Investigating the Role of Negative Data Selection in Deep Species Distribution Models

Diego Zuniga

Master of Science
Data Science
School of Informatics
University of Edinburgh
2025

Abstract

This dissertation examines how the selection of negative training data influences deep learning methods for species distribution modelling (SDM). Using a Spatial Implicit Neural Representation (SINR) trained on global presence-only observations from iNaturalist to evaluate how different pseudo-absence strategies impact model performance. Most existing methods treat all unobserved species–location combinations as negative examples, but this work proposes informed alternatives based on spatial absence areas and proximity regions. Experiments on 2,418 species validated against IUCN expert range maps reveal that although assumed negatives establish a solid performance baseline, hybrid strategies that combine absence-informed and proximity-based pseudo-absences consistently enhance accuracy. These hybrid approaches achieve better mean average precision while reducing the tendency to overpredict species ranges. These findings highlight the importance of carefully defining negatives in presence-only modelling and contribute methodological insights for improving global-scale biodiversity mapping.

Research Ethics Approval

This project was planned in accordance with the Informatics Research Ethics policy. It did not involve any aspects that required approval from the Informatics Research Ethics committee.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Diego Zuniga*)

Acknowledgements

I would like to express my gratitude to my supervisor, Oisin Mac Aodha, for providing me with such a compelling topic and for his invaluable guidance throughout the process.

I would also like to thank all the people who love and support me. My girlfriend, for providing unwavering support from afar, and the group of crazy people I was lucky enough to meet and now proudly call friends, for supporting me in person.

To a lesser extent, I want to thank my laptop, who has been wanting to retire since the first semester but was sympathetic enough to hear my pleas and hold on just long enough to get me through this dissertation. The Scottish summer, for providing me with grey and windy days to work on my research without having to think about the beautiful day I was missing. And all the fantastic horror movies that came out this year—nothing like a tense movie to distract from dissertation worries.

But above all, I would like to thank my family, who I think were equally sad to see me go and happy for the opportunity this represents for me. They might need a translator to read the rest so the next goes in spanish: Los amo.

Table of Contents

Chapter 1

Introduction

1.1 Background

Global biodiversity decline is a pressing ecological concern, with the diversity and numbers of animal and plant species declining at an unprecedented rate, being now a mere fraction of what they used to be. [35] report that current biodiversity levels are now a mere fraction of historical levels, with human-induced modifications of land and ocean environments having led to a loss of 83% in wild mammals biomass, and more species threatened with extinction than ever before. The Secretariat of the Convention on Biological Diversity (CBD) outlined a strategic plan with 20 biodiversity targets too be achieved by 2020, some of which put particular emphasis in safeguarding ecosystems and preventing the decline of endangered species [39]. However, none of this targets have been fully met, highlighting the urgency of effective conservation strategies [31, 35]. This crisis extends beyond species loss, to shifts in population structures and distributions, requiring improved monitoring to better understand the current defaunation [10].

A detailed comprehension of the state and evolution of ecosystems is indispensable for conservation efforts, making accurate mapping of species' geographical distributions essential for planning [15]. This is particularly critical for threatened species, where precise knowledge about their spatial ranges (i.e., the locations where they can be expected to be found) is crucial for developing habitat protection and species reintroduction policies [37]. Species Distribution Modelling (SDM) addresses this need by predicting where species are expected to be present or absent. Following [2], SDMs can be characterized by three core components: species data, location encoding methods, and functions mapping spatial information to presence classification. The field has evolved from traditional statistical tools and single-species models [32, 44] to deep learning approaches that jointly model multiple species [3, 7]. These advancements enable the capture of complex relations between species and environmental features [20, 3], and even the learning of sophisticated spatial representations from raw observation coordinates alone [7].

A significant trend in SDM is the increasing utilization of presence-only data, meaning researchers have records of where species have been observed but lack documentation of locations where species are definitively absent. This shift, driven by the availability of crowdsourced datasets like iNaturalist [26], a global citizen science platform where users contribute species observations, facilitates large-scale analysis but introduces challenges (see Figure 1.1 for an example of prediction based on presence-only data). Presence-only data is inherently biased, reflecting uneven sampling effort across species and regions [13]. Additionally, without confirmed absence data (true negatives), strategies must be implemented in terms of the loss function to incorporate some sense of negatives, a critical adaptation for effective model training. Traditional SDM methods often supplement this with random assumed pseudo-absences—randomly selected locations assumed to represent areas where the species is absent—, but modern approaches, especially those using coordinate-based learning, require innovative strategies to handle these biases and define negatives effectively, setting the stage for the research presented here.

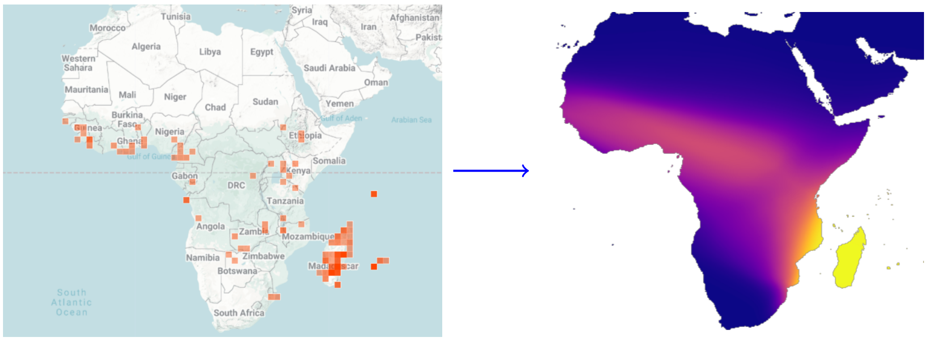


Figure 1.1: Species distribution prediction from presence-only data for the Mascarene Ridged Frog. *Left* Map of observed presences for the species in iNaturalist. *Right* Predicted range produced for the species by our implemented model.

1.2 Motivation

Research on negative label definition has predominantly relied on environmental covariates within regional studies [34] or has used simulated data [1, 4]. The work in this project advances the field by exploring negative selection within a worldwide dataset, and a model that learns implicitly geographical characteristics based solely on coordinates. This setting poses unique challenges in terms of negative selection. On the one hand, global presence-only data is highly susceptible to sampling bias, due to varying observation densities across species and locations. On the other hand, coordinate-based learning heightens the need to examine in detail strategies for modelling geographical absence.

Furthermore, our setting also provides a good lens to examine the generalization of SDM models in a context where sampling bias in species observations is strong, potentially causing predictions to be biased toward well-sampled areas and species. This is particularly important when considering that species with smaller and atypical geographic distributions are often the focus of conservation efforts [42]. Thus, this study addresses the critical need for SDM models that generalize across diverse sampling conditions, offering robust predictions for under-represented species and regions. By tackling these challenges, the research aims to enhance the applicability of machine learning in global conservation efforts.

1.3 Investigative Approach and Contributions

In terms of model specification, this study adopts the neural network architecture developed by [7], which learns directly from observed coordinates of presence-only data, in order to predict multi-species geographical distributions jointly. This approach leverages spatial implicit neural representations (SINR) to encode location data efficiently.

The primary contributions of this work are related to the negative selection strategy, and are:

1. Proposing alternatives strategies for defining absences using presence-only data, including informed negatives.

2. Assessing the impact of these strategies on coordinate-based model performance and generalization

3. Investigating factors that influence the effectiveness of negative definition approaches, such as size and singularity of species ranges
4. Introducing KL divergence as a metric to quantify the uniqueness of species distributions.
5. Developing a per-species negative selection strategy tailored to range characteristics.

The results of this study show that the way negatives are defined has a measurable impact on SDM performance. While assumed negatives remain a strong and simple baseline, informed strategies can provide complementary benefits. In particular, hybrid approaches that alternate between absence- and proximity-informed pseudo-absences outperform other methods, improving mean average precision and reducing the tendency to overpredict ranges. These improvements are most pronounced for species with limited geographic distributions, demonstrating the value of tailored negative selection for conservation-relevant taxa.

Chapter 2

Related Work

The present work intersects different domains. In this Chapter we are going to delve into the studies that are related to the present work, spanning methodological advancements and their application in the field of species distribution modelling (SDM).

2.1 Species Distribution Modelling (SDM)

Species distribution modelling (SDM) aims to predict the spatial distribution, and in some cases, the abundance of species across the world or at regional scales, playing a pivotal role in conservation planning [15, 37]. The need for accurate distribution maps has driven a constant development for new predictive approaches and methodologies. Readers interested in the subject can find more thorough description of the state of the art in SDM in the comprehensive reviews by [2] and [30]. SDM strategies typically diverge across three key components: model architecture, covariate selection, type of species data utilized, and evaluation task. Each of which will be discussed in the following paragraphs.

Early SDM approaches commonly relied on traditional statistical methods and supervised learning techniques, such as generalized linear models and random forests [32, 8, 44]. The popularisation of machine learning, coupled with the availability of large datasets, lead to new approaches in the field. While [3] were pioneers in the implementation of deep learning strategies in SDM, several studies shortly followed [29, 7] and now this methods are the state of the art, simultaneously modeling habitat suitability across multiple species. Emerging strategies, such as the use of graph neural networks (GNNs) implemented by [19], signal ongoing evolution in model architectures.

Covariates, particularly environmental and geographical variables, form the backbone of most SDM studies, linking the incidence of species to some selected covariates/features, like climate conditions, vegetation and land use [33, 44]. Extensive bioclimatic datasets, such as SoilGrids250m [21], support these efforts.

Even with the rise of deep learning, these covariates remain central, enabling the modelling of more complex relations between them and species presence [20, 3]. Notably, [20] demonstrated that these models can learn features that explain species distribution, even when they are not explicitly included as inputs. Following this line, other researchers have explored the idea of learning solely based on coordinates. [7], based on the architecture developed by [29], implement a neural network model that learns a high-dimensional spatial representation from raw coordinates of species observations, bypassing the need for environmental covariates.

Regarding species data, traditional studies often relied on presence-absence datasets, where species have defined areas of occurrence and non-occurrence. The issue is that this data is usually scarce, given the difficulty of defining absence with certainty, since the absence of an observation does not necessarily indicate that a species cannot be found in that location—it may simply reflect insufficient sampling effort. This shifted focus toward presence-only data, relying solely on observed locations. Detailed comparisons between the two approaches can be found on [16] and [12]. Initially sourced from natural history museum collections [13, 33, 46], presence data now benefits from large crowdsourced datasets, like the one provided by iNaturalist [26], facilitating large-scale SDM [7]. While this enhances species and spatial coverage, yielding state-of-the-art results, critics argue that presence-only models are prone to bias, being more susceptible to the generation of false positives and overestimating the ranges of suitable habitat. These limitations underscore the need for approaches that handle the effective use of presence-only data and for a thorough assessment of the predictions.

The use of presence-only data presents a complication in terms of model evaluation. Some papers have employed a withheld fraction of the same presence-only data for evaluation, using false negative rates and top-k classification accuracy as measures [5, 13]. The problem is that presence records are biased both in terms of species and locations [13], reason why the prevailing approach leverages independent presence-absence data to evaluate the model's performance [13, 7, 46]. The selection of evaluation metric has also received attention, [30] state that the use of accuracy or discrimination measures should be decided based on the prediction objectives.

2.2 Single-Positive Multi-Label (SPML) Learning

The use of presence-only data in species distribution modelling (SDM) aligns closely with the field of single-positive multi-label (SPML) learning. The research on SPML originated in the context of multi-label classification tasks in computer vision, which involve assigning multiple class labels to an image. While multi-label data can be used for training, this type of data requires exhaustive annotation of all present classes, which is difficult and time consuming [9]. This challenge has spurred research into partial-label learning [11], where only some classes are annotated, and the extreme case of single-positive learning, where each instance is associated with only one label [6, 45]. This setting, termed single-positive multi-label learning by [6], is particularly relevant to presence-only ecological data.

In the SPML framework, presence-only data mirrors this paradigm: each record reports the observation of a single species at a specific location, offering no information about the presence or absence of other species at this site. This limited information poses complications related to the loss function design. Traditional multi-label classification models are optimized to minimize a loss that contains both positive and negative values, comparing how accurate are the predictions in relation to the ground truth values. To address this, different approaches have been developed to train models with presence-only effectively. The most commonly used approach involves generating artificial absence points in areas where we expect the species to be absent [7], these negatives are commonly referred to as "pseudo-absences". However, this method introduces potential false negatives, prompting alternative techniques. For instance, [47] acknowledge that unseen observations might not necessarily imply absence, by applying a maximizing entropy approach that avoids enforcing pseudo-absences as zeros. Similarly, [45] define sets of expected positives and negatives to guide training, enhancing model robustness.

A critical aspect of SPML is the selection of negative labels, which significantly influences model performance. Uniformly sampled negatives might fail to provide a discriminative signal for the model to learn an appropriate representation of the space. This relates to the concept of ”hard negatives”, which are useful challenging to improve learning [14, 36]. This topics has also found its way into the domain of SDM. [1] examine where and how should pseudo-absences be generated, concluding that uniform sampled negatives are too distant to the presence data to be informative. This insight underscores the need for more sophisticated negative sampling strategies.

Furthermore, the sampling bias which is usually found in presence-only data, may play a role in the definition of negative labels. Studies such as [34, 17, 4, 44] have explored how these biases affect optimal negative selection, emphasizing the importance of tailored approaches. Being the main goal of this work to analyse the definition of negative labels in the context of multi-species joint modelling, these approaches and findings are deeply considered in the ideas behind the methodology, which seeks to determine the pseudo-absences that test the model's discriminative power more effectively.

Chapter 3

Methods

This chapter delves into the detailed methodology employed to estimate the species ranges, building on existing frameworks while introducing novel contributions. The implementation, that follows closely the works of [29] and [7], deals with the exclusive use of large, crowdsourced presence-only data and semi-supervised learning approaches to address species distribution modelling, data that is far more available than presence-absence data, both in terms of locations and species.

The implemented model can be outlined in two components: the model's architecture and the loss function. We begin by exploring the Spatial Implicit Neural Representation (SINR) model architecture [7], which forms the backbone of our approach. After this, we focus our attention to the loss function and to the negative definition, which is crucial when working with presence-only data. As previously mentioned, is here where the main contribution of this study falls into, investigating and comparing different strategies for modelling absence in presence-only datasets, adjusting the loss function accordingly.

Before going into more detail we start by framing our approach. We parametrize the setting as follows: $x = \{lon, lat\}$ denotes a physical coordinate on the Earth's surface and $y_x^s \in 0,1$ denotes true absence ($y_x^s = 0$) or presence ($y_x^s = 1$) of species s at location x. However, our study relies on presence-only data which we denote as $z_x^s \in \{\emptyset, 1\}$. Here, $z_x^s = 1$ signifies that species s has been observed at x, while $z_x^s = \emptyset$ indicates no observation. The absence of a species does not confirm true absence, it might just be an unreported presence, which is particularly common on presence-only data where sampling efforts are different across classes and locations. Our methodology addresses this by developing techniques to infer probable absences.

3.1 Model Architecture: SINR

This section outlines the model design and architecture used for estimating species ranges, providing a foundation for the subsequent loss function discussion. The Spatial Implicit Neural Representation (SINR) model, proposed by [29] and later adjusted by [7] serves as the core methodology. While we won't exhaustively detail every aspect due to its similarity to prior implementations, it is essential to highlight two significant advantages over traditional Species Distribution Modelling (SDM) approaches. Firstly, the model's neural network architecture allows direct learning from raw spatial coordinates, bypassing the need for environmental covariates variables, commonly used in SDM studies. These covariates are often difficult to obtain at a global scale and may not capture all relevant information for modelling species distributions. Secondly, the architecture's scalability makes it particularly suitable for handling large volumes of crowdsourced presence-only data, a key resource in this study.

The architecture consists of two interconnected components, a location encoder and multi-label classifier. Broadly speaking, the location encoder, that we denote as f_θ, is responsible for learning a representation of the spatial domain, transforming raw coordinates into meaningful patterns that reflect geographical relationships. The multi-label classifier, denoted as h_ϕ on the other hand, leverages these representations to predict presence or absence of multiple species across locations. The overall implementation is expressed by the following equation:

$$\hat{y} = h_\phi(f_\theta(x)), \tag{3.1}$$

where $\hat{y} \in [0,1]$ contained the estimated probability of being present at x for each species. In terms of their design, the encoder corresponds to a fully connected neural network and the classifier to a single fully connected layer with sigmoid activations, common in tasks of binary classification. Appendix A contains a diagram summarizing the architecture of the model. In summary, SINR learns implicit geographical information based solely on the coordinates of presence-only data and produces a high-quality mapping of them around the consider surface.

Assuming differentiability of the encoder on its parameters, the model can be trained by traditional optimization methods, such as stochastic gradient descent. As a result, SINR is a methodology that, based on where species have been observed, learning implicit geographical information based solely on the coordinates of presence-only data and produces a high-quality mapping of them around the consider surface.

Following the procedure of [29], we also need to handle geographical coordinates, $x = [lon, lat]$ properly since they wrap around the Earth (for example -180° and +180° represent the same longitude). To avoid this boundary issue, coordinates are applied a sinusoidal transformation that rescaled to a range $[-1, 1]$, which results in each coordinate being represented as:

$$[sin(\pi lon), cos(\pi lon), sin(\pi lat), cos(\pi lat)]. \tag{3.2}$$

3.2 Loss Function and Negative Definition

Building on the SINR architecture described in Section 3.1, this section shifts focus to the loss function and the critical challenge of defining negatives in presence-only data. In single-positive labels context, the loss implementation is not a straight choice, the solely inclusion of positive labels would lead to the trivial solution of predicting that every species is present at every location. Which makes necessary the inclusion of some sense of negative values, which in the context of presence-only data are known as pseudo-absences. Our main contribution lies in investigating and comparing various strategies for modelling these absences.

In this section, we will address two key decisions related to the definition and incorporation of negative values, which will be thoroughly examined in the subsequent subsections. First, we will explore the strategies for how negative are sampled and integrated into the loss function. Second, we will discuss the method by which the negatives are going to be defined and selected.

3.2.1 Negative Sampling Strategy and Loss Specification

This subsection delves into the specific strategies for incorporating negatives into the loss function during training, a direct extension of the methodological framework. The loss function, adapted from [29], shares similarities with the Binary Cross Entropy Loss (BCE), commonly used in classification tasks, but we are including always one positive and one negative label. This approach ensures the model learns to distinguish between presence and absence effectively. Mathematically, each loss term is expressed as:

$$l_s = -[log(\hat{y}_s) + log(1 - \hat{y}'_s)] \tag{3.3}$$

Here, $\hat{y}_s$ represents the predicted probability for the observed presence location, and

$\hat{y}'_s$ the predicted probability for the assumed absence location. The total batch loss is then calculated as:

$$L = \frac{1}{n} \sum_{s=1}^{S} l_s \tag{3.4}$$

where n correspond to the number of presence observations in the batch, and S to the species observed.

Negatives are usually sampled in one of two ways, randomly across the whole geographical space (random background points), or randomly among the locations where other species have been observed (target-group background points). These methods give rise to the SSDL and SLDS losses proposed by [7], which will be explored in the following subsections. This division allows us to compare their effectiveness in handling the absence data challenge.

3.2.1.1 Random Background Points: SSDL Loss

Building on the negative sampling strategies outlined in Section 3.2.1, this subsection focuses on the first method: random background points, which leads to the Same Species, Different Location (SSDL) loss. This approach involves randomly selecting one coordinate from the whole space of available locations to be used as pseudo-absence. For example, when using worldwide presence-only data, a random $\{lon, lat\}$ pair is drawn, and $\hat{y}'_s$ in Equation 3.3 estimates the probability that the species is present at that location. [7] coined the term SSDL loss because each loss element pairs one species at its observed location with a different, randomly chosen location, providing a contrast for learning.

However, this method is not without challenges. Presence-only data is often subject to sampling bias, where observations are more frequent in accessible or popular regions due to heterogeneous sampling effort influenced by geographical areas and environmental conditions [17]. This results in mismatched distributions of positives and negatives that could bias the model towards environmental features of over-sampled areas, potentially leading to inaccurate species range predictions [34, 22].

3.2.1.2 Target-group Background Points: SLDS Loss

The most common method to account for the sampling bias of the presence-only data, is to randomly select pseudo-absences from the pool of locations where other species have already been observed, aligning the distribution of negatives and positives [34, 7, 46].

In this work we are going to follow the approach of [7], which would require a small adjustment to our equation of loss elements:

$$l_s = -[log(\hat{y}_s) + log(1 - \hat{y}_{s'})] \tag{3.5}$$

where before we had the prediction for the same species in an alternative locations $\hat{y}'_s$, we now have the prediction for a different species (s') in the same location, $\hat{y}_{s'}$.

However, while this strategy effectively addresses sampling bias by grounding negatives in observed locations, it is not clear if this strategy is always superior than the alternative of random background points. The effectiveness of either approach remains context-dependent. Previous studies have highlighted that both strategies are prone to different biases when working with presence-only data in SDMs. For instance, [4] and [46] suggest that random background points may over-represent inaccessible regions, while target-group points might under-represent rare species due to their reliance on co-observed locations.

3.2.1.3 Full Loss

Following the discussion of individual sampling strategies and their associated biases in Sections 3.2.1.1 and 3.2.1.2, this subsection examines a more comprehensive solution by integrating both methods into a unified loss function. By doing so, we aim to leverage the strengths of both strategies—SSDL's broad geographical coverage and SLDS's alignment with sampling effort—while potentially mitigating their respective weaknesses. One approach is to include for each observed presence a negative coming from both sources, which mathematically can be expressed as:

$$l_s = -[2 \cdot log(\hat{y}_s) + log(1 - \hat{y}'_s) + log(1 - \hat{y}_{s'})] \tag{3.6}$$

where the term belonging to the positive label is given a weight of two to account for the new number of negatives included.

As an alternative, [7], design a loss function that efficiently uses all the entries in $\hat{y}$. They assume that only the observed species are present at the observed locations, with all other species considered absent, and then assume that all species are absent from randomly selected locations. The full batch loss following this design can be expressed as:

$$L = \frac{1}{n} \sum_{s=1}^{S} [1_{[z_s=1]}\, \lambda\, log(\hat{y}_s) + 1_{[z_s \neq 1]}\, log(1 - \hat{y}_s) + log(1 - \hat{y}'_s)] \tag{3.7}$$

where the first term in the sum accounts for the observed locations, and the second and third terms for the SLDS and SSDL negatives, respectively.

Up until now, our discussion has centred on how the negative values are incorporated into the loss specification, however, we have yet to address the critical question of how these negative values are defined. The next section, will delve into this topic, exploring the criteria and processes used to determine pseudo-absences.

3.2.2 Negative Definition

As discussed in Chapter 2, various studies have tackled the selection of negative labels in semi-supervised presence-only learning settings [1, 4, 46]. This section examines how negatives are defined and integrated into the loss function.

3.2.2.1 Assumed Negatives

One simple but effective strategy is to assume that unobserved species-location pairs are negative. This approach is grounded by the fact that most locations will contain only small fraction of all the species considered in the analysis. Therefore, there is a high probability that a randomly selected species is absent from an arbitrary coordinate. However, the possibility of false negatives persists, particularly for species with broad distribution ranges. Reducing this uncertainty is a central goal of our study, motivating the exploration of informed negatives in the next subsection.

3.2.2.2 Informed Negative Labels: Absence and Proximity

Here we propose an alternative approach where the definition of pseudo-absences is grounded in observed presence patterns, generating what we term as "informed negative values." The core idea is to define a probable presence area for each species, and avoid selecting pseudo-absences from within these areas to reduce the risk of false negatives. We implement this by identifying the probable presence area as the H3 cells (at a specified resolution) in which the species have been observed[1]. This spatial delineation provides a structured basis for refining our negative sampling strategy.

For the informed version of the SSDL loss, pseudo-absences correspond to coordinates located outside the H3 cells where the species have been observed. While, for the informed version of the SLDS loss, we randomly select a species between those

[1] Appendix B gives a brief description of the H3 indexing system and the related functions that were used in this project

that have not been observed within the same H3 cell as the observed coordinate. This approach leverages the co-occurrence patterns of other species to infer absence.

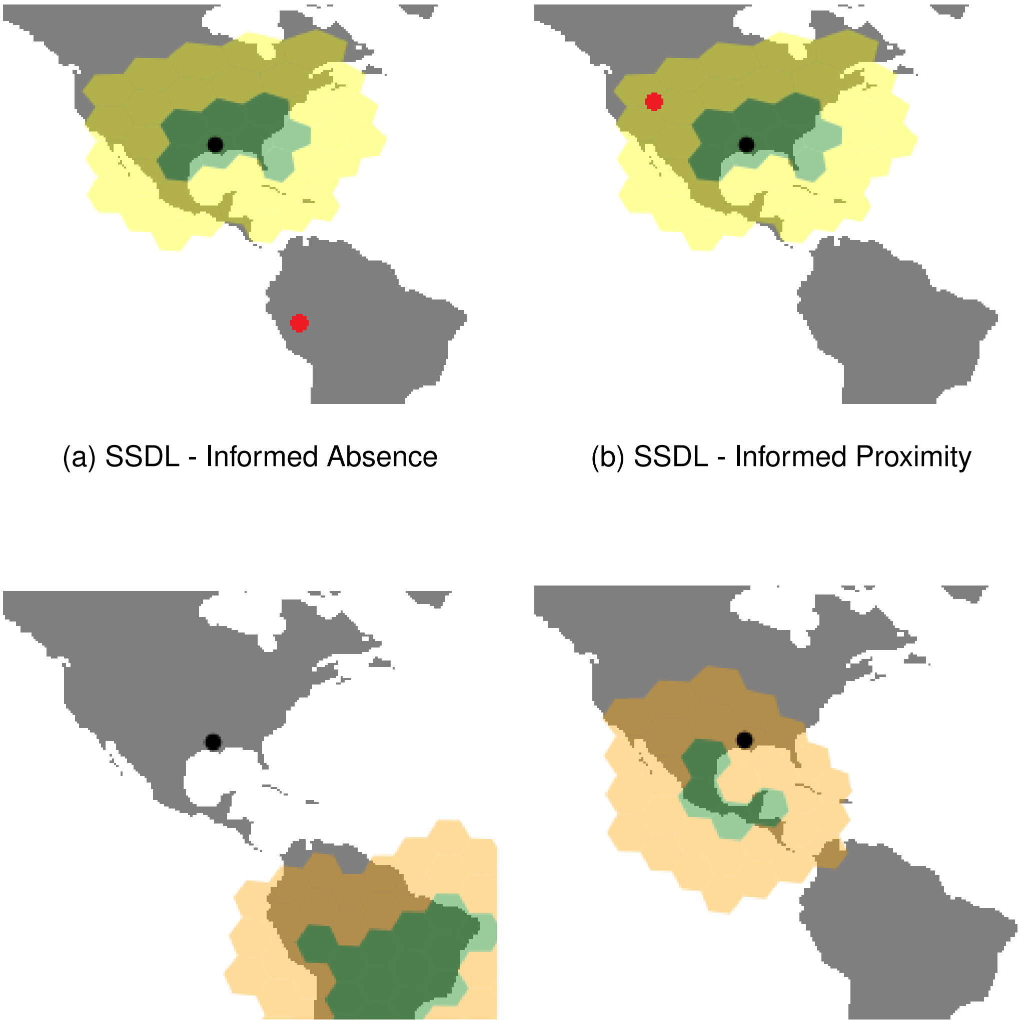
It is important to note that the negatives defined in this manner remain, in some sense, assumed negatives. We cannot entirely rule out the possibility that a species might truly be present outside its outlined presence area, especially for species with poorly documented ranges. However, the probability of introducing false negatives through this method is considerable lower compare to if we drawing locations completely at random from the entire geographical space.

However, the possibility of false negatives is not the only concern when selecting pseudo-absences randomly. We might inadvertently select location-species pairs that are geographically distant from the observed distribution of the species or that exhibit entirely different environmental features, such as climate or habitat type. These "easy" negatives may not challenge the model sufficiently, potentially limiting its ability to discern subtle boundaries or rare occurrences [28, 40]. This issue is tied to the concept of hard negatives, which has been explored in various studies, including those related to SDMs [1]. The general idea is to use more challenging negatives, which are similar to the observed presences (e.g., nearby locations or related species), to help the model learn better boundaries and improve its accuracy.

Considering these challenges, we explore an additional option for selecting pseudo-absences. Rather than using the the entire geographical area where the species is not expected to be found, we focus on locations that are close to the areas where the species have been observed. To achieve this, we define a proximity area, composed by the H3 cells that surround the species' presence area. This approach for selecting pseudo-absences aims to capture regions that are ecologically or geographically similar to the presence locations, potentially providing more relevant negative samples for training.

Table 3.1 summarizes the three discussed strategies for negative definition: assumed negatives, informed negatives based in absence area and informed negatives based in proximity area. While, Figure 3.1 illustrates how negative selection looks like for the same observation depending on the different approaches of informed negatives.

Based on the descriptions given in 3.2.1, adapting the SSDL and SLDS losses to incorporate informed negatives instead of assumed ones is relatively straightforward, as each observation (pair location-species) requires the selection of just one negative sample. On the other hand, for the Full loss we described two possibilities. We could be interested in adapting the loss developed by [7] to use the batch composition more efficiently. In summary generates two matrices: one for observed locations and one



(a) SSDL - Informed Absence

(b) SSDL - Informed Proximity

(c) SLDS - Informed Absence

(d) SLDS - Informed Proximity

Figure 3.1: Alternatives for Informed Negatives. This figure illustrates the pseudo-absences selection based on different informed negative strategies for the observation corresponding to the Broad-headed Skink in the coordinates $[-92.09, 32.33]$. Subfigures (a) and (b) show the SSDL model's response to absence and proximity conditions, respectively. Subfigures (c) and (d) present the corresponding results for the SLDS model, highlighting differences in their handling of negative information. The black dot accounts for the observed location. For (a) and (b) the green and yellow area show the presence and proximity area belonging to the observed species, while the red dot mark the selected negatives. For (c) and (d) the presence and proximity areas of the randomly selected background species are shown in green and orange, respectively. It can be seen how the proximity area covers the observed coordinate in (d).

Loss	**Positive Label**	**Negative Label SSDL**	**Negative Label SLDS**
Assumed Negatives	Observed locations (pair species/cell)	Random location	Random species
Informed Negatives - Absence	Observed locations (pair species/cell)	Random location excluding cells with presence	Random species excluding the ones that have been observed in the cell
Informed Negatives - Proximity	Observed locations (pair species/cell)	Random location in the proximity of the presence area	Random species whose proximity area includes the observed cell

Table 3.1: Summary of Negative Definition Strategies.

for the random sampled, each containing a number of elements equal to the product between the batch size and the number of species in training. One potential idea is to remove from the loss all the assumed negatives that correspond to locations within H3 cells where the species have been observed, as these are more likely to be false negatives. The problem with this approach is its inefficiency and time-consuming nature. It would require checking each of the millions of elements computed for the loss in every batch, a process that becomes impractical for large datasets. Undermining one of the primary benefits of using large presence-only datasets—scalability. Therefore, we will consider the simpler strategy described in Equation 3.10, where we include two negatives in the loss for each observation, one sampled in the fashion of SSDL (random location for observed species) and one in the fashion of SLDS (random alternative species for observed location).

3.2.2.3 Dual Losses

Following the exploration of informed negative strategies in Section 3.1, here we introduce a dual approach to define pseudo-absences, aiming to capitalize on the strengths of both absence and proximity methods. Each strategy offers distinct advantages. Selecting negatives from a wider absence area is beneficial for the model, as it enables the learning of distinctions between distant locations that may share certain features, such as similar weather patterns or ecological conditions. Conversely, providing negatives closer to the presence areas facilitates the model's ability to delineate the boundaries of a species' distribution more accurately, capturing fine-scale spatial variations.

With this in mind, we implement a Dual Loss function that includes one negative from each definition in the loss function. We can express the losses in this case as following:

Dual SSDL:

$$l_s = -[2 * log(\hat{y}_s) + \sum_{i=(a,p)} log(1 - \hat{y}_s^i)] \tag{3.8}$$

Dual SLDS:

$$l_s = -[2 * log(\hat{y}_s) + \sum_{i=(a,p)} log(1 - \hat{y}_{s^i})] \tag{3.9}$$

Dual Full:

$$l_s = -[4 * \cdot log(\hat{y}_s) + \sum_{i=(a,p)} (log(1 - \hat{y}_s^i) + log(1 - \hat{y}_{s^i}))] \tag{3.10}$$

where (a, p) represent the negatives selected based in absence and proximity and a factor of 4 is given to the positive label to account for the number of negatives included.

3.2.2.4 Hybrid Losses

While the model may benefit from considering both absence and proximity negatives, incorporating them simultaneously for each element in the loss can hinder the model's ability to interpret the loss signal. Receiving an aggregated loss composed of both absence and proximity negatives may prove detrimental in the mapping of clear geographical patterns. How the number of pseudo-absences used affects SDM model's performance has been a focus of research on its own [13, 1]. We are not getting involved in this subject, restricting our focus to study how the performance of a presence-only SDM is affected by the definition of negatives.

To address this, we implement a Hybrid Loss function that randomly selects, for each observation, whether the pseudo-absence to be included originates from the absence area or the proximity area. During a training batch, a presence observation of a species might be paired with a negative from a distant absence region in one iteration, and with a negative from an adjacent proximity cell in the next. Allowing the model to benefit from both broad geographical context and localized boundary information, but receiving a more transparent loss signal. In the Full version of the Hybrid Loss, both negatives are generated using the same definition strategy (either absence or proximity) to ensure consistency. Mathematically, this loss follows the same specifications described for the informed losses.

3.3 Optimization and Training Details

This section outlines the optimization and training procedures that bring the Spatial Implicit Neural Representation (SINR) model, described in Section 3.1, and the loss functions detailed in Section 3.2, to practical implementation. The training configuration is largely invariant to the approach established by [7]. All of the experiments were trained during 10 epochs, with batch size of 2048 and learning r of $5e^{-4}$. Adam was used as optimization algorithm [27] with an exponential learning rate decay schedule of 0.98 each epoch.

Additionally, the construction of presence and proximity areas, integral to the informed negative strategies, relied on an H3 resolution of one[2]. This resolution provides a coarse but manageable grid for global analysis, dividing the Earth's surface into large hexagonal cells that simplify spatial computations. The proximity area was defined using two layers of adjacent cells: the first layer consists of the cells that are adjacent to presence cells, and the second layer includes cells adjacent to the first.

[2] There are a total of 842 cells of resolution 1 that cover the whole world, each of them with an area of around 610 thousand square kilometres [43].

Chapter 4

Training Data and Processing

This chapter details the source of the training data and the necessary preprocessing steps applied prior to training the model.

4.1 iNaturalist Description and Preprocessing

The training dataset comprises presence-only observations from iNaturalist, an open and crowdsourced platform, where each record corresponds to a real-world observation of a species at a specific location. As of April 2025, Naturalist contains over 230 million observations across more than 528,000 species [25].

To maintain comparability with the baseline established by [7], we utilize their preprocessed training data, which applies filters including date validation, label consensus, and a minimum of 50 observations per species. However, we introduce two additional preprocessing decisions. First, we restrict the analysis to species that also come with expert range maps from the International Union for Conservation of Nature (IUCN), which is going to be used as evaluation data (see Section 5.1). Second, we cap the maximum number of observations per species at 100 during training, a threshold that already yielded strong performance in [7], reduces exposure to class imbalances issues, and reduces computational demands. Allowing us to centre our attention to the negative definition exploration. This results in a dataset of 2,418 species and 223,792 observations.

Despite these measures, sampling bias remains a concern. Figure 4.1 illustrates the geographical distribution of training locations, revealing a pronounced concentration in some regions (North America, Western Europe, the east coast of Australia, and parts of Southeast Africa). As discussed in Section 3.2.1, these imbalances significantly influence model performance, underscoring the importance of tailored negative selection strategies in addressing biased data distributions.

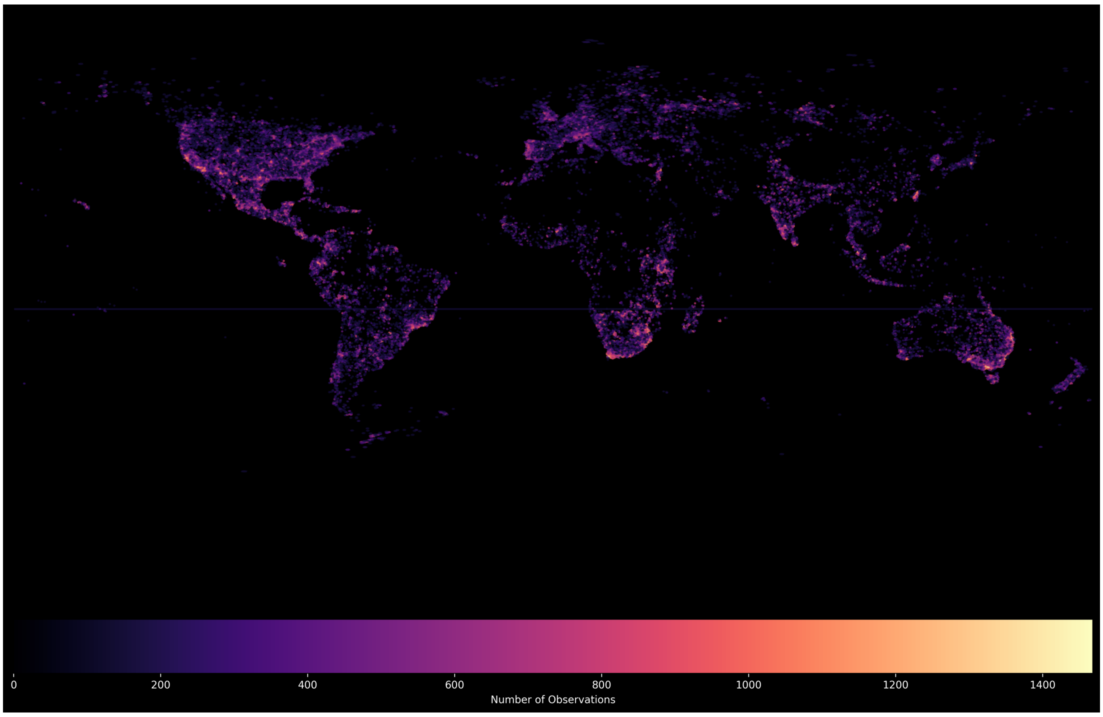


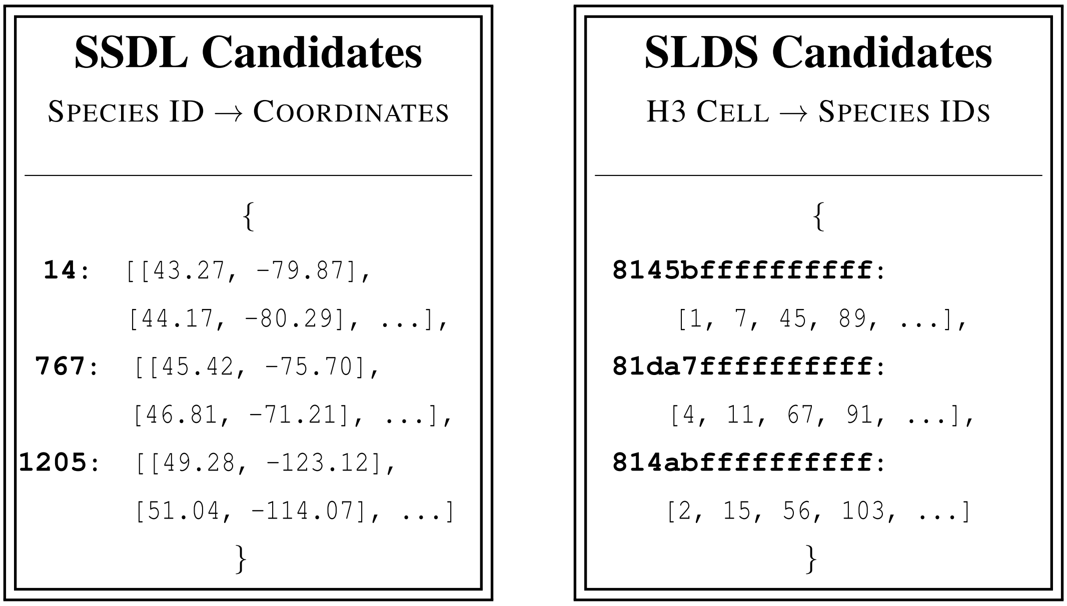
Figure 4.1: Geographical Distribution of Training Data.

4.2 Negative Candidates Generation

As outlined in the previous chapter, this study employs informed negatives derived from pseudo-absences in absence and proximity areas. In this section we are going to focus in the actual computation of these negatives. To avoid the inefficiency of generating pseudo-absences for each positive label during training, we precompute lists of negative candidates once, which are subsequently loaded with the training dataset. These candidates comprise coordinates for each species (for informed Single-Sample Distribution Loss, SSDL) and species for each location (aggregated at the H3 cell level for Spatial Location Distribution Loss, SLDS).

We start by computing and storing the H3 cells in which each species have been observed, considering a maximum of 1,000 observations per species. This cap is assumed sufficient to approximate geographical distributions accurately while reducing computational overhead compared to using the full dataset. For SLDS, negative candidates are directly derived from this mapping, containing all the species that have not been reported in each observed cell. For SSDL, we generate a list with 1,000 negatives for each species, by first selecting randomly a cell from the ones in which the species has not been observed, and then one random coordinate inside the boundaries of this cell.

A parallel procedure is done for obtaining the candidates based on the proximity areas. For SSDL, a list at species level is created with 1,000 negatives candidate locations within each species' proximity area. While for SLDS, a list is generated at cell level, including species whose proximity areas encompass the cell. Figure 4.2 exemplifies these candidate lists. During training, a single candidate is randomly selected from the precomputed lists based on the chosen negative sampling strategy (SSDL or SLDS) and definition (absence or proximity), as depicted in Figure 4.3, which outlines the loss computation workflow, including the integration of pseudo-absence candidates.



SSDL Candidates

Species ID → Coordinates

```
{
14:   [[43.27, -79.87],
       [44.17, -80.29], ...],
767:  [[45.42, -75.70],
       [46.81, -71.21], ...],
1205: [[49.28, -123.12],
       [51.04, -114.07], ...]
}
```

SLDS Candidates

H3 Cell → Species IDs

```
{
8145bffffffffff:
    [1, 7, 45, 89, ...],
81da7ffffffffff:
    [4, 11, 67, 91, ...],
814abffffffffff:
    [2, 15, 56, 103, ...]
}
```

Figure 4.2: Example structure of negative candidate lists for both loss functions.

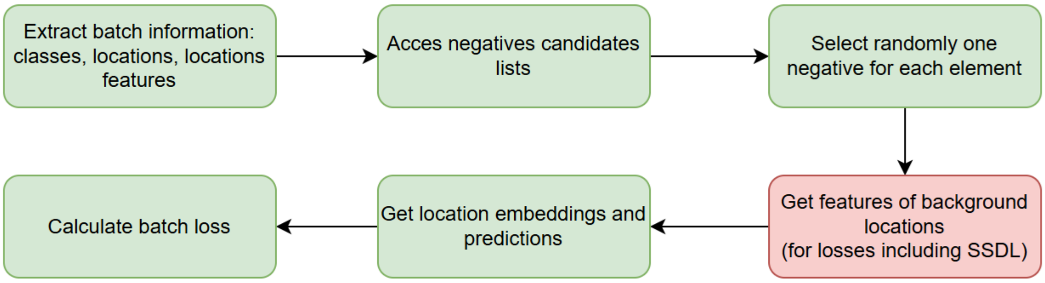

Extract batch information: classes, locations, locations features
Acces negatives candidates lists
Select randomly one negative for each element
Get features of background locations (for losses including SSDL)
Get location embeddings and predictions
Calculate batch loss

Figure 4.3: Loss computation procedure.

Chapter 5

Evaluation Task and Metrics

This chapter presents and discusses the evaluation of our proposed training losses, describing what is going to be the evaluation task to be resolves and which are the used metrics of performance.

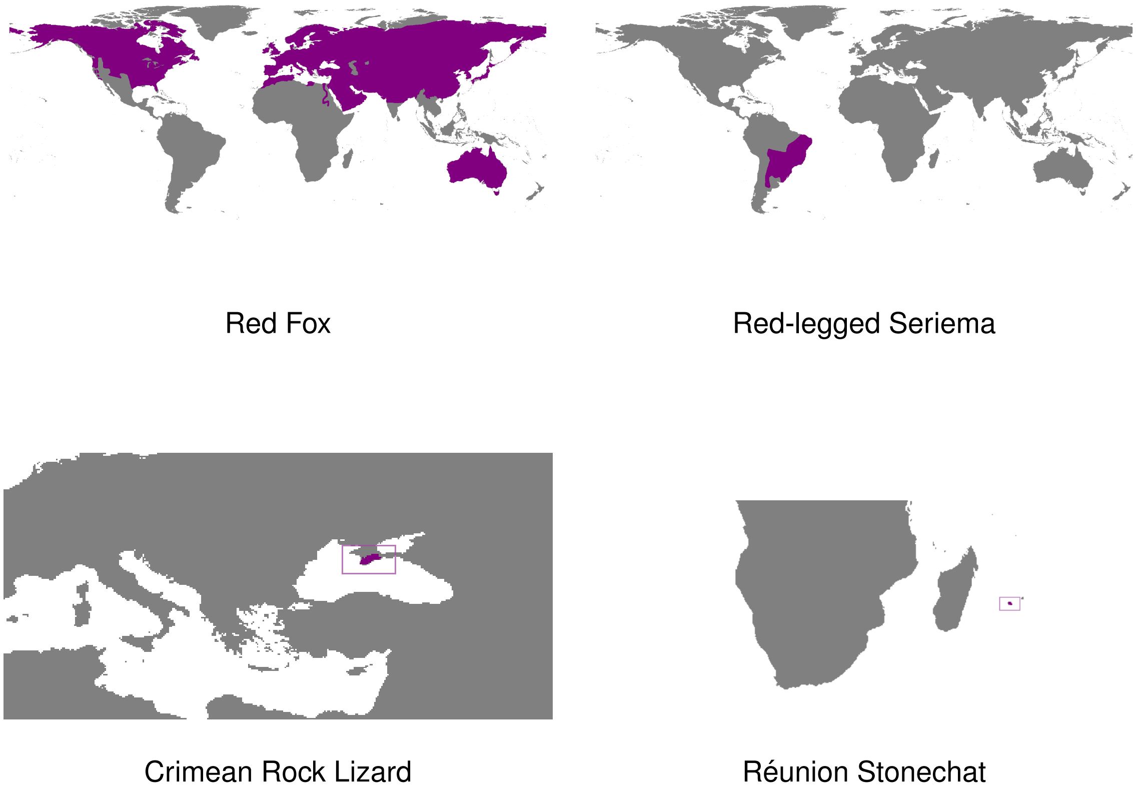
5.1 Evaluation Task

As mentioned in Chapter 2, the common strategy is to use presence-absence data to evaluate, given the sampling bias existing in presence-only observations. We choose to follow this line by evaluating the performance of our specifications against curated expert range maps from the International Union for Conservation of Nature (IUCN), which provide presence-absence areas for thousands of species around the world.

As for the training data, we are going to be using the processed IUCN dataset provided by [7]. This datasets contains 2,418 species which are present in both our training data and the IUCN dataset, belonging to taxonomically and geographically diverse set of species. Specifically, the species correspond to 1,368 birds, 438 reptiles, 330 mammals, and 282 amphibians. Figure 5.1 shows the expert created range map for four of the species contained in our training and evaluation data. It can be seen that the range of the species are quite varied, both in terms of range size and geographic location.

5.2 Performance metrics

The main metric to evaluate the quality of the experiments predictions is the Mean Average Precision (mAP). This evaluation metric, measures how well the model ranks



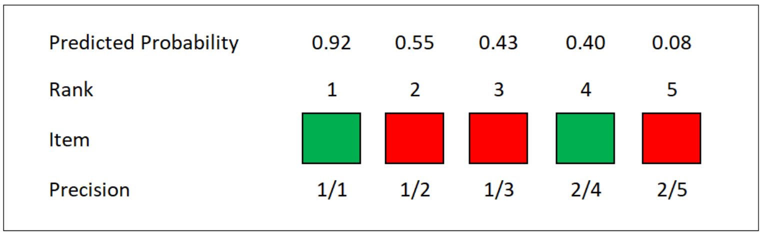
Red Fox

Red-legged Seriema

Crimean Rock Lizard

Réunion Stonechat

Figure 5.1: Expert ranges for four selected species. The figure shows the presence area mapped in IUCN for four selected species. This species have been seleted to illustrate the large differences that can exist between the distribution of different species, from species with intercontinental ranges (e.g. Red Fox), continental distribution (e.g. Red-Legged Seriema) and species found in reduced areas (e.g. Crimean Rock Lizard in Crimea and Réunion Stonechat in the Réunion island on the Indian Ocean).



Predicted Probability	0.92	0.55	0.43	0.40	0.08
Rank	1	2	3	4	5
Item	🟩	🟥	🟥	🟩	🟥
Precision	1/1	1/2	1/3	2/4	2/5

Figure 5.2: Example of Average Precision Computation.

positive groundtruth labels (presences) consistently higher than negative groundtruth labels (absences). Figure 5.2 shows a simple example with 2 positive and 3 negative labels for a class. The first step is to rank all the locations to be evaluated based on their calculated prediction probability in decreasing order. After this, a precision is calculated for each of the labels, given by the fraction between the number of positives with equal or higher ranking and the ranking belonging to the current label. For example, for our example, the positive label ranked with the fourth highest prediction has a precision of $2/4 = 0.5$, given that there are two positive labels with equal or higher ranking than 4. Finally the average precision corresponds to the average of the precision for the positive labels, which in the case of our example corresponds to $\frac{1}{2}(1+1/2) = 0.75$. This generates an average precision per class (species in our case), so then the mean is calculated to compute the mean average precision and having in this way an indicator representative of the complete evaluation set.

In addition to mAP, we evaluate model performance using Mean Precision (mP) and Mean Recall (mR). Unlike mAP, which works directly with prediction probabilities, these metrics require converting predictions to binary classifications (present/absent). For each species, we determine an optimal threshold that maximizes the F1 score—a balanced measure combining precision and recall. Precision measures the proportion of predicted presences that are actually correct, while recall measures the proportion of actual presences that the model successfully identifies. The mP and mR are calculated by averaging the precision and recall values across all species. Together, these metrics complement mAP by evaluating whether the model can make reliable binary predictions, not just rank locations effectively.

Chapter 6

Experiments and Results

This chapter presents and analyses the outcomes of the experiments conducted using the methodology outlined in Chapter 3. We begin by examining the results for assumed negative strategies, establishing a baseline, before proceeding to evaluate informed negative strategies.

6.1 Assumed Negative Labels

In this strategy, which will serve as our baseline, we assume that unobserved species-location pairs are negative, treating randomly selected locations and locations observed for other species as absences. Table 6.1 shows the results for this method for the three sampling strategies mentioned in Section 3.2.1: same-species different location (SSDL), same-location different species (SLDS) loss and the Full loss, combining both. These results are not directly comparable to those of [7], due to our model being trained on a constrained dataset of 2,418 species present in both our training and evaluation sets. Training with more species typically improves generalization, so our mean performance metrics may be somewhat lower because of this.

The results show that SSDL outperforms SLDS, likely because SLDS creates bias toward well-sampled regions, reducing accuracy for species in less-studied areas. However, incorporating both sampling strategies in the Full Loss yields the best overall performance, with notable improvements in mean precision (mP) and mean average precision (mAP). Suggesting that there is a benefit in considering different sources of negative labels, due to them providing complementary information.

Despite the improved performance observed for the Full loss, the key limitation still appears to be a tendency to overpredict presence areas. This is reflected in consistently

	SSDL			SLDS			FULL		
Method	**mAP**	**mP**	**mR**	**mAP**	**mP**	**mR**	**mAP**	**mP**	**mR**
Assumed Negatives	0.442	0.424	**0.730**	0.299	0.295	0.664	0.517	0.441	**0.772**
Informed - Absence	0.438	0.392	0.768	0.323	0.303	**0.672**	0.495	0.431	0.761
Informed - Proximity	0.328	0.333	0.614	0.175	0.194	0.622	0.352	0.360	0.591
Dual	0.512	0.489	0.700	**0.331**	**0.335**	0.640	0.504	0.476	0.690
Hybrid	**0.522**	**0.494**	0.696	0.319	0.326	0.638	**0.556**	**0.501**	0.735

Table 6.1: Results for Assumed Negatives and Informed Negatives strategies. SSDL (random background), SLDS (target-group background) and Full (both) correspond to the negative sampling strategies. Performance metrics include mean average precision (mAP), mean precision (mP), and mean recall (mR). Evaluation is based on predicting presence-absence from IUCN expert range maps. The same 2,418 species are used for both training and evaluation.

lower mean precision relative to recall values, indicating that while models successfully capture most actual presences, they also incorrectly classify many absence locations as presences. Incorporating more informed negatives could potentially improve model precision by refining the boundary between presence and absence.

6.2 Informed Negative Labels

In this section we examine the the performance of models incorporating informed negatives, to evaluate how the negative definition impacts over the predictive ability of our SDM model that learns based on coordinates alone.

6.2.1 Absence-Informed Loss

We begin with the model using pseudo-absences informed by absence areas, where random locations are selected from H3 cells where the species has not been observed, differing from the assumed negative approach. The results, shown in the second row of Table 6.1, reveal an improvement in mean average precision for SLDS but a decline for SSDL. This could be explained by the fact that, with assumed negatives, the possibility of falling on false negatives is greater when target-group background is used: assuming absence from locations where we know that species have already been observed is riskier than assuming absence from completely random locations across the world. While for SSDL, completely assumed pseudo-absences can belong to coordinates closer to the species distribution that, even though they are potentially false negatives, provide more informative negatives for the model.

Relating the Full loss, a decrease in performance is observed, which could be explained by two reasons. First, the negative effect when using an SSDL sampling strategy might be dominating the positive effect observed for SLDS. Second, the higher performance of assumed negatives for the Full loss could be explained by the fact that this approach leverages a larger volume of data per training batch, as noted in Section 3.2.2.2, whereas the Informed-Absence Full Loss includes only one negative of each type (SSDL and SLDS) for computational reasons. This constraint likely limits the diversity of negative samples, potentially exceeding the benefit of using a few informed pseudo-absences. Furthermore, assumed negatives have been found to yield better models in the presence of geographically biased presences [1].

6.2.2 Proximity-Informed Loss

We look now into the informed proximity approach, which defines negatives within the vicinity of observed presence areas. The third row of Table 6.1 shows that this method underperforms in all three negative sampling strategies, compared to both assumed and informed-absence negatives. This could be caused by the reduced range on the distribution of the negatives, as presence-only data is biased towards certain regions, causing proximity areas to likewise concentrate around these zones. This harms the model's ability to learn a representation that adapts well to most locations in the world.

Although the negatives defined within the whole absence area appear to be better overall, this is not true for every species. Figure 6.1 illustrates that some species benefit significantly from the use of proximity areas to define negatives. Moreover, even when one approach outperforms another for a particular species, there may still be value in incorporating negatives from both methods. Combining pseudo-absences from different sources could provide complementary information, allowing the model to learn a more robust geographical representation, as discussed in Section **??**. The next subsection will explore hybrid and dual approaches, which consider negatives defined from both informed approaches.

6.2.3 Hybrid and Dual Losses

The fourth row of Table 6.1 reports the results obtained for the experiments based on the Dual Loss, which simultaneously includes one absence and one proximity negative.

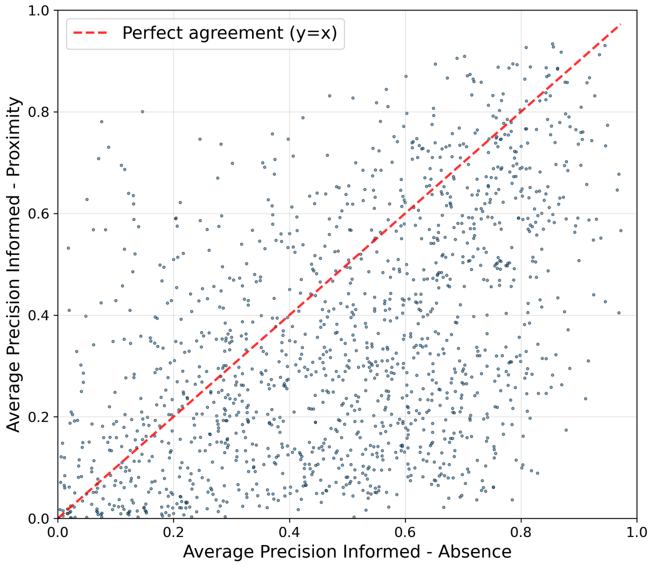


Figure 6.1: Species average precision: Full-Informed-Absence-Loss vs Full-Informed-Proximity-Loss.

This approach yields notable improvements in SSDL and SLDS, outperforming both the Informed and Assumed negatives approaches. However, this improvement does not hold for the Full Loss, whose performance is lower than that observed for the SSDL, suggesting a potential overload of conflicting signals that dilutes the model's ability to discern distribution patterns.

Considering the findings for the Dual Loss, we aim to evaluate the performance of the Hybrid Loss, which randomly selects between negatives defined via absence or proximity areas during training, with a 50% probability for each. The results for this approach, shown in the fifth row of Table 6.1, indicate a marked improvement for SSDL and, particularly, for the Full Loss, which achieves the best performance among all our specifications, offering an improvement of approximately four percentage points in mean average precision compared to our baseline of assumed negatives. The superior performance of the Hybrid model (mAP = 0.554) over the Dual model (mAP = 0.505) might be attributed to the use of a single negative per sampling strategy, providing a more stable loss signal compared to the Dual model's inclusion of two negatives.

Overall, the results in this section highlight the benefit of combining absence- and proximity-defined negatives, providing complementary information that enhances the model's discriminative power across species.

6.2.4 Hybrid and Dual Losses with Species-specific Preferences for Absence and Proximity Negatives

The initial experiments with hybrid and dual loss approaches provide a foundational exploration, given the arbitrary decision of giving equal preference to absence or proximity negatives. An idea could be increase the relevance of pseudo-absences based in absence, due to its superior performance for most species. However, we are going to try to understand which factors influence in the fact that one alternative is better than the other one for certain, so we can assign species specific values for these parameters.

Several studies have highlighted the influence of species-specific characteristics on SDM model outcomes and efficacy of negative selection strategies. [38] and [42] observed that geographic traits significantly affect performance, while [4] and [46] highlighted that optimal negative selection depends on the distribution of species observations. Furthermore, evidence suggests that SDM discriminative power diminishes for species with smaller range sizes and greater demographic or distributional uniqueness [1, 41, 42].

Motivated by these findings, our study explores whether range size and distributional rarity influence the preference for learning from absence or proximity negatives. To this end, we are going to adapt the Hybrid and Dual losses to reflect species-specific preferences for absence or proximity negatives. The Hybrid Loss, originally assuming equal probability for absence and proximity negatives, can be modified to incorporate per-species probabilities. Similarly, for the Dual Loss we can adjust the weight of the absence and proximity negatives differently.

We first start by estimating the mentioned distribution traits. To quantify species range size, we use the number of H3 cells that compose the species presence area, which were already used for the generation of the absence and proximity areas. This measure directly quantifies the spatial extent of a species' distribution, and is particularly advantageous for species observed across different continents, where more complex area calculations could require extensive adjusting and not generalize well to our 2,418 species. While, the uniqueness of the species distribution is quantified by the Kullback–Leibler (KL) divergence between its distribution and the average presence distribution across all species. Providing in this way a continuous indicator of how atypical a species' habitat preferences are relative to the broader ecological context. This is particularly valuable for distinguishing between species with similar range sizes but differing distribution patterns. Further details about the KL divergence estimation are provided in Appendix C. Figure 6.2 illustrates the distribution of these characteristics across the training species.

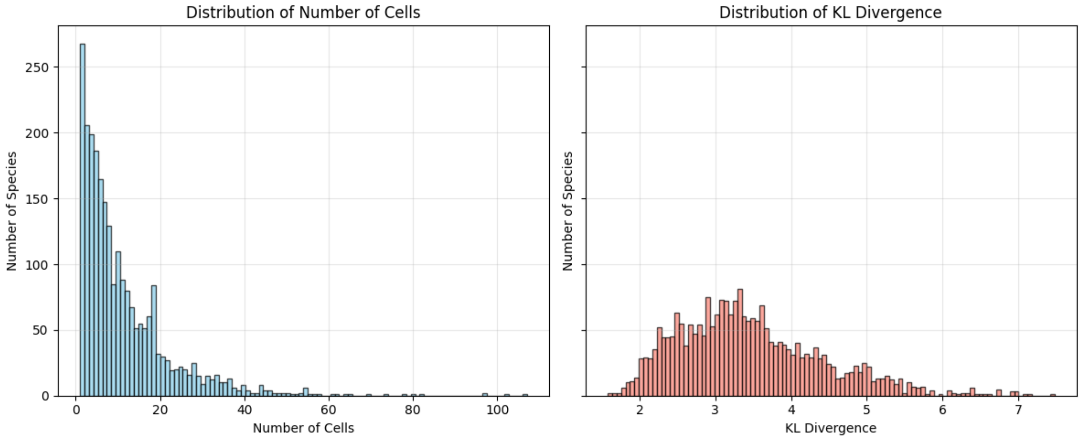


Figure 6.2: Distribution of species range sizes and distribution singularity.

We need to first estimate the species preferences for absence and proximity, and then use them in the training of our SDM model to examine their effect on the modelled species distributions. Therefore, we proceed to split the datasets based on species, 60% of them will be used to estimate preferences (validation set) and the remaining 40% to evaluate the accuracy of the predictions (test set). The left panel of 6.3 presents the difference in species average precision (AP) between the Full-Informed-Absence-Loss and Full-Informed-Proximity-Loss for the validation set, aggregated into hexagonal bins using the mean difference. As anticipated, the majority of species exhibit a positive difference, indicating that absence negatives generally outperform proximity negatives. However, a clear pattern emerges: species with smaller ranges (fewer cells) and with unusual distributions (higher KL divergence) benefit more from proximity negatives, whereas those with lower KL divergence and intermediate cell counts favour absence negatives more highly. Notably, it appears to be a non-linear relation between the number of cells and the preference for absence, observing smaller positive differences in AP.

The right panel of Figure 6.3 depicts a partition of the number of cells and KL divergence domain, assigning absence preferences for training based on K-means clustering of the observed differences in average precision (AP). A detailed methodology is provided in Appendix D. We can see how the method assigns probabilities closer to 0.5 for species with smaller and rarer distributions, signalling a balanced benefit from either absence or proximity negatives. While, species that follow a distribution more similar to the average are assigned a higher absence preference. The specified preference values will define the probability of selecting absence in the adapted Hybrid Loss, and the weight for absence negatives in the modified Dual Loss, while one minus this value will be the preference for proximity.

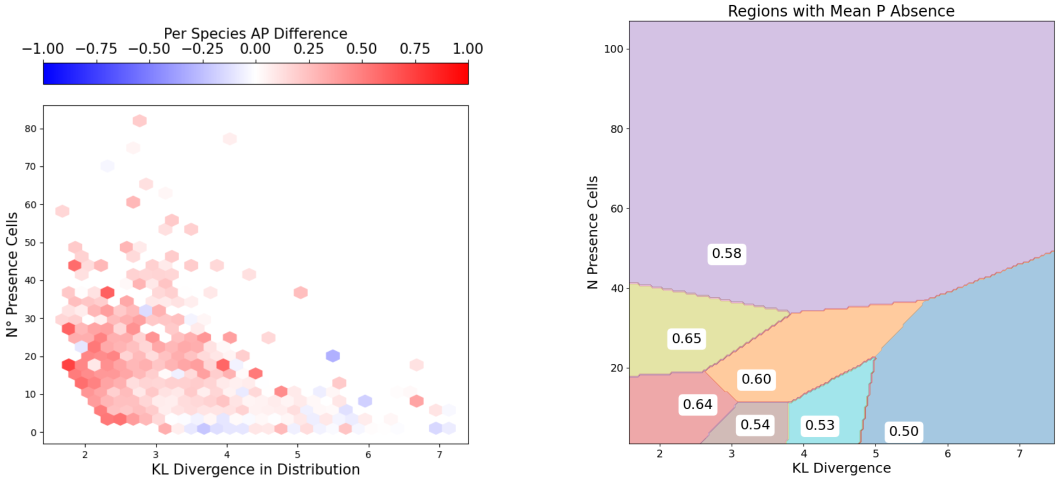


Figure 6.3: Estimation of preference for absence based on species AP difference as a function of number of presence cells and KL divergence. *(Left)* Distribution of the AP differences between Full-Informed-Absence-Loss and Full-Informed-Proximity-Loss. *(Right)* Regions and their corresponding preferences for absence.

Table 6.2 presents the results from experiments implementing the modified versions of the Dual and Hybrid loss, incorporating species preferences for absence based on their observed distribution. Baseline performances for the original Dual and Hybrid Losses on the same species set are also included. The results show no clear evidence of a generalized performance improvement from including informed preferences for the type of negative. For the Dual Loss, a marginal improvement is observed across all three performance metrics (mAP, mP, mR) when using SSDL or SLDS as the sampling strategy. However, the modification proves slightly detrimental to the Hybrid Loss, with its Full approach—previously the best model—experiencing a decline of approximately one percentage point in mAP.

Possible reasons for this lack of success include the method to estimate preferences not being optimal and a higher benefit of learning from absence and proximity in a balanced manner. This suggests that while species-specific preferences hold theoretical promise, the current implementation may not effectively capture the nuanced ecological differences driving negative selection efficacy.

	SSDL			SLDS			FULL		
Method	**mAP**	**mP**	**mR**	**mAP**	**mP**	**mR**	**mAP**	**mP**	**mR**
Dual	0.513	0.487	0.699	0.325	0.330	0.633	0.502	0.474	0.683
Dual - Species Weight	0.516	0.488	0.704	0.332	0.333	0.640	0.499	0.469	0.693
Hybrid	0.521	0.493	0.694	0.317	0.324	0.638	0.559	0.522	0.709
Hybrid - Species probability	0.515	0.492	0.694	0.320	0.328	0.634	0.550	0.514	0.706

Table 6.2: Results for Hybrid and Dual strategies considering species preference for absence/proximity negatives. SSDL (random background), SLDS (target-group background) and Full (both) correspond to the negative sampling strategies. Performance metrics include mean average precision (mAP), mean precision (mP), and mean recall (mR). Evaluation is based on predicting presence-absence from IUCN expert range maps. Results reported correspond to the 968 species in the test set, while all 2,418 species are used for training.

6.3 Predictions: Assumed Negatives vs Hybrid

In this section, we further compare the performance of our best model, the Hybrid Full Loss, against the baseline provided by the Assumed Negative Full Loss. As summarized in Table 6.1, the Hybrid Loss boosts mAP from 0.517 to 0.556, driven mainly by improved mean precision, indicating that combining absence and proximity negatives reduces overestimation of species ranges and improves habitat boundary accuracy.

To investigate whether the Hybrid Loss performs better for specific groups of species, Figure 6.4 illustrates how species average precision varies with range size (number of presence cells) and distributional singularity (KL divergence). Two key conclusions emerge: first, both approaches struggle to model distributions for species observed in fewer presence cells and with distributions more dissimilar to the average presence across all species, second, the improvement from the Hybrid approach appears to concentrate in certain group of species. A consistent increase in mAP is observed for species that have been reported in a limited number of H3 cells, provided their KL divergence is not excessively high. Additionally, average precision improves for species with intermediate range sizes but low divergence from the average distribution.

Figure 6.5 compares the predicted distributions for three selected species, highlighting cases of success and failure for the Hybrid Loss in comparison to our baseline of

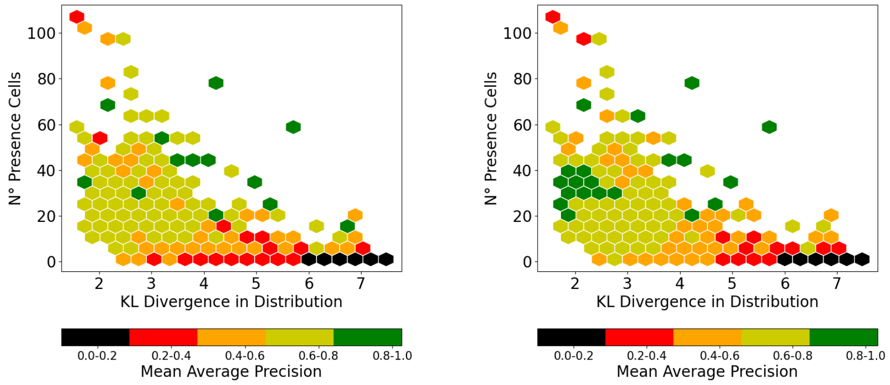


Figure 6.4: Mean average precision of Assumed Negative Full Loss (left) and Hybrid Full Loss (right) as a function of the number of presence cells and KL divergence of species distribution.

assumed negatives. The predicted ranges for the first species considered (Blue-breasted Bee-Eater) illustrate the main two effects observed overall regarding the use of the Hybrid approach: achieving better discrimination between the ground-truth presence area and regions close to it, but tending to incorrectly predict presences in regions further away due to a decrease in the use of arbitrarily selected pseudo-absences. In this case, however, the positive effect is far more relevant, with the average precision being 0.712 for the Hybrid strategy and 0.228 for the baseline. The second species considered (Puerto Rican Spindalis) corresponds to the group found to benefit most from the inclusion of the Hybrid Loss—species with limited range size. We can see more clearly here how the assumed negatives approach was not able to discern between presence and absence areas with great granularity, while the Hybrid approach is able to assign the highest predicted suitability more accurately to the presence locations (mAP Assumed Negative: 0.092, mAP Hybrid: 0.866). Conversely, for Species 3 (Madagascar Buzzard), a failure case, the Hybrid Loss struggles to predict larger habitats than it should, assigning non-negligible predicted probabilities to locations on entirely different continents (mAP Assumed Negative: 0.655, mAP Hybrid: 0.119).

These results underscore the Hybrid Loss's potential to refine SDM predictions, particularly for species with limited ranges, by leveraging complementary information from absence and proximity negatives. However, there is still room for further refinement to enhance model robustness across diverse species profiles.

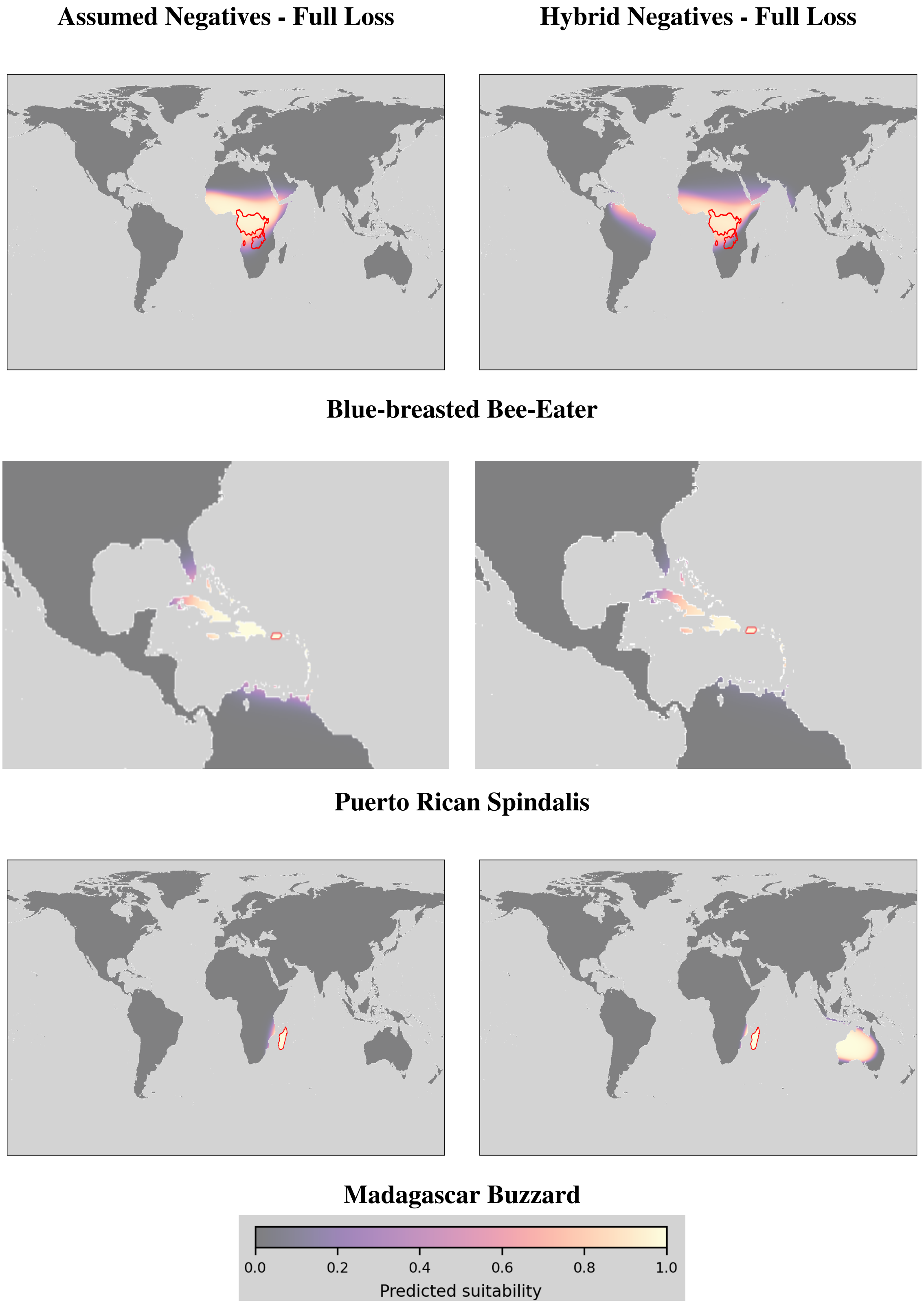


Figure 6.5: Comparison of predictions, for selected species, of Assumed Negative (left) and Hybrid loss (right) against the expert range maps provided by IUCN (delimited with a red line).

Chapter 7

Conclusions

7.1 Summary and Main Findings

This study examined how different strategies for defining negatives influence the performance of coordinate-based species distribution models trained on presence-only data. The results show that assumed negatives, while simple, provide a strong baseline but tend to overestimate species ranges. Informed approaches produced mixed outcomes: absence-informed negatives offered some benefits, whereas proximity-informed negatives alone underperformed due to their limited spatial scope. The most consistent improvements were achieved through hybrid strategies that combine absence- and proximity-informed negatives, which enhanced mean average precision and yielded more accurate habitat boundaries. Attempts to introduce species-specific weighting based on range size and distribution singularity did not consistently outperform simpler hybrid models, though they highlighted potential directions for refinement. Importantly, the improvements of the hybrid approach were most evident for species with small geographic ranges, a group of particular conservation concern. Overall, these findings demonstrate that negative selection is a crucial design choice in presence-only SDMs, and that hybrid definitions offer a practical and effective way to improve predictive performance at a global scale.

7.2 Limitations and Future Work

Our study is subject to some limitations stemming from methodological decisions. First, we restricted our analysis to a subset of species with valid data in both the training (iNaturalist) and evaluation (IUCN) datasets, limiting the total to 2,418 species. This choice, while practical, is detrimental in the sense that the model benefits from the inclusion of additional species even when they are not evaluated at test time [7]. Second, the results reported are the product of a single evaluation of the different experiments, lacking the robustness that multiple iterations could provide. Third, our evaluation relies solely on comparing predictions to the presence-absence data in IUCN expert range maps. These maps, however, are not foolproof; they have been shown to overestimate species habitats [23] and fail to accurately reflect the true distributions [24, 18].

The negative definition strategy employed in this work also presents limitations that motivate further investigation. For instance, the negative random background locations in SSDL were selected based on the species distributions and not the actual observed coordinates. For large-range species, a proximity negative located at the opposite extreme of its distribution might as well be an absence negative, diluting its discriminative signal. Future work could explore strategies that leverage observed locations to select more representative negatives, potentially improving model accuracy.

More sophisticated species-specific strategies is another promising study subject. The method carried out for determining absence preference offers a baseline but lacks the depth of more advanced techniques. Moreover, other aspects of the procedure could benefit from species-level customization. For example, we decided to impose a a uniform proximity area of two H3 cell layers around the presence area for all species. This fixed design might not be optimal for certain species, such as species that habitat secluded islands, where proximity areas predominantly consist of ocean locations. Adjusting the proximity area based on species could prove beneficial to offer a more accurate geographical representation.

Bibliography

[1] Morgane Barbet-Massin, Frédéric Jiguet, Cécile Hélène Albert, and Wilfried Thuiller. Selecting pseudo-absences for species distribution models: How, where and how many? *Methods in ecology and evolution*, 3(2):327–338, 2012.

[2] Sara Beery, Elijah Cole, Joseph Parker, Pietro Perona, and Kevin Winner. Species distribution modeling for machine learning practitioners: A review. In *Proceedings of the 4th ACM SIGCAS Conference on Computing and Sustainable Societies*, pages 329–348, 2021.

[3] Christophe Botella, Alexis Joly, Pierre Bonnet, Pascal Monestiez, and François Munoz. A deep learning approach to species distribution modelling. *Multimedia tools and applications for environmental & biodiversity informatics*, pages 169–199, 2018.

[4] Christophe Botella, Alexis Joly, Pascal Monestiez, Pierre Bonnet, and François Munoz. Bias in presence-only niche models related to sampling effort and species niches: Lessons for background point selection. *PLoS One*, 15(5):e0232078, 2020.

[5] Mark S Boyce, Pierre R Vernier, Scott E Nielsen, and Fiona KA Schmiegelow. Evaluating resource selection functions. *Ecological modelling*, 157(2-3):281–300, 2002.

[6] Elijah Cole, Oisin Mac Aodha, Titouan Lorieul, Pietro Perona, Dan Morris, and Nebojsa Jojic. Multi-label learning from single positive labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 933–942, 2021.

[7] Elijah Cole, Grant Van Horn, Christian Lange, Alexander Shepard, Patrick Leary, Pietro Perona, Scott Loarie, and Oisin Mac Aodha. Spatial implicit neural representations for global-scale species mapping. In *International conference on machine learning*, pages 6320–6342. PMLR, 2023.

[8] D Richard Cutler, Thomas C Edwards Jr, Karen H Beard, Adele Cutler, Kyle T Hess, Jacob Gibson, and Joshua J Lawler. Random forests for classification in ecology. *Ecology*, 88(11):2783–2792, 2007.

[9] Jia Deng, Olga Russakovsky, Jonathan Krause, Michael S Bernstein, Alex Berg, and Li Fei-Fei. Scalable multi-label annotation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3099–3102, 2014.

[10] Rodolfo Dirzo, Hillary S Young, Mauro Galetti, Gerardo Ceballos, Nick JB Isaac, and Ben Collen. Defaunation in the anthropocene. *science*, 345(6195):401–406, 2014.

[11] Thibaut Durand, Nazanin Mehrasa, and Greg Mori. Learning a deep convnet for multi-label classification with partial labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 647–657, 2019.

[12] Jane Elith, Catherine Graham, Roozbeh Valavi, Meinrad Abegg, Caroline Bruce, Simon Ferrier, Andrew Ford, Antoine Guisan, Robert J Hijmans, Falk Huettmann, et al. Presence-only and presence-absence data for comparing species distribution modeling methods. *Biodiversity informatics*, 15(2):69–80, 2020.

[13] Jane Elith, Catherine H Graham, Robert P Anderson, Miroslav Dudík, Simon Ferrier, Antoine Guisan, Robert J Hijmans, Falk Huettmann, John R Leathwick, Anthony Lehmann, et al. Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, 29(2):129–151, 2006.

[14] Pedro Felzenszwalb, David McAllester, and Deva Ramanan. A discriminatively trained, multiscale, deformable part model. In *2008 IEEE conference on computer vision and pattern recognition*, pages 1–8. Ieee, 2008.

[15] Simon Ferrier, Graham Watson, Jennie Pearce, and Michael Drielsma. Extended statistical approaches to modelling spatial pattern in biodiversity in northeast new south wales. i. species-level modelling. *Biodiversity & Conservation*, 11(12):2275–2307, 2002.

[16] Paul C Fiedler, Jessica V Redfern, Karin A Forney, Daniel M Palacios, Corey Sheredy, Kristin Rasmussen, Ignacio García-Godos, Luis Santillán, Michael J Tetley, Fernando Félix, et al. Prediction of large whale distributions: a comparison

of presence–absence and presence-only modeling techniques. *Frontiers in Marine Science*, 5:419, 2018.

[17] Yoan Fourcade, Jan O Engler, Dennis Rödder, and Jean 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.

[18] Catherine H Graham and Robert J Hijmans. A comparison of methods for mapping species ranges and species richness. *Global Ecology and biogeography*, 15(6):578–587, 2006.

[19] Lauren Harrell, Christine Kaeser-Chen, Burcu Karagol Ayan, Keith Anderson, Michelangelo Conserva, Elise Kleeman, Maxim Neumann, Matt Overlan, Melissa Chapman, and Drew Purves. Heterogenous graph neural networks for species distribution modeling. *arXiv preprint arXiv:2503.11900*, 2025.

[20] David J Harris. Generating realistic assemblages with a joint species distribution model. *Methods in Ecology and Evolution*, 6(4):465–473, 2015.

[21] Tomislav Hengl, Jorge Mendes de Jesus, Gerard BM Heuvelink, Maria Ruiperez Gonzalez, Milan Kilibarda, Aleksandar Blagotić, Wei Shangguan, Marvin N Wright, Xiaoyuan Geng, Bernhard Bauer-Marschallinger, et al. Soilgrids250m: Global gridded soil information based on machine learning. *PLoS one*, 12(2):e0169748, 2017.

[22] Lionel R Hertzog, Aurélien Besnard, and Pierre Jay-Robert. Field validation shows bias-corrected pseudo-absence selection is the best method for predictive species-distribution modelling. *Diversity and distributions*, 20(12):1403–1413, 2014.

[23] Allen H Hurlbert and Walter Jetz. Species richness, hotspots, and the scale dependence of range maps in ecology and conservation. *Proceedings of the National Academy of Sciences*, 104(33):13384–13389, 2007.

[24] Allen H Hurlbert and Ethan P White. Disparity between range map-and survey-based analyses of species richness: patterns, processes and implications. *Ecology Letters*, 8(3):319–327, 2005.

[25] iNaturalist. inaturalist: Explore species observations. https://www.inaturalist.org/observations?view=species, 2024. Accessed: 2025-04-11.

[26] iNaturalist Team. Celebrating 100,000 modeled taxa with the inaturalist open range map dataset. https://www.inaturalist.org/blog/106918-celebrating-100-000-modeled-taxa-with-the-inaturalist-open-range-map-dataset, 2024. Accessed: 2025-04-11.

[27] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[28] Jorge M Lobo, Alberto Jiménez-Valverde, and Joaquín Hortal. The uncertain nature of absences and their importance in species distribution modelling. *Ecography*, 33(1):103–114, 2010.

[29] Oisin Mac Aodha, Elijah Cole, and Pietro Perona. Presence-only geographical priors for fine-grained image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9596–9606, 2019.

[30] Anna Norberg, Nerea Abrego, F Guillaume Blanchet, Frederick R Adler, Barbara J Anderson, Jani Anttila, Miguel B Araújo, Tad Dallas, David Dunson, Jane Elith, et al. A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels. *Ecological monographs*, 89(3):e01370, 2019.

[31] Convention on Biological Diversity. Global biodiversity outlook 5: Summary for policymakers, 2020. Accessed: 2025-04-13.

[32] Otso Ovaskainen, Gleb Tikhonov, Anna Norberg, F Guillaume Blanchet, Leo Duan, David Dunson, Tomas Roslin, and Nerea Abrego. How to make more out of community data? a conceptual framework and its implementation as models and software. *Ecology letters*, 20(5):561–576, 2017.

[33] Steven J Phillips, Robert P Anderson, and Robert E Schapire. Maximum entropy modeling of species geographic distributions. *Ecological modelling*, 190(3-4):231–259, 2006.

[34] Steven J Phillips, Miroslav Dudík, Jane Elith, Catherine H Graham, Anthony Lehmann, John Leathwick, and Simon Ferrier. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological applications*, 19(1):181–197, 2009.

[35] Hans-Otto Pörtner, Robert J Scholes, John Agard, R Leemans, Emma Archer, Xuemei Bai, David Barnes, Michael Burrows, Lena Chan, William Cheung, et al. Ipbes-ipcc co-sponsored workshop report on biodiversity and climate change. 2021.

[36] Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. *arXiv preprint arXiv:2010.04592*, 2020.

[37] SP Rushton, Stephen James Ormerod, and G Kerby. New paradigms for modelling species distributions? *Journal of applied ecology*, 41(2):193–200, 2004.

[38] Truly Santika. Assessing the effect of prevalence on the predictive performance of species distribution models using simulated data. *Global Ecology and Biogeography*, 20(1):181–192, 2011.

[39] Secretariat of the Convention on Biological Diversity. Strategic plan for biodiversity 2011–2020 and the aichi targets. https://www.cbd.int/doc/strategic-plan/2011-2020/aichi-targets-en.pdf, 2011. Accessed: 2025-08-12.

[40] Senait D Senay, Susan P Worner, and Takayoshi Ikeda. Novel three-step pseudo-absence selection technique for improved species distribution modelling. *PloS one*, 8(8):e71218, 2013.

[41] Ratha Sor, Young-Seuk Park, Pieter Boets, Peter LM Goethals, and Sovan Lek. Effects of species prevalence on the performance of predictive models. *Ecological Modelling*, 354:11–19, 2017.

[42] Geiziane Tessarolo, Jorge M Lobo, Thiago Fernando Rangel, and Joaquín Hortal. High uncertainty in the effects of data characteristics on the performance of species distribution models. *Ecological Indicators*, 121:107147, 2021.

[43] Uber Technologies. H3: A hexagonal hierarchical geospatial indexing system. https://h3geo.org/, 2023. Accessed: August 21, 2025.

[44] Roozbeh Valavi, Jane Elith, José J Lahoz-Monfort, and Gurutzeta Guillera-Arroita. Modelling species presence-only data with random forests. *Ecography*, 44(12):1731–1742, 2021.

[45] Thomas Verelst, Paul K Rubenstein, Marcin Eichner, Tinne Tuytelaars, and Maxim Berman. Spatial consistency loss for training multi-label classifiers from single-label annotations. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3879–3889, 2023.

[46] Robin Zbinden, Nina Van Tiel, Benjamin Kellenberger, Lloyd Hughes, and Devis Tuia. On the selection and effectiveness of pseudo-absences for species distribution modeling with deep learning. *Ecological Informatics*, 81:102623, 2024.

[47] Donghao Zhou, Pengfei Chen, Qiong Wang, Guangyong Chen, and Pheng-Ann Heng. Acknowledging the unknown for multi-label learning with single positive labels. In *European Conference on Computer Vision*, pages 423–440. Springer, 2022.

Appendix A

Network Architecture

Figure A.1 illustrates the location encoder architecture. The right side shows the standard linear layer and the four residual layers that make up the the network structure. While, the left side shows the structure of a single residual layer. Every layer includes 256 nodes and a dropout probability of 0.5 is used [29].

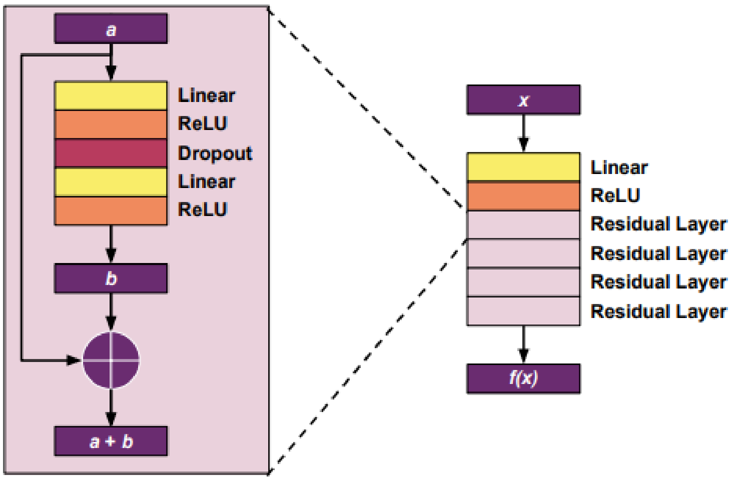


Figure A.1: Network diagram for the fully connected network (with residual connections) which we use for our location encoder f_θ. Image taken from [7].

Appendix B

H3 Global Grid System

H3 is a hierarchical grid system that partitions Earth's surface into hexagonal cells. The system hierarchy is achieved by different resolution levels, where every hexagonal cell has seven child cells of the next more detailed resolution [43]. This project uses the following functions included in the H3 Core Library:

`geo_to_h3(lat, lon, resolution):` Converts geographic coordinates to H3 cell index

`k_ring(cell, proximity_k):` Returns all cells within k distance of the given cell

`get_res0_indexes():` Gets all resolution 0 H3 indexes

`h3_to_children(cell, resolution):` Returns children cells at specified resolution

`h3_to_geo_boundary(cell, geo_json=True):` Gets boundary of H3 cell as coordinates

Appendix C

Estimation of Distribution Singularity with KL Divergence

This appendix provides detail about how KL divergence is computed to measure the singularity of species distributions. The KL divergence, also known as relative entropy, is a widely used metric of distance between two distributions. Mathematically is defined by:

$$D_{KL}(P \mid\mid Q) = \sum_{x \in X} P(x) ln \frac{P(x)}{Q(x)} \tag{C.1}$$

In our setting, P corresponds to the distribution of species observations, while Q corresponds to the distribution of the complete training dataset. The generation of these distributions and the KL computation is carried out following the next steps:

1. Estimate species distribution from training data, by counting for each species the number of observations within the H3 cells.

2. Estimate average distribution from training data, by counting the number of observations within the H3 cells for the whole training data.

3. Normalize species and average distribution so the density equals 1.

4. Calculate KL divergence for each species[1].

[1] The stats.entropy was used for this propose

Appendix D

Estimation of Preference for Absence

This appendix provides a detailed explanation for determining the preference for absence carried out in this work. The process can be summarized in the following steps:

1. Compute the difference in average precision (AP) for each species in the validation set (1,450 species) between the Full-Informed-Absence-Loss and Full-Informed-Proximity-Loss.

2. Apply a linear transformation to the difference in AP (ap_diff) to derive a preference for absence (p_abs):

   $$p\_abs = 0.5 \times (ap\_diff + 1)$$

   This mapping ensures assigning equal preference for absence and proximity if $ap\_diff = 0$, complete preference for absence if $ap\_diff = 1$ and complete preference for proximity if $ap\_diff = -1$.

3. Apply K-Means clustering to group species with similar input characteristics (number of presence cells and KL divergence) into a specified number of regions.

4. Calculate the mean p_abs within each cluster to establish an aggregated preference for absence

5. Store the resulting regions and their mean preferences in a dictionary, creating a lookup table for assigning preferences during training.

During model training, each species is assigned a preference for absence corresponding to the mean p_{abs} of the region it belongs to, determined by its number of presence cells and KL divergence. We evaluated a range of possible number of regions for clustering. Table D.1 reports the mean average precision (mAP) obtained for the validation set of 1,450 species when the different number of regions were specified. The analysis revealed that the best performance was attained with 7 regions, striking an optimal balance between granularity and generalization. The resulting regions are the ones shown in the right panel of Figure 6.3.

Number of Regions	**mAP (Validation)**
3	0.532
4	0.538
5	0.538
6	0.545
7	**0.555**
8	0.544
9	0.519

Table D.1: Mean Average Precision (mAP) on validation set for different numbers of regions in the preference estimation approach. Results are shown for 1,450 validation species across different region configurations.