, such as the species' adaptation to human landscapes, may have affected results. Additionally, the presence datasets had different sample sizes, which we did not account for in the study design but may have affected results, although data size is an inherent difference between VGI and authoritative data. Moreover, the datasets had different temporal resolutions and were sampled over different periods of time. This could result in mismatches with the environmental variables, like changing land covers, and so future work might consider different ways to filter observations based on their collection dates. Further, the use of counties as areal units may have issues regarding the modifiable areal unit problem that might be addressed by performing the experiment again with different areal units. Finally, the methods were based on the benchmark scientific model reclassified using the 10% training omission rate threshold. Because threshold selection has become a contentious discussion in SDM literature, future analyses might elucidate how results respond to changes in the threshold selected for the scientific model.

3.5 Conclusions

Our research highlighted the way that differences between VGI and authoritative data manifest themselves in different data products. We explored the ways that species distribution models developed with presence data from scientists differed from models developed with presence data from citizen scientists and volunteers. We did this by calibrating the model threshold of the eBird model to the scientific model in a majority of the study area and assessing its transferability to a minority section of the study area. Hence, we assessed the consistency of the selected model threshold and then analyzed the cells of model concordance and discordance. Based on the calibrated model threshold, we found that the consistency of model concordance varied through the study region. While in some counties the two models had a concordance

greater than 90%, in others more than 30% of cells in the model were discordant. After exploring the environmental characteristics of areas where the models agreed and disagreed, we discussed how the model mismatch might be rooted in different sampling protocols between the two datasets. For example, more eBird observations were recorded in urban areas and neighborhoods, resulting in models that predict more populated areas to be suitable compared with models developed with data from scientists who collect more datapoints in more pristine, less publicly-assessible areas. Researchers who employ VGI for species distribution models should be aware of the biases that may be present in data products and further research should be done to mitigate their effects.

Works Cited

Agresti, A., & Finlay, B. (1997). *Statistical methods for the social sciences*. Upper Saddle River, N.J: Prentice Hall.

Allouche, O., Tsoar, A., & Kadmon, R. (2006). Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). *Journal of applied ecology*, *43*(6), 1223-1232.

Amano, T., Lamming, J. D., & Sutherland, W. J. (2016). Spatial gaps in global biodiversity information and the role of citizen science. *Bioscience*, *66* (5), 393-400.

Capineri, C. (2016). The Nature of Volunteered Geographic Information. *European Handbook of Crowdsourced Geographic Information*, 15-33.

Clark, C. J. (2017). eBird records show substantial growth of the Allen's Hummingbird (Selasphorus sasin sedentarius) population in urban Southern California. *The Condor: Ornithological Applications*, *119* (1), 122-130.

Comber, A., See, L., Fritz, S., Van der Velde, M., Perger, C., & Foody, G. (2013). Using control data to determine the reliability of volunteered geographic information about land cover. *International Journal of Applied Earth Observation and Geoinformation*, *23*, 37-48.

[Dataset] eBird. 2019. eBird: An online database of bird distribution and abundance [web application]. eBird, Cornell Lab of Ornithology, Ithaca, New York. Available: http://www.ebird.org. (Accessed: Date [June 13, 2019]).

[Dataset] Florida Forest Service, 2013. Stand Age Raster Grid, https://freshfromflorida.s3.amazonaws.com/TimberStandAge.zip

[Dataset] Florida Natural Areas Inventory. Crested Caracara Presence Data, December 2018.

Fonte, C.C., Bastin, L., Foody, G., Kellenberger, T., Kerle, N., Mooney, P., Olteanu-Raimond, A.M. and See, L., 2015. VGI quality control. *ISPRS Geospatial Week 2015*, 317-324.

[Dataset] Gesch, D., Oimoen, M., Greenlee, S., Nelson, C., Steuck, M., & Tyler, D. 2002. The national elevation dataset. *Photogrammetric Engineering and Remote Sensing*, 5-32.

Hultquist, C., & Cervone, G. (2018). Citizen monitoring during hazards: validation of Fukushima radiation measurements. *GeoJournal*, *83* (2), 189-206.

Hurlbert, A. H., & Liang, Z. (2012). Spatiotemporal variation in avian migration phenology: citizen science reveals effects of climate change. *PLOS ONE*, *7* (2).

Kelling, S., Gerbracht, J., Fink, D., Lagoze, C., Wong, W.K., Yu, J., Damoulas, T. and Gomes, C., 2013. A human/computer learning network to improve biodiversity conservation and research. *AI magazine*, *34* (1), 10-20.

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied Linear Statistical Models* (Vol. 5). New York: McGraw-Hill Irwin.

Laurent, E. J., Drew, C. A., & Thogmartin, W. E. (2011). The role of assumptions in predictions of habitat availability and quality. In *Predictive Species and Habitat Modeling in Landscape Ecology* (pp. 71-90). Springer, New York, NY.

Li, E., Parker, S. S., Pauly, G. B., Randall, J. M., Brown, B. V., & Cohen, B. S. (2019). An urban biodiversity assessment framework that combines an urban habitat classification scheme and citizen science data. *Frontiers in Ecology and Evolution*, *7*, 277.

McCormack, J. E., Zellmer, A. J., & Knowles, L. L. (2010). Does niche divergence accompany allopatric divergence in Aphelocoma jays as predicted under ecological speciation?: insights from tests with niche models. *Evolution: International Journal of Organic Evolution*, *64* (5), 1231-1244.

Morrison, J. L. (2001). *Recommended management practices and survey protocols for Audubon's Crested Caracara (Caracara cheriway audubonii) in Florida* (No. 18). Technical Report.

Morrison, J. L., & Humphrey, S. R. (2001). Conservation value of private lands for Crested Caracaras in Florida. *Conservation Biology*, *15*(3), 675-684.

Muller, C.L., Chapman, L., Johnston, S., Kidd, C., Illingworth, S., Foody, G., Overeem, A. and Leigh, R.R., 2015. Crowdsourcing for climate and atmospheric sciences: current status and future potential. *International Journal of Climatology*, *35* (11), 3185-3203.

Munson, M.A., Caruana, R., Fink, D., Hochachka, W.M., Iliff, M., Rosenberg, K.V., Sheldon, D., Sullivan, B.L., Wood, C. and Kelling, S. (2010). A method for measuring the relative information content of data from different monitoring protocols. *Methods in Ecology and Evolution*, *1* (3), 263-273.

Phillips, S. J., Dudík, M., & Schapire, R. E. (2004). A maximum entropy approach to species distribution modeling. In *Proceedings of the twenty-first international conference on Machine learning* (p. 83).

Phillips, S. J., & Dudík, M. (2008). Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. Ecography, 31 (2), 161-175.

Ruete, A., & Leynaud, G. C. (2015). *Goal-oriented evaluation of species distribution models' accuracy and precision: true skill statistic profile and uncertainty maps* (No. e1478). PeerJ PrePrints.

Stewart, S. I., V. C. Radeloff, and R. Hammer. (2003). The Wildland-Urban Interface in US Metropolitan Areas. Pages 254 – 255 National Urban Forest Conference Proceedings.

Sui, D., Elwood, S., & Goodchild, M. (Eds.). (2012). *Crowdsourcing Geographic Knowledge: volunteered geographic information (VGI) in theory and practice*. Springer Science & Business Media.

Sullivan, B. L., Wood, C. L., Iliff, M. J., Bonney, R. E., Fink, D., & Kelling, S. (2009). eBird: A citizen-based bird observation network in the biological sciences. *Biological Conservation*, *142* (10), 2282-2292.

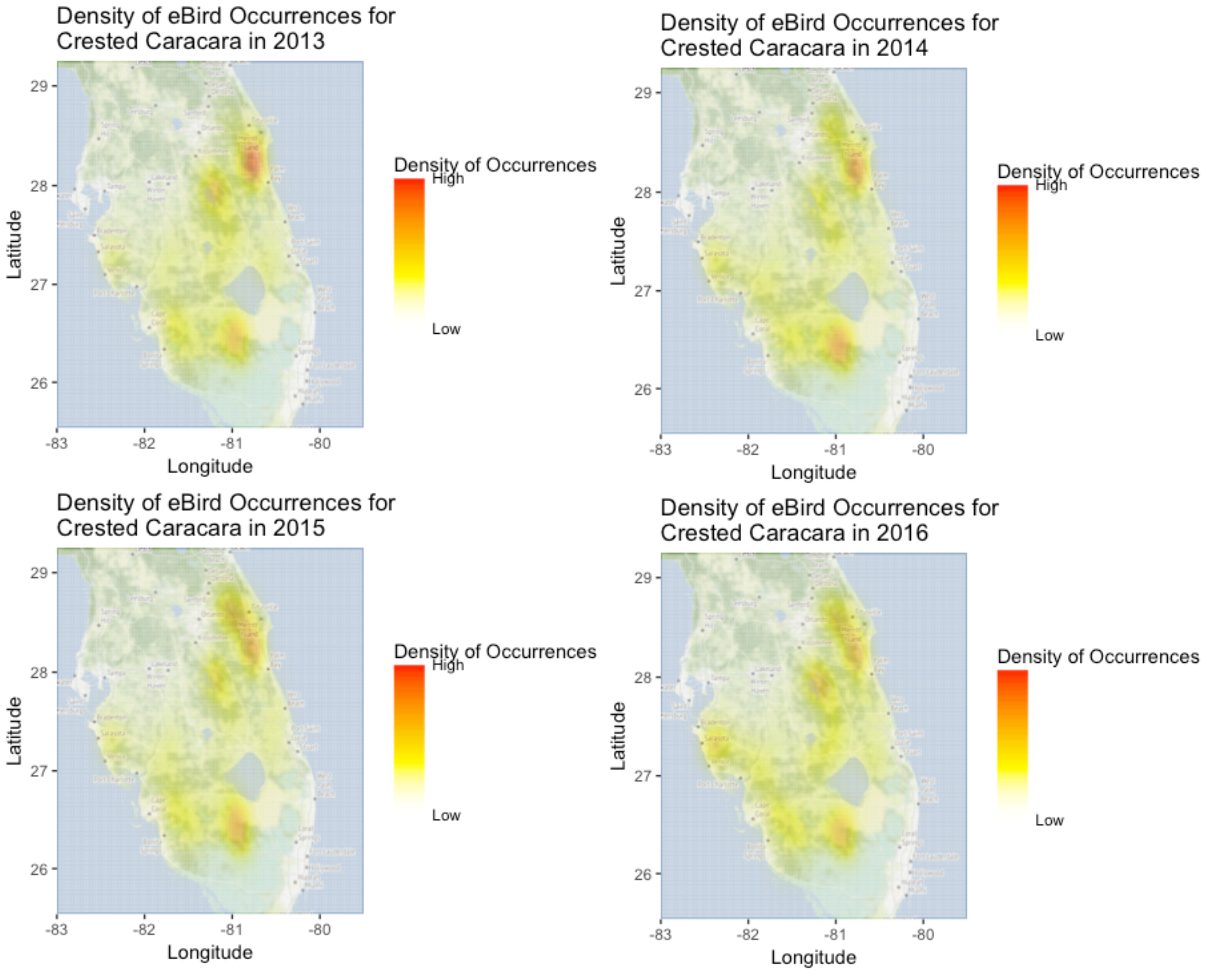
Trautmann, N., Fee, J., & Kahler, P. (2012). Flying into inquiry. *The Science Teacher*, *79* (9), 45.

Tye, C. A., McCleery, R. A., Fletcher Jr, R. J., Greene, D. U., & Butryn, R. S. (2017). Evaluating citizen vs. professional data for modelling distributions of a rare squirrel. *Journal of Applied Ecology*, *54* (2), 628-637.

[Dataset] U.S. Geological Survey Gap Analysis Program, GAP/LANDFIRE National Terrestrial Ecosystems 2011: U.S. Geological Survey, https://doi.org/10.5066/F7ZS2TM0.

Zhang, G. (2019). Enhancing VGI application semantics by accounting for spatial bias. *Big Earth Data*, *3* (3), 255-268.

Appendix



Appendix 3.1 Density plots demonstrating that the footprint of eBird presence points for crested caracara is relatively consistent at annual intervals

CHAPTER 4 EVALUATING AND FILTERING VOLUNTEERED GEOGRAPHIC INFORMATION USING SCIENTIFIC DATASETS FOR SPECIES DISTRIBUTION MODELS

My use of "we" in this chapter includes my co-authors, Liem T. Tran and Monica Papeş. As first author, I led experimental design, data analyses, and writing the manuscript.

Abstract

We used inferences from a species distribution model developed with a scientific presence dataset, along with the statistical concept of influence, to understand the utility of data from the citizen science platform eBird for species distribution modeling. We observed that Maxent models for the crested caracara (*Caracara cheriway)* developed with different presence datasets, one from eBird and the other from a scientific organization, had a 68% agreement. We then developed a logistic regression model using the eBird dataset as presence data, and with absence points sampled from areas predicted unsuitable by the Maxent model developed with the scientific dataset. From this logistic regression model, we measured the Cook's distance for each eBird point and, in conjunction with each point's predicted suitability value from the scientific model, we used K-means clustering analysis to categorize the eBird points. We then used the cluster containing points with the smallest Cook's distance and highest scientific predicted suitability to create a Maxent model, which we found had a 90% agreement with the scientific model and also performed better than a model developed with the eBird presence data filtered by existing data review mechanisms. We then analyzed the environmental and geographic attributes of the clusters to further understand how eBird data points were different from the scientific presence dataset. Finally, we posited explanations for the differences between the eBird and scientific presence datasets and provided suggestions for the utility of these methods and findings.

4.1 Introduction

4.1.1 *VGI in Environmental Analysis*

Wildlife sighting datasets provided by volunteers or citizen scientists are examples of Volunteered Geographic Information (VGI), data contributed by the public that contain geographic locations or spatial coverage (Sui, Elwood, & Goodchild, 2012). Data collection in a VGI framework is a cost-effective way to record useful information over wide geographic extents and long periods of time, and is also a way to support meaningful engagement between the public and the scientific community (Bonney et al., 2009). Foundationally, the collection of

data through a VGI framework is more decentralized and community-based, whereas authoritative or scientific datasets are gathered by trained and often paid personnel (Sui, Elwood, & Goodchild, 2012). Further, scientific/authoritative data are collected based on established methods, standards, specifications, and practices whereas in a VGI framework the roles of data users, producers, reviewers, and arbitrators are fluid, and formal training is not required (Bruns, 2008). In this research, we investigate how these two types of data fit different species distribution models (SDMs) and then seek to understand the additional information provided in presence data from citizen scientists.

Although few approaches to assess the accuracy of wildlife observation data collected by citizen scientists have been explored (Danielsen et al., 2005; Crall et al., 2011), species presence data have been employed in important and diverse scientific research. For example, Ramesh et al. (2017) used observation data from one of the most comprehensive wildlife sighting databases for citizen scientists, eBird, to critically evaluate results of niche models used for the IUCN Red List. Species presence datasets from eBird have also been employed to help understand the distribution of invasive species (Cardador et al., 2016), explore aspects of species evolution (McCormack et al., 2010), track distribution changes prompted by climate change (Hurlbert and Liang, 2012), and monitor populations of rare and endangered species (Clark, 2017).

There are several reasons why data collected by citizen scientists contain more error, noise, and bias compared with species presence data provided through traditional, more methodical collection protocols. At a broad scale, factors related to human population density and economic geography result in underdeveloped and sparsely-populated regions lacking in wildlife observations from citizen scientists, even though these areas often host rich biodiversity (Amano et al., 2016). Locally, there are sociopolitical factors like private landownership that might impact the spatial distribution of reported species observations, given how they affect the accessibility of areas to the public (Cooper et al., 2000). In the case of eBird, one explanation for variability of data quality is that there is a range of expertise among data contributors, as seasoned bird watchers input observations to the same dataset as newcomers to the hobby (Yu et al., 2010). The misidentification of a species or the incorrect submission of coordinates could also degrade the dataset (Kelling et al., 2012). For these reasons, eBird has developed a

comprehensive data review process that can identify submissions of questionable data quality (Kelling et al., 2012; Sullivan et al., 2014).

It is particularly important to assess the quality of eBird points (including their locational accuracy, species identification, and accurate representation of the species' habitat) when analyzing and modeling species' distributions. One technical reason for this is that eBird provides data that are appropriate for SDMs that do not require data on where the species is absent as an input to the model. For example, logistic regression models generate suitability surfaces by comparing environmental characteristics to both locations where the species has been observed and locations that are unsuitable for the species (Wisz and Guisan, 2009). By comparison, Maxent, a popular presence-background model, calculates a suitability surface for the species by contrasting the environmental characteristics at presence locations with those from the entire study area (Elith et al., 2011). Though Maxent models are in some ways robust to outliers, locations where a species is erroneously recorded as present can result in significant changes (Aubry et al., 2017). Thus, particularly when using VGI to develop SDMs that do not use absence data, data quality assessment and appropriate filtering methods are important to ensure the quality of data products.

4.1.2 *eBird Review Process*

Thousands of volunteers have contributed millions of data points to eBird. As both the misidentification of bird species and locational accuracy are possible sources of error, eBird operates a two-step automated data quality and review process to efficiently identify and flag unusual submissions for further review by an expert volunteer (Kelling et al., 2012; Sullivan et al., 2014). The first step of the process is an automated flag mechanism based on patterns that have emerged from the frequency of previously submitted data on the species in particular places at specific times. For example, a submission might be flagged if the majority of a species' previously submitted presence records for the species were in areas farther south because of migration patterns during that time of the year.

The second automated filter that submissions undergo in the data review process is called the Occupancy-Detection-Experience model, which considers the environmental attributes at the observation location, the frequency at which the species has been recorded at this location, and

the expertise of the data contributor (Kelling et al., 2012). The model distinguishes between novice and expert bird watchers, an important factor in assessing data validity, particularly in unusual sightings of rare species. Finally, submissions that are flagged by these automated filters are reviewed by eBird's international network of volunteer experts, who are knowledgeable about regional bird occurrences and will follow up with individuals who submit presence data requesting additional materials like photographs for validation (Kelling et al., 2012; Sullivan et al., 2014).

4.1.3 ***Research Questions***

An important facet of the eBird data review process in the context of this study is that submissions are validated based on inferences determined by past submissions to eBird. In our experiment, we explored properties of eBird submissions relative to inferences from a dataset developed by scientists and wildlife experts, to address the research question: *(i) Relative to observation points provided by scientists and wildlife professionals, what additional information can wildlife observation points from citizen scientists contribute for species distribution modeling?* Further, we explore (*ii) What can the differences in the environmental characteristics of eBird points tell us about the species' distribution and species presence data from citizen scientists?*

To better understand the additional information provided by eBird data points, we employed the statistical measure of influence. Data points are of particular importance when their inclusion in statistical models results in changes to the fit of the overall model or the model's parameter estimates. When these individual data points contribute unique or relatively rare information to models compared with the rest of the dataset, they can have important and disproportionate impacts on model outputs, and therefore they have high influence (Kutner et al., 2005). However, some of these points might also be candidates for a closer review of their properties and data quality assessment. In developing our experiment, we used measures of influence to characterize eBird points relative to their similarity with the larger eBird dataset and the scientific model.

Through logistic regression, the influence of each eBird presence point on the model could be measured relative to the rest of the dataset. However logistic regression models require

absence data which are unavailable from eBird, which is why Maxent is a useful SDM approach with eBird datasets. To overcome the lack of absence data for the species, we developed a Maxent model with data from a scientific research organization to identify areas predicted as unsuitable for our target species. By selecting background points only from these areas, we created pseudo-absence data (Wisz and Guisan, 2009) with scientific authority. Using those pseudo-absence points and eBird presence points, we developed a logistic regression model from which we calculated the influence of each eBird point. We then used that measure to help answer our research questions and as a basis to evaluate additional information provided by eBird points.

4.1.4 *Focal Species: Crested Caracara*

We modeled distributions of the crested caracara (*Caracara cheriway*), a threatened non-migratory bird of prey that occurs in an isolated population in southcentral Florida (Morrison and Humphrey, 2001). Historically the species was part of the dry prairie ecosystem, which has undergone significant land cover changes (Morrison and Humphrey, 2001). Caracaras are territorial, and, in contrast to most avian species, spend significant time on the ground, making them sensitive to local land cover conditions (Morrison, 2001). Though their natural habitat historically included marshes, grasslands, and prairies, increased urbanization and agricultural/pasture expansion mean the Florida caracara population is now often found in areas with scattered trees, short/ground vegetation, and with a minimal understory or shrub layer, such as pastures, cropland, and cattle ranches. The scavenger species is also found feeding on carrion along roadsides, waste facilities, and near slaughterhouses (Morrison, 2001).

4.2 Methods & Materials

4.2.1 *Species Presence Data and Environmental Data*

The scientific dataset used in our study included 225 presence records for the crested caracara with input dates ranging from 1978 to 2013. The dataset was provided by the Florida Natural Areas Inventory (FNAI), a non-profit research organization at Florida State University that maintains state biodiversity inventories. The original data were provided in the form of buffer polygons used

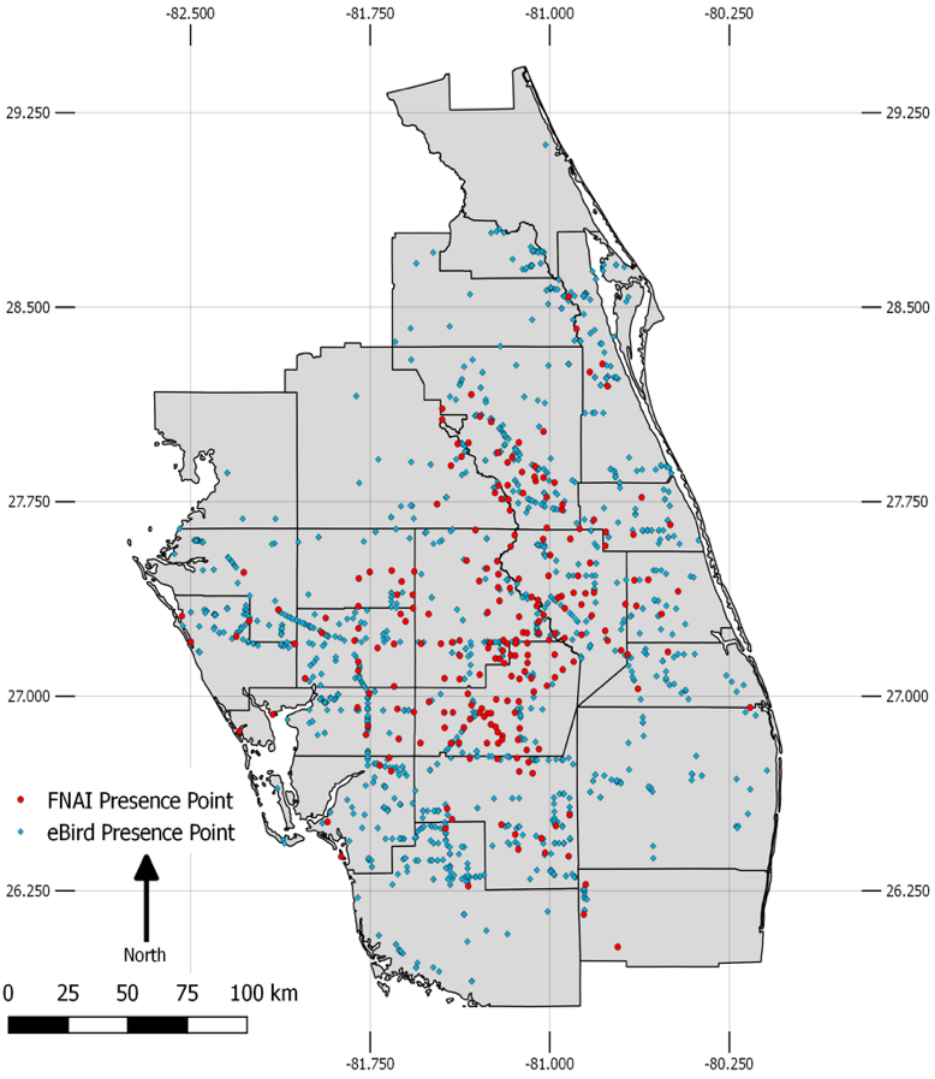


Figure 4.1 Distribution of eBird and scientific presence points

to provide a range of spatial uncertainty, but we used the centroid of these polygons as presence locations for use in Maxent. We download the crested caracara presence dataset directly from eBird. Because of the volume of data available for the species, we used all data points that were recorded in 2017, resulting in 2,831 presence points. Figure 4.1 displays the distribution of eBird and scientific presence points across the study area. We confirmed that subsetting the eBird data by year resulted in similar spatial patterns of points by creating heat maps of eBird points subsetted annually over a four-year period (Appendix 4.1). Table 4.1 outlines the variables used to develop the models in our experiment. Four of the independent variables used in the study were published by the US-Gap Analysis Program in 2011 (elevation, slope, aspect, and land cover), and had a native resolution of 30-m. The Florida Department of Agriculture and Consumer Services provided a raster of forest age from 2014 that also had a native resolution of 30-m.

4.2.2 *Modeling methods*

We first developed two SDMs using the same set of environmental variables; however, one model used the eBird presence dataset while the other model used presence data from the FNAI. Maxent models were developed using the stand-alone Java application with the cross-validation approach with 10 model replicates. The cross-validation method was chosen because it uses all presence data provided, randomly selected by Maxent as either a testing point or training point during model development (Phillips, Dudik, & Schapire, 2004). We used a commonly employed product of Maxent, the logistic output, a raster that has the same resolution as the model's input variables with values that range from 0.0 to 1.0 that correspond to the similarity of environmental characteristics of each raster cell to the environmental characteristics at locations where the species was recorded present (Phillips, Dudik, & Schapire, 2004). We assessed the quality of Maxent models using the Area Under the Receiver Operating Characteristic Curve (AUC), a frequently employed validation metric for threshold-based models. In Maxent, the AUC corresponds to the probability that a presence location chosen at random is ranked by the model as more suitable than a random background location (Phillips and Dudík, 2008). AUC ranges from 0 to 1.0 with larger numbers indicating a better model fit. We also compared the test gain of the Maxent models, which, when exponentiated, is the likelihood ratio of a typical presence point to a typical

Table 4.1 Input variables used in Maxent models

Variable	Variable Description	Data Source
Forest Age	Categorical raster corresponding to biomass stand age classified in 5-year intervals, derived from Landsat imagery (2014)	Florida Department of Agriculture and Consumer Services
Land cover	GAP National Terrestrial Ecosystems 2011 includes detailed vegetation and land cover information for the entirety of the United States, representing 590 land cover classifications nationally and 102 land cover classifications within Florida	USGS Gap Analysis Program
Elevation	Continuous raster, Digital elevation model derived from the National Elevation Dataset (2011)	USGS Gap Analysis Program
Slope	Dataset derived from the elevation dataset which reflects the slope (i.e., rise over run) of the area, reported in degrees	Derived from Elevation raster
Aspect	Categorical raster derived from elevation dataset which corresponds to the cardinal direction the area within each cell is facing	Derived from Elevation raster

background point. Through the model development in Maxent the gain is maximized such that the model solution best discriminates locations with observations from background locations (Merow et al., 2013). Hence, higher gains indicate models that can better discriminate between suitable habitat (as inferred by species occurrences) and the relative suitability of the rest of the study area.

To measure the influence of individual eBird observation points, we brought the data into a logistic regression context. To build the regression model and understand the characteristics of eBird points relative to the scientific data, we created pseudo-absence points by sampling 10,000 points, the number of pseudo-absence points recommended by Barbet-Massin et al. (2012) and the default number of background points used in Maxent. The pseudo-absence points were sampled from areas that were determined unsuitable by the scientific Maxent model after converting the continuous suitability output to a binary format (suitable-unsuitable) using the 10% training presence threshold, the threshold value at which 10% of training locations are predicted to be unsuitable. After developing the logistic regression and examining diagnostic plots, particularly the environmental attributes of the pseudo-absence points, to assess the model's quality we measured the influence of each observation on the model by calculating the Cook's distance of each point.

Cook's distance is a common measure of statistical influence in the regression context found by calculating the effect of deleting a single observation point on the rest of the model (Kutner et al., 2005). Cook's distance, as outlined by Kutner et al. (2005), is calculated by:

$$D_i = \frac{\sum_{j=1}^{n} (\hat{Y}_j - \hat{Y}_{j(i)})^2}{p\ MSE}$$

Equation 4.1 Calculation of Cook's Distance for observation $\boldsymbol{i}$ (for $\boldsymbol{i} = 1,\ldots, \boldsymbol{n}$) is found by summing all changes to a regression model after deleting the observation. $\widehat{\boldsymbol{Y}}_{\boldsymbol{j}}$ is the jth fitted response value from the entire regression model and $\widehat{\boldsymbol{Y}}_{\boldsymbol{j(i)}}$ is the jth fitted response value in the regression excluding observation $\boldsymbol{i}$. $\boldsymbol{p}$ is the number of parameters in the model, and $\boldsymbol{MSE}$ is the mean squared error of the full regression

We expected that eBird locations with more distinct environmental characteristics relative to the rest of the dataset would have larger Cook's distances, as their rare attributes would have a larger effect on variable coefficients leading to differences in the fitted values of the rest of presence points. Hence, we reasoned that presence points that had larger Cook's distances in the logistic regression would have more distinct environmental characteristics, and thus would be more

similar to the background points sampled for the scientific model and less similar to observations from the scientific dataset.

While Cook's distance can be used to identify rare and infrequent data characteristics, we reasoned that another important indicator of validity was the scientific model's suitability estimate of the eBird presence locations. Thus, we extracted the predicted suitability at eBird locations from the scientific model. We then had two measures for each eBird presence location: (1) the suitability predicted by the scientific model and (2) the Cook's distance from the logistic regression. By classifying all points based on these two measures, we aimed to isolate points that best matched the environmental profile of the scientific points, and then explore the points that deviated from those profiles (i.e., were more rare or infrequent, or in areas that disagreed with the scientific model) to better understand the additional information provided by the eBird dataset.

We determined the subset of eBird points most similar to the scientific points, and classified others, by partitioning the entire eBird dataset into clusters using the K-means clustering algorithm with GeoDa (Anselin et al., 2006). K-means clustering is a widely-employed data classification method in which n data points with d attributes are assigned to one of k clusters (Han et al., 2012). Clusters are determined through an iterative process, during which data points plotted with input variables as axes–in this case, scientific predicted suitability and Cook's distance in the logistic regression–are assigned to nearest cluster centers which are randomly selected in each iteration. Through each iteration, the mean value of the input variables for all points in each cluster is calculated, along with the within-group variation around the explanatory variables per cluster. Points are then reassigned until the within-cluster variation is minimized by the iterative cluster reassignment and cluster assignment is stable, or a maximum number of iterations is reached (Han et al., 2012). The number of clusters (k) is a parameter set by the researcher, and was determined using the elbow method (Ketchen and Shook, 1996), in which the ratio of the within to total sum of squares was plotted for a run of the K-means algorithm with k ranging from 2 to 10 clusters.

We then used the cluster of eBird points with the highest scientific suitability value and lowest Cook's distance value to develop a third Maxent model with the same Maxent settings and variables as our previous models. We measured the similarity of this model to the scientific

model and the model developed with all eBird points. The similarity between this model and the scientific model would provide evidence that the cluster of points was the most similar to the scientific dataset. We evaluated model similarity by calculating the overlap (i.e., percent of matching values) of the regression and Maxent binary outputs (suitable-unsuitable) based on a suitability threshold at which sensitivity and specificity of the test sample is maximized, chosen to address both the true positive and true negative rate. To understand how existing eBird data quality infrastructure changed the model outputs, we also removed the 150 flagged observations from the eBird dataset in a fourth Maxent model to compare its performance diagnostics and the output's similarity to the scientific model.

Finally, to most directly address our research goal of understanding the additional information provided by eBird points, we analyzed the environmental characteristics of points in each cluster and how they affected the Cook's distance and scientific suitability values of the locations. Although the K-means algorithm is non-spatial, we also mapped the points by clusters and explored spatial patterns along the scientific and eBird model to investigate the geographic attributes of the point clusters.

4.3 Results

The scientific model and full eBird (Maxent) model both performed reasonably well according to AUC (Table 4.2), with the scientific model (0.75) slightly outperforming, but within the margins of, the full eBird model (0.73). The scientific model also had a slightly higher test gain (0.405) than the eBird model (0.349). According to the full eBird model, the most important variables were land cover (52.3%), elevation (35.1%), and forest age (9.9%); compared with the scientific model which found land cover (64.8%) and elevation (33.3%) to have high importance in the model, while forest age (0.1%) did not substantially contribute to the model's solution. The logistic outputs of the models displayed similar spatial patterns, with some differences (Figure 4.2, top row). While the scientific model suggested higher suitability in the interior parts of the study area, the eBird model suggested higher suitability along coastal areas. Some regions have very different predicted suitability values, including the southeastern and northwestern

Table 4.2 Model diagnostics to compare Maxent models developed in our experiment

	Scientific Model	Full eBird Model	eBird model Minus Flagged	eBird model (Only Clusters 4)
AUC	0.75	0.73	0.73	0.85
Test Gain	0.405	0.349	0.38	0.933

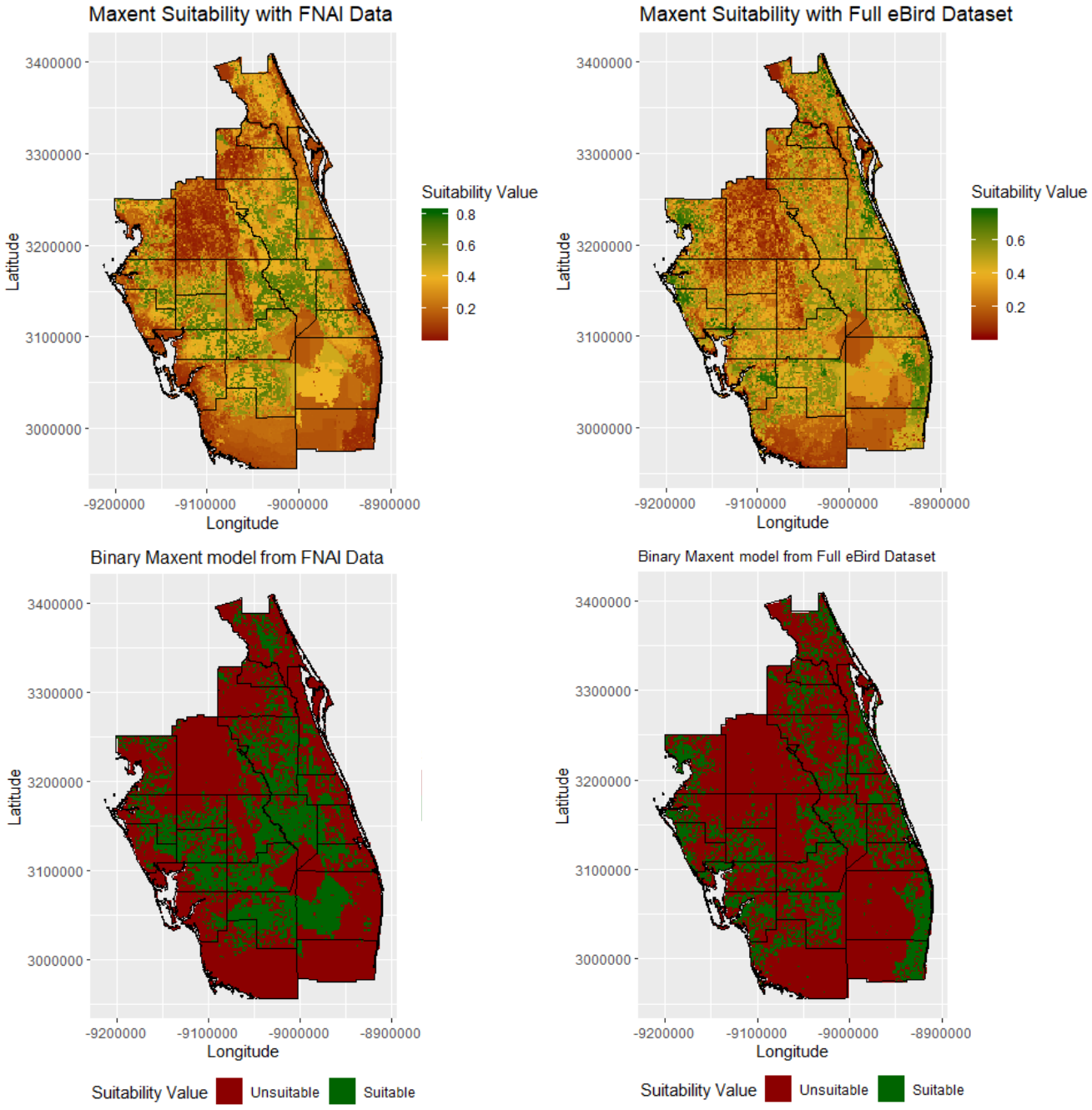


Figure 4.2 [Top Row] Maxent logistic outputs for crested caracara from models developed with scientific (left) and eBird presence data (right); [Bottom Row] thresholded scientific model output (left) and thresholded eBird model output (right)

Table 4.3 Percent of overlapping points when binary models were thresholded by the maximum sensitivity + specificity threshold

	Scientific Model	Full eBird Model	eBird model Minus Flagged	eBird model (Only Clusters 4)	Logistic Regression
Scientific Model	1.00	0.68	0.69	0.90	0.89
Full eBird Model	0.68	1.00	0.96	0.76	0.76
eBird Model minus Flagged	0.69	0.96	1.00	0.77	0.78
eBird model (Only Clusters 4)	0.90	0.76	0.77	1.00	0.96
Logistic Regression	0.89	0.76	0.78	0.96	1.00

coasts (where the eBird model predicted high suitability, yet the scientific model predicted low suitability). A general difference is that the eBird model predicted more areas were suitable than the scientific model, but the scientific model attributed higher values to a more limited area. Differences in the model overlap were more apparent when the threshold was applied (Figure 4.2, bottom row), as 68% of cells from the full eBird and scientific model overlapped (Table 4.3). When we plotted the eBird points on the scientific model output, there was a wide range of predicted values at eBird presence points (0.02 - 0.80; Figure 4.3), with many found in cells with low predicted suitability scores (75% of points had a suitability value less than or equal to 0.50).

The logistic regression model (Figures 4.4 and 4.5), developed with the full eBird presence data and absence points sampled using the scientific model output, had an AUC of 0.74. The thresholded model had an 89% overlap (Figures 4.2 & 4.4; Table 4.3) with the scientific model (compared with a 76% overlap with the full eBird model) and corroborated ecological knowledge about the species: known attributes of the species' habitat had higher effect sizes and the profile of independent variables (Figure 4.5) agreed with the scientific model and showed preference towards agricultural/pasture, generally but not always in areas of lower elevation with younger forest stands or nonforested cover. Moreover, the pseudo-absence data had attributes that corroborate knowledge about areas the species avoid such as developed areas and high elevation places (see Appendix 4.2). Given these results, we had confidence that the logistic regression was appropriate to base our measures of influence. When we calculated Cook's distance of the eBird presence points in the logistic regression, we observed the distribution of values was skewed (Figure 4.5). Hence, we log-transformed the values so their distribution was more normal and so that the resulting clusters would be less affected by extreme outliers and thus of similar sizes.

Using the (1) log-transformed Cook's distance and (2) scientific suitability values as independent variables, we determined via the elbow method that five was the ideal number of clusters for the K-means clustering of the eBird points. The algorithm determined clusters that ranged in number of points from 310 to 750 (Table 4.4), and had clear distinctions based on their Cook's distance values and predicted suitability by the scientific model (Figure 4.6). In K-means clustering, point membership to clusters is on a relative scale. We used group membership to

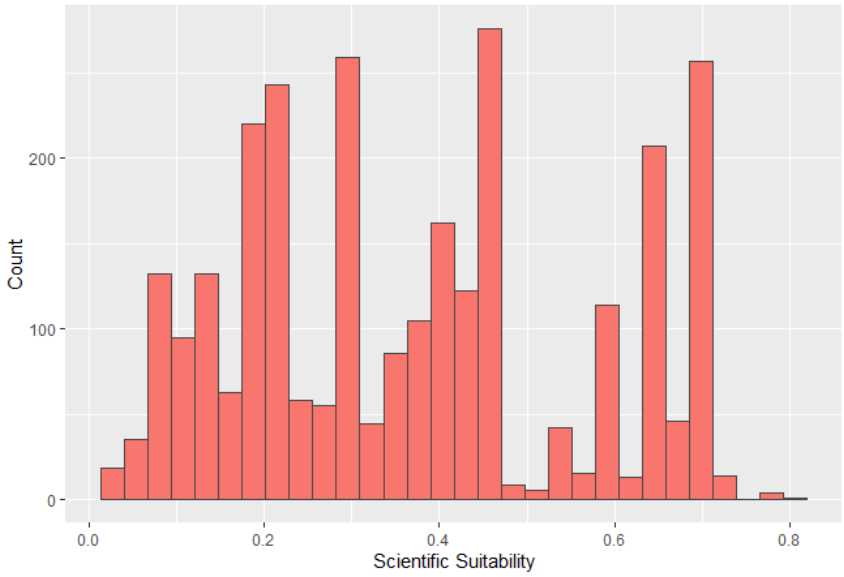


Figure 4.3 Histogram of predicted suitability values by the scientific model at all eBird points in the dataset

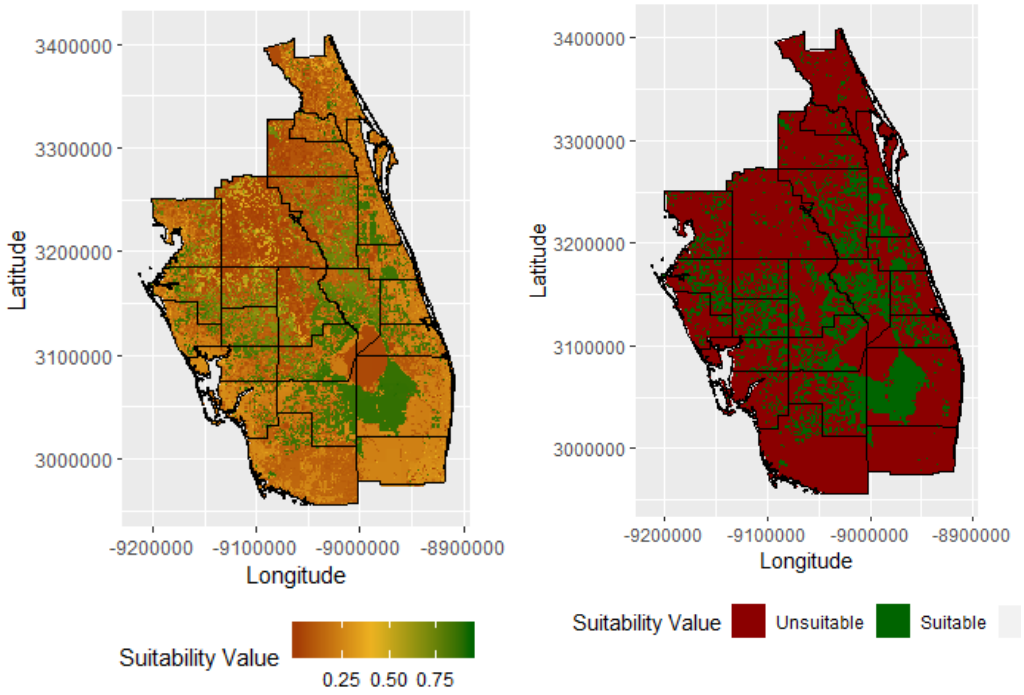


Figure 4.4 Logistic regression model output (left) and logistic regression model output thresholded at maximum sensitivity + specificity threshold (right)

Table 4.4 Cluster attributes from the K-means cluster analysis

	Number of Points	***Standardized Cook's D***	***Standardized Suitability***	***Within Cluster Sum of Squares***	***Interpretation***
Cluster 1	750	0.34	- 0.69	269.94	Low Suitability, Medium Cook's D
Cluster 2	639	-1.16	- 0.78	149.3	Low Suitability, Small Cook's D
Cluster 3	568	0.69	1.40	234.52	High Suitability, Medium Cook's D
Cluster 4	564	- 0.73	0.67	311.01	High Suitability, Small Cook's D
Cluster 5	310	1.64	-0.46	153.54	Low Suitability, Large Cook's D

Results

	Dependent variable	*Exponentiated Coefficient*
	Elevation	0.954***
	Forest Age	0.9***
	Slope	0.629***
Land cover	Shrub & Herb Vegetation	0.870
	Agricultural & Developed Vegetation	10.902***
	Developed & Other Human Use	0.943
	Recently Disturbed or Modified	1.910**
	Open Water	0.123***
	Nonvascular & Sparse Vascular Rock Vegetation	0.383
Aspect	N (0-22.5)	1.514
	NE (22.5-67.5)	1.562*
	E (67.5-112.5)	3.365 ***
	SE (112.5-157.5)	1.878**
	S (157.5-202.5)	0.923
	SW (202.5-247.5)	0.861
	W (247.5 – 292.5)	1.276
	NW (292.5-337.5)	1.554
	N (337.5-360)	0.847
	Constant	0.380***
	Observations	12,831
	AUC	0.74

*Note: *p<0.1;**p<0.05;***p<0.01*

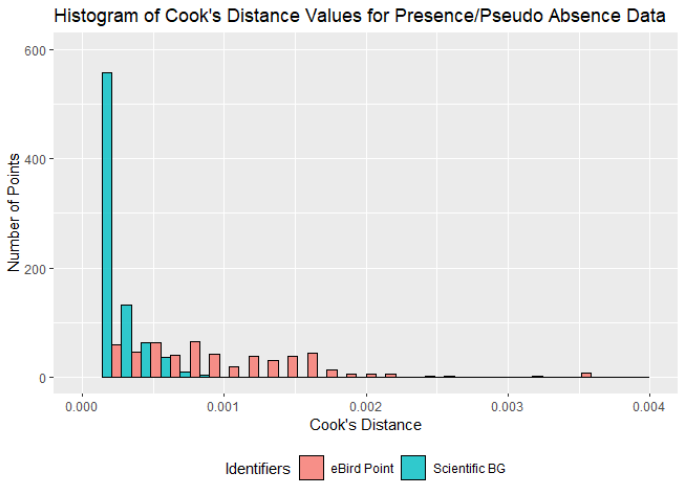


Figure 4.5 Summary from logistic regression (left) exponentiated parameter estimates from logistic regression, (right) histogram of Cook's distance values for presence and pseudo-absence data points

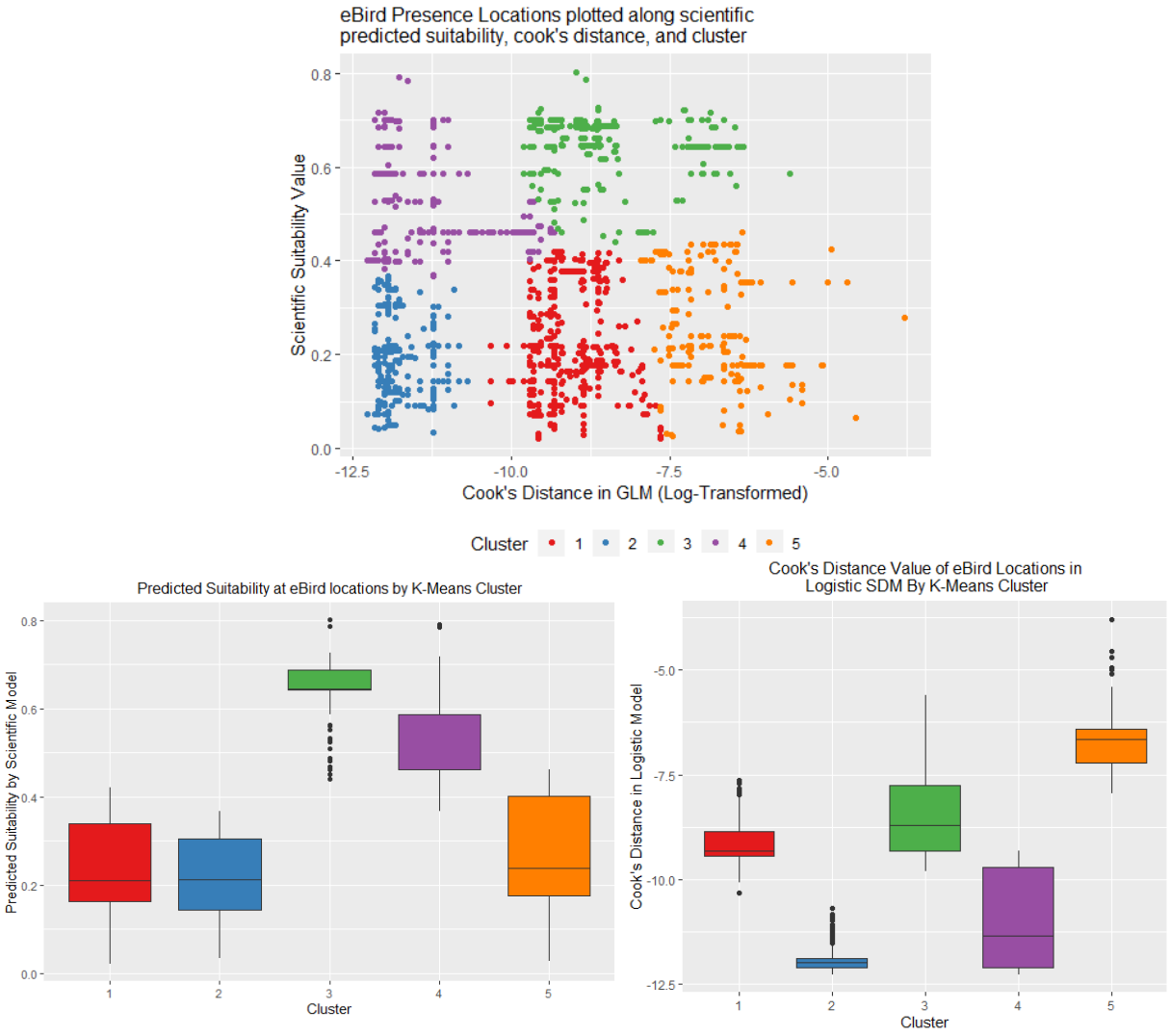


Figure 4.6 eBird points classified by clusters plotted with both explanatory variables: Scientific Suitability and Cook's Distance (top) showing the spread of values along scientific predicted suitability (bottom left) and showing the spread of values along Cook's distance (bottom right)

categorize and understand the cluster attributes, but at an individual level some points have stronger associations and representations of their cluster than others (i.e., points at the center of a cluster compared with those on the fringe near points in other clusters). Cluster 4 (which included ≈ 20% of all eBird points) contained the points with both high predicted suitability by the scientific model and the smallest Cook's distance values in the logistic regression, hence, we concluded Cluster 4 contained information most similar to the scientific dataset.

We confirmed this when we used the points in Cluster 4 to develop a Maxent model (Figure 4.7), which, by contrast to the 68% overlap between the full eBird and scientific model, had high overlap (90%) with the scientific Maxent model (Figures 4.2 & 4.8, Table 4.3). This indicated that, through our research methods, we could isolate eBird points with a relatively similar profile to the scientific points, and thus identify points that were relatively more distinct. When we compared the model developed with only eBird points from Cluster 4 with a model developed with the full eBird dataset but without points that existing eBird data quality filtering mechanisms had flagged, we observed that removing flagged records did not result in a model with better diagnostics (Flagged eBird Model AUC = 0.73 and Test Gain = 0.38; Cluster 4 Model AUC = 0.85 and Test Gain = 0.933; see Table 4.2) nor one that was more similar to the scientific model (overlap of 69%; Figures 4.2 & 4.8 and Table 4.3). Our results suggest that corroborating eBird points with authoritative/scientific data, compared with eBird data quality filters which are based on historic eBird submissions, result in stronger performing models that better align with scientific knowledge about the species.

When we analyzed the environmental characteristics (specifically those identified as important in the Maxent models) of the scientific points (Appendix 4.3) and the five eBird clusters (Figures 4.9, 4.10 & 4.11), we confirmed that presence locations in Cluster 4 had environmental attributes most similar to the scientific dataset (i.e., low-elevation and generally more unforested areas, in land covers including agricultural and developed vegetation, shrub & herb vegetation, and forests & woodlands). We further explored the environmental characteristics of eBird presence locations along their Cook's distance value (Figures 4.9 and 4.11) and predicted suitability by the scientific model (Figures 4.9 and 4.10) in order to understand how differences in these characteristics might provide insight regarding discrepancies between the eBird and scientific model outputs. For instance, points in Cluster 5, the smallest

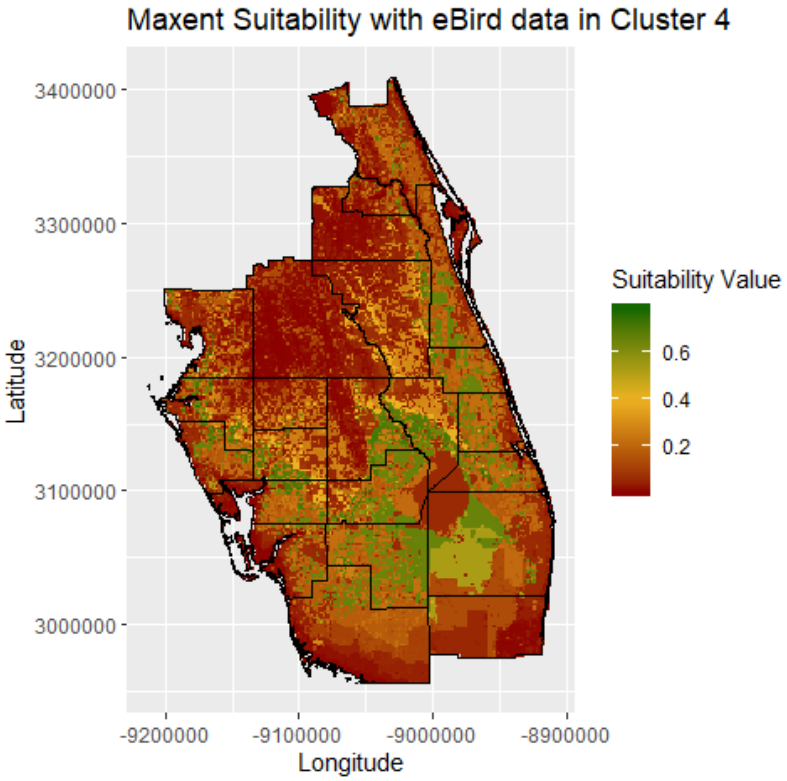


Figure 4.7 Maxent model output using eBird points in Cluster 4, with the highest scientific suitability and smallest Cook's distance in the logistic regression model

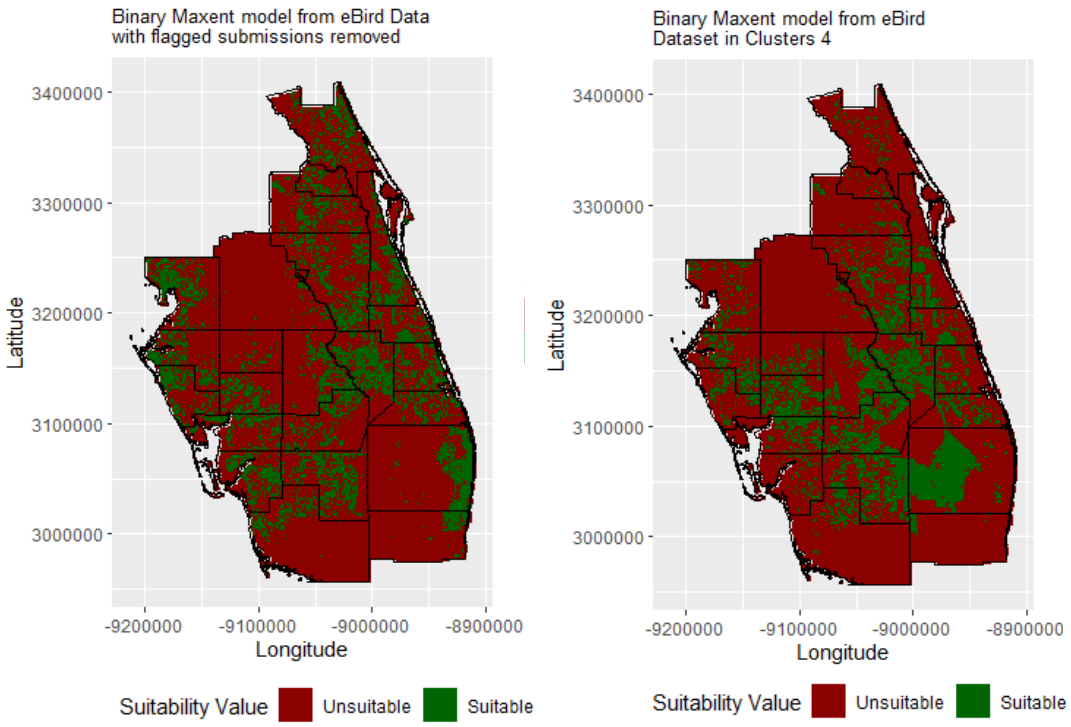


Figure 4.8 Binary model outputs from Maxent models developed with the eBird dataset without locations flagged by eBird (left) and with eBird points in Cluster 4 (right)

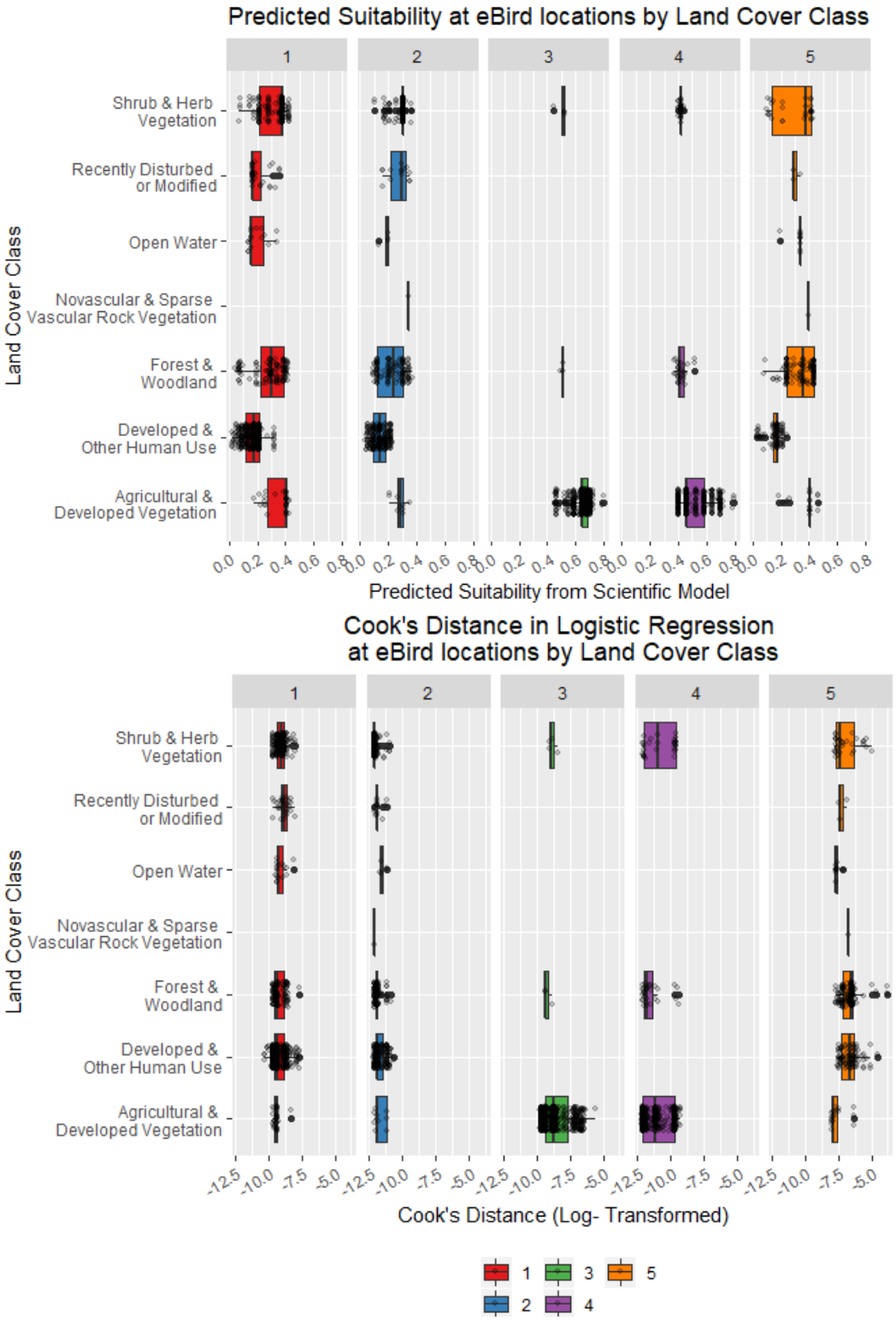


Figure 4.9 Cook's distance (in the logistic regression) and predicted suitability in the scientific model classified by clusters across land cover classes

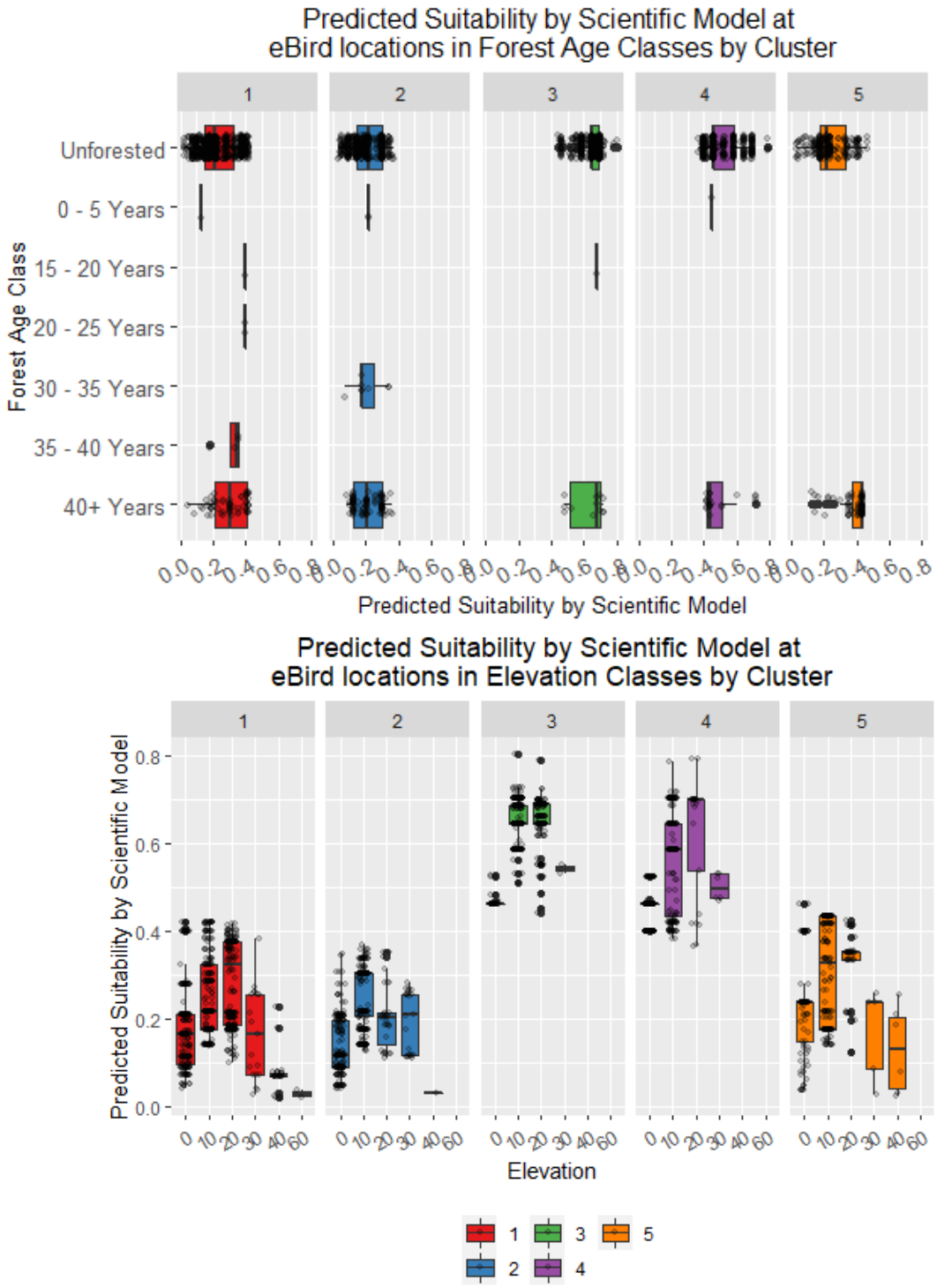


Figure 4.10 Predicted suitability by scientific model at eBird locations classified by cluster across elevation and forest age classes

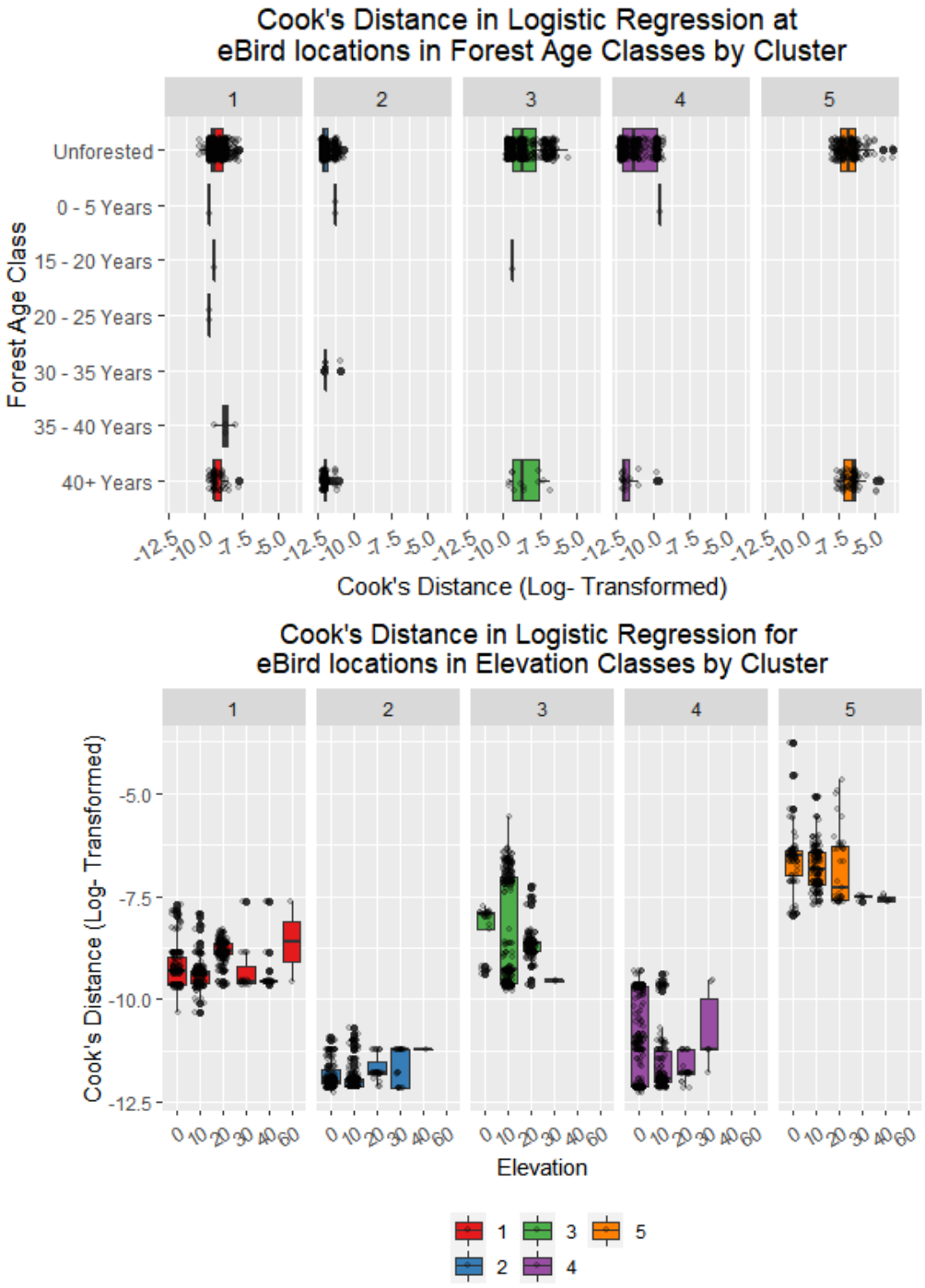


Figure 4.11 Cook's distance for eBird points in the logistic regression classified by cluster across elevation and forest age classes

cluster containing ≈ 11% of the dataset, were in areas predicted unsuitable by the scientific model and had larger Cook's distance values in the logistic regression, in part because points in this cluster had attributes most dissimilar to the scientific dataset—they were likely in areas with a combination of higher elevations, disproportionately older forests, and in a range of land cover classes that deviate from the known attributes of the species' habitat. Hence the relative infrequency of environmental attributes of points in Cluster 5–particularly land cover–explained why these locations were in areas of low suitability predicted by the scientific model and had a larger Cook's distance compared with the rest of the dataset.

Points in Cluster 3 (≈ 20% of the dataset) were the only ones in the dataset aside from points in Cluster 4 that were in areas predicted suitable by the scientific model; however, they had larger Cook's distance values in the regression model. All of the locations in Cluster 3 were in the preferred land cover of the species, and in many ways mirrored the attributes of points in Cluster 4, explaining the high suitability of presence locations. However, points in Cluster 3 did have proportionally more points in higher elevations than Cluster 4. Points in Cluster 1, the largest cluster with ≈ 26% of the dataset, had a similar profile to those in Cluster 5, as they were in areas of lower scientific predicted suitability and also had a relatively large Cook's distance. While Cluster 1 had the greatest number of points, these points had different environmental characteristics from one another: The cluster had the fullest distribution of forest age classes compared with other clusters, and also contained locations in the widest range of elevations in the dataset, that within-cluster environmental variation was relatively higher than in other clusters (Figures 4.10 & 4.11). Similar to points in Cluster 5, points in Cluster 1 were found in a wide range of land covers (Figure 4.9), including classes that were only represented by points in the unsuitable range, Clusters 1, 2, & 5, in areas classified as developed & other human use, open water, and recently disturbed or modified areas.

Cluster 2, containing ≈ 23% of the dataset, contained points with a low suitability value based on the scientific model and a relatively smaller Cook's distance in the logistic regression compared with Clusters 1, 3 and 5. We observed that the lower levels of influence may be affected by attributes of locations that were more unsuitable based on scientific inferences but also less distinct than those of Clusters 1 and 5. For example, points in Cluster 2 shared an elevation profile with

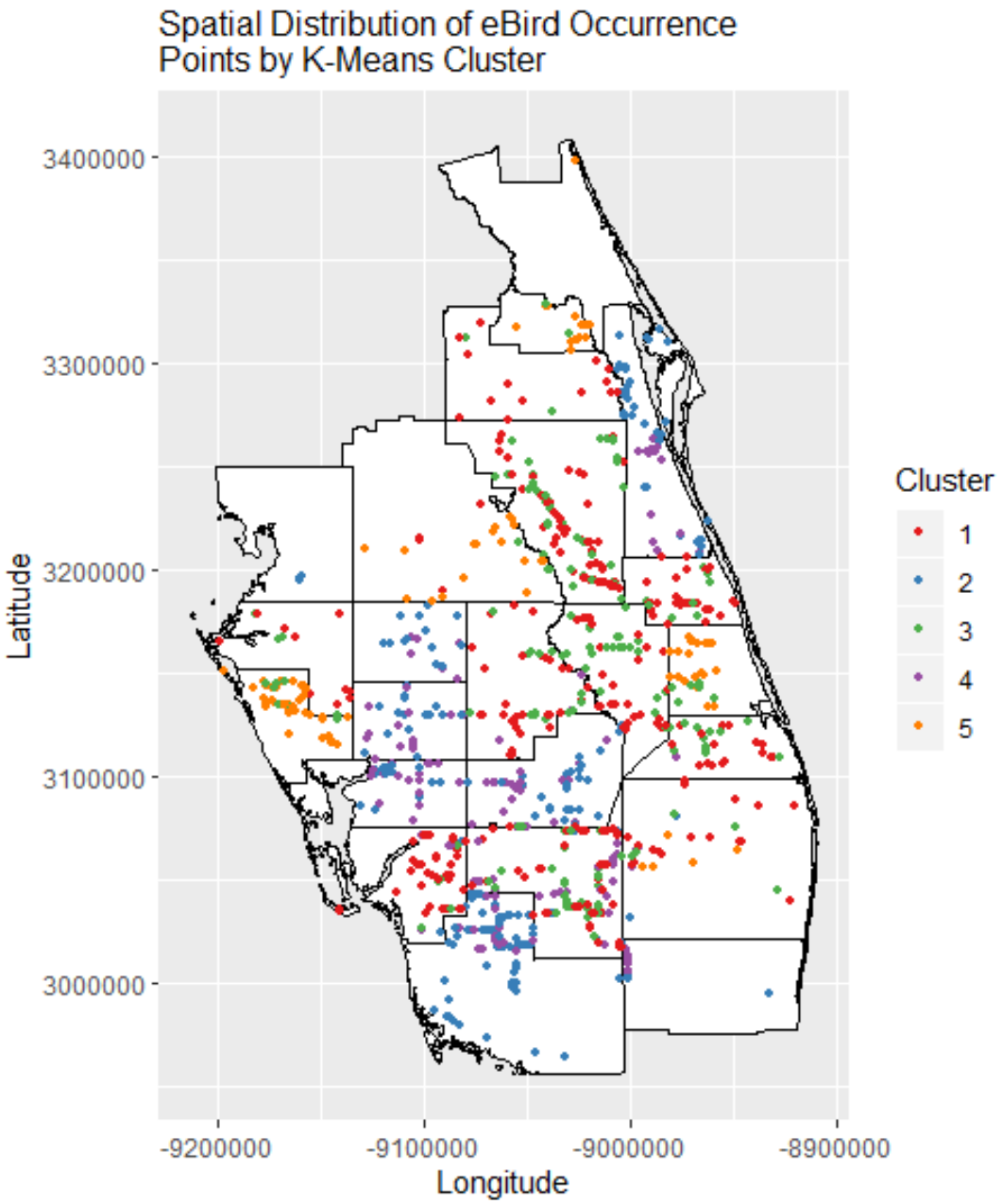


Figure 4.12 The spatial distribution of presence points from eBird classified by cluster

Cluster 4 that was more similar compared to points in Clusters 1 and 5. Land cover attributes of points in Cluster 2 may help explain the low predicted suitability by the scientific model considering that, in addition to Cluster 5, the cluster contained every land cover class in the dataset. Cluster 2 contained a relatively large portion of points in developed and other human use, and fewer points than other clusters in agricultural and developed vegetation.

Some spatial patterns emerged when these clusters were plotted geographically (Figure 4.12), which is notable as neither the models nor the clustering algorithm had explicit spatial parameters or components. The points with the largest influence and low suitability values according to the scientific model (in Cluster 5) had a distinct spatial pattern, concentrating in patches on the fringes of the study area. This indicated that their environmental attributes were unusual, but so were their spatial locations. A related aspect of Cook's distance is that points are less influential if they are more frequent, even if characteristics deviate from the norm. By comparing the spatial distribution of scientific points (Figure 4.1) and eBird points classified by cluster (Figure 4.12), we observed that points in Cluster 4 generally overlapped with the scientific points. However, there were also a number of areas that contained scientific points, with none of those classified in Cluster 4, but with points in Cluster 3, an indication that some points in Cluster 3 are perhaps more representative of observations in the scientific dataset and their spatial and environmental attributes than points in Cluster 4.

Additionally, we observed that points from Clusters 1 and 3 and points from Clusters 2 and 4, respectively, generally occurred within proximity of one another. This is notable as Clusters 1 and 2 were predicted unsuitable, while Clusters 3 and 4 were predicted suitable. We examined this further by exploring the proximity of predicted suitable habitat–defined by thresholding the scientific model to the lowest scientific predicted suitability value in Cluster 4–to points in clusters with low scientific predicted suitability (i.e., Clusters 1, 2, and 5) to understand the ways proximity and locational accuracy might affect the eBird dataset (Figure 4.13). We found that many of the points in these clusters were in close proximity (< 500 m) to suitable habitat, helping explain the occurrence; however, those that were not proximate to suitable habitat–determined by the scientific model–might be reviewed for errors given both their environmental characteristics and spatial attributes.

4.4 Discussion

Our results demonstrate that while models developed with scientific datasets and eBird datasets might have competitive performance diagnostics, the VGI dataset contained different information that resulted in differences between the two model outputs. To better understand the additional information provided by the eBird dataset, we further examined these differences at the observation level through the cluster analysis using the scientific predicted suitability at eBird presence locations and Cook's distance values in the regression model as input variables. Our analysis showed that 20% of eBird points, in Cluster 4, had similar attributes to the scientific dataset, and thus a high scientific suitability and small Cook's distance, while others deviated in varying degrees from the environmental profiles at scientific presence locations (Figures 4.9, 4.10 & 4.11). We found that the environmental characteristics of points in Cluster 4 corroborated scientific knowledge of the species' habitat, and were in agricultural and unforested landscapes, as well as in flat and low-lying elevation. Geographically, we observed that points in Cluster 4 (Figure 4.12) generally matched the spatial profile of the scientific points (Figure 4.1) and the thresholded scientific model (Figure 4.2).

More complex than simply reliable or unreliable, results from the cluster analysis highlighted that eBird points deviated from the scientific presence data in multiple and distinct ways. The two variables used to develop the clusters have different implications. The suitability predicted by the scientific model (Figures 4.9 & 4.10) is affected by the environmental characteristics at eBird presence locations and how they match the environmental characteristics of the scientific data. By comparison, individual eBird observations with high influence were those that had less common or more infrequent environmental characteristics compared to the larger eBird dataset, in addition to how they relate to the pseudo absence points sampled from the scientific model (Figures 4.9 & 4.11). Hence, the cluster analysis helped us categorize locations based on environmental suitability inferred from the scientific model, and also based on the frequency of their environmental attributes in the context of both datasets. Further exploration of these observations can indicate if these points are infrequent but nonetheless valid, and thus of interest for SDM research; or if the points are infrequent because, at best, they do not reflect the habitat

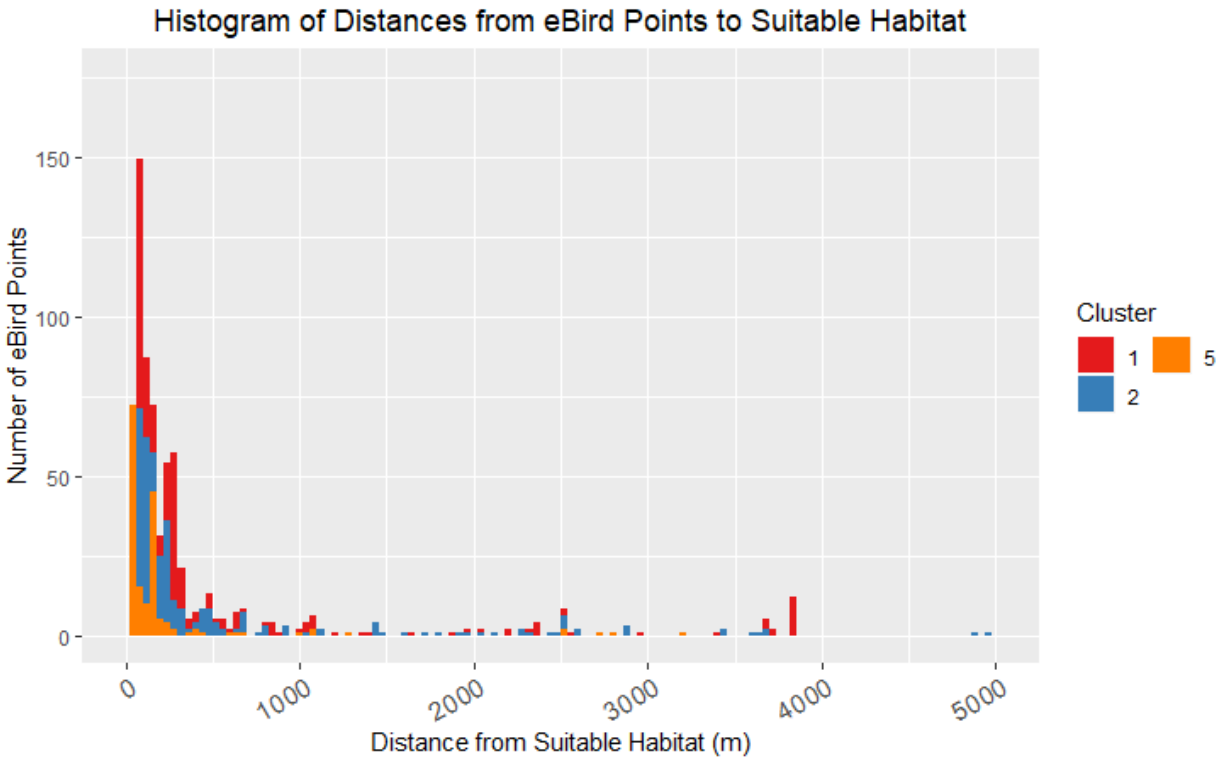


Figure 4.13 Histogram of distance from suitable habitat for clusters of eBird points with relatively low scientific suitability, the threshold used to define habitat was the lowest predicted suitability value from points in Cluster 4

of the species and are not useful, and at worst, they misrepresent species' habitat and degrade SDMs.

Points that were in areas predicted unsuitable by the scientific model in Clusters 1, 2, and 5 had environmental characteristics that were more similar to the pseudo-absence points in the logistic regression (Appendix 4.2) than points in the scientific dataset relative to the other clusters. For example, a number of eBird points in these clusters were found in land cover cells classified as Open Water (Figure 4.9). While it is possible these points might be valid occurrences of the species–for example, a birdwatcher identified the bird in flight on a boating trip–they also could be the result of locational accuracy issues from a volunteer–in habitat nearby these areas with a less precise or even faulty GPS. However, independent of the validity of the observations, including these points in SDMs would result in misleading models as this species is not part of coastal or aquatic ecosystems. While in our experiment we used a common resolution and environmental dataset for SDMs, some of these cases also could be related to the spatial resolution of environmental variables in model development– for example, locations were in suitable habitat nearby water but the size of cells led to presence locations classified in Open Water– which could indicate a scale-based issue in the utility of eBird data for SDMs.

Additionally, points with the largest Cook's distance and lowest scientific predicted suitability, in Cluster 5, had a more distinct spatial pattern than the rest of the eBird points as these points were located in patches on the edges of the study area (Figure 4.12). Thus, these points represented more unique environmental characteristics (Figures 4.9 & 4.10) and therefore had high influence (Figure 4.11). After corroborating citizen science data with a scientific dataset using these methods, eBird can use findings like these to further improve its data review process. For example, eBird could use this method identify observation locations with high influence and low scientific suitability, that were also spatial outliers, and evaluate their utility and/or validity. For scientists and wildlife professionals, occurrences with a higher scientific suitability value, relative to others within the cluster, but without previously recorded occurrences, could represent frontier individuals to a new environment or individuals moving in hospitable corridors among habitat patches. Therefore, more attention might be given to the suitability of these areas to better evaluate these observations and thus better understand the spatial distribution of the species.

One plausible hypothesis to explain the relatively large Cook's distance of points in Cluster 3, despite being in areas of high suitability predicted by the scientific model, is that their environmental characteristics only slightly deviated from those of the scientific presence points, and those of points in Cluster 4, perhaps by only one variable. For example, a larger portion of points in Cluster 3 were in areas of >20 m elevation whereas most points in Cluster 4 and in the scientific dataset were found in elevations of <20 m. Another plausible hypothesis involves trends and potential biases from the scientific data across the study region, which may have manifested themselves in the areas used to sample pseudo-absence points for the logistic regression. There were likely more points representative of some subsets of environmental conditions where the species is present in the scientific dataset than others. Thus, some conditions which are suitable for the species may have been underrepresented in the scientific dataset, possibly in areas where there were scientific points but only eBird points in Cluster 3, not Cluster 4. Hence, the relatively infrequent environmental characteristics of these points in the scientific dataset may have resulted in these observations having a relatively larger Cook's distance in the logistic regression, but their records in the scientific dataset resulted in the scientific model predicting these locations as suitable. Thus, many points in Cluster 3 likely have the same utility as points in Cluster 4 for SDMs and are more likely to be representative of locations where the species is present compared with points in Clusters 1, 2, and 5.

The relative infrequency of environmental attributes of points in Cluster 1 helps explain why these points had low scientific predicted suitability and relatively large Cook's distance values (Figures 4.9 & 4.11). Points in Cluster 1 were present at more forest age classes compared with other clusters, were located in higher elevations than other observations in the dataset and were found in a diverse range of land cover classes compared to the scientific presence points. Comparatively, the environmental attributes of points in Cluster 2 help explain why these points had a lower scientific suitability and smaller Cook's distance values, as the attributes did not match the environmental profile of scientific points and were more common (this cluster contained a large number of points in Developed & Other Human Use) compared to the rest of the dataset. However, a number of points in these clusters were found in land cover classes like Forest & Woodland or Shrub & Herb Vegetation and may represent areas that are not ideal, but still suitable, for the species. Hence, we hypothesized that eBird points with attributes in Cluster

1 and 2 could likely be explained by differences in sampling protocols, as wildlife professionals and scientists likely survey for the species explicitly in areas that contain known attributes of the species' habitat, while citizen scientists might log observations of individuals while on a walk in their neighborhood as well as nearby natural areas that are suitable to the species.

We also found points in Clusters 1 and 2 were in close spatial proximity to points in Clusters 3 and 4 respectively (Figure 4.12), and therefore many of these observations were nearby areas with a high scientific predicted suitability (Figure 4.13). Further, although there were considerable differences in environmental attributes, there were also many points in Clusters 1 and 2 that shared environmental characteristics with those in Clusters 3 and 4. Hence, as we reasoned that locations in Cluster 4 were most like the scientific observations, and Cluster 3 most closely resembled points in Cluster 4, we hypothesized that points in Clusters 1 and 2 were from volunteers who encountered the species in nearby ideal or near ideal areas for the species, some of which were relatively more suitable by certain environmental characteristics, particularly points at the fringe of these clusters nearby Cluster 4 in Figure 4.6. Plausible explanations for these observations include both irregularities involving the observed bird and the citizen scientist. Perhaps the individual bird moved between habitat patches or was disoriented in non-habitat areas, or perhaps the citizen scientists incorrectly submitted their location nearby suitable habitats where they observed the species. Based on our analysis of environmental characteristics, these submissions may represent areas where individuals are found but may not represent the areas with necessary conditions or resources for the population and thus, they may not be useful for SDMs.

In summation, the eBird dataset provided a variety of additional information about the species' distribution relative to the scientific dataset. Through exploring the environmental characteristics of eBird locations, clustered by their influence and scientific predicted suitability, we observed that eBird points fell into five different categories:

1. Observations that corroborate the scientific data spatially, and are found in locations with environmental characteristics that reflect known aspects of the species distribution
2. Observations in locations that almost match the spatial and environmental profile of scientific dataset, but to a lesser degree than others deviate from that profile (i.e., ideal, with exception of one variable)

3. Observations in locations that were hospitable but not ideal to individuals, and likely in areas lacking the necessary resources for the long-term survival of the species
4. Observations in locations that were nearby suitable areas, but based on scientific knowledge are unsuitable for the species (hence either locational accuracy is questionable, or the species is traversing a non-habitat area between habitat patches)
5. Observations in locations with unique spatial attributes and environmental characteristics that were the most distinct in the context of the full dataset and in areas that scientific knowledge suggest are unsuitable, and thus warrant data review

To reiterate a point about K-means clustering, the point membership in clusters is relative. As this method was the basis for these categories it indicates that while some eBird locations were identified most clearly in third category (hospitable but not ideal), some points in that category were nonetheless in relatively more ideal areas and more similar to the scientific dataset than others. Still, these categories can be useful to help scaffold discussions about the additional information that observation datasets from citizen scientists can provide about species distributions.

The differences that we explored between scientific and volunteered presence datasets intersect central conceptual questions involved when developing SDMs. Researchers have developed important theoretical frameworks to understand how different modeling approaches, along with their inputs, mechanics, and assumptions, might represent different aspects of species distributions. One assumption related to this experiment regards *Perfect versus Imperfect Habitat Use* (Laurent et al., 2011). The assumption refers to the tendency of SDMs to dismiss areas where the species might be present because these conditions were not perfect: They are not the ideal or best areas incorporating all necessary factors for the survival of the species. However, the crested caracara, particularly through the loss of natural prairie habitat, has become adapted and more common in human landscapes (i.e., agricultural), which would generally be considered imperfect habitat. As our findings indicate, the crested caracara might be found foraging along highway, corroborating findings from Morrison and Pias (2006) or in a neighborhood with non-natural and imperfect environmental characteristics compared to the ecological knowledge of the species' perfect habitat conditions. In consideration of this assumption, the eBird dataset might be employed when trying to model the broader spatial distribution of the species, areas of both

perfect and imperfect habitat, while the scientific dataset might be best used when trying to model areas specific to the species' natural niche requirements, only perfect habitat.

The methods employed in this research can help broaden questions of data quality for species presence dataset by offering multiple ways to characterize citizen science datasets. For instance, instead of assessing the reliability of eBird observations, we assessed their unique information based on inferences from the scientific model and relative to their frequency in the full dataset for a more specific characterization of the data. Broadly, the use of scientific data points as a benchmark to assess the quality of volunteered wildlife data points can both help guide the practice of citizen science data collection and provide a means to qualify knowledge from citizen science data for use by the scientific community. For example, based on our findings, eBird could require photos of occurrences from citizen scientists in areas where there were points with high influence and low scientific suitability. Conversely, scientists might be guided by areas with high influence to evaluate the sensitivity of the species to environmental variables, or to better understand the anthropogenic environments (i.e., suburban versus rural) where crested caracara are more likely to be present.

There were notable limitations in our experiment. First, to ensure the method works across taxa, this analysis could be corroborated with a scientific dataset with occurrences of other species with different spatial ecologies. Second, the different temporal resolutions of the two datasets might have affected our results—the scientific dataset was compiled over several years with data input at inconsistent time intervals, compared to near daily input of data input from eBird over a single year. This could result in mismatches with the environmental variables, like changing land covers, and so future work might consider different ways to filter observations based on their collection dates. Finally, though often necessary in practice, model thresholds are an arbitrary and debated aspect of SDMs. The threshold used to determine the area to sample pseudo-absence points, as well as the threshold used to compare the model outputs, could be reexamined, and future analyses might examine a range of thresholds to test if the results are consistent or optimized at any particular threshold.

4.5 Conclusions

In the case of SDMs, data collection protocols in a VGI context have many differences compared with those for authoritative data sources that manifest themselves in different data products. We found that inferences from a scientific SDM and the statistical concept of inference could help identify eBird presence points that were most similar to a scientific presence dataset. By exploring the remaining points, we could better understand these differences and the additional information provided by the eBird dataset. We found that eBird points deviated in many ways. For example, while one segment of points were outliers spatially in rare locations for the species to be seen, another subset of points was in areas that were likely hospitable for individuals to traverse but not the conditions necessary for the long-term survival of a population. This information can be used by eBird to build on their data filtering mechanisms or their data input infrastructure, by ecologists to better understand species' sensitivity to nonhabitat areas and different environmental gradients, and by data scientists to better understand the differences between VGI and authoritative data.

Works Cited

Amano, T., Lamming, J. D., & Sutherland, W. J. (2016). Spatial gaps in global biodiversity information and the role of citizen science. *Bioscience*, *66* (5), 393-400.

Anselin, L., Syabri, I., & Kho, Y. (2010). GeoDa: an introduction to spatial data analysis. In *Handbook of Applied Spatial Analysis* (pp. 73-89). Springer, Berlin, Heidelberg.

Aubry, K. B., Raley, C. M., & McKelvey, K. S. (2017). The importance of data quality for generating reliable distribution models for rare, elusive, and cryptic species. *PLOS ONE*, *12* (6).

Barbet-Massin, M., Jiguet, F., Albert, C. H., & Thuiller, W. (2012). Selecting pseudo-absences for species distribution models: how, where and how many? *Methods in Ecology and Evolution*, *3* (2), 327-338.

Bonney, R., Cooper, C. B., Dickinson, J., Kelling, S., Phillips, T., Rosenberg, K. V., & Shirk, J. (2009). Citizen science: a developing tool for expanding science knowledge and scientific literacy. *BioScience*, *59* (11), 977-984.

Bruns, A. (2008). *Blogs, Wikipedia, Second Life, and Beyond: From Production to Produsage.* New York: Peter Lang.

Cardador, L., Carrete, M., Gallardo, B., & Tella, J. L. (2016). Combining trade data and niche modelling improves predictions of the origin and distribution of non-native European populations of a globally invasive species. *Journal of Biogeography*, *43* (5), 967-978.

Clark, C. J. (2017). eBird records show substantial growth of the Allen's Hummingbird (Selasphorus sasin sedentarius) population in urban Southern California. *The Condor: Ornithological Applications*, *119* (1), 122-130.

Cooper, C. B., Dickinson, J., Phillips, T., & Bonney, R. (2007). Citizen science as a tool for conservation in residential ecosystems. *Ecology and Society*, *12* (2).

Crall, A. W., Newman, G. J., Stohlgren, T. J., Holfelder, K. A., Graham, J., & Waller, D. M. (2011). Assessing citizen science data quality: an invasive species case study. *Conservation Letters*, *4* (6), 433-442.

Danielsen, F., Burgess, N. D., & Balmford, A. (2005). Monitoring matters: examining the potential of locally-based approaches. *Biodiversity & Conservation*, *14* (11), 2507-2542.

Elith, J., Phillips, S. J., Hastie, T., Dudík, M., Chee, Y. E., & Yates, C. J. (2011). A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions*, *17* (1), 43-57.

[Dataset] eBird. 2019. eBird: An online database of bird distribution and abundance [web application]. eBird, Cornell Lab of Ornithology, Ithaca, New York. Available: http://www.ebird.org. (Accessed: Date [June 13, 2019]).

[Dataset] Florida Forest Service, 2013. Stand Age Raster Grid, https://freshfromflorida.s3.amazonaws.com/TimberStandAge.zip

[Dataset] Florida Natural Areas Inventory. Crested Caracara Presence Data, December 2018.

[Dataset] Gesch, D., Oimoen, M., Greenlee, S., Nelson, C., Steuck, M., & Tyler, D. 2002. The National Elevation Dataset. *Photogrammetric engineering and remote sensing*, 5-32.

Elith, J., Phillips, S. J., Hastie, T., Dudík, M., Chee, Y. E., & Yates, C. J. (2011). A statistical explanation of MaxEnt for ecologists. *Diversity and Distributions*, *17* (1), 43-57.

Hurlbert, A. H., & Liang, Z. (2012). Spatiotemporal variation in avian migration phenology: citizen science reveals effects of climate change. *PLOS ONE*, *7* (2).

Han, J., Pei, J., & Kamber, M. (2011). *Data Mining: Concepts and Techniques*. Elsevier.

Kelling, S., Gerbracht, J., Fink, D., Lagoze, C., Wong, W.K., Yu, J., Damoulas, T. and Gomes, C., 2013. A human/computer learning network to improve biodiversity conservation and research. *AI magazine*, *34* (1), 10-20.

Ketchen, D. J., & Shook, C. L. (1996). The application of cluster analysis in strategic management research: an analysis and critique. *Strategic Management Journal*, *17* (6), 441-458.

Kutner, M. H., C. J. Nachtsheim, J. Neter, and W. Li. (2005). Applied Linear Statistical Models. Page IOP Conference Series: Materials Science and Engineering. Fifth Edit. McGraw-Hill Irwin, Boston.

Laurent, E. J., Drew, C. A., & Thogmartin, W. E. (2011). The role of assumptions in predictions of habitat availability and quality. In *Predictive Species and Habitat Modeling in Landscape Ecology* (pp. 71-90). Springer, New York, NY.

McCormack, J. E., Zellmer, A. J., & Knowles, L. L. (2010). Does niche divergence accompany allopatric divergence in Aphelocoma jays as predicted under ecological speciation?: insights from tests with niche models. *Evolution: International Journal of Organic Evolution*, *64* (5), 1231-1244.

Merow, C., Smith, M. J., & Silander Jr, J. A. (2013). A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. *Ecography*, *36* (10), 1058-1069.

Morrison, J. L. (2001). *Recommended management practices and survey protocols for Audubon's Crested Caracara (Caracara cheriway audubonii) in Florida* (No. 18). Technical Report.

Morrison, J. L., & Humphrey, S. R. (2001). Conservation value of private lands for Crested Caracaras in Florida. *Conservation Biology*, *15* (3), 675-684.

Morrison, J. L., & Pias, K. E. (2006). Assessing the vertebrate component of the diet of Florida's crested caracaras (Caracara cheriway). *Florida Scientist*, 36-43.

Phillips, S. J., & Dudík, M. (2008). Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. *Ecography*, *31* (2), 161-175.

Phillips, S. J., Dudík, M., & Schapire, R. E. (2004). A maximum entropy approach to species distribution modeling. In *Proceedings of the twenty-first international conference on Machine learning* (p. 83).

Ramesh, V., Gopalakrishna, T., Barve, S., & Melnick, D. J. (2017). IUCN greatly underestimates threat levels of endemic birds in the Western Ghats. *Biological Conservation*, *210*, 205-221.

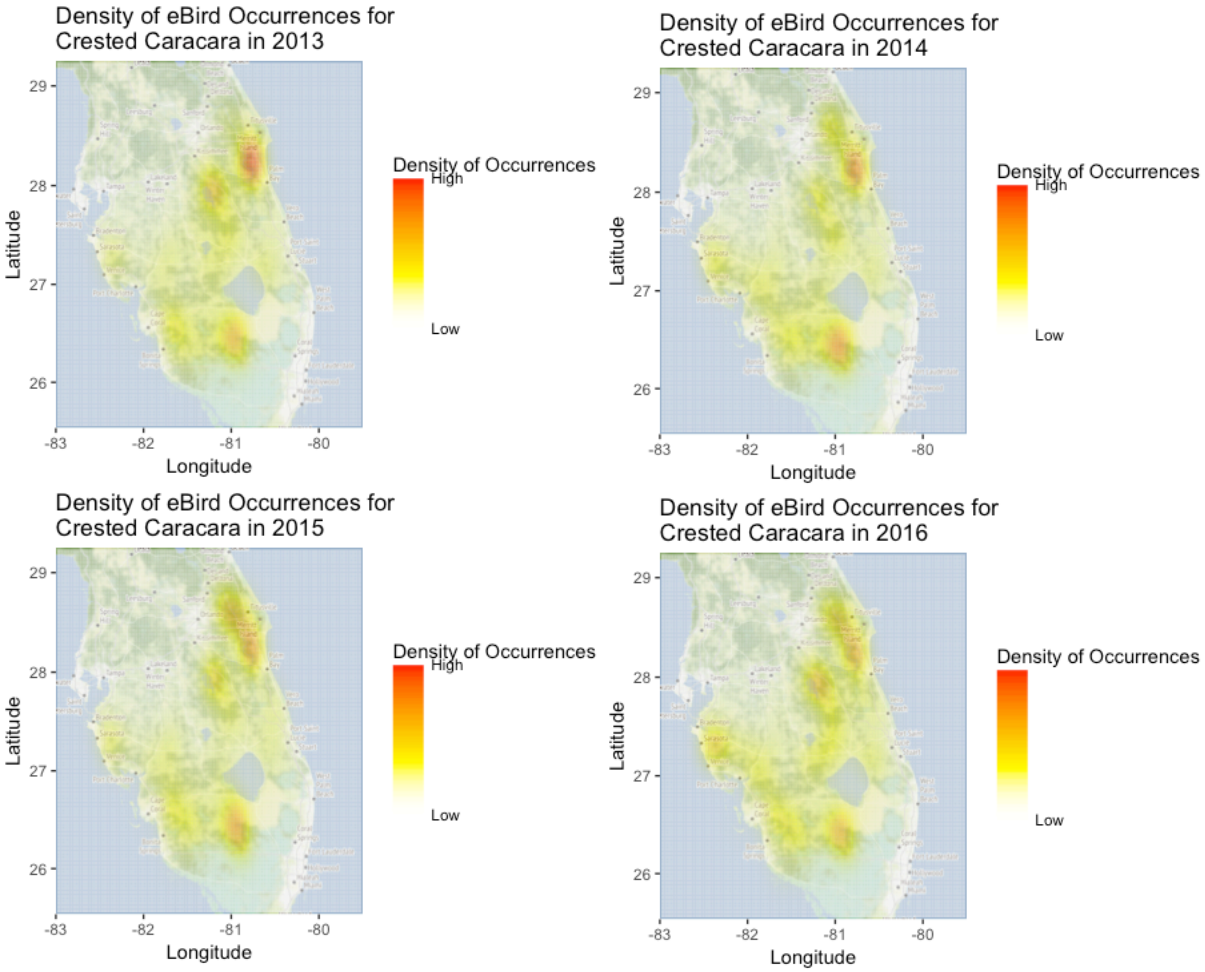
Sullivan, B.L., Aycrigg, J.L., Barry, J.H., Bonney, R.E., Bruns, N., Cooper, C.B., Damoulas, T., Dhondt, A.A., Dietterich, T., Farnsworth, A. and Fink, D. (2014). The eBird enterprise: an integrated approach to development and application of citizen science. *Biological Conservation*, *169*, 31-40.

[Dataset] U.S. Geological Survey Gap Analysis Program, GAP/LANDFIRE National Terrestrial Ecosystems 2011: U.S. Geological Survey, https://doi.org/10.5066/F7ZS2TM0.

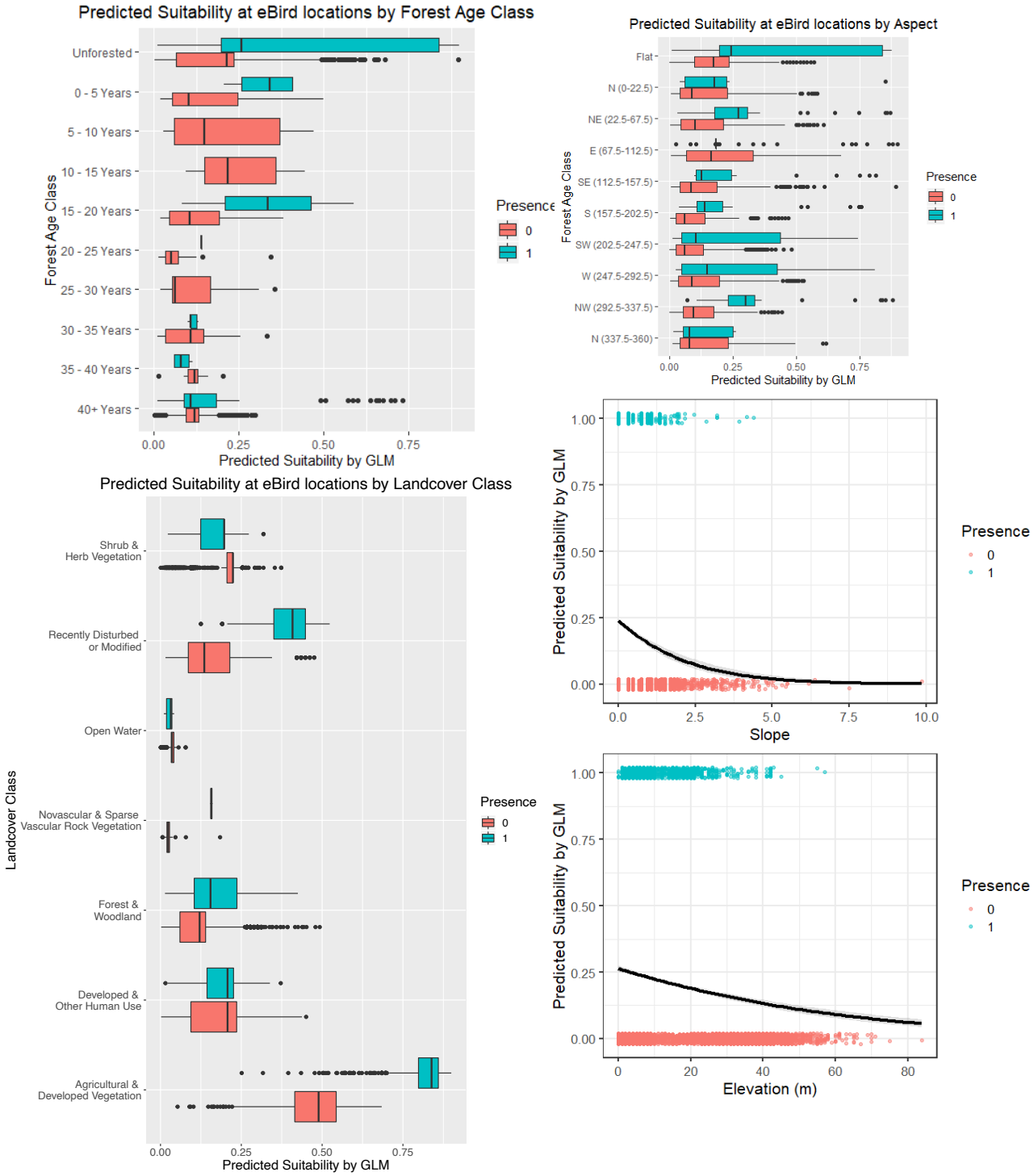
Wisz, M. S., & Guisan, A. (2009). Do pseudo-absence selection strategies influence species distribution models and their predictions? An information-theoretic approach based on simulated data. *BMC ecology*, *9* (1), 8.

Yu, J., Wong, W. K., & Hutchinson, R. A. (2010). Modeling experts and novices in citizen science data for species distribution modeling. In *2010 IEEE International Conference on Data Mining* (pp. 1157-1162). IEEE.

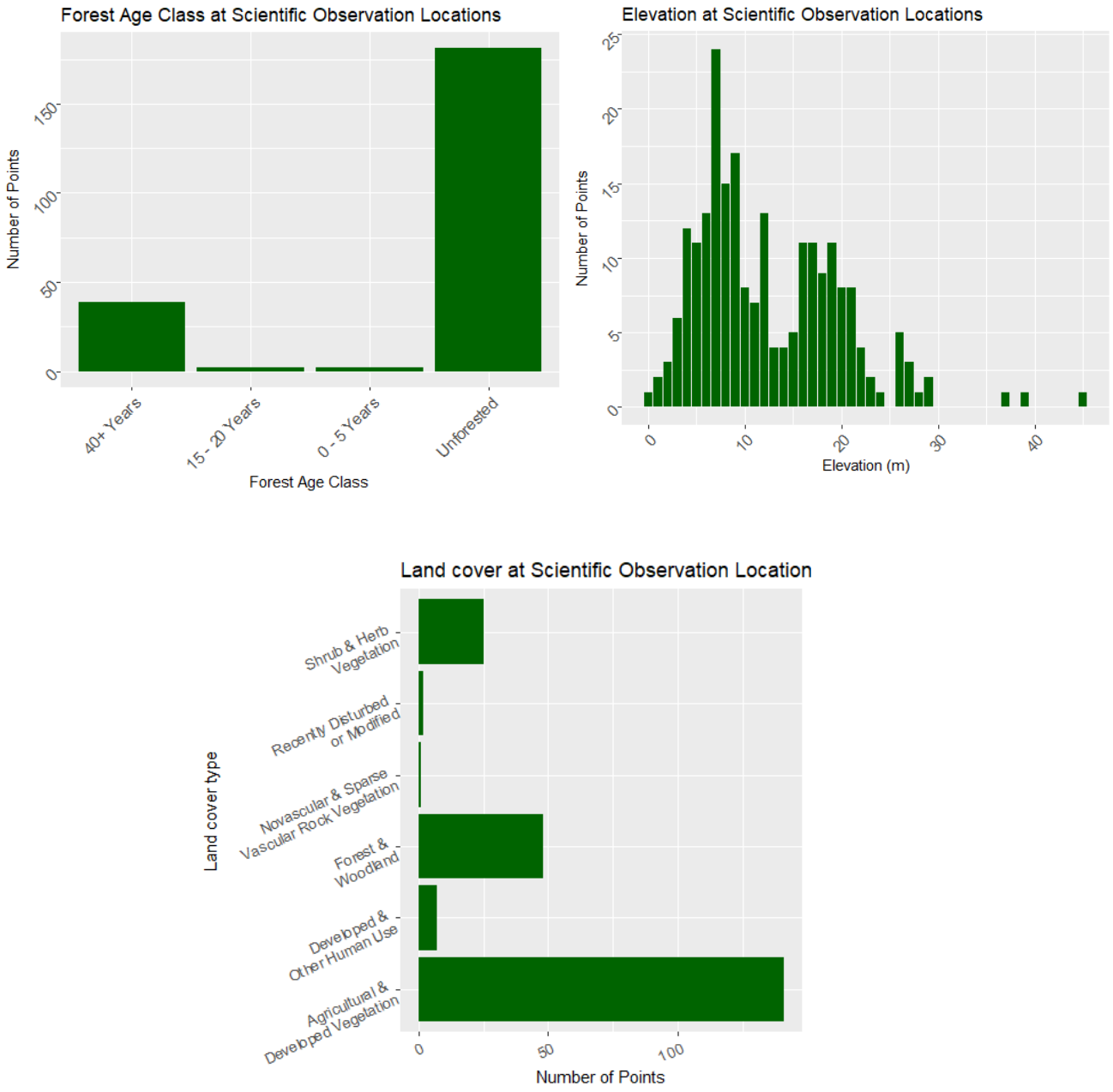
Appendix



Appendix 4.1 Density plots demonstrating that the footprint of eBird presence points for crested caracara is relatively consistent at annual intervals



Appendix 4.2 Environmental characteristics at presence and pseudo-absence points from the regression model



Appendix 4.3 Environmental characteristics of Crested Caracara presence points from Florida Natural Areas Inventory (FNAI)

CHAPTER 5 SUMMARY AND FUTURE RESEARCH DIRECTIONS

4.6 Summary of Research Questions and Study Objectives

In Chapter 1, I detailed my research goals in this dissertation and contextualized the work within the field of Geography. I explored the following research questions in this dissertation:

(1) **How can spatial data with different resolutions be used to improve species distribution models?**
(2) **Are wildlife observation data from citizen scientists valid, and what additional information can they contribute to species distribution models?**

I suggested that this research intersected two subfields in Geography: GIScience (basing our methodologies on geographic data analysis) and Biogeography (providing the context to link ecological theory, spatial data methods, and the environmental characteristics of species' niche). I aimed to contribute knowledge about the use of two emerging big data sources in Geography. In the first chapter, I presented a method of working with fine-resolution environmental data from lidar that additionally approached issues of scale in Ecology and Environmental Modeling. In the second and third chapters, I explored characteristics of VGI in species distribution models through the manipulation of spatial models and their individual components (i.e., thresholds, absence data). While these studies contributed knowledge about spatial modeling and GIScience for Biogeography, I also suggest that the methodologies employed in these studies helped us better understand big data sources in ways that can be broadly applied in Geography.

4.7 Contextualizing Research within GIScience and Biogeography

My study design and research findings have made contributions to the fields of Biogeography and GIScience in the following ways:

A. In Chapter 2, I introduced a new modeling method that approached a common scale-related issue, input variables with different spatial resolutions, in species distribution modeling. These issues are particularly salient given the increased accessibility to fine-scale environmental data from lidar. This method, which I called scale-based background sampling (SBBS), involved manipulating inputs (i.e., the sampling area for background points) for Maxent models developed at a fine scale based on inferences found at a

broader scale of analysis. Additionally, this study involved the inclusion of biotic components of species niche in SDMs, which has been a particularly relevant challenge given that biotic interactions influence species distributions at finer scales than abiotic ones. Employing this method improved models compared with conventional approaches of working with data at varying resolutions and an understanding of the species' spatial ecology helped corroborate our results and give us confidence in the model improvement mechanisms. Finally, while many have suggested there is a disconnect between various methods to improve SDMs and ecological theory, from my perspective SBBS is rooted in both.

B. In Chapter 3, I evaluated differences between a species distribution model developed with citizen science data and one developed with data collected by scientists and wildlife professionals. To compare the differences and similarities between these two model outputs, I focused on an integral aspect of SDMs: the threshold used to binarize model outputs. By determining the model threshold at which the binarized eBird model had the highest overlap with the scientific model, I was able to analyze model differences across space at the, for lack of a better term, best-case scenario for the VGI model. Thus, I used a model based on scientific, or authoritative, data as a benchmark to understand the type of areas where VGI-based models perform similarly or dissimilarly. Based on prior scientific knowledge about the species, I could confirm areas the scientific model predicted suitable. However, given that past work has found the species has become adapted to human environments, I also recognized the eBird model could be uniquely useful to understand the species' sensitivity to varying degrees of human development and anthropogenic landscapes. Thus, connecting domain knowledge to spatial modeling methods allowed me to understand more about the utility of species distribution models developed with citizen science data.

C. In Chapter 4, I began with a premise that the statistical measure of influence can help shed light on the quality of eBird points for species distribution models. However, I was limited in using this measure because I did not have absence data for the species to develop a regression model. To overcome this issue, I created pseudo-absence points by sampling locations from areas predicted unsuitable by the scientific model. Hence, I was

able to create pseudo-absence points with scientific authority to use in a regression model with the eBird presence points. After I categorized the points based on each eBird location's influence (measured by Cook's distance) and its predicted suitability from the scientific model, I investigated the environmental characteristics of each group. I concluded that the environmental characteristics of some eBird locations matched or closely matched the environmental profile of observation locations from scientists or wildlife professionals, while some of the locations warrant data review, and others were likely in hospitable areas but without necessary attributes of niche. Thus, I used knowledge from Biogeography fused with spatial modeling methods to help understand information about VGI data at the individual level.

4.8 Broader Contributions

This dissertation work also contributes more broadly than to the domains of Ecology, GIScience, and Biogeography. While I provided a practical approach to integrate lidar data in SDMs, I also developed a method in which inferences were employed across scales to enhance models. Thus, the merit of the study is in the production of knowledge through the synthesis of information from modeling environmental phenomenon at different scales. As many phenomena occur at multiple spatial scales, this type of approach could be implemented to better understand a range of environmental, social, and economic issues.

I approached important data quality issues with VGI at two levels of analysis. First, I explored VGI at the aggregate level, by investigating the overall difference between models made with scientific data and those made with eBird data. In Chapter 3, I found that on a per county basis, there was an average of 80% concordance between models made with scientific data and those made with eBird data. Hence, in some areas eBird models performed very well, as more than 90% of cells were concordant with the scientific model, but in others the models performed worse as less than 70% of cells were concordant. Upon further analysis, I identified particular environmental characteristics–percent of developed land cover, elevation ranges, habitat fragmentation–that led to model discordance. In my methods to understand concordance between VGI and scientific based SDMs, I recognized eBird based models performed well in certain areas and that particular geographic characteristics made model performance variable

across space. Thus, when developing models using VGI data, researchers could cite work in this chapter to frame how models are performing given characteristics of study areas.

Second, I investigate the quality of VGI with regards to species distribution models at the observation level of analysis. In Chapter 4, my methods prompted a comprehensive discussion of different classes of eBird points, particularly as I did not solely evaluate their quality with a binary (valid/invalid) lens, but instead I identified a number of plausible interpretations. I found that many eBird points defied inferences from the scientific model, and some were so rare they had disproportionate effects on the model outputs. However, though I did recommend those points might go under data review, they were not immediately dismissed as invalid and contextualized within the various differences between eBird and scientific data collection. I anticipate that these frameworks are transferrable to other domains, which is to say that a broader impact of Chapter 4 was the employment of methods to understand individual VGI data quality from a complex perspective that recognizes the differences in data collection protocols and the effects these might have on model outputs.

4.9 Concluding Thoughts and Future Research Directions

I will end by stating some of the future research directions that I intend to follow after earning my Ph.D. First, there are a number of other studies I plan to pursue that build off the work presented in this dissertation. These include a study in which I relate our measures of data quality of individual eBird points to traits of individual birdwatchers (for example, novice or expert), as well as a study in which eBird models and scientific models developed with input data that have different resolutions to understand how the quality of model changes based on the scale of input variables. I have additional ideas that involve more manipulation of model parameters based on species' spatial ecology on which I plan to base future experiments and research articles. I also look forward to exploring further biogeographical phenomena with the red-cockaded woodpecker and crested caracara, especially as their spatial ecology embody important ecological concepts such as niche, mutualism, and human adaptation. I am also particularly interested in the coastal ecosystems along the southeastern United States and plan to contribute to conservation research in these unique environments.

Within the broader realms of Biogeography and Ecology, I intend to continue engaging with research in the realm of conservation science. In particular, I plan to engage with research that promotes the conservation of habitat for rare species and the management of species found along the human-wildland interface using spatial analysis. As a means to this end, I plan to continue contributing research studying species distribution models, and plan to design future studies that focus on landscape ecology. Landscape pattern analysis has proven to be essential for understanding a host of ecological phenomenon in the Anthropocene including population dynamics and human and environment interactions. I am particularly interested in the integration of spatial data at different resolutions in network-based models used to understand habitat connectivity and ecological relationships among habitat patches.

In addition to these focuses at the nexus of Biogeography and GIScience, I also plan to pursue opportunities that allow me to contribute to future research with lidar and VGI. Given the promise that lidar datasets have for environmental analyses by providing fine-scale, 3D information about geography, I plan to continue developing methods to integrate lidar-derived environmental characteristics, including forest stand and patch attributes, in spatial models. I am particularly interested in better understanding the ecological dynamics related to changes to structural forest ecology in the context of wildlife management and conservation. I have also developed a keen academic interest in scale conceptually, and plan to continue exploring the contributions of lidar data to multiscale modeling and an understanding of ecological phenomenon across spatiotemporal scales. Lidar analysis can inform us about important ecological characteristics essential for data-centered wildlife conservation, and I plan to continue highlighting ways it can do so in my future research.

This dissertation research allowed me to develop a more comprehensive understanding of VGI. I plan to continue researching the utility of data provided by citizen scientists in part because of the meaningful engagement VGI facilitates between the public and scientists, but also given the potential for data collection across spatiotemporal scales. With the increased ownership of personal smart devices that enable humans to act as data sensors, there will be a greater need for research involving data management as well as data quality assurance/quality control. Through this lens, I intend to contribute research that supports the increased use of VGI in environmental research so that the potential of these data can be met. I have appreciated learning

about not only the ways VGI is different from authoritative data, but also the ways information from this data source challenges our understanding about phenomena. By helping to better understand the utility of massive sums of data contributed and processed by human beings through a VGI framework—whether it be sightings of endangered species or public infrastructure problems through Open Street Map—I look forward to contributing towards a better understanding about the ways VGI can be used in scientific analysis through my career.

Geography, and specifically work using GIScience, has proven to be an important basis for decision-making during extraordinary circumstances, ensuring not only a more sustainable and equitable society, but a secure one. Through my career, I look forward to using the skills that I developed while earning my Ph.D., including in experimental design and data analysis, to conduct scientific research that addresses the most important issues facing our country and world, from economic inequality to environmental stewardship to public health and emergency management. Thus, my broadest but most certain plan after finishing this dissertation is to use my academic background in Geography and technical skillset in GIScience to address our world's most critical and pressing challenges.

VITA

Adam Guy Alsamadisi was born in Brooklyn, New York, and raised in Bridgewater, New Jersey. Adam received his B.A. from Rhodes College (2012) and M.S. from The University of Tennessee (2015). Adam completed his Ph.D. in Geography from The University of Tennessee in August 2020. Through his time as a Ph.D. student, Adam served as an Academic Advisor in the College of Arts & Sciences and a Graduate Teaching Associate at UTK's Department of Geography, and later taught at The Buckley School and worked as a GIS Analyst in different research contexts. He was awarded the McCroskey Memorial Grant, the Chancellor Top-off Fellowship, and the Outstanding Teaching Associate Award, and was funded to present his research at conferences including AAG, SEDAAG, and the Society of Conservation GIS.