Spatial Implicit Neural Representations for Global-Scale Species Mapping

Elijah Cole [1] **Grant Van Horn** [2] **Christian Lange** [3] **Alexander Shepard** [4] **Patrick Leary** [4] **Pietro Perona** [1] **Scott Loarie** [4] **Oisin Mac Aodha** [3]

Abstract

Estimating the geographical range of a species from sparse observations is a challenging and important geospatial prediction problem. Given a set of locations where a species has been observed, the goal is to build a model to predict whether the species is present or absent at any location. This problem has a long history in ecology, but traditional methods struggle to take advantage of emerging large-scale crowdsourced datasets which can include tens of millions of records for hundreds of thousands of species. In this work, we use Spatial Implicit Neural Representations (SINRs) to jointly estimate the geographical range of 47k species simultaneously. We find that our approach scales gracefully, making increasingly better predictions as we increase the number of species and the amount of data per species when training. To make this problem accessible to machine learning researchers, we provide four new benchmarks that measure different aspects of species range estimation and spatial representation learning. Using these benchmarks, we demonstrate that noisy and biased crowdsourced data can be combined with implicit neural representations to approximate expert-developed range maps for many species.

1. Introduction

We are currently observing a dramatic decline in global biodiversity, which has severe ramifications for natural resource management, food security, and ecosystem services that are crucial to human health (Watson et al., 2019; Rosenberg et al., 2019). In order to take effective conservation action we must understand species' ranges, i.e. where they live. However, we only have estimated ranges for a relatively small number of species in limited areas, many of which are already out of date by the time they are released.

The range of a species is typically estimated through *Species Distribution Modeling* (SDM) (Elith & Leathwick, 2009), the process of using species observation records to develop a statistical model for predicting whether a species is present or absent at any location. With enough *presence-absence* data (i.e. records of where a species has been confirmed to be present and absent) this problem can be approached using standard statistical learning methods (Beery et al., 2021).[1] However, presence-absence data is scarce due to the difficulty of verifying that a species is truly absent from an area. *Presence-only* data (i.e. verified observation locations, with no confirmed absences) is much more abundant as it is easier to collect. For instance, the community science platform iNaturalist (iNa) has collected over 141M presence-only observations to date across 429k species. Though presence-only data is not without drawbacks (Hastie & Fithian, 2013), it is important to develop methods that can take advantage of this vast supply of data.

Deep learning is one of our best tools for making use of large-scale datasets. Deep neural networks also have a key advantage over many existing SDM methods because they can *jointly* learn the distribution of many species in the same model (Chen et al., 2017; Tang et al., 2018; Mac Aodha et al., 2019). By learning representations that share information across species, the models can make improved predictions (Chen et al., 2017). However, the majority of current deep learning approaches need presence-absence data for training, which prevents them from scaling beyond the small number of species and regions for which sufficient presence-absence data is available.

Our work makes the following contributions:
(i) We show that implicit neural representations trained with noisy crowdsourced presence-only data can be used to estimate dense species' ranges. We call these models Spatial Implicit Neural Representations (SINRs).[2]

[1]Caltech [2]Cornell [3]University of Edinburgh [4]iNaturalist. Correspondence to: Elijah Cole <ecole@caltech.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

[1]The term "presence-absence" should not be taken to convey absolute certainty about whether a species is present or absent. False absences (i.e. non-detections) and, to a lesser extent, false presences are a serious concern in SDM (MacKenzie et al., 2002).

[2]We slightly abuse the terminology by using "SINR" to refer to both the model and the representation it parameterizes.

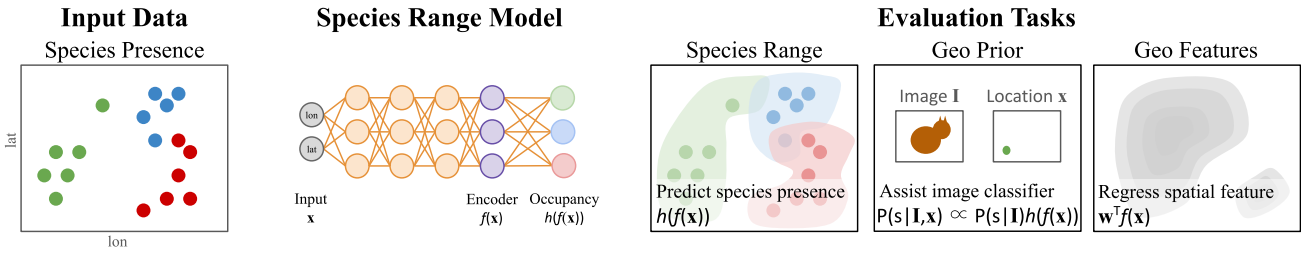


Figure 1. We show that sparse species observation data can be used to train Spatial Implicit Neural Representations (SINRs) which are transferable to other geospatial tasks. *(Left)* Here we show sparse, presence-only, spatial observations for three toy species (red, green, and blue). *(Middle)* The species observations are used to train a neural network that consists of a spatial feature encoder and per-species presence predictors. *(Right)* We evaluate on three diverse tasks: (i) estimating species ranges, (ii) assisting image classifiers using geographical range priors, and (iii) regressing geospatial features via our learned SINR.

(ii) We conduct a detailed investigation of loss functions for learning from presence-only data, their scaling properties, and the resulting geospatial representations.
(iii) We provide a suite of four geospatial benchmark tasks – ranging from species mapping to fine-grained image classification – which will facilitate future research on spatially sparse high-dimensional implicit neural representations, large-scale SDM, and geospatial representation learning.

Training and evaluation code is available at:
`https://github.com/elijahcole/sinr`

2. Related Work

Species distribution modeling (SDM) refers to a set of methods that aim to predict where (and sometimes when, and in what quantities) species of interest are likely to be found (Elith & Leathwick, 2009). The literature on SDM is vast. Readers interested in an overview should consult the review by Elith & Leathwick (2009) or the recent review of SDM for computer scientists by Beery et al. (2021). Note that we focus narrowly on the problem of predicting the occurrence of a species at a location, i.e. we do not consider more complex problems like trend or abundance estimation (Potts & Elith, 2006).

Traditional approaches to SDM train conventional supervised learning models (e.g. logistic regressors (Pearce & Ferrier, 2000), random forests (Cutler et al., 2007), etc.) to learn a mapping between hand-selected sets of environmental features (e.g. altitude, average rainfall, etc.) and species presence or absence (Phillips et al., 2004; Elith et al., 2006). Readers interested in these approaches should consult Norberg et al. (2019); Valavi et al. (2021; 2022), and the references therein. More recently, deep learning methods have been introduced that instead *jointly* represent multiple different species within the same model (Chen et al., 2017; Botella et al., 2018b; Tang et al., 2018; Mac Aodha et al., 2019; Teng et al., 2023). These models are typically trained on crowdsourced data, which can introduce additional challenges and biases that need to be accounted for during training (Fink et al., 2010; Chen & Gomes, 2019; Johnston et al., 2020; Botella et al., 2021). We build on the work of Mac Aodha et al. (2019), who proposed a neural network approach that forgoes the need for environmental features (as used by e.g. Botella et al. (2018b); Tang et al. (2018)) by learning to predict species presence from geographical location alone.

The problem of joint SDM with presence-only data can be viewed as an instance of multi-label classification with incomplete supervision. In particular, it is an example of Single Positive Multi-Label (SPML) learning (Cole et al., 2021; Verelst et al., 2023; Zhou et al., 2022). The goal is to train a model that is capable of making multi-label predictions at test time, despite having only ever observed one positive label per training instance (i.e. no confirmed negative training labels). Our work connects the SPML literature and SDM literature, and sets up large-scale joint species distribution modeling as a challenging real-world SPML task. This setting presents significant new difficulties for SPML, which has largely been limited to artificial label bias patterns (Arroyo et al., 2023) and relatively small label spaces (< 100 categories). Some SPML methods such as ROLE (Cole et al., 2021) are not computationally viable when the label space is large. One of our baselines is based on the SPML method of Zhou et al. (2022), which is scalable and obtains nearly state-of-the-art performance on the standard SPML benchmarks (Cole et al., 2021), but it is not a top performer on our new benchmark tasks.

Our work is related to the growing number of papers that use coordinate neural networks for implicitly representing images (Tancik et al., 2020) and 3D scenes (Sitzmann et al., 2019; Mildenhall et al., 2020). There are many design choices in these methods that are being actively studied, including the impact of the activation functions in the network (Sitzmann et al., 2019; Ramasinghe & Lucey, 2022) and the effect of different input encodings (Tancik et al., 2020; Zheng et al., 2022). In most research on implicit neural representations, there is an obvious choice of training objective, e.g. mean squared error between the predictions and the data. In the context of presence-only species estima-

tion, this choice is less clear. We systematically investigate this question in our experiments. Our benchmark also facilitates investigations of implicit neural representations with high-dimensional output spaces and sparse supervision.

Quantifying the performance of SDM at scale is notoriously difficult due to the fact that we lack confirmed presence-absence data for most species and locations (Beery et al., 2021). One approach is to evaluate performance on a small set of species from limited geographical regions where it is feasible to collect presence-absence data, as done in e.g. Potts & Elith (2006); Norberg et al. (2019); Valavi et al. (2022). Two of our evaluation tasks are larger-scale versions of this idea, in which we compare the performance of our models against expert range maps. An alternative evaluation approach is to measure the performance on a related "proxy" task. For example, there have been a number of works that use models trained for species range estimation to assist deep image classifiers (Berg et al., 2014; Tang et al., 2015; Mac Aodha et al., 2019; Chu et al., 2019; Mai et al., 2020; Terry et al., 2020; Skreta et al., 2020; Yang et al., 2022). By using images from platforms like iNaturalist, we can evaluate different range estimation methods on the task of aiding fine-grained image classification across tens of thousands of species. Finally, we also evaluate the spatial representations learned by our models via transfer learning, using them as inputs for a set of geospatial regression tasks. These complementary benchmark tasks capture different aspects of performance, and provide a starting point for large-scale SDM evaluation. See Figure 1 for an overview of our tasks.

3. Methods

3.1. Preliminaries

Problem statement. Let $\mathbf{x} = [lon, lat]$ denote a geographical location (i.e. longitude and latitude). Let $\mathbf{y} \in \{0, 1\}^S$ denote the true presence (1) or absence (0) of S different species at location $\mathbf{x}$. Following Cole et al. (2021), we introduce $\mathbf{z} \in \{0, 1, \varnothing\}^S$ to represent our observed data at $\mathbf{x}$, where $z_j = 1$ if species j is present, $z_j = 0$ if species j is absent, and $z_j = \varnothing$ if we do not know whether species j is present or absent. Our goal is to develop a model that produces an estimate of $\mathbf{y}$ at any location $\mathbf{x}$ over some spatial domain $\mathcal{X}$, given observed data $\{(\mathbf{x}_i, \mathbf{z}_i)\}_{i=1}^N$. We parameterize this model as $\hat{\boldsymbol{y}} = h_\phi(f_\theta(\boldsymbol{x}))$, where $f_\theta : \mathcal{X} \rightarrow \mathbb{R}^k$ is a location encoder with parameters θ and $h_\phi : \mathbb{R}^k \rightarrow [0, 1]^S$ is a multi-label classifier with parameters ϕ. The prediction $\hat{\boldsymbol{y}} \in [0, 1]^S$ is our estimate of how likely each species is to be present at $\mathbf{x}$.

Intuitively, the location encoder f_θ provides a representation of geographical space that is used by the multi-label classifier h_ϕ to predict species presence at each location. If θ is fixed or if f is a differentiable function of θ, then we can use standard methods like stochastic gradient descent to approximately solve

$$\theta^*, \phi^* = \text{argmin}_{\theta,\phi} \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(\hat{\mathbf{y}}_i, \mathbf{z}_i) \tag{1}$$

where $\hat{\boldsymbol{y}}_i = h_\phi(f_\theta(\boldsymbol{x}_i))$ and $\mathcal{L}$ is a suitably chosen loss function. Once trained, we say that $h_\phi \circ f_\theta$ has learned a Spatial Implicit Neural Representation (SINR) for the distribution of each species in the training set. Along the way we can learn f_θ, which produces a representation for any location on earth. See Figure 3 for visualizations of some of these geospatial representations.

Input encoding. Each species observation is associated with spatial coordinates $\mathbf{x} = [lon, lat]$. In practice, we rescale these values so that $lon, lat \in [-1, 1]$ and, following Mac Aodha et al. (2019), we guard against boundary effects using a sinusoidal encoding. The results is an input vector

$$\mathbf{x} = [\sin(\pi\ lon), \cos(\pi\ lon), \sin(\pi\ lat), \cos(\pi\ lat)]\ . \tag{2}$$

Alternative input encodings for related coordinate networks have been explored in the existing literature (Mai et al., 2020; Tancik et al., 2020; Mai et al., 2022; Zheng et al., 2022). This choice is orthogonal to the losses we explore, so we leave the evaluation of input encodings to future work.

Implicit neural representations. Traditionally, representation learning aims to transform complex objects (e.g. images, text) into simpler objects (e.g. low-dimensional vectors) that facilitate downstream tasks like classification or regression (Goodfellow et al., 2016). Implicit neural representations offer a different perspective, in which a signal is represented by a neural network that maps the signal domain (e.g. $\mathbb{R}$ for audio, $\mathbb{R}^2$ for images) to the signal values (Sitzmann et al., 2019; Tancik et al., 2020). In this work we learn implicit neural representations from a large collection of crowdsourced data containing observations of many species. This yields an implicit neural representation for the geospatial distribution of each species, as well as a representation for any location on earth.

Presence-absence vs. presence-only data. Species observation datasets come in two varieties: (i) *Presence-absence* data consists of locations where a species has been observed to be present and locations where it has been confirmed to be absent. That is, we say we have presence-absence data for species j if $|\{\mathbf{z}_i : z_{ij} = 0\}| > 0$ and $|\{\mathbf{z}_i : z_{ij} = 1\}| > 0$. Unfortunately, presence-absence data is costly to obtain at scale because confirming absence requires skilled observers to exhaustively search an area. (ii) *Presence-only* data is easier to acquire and thus more abundant because absences are not collected, i.e. $z_{ij} \in \{1, \varnothing\}$, for $i \in [N]$ and $j \in [S]$.

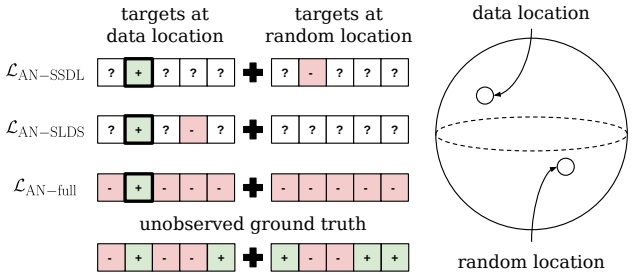


Figure 2. Illustration of the data used by three loss functions from Section 3.2. For each loss, we visualize the targets that the network is trained to predict. Each loss can be broken into two parts: one part that updates the network's predictions at the location of a training example (*data location*) and one part that updates the network's predictions at another location chosen randomly (*random location*). Each loss has access to one confirmed positive label (bold boxes). The rest of the labels are unobserved (non-bold boxes), and the losses make different, imperfect, assumptions about those unobserved labels.

3.2. Learning from Large-Scale Presence-Only Data

In the context of training SPML *image* classifiers, a simple but effective approach is to assume that unobserved labels are negative (Cole et al., 2021). This approach is based on a probabilistic argument: since natural images tend to contain a small number of categories compared to the size of the label set, the vast majority of the labels will be negative. This is also true for species distribution modeling. Given an arbitrary location and a large set of candidate species, nearly all of them will be absent. In this section we describe several simple and scalable loss functions based on this idea. We illustrate three of our losses in Figure 2.

"Assume negative" loss (same species, different location). As confirmed absences are not available in the presence-only setting, a common approach is to use randomly generated "pseudo-negatives" (Phillips et al., 2009). This first loss pairs each observation of a species with a pseudo-negative for that species at another location chosen uniformly at random:

$$\mathcal{L}_{\text{AN-SSDL}}(\hat{\mathbf{y}}, \mathbf{z}) = -\frac{1}{n_{\text{pos}}} \sum_{j=1}^{S} \mathbb{1}_{[z_j=1]} [\log(\hat{y}_j) + \log(1 - \hat{y}'_j)] \quad (3)$$

where $\hat{\mathbf{y}}' = h_\phi(f_\theta(\mathbf{r}))$ with $\mathbf{r} \sim \text{Uniform}(\mathcal{X})$ and $n_{\text{pos}} = \sum_{j=1}^{S} \mathbb{1}_{[z_j=1]}$. This approach generates pseudo-negatives (i.e. random absences) across the globe, but many of them are likely to be "easy" because they are far from the true species range.

"Assume negative" loss (same location, different species). This loss pairs each observation of a species with a pseudo-negative at the same location for a different species:

$$\mathcal{L}_{\text{AN-SLDS}}(\hat{\mathbf{y}}, \mathbf{z}) = -\frac{1}{n_{\text{pos}}} \sum_{j=1}^{S} \mathbb{1}_{[z_j=1]} [\log(\hat{y}_j) + \log(1 - \hat{y}_{j'})] \quad (4)$$

where $j' \sim \text{Uniform}(\{j : z_j \neq 1\})$. Intuitively, this approach generates pseudo-negatives that are aligned with the spatial distribution of the observed data.

Full "assume negative" loss. The previous two losses are inefficient in the sense that they do not use all of the entries in $\hat{\mathbf{y}}$. We can combine the pseudo-negative sampling strategies of $\mathcal{L}_{\text{AN-SSDL}}$ and $\mathcal{L}_{\text{AN-SLDS}}$ and use all available predictions as follows:

$$\mathcal{L}_{\text{AN-full}}(\hat{\mathbf{y}}, \mathbf{z}) = -\frac{1}{S} \sum_{j=1}^{S} [\mathbb{1}_{[z_j=1]} \lambda \log(\hat{y}_j) + \mathbb{1}_{[z_j \neq 1]} \log(1 - \hat{y}_j) + \log(1 - \hat{y}'_j)] \quad (5)$$

where $\hat{\mathbf{y}}' = h_\phi(f_\theta(\mathbf{r}))$ with $\mathbf{r} \sim \text{Unif}(\mathcal{X})$. The hyperparameter $\lambda > 0$ can be used to prevent the negative labels from dominating the loss. This is equivalent to the loss from Mac Aodha et al. (2019), but without their user modeling terms. Their version (including user modeling terms) is $\mathcal{L}_{\text{GP}}$ in Table 1 ("GP" = "Geo Prior").

Maximum entropy loss. Zhou et al. (2022) recently proposed a simple but effective and scalable technique for SPML image classification. Their approach encourages predictions for unobserved labels to maximize entropy instead of forcing them to zero like the "assume negative" approaches we have been discussing. We can apply this idea to $\mathcal{L}_{\text{AN-SSDL}}$, $\mathcal{L}_{\text{AN-SLDS}}$, and $\mathcal{L}_{\text{AN-full}}$ by replacing all terms of the form "$-\log(1-p)$" with terms of the form "$H(p)$", where $H(p) = -(p\log(p) + (1-p)\log(1-p))$ is the Bernoulli entropy. We write these "maximum entropy" (ME) variants as $\mathcal{L}_{\text{ME-SSDL}}$, $\mathcal{L}_{\text{ME-SLDS}}$, and $\mathcal{L}_{\text{ME-full}}$. (Zhou et al. (2022) also includes a pseudo-labeling component, but we omit this because Zhou et al. (2022) shows that it provides only a small improvement.)

4. Experiments

In this section we investigate the performance of SINR models on four species and environmental prediction tasks.

4.1. Models

As described in Section 3.1, our SINR models consist of a location encoder f_θ and a multi-label classifier h_ϕ which produce a vector of predictions $\hat{y} = h_\phi(f_\theta(\mathbf{x}))$ for a location $\mathbf{x}$. The location encoder f_θ is implemented as the fully connected neural network shown in Figure A3. We implement the multi-label classifier h_ϕ as a single fully connected

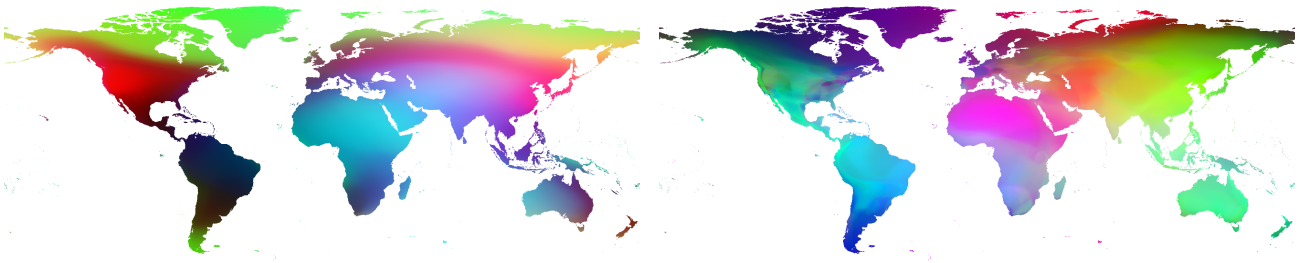


Figure 3. Visualization of the 256-dimensional features from learned location encoders f_θ projected to three dimensions using Independent Component Analysis (ICA). All models use the $\mathcal{L}_{\text{AN}-\text{full}}$ loss and take coordinates as input. *(Left)* This corresponds to a SINR model trained with a maximum of *10* examples per class. The features are smooth and do not appear to encode much high frequency spatial information. *(Right)* In contrast, the SINR model trained with a maximum of *1000* examples per class contains more high frequency information. The increase in training data appears to enable this model to better encode spatially varying environmental properties. Note, ICA is performed independently per-model, so similar colors do not indicate correspondence between the two images.

layer with sigmoid activations. For fair companions, we follow a similar architecture to Mac Aodha et al. (2019). Full implementation details can be found in Appendix C.

Besides SINR, we study two other model types. The first is logistic regression (Pearce & Ferrier, 2000), in which the location encoder f_θ is replaced with the identity function and h_ϕ is unchanged. Logistic regression is commonly used for SDM in the ecology literature. It also has the virtue of being highly scalable since it can be trained using GPU-accelerated batch-based optimization. The second type of non-SINR model is the discretized grid model. These models do not use a location encoder at all, but instead make predictions based on binning the training data (Berg et al., 2014). Full details for these models can be found in Appendix C. These baselines allow us to quantify the importance of the deep location encoder in our SINR models.

4.2. Training Data

We train our models on presence-only species observation data obtained from the community science platform iNaturalist (iNa). The training set consists of 35.5 million observations covering 47,375 species observed prior to 2022. Each species observation includes the geographical coordinate where the species was observed. We only included species in the training set if they had at least 50 observations. Some species are far more common than others, and thus the dataset is heavily imbalanced (see Figure A5). Later we use this data in its entirety during training ("All"), with different maximum observations per class ("X / Class"), or with different subsets of classes. See Appendix D for more details on the training dataset.

4.3. Evaluation Tasks and Metrics

We propose four tasks for evaluating large-scale species range estimation models. We give brief descriptions here, and provide further details in Appendix E.

S&T: eBird Status and Trends. This task quantifies the agreement between our presence-only predictions and expert-derived range maps from the *eBird Status & Trends* dataset (Fink et al., 2020), covering 535 bird species with a focus on North America. The spatial extent of this task is visualized in Figure A6. Performance is measured using mean average precision (MAP), i.e. computing the per-species average precision (AP) and averaging across species.

IUCN: Expert Range Maps. This task compares our predictions against expert range maps from the International Union for Conservation of Nature (IUCN) Red List (IUC). Unlike the bird-centric *S&T*, this task covers 2,418 species from different taxonomic groups, including birds, from all over the world. The spatial extent of this task is visualized in Figure A6. Performance is measured using MAP.

Geo Prior: Geographical Priors for Image Classification. This task measures the utility of our range maps as priors for fine-grained image classification (Berg et al., 2014; Mac Aodha et al., 2019). As illustrated in Figure 1, we combine the output of an image classifier with a range estimation model and measure the improvement in classification accuracy. The intuition is that an accurate range model can downweight the probability of a species if it is not typically found at the location where the image was taken. For this task we collect 282,974 images from iNaturalist, covering 39,444 species from our training set. Each image is accompanied by the latitude and longitude at which the image was taken. The performance metric for this task ("Δ Top-1") is the change in image classifier top-1 accuracy when using our range predictions as a geographical prior. Note that the geographical prior is applied to the classifier at test time – the image classifier is not trained with any geographical information. A positive value indicates that the prior improves classifier performance. Unlike *S&T* and *IUCN*, this is an *indirect* evaluation of range map quality since we assess how useful the range predictions are for a downstream task.

Table 1. Results for four geospatial tasks: **S&T** (eBird Status & Trends species mapping), **IUCN** (IUCN species mapping), **Geo Prior** (fine-grained image classification with a geographical prior), and **Geo Feature** (geographical feature regression). Tasks and metrics are defined in Section 4.3.We assess performance as a function of the loss function and the amount of training data ("# / Class"). Model inputs may be coordinates ("Coords."), environmental features ("Env.") or both ("Env. + Coords."). The logistic regression ("LR") and "Best Discretized Grid" baselines do not have an entry for the **Geo Feature** task as they do not learn a location encoder. We also do not evaluate models tagged with "Env." on the **Geo Feature** task because they are trained on closely related environmental features. Higher values are better for all tasks.

Loss	Model Type	# / Class	**S&T** (MAP)	**IUCN** (MAP)	**Geo Prior** (Δ Top-1)	**Geo Feature** (Mean R^2)
Baselines:						
N/A	Best Discretized Grid (Berg et al., 2014)	All	61.56	37.13	+4.1	-
$\mathcal{L}_{\text{AN}-\text{full}}$	LR (Pearce & Ferrier, 2000) - Coords.	1000	26.41	0.93	-0.6	-
$\mathcal{L}_{\text{AN}-\text{full}}$	LR (Pearce & Ferrier, 2000) - Env.	1000	32.91	1.23	-5.6	-
$\mathcal{L}_{\text{AN}-\text{full}}$	LR (Pearce & Ferrier, 2000) - Env. + Coords.	1000	35.42	1.11	-3.9	-
$\mathcal{L}_{\text{ME}-\text{SSDL}}$ (Zhou et al., 2022)	SINR - Coords.	1000	62.74	42.55	+1.6	0.726
$\mathcal{L}_{\text{ME}-\text{SLDS}}$ (Zhou et al., 2022)	SINR - Coords.	1000	74.37	32.22	+2.1	0.734
$\mathcal{L}_{\text{ME}-\text{full}}$ (Zhou et al., 2022)	SINR - Coords.	1000	73.61	58.60	+1.5	0.749
$\mathcal{L}_{\text{GP}}$ (Mac Aodha et al., 2019)	SINR - Coords.	1000	73.14	59.51	+5.2	0.724
$\mathcal{L}_{\text{AN}-\text{SSDL}}$	SINR - Coords.	10	51.12	27.63	+3.4	0.631
$\mathcal{L}_{\text{AN}-\text{SSDL}}$	SINR - Coords.	100	63.98	47.42	+4.7	0.721
$\mathcal{L}_{\text{AN}-\text{SSDL}}$	SINR - Coords.	1000	66.99	53.47	+4.9	0.744
$\mathcal{L}_{\text{AN}-\text{SSDL}}$	SINR - Coords.	All	68.36	55.75	+4.8	0.739
$\mathcal{L}_{\text{AN}-\text{SLDS}}$	SINR - Coords.	10	63.73	27.14	+4.6	0.693
$\mathcal{L}_{\text{AN}-\text{SLDS}}$	SINR - Coords.	100	72.18	38.40	+6.1	0.731
$\mathcal{L}_{\text{AN}-\text{SLDS}}$	SINR - Coords.	1000	76.19	42.26	+6.2	0.739
$\mathcal{L}_{\text{AN}-\text{SLDS}}$	SINR - Coords.	All	75.78	41.11	+6.1	0.748
$\mathcal{L}_{\text{AN}-\text{full}}$	SINR - Coords.	10	65.36	49.02	+4.3	0.712
$\mathcal{L}_{\text{AN}-\text{full}}$	SINR - Coords.	100	72.82	62.00	+6.6	0.736
$\mathcal{L}_{\text{AN}-\text{full}}$	SINR - Coords.	1000	77.15	65.84	+6.1	0.755
$\mathcal{L}_{\text{AN}-\text{full}}$	SINR - Coords.	All	77.94	65.59	+5.0	0.759
$\mathcal{L}_{\text{AN}-\text{full}}$	SINR - Env.	10	60.10	41.68	+3.8	-
$\mathcal{L}_{\text{AN}-\text{full}}$	SINR - Env.	100	74.54	66.64	+6.7	-
$\mathcal{L}_{\text{AN}-\text{full}}$	SINR - Env.	1000	79.65	70.54	+6.4	-
$\mathcal{L}_{\text{AN}-\text{full}}$	SINR - Env.	All	80.54	69.25	+5.3	-
$\mathcal{L}_{\text{AN}-\text{full}}$	SINR - Env. + Coords.	10	67.12	62.99	+4.7	-
$\mathcal{L}_{\text{AN}-\text{full}}$	SINR - Env. + Coords.	100	76.88	74.49	+6.8	-
$\mathcal{L}_{\text{AN}-\text{full}}$	SINR - Env. + Coords.	1000	80.48	76.07	+6.5	-
$\mathcal{L}_{\text{AN}-\text{full}}$	SINR - Env. + Coords.	All	81.39	74.67	+5.5	-

Geo Feature: Environmental Representation Learning. Instead of evaluating the species predictions, this transfer learning task evaluates the quality of the underlying geospatial representation learned by a SINR. The task is to predict nine different geospatial characteristics of the environment, e.g. above-ground carbon, elevation, etc. First, we use the location encoder f_θ to extract features for a grid of evenly spaced locations across the contiguous United States. After splitting the locations into train and test data, we use ridge regression to predict the geospatial characteristics from the extracted features. Performance is evaluated using the coefficient of determination R^2 on the test set, averaged across the nine geospatial characteristics.

4.4. Results

Which loss is best? No loss is best in every setting we consider. However, some losses do tend to perform better than others. In Table 1 we observe that, when we control for input type and the amount of training data, $\mathcal{L}_{\text{AN}-\text{full}}$ outperforms $\mathcal{L}_{\text{AN}-\text{SSDL}}$ and $\mathcal{L}_{\text{AN}-\text{SLDS}}$ most of the time. $\mathcal{L}_{\text{AN}-\text{full}}$ has a decisive advantage on the *S&T* and *IUCN* tasks and a consistent but small advantage on the *Geo Feature* task. Both $\mathcal{L}_{\text{AN}-\text{full}}$ and $\mathcal{L}_{\text{AN}-\text{SLDS}}$ perform well on the *Geo Prior* task, significantly outperforming $\mathcal{L}_{\text{AN}-\text{SSDL}}$. We note that $\mathcal{L}_{\text{AN}-\text{full}}$ is a simplified version of $\mathcal{L}_{\text{GP}}$ from Mac Aodha et al. (2019), but $\mathcal{L}_{\text{AN}-\text{full}}$ outperforms $\mathcal{L}_{\text{GP}}$ on every task.

Pseudo-negatives that follow the data distribution are usually better. $\mathcal{L}_{\text{AN}-\text{SSDL}}$ and $\mathcal{L}_{\text{AN}-\text{SLDS}}$ differ only in the fact that $\mathcal{L}_{\text{AN}-\text{SSDL}}$ samples pseudo-negatives from random locations while $\mathcal{L}_{\text{AN}-\text{SLDS}}$ samples pseudo-negatives from data locations (see Figure 2). In Table 1 we see that $\mathcal{L}_{\text{AN}-\text{SLDS}}$ outperforms $\mathcal{L}_{\text{AN}-\text{SSDL}}$ for all tasks except *IUCN*. This could be due to the fact that some *IUCN* species have ranges far from areas that are well-sampled by iNaturalist. As we can see in Figure A2 (Black Oystercatcher), $\mathcal{L}_{\text{AN}-\text{SSDL}}$ can behave poorly in areas with little training

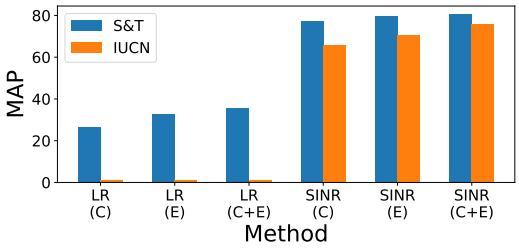


Figure 4. Results for the *S&T* and *IUCN* tasks. All models are trained with 1000 examples per class using the $\mathcal{L}_{\text{AN}-\text{full}}$ loss. We compare logistic regression ("LR") models against SINR models, using either coordinates (C), environmental covariates (E), or both (C+E) as inputs. These values can also be found in Table 1.

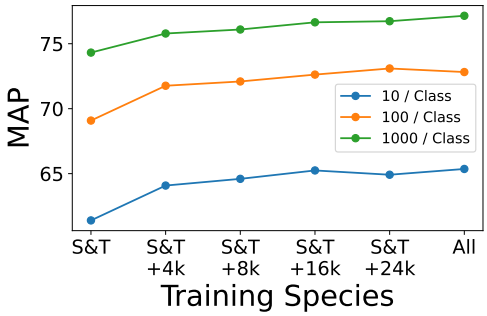


Figure 5. *S&T* task performance with $\mathcal{L}_{\text{AN}-\text{full}}$ as a function of the number of training examples per class (i.e. species) and number of classes. The horizontal axis gives the set of species used for training. "S&T" indicates that we only train on the 535 species in the S&T task. For "S&T + X" we add in X species chosen uniformly at random. For "All" we train on all 47k species. Note that the "10 / Class" point for "S&T" is trained with a higher learning rate than usual ($5e-3$ instead of $5e-4$) due to the small number of training examples per epoch. The values for "All" are also present in Table 1. All models use coordinates as input.

data. This highlights the importance of using diverse tasks to study range estimation methods.

Implicit neural representations significantly improve performance. We can assess the impact of the deep location encoder by comparing SINR and LR in models Table 1. For instance, if we use the $\mathcal{L}_{\text{AN}-\text{full}}$ loss with 1000 examples per class and coordinates as input, SINR outperforms LR by over 50 MAP on the *S&T* task. Both methods use the same inputs and training loss – the only difference is that SINR uses a deep location encoder while LR does not. Figure 4 shows that same pattern holds whether we use coordinates, environmental features, or both as inputs. For each input type, a deep location encoder provides significant benefits.

Environmental features are not necessary for good performance. In Figure 4 we show the *S&T* and *IUCN* performance of different models trained with coordinates only, environmental features only, or both. We see that SINR models trained with coordinates perform nearly as well as SINR models trained with environmental features. For the SINR models in Figure 4, coordinates are 97% as good as environmental features for the *S&T* task, 93% as good for the *IUCN* task, and 95% as good for the *Geo Prior* task. This suggests that SINRs can successfully use sparse presence-only data to learn about the environment, so that using environmental features as input provides only a marginal benefit.

Coordinates and environmental features are complementary. Figure 4 shows that it is better to use the concatenation of coordinates and environmental features than it is to use either coordinates or environmental features alone. This is true for LR and SINR. This indicates that the coordinates and environmental features are carrying some complementary information. However, as we discuss in Appendix B.2, environmental features introduce an additional layer of complexity compared to models that use only coordinates.

Joint learning across categories is beneficial, but more data is better. In Figure 5 we study the effect of the amount of training data on performance for the *S&T* task. We first note that, unsurprisingly, increasing the number of training examples per species reliably and significantly improves performance. One possible mechanism for this is suggested by Figure 3, which shows a more spatially detailed representation emerging with more training data. More interestingly, Figure 5 also shows that adding training data for additional species (which are not evaluated at test time) improves performance as well. That is, the model can better predict the distributions of the *S&T* birds by also learning the distributions of other birds, plants, insects, etc. Intuitively, it seems reasonable that training on more species could lead to a richer and more useful geospatial representation. However, the direct benefit of additional training data for the species of interest is far larger. If we were given a fixed budget of training examples to allocate among species as we wished, we should prefer to have a larger number of training examples per species (instead of fewer training examples per species, but spread across a greater number of species).

Low-shot performance is surprisingly good. In Table 1 we see that a SINR trained with $\mathcal{L}_{\text{AN}-\text{full}}$ and only 10 examples per category (i.e. $\sim$1% of the training data) beats the "Best Discretized Grid" baseline (which uses all of the training data) on every task. SINRs seem to be capable of capturing general spatial patterns using relatively little data. While this is encouraging, we expect that more data is necessary to capture fine detail as suggested by Figure 3 and Figure 7.

How are our tasks related? In this work we study four spatial prediction tasks. This tasks differ in their spatial domains, evaluation metrics, and categories of interest, but it is reasonable to wonder to what extent they may be related. In Figure 6 we show the pairwise correlations between scores on our tasks. Some tasks are highly correlated (e.g. *S&T* and *Geo Features*, 0.92) while others are not (e.g. *IUCN* and *Geo Prior*, 0.39).

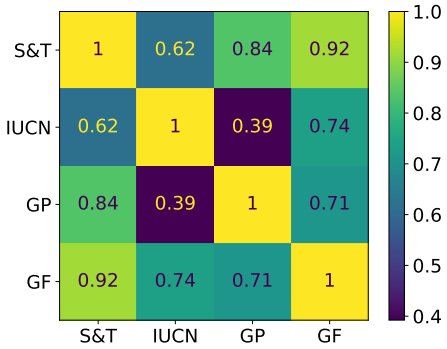


Figure 6. Performance correlations across our four tasks: *S&T*, *IUCN*, *Geo Prior* (GP), and *Geo Feature* (GF). Values are Pearson product-moment correlation coefficients. The correlations are computed across 12 SINR models: $\mathcal{L}_{\text{AN}-\text{SSDL}}$, $\mathcal{L}_{\text{AN}-\text{SLDS}}$, and $\mathcal{L}_{\text{AN}-\text{full}}$ for 10, 100, 1000, and All training examples per class. All models use coordinates as input.

Imbalance hurts performance, but not too much. In Table 1 we notice that a SINR trained with all of the training data often performs worse than a SINR trained on up to 1000 examples per class. This pattern is clearest for the *IUCN* and *Geo Prior* tasks. Capping the number of training examples per class reduces the amount of training data, but it also reduces class imbalance in the training set (some categories have as many as $\sim 10^5$ training examples). It seems that the benefit of reducing class imbalance outweighs the benefit of additional training data in these cases. However, it is important to keep in mind that the performance drops we are discussing are small. For instance, for a SINR trained with $\mathcal{L}_{\text{AN}-\text{full}}$ and coordinates as input, switching from 1000 training examples to all of the training data changes performance by -0.79 MAP for the *S&T* task, -0.25 MAP for the *IUCN* task, -1.1 Δ Top-1 for the *Geo Prior* task, and +0.004 for the *Geo Feature* task. Given the extreme imbalance in the training set and the fact that we do not explicitly handle class imbalance during training, it may be surprising that the performance drops are not larger.

Loss function rankings may not generalize across domains. The presence-only SDM problem in this work and the single positive image classification problem in Cole et al. (2021) are both SPML problems. Despite this formal equivalence, it does not seem that the best methods for SPML image classification are also the best methods for presence-only SDM. Zhou et al. (2022) show that their "maximum entropy" loss performs much better than the "assume negative" loss across a number of image classification datasets. However, all of the "maximum entropy" losses in Table 1 ($\mathcal{L}_{\text{ME}-\text{SSDL}}$, $\mathcal{L}_{\text{ME}-\text{SLDS}}$, $\mathcal{L}_{\text{ME}-\text{full}}$) underperform their "assume negative" counterparts ($\mathcal{L}_{\text{AN}-\text{SSDL}}$, $\mathcal{L}_{\text{AN}-\text{SLDS}}$, $\mathcal{L}_{\text{AN}-\text{full}}$). Thus, the benchmarks in this paper are complementary to those in Cole et al. (2021) and may be useful in developing a more holistic understanding of SPML learning.

4.5. Limitations

It is important to be aware of the limitations associated with our analysis. As noted, the training set is heavily imbalanced, both in terms of the species themselves and where the data was collected. In practice, some of the most biodiverse regions are underrepresented. This is partially because some species are more common and thus more likely to be observed than others by iNaturalist users. We do not explicitly deal with species imbalance in the training data, other than by showing that the ranking of methods does not significantly vary even when the training data for each species is capped to the same upper limit (see Table 1).

Reliably evaluating the performance of SDMs for many species and locations is a long standing challenge. To address this issue, we present a suite of complementary benchmarks that attempt to evaluate different facets of this spatial prediction problem. However, obtaining ground truth range data for thousands of species remains very difficult. While we believe our benchmarks to be a significant step forward, they are likely to have blind spots, e.g. they are limited to well-described species and can contain inaccuracies.

Finally, care should be taken before making conservation decisions based on the outputs of models such as the ones presented here. Our goal in this work is to demonstrate the promise of large-scale representation learning for species distribution modeling. Our models have not been calibrated or validated beyond the experiments illustrated above.

5. Conclusion

We explored the problem of species range mapping through the lens of learning spatial implicit neural representations (SINRs). In doing so, we connected recent work on implicit coordinate networks and learning multi-label classifiers from limited supervision. We hope our contributions encourage more machine learning researchers to work on this important problem. While the initial results are encouraging, there are many avenues for future work. For example, our models make no use of time (Mac Aodha et al., 2019), do not account for spatial bias (Chen & Gomes, 2019), and have no inductive biases for encoding spatially varying signals (Ramasinghe & Lucey, 2022).

Acknowledgments. We thank the iNaturalist and eBird communities for their data collection efforts, as well as Matt Stimas-Mackey and Sam Heinrich for help with data curation. This project was funded by the Climate Change AI Innovation Grants program, hosted by Climate Change AI with the support of the Quadrature Climate Foundation, Schmidt Futures, and the Canada Hub of Future Earth. This work was also supported by the Caltech Resnick Sustainability Institute and an NSF Graduate Research Fellowship (grant number DGE1745301).

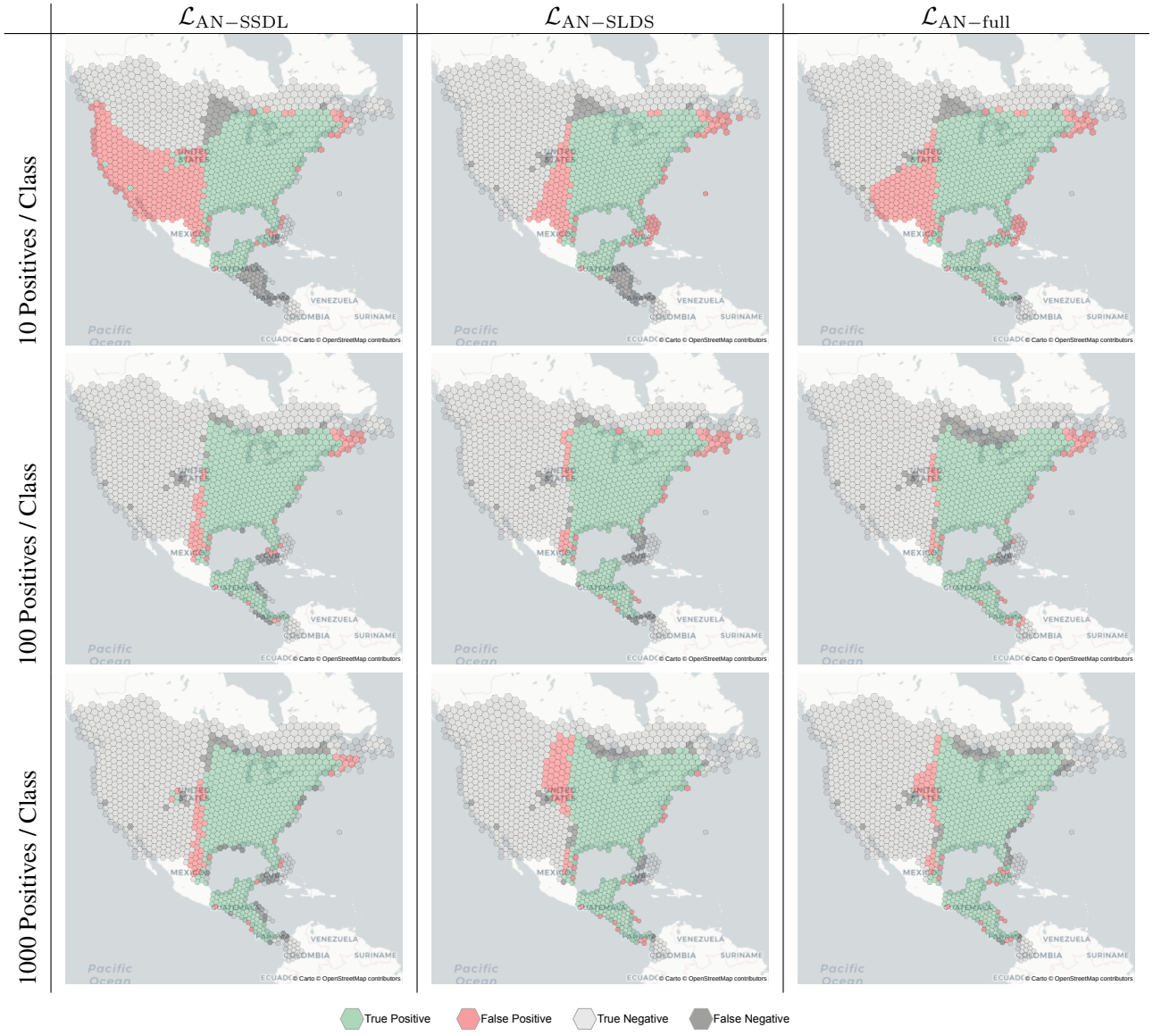


Figure 7. Visualization of SINR predictions for Wood Thrush when varying the amount of training data (rows) for different loss functions (columns). Model predictions are generated at the centroid of the rendered hexagons for a coarse H3 grid (resolution three), signifying locations where we can evaluate the model outputs for the *S&T* task. We convert the predictions to binary values using the threshold that maximizes the F1 score on the *S&T* data. This is done for each configuration independently. In practice this threshold would be chosen by a practitioner to meet particular project requirements. A model that matches the *S&T* task exactly would show only green and light grey hexagons. All models improve their range maps when given access to more data, as expected. $\mathcal{L}_{\text{AN}-\text{SSDL}}$ overestimates the western range extent and misses the southern extent with few examples, but refines these extents with additional data. $\mathcal{L}_{\text{AN}-\text{full}}$ starts off with most of the range covered (few "False Negative" hexagons) and proceeds to tighten the boundaries with more data. The range predicted by $\mathcal{L}_{\text{AN}-\text{SLDS}}$ is somewhere in between. All models use coordinates as input.

References

Birdlife international and handbook of the birds of the world (2022) bird species distribution maps of the world. version 2022.2. http://datazone.birdlife.org/species/requestdis, accessed 9 May 2023.

H3. https://h3geo.org/, accessed 9 May 2023.

IUCN 2022. The IUCN Red List of Threatened Species. 2022-2. https://www.iucnredlist.org, accessed 9 May 2023.

iNaturalist. www.inaturalist.org, accessed 9 May 2023.

Arroyo, J., Perona, P., and Cole, E. Understanding label bias in single positive multi-label learning. 2023.

Beery, S., Cole, E., Parker, J., Perona, P., and Winner, K. Species distribution modeling for machine learning practitioners: A review. In *Conference on Computing and Sustainable Societies*, 2021.

Berg, T., Liu, J., Woo Lee, S., Alexander, M. L., Jacobs, D. W., and Belhumeur, P. N. Birdsnap: Large-scale fine-grained visual categorization of birds. In *CVPR*, 2014.

Botella, C., Bonnet, P., Munoz, F., Monestiez, P. P., and Joly, A. Overview of geolifeclef 2018: location-based species recommendation. In *Working Notes of CLEF 2018-Conference and Labs of the Evaluation Forum*, 2018a.

Botella, C., Joly, A., Bonnet, P., Monestiez, P., and Munoz, F. A deep learning approach to species distribution modelling. In *Multimedia Tools and Applications for Environmental & Biodiversity Informatics*. 2018b.

Botella, C., Servajean, M., Bonnet, P., and Joly, A. Overview of geolifeclef 2019: plant species prediction using environment and animal occurrences. In *Working Notes of CLEF 2019-Conference and Labs of the Evaluation Forum*, 2019.

Botella, C., Joly, A., Bonnet, P., Munoz, F., and Monestiez, P. Jointly estimating spatial sampling effort and habitat suitability for multiple species from opportunistic presence-only data. *Methods in Ecology and Evolution*, 2021.

Chen, D. and Gomes, C. P. Bias reduction via end-to-end shift learning: Application to citizen science. In *AAAI*, 2019.

Chen, D., Xue, Y., Fink, D., Chen, S., and Gomes, C. P. Deep multi-species embedding. In *IJCAI*, 2017.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.

Chollet, F. Xception: Deep learning with depthwise separable convolutions. In *CVPR*, 2017.

Chu, G., Potetz, B., Wang, W., Howard, A., Song, Y., Brucher, F., Leung, T., and Adam, H. Geo-aware networks for fine-grained recognition. In *ICCV Workshops*, 2019.

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

Cole, E., Mac Aodha, O., Lorieul, T., Perona, P., Morris, D., and Jojic, N. Multi-label learning from single positive labels. In *CVPR*, 2021.

Collins, S. L. and Glenn, S. M. Importance of spatial and temporal dynamics in species regional abundance and distribution. *Ecology*, 72(2):654–664, 1991.

Cutler, D. R., Edwards Jr, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., and Lawler, J. J. Random forests for classification in ecology. *Ecology*, 2007.

Deneu, B., Lorieul, T., Cole, E., Servajean, M., Botella, C., Bonnet, P., and Joly, A. Overview of lifeclef location-based species prediction task 2020 (geolifeclef). CEUR-WS, 2020.

Domisch, S., Friedrichs, M., Hein, T., Borgwardt, F., Wetzig, A., Jähnig, S. C., and Langhans, S. D. Spatially explicit species distribution models: A missed opportunity in conservation planning? *Diversity and Distributions*, 2019.

Elith, J. and Leathwick, J. R. Species distribution models: ecological explanation and prediction across space and time. *Annual review of ecology, evolution, and systematics*, 2009.

Elith, J., H. Graham, C., P. Anderson, R., Dudík, M., Ferrier, S., Guisan, A., J. Hijmans, R., Huettmann, F., R. Leathwick, J., Lehmann, A., et al. Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, 2006.

Elith, J., Graham, C., Valavi, R., Abegg, M., Bruce, C., Ferrier, S., Ford, A., Guisan, A., Hijmans, R. J., Huettmann, F., et al. Presence-only and presence-absence data for comparing species distribution modeling methods. *Biodiversity informatics*, 2020.

Fick, S. E. and Hijmans, R. J. Worldclim 2: new 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, 2017.

Fink, D., Hochachka, W. M., Zuckerberg, B., Winkler, D. W., Shaby, B., Munson, M. A., Hooker, G., Riedewald, M., Sheldon, D., and Kelling, S. Spatiotemporal

exploratory models for broad-scale survey data. *Ecological Applications*, 2010.

Fink, D., Auer, T., Johnston, A., Strimas-Mackey, M., Robinson, O., Ligocki, S., Hochachka, W., Jaromczyk, L., Wood, C., Davies, I., Iliff, M., and Seitz, L. ebird status and trends, data version: 2020; released: 2021. *Cornell Lab of Ornithology, Ithaca, New York*, 10, 2020.

Goodfellow, I., Bengio, Y., and Courville, A. *Deep Learning*. MIT Press, 2016.

Guisan, A. and Rahbek, C. Sesam–a new framework integrating macroecological and species distribution models for predicting spatio-temporal patterns of species assemblages. *Journal of Biogeography*, 2011.

Hastie, T. and Fithian, W. Inference from presence-only data; the ongoing controversy. *Ecography*, 36(8):864–867, 2013.

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., and Jarvis, A. Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, 2005.

Johnston, A., Moran, N., Musgrove, A., Fink, D., and Baillie, S. R. Estimating species distributions from spatially biased citizen science data. *Ecological Modelling*, 2020.

Lorieul, T., Cole, E., Deneu, B., Servajean, M., Bonnet, P., and Joly, A. Overview of geolifeclef 2021: Predicting species distribution from 2 million remote sensing images. In *Working Notes of CLEF 2021-Conference and Labs of the Evaluation Forum*, 2021.

Lorieul, T., Cole, E., Deneu, B., Servajean, M., Bonnet, P., and Joly, A. Overview of geolifeclef 2022: Predicting species presence from multi-modal remote sensing, bioclimatic and pedologic data. In *CLEF 2022-Conference and Labs of the Evaluation Forum*, volume 3180, pp. 1940–1956, 2022.

Mac Aodha, O., Cole, E., and Perona, P. Presence-only geographical priors for fine-grained image classification. In *ICCV*, 2019.

MacKenzie, D. I., Nichols, J. D., Lachman, G. B., Droege, S., Andrew Royle, J., and Langtimm, C. A. Estimating site occupancy rates when detection probabilities are less than one. *Ecology*, 2002.

Mai, G., Janowicz, K., Yan, B., Zhu, R., Cai, L., and Lao, N. Multi-scale representation learning for spatial feature distributions using grid cells. In *ICLR*, 2020.

Mai, G., Xuan, Y., Zuo, W., Janowicz, K., and Lao, N. Sphere2vec: Multi-scale representation learning over a spherical surface for geospatial predictions. *arXiv:2201.10489*, 2022.

Martinez, J., Hossain, R., Romero, J., and Little, J. J. A simple yet effective baseline for 3d human pose estimation. In *ICCV*, 2017.

Merow, C., Smith, M. J., Edwards Jr, T. C., Guisan, A., McMahon, S. M., Normand, S., Thuiller, W., Wüest, R. O., Zimmermann, N. E., and Elith, J. What do we gain from simplicity versus complexity in species distribution models? *Ecography*, 37(12):1267–1281, 2014.

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., and Ng, R. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.

Norberg, A., Abrego, N., Blanchet, F. G., Adler, F. R., Anderson, B. J., Anttila, J., Araújo, M. B., Dallas, T., Dunson, D., Elith, J., et al. A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels. *Ecological monographs*, 2019.

Pearce, J. and Ferrier, S. An evaluation of alternative algorithms for fitting species distribution models using logistic regression. *Ecological modelling*, 2000.

Phillips, S. J., Dudík, M., and Schapire, R. E. A maximum entropy approach to species distribution modeling. In *ICML*, 2004.

Phillips, S. J., Dudík, M., Elith, J., Graham, C. H., Lehmann, A., Leathwick, J., and Ferrier, S. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological applications*, 2009.

Potts, J. M. and Elith, J. Comparing species abundance models. *Ecological modelling*, 2006.

Ramasinghe, S. and Lucey, S. Beyond periodicity: Towards a unifying framework for activations in coordinate-mlps. In *ECCV*, 2022.

Rolf, E., Proctor, J., Carleton, T., Bolliger, I., Shankar, V., Ishihara, M., Recht, B., and Hsiang, S. A generalizable and accessible approach to machine learning with global satellite imagery. *Nature communications*, 2021.

Rosenberg, K. V., Dokter, A. M., Blancher, P. J., Sauer, J. R., Smith, A. C., Smith, P. A., Stanton, J. C., Panjabi, A., Helft, L., Parr, M., et al. Decline of the north american avifauna. *Science*, 2019.

Sitzmann, V., Zollhöfer, M., and Wetzstein, G. Scene representation networks: Continuous 3d-structure-aware neural scene representations. *NeurIPS*, 2019.

Skreta, M., Luccioni, A., and Rolnick, D. Spatiotemporal features improve fine-grained butterfly image classification. In *Tackling Climate Change with Machine Learning Workshop at NeurIPS*, 2020.

Tancik, M., Srinivasan, P., Mildenhall, B., Fridovich-Keil, S., Raghavan, N., Singhal, U., Ramamoorthi, R., Barron, J., and Ng, R. Fourier features let networks learn high frequency functions in low dimensional domains. In *NeurIPS*, 2020.

Tang, K., Paluri, M., Fei-Fei, L., Fergus, R., and Bourdev, L. Improving image classification with location context. In *ICCV*, 2015.

Tang, L., Xue, Y., Chen, D., and Gomes, C. Multi-entity dependence learning with rich context via conditional variational auto-encoder. In *AAAI*, 2018.

Teng, M., Elmustafa, A., Akera, B., Larochelle, H., and Rolnick, D. Bird distribution modelling using remote sensing and citizen science data. *Tackling Climate Change with Machine Learning Workshop, ICLR*, 2023.

Terry, J. C. D., Roy, H. E., and August, T. A. Thinking like a naturalist: Enhancing computer vision of citizen science images by harnessing contextual data. *Methods in Ecology and Evolution*, 2020.

Valavi, R., Elith, J., Lahoz-Monfort, J. J., and Guillera-Arroita, G. Modelling species presence-only data with random forests. *Ecography*, 2021.

Valavi, R., Guillera-Arroita, G., Lahoz-Monfort, J. J., and Elith, J. Predictive performance of presence-only species distribution models: a benchmark study with reproducible code. *Ecological Monographs*, 2022.

Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., and Belongie, S. The inaturalist species classification and detection dataset. In *CVPR*, 2018.

Verelst, T., Rubenstein, P. K., Eichner, M., Tuytelaars, T., and Berman, M. Spatial consistency loss for training multi-label classifiers from single-label annotations. In *WACV*, 2023.

Watson, R., Baste, I., Larigauderie, A., Leadley, P., Pascual, U., Baptiste, B., Demissew, S., Dziba, L., Erpul, G., Fazel, A., et al. *Summary for policymakers of the global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services*. IPBES Secretariat, 2019.

Yang, L., Li, X., Song, R., Zhao, B., Tao, J., Zhou, S., Liang, J., and Yang, J. Dynamic mlp for fine-grained image classification by leveraging geographical and temporal information. In *CVPR*, 2022.

Zheng, J., Ramasinghe, S., Li, X., and Lucey, S. Trading positional complexity vs. deepness in coordinate networks. In *ECCV*, 2022.

Zhou, D., Chen, P., Wang, Q., Chen, G., and Heng, P.-A. Acknowledging the unknown for multi-label learning with single positive labels. In *ECCV*, 2022.

Appendix

A. Additional Results

A.1. How much does performance vary when we re-train a SINR?

The goal of this section is to provide a sense for how much variance in the performance of a SINR is due to randomness in the training process. We show *S&T* results for multiple independently trained SINRs in Figure A1. First, we observe that (as expected) performance varies more when training on 10 examples per class than it does when training on 100 or 1000 examples per class. Second, we note that deviation from the mean is typically less than 0.5 MAP and always less than 1.0 MAP. This provides some context for understanding whether a difference between two models is likely to be "real" or merely due to randomness.

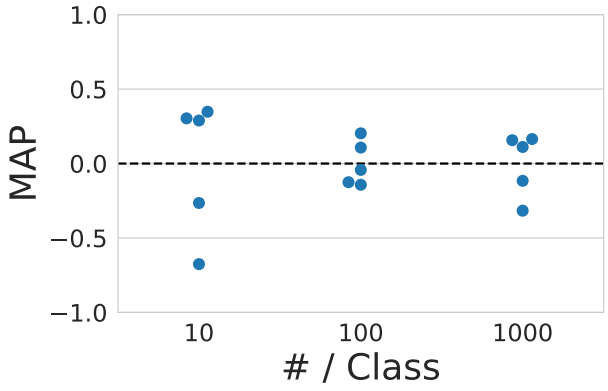


Figure A1. *S&T* results for SINRs trained with the $\mathcal{L}_{\mathrm{AN-full}}$ loss and varying amounts of training data. For each training data level, we show the mean-subtracted performance for 5 independent training runs. For this figure, the training examples selected for each class are re-sampled for each run. Thus, the randomness we see in this figure combines the randomness due to retraining and the randomness due to training data selection. Deviation from the mean is typically less the 0.5 MAP, and is always less than 1.0 MAP. All models use coordinates as input.

A.2. Additional Qualitative Results

To build some intuition for the behavior of $\mathcal{L}_{\mathrm{AN-SSDL}}$, $\mathcal{L}_{\mathrm{AN-SLDS}}$, and $\mathcal{L}_{\mathrm{AN-full}}$, we compare these losses on three species that are known to have interesting ranges in Figure A2.

B. Additional Discussion

B.1. How do the benchmark tasks proposed in this paper compare to existing SDM benchmarks?

Presence-only SDM is notoriously tricky to evaluate (Beery et al., 2021), and there are few public benchmark datasets available for the task. Here we will discuss the two most relevant lines of prior work that have approached this evaluation problem (one from the ecology community and one from the machine learning community), and discuss where our benchmark is similar and different.

To the best of our knowledge, Elith et al. (2006) was the first attempt to systematically compare presence-only SDM algorithms across many species and locations. That work compared 16 SDM algorithms on a collection of taxonomically-specific datasets from 6 different regions, covering a total of 226 species. Presence-only data was used for training and presence-absence data was used for evaluation. Unfortunately the data was not made publicly available until Elith et al. (2020). There are two main issues with this benchmark. First, the benchmark is not suitable for studying large-scale joint SDM. It has a small number of species overall, and there are at most 54 covered in any region. Second, the species in the dataset are anonymized. This makes it impossible to use their dataset to study large-scale SDM, because we cannot increase the size of their training with external data nor can we evaluate our trained models on their test data.

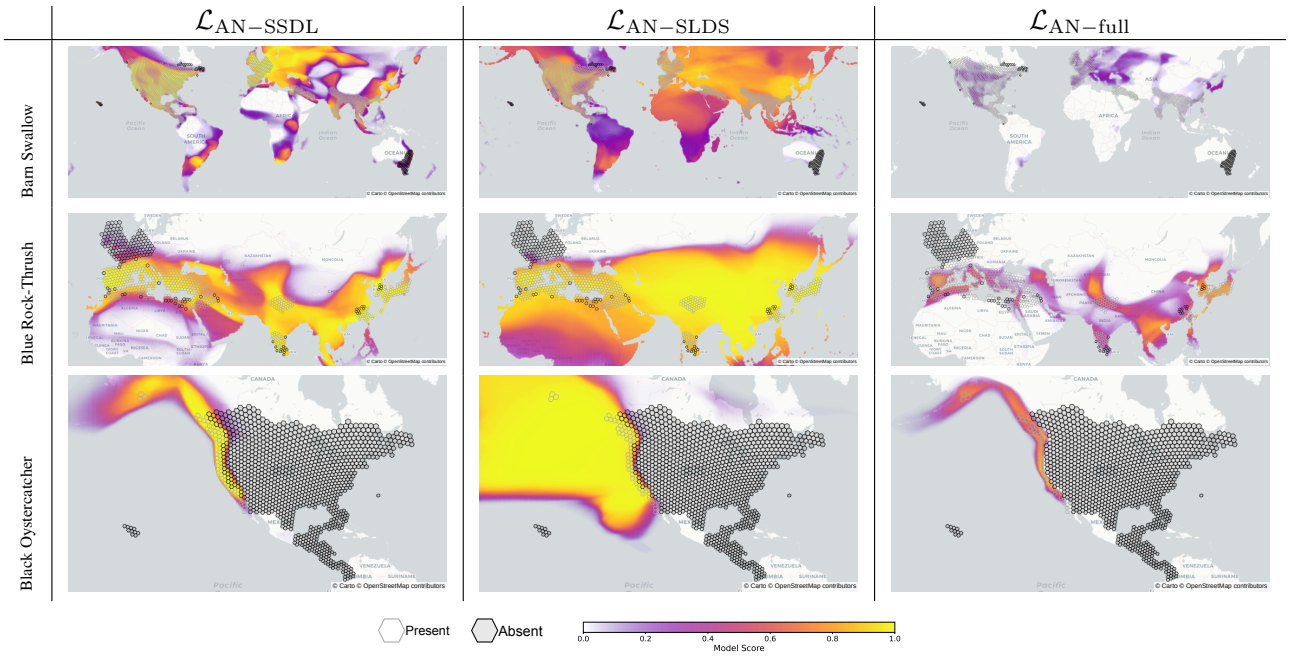


Figure A2. Loss function comparison (columns) for three different species of birds (rows). All models were trained with 1000 examples per class. See Figure 7 for an explanation of the plots. These species were chosen for visualization because their ranges have interesting complementary properties. (*Top Row*) Barn Swallow is a species that occurs across the globe. (*Middle Row*) Blue Rock-Thrush is a species whose range goes from the data rich *Western Palearctic biogeographic realm*, through a data sparse area, and back to a more data rich area of eastern Asia. (*Bottom Row*) Black Oystercatcher is a species whose range hugs the west coast of the United States. Unlike in other visualizations, these maps for Black Oystercatcher do not use a mask to filter out predictions from non-land regions. Here, we specifically wanted to see if the models learned to follow the coastline. We observe that $\mathcal{L}_{\text{AN}-\text{SLDS}}$ incorrectly expands the range into the Pacific. All models use coordinates as input.

Another line of work comes from the GeoLifeCLEF series of datasets and competitions (Botella et al., 2018a; 2019; Deneu et al., 2020; Cole et al., 2020; Lorieul et al., 2021; 2022). These benchmarks represent an attempt to scale up presence-only SDM, with the 2022 dataset covering 17k species with 1.6M species observations the U.S. and France. As in our benchmark, all of their training data is drawn from community science projects. The primary limitation of the GeoLifeCLEF benchmarks is that they use spatially biased presence-only data at test time, evaluating the problem as an information retrieval task instead of a spatial prediction task.

Our benchmark can be viewed as a significant expansion of the GeoLifeCLEF line of work. Instead of being geographically limited to France and the U.S., we allow data from anywhere in the world. (See Figure A6 for a visualization of the spatial coverage of the *S&T* and *IUCN* tasks.) Instead of evaluating with presence-only data, we use presence-absence data like Elith et al. (2020). However, unlike Elith et al. (2020), we work at a large scale that allows us to study data scaling in SDM. Our indirect evaluation tasks (*Geo Prior* and *Geo Feature*) add complementary dimensions to presence-absence evaluation, and have no counterpart in Cole et al. (2020) or Elith et al. (2020).

B.2. Environmental Covariates vs. Coordinates

One important characteristic of any SDM is whether or not it is *spatially explicit*. Spatially explicit SDMs include geospatial coordinates as part of the model input (Domisch et al., 2019). Traditional covariate-based SDMs include only environmental features (e.g. altitude, distance to roads, average temperature, etc.) in the input (Elith & Leathwick, 2009).

Covariate-based SDMs are often understood to reflect *habitat suitability*, because they learn a relationship between environmental characteristics and observed species occurrence patterns. A covariate-based SDM will make the same predictions for all locations with same covariates, even if those locations are on different continents. Furthermore, covariates sets must be selected by hand and are often limited in their spatial resolution and coverage.

By contrast, spatially explicit SDMs can model the fact that a species may be present in one location and absent in another,

even if those two locations have similar characteristics. However, spatially explicit models are unlikely to generalize to locations that are spatially distant from the training data – such locations are simply out of distribution.

SINRs trained with coordinates are spatially explicit, so our goal is not to learn from data in one location and extrapolate to distant locations. Instead, our goal is to use abundant (but noisy and biased) species observation data to approximate high-quality expert range maps. Our locations of interest are the same during training and testing. The difference is the training data source and quality. See Merow et al. (2014) for a more nuanced discussion of extrapolation vs. interpolation and the role of model complexity in SDM.

B.3. The Role of Time

Some species are immobile (e.g. trees), while others (e.g. birds) may occupy different areas at different times of the year. For this reason, there has long been interest in the temporal dynamics of species distributions (Collins & Glenn, 1991; Guisan & Rahbek, 2011). However, traditional SDMs use environmental features as input, which seldom include temporal structure (Elith et al., 2020; Norberg et al., 2019). For instance, the popular WorldClim bioclimatic variables used in many SDM papers are non-temporal (Hijmans et al., 2005). It is therefore not unusual for papers on SDM to make no explicit considerations for temporal information. Similarly, in this work we do not use temporal information during training or evaluation. However, we consider this to be an interesting area for future work.

C. Implementation Details

C.1. Network Architecture

We use the network in Figure A3 for our location encoder f_θ. This is identical to the architecture in Mac Aodha et al. (2019), and similar architectures have been used for other tasks (Martinez et al., 2017). The right side of the figure shows the network structure, consisting of one standard linear layer and four residual layers. The left side of the figure shows the structure of a single residual layer. Note that all layers have the same number of nodes. Every layer of the network has 256 nodes and we use $p = 0.5$ for the dropout probability.

C.2. Training Details

Environment. All models were trained on an Amazon AWS `p3.2xlarge` instance with a Tesla V100 GPU and 60 GB RAM. The model training code was written in PyTorch (v1.7.0).

Hyperparameters. All models were trained for 10 epochs using a batch size of 2048 and a learning rate of $5e-4$. We used the Adam optimizer with an exponential learning rate decay schedule of

$$\texttt{learning\_rate} = \texttt{initial\_learning\_rate} \times \texttt{epoch}^{0.98}$$

where $\texttt{epoch} \in \{0, 1, \ldots, 9\}$. For $\mathcal{L}_{\text{AN}-\text{full}}$ and $\mathcal{L}_{\text{GP}}$ we set $\lambda = 2048$.

C.3. Environmental Features

When environmental features are required for model inputs, we use the elevation and bioclimatic rasters from WorldClim 2.1 (Fick & Hijmans, 2017) at the 5 arc-minute spatial resolution. We normalize each covariate independently by subtracting the mean and dividing by standard deviation (ignoring `NaN` values). We then replace `NaN` values with zeros i.e. the new mean value.

C.4. Baselines

C.4.1. Logistic Regression

This section discusses our implementation of logistic regression with environmental covariates. The architecture for this approach is equivalent to a SINR but replacing the location encoder f_θ with the identity function. Then we can in principle use any of our loss functions for training. All other training details follow Appendix C.2.

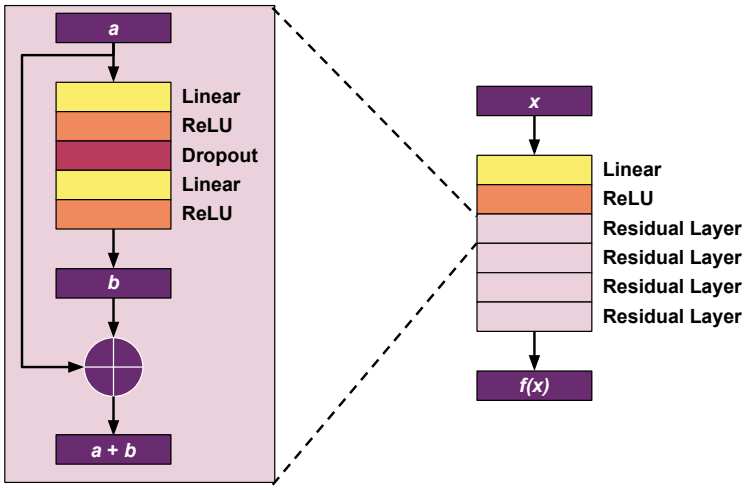


Figure A3. Network diagram for the fully connected network (with residual connections) which we use for our location encoder f_θ.

C.4.2. Discretized Grid

In this section we describe our discretized grid baseline for the *S&T*, *IUCN*, and *Geo Prior* tasks, which is a simple spatial binning method. Once we choose a resolution level, the H3 geospatial indexing library (H3W) defines a collection of W cells $\{H_1, \ldots, H_W\}$ that partition the globe. For instance, $W = 2{,}016{,}842$ at resolution level five. We show discretized grid results for a few different resolution choices in Table A1. Below we describe the discretized grid baseline in more detail.

For the *S&T* and *IUCN* tasks, we can compute a score for any hexagon and species as follows:

1. We compute the number of occurrences of species j in hex w as

$$n_{wj} = \sum_{i=1}^{N} \mathbb{1}_{[\mathbf{x}_i \in H_w]} \mathbb{1}_{[z_{ij}=1]} \tag{6}$$

   for $1 \leq w \leq W$ and $1 \leq j \leq S$.

2. Let H_t be a hexagon we wish to evaluate at test time. For any species $1 \leq j \leq S$, we compute a prediction for H_t as

$$\hat{y}_j = \frac{n_{tj}}{\max_{1 \leq w \leq W} n_{wj}}. \tag{7}$$

That is, $\hat{y}_j$ measures how often species j was observed in H_t (relative to how often species j occurred in the location where it was observed most often). These predictions always fall between 0 and 1, which ensures that they are compatible with the average precision metrics we use for *S&T* and *IUCN* evaluation.

For the *Geo Prior* task, the first step is the same but the second step is different:

1. We compute the number of occurrences of species j in hex w as

$$n_{wj} = \sum_{i=1}^{N} \mathbb{1}_{[\mathbf{x}_i \in H_w]} \mathbb{1}_{[z_{ij}=1]} \tag{8}$$

   for $1 \leq w \leq W$ and $1 \leq j \leq S$.

2. Let H_t be a hexagon we wish to evaluate at test time. For any species $1 \leq j \leq S$, we compute a prediction for H_t as

$$\hat{y}_j = \mathbb{1}_{[n_{wj} > 0]}. \tag{9}$$

That is, any species which were not observed in H_t are "ruled out" for the downstream image classification problem.

Table A1. Discretized Grid baseline results on test data when using various hexagon resolution for "training" the model. As this baseline does not learn a location encoder, it is not possible to evaluate on the *Geo Feature* task.

		Species Range	IUCN	Geo Prior
Hex Res	# / Cls.	MAP	MAP	Δ Top-1
0	All	54.67	21.89	3.5
1	All	61.56	37.13	4.1
2	All	61.03	36.92	3.1
3	All	51.09	26.57	-0.9

D. Training Dataset

D.1. Dataset Construction

Our training data was collected by the users of the community science platform iNaturalist (iNa). iNaturalist users take photographs of plants and animals, which they then upload to the platform. Other users review these images and attempt to identify the species. The final species labels are decided by the consensus of the community. Each species observation consists of an image and associated metadata indicating when, where, and by whom the observation was made. iNaturalist data only contains presence observations, i.e. we do not have access to any confirmed absences in our training data.

Specifically, our training data was sourced from the iNaturalist AWS Open Dataset[3] in May 2022. This public split of the data does not include location data for sensitive species if they are deemed to be threatened by location disclosure. We began by filtering the species observations according to the following rules:

1. Observations must have valid valid longitude and latitude data.
2. Observations must be identified to the *species* level by the iNaturalist community. Observations which can only be identified to coarser levels of specificity are discarded.
3. Observations must have *research grade* status, which indicates that there is a consensus from the iNaturalist community regarding their taxonomic identity.

After this filtering process, species with fewer than 50 observations were removed from the dataset. We also remove any species which are marked as *inactive*[4]. Finally, we only included observations made prior to 2022. This will enable a temporal split from 2022 onward to be used as a validation set in the future. After filtering, we were left with 35,500,262 valid observations from 47,375 distinct species. A visualization of the geographical distribution of the resulting data can be seen in Figure A4. As Figure A5 shows, our training data is heavily imbalanced, reflecting the natural frequency with which they are reported to iNaturalist (Van Horn et al., 2018).

D.2. Changing the Number of Training Examples per Category

In the main paper we consider the impact of the number of *observations* per species in the training set by training on different sub-sampled datasets. We construct these datasets by choosing k observations per species, uniformly at random. We set a seed to ensure that we are always using the same k observations per category. We also make certain that sampled datasets are nested, so the dataset with k_1 examples per category is a superset of the dataset with $k_2 < k_1$ examples per category. If a category has fewer than k observations, we use them all.

D.3. Changing the Number of Training Categories

In the main paper we also consider the impact of the number of *species* in the training set. In particular, we consider the following nested subsets:

- The set of 535 bird species in the eBird Status & Trends dataset (Fink et al., 2020).
- The eBird Status & Trends species plus an additional A randomly selected species, where $A \in \{4000, 8000, 16000, 24000\}$.

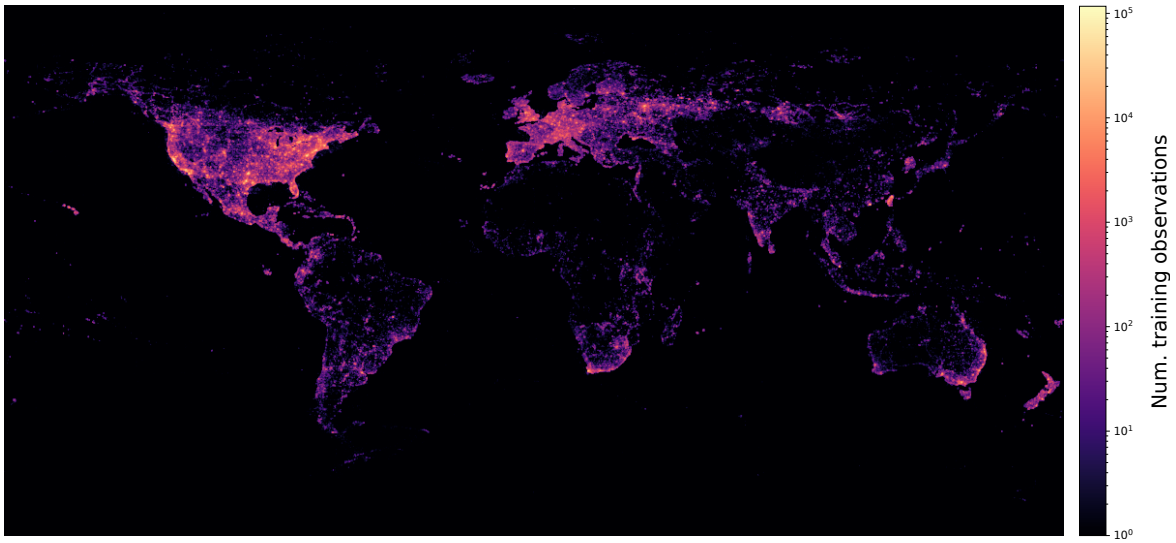


Figure A4. Histogram of the locations of the observations from our iNaturalist training set. Here we bin the data for all 35 million observations across all 47,375 species. Darker colors indicate fewer observations, brighter colors indicate more. The training data is biased towards North America, Europe, and New Zealand.

E. Evaluation Tasks

Here we provide additional details on the benchmark tasks used in the main paper. For each task, we outline the dataset properties, how it was collected, and the evaluation metrics used. In Figure A6 we visualize the spatial coverage of the *S&T* and *IUCN* tasks.

E.1. S&T: eBird Stats and Trends Range Maps

Task: The goal of this task is to evaluate the effectiveness of models trained on noisy crowdsourced data from iNaturalist for predicting species range maps. We use the *eBird Status & Trends* data from Fink et al. (2020) to evaluate our range predictions. This dataset consists of estimated relative abundance maps for 535 species of birds predominately found in North America, but also other regions. The relative abundance maps are computed at a spatial resolution of 3×3 km. The predictions are the output of an expert crafted model (Fink et al., 2020) that has been trained on tens of millions of presence and absence observations, makes use of additional expert knowledge to perform data filtering, and uses rich environmental covariates as input. While not without its own limitations, we treat this data as the ground truth for evaluation purposes as it is developed using much higher quality data and expert knowledge compared to what we use to train our models.

Dataset: We first download the rasterized abundance data for each species for all weeks of 2021 using the eBird API. We next reprojected each species' raster stack into latitude and longitude coordinates. We then spatially binned the data using H3 hexagons (i.e. cells) at resolution five[5]. 2,016,842 cells cover the world at this resolution, each with an average area of 252.9km^2. We finally sum all the relative abundance values for each cell, for each week of the year, for each species. Cells with nonzero values are considered present locations, cells with zero values are considered absent locations.

Our goal is to predict the presence or absence of a given species in each hexagon using the *eBird Status & Trends* output as the (psuedo) ground truth. The evaluation regions are restricted to those where the *eBird Status & Trends* models have determined that there is sufficient data to make a prediction for a given species. Thus, the set of evaluation regions can vary from species to species. For example, `Melozone aberti` has 127,270 locations with known presence or absence, of which 549 are deemed present. On the other hand, `Columba livia` has 499,406 locations with known presence or absence, of which 132,807 are deemed present.

The *eBird Status & Trends* data provides species presence and absence information for each location over the course of the

[3] https://github.com/inaturalist/inaturalist-open-data
[4] https://www.inaturalist.org/pages/how+taxon+changes+work
[5] https://h3geo.org

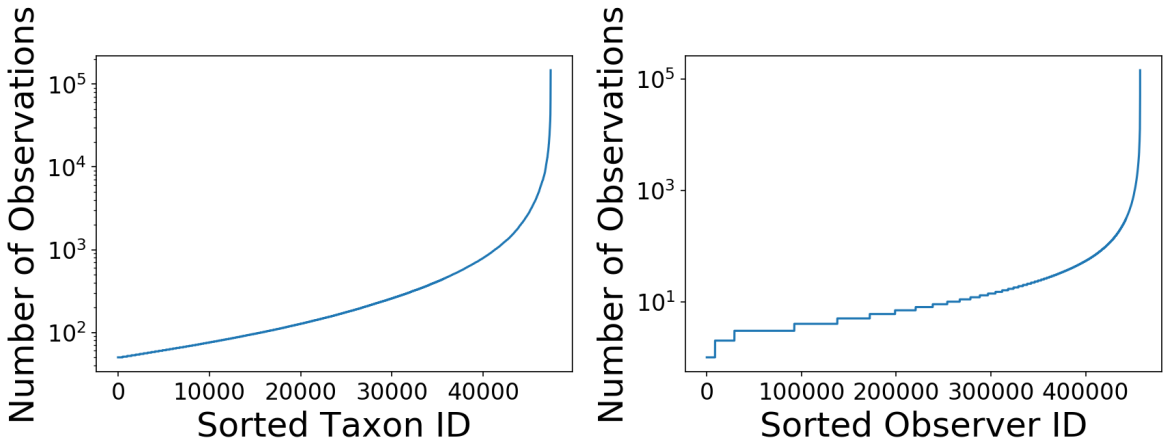


Figure A5. Summary statistics for our training observations data from iNaturalist. (*Left*) Distribution of observations over species. (*Right*) Distribution of observations over users (i.e. observers).

year. For the purposes of our evaluation, we collapse the time dimension and count a hexagon region as being a presence for a given bird if the output of their model is greater than zero for any week in the year for that species.

Evaluation: We use mean average precision (MAP) for evaluation, only evaluating on valid regions for a given species.

E.2. IUCN: Range Maps

Task: The goal of this task is similar to the previous one, i.e. to predict the geographical range of a set of species. However, instead of target range maps that are estimated by another model, here we use expert curated range maps (encoded as geospatial polygons) from the International Union for Conservation of Nature (IUCN) (IUC). This set of data contains a more taxonomically and geographically diverse set of species compared to the *Stats and Trends* task, as it contains mammals, reptiles, and amphibians, in addition to birds. The bird data in this task comes from BirdLife International (Bir). The IUCN data is from the '2022-2 update", last updated on the 9th of December 2022, and the BirdLife data is the "2022.2" version.

Dataset: Of the 47k species in our training set, we first exclude all species where more than 10% of the iNaturalist observations fall outside of the expert defined ranges and where there a taxonomic difference between IUCN and iNaturalist. This leaves 2,418 species that overlap with our training set. The data is contains 1,368 birds, 438 reptiles, 330 mammals, and 282 amphibians. Note our filtering cannot account for false positive regions from the IUCN data as we have no mechanism of extracting true absence from the iNaturalist source data.

Using the H3 geospatial indexing library (H3W), we sample all locations (i.e. latitude and longitude coordinates) at resolution five to determine if each location is contained within the IUCN range polygon(s) for a given species. This results in 2,016,842 locations for each species, where each location denotes the centroid of the corresponding H3 cell. Each location is either marked as a true presence (if the cell centroid is contained within an IUCN polygon(s)) or a true absence (if the cell is *not* contained within a polygon). Note, these expert range maps cannot necessarily be assumed to be the objective "ground truth" (i.e. species ranges can shift over time), but serve as strong proxy for it. A visualization of the expert provided ranges for a subset of species is shown in Figure A7.

Evaluation: As for the *S&T* task, we use mean average precision (MAP) as the evaluation metric, which results in a single score for a model averaged across all species.

E.3. Geo Prior: Geographical Priors for Image Classification

Task: The goal of this task is to combine the outputs of the models trained for species range estimation on the iNaturalist dataset with computer vision image classifier predictions. This evaluation protocol has also been explored in other work, e.g. Berg et al. (2014); Mac Aodha et al. (2019). We simply weight the probabilistic image classifier predictions for a given image with the species presence predictions from the location where that image was taken. The intuition is that the range prediction reduces the probability of a given species being predicted by the vision model if the range estimation model

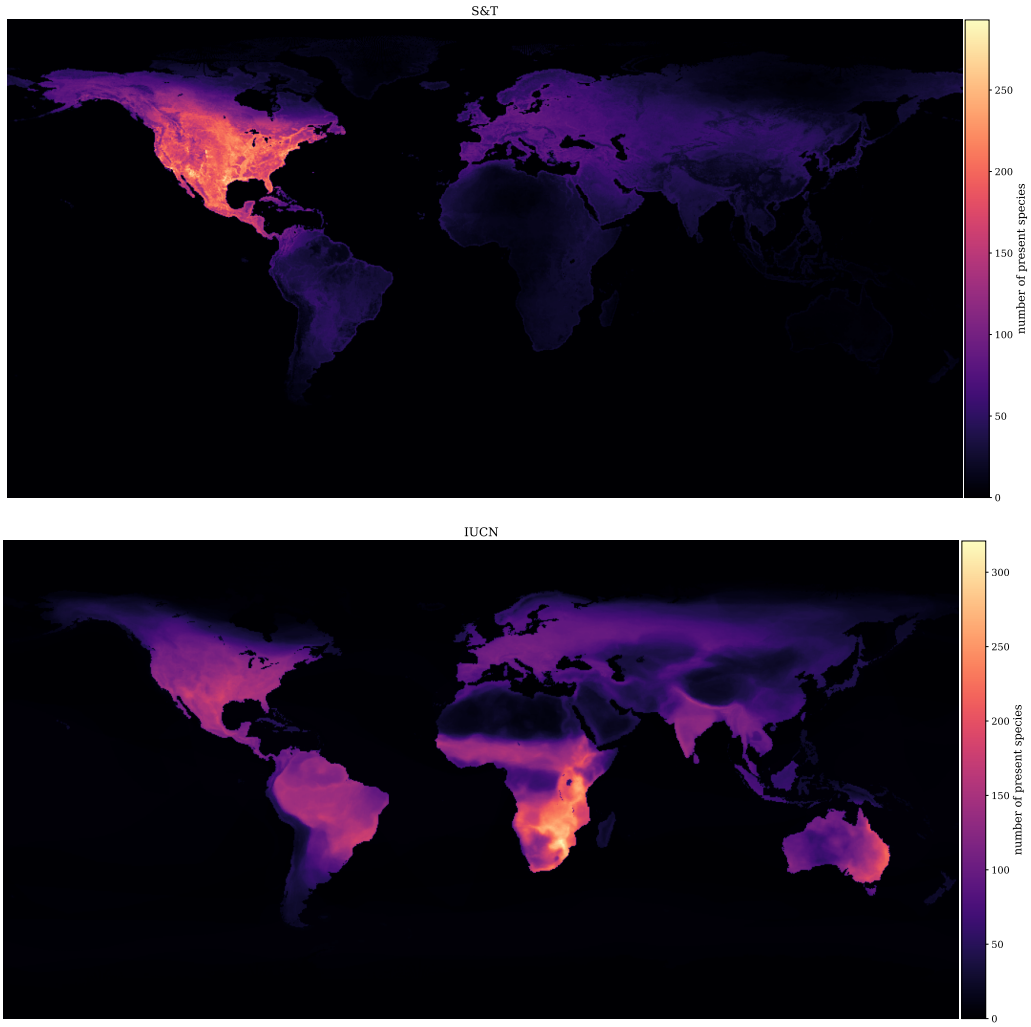


Figure A6. Visualization of the number of species present at different locations for the *S&T* task (top) and the *IUCN* task (bottom). Darker colors indicate fewer species, brighter colors indicate more species. The *IUCN* task has much broader coverage than the *S&T* task.

predicts that the species is *not* likely to be present at that location.

Dataset: For the vision classifier, we use an image classification model developed by the iNaturalist community science platform [6]. This model is an Xception network (Chollet, 2017) that has been trained on 55,000 different taxonomic entities (i.e. classes) from over 27 million images. We take the predictions from the final classification layer of the classifier, and do not apply any of their sophisticated taxonomic post-processing. There are a total of 49,333 species in the set of 55,000 classes – the others are higher levels in the taxonomy, e.g. genera. The images used to train the image classifier come from observations that were added to iNaturalist prior to December 2021.

We then constructed a test set consisting of all research grade observations (i.e. those observations for which there is a consensus from the iNaturalist community as to which species is present in the image). The images in the test set only contain the set of species that were observed at training time, i.e. we do not consider the open-set prediction problem. The observations were selected from between January and May 2022 to ensure that they did not overlap with the training set. We take at most ten observations per species, which results in 282,974 total observations from 39,444 species. In practice, many species do not have 10 observations. In total there are 2,721 species (with 9,808 total images) that are not present in our range estimation training set. For each of the 282,974 observations, we extract the predictions from the deep image classifier

[6] https://www.inaturalist.org/blog/63931-the-latest-computer-vision-model-updates

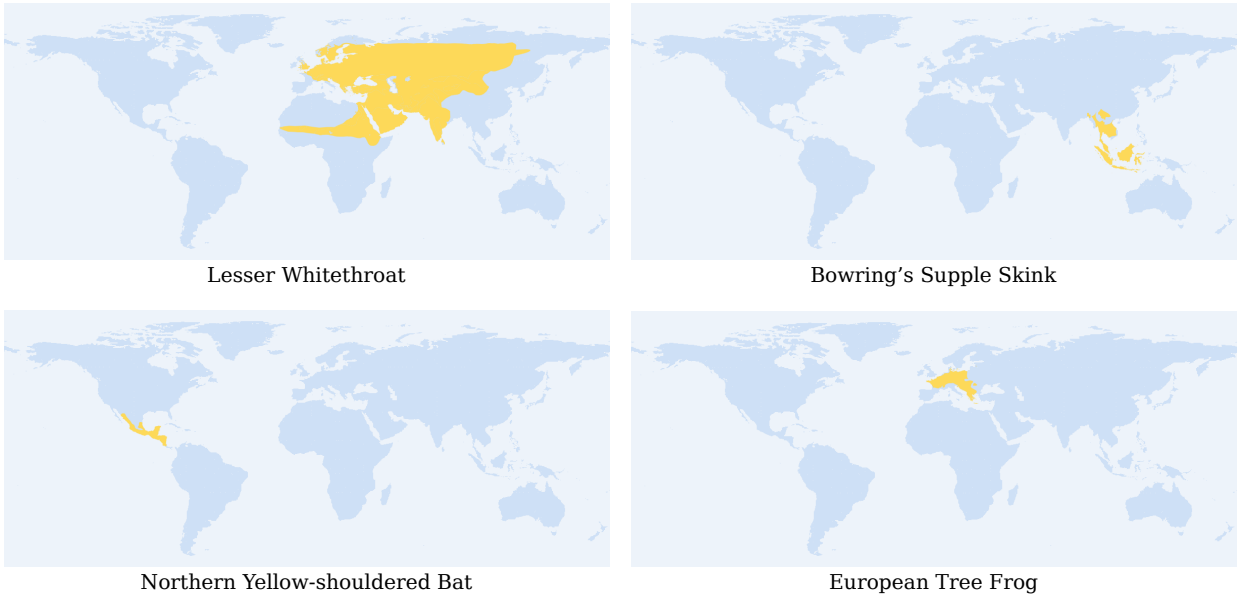


Figure A7. Expert defined ranges for four randomly selected species from our IUCN range evaluation task. The yellow regions indicate locations where the species is said to be present, otherwise they are absent. Light blue and darker blue indicate ocean and land respectively, and are only included for visualization purposes.

across all 39,444 remaining species.

Evaluation: Performance is evaluated in terms of top-1 accuracy, where the ground truth species label is provided by the iNaturalist community. Without using any information about where an image was taken, the computer vision model alone obtains an accuracy of 75.4%, which increases to 90.4% for top-5 accuracy. During evaluation, if a species is not present in a range model, we simply set the output for the range model for that species to 1.0.

E.4. Geo Feature: Environmental Representation Learning

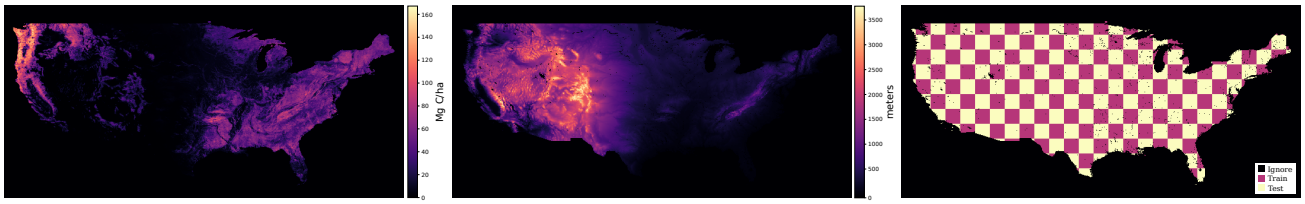
Task: This task aims to evaluate how well features extracted from deep models trained to perform species range estimation can generalize to *other* dense spatial prediction tasks. Unlike the other benchmark tasks that use the species occupancy outputs directly, this is a transfer learning task. We remove the classification head h_ϕ and evaluate the trained location encoder f_θ in terms of downstream environmental prediction tasks. The intuition is that a model that is effective at range estimation may have learned a good representation of the local environment. If so, that representation should be transferable to other environmental tasks with minimal adaptation.

This task is inspired by the linear evaluation protocol that is commonly used in self-supervised learning, e.g. Chen et al. (2020). In that setting, the features of the backbone model are frozen and a linear classifier is trained on them to evaluate how effective they are on various downstream classification tasks. In our case, instead of classification, we aim to *regress* various continuous environmental properties from the learned feature representations of our range estimation models. A related evaluation protocol was recently used in Rolf et al. (2021) for the case of evaluating models trained on remote sensing data.

Dataset: The task contains nine different environmental data layers which have been collected using Google Earth Engine [7]. The nine data layers are described in Table A2. For each of the layers, we have rasterized the data so that the entire globe is represented as a 2004×4008 pixel image. Each pixel represents the measured value for a given layer for the geographical region encompassed by the pixel. Example images can be seen in Figure A8.

Evaluation: For evaluation, we crop the region of interest to the contiguous United States and grid it into training and test cells. The spatial resolution of the training and testing cells are illustrated in Figure A8 (right). Note, we simply ignore locations that are not in the training or test sets, e.g. the ocean. This results in 51,422 training points and 50,140 test points. Features are then extracted from the location encoder for the spatial coordinates specified in the training split, and then a

[7] https://earthengine.google.com



AGC: Above Ground Carbon ELE: Elevation Train / Test Split

Figure A8. Here we illustrate two of the nine evaluation layers used in the *Geo Feature* prediction task (left and middle). On the right we indicate which regions contain pixels that are in the train or test split, or simply ignored during evaluation.

linear ridge regressor is trained on the train pixels and evaluated on the held out test pixels. The input features are normalized to the range $[0, 1]$. We cross validate the regularization weighting term α of the regressor on the training set, exploring the set $\alpha \in \{0.1, 1.0, 10.0\}$. Performance is reported as the coefficient of determination R^2 on the test pixels, averaged across all nine layers.

F. Reproducibility Statement

The information needed to implement and train the models outlined in this paper is provided in Appendix C. In addition, the different training losses we study are described in Section 3.2. Training and evaluation code is available at:

https://github.com/elijahcole/sinr

G. Ethics

This work makes use of species observation data provided by the iNaturalist community. We only use the public data exports from iNaturalist, ensuring that sensitive data (e.g. data related to species at risk of extinction) is not used by our models.

As noted in the limitations section in the main paper, extreme care must be taken when attempting to interpret any species range predictions from the models presented in this paper. Our work is intended to provide (i) a proof-of-concept for large-scale joint species distribution modeling with SINRs and (ii) benchmarks for further model development and analysis. However, our models have failure modes and our benchmarks have blind spots. Further validation is necessary before using these models for conservation planning or other consequential use cases.

Table A2. Description and sources of the nine environmental spatial layers that are part of our *Geo Feature* prediction task.

Name	Task	Units	Range
AGC	Above ground living biomass carbon stock density of combined woody and herbaceous cover in 2010. NASA/ORNL/biomass_carbon_density/v1 - agb	Mg C/ha	0 to 129
ELE	GMTED2010: Global multi-resolution terrain elevation data 2010. Masked to land only. USGS/GMTED2010 - be75	meters	-457 to 8746
LAI	The sum of the one-sided green leaf area per unit ground area. JAXA/GCOM-C/L3/LAND/LAI/V2 - LAI_AVE - 2020	(leaf area per ground area)	0 to 65531
NTV	Percent of a pixel which is covered by non-tree vegetation. JAXA/GCOM-C/L3/LAND/LAI/V2 - LAI_AVE - 20202	%	0 to 100
NOV	Percent of a pixel which is not vegetated. MODIS/006/MOD44B - Percent_NonVegetated	%	0 to 100
POD	UN adjusted estimated population density. CIESIN/GPWv411/GPW_UNWPP-Adjusted _Population_Density - unwpp-adjusted_population_density	# of persons / km^2	0 to 778120
SNC	Normalized difference snow index snow cover. MODIS/006/MOD10A1 - NDSI_Snow_Cover - 2019	(amount snow cover)	0 to 100
SOM	Soil moisture, derived using a one-dimensional soil water balance model. IDAHO_EPSCORTERRACLIMATE - soil - 2020	mm	0 to 8882
TRC	The percentage of pixel area covered by trees. NASA/MEASURES/GFCC/TC/v3 - tree_canopy_cover - 2000-2020	%	0 to 100