Machine Learning for Biodiversity Monitoring from Remote Sensing and Citizen Science Data

Benjamin Binen Akera

School of Computer Science
McGill University
Montreal, Quebec, Canada

Wednesday 6th December, 2023

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Master of Computer Science

Abstract

Biodiversity loss is occurring at an unprecedented rate, threatening ecosystem services critical to food, water, and human health and well-being. Understanding species distributions is crucial for conservation policy. However, traditional species distribution modeling (SDM) methods focus on limited species or regions, leaving major knowledge gaps. A key barrier is the extensive effort needed for traditional monitoring. Remote sensing and citizen science offer opportunities to transform biodiversity monitoring and enable modeling complex ecosystems.

This thesis introduces the task of mapping bird species to habitats by predicting encounter rates from satellite images and crowd-sourced citizen science data. We create a dataset with satellite images from the US and Kenya with labels derived from presence-absence observation data from citizen science database eBird. We train baseline models and show that we can learn specific distribution patterns from these data. We also show that we can utilize the trained models to improve predictions in areas where data may be limited, specifically, where eBird checklists are limited. The released dataset - SatBird and pre-trained models enable scalable ecosystem modeling worldwide.

Abrégé

La perte de biodiversité se produit à un rythme sans précédent, menaçant les services écosystémiques essentiels à la nourriture, à l'eau et à la santé et au bien-être humains. Comprendre la distribution des espèces est crucial pour la politique de conservation. Cependant, les méthodes traditionnelles de modélisation de la distribution des espèces (MDS) se concentrent sur des espèces ou des régions limitées, engendrant d'importants lacunes de connaissance. Un obstacle majeur est l'effort considérable nécessaire pour une surveillance traditionnelle. La télé-détection et la science citoyenne offrent des opportunités de transformer la surveillance de la biodiversité et de permettre la modélisation d'écosystèmes complexes.

Cette thèse introduit une technique afin de cartographer les espèces d'oiseaux selon les habitats en prédisant les taux de rencontres, et ce à partir d'images satellite et de données de science citoyenne. Nous créons un ensemble de données avec des images satellites des États-Unis et du Kenya avec des étiquettes dérivées des données d'observation de présence-absence de la base de données de science citoyenne eBird. Nous entraînons des modèles de base et montrons que nous pouvons apprendre des schémas de distribution spécifiques à partir de ces données. Nous montrons également que nous pouvons utiliser les modèles formés pour améliorer les prédictions dans les zones où les données peuvent être limitées, en particulier, là où les listes de contrôle d'eBird sont restreintes. L'ensemble de données publié - SatBird et les modèles pré-entraînés permettent une modélisation écosystémique à grande échelle dans le monde entier.

Acknowledgments

I wish to express my sincere gratitude to my advisor, Prof. David Rolnick, for his invaluable mentorship and guidance throughout this research. His extensive expertise and unwavering support were instrumental in shaping this work. I also thank Mila, IVADO, and CIFAR for generously providing funding and computational resources that enabled me to successfully carry out this project.

I am grateful to my colleagues at Mila and CliMath for engaging in stimulating discussions and providing helpful perspectives that assisted in developing my ideas. I would like to extend special thanks to Mélisande Teng, Hager Radi, Amna Elmustafa, Sal Elkafrawy and Shrishti Yadav, with whom I co-authored works presented at NeurIPS and ICLR 2023. This thesis builds upon and extends the research conducted in collaboration with them.

On a personal note, I am eternally grateful for the unconditional love and encouragement of my family. I especially want to thank my dad, Dr. Acer Godfrey Okot, for showing me the way and guiding me to follow in his footsteps. My beloved late mother, Susan, whose memory continues to inspire me. My brothers and sisters have been a pillar of strength, keeping me motivated through this challenging journey. Special thanks to Hugh and Heather Cameron for their unwavering support. And to Cathy, who was a voice of reason when I needed perspective. This accomplishment would not have been possible without my family cheering me on every step of the way.

I extend my appreciation to all who contributed in various ways to this thesis. Friends I made along this journey, near and far, Your support made a difference. *Thank you!*

Dedication

To the future, where endless possibilities await and where our collective efforts will shape a better world for generations to come

To my dad, Dr. Godfrey Acer Okot

To my mum, Susan Akol, may her soul continue to rest in peace

Table of Contents

List of Figures

List of Tables

List of Algorithms

1

Introduction

1.1 Motivation

The signs of climate change are more apparent than ever. We see more powerful and frequent storms, droughts, fires, and floods worldwide. These changes are reshaping the composition and distribution of global ecosystems which include the natural resources and farming systems we rely on and biodiversity. Biodiversity refers to the variety of life on Earth, including all species, genes, and ecosystems that comprise our natural world [9]. Biodiversity is critical for human well-being, as it provides essential ecosystem services such as food, water, medicine, climate regulation, and cultural value [87]. However, biodiversity faces threats from human activities including habitat loss, overexploitation, invasive species, pollution, and climate change [87]. Biodiversity change describes alterations in the composition, structure, and function of biological communities over time and space [42]. Biodiversity change can positively or negatively impact ecosystems and human societies, contingent on context and scale [87]. As climate change continues to threaten global biodiversity, understanding its effects on species distribution becomes paramount. This threat extends to bird biodiversity, a crucial component of healthy ecosystems, thus underscoring the need to close significant knowledge gaps regarding species distribution and habitat suitability. However, the challenge lies in the inadequacies of traditional species distribution models (SDMs), which often focus on a narrow set of species or geographical areas due to computational costs, limited data availability, and an inability to account for complex inter-variable relationships. Furthermore, the task is complicated by the varying scales of species habitats, such as the broad-ranging song sparrows compared to the specific pine forest habitats of Kirtland's warblers. To effectively inform policy decisions, including those related to land use and conservation, a more comprehensive approach to SDMs is necessary—one that considers the extensive and interactive nature of global ecosystems. Machine Learning methods pose a solution to capture these patterns in global ecosystems. Machine learning has seen wide applicability across various solutions to tackle climate change, and more specifically in biodiversity.

In this study, we utilize remote sensing data and citizen science observations to analyze the joint distribution of bird species across various geographical locations. Our approach is grounded in the established principle that a species' presence is influenced by the characteristics of its local ecosystem, leading to dependencies among different species' abundances. We present end-to-end machine learning pipelines to build predictive models of biodiversity. Specifically, we create a comprehensive dataset and develop methods to estimate encounter rates for 684 and 1054 bird species at sites across the continental USA and Kenya respectively. By integrating publicly available remote sensing imagery and citizen science data, we provide a foundation for multimodal species distribution modelling. This is a crucial tool for understanding changing species ranges globally, informing policies on land use and conservation choices.

1.2 Related work

1.2.1 Tackling climate change with Machine Learning

Machine learning's utility in addressing climate change is well-established, extending across multiple domains to aid both in mitigation and adaptation. Mitigation strategies involve human-led efforts aimed at minimizing or preventing greenhouse gas emissions, a primary driver of climate change. Machine learning aids these initiatives, for example, by optimizing electrical systems through precise forecasting of supply and demand [81], enhancing scheduling for flexible electricity demand [52], or streamlining transportation via reduced activity, improved vehicle efficiency, and alternative fuel sources exploration [104].

Machine learning is being applied to increase resilience and enable adaptive measures in preparation for expected climate-related changes. For example, in climate science, different models often use varying parameterizations for the same processes, introducing uncertainty in predictions, especially for rainfall [22]. To address this issue, uncertain climate predictions can be improved by refining them with machine learning. Bretherton et al. (2022) demonstrated this by using Machine Learning to analyze climatic data, reducing errors in precipitation forecasts compared to the raw model outputs [22]. They demonstrated that machine learning can correct errors from traditional modeling methods that rely on parameterizations to approximate small or complex features like cumulus cloud convection. The approach showed promising results in improving precipitation and land surface temperature estimates across various climates by learning complex parameterization processes directly from data. This highlights the potential of machine learning to reduce uncertainty in climate models by bypassing manual parameterization. Studies have also shown machine learning can improve forecasts of extreme weather events [88]. In ecology, machine learning has shown potential to enable effective ecosystem monitoring, for example through using deep learning to monitor animal populations from satellite data and conduct biodiversity surveillance [99]. This helps guide societal adaptations by providing actionable insights into ecological changes. Further applications include using machine learning to assist the development and maintenance of resilient infrastructure [61]. By and large, we can see that machine learning is facilitating adaptive and mitigation strategies for resilience across domains from climate modeling to infrastructure. By leveraging large datasets and computational power, machine learning provides key tools to strengthen preparation and response capabilities for climate change impacts.

In Africa particularly, machine learning has in many ways been adopted as a tool to help tackle climate change. From efficient mapping of croplands in Togo where little to no ground data is available [64] [73] to floods inundation mapping in the Baro River Basin of Ethiopia [90]. While promising applications have emerged, there remains substantial scope for further development of Machine Learning solutions tailored to the African context. In sub-saharan Africa, for example, previous studies on the effects of climate change on terrestrial biodiversity have predominantly focused on Southern Africa [57,66,80,102] or Madagascar [23, 24]. In contrast, there's a noticeable gap in research regarding the impacts of climate change on terrestrial biodiversity in East African nations, particularly Kenya and Uganda, despite their rich biodiversity. This could be attributed to the large data requirements required to perform effective machine learning. Additionally, there is a significant gap in incorporating citizen science observations, notably from platforms like eBird [63], when studying the ramifications of climate change. This type of data can play a crucial role in formulating detailed species distribution models. Presently, Kenya's diverse biodiversity is facing a myriad of threats, including ecosystem degradation, water scarcity, and habitat fragmentation [13]. These existing challenges, when coupled with climate change, can intensify the adversities the regional ecosystems face, underscoring the urgency for relevant research. Moreover, with the Kenyan government's proactive measures to protect its wildlife ecosystems [13] and its Vision 2030 [48] striving to secure wildlife corridors, understanding the conjoined impacts of land-use changes and climate change on the country's biodiversity is essential to augment these conservation initiatives.

1.2.2 Species Distribution Modeling from citizen science data

There have been an increasing number of citizen science initiatives to collect species observation data for a variety of taxa in the past decade, from butterflies, to plants, to sharks [49, 82, 94]. Indeed, by crowdsourcing data collection efforts, it is possible to gather data not only over a larger temporal and geographic extent but also at a fine resolution [31]. As a result, the number of papers using citizen science for SDMs has increased at approximately double the rate of the overall number of SDM papers [40]. eBird data has been increasingly used for scientific research on birds, from assessing climate change-driven vulnerability of species, to modelling population change to predicting virus transmission [28,60,96]. The eBird status and trends project [41] from the eBird team combines satellite images with raw eBird data and uses statistical models and machine learning to build visualizations and tools to better understand migration, abundance patterns, range boundaries, and other patterns around the world. Other large scale citizen science databases used for SDM include iNaturalist [58] in which users record observations of species by taking geolocated pictures. In total, it has 113 million species observations. However, because it is limited to observations with pictures, it is more prone to sampling biases, both geographically and towards certain species (e.g. larger, more brightly coloured species [20,25]). It is also *presence-only* – that is, while there is data on the presence of certain species at a given location, it is assumed that other species may also be present.

1.2.3 Remote sensing for biodiversity monitoring

Remote sensing [1] data has been used for a variety of biodiversity monitoring applications, including predicting land cover classes for downstream ecological modeling [8], measuring the size of groups of animals [100], identifying tree species from crowns [97], and localizing bird nesting sites [55]. Bioacoustics has been used for bird monitoring [7] but

[1] Remote sensing used here and across this thesis refers to data collected about the earth's surface from aerial or satellite platforms, allowing observations of large regions difficult to survey on the ground.

one of the main limitations relates to the detectability of species, for example in densely populated areas with many anthropogenic sound sources [44]. Therefore, we will focus on discussing related work with remotely sensed imagery. In the GeoLifeCLEF 2020 challenge [29], they proposed a task combining remote sensing and citizen science data, where the goal is to predict the localization of plant and animal species using 1.6M geo-localized observations from France and the USA of 17,000 plant and animal species from aerial images and environmental features. The labels are derived from citizen science databases of presence-only species observations and each location is associated with only one species, limiting the ability to accurately model multiple species at a time. A newer version (2023 edition) of this classification challenge proposes to focus only on plant species and to use Sentinel-2 satellite images rather than aerial images. While each location is still associated with one species, a validation set with presence-absence labels is provided, highlighting the importance of presence-absence data, at least for the evaluation of such species modeling tasks. Methods deriving information from satellite images have also been developed in the context of avian studies but they usually use a suite of carefully selected measures calculated from the imagery rather than leveraging its full potential [11, 39].

A particular field of application that saw the development of large remote sensing images datasets is agriculture, with an emphasis on the temporal resolution for crop monitoring [54, 70, 98]. Recently, a number of methods have been introduced to learn robust representations of remote sensing data, with the goal of using them for a variety of downstream tasks [70, 83, 85]. In particular, MOSAIKS [86] proposes to use a single encoding of satellite imagery for diverse prediction tasks, and Satlas [16] and SatMAE [30] provide pre-trained models on a large number of remote sensing images, arguing that they can be useful feature extractors for remote sensing tasks.

1.2.4 Advances Over Prior Work

Our proposed approach makes several key advances over existing species distribution modeling methods:

- We fully leverage presence-absence data from a large citizen science database and compute encounter rates such that biases due to the different number of observers visiting locations are mitigated.
- We enable modelling of a wide range of species at a time across large geographical areas. To our knowledge, this is the first attempt at predicting encounter rates for many species jointly, building on recent advances in machine learning for remote sensing.
- Our proposed pipeline makes it easily extendable to other regions in the world as all data sources are open; eBird [63] data is available in all countries in the world and satellite Sentinel-2 data is publicly available.

1.3 Purpose and Research Questions

In this thesis, a machine learning technique will be used in a knowledge discovery process to infer the joint distribution of many species for a given location, using publicly available citizen science observation records as ground truth. This work uses techniques from remote sensing, computer vision, and citizen science. This approach leverages the hypothesis that a species' presence or absence at a location depends on the ecosystem present there. Therefore, an abundance of different species is highly correlated.

Through this work, we will address the following main research questions;

1. How can remote sensing and computer vision methods be used in predicting encounter rates of different bird species at specific geographical locations using publicly available bird observation and satellite data sources?
2. How well do these models transfer to low data regimes especially when we do not have reliable data sources?

1.4 Scope and Limitation

- **Geographical Focus:** The primary focus of this work is on species distribution patterns within the continental USA and Kenya. We acknowledge that there are other regions with unique characteristics but these two countries offer a solid foundation for our analysis. We chose these two regions due to their distinct geography, diverse species populations and amount of remote sensing and eBird data available.

- **Encounter Rate:** In this work, we adopt the encounter rate as the main measure of species presence. This approach aligns with the established ebird best practices and allows us to assess the distribution and relative abundance of different species effectively. While we recognize that there may be additional techniques and measures to explore species distribution [17] we are only focusing on Encounter Rate.

1.5 Target Group

This work and the research findings presented are relevant across multiple fields, as they demonstrate novel techniques and applications for biodiversity monitoring and conservation. Machine learning researchers, especially those specializing in computer vision and remote sensing, will find interest in the methods for integrating machine learning into species distribution modelling using techniques from computer vision. These methods can enhance the accuracy and efficiency of predicting and mapping species habitats from satellite imagery. Similarly, biodiversity and ecology researchers can utilize the computational methods herein to further study species conservation, especially in areas where field surveys are difficult. Climate scientists and activists can leverage the biodiversity data used in this research to understand the ecological impacts of climate change, such as how species distributions may shift or decline under different scenarios. Remote sensing experts can apply satellite image analysis for biodiversity monitoring, as well as other environmental applications that require high-resolution and large-scale data. Researchers in computational sustainability can build upon the methods presented to address environmental challenges, such as optimizing conservation planning and management.

Additionally, this work may appeal to data scientists and engineers for the large-scale modelling approaches; conservation organizations who can employ the predictive models and pipelines for wildlife tracking; citizen science communities as the research builds on citizen science data for model development; industry professionals to demonstrate business and technology applications; Policymakers, to guide biodiversity conservation efforts and climate adaptation; and curious individuals for an overview of how data science aids ecological research.

2

Background

2.1 Data Sources for Biodiversity Monitoring

Effective machine-learning models rely on high-quality data. Biodiversity data can be sourced from a variety of ecosystems and taxa. In this chapter, we describe some of the data sources that have been successfully used with machine learning methods to improve biodiversity monitoring. This data is useful for monitoring the state of biodiversity, identifying pressures and threats in a location, and developing conservation responses, especially in the face of climate change. By analyzing diverse biodiversity data, these techniques facilitate species monitoring, population modeling, and conservation insights.

In this chapter, we provide a detailed overview of key machine-learning applications across a variety of taxa and data modalities. To navigate the extensive array of biodiversity datasets and data sources, we will adopt the classifications provided by the Essential Biodiversity Variables (EBVs). Developed by the Group on Earth Observations Biodiversity Observation Network (GEO BON) [78], these EBVs help evaluate of the geographical distribution of various aspects of biodiversity and their temporal changes.

We begin by exploring different datasets generated from various taxa and highlight how machine learning contributes to biodiversity monitoring within these domains.

2.1.1 Raw Species Occurrence Data

Raw species occurrence data [1] is useful for understanding the presence or absence of a species in a particular habitat or location. It is useful for understanding the general populations of a species, their movement and patterns that provide critical insight into biodiversity. Useful resources include the Global Biodiversity Information Facility (GBIF) [45] aggregates species occurrence data from across the world. eBird [63] does the same specifically for bird sightings and observations from citizen scientists, describing their ranges, abundances and trends. Similarly, iNaturalist [58] crowdsources sightings and observations of different plant and animal species worldwide.

2.1.2 Genetic Data

Genetic data is crucial for understanding key species traits that lie at the heart of biodiversity. Over the years, genetic data has been collected from different taxonomies to understand these vital traits. This data is particularly important for understanding evolutionary processes and predicting future changes at the DNA level. Notable examples include GenBank [18], European Nucleotide Archive [67] and the Barcode of Life Data Systems [84]. These data sources are useful for understanding patterns of natural selection is useful for understanding ways in which species can adapt to changes in their habitats, especially due to climate change. Many studies already utilize this data for biodiversity monitoring for example to understand how human land use and climate change will have a significant impact on animal genetic diversity [91]. More recent work is also being done with multimodal taxonomic data, including genetic data, to accelerate insect biodiversity monitoring using machine learning. One example is the BIOSCAN-1M dataset, which

[1] Note: Every other one of the other data modalities may also include raw species occurrence data

contains hand-labeled insect images along with associated genetic information such as raw nucleotide barcode sequences and assigned barcode index numbers [46]. Datasets like BIOSCAN-1M that integrate visual and genomic insect data enable new machine-learning capabilities for automated species identification and biodiversity assessment. This multimodal approach combining computer vision and genomics holds promise for scaling up and improving the accuracy of biodiversity monitoring.

In addition, the emerging field of environmental DNA (eDNA) [92] sequencing allows estimation of biodiversity by sampling organisms' DNA from environmental samples like water or soil, rather than directly observing individuals. By detecting tissue or cells shed into the environment, eDNA enables broader taxonomic evaluation than possible through visual surveys or specimen collection alone.

2.1.3 Camera Traps

Camera traps have also been employed in more specific habitats such as national parks to collect imagery for population monitoring and species occurrence, in particular, have emerged as an inexpensive and easy way to collect ecological data through high-resolution, motion-triggered photography [89]. Deployed in the field, camera traps can capture images of wildlife to provide insights into behaviours and interactions. Sites like the Serengeti have used camera traps effectively to enable accurate models of dynamic species populations [89].

2.1.4 Species Traits Data

Species traits data offers another dimension for biodiversity monitoring. Species traits are measurable characteristics of organisms that can affect their ecological performance. Understanding the traits and characteristics of individual species is vital for modeling biodiversity. Key trait data resources include TraitBank [75] by the Encyclopedia of Life, a public database compiling species traits; the TRY Plant Trait Database [62], an initiative aggregating plant trait data from diverse sources and the coral trait database [69] – a research initiative that aims at making all observations and measurements of corals accessible for coral reef conservation research. Combining trait data from specialized databases like these allows researchers to better analyze interspecific variation and model species roles within ecological communities. With machine learning, trait data can be used to assess the risk of extinction, design and prioritize conservation actions, evaluate their effectiveness, and monitor how species respond to global change.

2.1.5 Time series Data

Time series data capturing changes over time provides valuable insights for biodiversity conservation research and monitoring. BioTIME is an open-access global database containing time series assemblage data to quantify and analyze biodiversity change [33]. It includes species abundance, biomass, and diversity information across ecosystems. Researchers can utilize BioTIME to identify complex biodiversity trends and gain insights into species growth and reproduction strategies. Other sources like satellite remote sensing provide consistent time series measurements of environmental and vegetation variables (e.g. temperature, precipitation, canopy height) to assess habitat change.

Time series data enables identification of meaningful biodiversity trends and changes. Field observations and remote sensing provide complementary perspectives for a comprehensive understanding of biodiversity dynamics. Open access resources like BioTIME accelerate biodiversity research and conservation efforts. Established machine learning methods for time series analysis have been extended to biodiversity applications.

2.1.6 Ecosystem Function Data

One of the key aspects of understanding ecosystem function is to measure and analyze various data variables that reflect the interactions and exchanges of matter and energy within and across ecosystems. Some sources of such data variables are; Fluxnet [14] – a network of micro-meteorological tower sites that measure ecosystem gas exchanges of carbon dioxide, water vapor, and energy. These data can help monitor how ecosystems respond to climate change and human disturbances, Global Carbon Project [4] – A project that provides information on global carbon budgets, which are essential for assessing the sources and sinks of greenhouse gases and their impacts on biodiversity and the World Ocean Database [5], a portal which provides access to oceanographic datasets and information, such as temperature, salinity, oxygen, nutrients, and plankton. These data can help monitor the health and productivity of marine ecosystems and their biodiversity.

2.1.7 Remote Sensing Data

Advancements in satellite and remote sensor technologies have opened innovative avenues for ecological data collection through consistent, widespread monitoring at multiple spatial and temporal scales. Space agencies including NASA [74] and ESA [37] provide near-daily satellite imagery, offering insights into species populations, climate and weather patterns, crop yields, and ecosystem changes over time. Remote sensing proves invaluable for observing phenomena like glacier melting [15] as well as detecting habitat degradation and encroachment of invasive species by comparing images longitudinally [47]. The integration of data across multiple satellite channels, including RGB, near-infrared and elevation facilitates bioclimatic modelling. When incorporated into ecological models, these multimodal remote sensing datasets shed light on the complex mechanisms driving shifts in biodiversity patterns, in turn informing conservation strategies. Remote sensing via satellite imagery has become a very useful tool for ecological research and management by enabling consistent, scalable data collection that reveals biodiversity changes that would otherwise be difficult to observe.

2.1.8 BioAcoustics Data

Bioacoustic sensors offer a promising alternative to visual data collection for biodiversity monitoring. Unlike cameras, which can be obstructed by vegetation or weather, microphones can capture sounds over large spatial and temporal scales. For instance, [95] used audio recordings from the Mt. Kenya ecosystem to identify bird species using machine learning techniques. They collected and annotated hundreds of recordings from different habitats and seasons, and trained various models to automatically recognize the vocalizations of the birds. Their work demonstrates the potential of bioacoustic sensors for assessing the diversity and distribution of wildlife in tropical regions.

With low technical barriers, microphones have been deployed in diverse environments from deep seas to tropical forests. As a complement to other methods like remote sensing, ecoacoustics generates datasets on vocalizing species like birds, whales, dolphins, and elephants. These datasets enable machine learning techniques to extract valuable insights into population structures, migration patterns, and effective species monitoring. By leveraging the sounds that animals naturally produce, passive acoustic monitoring offers an innovative data stream for biodiversity studies.

2.2 Species Distribution Modelling

A primary challenge in studying biodiversity change involves measuring and predicting species distribution across landscapes and regions.

Species Distribution Models (SDMs) [2] are numerical tools that combine observations of species occurrence or abundance with environmental data to estimate species' geographical range. These models can provide an understanding of or predict species distribution across a landscape [17]. Intuitively, [17] defines an SDM as a function that uses the characteristics of a location to predict whether or not a species is present at that location. It therefore can be understood as a supervised learning problem. The input is a vector of environmental characteristics for a location and the output is species' presence or absence.

A simple species distribution modeling pipeline consists of three key components [17]:

1. **Species observation data** - Records of where species have been observed and collected. This serves as the target data for modelling.

2. **Location encoding**- A method to encode geographic locations into numerical values that can be used as inputs to a model. This allows generalizing to new locations.

[2] Names for such models vary. These models are also called bioclimatic models, Ecological Niche Models (ENMs)

3. **Prediction function** - A function that maps from the encoded location inputs to predicted species occurrence values. This is the model that is trained on observation data.

2.2.1 Early SDMs

Predecessors of SDMs include prior research that highlighted the correlations between species patterns and environmental or geographic factors. A notable example is Joseph Grinnel's 1904 analysis on the chestnut-backed chickadee's distribution [50], Analysis of population ecology of some warblers of northeastern coniferous forests [68] among others.

Early SDM approaches relied on simple statistical methods like multiple linear regression and linear discriminant function analysis to relate species occurrences to environmental variables [26]. These methods provided coherent error distribution treatments of presence-absence and abundance data. While pioneering, however, these techniques had limited flexibility to model complex ecological relationships. They assume a linear relationship between the predictor variables and the response variables which may not hold true for complex ecological data. They also require a large number of observations compared to the number of predictor variables which may not be feasible for rate or endemic species. Additionally, they do not account for spatial autocorrelations or non-stationarity in the data, which may affect the reliability of the models.

2.2.2 Generalized Linear Models (GLMs)

Generalized linear models (GLMs) represented a major advance, enabling the accommodation of non-normal errors, additive components, and non-linear species-environment correlations [71]. The advent of GLMs expanded the modelling capabilities of SDMs. They enabled regression-based SDMs that had more sophistication than possible earlier. They continue to be useful and are part of many current methods including MaxEnt [79]. GLMs overcame some of the limitations of Linear Discriminant Analysis which failed to capture the complexity and variability of ecological data. However, GLMs also had some drawbacks, such as requiring the assumptions that errors are identically, independently, and normally distributed [21], and being sensitive to the choice of predictor variables and their interactions [10]. Moreover, GLMs based on climate alone may not be reliable, as topographic variables may also influence species distributions [34].

2.2.3 Maximum Entropy Modeling (MaxEnt)

More recently, the growth of large presence-only species datasets has driven innovation in species distribution modelling (SDM) methods that utilize presence-only data. A prominent technique is Maximum Entropy Modeling (MaxEnt). MaxEnt leverages point process theory to model species ranges using only presence data [79]. Its key advantages are requiring just presence points, accounting for interactions between predictors, and generating smooth response curves. The MaxEnt algorithm works by finding the probability distribution with maximum entropy (most spread out) while constrained by the environmental conditions at known occurrence locations [2]. It iteratively tests different models, selecting the one with the highest entropy [3]. While effective, MaxEnt has some limitations. It produces only point estimates of occurrence probability, lacking measures of uncertainty [43]. Results can also be sensitive to the choice of regularization parameters and feature types, potentially affecting model performance and interpretation [72].
While MaxEnt provides a flexible and powerful framework for SDM using presence-only data, users should be aware of its limitations when applying and interpreting the MaxEnt model.

2.3 Summary

In this chapter, we have provided a background for various data sources crucial for biodiversity monitoring, ranging from raw species occurrence data to genetic data, camera traps, species traits data, time series data, ecosystem function data, remote sensing data, and bioacoustics data. These diverse sources collectively provide valuable insights into the distribution, behaviour, and health of species and ecosystems. Furthermore, we have introduced the concept of Species Distribution Modeling (SDM) as a fundamental approach to predicting and understanding species distributions across landscapes. The next sections will describe our data, methods and models utilized.

3

Data and Methods

In this chapter, we outline the data sources and methodologies employed in our experiments. We first describe the eBird dataset in section 3.1 – which is the main source of bird occurrence data, and how we selected regions of interest for our study. Next, we describe the environmental datasets we used, which include satellite image-based remote sensing data and bioclimatic variables, and how we extracted and processed them to match the eBird data. Section 3.2 explains the steps involved in this procedure. Then, in Section 3.4 we discuss the rationale and implementation of the techniques we used to split the final dataset into training, validation, and test sets, and how we ensured that the splits were representative and unbiased. Finally, we describe the methods we used to model species distributions from the environmental data, including machine learning models and evaluation metrics. Section 5.2 presents the details of these methods.

3.1 eBird Dataset

eBird [63] is a crowdsourced database of bird observations collected and maintained by the Cornell Lab of Ornithology. It contains millions of birds submitted by citizen scientists around the world. eBird users submit complete checklists of all the birds they have identified during an outing at a specific location and time. A *complete checklist* records both the species detected and not detected hence providing data on presences as well as absences. *Hotspots* on eBird refer to locations with high birding activity and usually contain a high number of complete checklists. *Encounter Rates* on eBird refers to a measuring of probability of an eBirder encountering a species on a standard eBird checklist, and is a proxy for species abundance in that hotspot.

For our study, we analyzed complete checklists from eBird hotspots in the continental US and Kenya.

3.1.1 USA eBird dataset

For the continental US, we extracted checklists recorded between $2010 - 2023$, targeting summer (June-July) and winter (December-January) seasons. Hotspots were filtered to have at least 5 complete checklists, the minimum threshold set for calculating statistically meaningful encounter rates. Additionally, marine locations were excluded using 5-meter US geographic boundaries from the Census Bureau's MAF/TIGER database [1].

In the USA species selection, we considered ABA (American Birding Association) Codes[1] 1 and 2, representing regular breeders/visitors and less widespread but relatively common species respectively [6]. We excluded species found exclusively in Hawaii and Alaska, as well as a Code-2 seabird species with rare oceanic observations. The final species dataset for the USA resulted in a total of 684 species.

[1] ABA Code 1 refers to widespread, common species. Code 2 refers to less widespread but relatively common species.

3.1.2 eBird Kenya dataset

For Kenya, we gathered all complete checklists recorded between 2010 and 2023, focusing on the 1,054 species regularly found in the region according to Avibase [12]. Given the scarcity of data in Kenya, we did not impose any criteria regarding the minimum number of checklists per hotspot or specific observation months. Additionally, the Kenya dataset comprises a larger number of non-migratory bird species, prompting our decision to aggregate records across seasons.

3.1.3 eBird data preparation

As part of data preparation, we merged hotspots with identical latitudes and longitude but different IDs in eBird, combining their checklists, except in cases where all checklists originated from the same observer on the same day, in which case we retained only one hotspot. Next, we computed the encounter rates for each hotspot as will be explained in detail in section 4.1. To address vagrants (species observed in hotspots outside their typical geographic range), we referred to eBird's range maps. When available, we set the target encounter rates to zero for such species in the respective hotspots. To maintain consistency, we aggregated species observations over a span of 13 years, considering seasonal changes in distributions for the USA dataset and leaving annual temporal changes for future iterations of this research.

3.2 Environmental Data

For our analysis, we integrated essential environmental data, drawing inspiration from the GeoLifeCLEF 2020 dataset [29]. Specifically, we extracted a total of 19 bioclimatic variables in raster format, each with a size of 50×50 pixels and a spatial resolution of approximately 1 km, centered on each hotspot. These variables were sourced from WorldClim 1.4 for both the USA and Kenya datasets. Bioclimatic variables are commonly used to model species distributions as they provide valuable insights into climate-related factors [59]. They encompass annual trends related to temperature, precipitation, solar radiation, wind speed, and water vapour pressure, all of which play crucial roles in influencing species habitats.

Additionally, for the USA dataset, we extracted an additional 8 pedologic (soil) variables, each with a resolution of 250 meters, from SoilGrids [56]. SoilGrids offers global soil properties maps, including pH levels, soil organic carbon content, and stocks. These detailed soil property maps are derived using machine learning techniques trained on extensive soil profile observations and environmental covariates derived from remote sensing data.

The selection of environmental variables serves as a good foundation for understanding the ecological context in which species are distributed and how they respond to various climatic and soil conditions. More comprehensive details regarding the environmental variables can be found in the Appendix.

3.3 Remote Sensing data

To acquire remote sensing data for each hotspot, we extracted RGB and NIR reflectance data at a resolution of 10 meters from Sentinel-2 satellite tiles. The extracted data covered a square region of approximately 5 km^2, centered around each hotspot. Additionally, we obtained true color image RGB bands.

To ensure the quality of the data, we selected images with cloud coverage of at most 10%. For the USA-summer dataset, we considered the time window between June 1 and July 31, 2022, while for the USA-winter dataset, we chose the time window between December 1, 2022, and January 31, 2023. For the Kenya dataset, we utilized images from the time window between January 1, 2022, and January 1, 2023.

To minimize any temporal bias in our data, we associated a single image per hotspot. This approach allowed us to represent species data with the most recent satellite image available since recent years generally have more checklists compared to earlier years. For images that covered the entire 5 km^2 region, we directly used them. For images that covered a smaller area, we considered either composing a mosaic with the extracted images or discarding them to minimize seams in our dataset.

3.4 Dataset Splits

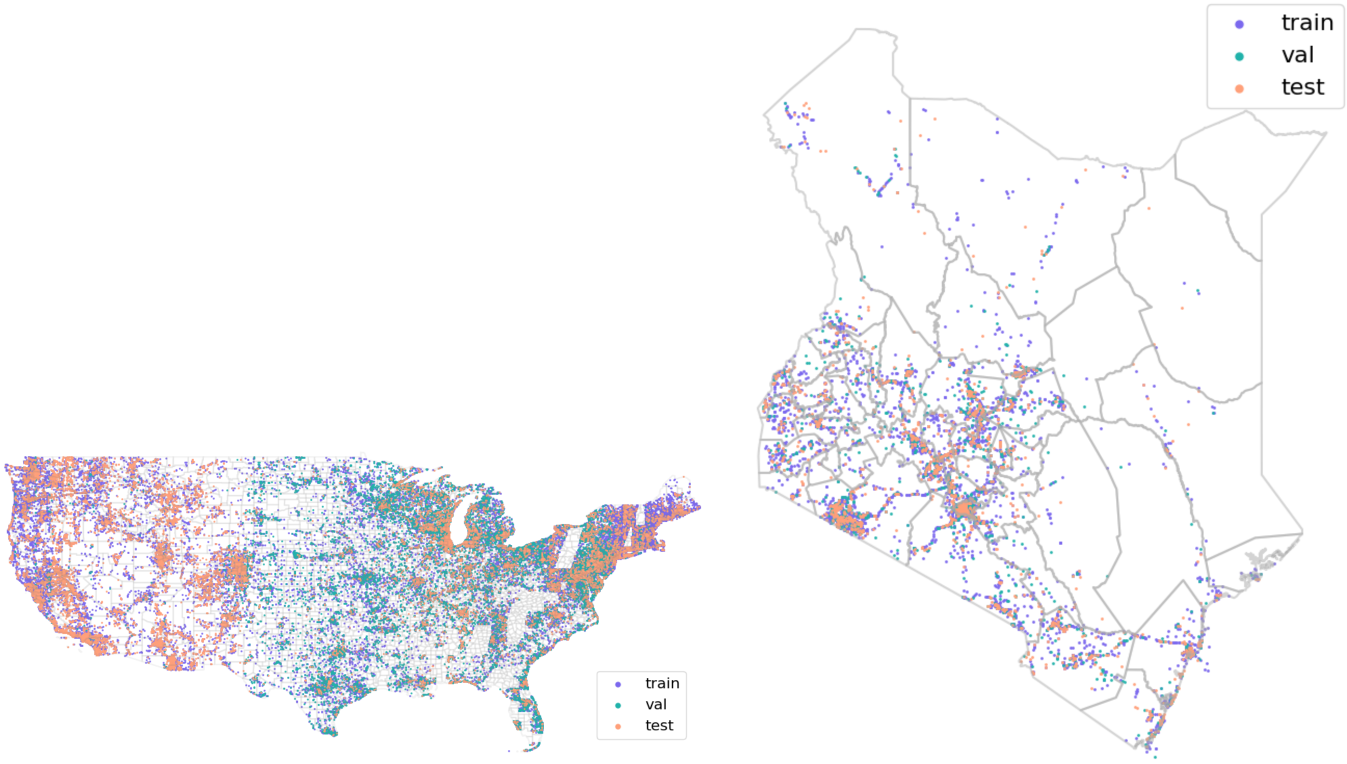
To account for spatial autocorrelation that may arise from random splits of geospatial data, we utilized the sklearn's Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm [77]. DBSCAN performs clustering from from vector array or distance matrix. It finds core samples of high density and expands clusters from them. It is good for data that contains clusters of similar density.

Table 3.1 describes the number of hotspots present in each split for each dataset of SatBird. The splits were obtained by following the process described in Section 3.4.

Split	**USA summer**	**USA winter**	**Kenya**
Train	76,590	12,102	6,537
Validation	22,395	3,197	1,932
Test	17,469	2,595	1,633

Table 3.1: Number of hotspots in each split for the three datasets of SatBird.

We specified that core samples must have at least 2 hotspots with a maximum allowed distance of 5 km between them. After applying the DBSCAN algorithm, we obtained 217 clusters with a total of 12,650 hotspots. These clusters were then randomly assigned to the train, validation, and test splits, with proportions of 65%, 20%, and 15% respectively for the USA-summer dataset. For the USA-winter dataset, we maintained the same split assignment as for the summer, resulting in a repartition of 70%, 15%, and 15% for the train, validation, and test sets. A similar procedure was followed for the Kenya dataset, leading to a split distribution of 65%, 20%, and 15%.



(a) SatBird-USA-summer **(b)** SatBird-Kenya

Figure 3.1: Distribution of hotspots across the training, validation, and test sets.

Table 3.2: Summary of the data provided for each of the SatBird subsets: USA-summer, USA-winter, and Kenya.

	USA-summer	USA-winter	Kenya
Number of hotspots	122593	55497	9975
Number of species	670	670	1,054
Satellite RGBNIR reflectance	✓	✓	✓
Satellite true color image	✓	✓	✓
Bioclimatic rasters	✓	✓	✓
Pedologic rasters	✓	✓	
Range maps	✓ (586 species)	✓ (620 species)	

The map in Figure 3.1 visually represents the data splits for reference.

3.5 Methods

3.5.1 Machine Learning Methods for Biodiversity Monitoring

With the diversity of data types and modalities required for effective biodiversity monitoring and species distribution modelling, machine learning provides new opportunities to fully utilize these datasets. In this work, we employ several machine learning techniques to model species distributions across different stages of our pipeline.

Unsupervised clustering methods (detailed in Section 3.4) enabled informed splitting of our datasets for training and evaluation. By clustering based on features of the data, we obtained splits that better represent the full distribution of the data. For baseline modelling, we utilized random forests, a versatile algorithm well-suited for ecological data. Random forests model complex interactions through ensemble decision trees. They naturally handle mixed data types and capture nonlinear relationships. Boosting methods like XGBoost allowed us to model particularly complex and nonlinear species-environment interactions. By combining many weak learners, boosting algorithms build up to highly accurate predictions. To further improve on the baseline, we implemented neural networks and other deep learning architectures. These state-of-the-art techniques can learn subtle features and patterns beyond simpler models, providing the top performance in our experiments. With this multi-pronged approach utilizing diverse machine learning methods, we aimed to effectively extract as much information as possible from the available data for enhanced biodiversity monitoring. The following sections provide further details on the implementation and results for each technique

Random forests

A Random Forest is a classification and regression tree model. It is a combination of tree predictors where every tree can depend on the values of a random vector sampled independently with the same distribution for all trees in the forest. Random Forests consist of an ensemble of decision trees trained on different subsets of the data. They can model complex interactions and have built-in ways to avoid overfitting. Random forests have been shown to have high predictive performance for SDMs across diverse taxa and regions. [27,93,101]. In our task, we particularly used them as a baseline model to relate the bioclimatic variables and the encounter rates of all birds in our checklist.

Boosted Regression Trees

Boosted Regression Trees (BRT) combine regression trees and a boosting technique that iteratively fits tree models using binary splits of predictor variables [36]. This approach incorporates key advantages of tree-based methods, including the ability to handle different types of predictor variables, accommodate missing data, fit complex nonlinear relationships, and automatically capture interaction effects between predictors. BRT models require no prior data transformation or elimination of outliers. By fitting an ensemble of simple tree models, boosting allows BRT to model complex response surfaces efficiently. BRTs have demonstrated high accuracy for biodiversity research across a variety of taxa [51, 103]. Following the GeoLifeCLEF challenge [29], we propose an environmental baseline, using Gradient Boosted Regression Trees on the bioclimatic and pedological variables extracted at each of the hotspots.

Neural Networks and Deep Learning

Neural Networks are machine learning algorithms that mimic the structure and function of biological neural networks. They consist of layers of artificial neurons that perform computations on the input data and pass the output to the next layer. Neural Networks can model complex nonlinear relationships and have the universal approximation property, which means they can approximate any continuous function with arbitrary accuracy. Deep learning is a branch of machine learning that uses multi-layer neural networks to learn high-level representations from large amounts of raw data. Convolutional neural networks (CNNs) are a specific type of neural network that are especially suited for image-based data. They use convolutional filters to extract local features from the input images and pool them to reduce the dimensionality. In our task, we leverage the power of deep learning to learn such complex features from the eBird presence-only data combined with the environmental data and remote sensing data. Previous studies have shown that deep learning SDMs can outperform other methods for presence-only data [32,35,38].

3.5.2 Bayesian Models

In scenarios where we have limited checklists, we utilize Bayesian approaches to improve the predictions for locations where limited data exists. Bayesian models allow us to incorporate prior knowledge and update it with new evidence from the data. In our case, we use the mean of the predictions of our base models as a prior distribution, which represents our initial belief about the species distribution. Then, we perform a sampling of additional checklists in the region and use the ground truth to provide a likelihood function, which measures how well the data fits the prior. By applying Bayes' theorem, we obtain a posterior distribution, which represents our updated belief about the species distribution after observing the data. After successive sampling of checklists, the posterior distribution converges to the true distribution, hence providing a better estimate than just the predictions. We provide more details on this approach in chapter 5

4

Experiment Design

In this chapter, we describe the experimental design of our study. We first define the task and the objective of our work in Section 4.1. Then, we present the baseline models that we compare our proposed methods with in Section 4.2. Next, we introduce the image-based models that we use to leverage the remote sensing data such as ResNet, SATMAE, and MOSAIKS, in Section 4.3. In Section 4.4, we explain how we incorporate geographical splits to account for spatial autocorrelation and sampling bias. In Section 4.5, we discuss the evaluation metrics that we use to measure the performance of our models. In Section 4.6, we describe the experiments that we conduct to answer our research questions. Finally, we report and analyze the results of our experiments in Section 4.7.

4.1 Task Definition

The main objective of our research is to predict bird encounter rates using remote sensing data, aiming to complete species distribution mapping in unexplored regions. To accomplish this, we draw insights from the rich bird sighting records available in the eBird citizen science database [63], which comprises an impressive collection of approximately 80 million records covering nearly $10,000$ bird species worldwide. Our focus lies on leveraging the valuable observation reports known as ***complete checklists***.

In the eBird database, a ***hotspot*** refers to a specific location where birdwatchers have submitted checklists. Each hotspot is associated with a set of complete checklists, documenting all species observed by birdwatchers on specific dates and times at that location. These checklists serve as *presence-absence* data, providing information not only about the presence of reported species but also the absence of non-reported species. Consequently, they are highly informative and serve as a robust substitute for expert field surveys. This modeling of the complete checklists associated with eBird hotspots make it possible for us to predict species encounter rates for unexplored regions lacking in expert bird surveys.

Given a hotspot h and a list of species $s_1, \ldots, s_n$ of interest, our ultimate goal is to build a machine learning model that takes a satellite image of the hotspot (and optionally other relevant data) as input and outputs a vector:

$$\mathbf{y^h} = (y^h_{s_1}, ..., y^h_{s_n}) \tag{4.1}$$

where:

- $\mathbf{y^h}$ represents a vector containing encounter rates for each species $s_1, \ldots, s_n$ at hotspot h.
- y^h_s denotes the encounter rate of species s at hotspot h.

- The encounter rate y_s^h is computed as the number of complete checklists reporting species s at hotspot h divided by the total number of complete checklists recorded at that location.

This ratio y_s^h represents an *encounter rate*, corresponding to the probability of a visitor observing a particular species when visiting the same location (hotspot). The key focus of our task is to jointly predict encounter rates for multiple relevant bird species, transforming our problem into a supervised multi-output regression challenge.

We specifically chose to predict encounter rates due to their ecological significance and widespread use in the eBird platform as "hotspot bar charts". These bar charts serve as summaries of species present in a specific location, aiding birdwatchers and ornithologists in understanding the expected bird species. However, they are limited to past data and cannot be extended to unexplored regions. Furthermore, encounter rates from eBird have not been previously modeled jointly to account for species interactions within the same habitat, an essential aspect for accurate species presence estimates. We aim to bridge this gap by simultaneously modeling encounter rates for multiple species and extending predictions to locations with limited or no recorded observations

By predicting encounter rates and considering species interactions, we intend to provide more informative and comprehensive species distribution mapping. This advancement can significantly contribute to biodiversity research, conservation planning, and support efforts to understand and protect avian populations in uncharted regions. The interaction between our data and models is illustrated in Figure 4.1. The satellite data in figure 4.1 represent a single hotspot, we use patches corresponding to $640m^2$ of a location, combined with pedologic and bioclimatic rasters. This area of interest is generated by using the center of the lat/lon of a hotspot, and then extending it to $640m^2$. In practice, this results in a large raster. During training, we resize it to 64px to allow it fit into memory; see Section 4.3

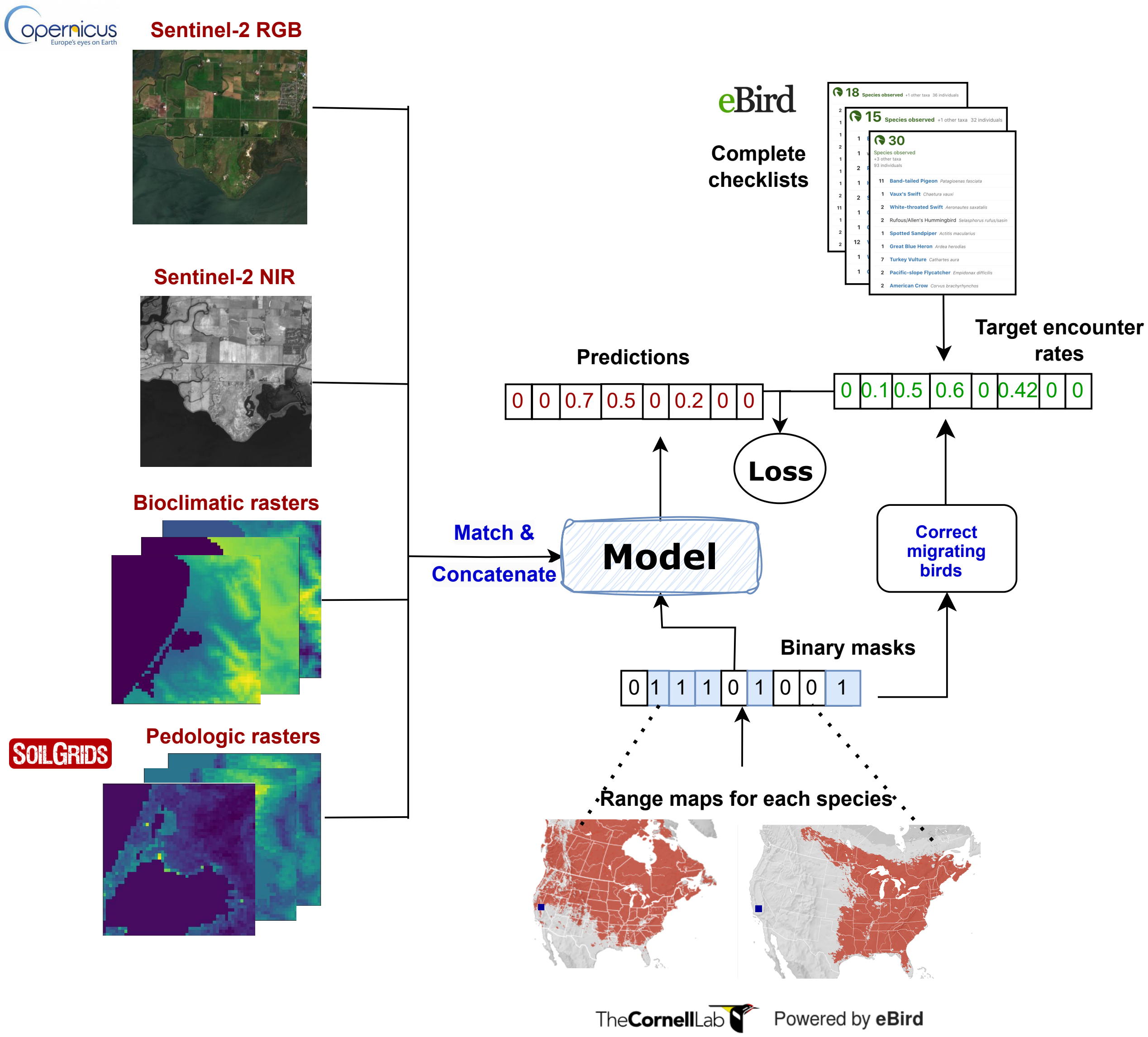


Figure 4.1: An overview of data streams for SatBird, as well as inputs and outputs for the task for predicting species encounter rates. The Sentinel-2 10m-resolution satellite data, bioclimatic and peologic rasters cover a single hotspot ($640m^2$) patches, and can be used along low resolution environmental data as input to a model after matching their resolutions. Labels are derived from eBird complete checklists. Observations of vagrants (migrating birds) in the labels are corrected with range maps from eBird, which can also be incorporated in the model to make it geography-aware.

4.2 Baseline Models

- **Mean Encounter Rates**: The initial fundamental employed in this study is the mean encounter rate. It is established by calculating the average encounter rates of each species over the training set. This baseline gives us insight into the overall distribution and relative abundance of species,forming a reference for further analysis.

- **Environmental Baseline**: Building upon the GeolifeCLEF challenge methodology [29], the second baseline leverages gradient-boosted regression trees to predict species distribution based on bioclimatic and pedological variables. This approach allows us to explore the influence of environmental factors on species presence.

4.3 Image Based Models

- **Resnet18**: In addition to the baseline models, we incorporate Convolutional Neural Networks (CNNs) as feature extractors from the satellite images. Specifically, we use the ResNet-18 model [53] and input both the RGB and NIR band's reflectance data. To initialize the network, we use pre-trained weights from the ImageNet pretrained model, finetuning the model for our specific task. For the first layer corresponding to the NIR band, we adopt an initialization process, sampling from a normal distribution based on the mean and standard deviation of the layer's weights for the other bands. Additionally, we conduct experiments using RGB true-color images as input and compare the results against the reflectance data approach. To maintain consistency across the dataset, we define a region of interest of $640m^2$ and apply center cropping to satellite patches, resulting in a size of 64 x 64 around the hotspot. Augmentation techniques such as random vertical and horizontal flipping are employed to enhance model robustness. The Resnet18 model is trained with a cross-entropy loss function to optimize performance.

- **Multi-Task Observation Using Satellites and Kitchen Sinks (MOSAIKS)**: The MOSAIKS model, proposed by [86], is an accessible and versatile method with wide-ranging applicability across various tasks. Its adaptability allows us to transform satellite imagery from diverse geographical locations on Earth into meaningful summary information.It works by collecting satellite images on servers, extracting color, texture, and spatial ”features”, and putting those features into a regression model to predict the task of interest. For our specific research, we utilize MOSAIKS to predict the encounter rate for each hotspot, leveraging its capability to handle a broad spectrum of prediction outcomes.

- **SatMAE**: SatMAE [30] serves as a pre-training framework designed for temporal or multispectral satellite imagery, relying on the Masked Autoencoder (MAE) approach. This framework incorporates spectral embedding and employs independent masking of image patches across different time points [30]. The utilization of SatMAE allows us to harness the potential of temporal and multispectral satellite data for our analysis, contributing to a more comprehensive understanding of the underlying patterns and dynamics within the hotspot regions.

4.4 Including geographical information

To enhance the accuracy of our vision-based models regarding the geographic ranges of specific species, we integrate range maps sourced from Ebird. By incorporating these maps into our models, we enable them to gain insights into species that inhabit similar habitats but may be distributed across different geographical regions, potentially not coexisting together.

The range maps obtained from Ebird are created through the construction of binary masks. These masks effectively zero out the predicted encounter rates for a particular species when the model’s location falls outside the species’ known distribution range. By incorporating this information, our models become better equipped to make informed predictions based on species-specific geographic preferences, facilitating a more comprehensive understanding of their distribution patterns and potential presence in various regions.

4.5 Evaluation metrics

To analyze the quality of our species distribution predictions, we employ several quantitative evaluation metrics to compare our model outputs against the ground truth data. These metrics provide numerical scores that give insight into different aspects of predictive performance. These include;

- **MSE (Mean Squared Error)**: This metric measures the average squared difference between the predicted values and the actual ground truth values. It provides a quantitative measure of how well our model's predictions align with the true values.

- **MAE (Mean Absolute Error)**: The MAE calculates the average absolute difference between the predicted values and the true values. It gives us an indication of the overall accuracy of our model's predictions.

- **Custom KL (Kullback-Leibler Divergence)**: This metric quantifies the difference between two probability distributions – the predicted distribution and a reference distribution. The KL divergence is particularly useful in comparing the similarities or differences in probability distributions.

- **Presence (K) Threshold Accuracy**: This metric measures how accurately the model predicts whether species are present or absent at a site. To do so, we convert the model's predicted probabilities to binary presence/absence values using different probability thresholds (k values). For each species at each site, if the predicted probability exceeds the threshold k, we consider that a prediction of ”presence,” while lower probabilities predict ”absence.” We can then compute the accuracy for these binary predictions across all species and sites. Sweeping over different k threshold values provides insights into species presence/absence prediction quality at different probability cutoff points. This helps evaluate how well the model identifies whether a species occurs at a given site based on the predicted probability.

- **Adaptive Top K Accuracy:** This metric measures how accurately the model predicts the top k species at each location, where k is based on what's actually observed at that location. Rather than fixing k across all sites, we adapt it on a per-site basis to focus on the most dominant species expected locally. Specifically, we set k to be the number of species actually observed at that same site in the ground truth data. This provides insights into how well the model is capturing the most significant species expected across different locations, not just the overall most common species. By adapting k per-site, this metric evaluates the model's performance relative to what is actually seen at each specific location of interest.

- **Top 10 Accuracy** This metric evaluates the accuracy of the model's top 10 highest predicted probability species, regardless of how many actual observed species there are at each location. For each site, we take the 10 species with the highest predicted probabilities and compare to the actual species listed for that site in the ground truth data. The accuracy is then calculated across all sites. Unlike the Adaptive Top K metric which bases the value of k on observed species counts per site, this fixes k=10 globally. By isolating consistent top-of rankings, this specifically examines how precisely the model can predict the 10 most likely species overall, highlighting any systematic tendencies to over or under predict the prevalence of common species.

- **Top K=30 Accuracy** Similar to Top 10 Accuracy, this metric evaluates the model's accuracy for the top 30 highest predicted probability species. In this case, the top 30 species predictions are compared to ground truth across all sites regardless of actual observed counts. Fixing k=30 provides a broader perspective on accuracy for near-top predictions compared to the Top 10 metric focused just on the very most likely species.

4.6 Experiments

In our experimental setup, we focus on a region of interest covering 640 m^2 and center-crop the satellite patches to a size of 64×64 around each hotspot. The bands of the satellite images are normalized using statistics from our training set. When incorporating environmental data, we ensure that their resolution matches that of the satellite images. To augment our data, we randomly perform vertical and horizontal flipping. To train the models, we use the cross-entropy loss function:

$$\mathcal{L}_{CE} = \frac{1}{N_h} \sum_h \mathcal{L}_h = \frac{1}{N_h} \sum_h \sum_{s(\text{species})} -y_h^s \log(\hat{y}_h^s) - (1 - y_h^s) \log(1 - \hat{y}_h^s) \tag{4.2}$$

where

- N_h represents the number of hotspots h
- y denotes the model predictions
- $\hat{y}$ represents the ground truth encounter rates

The Satlas and SatMAE models utilize true color images as input and do not support the use of environmental data. The input true color images are normalized to ensure that their pixel values fall within the range of 0 to 1. On the other hand, the MOSAIKS model involves extracting 1024 features from each true color image, combining them with environmental data, and training an XGBoost regressor on the combined features.

For the ResNet-18-based models, we conduct experiments with different inputs, including RGB true color images, RGBNIR reflectance values, and RGBNIR reflectance values combined with bioclimatic and pedologic data. The bioclimatic and pedologic data are normalized based on variable-wise statistics from the training set. We align these data to the resolution of the satellite images and stack the corresponding patches to the images, facilitating efficient training of the models.

We employ these various input configurations and combine environmental data so that our models leverage both visual and ecological information to accurately predict encounter rates and facilitate species distribution mapping in unexplored locations.

Weighted Loss: In addition to the aforementioned experiments, we attempted to improve the training strategy for the regions with fewer complete checklists such as Kenya and some regions of the US by taking into account the number of complete checklists recorded at each hotspot. Our motivation for this approach stems from the observation that Kenya exhibits a relatively smaller number of hotspots, and some of these hotspots are disproportionately represented by the number of complete checklists reported. To address this issue and promote a more balanced representation of encounter rates across the entire dataset, we introduced a weighted loss function.

The weighted loss builds upon our original cross-entropy loss but incorporates an additional weight, which considers the number of complete checklists per hotspot. By applying this weight, we aim to balance the contribution of each hotspot to the overall loss, thereby allowing the model to focus on hotspots with varying numbers of observations more effectively.

The weighted loss function is defined as follows:

$$\mathcal{L}_{WCE} = \frac{1}{N_h} \sum_h w_h \mathcal{L}_h = \frac{1}{N_h} \sum_h w_h \sum_{s(\text{species})} -y_h^s \log(\hat{y}_h^s) - (1 - y_h^s) \log(1 - \hat{y}_h^s) \tag{4.3}$$

where:

- N_h is the total number of hotspots.
- w_h represents the weight associated with the number of complete checklists for hotspot h.
- y denotes the model predictions for encounter rates.
- $\hat{y}$ represents the ground truth encounter rates.

- The summation over s indicates a sum across all species.

By incorporating the weighted loss into our training strategy, we aim to improve the model's ability to learn from hotspots with varying degrees of data availability. The weight assigned to each hotspot dynamically adjusts the contribution of the hotspot to the overall loss, providing a more refined learning process. We hypothesize that this balanced approach will enable our model to generate more accurate and representative predictions across the entire Kenya dataset, ultimately contributing to better species distribution mapping and encounter rate estimations

4.6.1 Experiment setup

The models presented in Section 4.6 establish baseline performance on the task of regressing encounter rates using different model architectures and input features. Hyperparameter tuning was not performed to focus on demonstrating feasibility rather than maximizing performance. All results should be considered preliminary baselines for future work.

Compute Constraints

Experiments were run on a GPU compute cluster at McGill University and Mila [1] equipped with Nvidia A100 GPUs. For the ResNet deep learning baselines, training took up to 54 minutes per epoch using a batch size of 128, 16GB RAM, and the ResNet-18(refl+env+RM) model on a single A100 GPU. Other ResNet-18 variants required less time. Some experiments utilized more powerful GPUs, reducing training times to under 1 day for 50 epochs. Each baseline was trained for 3 random seeds. The Satlas and SatMAE baselines required approximately 2 days of training per experiment using a single GPU and the PyTorch framework [76].

[1] Mila Quebec AI Institute cluster

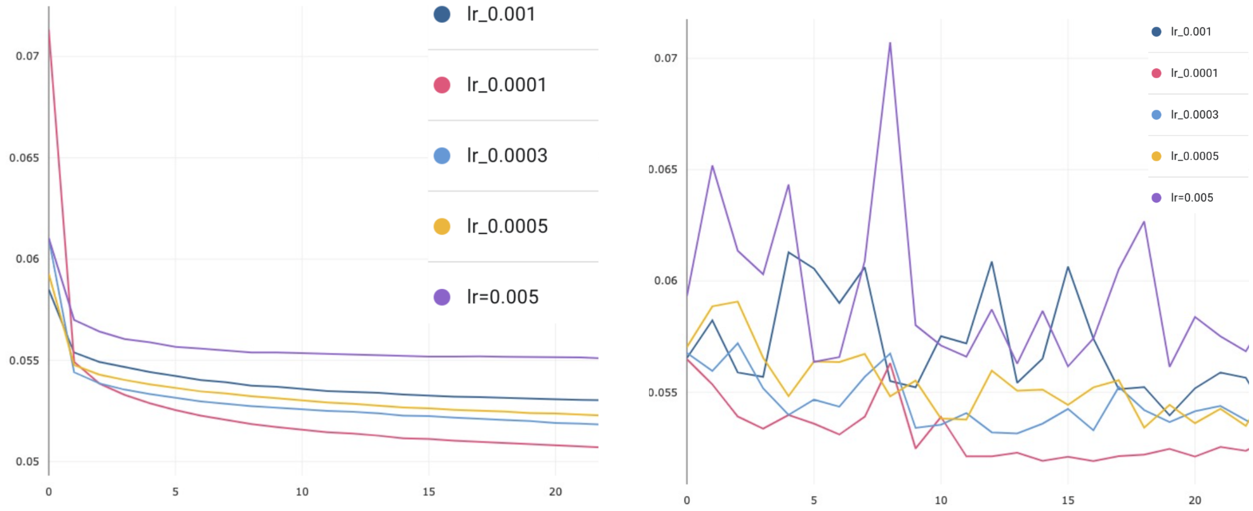
4.6.2 Hyperparameters

In our experiments, we initially set the learning rate to 3×10^{-5}. However, as part of a detailed analysis, we conducted a hyperparameter search to fine-tune the learning rate for the ResNet-18 (RGBNIR + env + RM) baseline model on the SatBird-USA-summer dataset. This search involved exploring learning rates in the range of 10^{-4} to 5×10^{-3}. Interestingly, we found that a learning rate of 10^{-4} yielded slightly better results than the initial 3×10^{-5} setting. This suggests that further improvements in performance could be achieved with more comprehensive hyperparameter tuning.

To provide a visual representation of this analysis, we present the loss curves for several models resulting from our hyperparameter search in Figure 4.2 below.

All the models in our experiments were trained using a batch size of 128 and the Adam optimizer [65]. We ran each experiment with three different random seeds and report the average results on the test set across these seeds.

The data splitting strategy described in Section 3.4 was employed for training, validation, and test sets. Specifically, all ResNet-18 models were trained for 50 epochs, while Satlas and SatMAE models were trained for 100 epochs, with the pre-trained model's layers frozen and only the last layer being fine-tuned.



(a) Train loss vs. learning rate

(b) Validation loss vs. learning rate

Figure 4.2: Training and validation loss curves for different learning rates. Lower learning rates converge more slowly but achieve lower loss, indicating potential for further optimization

5

Integrating Prior Knowledge to Address Data Scarcity

eBird data exhibits intrinsic sparsity similar to many other biodiversity datasets, following a heavy-tailed distribution [84]. Consequently, our models do not achieve optimal performance in regions with fewer complete checklists per hotspot, referred to as *low data regimes*. In this chapter, we present approaches for developing more robust machine-learning models that can overcome these data-scarce areas. We start by defining what constitutes sparse hotspots and detailing our proposed solution for addressing data scarcity in Section 5.1. Next, we discuss the primary factors that lead to sparse hotspots in eBird in Section 5.1.1. Finally, in Section 6.1.2, we present results after applying a Bayesian approach to handle sparse hotspots demonstrating improved model robustness.

5.1 Addressing Data Sparsity in eBird Hotspots

eBird hotspots are critical data points for understanding bird species distribution and occurrence. However, the volume of data available for each hotspot varies significantly. Some hotspots, enriched by years of birder observations, boast over 1000 complete checklists. Conversely, other hotspots remain relatively underexplored, with a mere 5 to 10 complete checklists submitted. This disparity in data distribution poses challenges in accurately characterizing the full spectrum of avian species inhabiting underrepresented hotspot locations. Hotspots with fewer submitted checklists are likely to exhibit substantial gaps in species inventories, failing to capture the complete biodiversity present at the site. Consequently, models that rely solely on raw observational data from these sparse hotspots may generate biased predictions about species occurrence and distribution.

This chapter delves into Bayesian approaches to amalgamate limited ground truth data from underrepresented hotspots with model-generated probability distributions. The objective is to derive more robust and reliable predictions of species occupancy, particularly for hotspots with minimal observational data. By integrating these two sources of information, we can enhance the accuracy of our models and provide a more comprehensive understanding of avian biodiversity across all eBird hotspots. First, we begin by discussing some of the factors leading to data sparsity. Section 5.1.2 Introduces a Bayesian approach we adapted for this problem.

5.1.1 Factors Contributing to eBird Sparsity

Several factors contribute to the sparsity and potential bias present in eBird data:

Taxonomic Bias

Birders may not uniformly detect or report all species. There is a tendency to prioritize recording sightings of rare, exotic, or charismatic species, often at the expense of reporting more common birds. This selective recording can filter the data, leading to species inventories that do not accurately reflect the full avian diversity at a hotspot.

Temporal Bias

The activity and detectability of species fluctuate considerably throughout the day and across different seasons. Sparse sampling during certain times of day or year may result in failure to record species that are only present or vocal during limited periods. For instance, nocturnal birds like owls and nightjars active after dusk are likely to be missed by daytime birding. Similarly, transient migrants that are present only briefly during spring or fall may be overlooked.

Spatial Bias

Birders often concentrate their efforts near trails, lookouts, parking areas, and other easily accessible locations. As a result, species inhabiting remote or challenging-to-navigate areas may receive less sampling and documentation.

Class Imbalance

Some species naturally occur in higher abundances and are easier to detect than rarer species. With limited sampling, common species will be observed and reported frequently, while rare birds are more likely to be overlooked.

5.1.2 Integrating Limited Ground Truth with Predictive Modeling

At data-sparse hotspots, we lack sufficient checklists to definitively characterize community composition. However, we can still derive useful information from two approximate probability distributions:

1. The *ground truth* [1] distribution based on limited observational data
2. The *model-predicted* probability distribution

The ground truth distribution represents the best available, yet limited, data on species occurrence derived from birder observations. While prone to sampling bias, these records

[1] *Throughout this section, we shall refer to the sparse observations as the "ground truth", as they reflect actual on-the-ground conditions*

reflect on-the-ground conditions. In contrast, our model predictions estimate species probability of occurrence based solely on environmental conditions captured in satellite imagery. However, models can misrepresent fine-scale conditions not perceivable through imagery alone. By integrating these two distributions using Bayesian methods, we can leverage the strengths of both to generate more robust predictions, especially for sparsely sampled sites. Bayesian approaches allow us to incorporate prior knowledge (the ground truth data) to refine model-based probability estimates. The following section details our proposed Bayesian integration methodology for combining ground truth and model predictions.

To integrate ground truth with model predictions, we employ a Bayesian framework to derive posterior probability distributions that incorporate both sources of information through Bayes' theorem:

$$P(S|G, M) = \frac{P(G|S)P(S|M)}{P(G|M)} \tag{5.1}$$

Where:

- $P(S|G, M)$ = Posterior probability of species S occurring given ground truth G and model prediction M
- $P(G|S)$ = Likelihood of observed ground truth G given species S is present
- $P(S|M)$ = Model-predicted probability for species S
- $P(G|M)$ = Normalizing constant

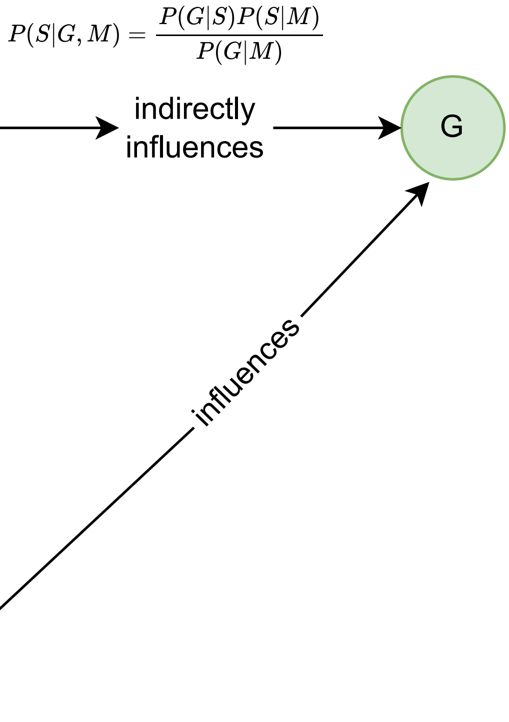


Figure 5.1: Dependency Graph: Illustrates the influence of Model Predictions (M) on both Species Occurrence (S) and Ground Truth (G), and the indirect dependency of $P(G|M)$ in the Bayesian framework for equation 5.1

The resulting $P(S|G,M)$ posterior probability distribution integrates our model's predicted probability for each species with the ground truth data observed at a site. This allows limited ground truth records to refine and improve model-based predictions.

5.1.3 Factorization

The joint distribution $P(S|G,M)$ can be factorized using the chain rule of probability along with a conditional independence assumption that the model predictions M depend only on the species S, not directly on the ground truth data G. This gives: $P(S|G,M) = P(G|S)P(S|M)P(M)$. Here $P(S|M)$ follows from Bayes' theorem and $P(M)$ serves as a normalization constant. This derived form aligns with the Bayesian update equation structure used to integrate ground truth and model predictions. Specifically, it decomposes the joint into the likelihood of G given S, model-based probabilities for S, and a constant term—enabling formal information fusion under a Bayesian framework. While simplifying assumptions are made for computational ease, this captures the core statistical relationships between data and models needed to generate posterior probability estimates that combine both sources of knowledge about species distributions.

5.1.4 Benefits to Bayesian integration

Bayesian integration provides several benefits:

- Leverages strengths of both ground truth data and predictive modelling. The ground truth provides real-world observations that capture on-the-ground conditions. Meanwhile, the model provides complete coverage and environmental context. Integrating them gives us the best of both worlds.

- Accounts for sampling bias and gaps in ground truth data. The model predictions help fill in gaps where ground truth data is sparsely sampled or absent. This overcomes the limitations of incomplete observational records.

- Constrains model predictions using on-the-ground conditions. Ground truth acts as a check on model estimates, anchoring them to real observations that reflect fine-scale environmental factors. This enhances reliability.

- Can emphasize ground truth data in sparsely sampled sites and model predictions where data is abundant. The Bayesian approach allows the influence of ground truth vs. model predictions to vary based on the amount of data available.

- Quantifies uncertainty in the integrated predictions. Bayesian methods provide uncertainty estimates alongside the probability predictions. This allows us to assess reliability and account for uncertainty in downstream analyses.

In the next section, we detail Bayesian methods to implement the integration and assess resulting improvements in model accuracy. Bayesian methods allow us to effectively integrate the model and ground truth data to improve model accuracy, even for hotspots with minimal ground truth. A key challenge is determining how to combine the model and ground truth probability distributions when both have limitations. The model estimates can be inaccurate, while the ground truth data is imprecise until sufficient observations are made at a location. Finding an optimal way to merge these two probability distributions is critical, as it would yield a superior estimate compared to using either alone. This issue warrants further investigation given its importance.

5.2 Bayesian Pipeline

We attempt to utilize Bayesian Evaluation Strategies, which consist of a *prior distribution*, *posterior distribution* and a *likelihood function*. A prior distribution is a probability distribution used to represent an initial belief about the probability of a certain event or parameter before any new evidence is taken into account. The likelihood function is a function that describes the probability of observing a certain set of data given a particular set of parameters.

Here, Bayesian inference is used to estimate the probability that a certain bird species is present in a certain location (hotspot) using data collected in the form of checklists. The prior distribution represents our initial belief about the probability before any observations are made, while the likelihood function describes the probability of observing a certain checklist given a certain probability of the presence of the species in the hotspot. As new observations are made, the likelihood function is used to update the prior distribution and our belief about the probability of the species presence in the hotspot converges towards the true probability. Figure 5.2 gives an overview of the pipeline.

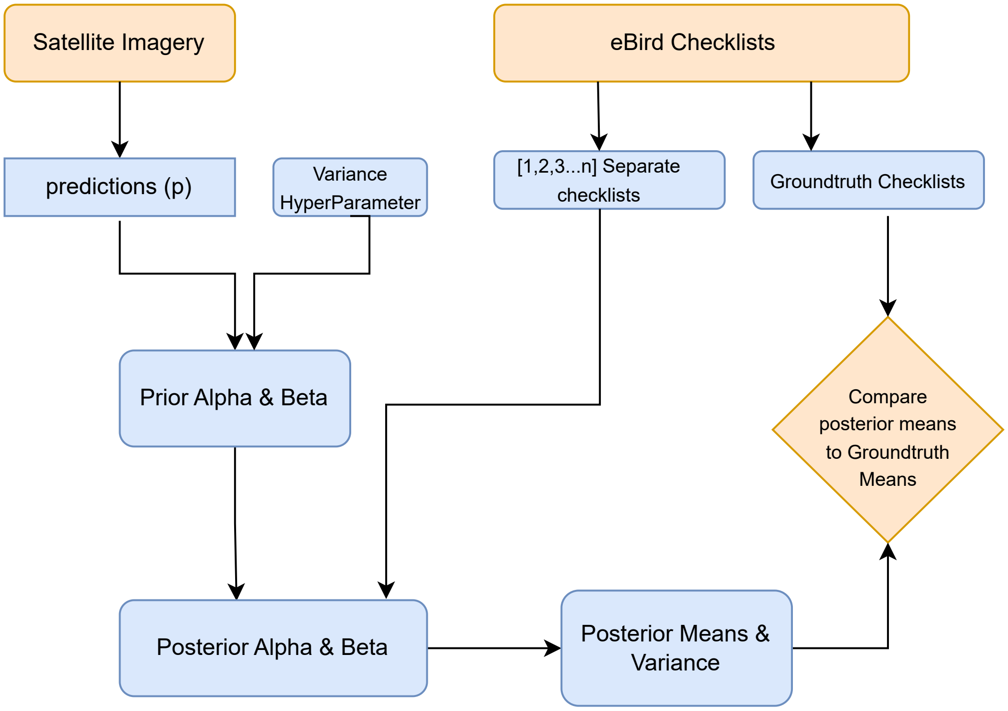


Figure 5.2: The pipeline integrates predicted and observed encounter probabilities in a Bayesian framework. Satellite imagery generates initial predictions $p_{h,s}$ for each species s and hotspot h. These inform a beta prior distribution, which is updated by sampling checklist data $C'h, s$ to a posterior distribution. The posterior mean $p'h, s$ is compared to the ground truth $p_{h,s,true}$ from full checklists to assess accuracy. This combines the modeled priors and observational data to derive robust probability estimates.

5.2.1 Estimating Species Presence Probabilities with Bayes' Theorem

eBird survey data is often collected from a set of sites or *hotspots* that are repeatedly surveyed over time. Let's assume we have a set of hotspots:

$$H = \{h_1, h_2, ..., h_n\} \tag{5.2}$$

where n is the number of hotspots.

For each hotspot h_i, there exists a set of checklists

$$C_i = \{c_1, c_2, ..., c_m\} \tag{5.3}$$

where m is the number of checklists for hotspot i.

For each checklist c_j: A binary variable $x_{i,j,k}$ indicating whether species k is observed in checklist j of hotspot i or not.

Our goal is to estimate the true probability θ_{ih} that each species i is present at each hotspot h. Specifically, θ_i represents the underlying geographic distribution or range map for species i, indicating the probability of occurrence across all locations based on that species' habitat preferences and ranges. We aim to model each species' complete range map θ_i. We can model the observation process using a likelihood function:

We can write the likelihood function as:

$$P(C_i|\theta_i) = \prod_{j=1}^{m} \prod_{k=1}^{20} (\theta_i)^{x_{i,j,k}} (1 - \theta_i)^{1 - x_{i,j,k}} \tag{5.4}$$

where θ_i is the probability that species k is present in hotspot i.

The prior probability distribution of the underlying rangemap of a species at a hotspot h is given by:

$$P(\theta_{ih}) \tag{5.5}$$

By combining the likelihood function and the prior distribution, we can update our beliefs about the probability that a certain bird species is present in a certain location (hotspot) using data collected in the form of checklists.

We use the Bayes' theorem to update our beliefs as follows:

$$P(\theta_{ih}|C_i) = \frac{P(C_i|\theta_{ih})P(\theta_{ih})}{P(C_i)} \tag{5.6}$$

where $P(\theta_{ih}|C_i)$ represents our updated knowledge about the species presence probability θ_i after considering the survey data collected in that hotspot h_i in the form of checklists C_i. In this way, Bayes' theorem provides a principled approach to estimating species presence from checklist data.

5.2.2 Conjugate Priors and Beta distribution theory

For the likelihood function defined in equation 5.4, the Beta distribution is a conjugate prior for this binomial likelihood function. The Beta distribution is:

$$P(\theta_i|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta_i^{\alpha-1}(1-\theta_i)^{\beta-1} \tag{5.7}$$

where: $\Gamma()$ denotes the gamma function. α and β can be interpreted as prior observations of "successes" and "failures" respectively, providing a natural way to incorporate prior knowledge or beliefs.

If we use a beta prior for θ_i, then the posterior distribution is also a beta distribution:

$$P(\theta_i|C_i,\alpha,\beta) = \text{Beta}\left(\alpha + \sum_j \sum_k x_{i,j,k}, \beta + mK - \sum_j \sum_k x_{i,j,k}\right) \tag{5.8}$$

The posterior parameters are simply the prior parameters plus the number of "successes" and "failures" observed in the data.

The beta-binomial model is simple to implement, fast to compute, and provides a natural interpretation. The posterior means $E[\theta_i|C_i]$ can be used as the estimates of the species presence probabilities at each hotspot. The posterior variances quantify the remaining uncertainty.

So by using a conjugate beta prior, we obtain a tractable Bayesian update for the species presence probabilities that leverages both the prior information and observed data.

5.3 Incorporating Conjugate Priors

We develop a Bayesian model to predict encounter rate probabilities for different species at biodiversity hotspots, especially leveraging even the few checklists in some hotspots. We model the probabilities using a beta distribution, which is suitable for proportions ranging between [0,1]. The distribution is defined by two positive shape parameters: α and β, which can be interpreted as ″presence″ and ″absence″ respectively of a species in a checklist. (This is the equivalent of the successes or failures of a coin toss) [19]

Given the initial encounter probabilities, denoted by $p_{h,s}$, derived from our preliminary predictions, we can parameterize the beta prior. Here, we introduce a hyperparameter τ to modulate the variance $v_{h,s}$ intrinsic to the binomial distribution as:

$$v_{h,s} = \tau \times p_{h,s} \times (1 - p_{h,s})$$

With the above variance, we compute the initial $\alpha_{h,s}$ and $\beta_{h,s}$ utilizing both $p_{h,s}$ and $v_{h,s}$ as follows:

$$\alpha_{h,s} = p_{h,s} \times \left(\frac{p_{h,s}(1 - p_{h,s})}{v_{h,s}} - 1 \right) \tag{5.9}$$

$$\beta_{h,s} = \alpha_{h,s} \times \left(\frac{1}{p_{h,s}} - 1 \right) \tag{5.10}$$

To infuse the checklist data $C_{h,s}$ into our model, we select subsets $C'_{h,s} \subseteq C_{h,s}$ of sizes 2, 5, or 10 checklists. The parameters $\alpha_{h,s}$ and $\beta_{h,s}$ are updated by accumulating ″successes″

and "failures" from $C'_{h,s}$:

$$\alpha'_{h,s} = \alpha_{h,s} + \sum_{c \in C'_{h,s}} c \tag{5.11}$$

$$\beta'_{h,s} = \beta_{h,s} + |C'_{h,s}| - \sum_{c \in C'_{h,s}} c \tag{5.12}$$

Subsequently, the posterior mean $p'_{h,s}$ and variance $v'_{h,s}$ are computed:

$$p'_{h,s} = \frac{\alpha'_{h,s}}{\alpha'_{h,s} + \beta'_{h,s}} \tag{5.13}$$

$$v'_{h,s} = \frac{\alpha'_{h,s}\beta'_{h,s}}{(\alpha'_{h,s} + \beta'_{h,s})^2(\alpha'_{h,s} + \beta'_{h,s} + 1)} \tag{5.14}$$

In our concluding analysis, the model's accuracy is ascertained by combining with the ground truth mean $p_{h,s,\text{true}}$ obtained from all checklists. The accuracy metric, $a_{h,s}$, is thus the absolute difference between $p_{h,s,\text{true}}$ and $p'_{h,s}$.

We incorporate these conjugate priors into our approach and show that we can improve the confidence of the model even better by sampling more checklists from the ground truth and updating the posterior alpha and Beta parameters with the predictions from the satellite imagery. Section 6.1.2 shows an overview of these results.

We discuss these findings in detail in the next section.

6

Results and Discussion

This chapter will address the evaluation and discussion of the achieved results, the chosen methodology and the validity as well as reliability of the experiments. We begin in Section 6.2.1 by discussing the different data modalities and how they influenced the model performance. Then we give an overview of the architectures in section 6.2.2. In section 6.2.3, we discuss the effect of the different loss functions used, and finally in section 6.4 we discuss the Bayesian approach to improve predictions for hotspots with fewer complete checklists.

6.1 Results

6.1.1 Baseline Model Results

In this section, we present the results obtained from our baseline models, using the USA-summer as our primary dataset for experiments due to its extensive number of hotspots. The evaluation of the models on the test set is summarized in Table 6.1, while the results. Furthermore, we provide additional baseline results for USA-winter dataset in Table 6.2 and Kenya in Table 6.3.

In the tables below; *img* refers to using the RGB true color image. All ResNet-18 baselines use *refl+env*, which corresponds to using RGBNIR reflectance bands and the environmental data (bioclimatic and pedologic variables), and WL refers to the weighted Loss. *scratch* refers to training ResNet-18 model from scratch. *finetune-USA* refers to fine-tuning a ResNet-18 model with weights transferred from USA-summer. *freeze-USA* refers to using ResNet-18 weights transferred from the USA-summer dataset while training the last layer only.

Table 6.1 showcases the performance of our baseline models on the test set of USA-summer. The baseline models establish a solid foundation for further investigations. The test results serve as benchmarks for evaluating the effectiveness of our proposed approach. Top-K refers to the Adaptive Top-K accuracy defined in 4.5 and the rest of the metrics are defined in 4.5

Using USA-summer as our main dataset allows us to capture a diverse range of hotspots and species, enabling comprehensive evaluations of our models' capabilities.

Model	**MAE**[10^{-2}]	**MSE**[10^{-2}]	**Top-10**	**Top-30**	**Top-k**
Mean encounter rates	3.07 ± 0.0	0.9 ± 0.0	24.7 ± 0.0	34.9 ± 0.0	40.5 ± 0.0
Environmental baseline	2.5 ± 0.0	0.7 ± 0.0	38.4 ± 0.1	51.3 ± 0.1	56.6 ± 0.1
Satlas (img)	2.8 ± 0.0	0.8 ± 0.0	29.1 ± 0.01	44.7 ± 0.0	46.1 ± 0.0
SatMAE (img)	3.1 ± 0.0	0.9 ± 0.0	24.7 ± 0.0	38.0 ± 0.3	40.5 ± 0.0
MOSAIKS (img+env)	2.2 ± 0.0	0.6 ± 0.0	45.9 ± 0.0	59.4 ± 0.0	64.3 ± 0.0
ResNet-18 (img)	2.7 ± 0.0	1.0 ± 0.0	20.4 ± 2.2	35.7 ± 0.6	38.7 ± 0.7
ResNet-18 (refl)	3.2 ± 0.0	0.9 ± 0.0	23.8 ± 0.9	36.4 ± 0.2	39.1 ± 0.6
ResNet-18 (img+env)	2.0 ± 0.0	0.6 ± 0.0	45.0 ± 1.1	64.8 ± 1.2	65.2 ± 1.0
ResNet-18 (refl+env)	**2.1** ± 0.0	**0.6** ± 0.0	**46.4** ± 0.8	**65.9** ± 0.7	**66.2** ± 0.6
ResNet-18 (refl+env+RM)	**2.0** ± 0.0	**0.6** ± 0.0	**46.2** ± 0.4	**65.9** ± 0.3	**66.3** ± 0.2

Table 6.1: Test results on USA-Summer: Best results are shown in bold. *img* refers to using the RGB true color image, *refl* refers to using RGBNIR reflectance bands, *env* refers to using the environmental data (bioclimatic and pedologic variables). *RM* refers to the use of range maps. Top-K refers to the Adaptive Top-K accuracy defined in 4.5

Model	**MAE**[10^{-2}]	**MSE**[10^{-2}]	**Top-10**	**Top-30**	**Top-k**
Mean encounter rates	2.4 ± 0.0	0.7 ± 0.0	25.9 ± 0.0	44.5 ± 0.0	50.5 ± 0.0
Environmental baseline	1.7 ± 0.0	0.44 ± 0.0	51.2 ± 0.0	63.7 ± 0.0	67.7 ± 0.0
Satlas (img)	2.2 ± 0.01	0.7 ± 0.0	30.4 ± 0.3	49.2 ± 0.2	51.3 ± 0.2
SatMAE (img)	2.4 ± 0.01	0.7 ± 0.0	25.7 ± 0.4	45.5 ± 0.3	48.1 ± 0.1
MOSAIKS (img+env)	2.2 ± 0.0	0.6 ± 0.0	47.8 ± 0.0	62.3 ± 0.0	67.7 ± 0.0
ResNet-18 (img)	2.0 ± 0.22	1.1 ± 0.22	10.0 ± 3.3	25.5 ± 4.9	31.3 ± 3.8
ResNet-18 (refl)	2.3 ± 0.13	0.7 ± 0.01	27.2 ± 2.5	47.5 ± 1.7	49.7 ± 1.9
ResNet-18 (img+env)	1.6 ± 0.01	0.4 ± 0.0	52.3 ± 0.36	**69.9** ± 0.25	69.9 ± 0.2
ResNet-18 (refl+env)	1.6 ± 0.0	0.4 ± 0.01	52.0 ± 0.16	69.5 ± 0.2	69.9 ± 0.2
ResNet-18 (refl+env+RM)	**1.6** ± 0.01	**0.4** ±0.01	**52.3** ± 0.15	69.5 ± 0.14	**70.1** ± 0.03

Table 6.2: Test results on the SatBird-USA-winter: Best results are shown in bold. *img* refers to using the RGB true color image, *refl* refers to using RGBNIR reflectance bands, *env* refers to using the environmental data (bioclimatic and pedologic variables). *RM* refers to the use of range maps.

6.1.2 Bayesian Model Results

We present the results of the Bayesian Sampling strategy in the table 6.4 below. We see that as we sample more and more checklists, the overall Mean Absolute Error Reduces significantly reduces.

Model	**MAE**[10^{-2}]	**MSE**[10^{-2}]	**Top-10**	**Top-30**	**Top-k**
Mean encounter rates	3.2 ± 0.0	1.4 ± 0.0	14.1 ± 0.0	18.1 ± 0.0	21.3 ± 0.0
Environmental Baseline	2.8 ± 0.0	1.3 ± 0.00	21.4 ± 0.0	29.5 ± 0.0	33.3 ± 0.0
Satlas (img)	2.9 ± 0.0	1.3 ± 0.0	23.8 ± 0.2	44.0 ± 0.2	24.5 ± 0.1
SatMAE (img)	3.1 ± 0.0	1.4 ± 0.0	21.6 ± 0.4	40.7 ± 0.0	22.4 ± 0.5
MOSAIKS (img+env)	3.4 ± 0.0	1.4 ± 0.0	12.8 ± 0.0	16.9 ± 0.0	19.8 ± 0.0
ResNet-18(scratch)	3.3 ± 0.2	1.4 ± 0.0	18.8 ± 0.2	33.0 ± 0.2	23.2 ± 0.2
ResNet-18(finetune-USA)	3.2 ± 0.0	1.4 ± 0.0	18.8 ± 0.0	32.6 ± 0.3	23.5 ± 0.1
ResNet-18(freeze-USA)	3.2 ± 0.0	1.4 ± 0.0	18.8 ± 0.0	32.8 ± 0.5	23.5 ± 0.0
WL-ResNet-18(scratch)	4.6 ± 0.1	1.5 ± 0.0	18.6 ± 0.2	32.7 ± 0.7	23.0 ± 0.7
WL-ResNet-18(finetune-USA)	4.3 ± 0.1	1.5 ± 0.0	18.8 ± 0.1	33.2 ± 0.1	23.5 ± 0.1
WL-ResNet-18(freeze-USA)	4.2 ± 0.0	1.4 ± 0.0	18.5 ± 0.0	32.8 ± 0.5	23.5 ± 0.0

Table 6.3: Test results on the Kenya dataset: *img* refers to using the RGB true color image. All ResNet-18 baselines use *refl+env*, which corresponds to using RGBNIR reflectance bands and the environmental data (bioclimatic and pedologic variables), and WL refers to the Weighted Loss. *scratch* refers to training ResNet-18 model from scratch. *finetune-USA* refers to fine-tuning a ResNet-18 model with weights transferred from USA-summer. *freeze-USA* refers to using ResNet-18 weights transferred from the USA-summer dataset while training the last layer only.

Model	**Test MAE**	**Test MSE**	**Test Top10**	**Test Top30**
RGBNIR + ENV + RM	0.020188	0.005955	0.466101	0.662649
Beta_updated				
5 Checklists	0.011428	0.002287	0.675884	0.727243
10 Checklists	0.008104	0.001278	0.759780	0.787577
Beta_updated, Var=1				
5 Checklists	0.012230	0.003139	0.636751	0.664888
10 Checklists	0.008156	0.001518	0.744447	0.764256
Beta_updated, Var=0.25				
5 Checklists	0.011655	0.001874	0.679729	0.728702
10 Checklists	**0.008592**	**0.001126**	**0.759885**	**0.787248**

Table 6.4: Summary of Test Results for Different Model Configurations. Our results suggest that providing more checklist information enhances the model's performance. The most effective configuration utilized 10 checklists and a variance parameter of 0.25. This increase in checklist sampling boosts the model's overall confidence, yielding predictions superior to those of the basic model (RGBNIR, ENV + RM).

6.2 Discussion of Results

6.2.1 Data Modalities

We observe that models trained with satellite images only, either true color images or reflectance values do not outperform the environmental baseline. However, combining the satellite images with environmental rasters results in significant improvement in all metrics, with the highest improvement in the top-30 (14%) top-K (10%), and top-10 (8%) metrics. This highlights the value of satellite data for our task, and also the importance of using both environmental and remote sensing data. The best models utilize RGBNIR reflectance images, along with environmental data. We suggest incorporating other information as input such as landcover and alitutude data which can be obtained from publicly available sources for further improvements.

6.2.2 Architectures

The Resnet architecture excelled at the primary task. Resnet18 worked across all the data modalities and showed it's utility for the task at hand. The transformer based models, both Satlas and SatMAE both outperformed the ResNet-18 model on the true colored image, this suggests the potential of transformer-based models. We note that Satlas outperforms SatMAE, likely because the former was pretrained on a very large dataset of Sentinel-2 images which is the same source (and has the same resolution) as our images. For SatMAE, the model was trained on the fMoW dataset with a higher resolution (0.5m) than the images in our dataset. MOSAIKS also performs reasonably well compared to other deep learning models. for the true color image and environmental rasters input especially considering it is is lightweight model.

6.2.3 Comparison of different loss functions

Throughout this research, our experiments use cross entropy as the loss function as mentioned in Sec 5.2. in addition, we conducted further experiments using other regression loss functions such as L1 loss and L2 loss. We also extended our cross-entropy loss to focal loss to address the class imbalance. Table 6.5 shows that cross entropy loss achieves the best accuracies, and even the best MSE over the test set.

Model	**MAE**$[10^{-2}]$	**MSE**$[10^{-2}]$	**Top-10**	**Top-30**	**Top-k**
Cross Entropy	2.1 ± 0.02	**0.59** ± 0.0	**45.68** ± 0.36	**65.39** ± 0.5	**65.85** ± 0.37
L1 Loss	**1.78** ± 0.01	0.7 ± 0.0	41.87 ± 0.42	58.5 ± 0.16	57.28 ± 0.15
L2 Loss	2.59 ± 0.05	0.6 ± 0.0	43.79 ± 0.49	62.86 ± 0.5	63.12 ± 0.41
Focal Loss	3.2 ± 0.11	0.7 ± 0.0	36.23 ± 0.33	54.79 ± 0.3	56.54 ± 0.14

Table 6.5: Using different loss functions for Resnet-18 baseline: this table compares the performance of Resnet-18 baseline when using different loss functions. Cross entropy Loss outperforms all other losses.

6.2.4 Model inference

During inference, the best-performing models were the Resnet models, however, we noticed a degradation in performance, especially in regions that do not have sufficient complete checklists. Hence We employed Bayesian models for better inference and this proved to perform better. The Bayesian inference strategy is discussed in section 6.4

6.3 Analysis of results

On the test set of SatBird-USA-summer, the top 50 species with lowest test-MSE are unsurprisingly species with very low number of occurrences in the training set (< 11) (so the model can predict a zero encounter rate and have good MSE for these species). Among those species we find species breeding in arctic regions (e.g. Bombycilla garrulus, Calcarius lapponicus) and seabirds (e.g. Uria lomvia). By design of our dataset, we don't have many observations of seabirds since we excluded all hotspots outside the boundaries of the continental US). Species with largest MSE have $>$ 30k hotspots where they were ob-

Species scientific name	Training occurrences	MSE ranking	MSE
Uria lomvia	2	1	6.488607e-08
Calcarius lapponicus	2	1	7.974181e-08
Calidris maritima	6	3	8.040372e-08
Limosa lapponica	11	4	8.041987e-08
Geopelia striata	1	5	8.050395e-08
Zonotrichia querula	6	5	8.085927e-08
Acanthis flammea	2	7	8.093858e-08
Stercorarius pomarinus	4	8	8.108264e-08
Bombycilla garrulus	2	9	8.149682e-08
Sicalis flaveola	3	10	8.168477e-08

Table 6.6: Top 10 predicted species on the USA-summer test set according to MSE. *Occurrences* refers to the number of training hotspots for which the encounter rate (target value) is non-zero.

Species scientific name	Training occurrences	MSE ranking	MSE
Agelaius phoeniceus	45670	684	0.083132
Turdus migratorius	60637	683	0.078335
Melospiza melodia	46416	682	0.074919
Zenaida macroura	60914	681	0.072872
Haemorhous mexicanus	43441	680	0.062781
Passer domesticus	36534	679	0.062168
Hirundo rustica	42716	678	0.060854
Anas platyrhynchos	31729	677	0.056884
Corvus brachyrhynchos	53900	676	0.054558
Setophaga petechia	29713	675	0.053734

Table 6.7: Bottom 10 predicted species on SatBird-USA-summer test set according to MSE in descending order. *Occurrences* refers to the number of training hotspots for which the encounter rate (target value) is non-zero.

served in the training set. Tables 6.6 and 6.7 show the 10 best and worst predicted species on the USA-summer test set according to MSE.

6.3.1 Top Predicted Species

We show the squared error distribution by target value (encounter rates) for the USA-summer test set in Fig. 6.2. We binned the target values 19 bins and for each bin averaged the squared errors. The squared error is higher on average for higher target values. In

Fig. 6.1, we show the predicted values against the target values. The target values were binned in 19 bins. The model generally underestimates the target values.

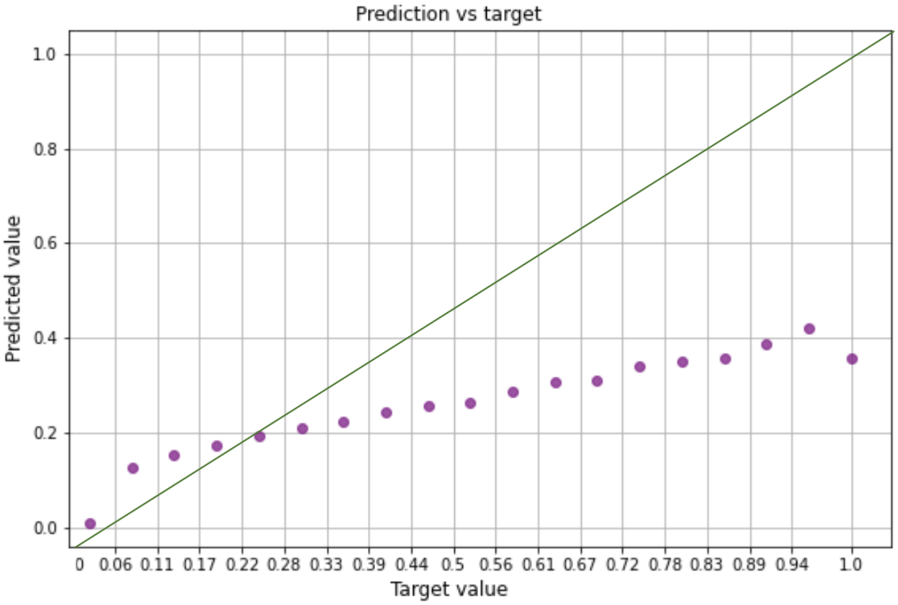


Figure 6.1: Predicted vs target values (binned in 19 bins) for the USA-summer test set.

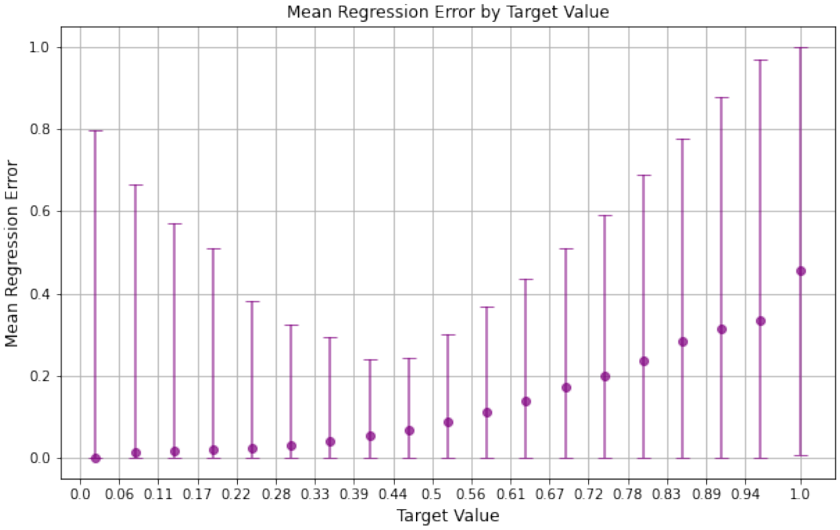


Figure 6.2: MSE value distribution by target value (binned in 19 bins) for the USA-summer test set. The bars show the minimum and maximum squared error for targets in a given bin.

We show the geographic distribution of the top-k error in the predictions for the USA-summer test set in Fig. 6.3

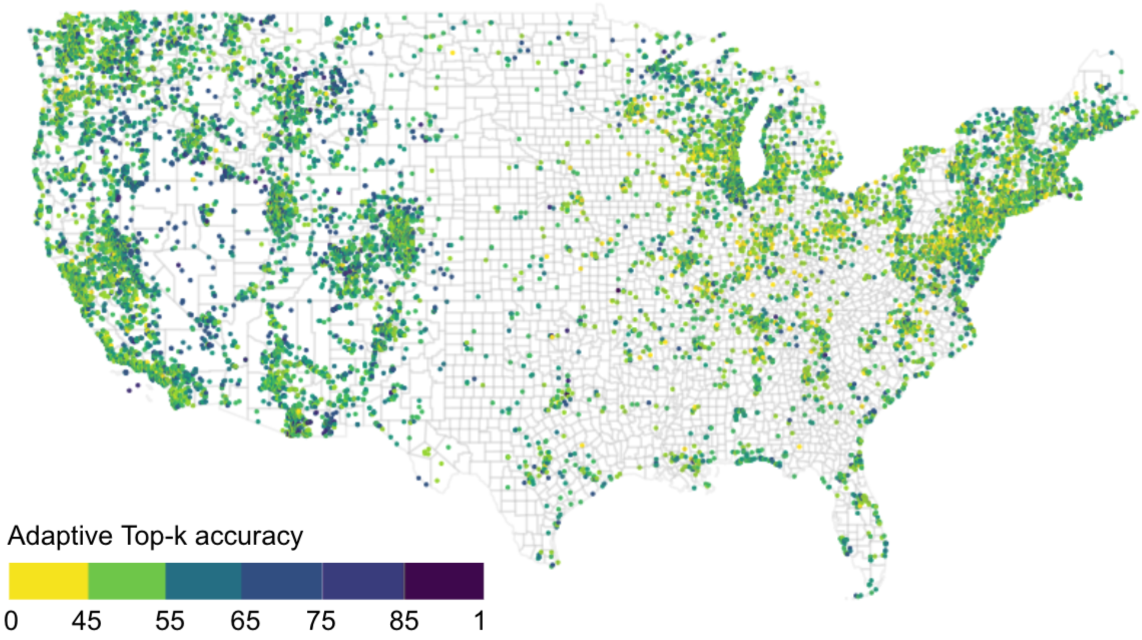


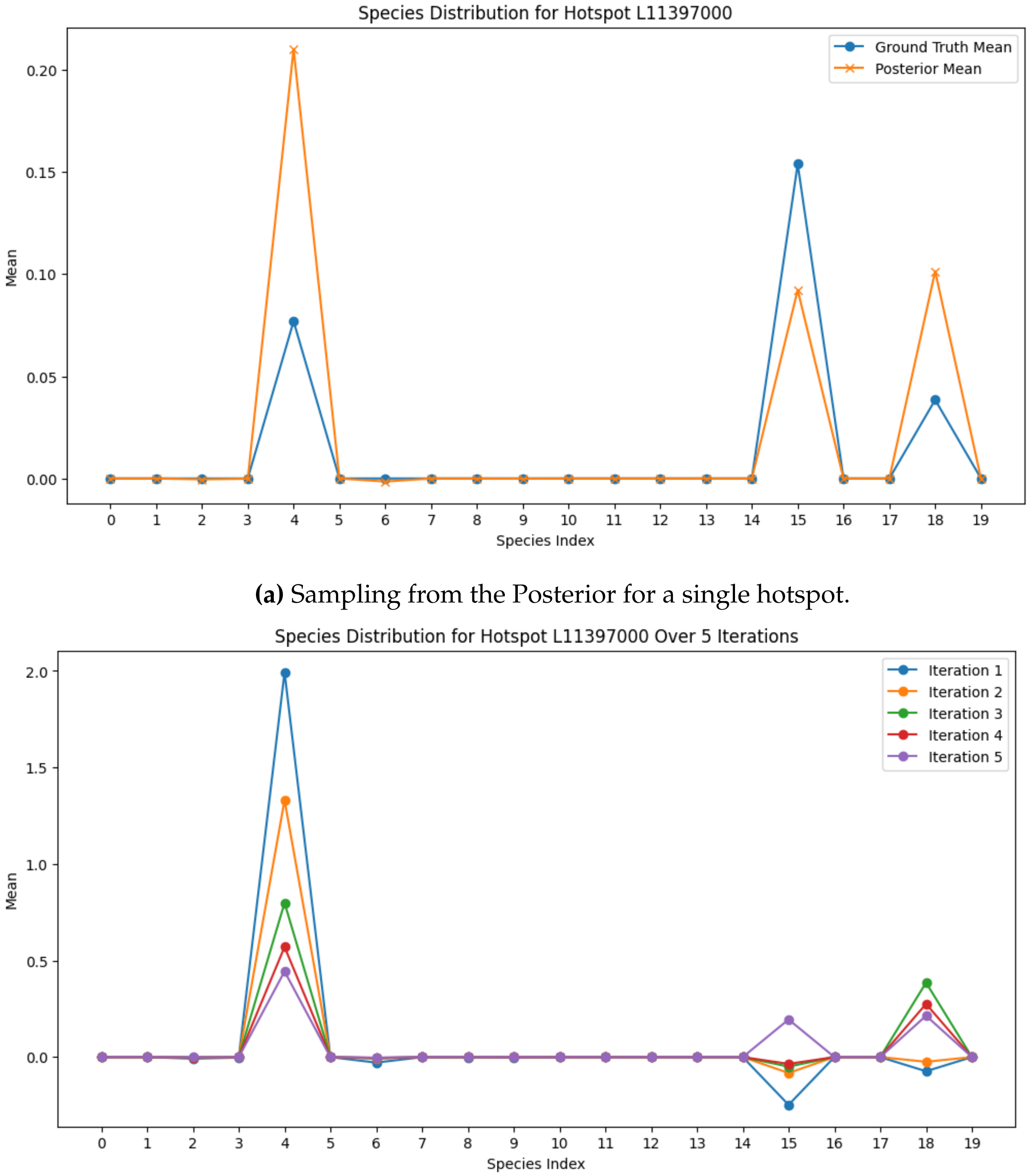
Figure 6.3: USA-summer test set hotspots colored by adaptive top-k performance for the ResNet-18 (refl+env+RM) model.

6.4 Improving Predictions with Bayesian Approaches

6.4.1 Single Hotspot Scenario

We first evaluated how well the posterior mean estimates match the ground truth mean across all species in our test set. Figure 6.4a shows the results for the top 20 species (out of 684 total). We see that the posterior means closely follow the ground truth, as expected when sampling from the posterior. We also examined how additional checklists improve the posterior estimates by performing successive Bayesian inference iterations, each adding 1, 2, 3, 4, or 5 additional checklists. Figure 6.4b illustrates that with more checklists, the posterior means converge towards the ground truth means, especially for under-represented species. This demonstrates that the Bayesian approach successfully leverages additional checklists to improve model confidence, particularly for species with limited data in the initial few checklists for a given hotspot.

The results, for a single hotspot, show that the Bayesian methodology produces sensible posteriors that align with ground truth and improve with more data, validating its ability to enhance predictions from limited observational data. Next, we extend this to all the hotspots in our test set and discuss the results.



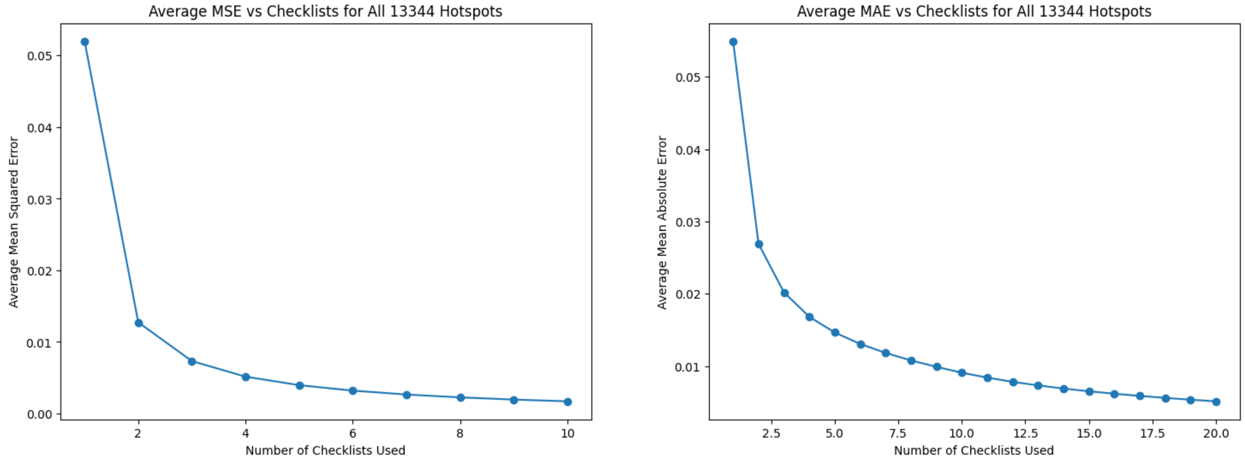
(a) Sampling from the Posterior for a single hotspot.

(b) Sampling from the Posterior for a single hotspot after 5 iterations

Figure 6.4: Sampling from the Posterior

6.4.2 Multiple Hotspots scenario

We extend the analysis across all the hotspots in our test set. Assessing numerous hotspots, rather than a single case, provides a robust test of the effect of additional complete checklists on the performance of our model. We note in 6.6 that the Average Mean squared Error (MSE) significantly diminishes for each additional checklist added for all the hotspots.



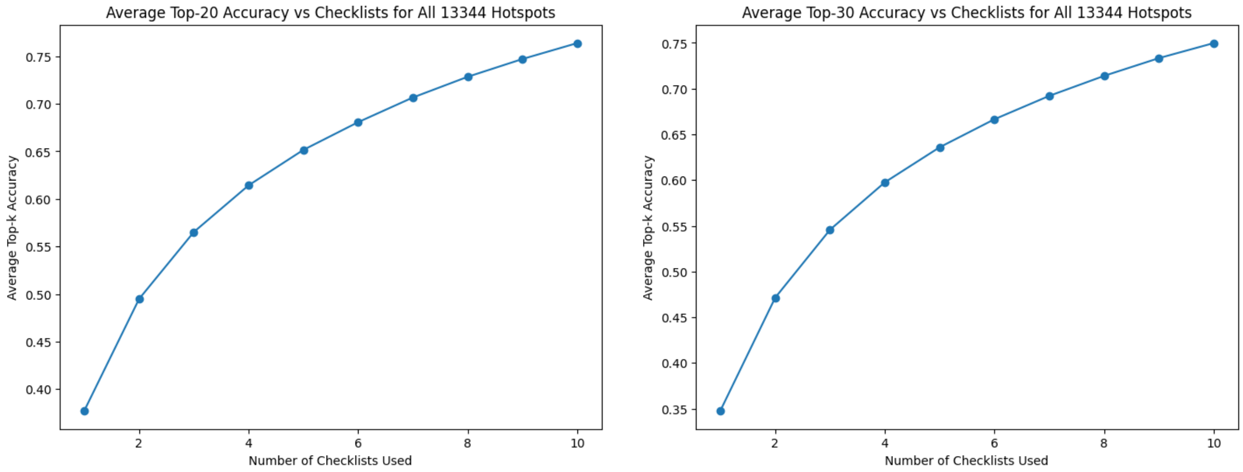
(a) MSE vs number of complete checklists for all hotspots

(b) MAE vs number of complete checklists for all hotspots

Figure 6.5: Comparison of Mean Squared Error (MSE) and Mean Absolute Error (MAE) versus the number of checklists (nchklists) for all hotspots. The plots demonstrate significant model improvement with increased checklists. (**Left:**) The MSE plot reveals a substantial decrease in error with the addition of each subsequent checklist, evident up to 10 additional checklists. (**Right:**) A similar trend is observed for MAE, confirming the model's enhanced accuracy with more checklists.

Top-K Errors

Next, we explore the effect of an additional checklist on the top 10, 20 and 30 accuracy



(a) Top 20 Accuracy vs number of complete checklists for all hotspots

(b) Top 30 Accuracy vs number of complete checklists for all hotspots

Figure 6.6: Examination of Top-K Error. (**Right:**) The plot illustrates the consistent improvement in top-20 accuracy with each additional checklist, advancing from 40% with one checklist to a peak of 75% with ten additional checklists. A similar progressive enhancement is observed for top-30 accuracy, reaffirming the positive impact of additional checklists on model performance.

7

Conclusion

In this work, we present SatBird, a novel large-scale dataset for species distribution modeling that integrates remote sensing and citizen science data. SatBird comprises seasonal datasets for summer and winter in the USA, as well as Kenya. We outline the methodology for constructing SatBird and demonstrate its utility by training different models to predict species encounter rates directly from satellite imagery and bioclimatic variables. Our benchmark results validate the feasibility of this approach and highlight the potential of SatBird to advance ecological deep learning. By releasing this dataset, we aim to provide a valuable resource that spurs new innovations in biodiversity monitoring and conservation. The creation of SatBird also offers a template for assembling multimodal species occurrence data that could be extended to other organisms and geographies. Two

limitations of the present work include using a single satellite image to represent a season, while the landscape can change over time, and that satellite images are extracted from one recent year, while eBird data is aggregated across multiple recent years. We are planning a next version of SatBird that will include multiple satellite images for each hotspot to account for changes over time. We also aim to expand SatBird using citizen science data for other organisms, not merely birds (this is more challenging since most such datasets are presence-only).

As SatBird is intended to directly impact biodiversity monitoring, we look forward to integrating the approaches we have introduced into existing tools for ecology and policy. We invite ecologists and researchers in AI for climate change and biodiversity to use and build upon our dataset and experiments on SatBird-Kenya and SatBird-USA-winter by transferring pretrained models from SatBird-USA-summer to test for the generalization of models on different seasons and locations. One immediate application for models trained on SatBird is eBird's existing tool that lists the "likely species" in a given area, to be available for poorly monitored locations. This tool currently relies on encounter rates from past checklists and is therefore only available in well-monitored locations. Models trained on SatBird could estimate such encounter rates for poorly monitored locations via remote sensing. We hope such input will be valuable to researchers seeking to understand biodiversity and climate change, as well as policymakers interested in evaluating conservation priorities across different areas of land.

7.1 Applications

This work enables several important biodiversity monitoring applications, although it also has some limitations on potential use cases.

- **Species range shift detection** - Models can identify changes in species distributions over time, critical for tracking range shifts due to climate change or habitat loss. Comparing predicted distribution maps over multiple years could reveal range expansions, contractions, or shifts.

- **Biodiversity hotspot mapping** - Predicted species richness maps can highlight areas of high biodiversity that should be priority targets for conservation. Hotspots with many overlapping species ranges can be identified.

- **Ecological forecasting** - Models can make projections of future species distributions under different climate change scenarios. These forecasts guide mitigation strategies by revealing which species are likely to gain or lose suitable habitat.

- **Revealing data gaps** - Model uncertainty maps can identify regions that require more sampling effort to improve predictions. Citizen science campaigns could then target those high uncertainty areas for data collection.

- **Optimizing wildlife surveys** - By predicting habitat suitability and species occurrence probabilities, surveys can be made more efficient by focusing sampling effort on high probability sites.

- **Informing policy** - Conservation decisions and land use planning can leverage model predictions to better protect biodiversity and identify regions needing protection.

- **Improving citizen science** - Model predictions can help guide citizen scientists on where and when to look for certain species and provide educational resources to improve species identification abilities.

7.2 Future work

More research is needed to develop Bayesian neural networks that can directly predict both the mean and variance of species distributions directly. As discussed in Section 5.3, our current approach estimates variance $v_{h,s}$ from the predicted mean $p_{h,s}$. However, enabling models to independently learn the variance could improve performance.

Additional modalities beyond those explored here may also enhance species distribution predictions for birds and other taxa. Integrating data like land cover(urban areas, waterbodies etc), tree canopy metrics, bioacoustics, and butterfly occurrences could provide useful ecological context alongside remote sensing and climate variables. Deep learning methods are well-suited to fuse such multivariate data.

It would also be interesting to see how varying the input patch size described in 4.3 of $64px$ depending on the density of observations in that location affects the performance of the model.

This work establishes a foundation for further advances in ecological deep learning and multimodal species distribution modeling. There are tremendous opportunities to refine model architectures, incorporate new data streams, and tailor methods to different organisms and settings. We hope our study catalyzes future efforts to leverage deep learning for biodiversity monitoring and conservation.

We provide the data, code, and other resources needed to reproduce our experiments in Appendix A.

A

Appendix A: Data and Code

A.1 Dataset

The dataset presented in this work is available for download at the following link:

Dataset Download

A.2 Code

Additionally, you can find the companion code for the dataset preparation pipeline and benchmark here:

Companion Code Download

B

Algorithms

B.1 Algorithm 1: Predicting Species Encounter Rate

In this algorithm, species encounter probabilities for each hotspot are predicted using a Bayesian approach. The algorithm loops through each hotspot, collecting the initial predicted probabilities and observed encounter checklists for all species present. Hotspots with insufficient checklists are skipped. For each species, the sample variance is computed from the checklists. This variance is used along with the initial prediction to determine the parameters α and β of a Beta prior distribution.

The algorithm then samples multiple sets of 2, 5, and 10 checklists. For each checklist sample, the 0s and 1s are added to the prior α and β to compute the posterior parameters. The posterior mean and variance are then calculated. As the ground truth, the mean and variance are also calculated from all available checklists.

Algorithm 1 Predicting Species Encounter Probabilities

```
for each hotspot do
    Collect initial predictions and all checklists for species in the hotspot
    if there are less than 5 checklists or no checklists then
        Continue to the next hotspot
    end if
    for each species do
        if all checklist values are 0 then
            Skip this species
        end if
        Compute sample variance $v$ from all checklists
        Compute prior parameters alpha and beta based on initial prediction $p$ and sample variance $v$
        for each of 2, 5, and 10 checklists do
            Select checklists
            Compute posterior alpha and beta by adding 0s and 1s from the checklists to prior alpha and beta
            if division by 0 error then
                Handle the error
            end if
            Compute posterior mean $p_{post}$ and variance $v_{post}$
            Compute ground truth mean $p_{true}$ and variance $v_{true}$ from ALL checklists
            Compute accuracy as $1 - |p_{true} - p_{post}|$
            Compute average accuracy for the hotspot
        end for
    end for
end for
```

The accuracy of the posterior mean compared to the ground truth is computed for each sample. Finally, the accuracy is averaged across samples and species to quantify the overall performance on that hotspot. This provides a data-driven framework for updating the initially predicted probabilities to posterior probabilities that reflect the observed encounters, while properly quantifying the uncertainty. The performance on standardized checklist samples evaluates the real-world applicability of the probabilistic species distribution modelling.

B.2 Predicting Species Encounter Probabilities with Hyperparameter

This algorithm extends the previous one by introducing a hyperparameter that controls the variance calculation. The key difference is that instead of computing the variance directly from the checklists, the variance is calculated as a function of the initial prediction and the hyperparameter.

Specifically, for each species, a hyperparameter value is set or optimized. The variance is then calculated as $v = hyperparameter * p * (1 - p)$, where p is the initial prediction. This links the variance to the prediction through the hyperparameter.

The calculated variance is then used along with the initial prediction to determine the prior distribution parameters. The rest of the algorithm follows the same Bayesian updating approach as before - sampling checklists, computing posteriors, determining accuracy compared to ground truth, and averaging across species and samples.

The addition of the hyperparameter provides more control over the variance and uncertainty quantification of the probabilistic model. It can be tuned as a form of regularization to improve generalizability. The standardized checklist sampling evaluates how robust the predictions are to limited observations, for different hyperparameter values.

B.2.1 More Detailed extension of Algortithm 1

Algorithm 2 Predicting Species Encounter Probabilities with Hyperparameter

```
for each hotspot do
    Collect initial predictions and all checklists for species in the hotspot
    if there are less than 5 checklists or no checklists then
        Continue to the next hotspot
    end if
    for each species do
        if all checklist values are 0 then
            Skip this species
        end if
        Set or optimize the hyperparameter value
        Calculate variance $v$ as $v = \mathit{hyperparameter} \times p \times (1 - p)$
        Compute prior parameters alpha and beta based on initial prediction $p$ and calculated variance $v$
        for each of 2, 5, and 10 checklists do
            Select checklists
            Compute posterior alpha and beta by adding 0s and 1s from the checklists to prior alpha and beta
            if division by 0 error then
                Handle the error
            end if
            Compute posterior mean $p_{post}$ and variance $v_{post}$
            Compute ground truth mean $p_{true}$ and variance $v_{true}$ from ALL checklists
            Compute accuracy as $1 - |p_{true} - p_{post}|$
            Compute average accuracy for the hotspot
        end for
    end for
end for
```

Algorithm 3 Predicting Species Encounter Probabilities with Hyperparameter

```
for each hotspot $h$ do
    Collect predictions $P_h$ and all checklists $C_h$ for species in $h$
    if $|C_h| < 5$ then
        Continue to the next hotspot
    end if
    for each species $s$ do
        if $\max(C_{h,s}) = 0$ then
            Skip this species
        end if
        Set or optimize hyperparameter $\tau$
        Calculate $v_{h,s} = \tau \times p_{h,s} \times (1 - p_{h,s})$
        Compute $\alpha_{h,s}$ and $\beta_{h,s}$ using $p_{h,s}$ and $v_{h,s}$
        for each $n \in \{2, 5, 10\}$ do
            Select $n$ checklists $C'_{h,s} \subseteq C_{h,s}$
            Compute $\alpha'_{h,s} = \alpha_{h,s} + \sum_{c \in C'_{h,s}} c$
            Compute $\beta'_{h,s} = \beta_{h,s} + n - \sum_{c \in C'_{h,s}} c$
            Compute $p'_{h,s} = \frac{\alpha'_{h,s}}{\alpha'_{h,s} + \beta'_{h,s}}$
            Compute $v'_{h,s} = \frac{\alpha'_{h,s}\beta'_{h,s}}{(\alpha'_{h,s} + \beta'_{h,s})^2(\alpha'_{h,s} + \beta'_{h,s} + 1)}$
            Compute $p_{h,s,true} = \frac{\sum_{c \in C_{h,s}} c}{|C_{h,s}|}$
            Compute accuracy $a_{h,s} = 1 - |p_{h,s,true} - p'_{h,s}|$
        end for
        Compute average accuracy $a_h = \frac{\sum_s a_{h,s}}{|S_h|}$
    end for
end for
```

C

Appendix C: Hotspot Visualizations

C.0.1 Species distribution per hotspot

A well-known issue in species distribution modeling when looking at many species at a time is the zero-inflated nature of the targets, which also appears in our dataset. Indeed, all 684 considered species in the USA and 1054 species in Kenya are never found in the same place together. Figure C.1 shows the distribution of the number of species with non-zero encounter rates per hotspot in the SatBird-US-summer and the SatBird-Kenya datasets.

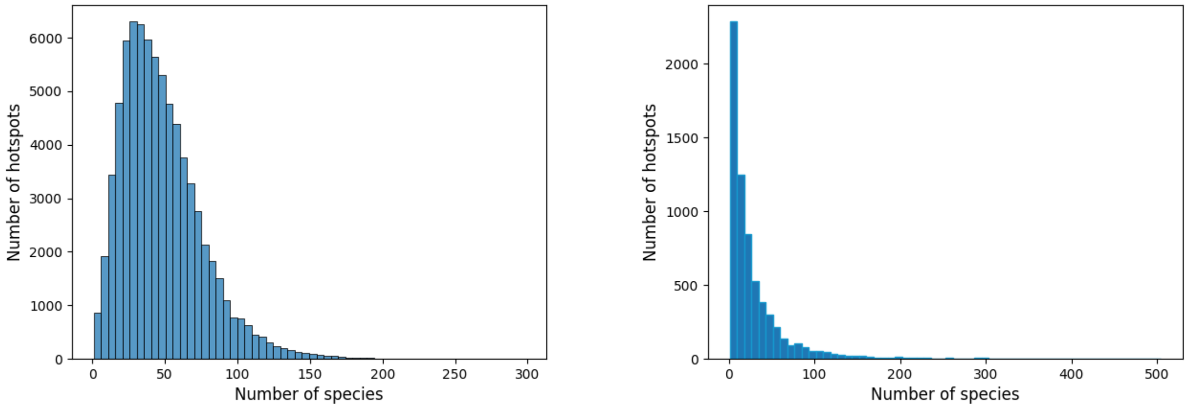
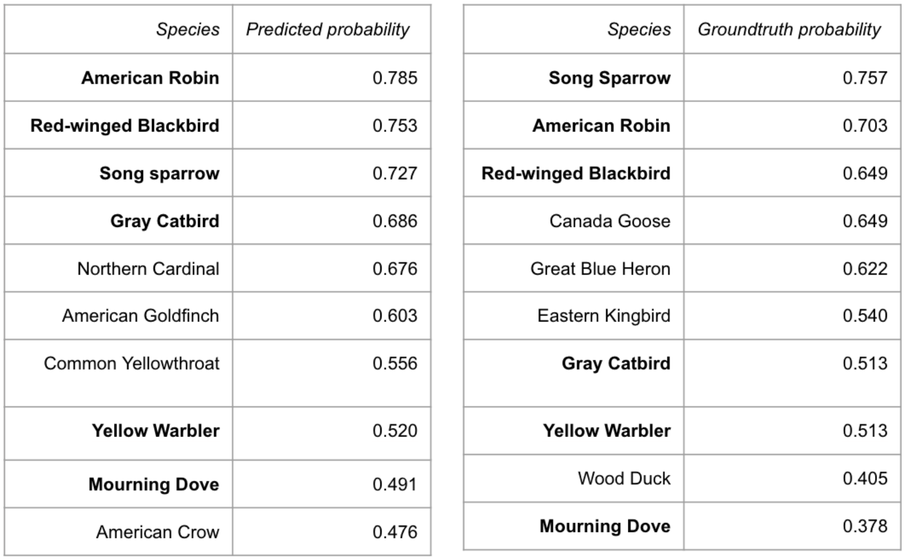


Figure C.1: Distribution of the number of species encountered per hotspot for the USA-summer (left) and Kenya (right) training sets. On average a hotspot in these datasets has resp. 48 and 30 different species were reported.

C.0.2 Predictions vs Groundtruths

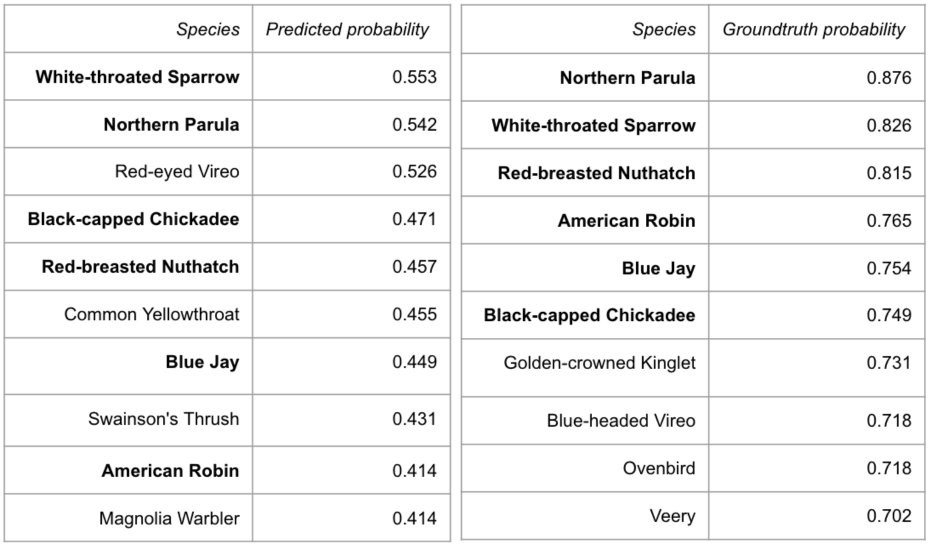
We present the predictions from the model versus what the groundtruth data contained. Essentially, we can see that the model learned some of the species present in the groundtruth dataset as intended, showing the potential of the models to learn the underlying rangemaps. We visualize three regions: Figure C.2 shows Lampson Reservoir in Ohio; Figure C.3 shows a hotspot in California; and Figure C.4 shows species in a hotspot in San Bernardino County, California.



Species	Predicted probability
American Robin	0.785
Red-winged Blackbird	0.753
Song sparrow	0.727
Gray Catbird	0.686
Northern Cardinal	0.676
American Goldfinch	0.603
Common Yellowthroat	0.556
Yellow Warbler	0.520
Mourning Dove	0.491
American Crow	0.476

Species	Groundtruth probability
Song Sparrow	0.757
American Robin	0.703
Red-winged Blackbird	0.649
Canada Goose	0.649
Great Blue Heron	0.622
Eastern Kingbird	0.540
Gray Catbird	0.513
Yellow Warbler	0.513
Wood Duck	0.405
Mourning Dove	0.378

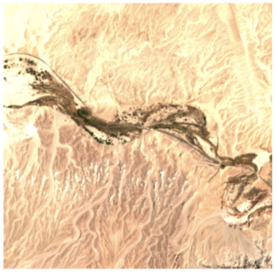
Figure C.2: Hotspot and species in Lampson Reservoir,Ashtabula County, Ohio: Our model predicts the presence of various species in this region, including the song sparrow, red-winged blackbird and gray catbird. These species are consistent with the ground truth and are highly reported on eBird for this location.

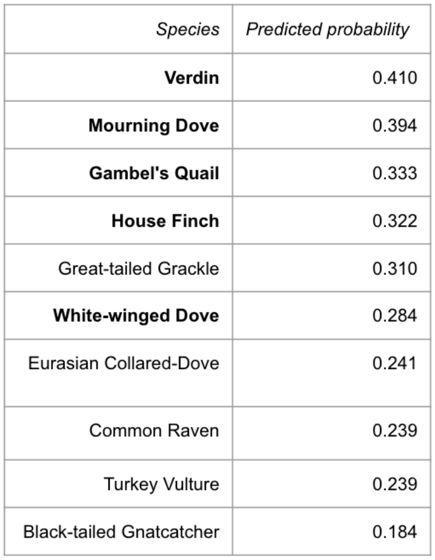


Species	Predicted probability
White-throated Sparrow	0.553
Northern Parula	0.542
Red-eyed Vireo	0.526
Black-capped Chickadee	0.471
Red-breasted Nuthatch	0.457
Common Yellowthroat	0.455
Blue Jay	0.449
Swainson's Thrush	0.431
American Robin	0.414
Magnolia Warbler	0.414

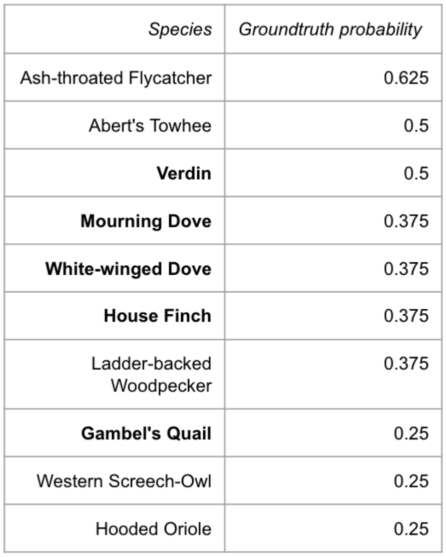
Species	Groundtruth probability
Northern Parula	0.876
White-throated Sparrow	0.826
Red-breasted Nuthatch	0.815
American Robin	0.765
Blue Jay	0.754
Black-capped Chickadee	0.749
Golden-crowned Kinglet	0.731
Blue-headed Vireo	0.718
Ovenbird	0.718
Veery	0.702

Figure C.3: comparison of predictions in California versus the groundtruths for top 10 species, we see that the white throated sparrow, Northern Parula, Black-capped chickadee, Blue Jay and American Robin were correctly predicted by the model, making it six out of 10 correct predictions.





Species	*Predicted probability*
Verdin	0.410
Mourning Dove	0.394
Gambel's Quail	0.333
House Finch	0.322
Great-tailed Grackle	0.310
White-winged Dove	0.284
Eurasian Collared-Dove	0.241
Common Raven	0.239
Turkey Vulture	0.239
Black-tailed Gnatcatcher	0.184



Species	*Groundtruth probability*
Ash-throated Flycatcher	0.625
Abert's Towhee	0.5
Verdin	0.5
Mourning Dove	0.375
White-winged Dove	0.375
House Finch	0.375
Ladder-backed Woodpecker	0.375
Gambel's Quail	0.25
Western Screech-Owl	0.25
Hooded Oriole	0.25

Figure C.4: Hotspot and species at Afton Canyon, San Bernardino County in California: Our model correctly identified the Verdin in this region in the top-10 species predicted list. It's worth noting that the Verdin's habitat is a fairly rare species whose habitat is mainly restricted to shrublands.

Bibliography

[1] Census Bureau of USA. https://www.census.gov/geographies/mapping-files/time-series/geo/carto-boundary-file.html. Accessed: 2023-06-06.

[2] Maxent. https://support.bccvl.org.au/support/solutions/articles/6000083216-maxent. Accessed: 2023-08-21.

[3] Maxent. https://biodiversityinformatics.amnh.org/open_source/maxent/. Accessed: 2023-08-21.

[4] Global carbon project. https://www.globalcarbonproject.org/, 2023. Accessed: Wednesday 6th December, 2023.

[5] Noaa national centers for environmental information (ncei) world ocean database. https://www.ncei.noaa.gov/products/world-ocean-database, Year. Accessed: Wednesday 6th December, 2023.

[6] ABA. *American Birding Association checklist: Birds of the continental United States and Canada*. 2022.

[7] C. Abrahams and M. Geary. Combining bioacoustics and occupancy modelling for improved monitoring of rare breeding bird populations. *Ecological Indicators*, 112:106131, 2020.

[8] E. A. Alshari, M. B. Abdulkareem, and B. W. Gawali. Classification of land use/land cover using artificial intelligence (ann-rf). *Frontiers in Artificial Intelligence*, 5, 2023.

[9] American Museum of Natural History. What is biodiversity? why is it important?, 2021.

[10] M. P. Austin. Species distribution models and ecological theory: A critical assessment and some possible new approaches. *Ecological Modelling*, 200:1–19, 2007.

[11] T. K. G. C. author, F. Huettmann, and M. Ehlers. Review article: Thirty years of analysing and modelling avian habitat relationships using satellite imagery data: a review. *International Journal of Remote Sensing*, 26(12):2631–2656, 2005.

[12] avibase. Wavibase - the world bird database, 2023.

[13] T. Baker, J. Kiptala, L. Olaka, N. Oates, A. Hussain, and M. McCartney. Baseline review and ecosystem services assessment of the tana river basin, kenya. Technical report, International Water Management Institute (IWMI), 2015.

[14] D. Baldocchi, E. Falge, L. Gu, R. Olson, D. Hollinger, S. Running, P. Anthoni, C. Bernhofer, K. Davis, R. Evans, et al. Fluxnet: A new tool to study the temporal and spatial variability of ecosystem-scale carbon dioxide, water vapor, and energy flux densities. *Bulletin of the American Meteorological Society*, 82(11):2415–2434, 2001.

[15] S. Baraka, B. Akera, B. Aryal, T. Sherpa, F. Shresta, A. Ortiz, K. Sankaran, J. L. Ferres, M. Matin, and Y. Bengio. Machine learning for glacier monitoring in the hindu kush himalaya. *arXiv preprint arXiv:2012.05013*, 2020.

[16] F. Bastani, P. Wolters, R. Gupta, J. Ferdinando, and A. Kembhavi. Satlas: A large-scale, multi-task dataset for remote sensing image understanding. 11 2022.

[17] S. Beery, E. Cole, J. Parker, P. Perona, and K. Winner. Species distribution modeling for machine learning practitioners: A review. In *ACM SIGCAS conference on computing and sustainable societies*, pages 329–348, 2021.

[18] D. A. Benson, M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers. Genbank. *Nucleic acids research*, 41(D1):D36–D42, 2012.

[19] J. Blitzstein and J. Hwang. *Probability! An Interactive Introduction*. bookdown.org, 2023. Accessed: 2023.

[20] E. H. Boakes, G. Gliozzo, V. Seymour, M. Harvey, C. Smith, D. B. Roy, and M. Haklay. Patterns of contribution to citizen science biodiversity projects increase understanding of volunteers' recording behaviour. *Scientific reports*, 6(1):1–11, 2016.

[21] C. M. Bourne, P. M. Regular, B. Sun, S. P. Thompson, A. J. Trant, and J. A. Wheeler. Generalized linear model analysis.

[22] C. S. Bretherton, B. Henn, A. Kwa, N. D. Brenowitz, O. Watt-Meyer, J. McGibbon, W. A. Perkins, S. K. Clark, and L. Harris. Correcting coarse-grid weather and climate models by machine learning from global storm-resolving simulations. *Journal of Advances in Modeling Earth Systems*, 14(2):e2021MS002794, 2022.

[23] J. L. Brown and A. D. Yoder. Shifting ranges and conservation challenges for lemurs in the face of climate change. *Ecol Evol*, 5:1131–1142, 2015.

[24] K. A. Brown et al. Predicting plant diversity patterns in madagascar: understanding the effects of climate and land cover change in a biodiversity hotspot. *PLoS one*, 10(4):e0122721, 2015.

[25] C. T. Callaghan, A. G. Poore, M. Hofmann, C. J. Roberts, and H. M. Pereira. Large-bodied birds are over-represented in unstructured citizen science data. *Scientific reports*, 11(1):1–11, 2021.

[26] D. E. Capen. The use of multivariate statistics in studies of wildlife habitat. 1981.

[27] N. D. Charney, S. Record, B. E. Gerstner, C. Merow, P. L. Zarnetske, and B. J. Enquist. A test of species distribution model transferability across environmental and geographic space for 108 western north american tree species. In *Frontiers in Ecology and Evolution*, 2021.

[28] Y. Chen, Z. Jiang, P. Fan, P. Ericson, G. Song, X. Luo, F. Lei, and Y. Qu. The combination of genomic offset and niche modelling provides insights into climate change-driven vulnerability. *Nature Communications*, 13, 08 2022.

[29] E. Cole, B. Deneu, T. Lorieul, M. Servajean, C. Botella, D. Morris, N. Jojic, P. Bonnet, and A. Joly. The GeoLifeCLEF 2020 dataset. *Preprint arXiv:2004.04192*, 2020.

[30] Y. Cong, S. Khanna, C. Meng, P. Liu, E. Rozi, Y. He, M. Burke, D. B. Lobell, and S. Ermon. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery, 2023.

[31] C. B. Cooper, W. M. Hochachka, A. A. Dhondt, R. Louv, and J. W. Fitzpatrick. *The Opportunities and Challenges of Citizen Science as a Tool for Ecological Research*, pages 99–113. Cornell University Press, 1 edition, 2012.

[32] B. Deneu, M. Servajean, P. Bonnet, C. Botella, F. Munoz, and A. Joly. Convolutional neural networks improve species distribution modelling by capturing the spatial structure of the environment. *PLoS Computational Biology*, 17, 2021.

[33] M. Dornelas, L. H. Antao, F. Moyes, A. E. Bates, A. E. Magurran, D. Adam, A. A. Akhmetzhanova, W. Appeltans, J. M. Arcos, H. Arnold, et al. Biotime: A database of biodiversity time series for the anthropocene. *Global Ecology and Biogeography*, 27(7):760–786, 2018.

[34] L. Dutra Silva, E. Brito de Azevedo, F. Vieira Reis, R. Bento Elias, and L. Silva. Limitations of species distribution models based on available climate change data: a case study in the azorean forest. *Forests*, 10(7):575, 2019.

[35] J. Elith, C. Graham, R. Valavi, M. Abegg, C. Bruce, S. Ferrier, A. Ford, A. Guisan, R. J. Hijmans, F. Huettmann, et al. Presence-only and presence-absence data for comparing species distribution modeling methods. *Biodiversity informatics*, 15(2):69–80, 2020.

[36] J. Elith, J. R. Leathwick, and T. Hastie. A working guide to boosted regression trees. *Journal of animal ecology*, 77(4):802–813, 2008.

[37] ESA. European space agency, Year. Accessed: Wednesday 6th December, 2023.

[38] J. Estopinan, M. Servajean, P. Bonnet, F. Munoz, and A. Joly. Deep species distribution modeling from sentinel-2 image time-series: A global scale analysis on the orchid family. *Frontiers in Plant Science*, 13, 2022.

[39] L. S. Farwell, P. R. Elsen, E. Razenkova, A. M. Pidgeon, and V. C. Radeloff. Habitat heterogeneity captured by 30-m resolution satellite image texture predicts bird richness across the united states. *Ecological Applications*, 30(8):e02157, 2020.

[40] M. J. Feldman, L. Imbeau, P. Marchand, M. J. Mazerolle, M. Darveau, and N. J. Fenton. Trends and gaps in the use of citizen science derived data as input for species distribution models: A quantitative review. *PloS one*, 16(3):e0234587, 2021.

[41] D. Fink, T. Auer, A. Johnston, M. Strimas-Mackey, S. Ligocki, O. Robinson, W. Hochachka, L. Jaromczyk, A. Rodewald, C. Wood, I. Davies, and A. Spencer. ebird status and trends, data version: 2021, cornell lab of ornithology, ithaca, new york, 2022.

[42] M. R. Fisher. *Environmental Biology*. Open Oregon Educational Resources, 2017.

[43] Y. Fourcade, J. O. Engler, D. Rödder, and J. Secondi. Mapping species distributions with maxent using a geographically biased sample of presence data: a performance assessment of methods for correcting sampling bias. *PloS one*, 9(5):e97122, 2014.

[44] K.-H. Frommolt, K.-H. Tauchert, M. Koch, et al. Advantages and disadvantages of acoustic monitoring of birds–realistic scenarios for automated bioacoustic monitoring in a densely populated region. In *Computational Bioacoustics for Assessing Biodiversity. Proc. of the Internat. Expert Meeting on IT-based Detection of Bioacoustical Patterns. BfN-Skripten*, volume 234, pages 83–92, 2008.

[45] GBIF. Global biodiversity information facility, 2023. Accessed: 2023-09-05.

[46] Z. Gharaee, Z. Gong, N. Pellegrino, I. Zarubiieva, J. B. Haurum, S. C. Lowe, J. T. McKeown, C. C. Ho, J. McLeod, Y.-Y. C. Wei, et al. A step towards worldwide biodiversity assessment: The bioscan-1m insect dataset. *arXiv preprint arXiv:2307.10455*, 2023.

[47] A. Ghulam, I. Porton, and K. Freeman. Detecting subcanopy invasive plant species in tropical rainforest by integrating optical and microwave (insar/polinsar) remote sensing data, and a decision tree algorithm. *ISPRS journal of photogrammetry and remote sensing*, 88:174–192, 2014.

[48] GoK. Kenya vision. 2030. a globally competitive and prosperous kenya. Technical report, Government of Kenya, Nairobi, 2007.

[49] H. Goëau, P. Bonnet, A. Joly, V. Bakić, J. Barbe, I. Yahiaoui, S. Selmi, J. Carré, D. Barthélémy, N. Boujemaa, J.-F. Molino, G. Duché, and A. Péronnet. Pl@ntnet mobile app. pages 423–424, 10 2013.

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

[51] T. A. Hallman and W. D. Robinson. Comparing multi- and single-scale species distribution and abundance models built with the boosted regression tree algorithm. *Landscape Ecology*, 35:1161–1174, 2020.

[52] M. A. Hammad, B. Jereb, B. Rosi, and D. Dragan. Methods and models for electric load forecasting: A comprehensive review. *Logistics & Sustainable Transport*, 11:51 – 76, 2020.

[53] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[54] P. Helber, B. Bischke, A. Dengel, and D. Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

[55] M. Helvey, M. Ryckman, S. Ellis-Felege, J. Van Aardt, and C. Salvagio. Duck nest detection through remote sensing. In *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*, pages 6321–6324, 2020.

[56] T. Hengl, J. Mendes de Jesus, G. B. Heuvelink, M. Ruiperez Gonzalez, M. Kilibarda, A. Blagotić, W. Shangguan, M. N. Wright, X. Geng, B. Bauer-Marschallinger, et al. Soilgrids250m: Global gridded soil information based on machine learning. *PLoS one*, 12(2):e0169748, 2017.

[57] B. Huntley and P. Barnard. Potential impacts of climatic change on southern african birds of fynbos and grassland biodiversity hotspots. *Diversity and Distributions*, 18:769–781, 2012.

[58] iNaturalist. inaturalist.org, 2023.

[59] J. M. Jeschke and D. L. Strayer. Usefulness of bioclimatic models for studying climate change and invasive species. *Annals of the New york Academy of Sciences*, 1134(1):1–24, 2008.

[60] M. P. Kain and B. M. Bolker. Predicting west nile virus transmission in north american bird communities using phylogenetic mixed effects models and ebird citizen science data. *Parasites & vectors*, 12:1–22, 2019.

[61] K. S. Kaswan and J. S. Dhatterwal. The use of machine learning for sustainable and resilient buildings. *Digital Cities Roadmap: IoT-Based Architecture and Sustainable Buildings*, pages 1–62, 2021.

[62] J. Kattge, G. Bönisch, S. Díaz, S. Lavorel, I. C. Prentice, P. Leadley, S. Tautenhahn, G. D. Werner, T. Aakala, M. Abedi, et al. Try plant trait database–enhanced coverage and open access. *Global change biology*, 26(1):119–188, 2020.

[63] S. Kelling, J. Gerbracht, D. Fink, C. Lagoze, W.-K. Wong, J. Yu, T. Damoulas, and C. Gomes. eBird: A human/computer learning network for biodiversity conservation and research. *AI Magazine*, 34, 03 2013.

[64] H. Kerner, G. Tseng, I. Becker-Reshef, C. Nakalembe, B. Barker, B. Munshell, M. Paliyam, and M. Hosseini. Rapid response crop maps in data sparse regions. *arXiv preprint arXiv:2006.16866*, 2020.

[65] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[66] A. LEE and P. BARNARD. Endemic birds of the fynbos biome: a conservation assessment and impacts of climate change. *Bird Conservation International*, 26:52–68, 2016.

[67] R. Leinonen, R. Akhtar, E. Birney, L. Bower, A. Cerdeno-Tárraga, Y. Cheng, I. Cleland, N. Faruque, N. Goodgame, R. Gibson, et al. The european nucleotide archive. *Nucleic acids research*, 39(suppl_1):D28–D31, 2010.

[68] R. H. MacArthur. Population ecology of some warblers of northeastern coniferous forests. *Ecology*, 39(4):599–619, 1958.

[69] J. S. Madin, K. D. Anderson, M. H. Andreasen, T. C. Bridge, S. D. Cairns, S. R. Connolly, E. S. Darling, M. Diaz, D. S. Falster, E. C. Franklin, et al. The coral trait database, a curated database of trait information for coral species from the global oceans. *Scientific Data*, 3(1):1–22, 2016.

[70] O. Mañas, A. Lacoste, X. Giró-i-Nieto, D. Vázquez, and P. Rodríguez. Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data. *CoRR*, abs/2103.16607, 2021.

[71] P. McCullagh. Introduction to nelder and wedderburn (1972) generalized linear models. 1992.

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

[73] W. Mupangwa, L. Chipindu, I. Nyagumbo, S. Mkuhlani, and G. Sisito. Evaluating machine learning algorithms for predicting maize yield under conservation agriculture in eastern and southern africa. *SN Applied Sciences*, 2:1–14, 2020.

[74] NASA. National aeronautics and space administration, Year. Accessed: Wednesday 6th December, 2023.

[75] C. S. Parr, K. S. Schulz, J. Hammock, N. Wilson, P. Leary, J. Rice, and R. J. Corrigan Jr. Traitbank: Practical semantics for organism attribute data. *Semantic Web*, 7(6):577–588, 2016.

[76] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. 2017.

[77] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[78] H. M. Pereira, S. Ferrier, M. Walters, G. N. Geller, R. H. Jongman, R. J. Scholes, M. W. Bruford, N. Brummitt, S. H. Butchart, A. Cardoso, et al. Essential biodiversity variables. *Science*, 339(6117):277–278, 2013.

[79] S. J. Phillips, R. P. Anderson, and R. E. Schapire. Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, 190:231–259, 2006.

[80] B. Pienaar, D. Thompson, B. Erasmus, T. Hill, and E. Witkowski. Evidence for climate-induced range shift in brachystegia (miombo) woodland. *South African Journal of Science*, 111:1–9, 2015.

[81] A. R. Portabales, M. L. Nores, and J. J. Pazos-Arias. Systematic review of electricity demand forecast using ann-based machine learning algorithms. *Sensors (Basel, Switzerland)*, 21, 2021.

[82] K. L. Prudic, K. P. McFarland, J. C. Oliver, R. A. Hutchinson, E. C. Long, J. T. Kerr, and M. Larrivée. ebutterfly: Leveraging massive online citizen science for butterfly conservation. *Insects*, 8, 2017.

[83] N. Rahaman, M. Weiss, F. Träuble, F. Locatello, A. Lacoste, Y. Bengio, C. Pal, L. E. Li, and B. Schölkopf. A general purpose neural architecture for geospatial systems, 2022.

[84] S. Ratnasingham and P. D. Hebert. Bold: The barcode of life data system (http://www. barcodinglife. org). *Molecular ecology notes*, 7(3):355–364, 2007.

[85] C. J. Reed, R. Gupta, S. Li, S. Brockman, C. Funk, B. Clipp, K. Keutzer, S. Candido, M. Uyttendaele, and T. Darrell. Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning, 2023.

[86] E. Rolf, J. Proctor, T. Carleton, I. Bolliger, V. Shankar, M. Ishihara, B. Recht, and S. Hsiang. A generalizable and accessible approach to machine learning with global satellite imagery. *Nature communications*, 12(1):1–11, 2021.

[87] Royal Society. Why is biodiversity important?, 2021.

[88] S. Shamekh, K. D. Lamb, Y. Huang, and P. Gentine. Implicit learning of convective organization explains precipitation stochasticity. *Proceedings of the National Academy of Sciences*, 120(20):e2216158120, 2023.

[89] A. Swanson, M. Kosmala, C. Lintott, R. Simpson, A. Smith, and C. Packer. Snapshot serengeti, high-frequency annotated camera trap images of 40 mammalian species in an african savanna. *Scientific data*, 2(1):1–14, 2015.

[90] H. Tamiru and M. Wagari. Machine-learning and hec-ras integrated models for flood inundation mapping in baro river basin, ethiopia. *Modeling Earth Systems and Environment*, 8(2):2291–2303, 2022.

[91] S. Theodoridis, C. Rahbek, and D. Nogues-Bravo. Exposure of mammal genetic diversity to mid-21st century global change. *Ecography*, 44(6):817–831, 2021.

[92] P. F. Thomsen and E. Willerslev. Environmental dna–an emerging tool in conservation for monitoring past and present biodiversity. *Biological conservation*, 183:4–18, 2015.

[93] R. Valavi, G. Guillera-Arroita, J. J. Lahoz-Monfort, and J. Elith. Predictive performance of presence-only species distribution models: a benchmark study with reproducible code. *Ecological Monographs*, 2021.

[94] G. Vianna, M. Meekan, T. Bornovski, and J. Meeuwig. Acoustic telemetry validates a citizen science approach for monitoring sharks on coral reefs. *PloS one*, 9:e95565, 04 2014.

[95] C. wa MAINA. Bioacoustic approaches to biodiversity monitoring and conservation in kenya. In *2015 IST-Africa Conference*, pages 1–8. IEEE, 2015.

[96] J. Walker and P. Taylor. Using ebird data to model population change of migratory bird species. *Avian Conservation and Ecology*, 12, 06 2017.

[97] B. Weinstein, S. Marconi, S. Bohlman, A. Zare, A. Singh, S. Graves, and E. White. A remote sensing derived data set of 100 million individual tree crowns for the national ecological observatory network. *eLife*, 10, 02 2021.

[98] J. Wu, D. Pichler, D. Marley, D. Wilson, N. Hovakimyan, and J. Hobbs. Extended agriculture-vision: An extension of a large aerial image dataset for agricultural pattern analysis. *arXiv preprint arXiv:2303.02460*, 2023.

[99] Z. Wu, C. Zhang, X. Gu, I. Duporge, L. F. Hughey, J. A. Stabach, A. K. Skidmore, J. G. C. Hopcraft, S. J. Lee, P. M. Atkinson, et al. Deep learning enables satellite-based monitoring of large populations of terrestrial mammals across heterogeneous landscape. *Nature communications*, 14(1):3072, 2023.

[100] Y. Xue and A. Skidmore. Automatic counting of large mammals from very high resolution panchromatic satellite imagery. *Remote Sensing*, 9, 08 2017.

[101] R. yan Duan, X.-Q. Kong, M. yi Huang, W. Fan, and Z. Wang. The predictive performance and stability of six species distribution models. *PLoS ONE*, 9, 2014.

[102] A. Young, D. Guo, P. Desmet, and G. Midgley. Biodiversity and climate change: Risks to dwarf succulents in southern africa. *Journal of Arid Environments*, 129:16–24, 2016.

[103] H. Yu, A. R. Cooper, and D. M. Infante. Improving species distribution model predictive accuracy using species abundance: Application with boosted regression trees. *Ecological Modelling*, 432:109202, 2020.

[104] F. Zantalis, G. E. Koulouras, S. Karabetsos, and D. Kandris. A review of machine learning and iot in smart transportation. *Future Internet*, 11:94, 2019.