Feedforward Few-shot Species Range Estimation

Christian Lange [1] **Max Hamilton** [2] **Elijah Cole** [3] **Alexander Shepard** [4] **Samuel Heinrich** [5] **Angela Zhu** [2]
Subhransu Maji [2] **Grant Van Horn** [2] **Oisin Mac Aodha** [1]

Abstract

Knowing where a particular species can or cannot be found on Earth is crucial for ecological research and conservation efforts. By mapping the spatial ranges of all species, we would obtain deeper insights into how global biodiversity is affected by climate change and habitat loss. However, accurate range estimates are only available for a relatively small proportion of all known species. For the majority of the remaining species, we typically only have a small number of records denoting the spatial locations where they have previously been observed. We outline a new approach for few-shot species range estimation to address the challenge of accurately estimating the range of a species from limited data. During inference, our model takes a set of spatial locations as input, along with optional metadata such as text or an image, and outputs a species encoding that can be used to predict the range of a previously unseen species in a feedforward manner. We evaluate our approach on two challenging benchmarks, where we obtain state-of-the-art range estimation performance, in a fraction of the compute time, compared to recent alternative approaches.

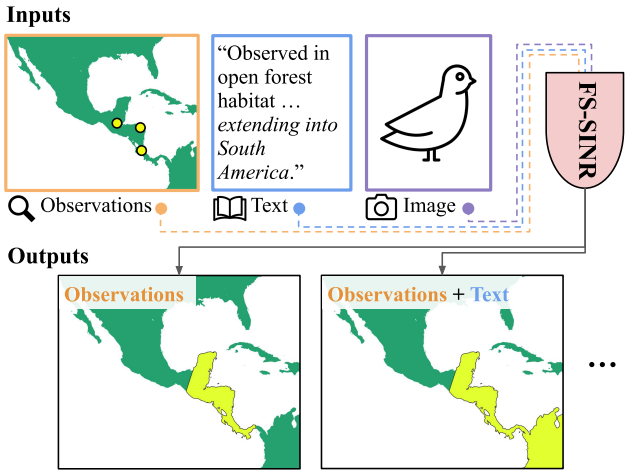


Figure 1. **Few-shot species range estimation with FS-SINR**. Our FS-SINR approach is trained on citizen science collected species observation data (i.e., locations where a species has been observed), and once trained, can estimate the spatial range of a previously *unseen* species with a single forward pass through the model, with no retraining required at inference time. It supports different input modalities such as variable length sequences of location observations, in addition to other metadata such as text or images. In this illustration, we show two different range predictions: one using only location observations (bottom left) and the other using observations and text (bottom right).

1. Introduction

Understanding the spatial distribution of plant and animal species is essential to mitigate the ongoing decline in global biodiversity (Jetz et al., 2019). Monitoring these distributions over time allows us to quantify the impacts of climate change, habitat loss, and conservation interventions (Mantyka-pringle et al., 2012). Estimating a species' spatial distribution typically starts with collecting a set of observations that denote the locations where the species has been confirmed to be present or absent. Traditionally, this data is used to train models that can then generate detailed predictions over a spatial region of interest (Elith et al., 2006; Beery et al., 2021). When sufficient data is available, these models enable practitioners to estimate important quantities such as the spatial range (i.e., where a species can be found) or abundance (i.e., the total number of individuals) of a species, in addition to quantifying how these quantities are changing over time.

Despite the availability of well-established modeling techniques, our current understanding of species' distributions is extremely limited as little or no observational data is available for most species. For example, iNaturalist, one of the largest citizen science platform documenting global biodiversity, has collected over 130 million research quality observations for approximately 373,000 species globally (iNaturalist, 2025). However, the data is severely long-tailed, i.e., a small percentage of common species account

[1]University of Edinburgh [2]UMass Amherst [3]GenBio AI [4]iNaturalist [5]Cornell. Correspondence to: Christian Lange <c.p.lange@sms.ed.ac.uk>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

for the majority of the observations, while many species have very few observations. In fact, over half of the 373,000 species cataloged by iNaturalist have been observed fewer than ten times to date. This data limitation is amplified by the fact that the vast majority of the several million species that are thought to exist have not yet even been documented by science (Mora et al., 2011). Identifying locations where under-observed species can be found is a time-consuming and laborious process, often requiring long expeditions to remote locations to search for species that are hard to find. Consequently, there is a pressing need for computational methods that can reliably estimate the spatial distributions of species using only a small number of observations.

Knowing the range of one species can help predict the range of another due to shared ecological, environmental, and geographic contexts. Recent advances in range estimation, such as Spatial Implicit Neural Representations (SINR), have leveraged this idea by training on millions of observations, across tens of thousands of species, inside one model (Cole et al., 2023). However, these models still rely on relatively large numbers of training observations for individual species, which limits their applicability to species with limited observations. In this work, we introduce Few-shot Spatial Implicit Neural Representations (FS-SINR), a novel Transformer-based model that overcomes this limitation and offers two key advantages over previous approaches. First, we obtain improved performance in the few-shot regime, a scenario that represents the reality for the majority of species, yet remains underexplored in prior work. Second, we make accurate predictions for species not present in the training set without any additional training, which can enable interactive exploration and modeling. At inference time, we only require a set of observed locations for the unseen species to generate reliable range estimates. Furthermore, we show we can flexibly incorporate additional non-geographic context information (e.g., a text summary of the species' habitat or range preferences or an image of the species) to further improve prediction quality. Figure 1 illustrates how FS-SINR can be used at inference time.

In summary, we make the following contributions: (i) We introduce FS-SINR, a new approach for few-shot species range estimation. FS-SINR has novel capabilities, including the ability to predict the spatial range of a previously unseen species at inference time without requiring any retraining. (ii) We demonstrate that FS-SINR achieves state-of-the-art performance in the few-shot setting on the challenging IUCN and S&T benchmark datasets. (iii) We provide detailed ablation studies and visualizations to highlight the benefits of integrating observational data with textual and visual context, as well as to compare our approach with alternative methods.

2. Related Work

Species Distribution Modeling. Estimating the spatial distribution of a species is a widely explored topic in both statistical ecology and machine learning (Beery et al., 2021). The goal is to develop models that can predict the distribution of species over space, and possibly time, given sparse observation data. Different machine learning approaches, initially using traditional techniques, such as decision trees among others have been extensively explored, e.g., (Phillips et al., 2004; Elith et al., 2006). More recently, deep learning-based methods have been introduced (Botella et al., 2018; Mac Aodha et al., 2019; Cole et al., 2023; Kellenberger et al., 2024). One of the strengths of these deep methods is that they can jointly represent thousands of different species within the same model and have been shown to improve as more training data is added, even when the data is from different species (Cole et al., 2023).

There has also been work investigating different approaches to address some of the challenges associated with training and evaluating these models. Examples include attempts to address imbalances across species in the training observation data (Zbinden et al., 2024b), sampling pseudo-absence data (Zbinden et al., 2024a), biases in training locations (Chen & Gomes, 2019), representing location information (Rußwurm et al., 2024), discretizing continuous model predictions (Dorm et al., 2024), active learning approaches (Lange et al., 2023), using additional metadata such as images (Teng et al., 2023; Dollinger et al., 2024; Picek et al., 2024) or text (Sastry et al., 2023; 2025; Hamilton et al., 2024), and designing new evaluation datasets to benchmark performance (Cole et al., 2023; Picek et al., 2024). In our work, we investigate the underexplored *few-shot* setting, where only limited observations (e.g., fewer than ten) are available for each species at training time.

Few-shot Species Range Estimation. There are several aspects of the species range estimation task in the low-data regime that make it different from other few-shot problems more commonly explored in the literature (Parnami & Lee, 2022; Wang et al., 2020). For example, the input domain is fixed (i.e., all locations on earth), each location can support more than one species (i.e., multi-label instead of multi-class), the label space is much larger (i.e., tens of thousands of species as opposed to hundreds of classes in image classification), and only partial supervision is available (e.g., presence-only data, with no confirmed absences).

Lange et al. (2023) introduced an active learning-based approach for species range estimation which makes predictions based on linear combinations of learned species embeddings and showed its effectiveness in the few-shot regime. LE-SINR (Hamilton et al., 2024) showed that internet sourced free-form text descriptions of species' ranges can be used when training models for zero-shot range esti-

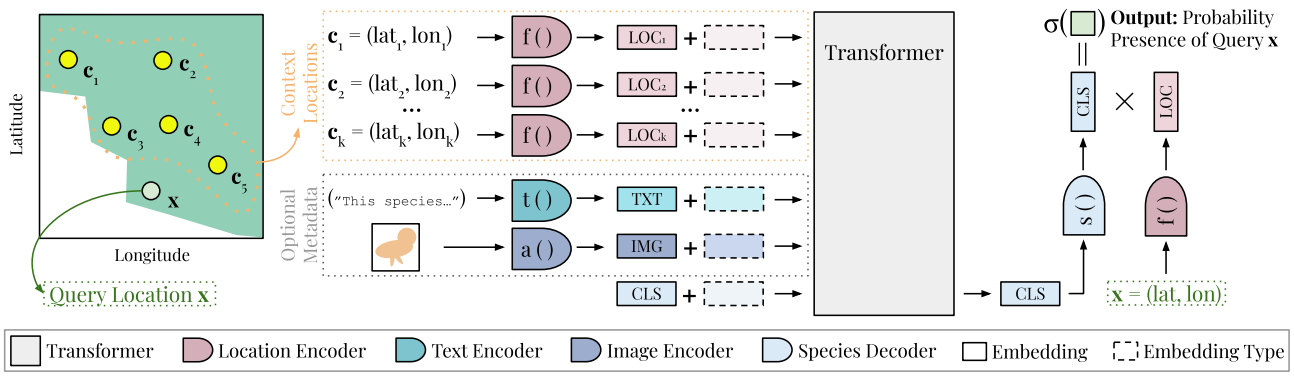


Figure 2. **FS-SINR overview**. Here we depict our few-shot species range estimation model. The input consists of an arbitrary number of context locations $\mathcal{C}^t$ for target species t that are each independently tokenized using a location encoder $f_{\boldsymbol{\theta}}()$, and optional auxiliary context information like text or an image. A class token (CLS) is also appended to the input. All input tokens are processed by a Transformer $m_{\boldsymbol{\psi}}()$. Given the set of input context locations, we estimate the probability that a species is present at a query location $\mathbf{x}$ by multiplying the location encoder's embedding of $\mathbf{x}$ with the projected embedding of the CLS token which is output from the species decoder.

mation. They applied their approach to the few-shot setting, but it requires retraining a classifier for each new species observation added. In our evaluation, we demonstrate that our FS-SINR approach, which can incorporate additional metadata at training time and does not require retraining during inference, outperforms existing methods.

3. Methods

We first set up the species range estimation problem and then describe our approach for few-shot range estimation.

3.1. Species Range Estimation

We start by describing the SINR approach from Cole et al. (2023). Let $\boldsymbol{x} = (\text{lat}, \text{lon}) \in \mathcal{X}$ be a location of interest sampled from a spatial domain $\mathcal{X}$ (e.g., a location on earth). Our goal is to train a model $g() : \mathcal{X} \rightarrow [0, 1]^s$ to predict the probabilities of s different species of interest occurring at $\boldsymbol{x}$. We let $\hat{\boldsymbol{y}} = g(\boldsymbol{x})$, where $\hat{y}_j \in [0, 1]$ (i.e., the j^{th} entry of $\hat{\boldsymbol{y}}$) represents the probability that species j occurs at location $\boldsymbol{x}$.

We can decompose the model as $g() = h_{\boldsymbol{\phi}}() \circ f_{\boldsymbol{\theta}}()$, where $f_{\boldsymbol{\theta}}() : \mathcal{X} \rightarrow \mathbb{R}^d$ is a location encoder with parameters $\boldsymbol{\theta}$ and $h_{\boldsymbol{\phi}}() : \mathbb{R}^d \rightarrow [0, 1]^s$ is a multi-label classifier with parameters $\boldsymbol{\phi}$. The location encoder $f_{\boldsymbol{\theta}}()$ maps a location $\boldsymbol{x}$ to a d-dimensional latent embedding $f_{\boldsymbol{\theta}}(\boldsymbol{x})$. The multi-label classifier $h()$ is implemented as a per-species linear projection followed by an element-wise sigmoid non-linearity, meaning that $\hat{\boldsymbol{y}} = \sigma(f_{\boldsymbol{\theta}}(\boldsymbol{x})\boldsymbol{W})$, where $\boldsymbol{W} \in \mathbb{R}^{d \times s}$ (i.e., $h_{\boldsymbol{\phi}}() = \boldsymbol{\phi} = \boldsymbol{W}$) and $\sigma()$ is the sigmoid function. Thus, each column vector $\boldsymbol{w}_j$ of $\boldsymbol{W}$ can be viewed as a species embedding, which we can combine with a location embedding $f_{\boldsymbol{\theta}}(\boldsymbol{x})$ via an inner product to compute the probability that the species j is present at $\boldsymbol{x}$. Importantly, the location embedding is shared across all species. Once trained, it is possible to generate a prediction for a given species for all locations of interest by evaluating the model for all locations (i.e., $\boldsymbol{x} \in \mathcal{X}$).

One of the main challenges associated with training models for species range estimation is that there is a dramatic asymmetry in the available training data. Specifically, it is much easier to collect presence observations (i.e., confirmed sightings of a species) compared to absence observations (i.e., confirmation that a species is not present at a specific location). As a result, many methods have been developed to train models using *presence-only* data. In the presence-only setting, we have access to training pairs $(\boldsymbol{x}, z)$, where $\boldsymbol{x}$ is a geographic location, and $z \in \{1, \ldots, s\}$ is an integer indicating which species was observed there. To overcome the lack of confirmed absence data, one common approach is to generate *pseudo-absences* by sampling random locations on the surface of the earth (Phillips et al., 2009). Given these pseudo-absences, the parameters of $g()$ can be trained in an end-to-end manner using variants of the cross-entropy loss. Specifically, we use *full assume negative loss* from Cole et al. (2023) to train the SINR baseline:

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

where z is the index of the species present for a given training instance, $\hat{y}_j$ is the predicted probability of the presence of species j, $\hat{y}'_j$ is the model prediction for a randomly sampled pseudo-absence location, and the hyperparameter λ balances the presence and pseudo-absence loss terms.

3.2. Few-shot Range Estimation

For the SINR model to make predictions for a new species, it is necessary to learn a new embedding vector $\boldsymbol{w}_j$ for that

species. If additional location data is later observed for that species, the model must be updated again. However, the number of observations for rarer species can be limited and thus it is necessary to have methods that can be updated efficiently with less training data.

We address this challenge by proposing a new approach for few-shot species range estimation called FS-SINR. Our model can predict the probability of presence for a previously unobserved species directly at inference time given only the set of confirmed presence locations available, without any retraining or parameter updates. At inference time, we assume we have access to a set of context locations $\mathcal{C}^t = \{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_k\}$, which represent a set of k locations where the species j has been confirmed to be present. Each entry in this set denotes a geographic location, i.e., $\boldsymbol{c} = (\text{lat}, \text{lon})$. Like SINR, our model is also conditioned on a location $\boldsymbol{x}$ of interest (i.e., the 'query' location), but uses the context locations to inform the prediction for the query location. Note, the context locations can come from a species not previously observed during training.

We represent our model as $g(\boldsymbol{x}) = m_{\boldsymbol{\psi}}(f_{\boldsymbol{\theta}}(\boldsymbol{x}), \mathcal{C}^t)$. Unlike in SINR, where the classifier head $h_{\boldsymbol{\phi}}()$ is a simple multi-label classifier and sigmoid non-linearity, in our case, the 'head' of the model $m_{\boldsymbol{\psi}}()$ is a Transformer-based encoder (Vaswani et al., 2017). FS-SINR takes an unordered set of context locations $\mathcal{C}^t$ as input, where each location is encoded into an embedding vector (i.e., a token) via a SINR-style multi-layer perceptron location encoder – see Figure 2 for an illustration. Importantly, our model can accept a variable number of context locations and is invariant to their ordering as we do not append any positional embeddings. This flexibility ensures that it can process a variable number of context locations during inference. We also append an additional register token (`REG`) as in Darcet et al. (2024) to provide the model with an additional token to 'store' information. Given that the input sequence is unordered and may or may not include additional context information, we add learned 'embedding type' vectors to each token such that the Transformer knows if a given input token is a location, register, text, image, etc.

We represent the species embedding vector (i.e., $\boldsymbol{w}_j$ in SINR) as the class token `CLS` of the Transformer after passing it through a small species decoder MLP $s()$. To make a final prediction, we simply compute the inner product between the location embedding of the query location $\boldsymbol{x}$ and the species embedding vector, and pass it through a sigmoid. Our approach is computationally efficient in that once the species embedding is generated it can then be efficiently multiplied by the embeddings for all locations of interest to generate a prediction for a species' range.

FS-SINR uses a similar training loss to $\mathcal{L}_{\text{AN}-\text{full}}$. However, since it has no equivalent to $h_{\boldsymbol{\phi}}()$ we cannot easily include all species in the loss, and instead consider only those within the same batch of training examples of size s^b. We obtain a predicted species embedding vector for a given species during the forward pass which can be used to estimate the probabilities of presence of that species for all locations sampled in the batch. We denote this new loss as $\mathcal{L}_{\text{AN}-\text{full}-\text{b}}$, which indicates that we are considering only those elements contained within the current batch b:

$$\mathcal{L}_{\text{AN-full-b}}(\hat{\boldsymbol{y}}, z^b) = -\frac{1}{s^b} \sum_{j=1}^{s^b} [\mathbb{1}_{[z^b=j]} \lambda \log(\hat{y}_j) + \mathbb{1}_{[z^b \neq j]} \log(1 - \hat{y}_j) + \log(1 - \hat{y}'_j)]. \quad (2)$$

3.2.1. Additional Context Information

The design of FS-SINR is flexible in that we can also provide additional context information to the model if it is available. For example, if there is additional text (e.g., a range description) or visual (i.e., images) information available for a novel species, it can be added to the context, assuming that such information was also available at training time for other species. This observation is inspired by recent work that also uses language-derived information to improve range predictions (Sastry et al., 2023; Hamilton et al., 2024) and work that uses species images and observations (Sastry et al., 2025). This additional information can provide a rich source of metadata encoding aspects of a species' habitat preferences, even when there might only be a limited number of location observations available for it. We can represent the expanded contextual input tokens as $\{\boldsymbol{t}_j, \boldsymbol{a}_j, f_{\boldsymbol{\theta}}(\boldsymbol{c}_1), \ldots, f_{\boldsymbol{\theta}}(\boldsymbol{c}_k)\}$, where $\boldsymbol{t}_j$ denotes a fixed-length text embedding from a large language model and $\boldsymbol{a}_j$ an image embedding obtained from a pre-trained vision model for species j – see Figure 2. Note that we train FS-SINR so that it can use arbitrary subsets, including none, of these input tokens during inference.

4. Experiments

Here we evaluate FS-SINR on the task of species range estimation and compare it to alternative methods.

4.1. Implementation Details

Architecture. Our location encoders use the same fully connected neural network with residual connections as in Cole et al. (2023). Each of the context locations is processed by the same shared location encoder which is first pre-trained as in SINR after which the multi-label classifier head is discarded. Importantly, this pre-trained encoder is only trained on species from the training set, and does not observe any data from the evaluation species during training. The text embedding backbone is a frozen GritLM (Muennighoff et al., 2025) and the default image embedding backbone is

a frozen EVA-02 ViT (Fang et al., 2024) pre-trained on the iNaturalist species image classification dataset (Van Horn et al., 2021). Both backbones provide a fixed length embedding vector, and we train two-layer fully connected text and image encoders to transform these embeddings into their context tokens. FS-SINR's Transformer contains four encoder layers and the parameters are updated jointly with the location, text, and image encoders and species decoder during training. In total, FS-SINR has 8.2M learnable parameters compared to 11.9M for SINR. This reduction is due to the fact that we do not have to learn a per-species embedding vector as in SINR. We train with a batch size of 2,048 instances and randomly drop-out text/image or location tokens during training with a probability of 0.5 and 0.1 respectively to enhance robustness. See Appendix C.1 for more details. Code for FS-SINR is available at: `https://github.com/Chris-lange/fs-sinr`

Data. We train FS-SINR on the presence-only dataset from Cole et al. (2023), which comprises 35.5 million citizen-science records—each annotated with latitude, longitude, and species label—for 47,375 diverse species including plants, fungi, and animals from the iNaturalist platform (iNaturalist, 2025). We also leverage 127 thousand text descriptions of these species used in Hamilton et al. (2024) and 200 thousand images obtained from iNaturalist (iNaturalist, 2025). The text provided during training is composed of sections of Wikipedia (Wikipedia, 2025) articles of the target species. During training, we supply FS-SINR with 20 context locations per training example, although we find that model performance is robust to changes in the number of context locations provided during training.

We evaluate models using the IUCN and S&T datasets also from Cole et al. (2023), which contain expert and model-derived range maps for 2,418 and 535 different species, respectively. The IUCN dataset is more globally distributed and contains a larger variation in range size and more diverse animal species, while the S&T dataset only contains bird species that are found primarily, but not always, in North America and have a larger average range size. We follow the same preprocessing steps for these datasets as in Cole et al. (2023). While not perfect, these datasets represent the best evaluation data currently available and contain large variety in terms of range sizes and locations. The text used during evaluation consists of pre-trained large language model generated summaries of the range or habitat of the target species as used in Hamilton et al. (2024). Importantly, we hold out any species from the union of these two datasets from the training set so that species from the evaluation set are not observed during training. As a result, by default, FS-SINR is trained on data from 44,422 species. Performance is reported as mean average precision (MAP) for different numbers of input (i.e., context) locations.

Baselines. Generating a species' range from FS-SINR for a held-out species at inference time only requires a single forward pass through the model to obtain an embedding vector for the species. Current methods (e.g., LE-SINR or SINR) cannot be used in such a feedforward manner and need to be retrained for each species that was not observed at training time. To obtain an equivalent embedding for the SINR and LE-SINR baselines we train a per-species binary logistic regression classifier using any few-shot presence observations that are available, in addition to adding 10,000 uniformly random and 10,000 target (i.e., in locations where species are) pseudo-absences as in LE-SINR. We also compare to the species embedding combination method from Lange et al. (2023) and a Prototypical Network-style baseline (Snell et al., 2017), denoted Active SINR and Prototype SINR, respectively. These baselines do not require retraining. For fairness, we use the same presence observations across each method, and the larger number of presences are supersets of the smaller ones. Implementation details of the baseline methods can be found in Appendix C.2.

4.2. Few-shot Evaluation

First, we evaluate how effective different range estimation models are at few-shot range estimation. The goal for each model is to generate a plausible prediction for a previously unseen species' range given limited location observations. Quantitative results are presented in Figure 3, and additional results can be found in Appendix A.

The SINR baseline performs poorly in the low-data regime, but as more data is added performance improves. As noted earlier, here a per-species embedding vector is learned using logistic regression using the provided presence locations and generated pseudo-absences. The recently introduced LE-SINR approach extends the basic SINR model to use text information (here range text), when available, at inference time. LE-SINR tends to outperform SINR, particularly when text data is available. Like our FS-SINR approach, neither the Active SINR or Prototype SINR baselines require retraining at inference time, but perform much worse than FS-SINR.

In all instances, when the same metadata is available, FS-SINR outperforms existing methods. Furthermore, we also outperform SINR in the larger data regime (i.e., when 50 observations are available). In general, we observe that image information is not as informative as text and does not help on average outside of the zero-shot case. A range text description provides much more context than an image of a previously unseen species. However, range estimates for some species benefit significantly from using images, but other species actually see no change or decreased performance. Making use of images to improve zero-shot range estimates is rewarded during training, but this can harm

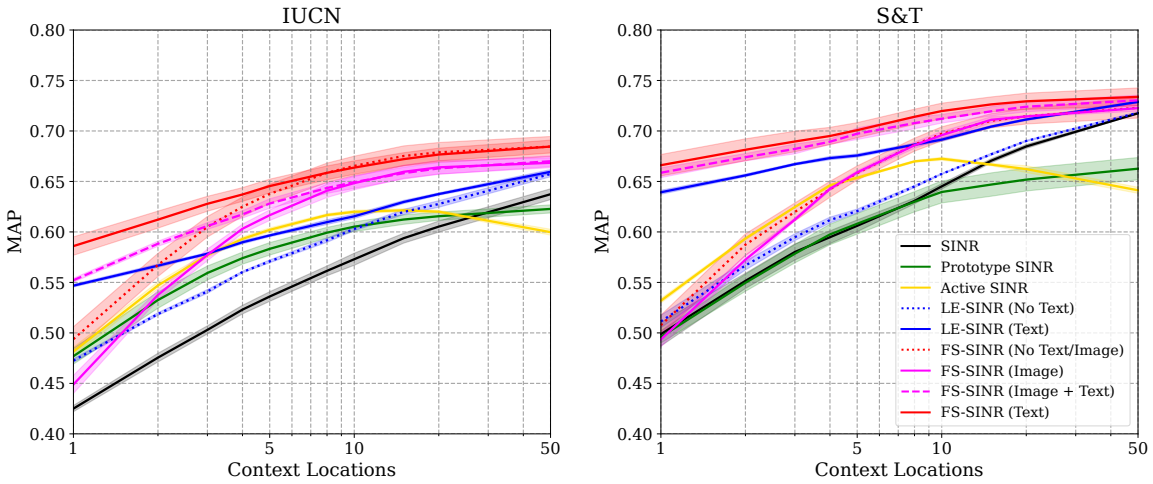


Figure 3. **Few-shot results**. Here we evaluate different models on the task of species range estimation on the IUCN (left) and S&T (right) datasets. On the x-axis we vary the number of context locations seen at inference time for the held-out evaluation species. The y-axis represents Mean Average Precision (MAP), where higher values are better. The error bars display the standard deviation of three different runs. Our FS-SINR approach outperforms existing methods, especially in the very low-data setting (i.e., $<$ five context locations). Note that LE-SINR and SINR need to be retrained during evaluation when more observations are provided. Tables A3 and A4 report expanded results including larger numbers of context locations.

performance for some species at inference time by biasing the ranges produced toward the rough estimates made using images, which can provide limited information. Importantly, unlike SINR and LE-SINR, FS-SINR does not need to be retrained at inference time. Instead, it can make predictions in a feedforward manner irrespective of the context data available. This is advantageous in interactive settings, whereby the model can compute the location embeddings for all query locations on earth once, and then a user could experiment by adding different context information interactively. Removing the retraining step also allows FS-SINR to produce estimated ranges in a fraction of the compute time compared to other approaches. Compared to the publicly released implementation of LE-SINR on the same hardware, FS-SINR generates range estimates from one context location and text for all species in the IUCN and S&T datasets in 2% of the time on CPU, and 6% of the time on GPU.

We present qualitative results for three different species in Figure 7 where we visualize FS-SINR's predictions as we change the number of context locations. Given only a single context location, the model does a sensible job of localizing the species on Earth. This supports the findings from Figure 3 where we observe strong performance even when only one context location is available. When more information is provided, the predicted range more closely resembles the expert-derived range shown in the first row. However, we do note that the model can still make mistakes in our low data setting, such as the erroneous predictions for the 'Black and White Warbler' in South America. In Figure 4 we illustrate some examples of how text information, when paired with one single context location, can influence the model predictions. We observe dramatically different predicted ranges when the text prompt encourages the model to focus on different habitat types. We note that each of the predicted ranges is still consistent with the location of the single context location provided. Finally, in Figure 6 we compare FS-SINR range predictions to other approaches, namely SINR, LS-SINR, and Active SINR. We see that for this species FS-SINR more closely resembles the expert range when only three context locations are provided. Additional qualitative examples are provided in Appendix E.

4.3. Zero-shot Evaluation

In addition to being able to generate range predictions in the few-shot setting when limited location observations are provided, FS-SINR can also make predictions when no location information is provided but only additional metadata such as an image or text describing a previously unseen species is given, i.e., the *zero*-shot setting. These zero-shot results are presented in Table 1 for both the IUCN and S&T datasets.

We report results for several variants of FS-SINR where different types of metadata are used. As a baseline, we also present the performance of SINR (row 1) where the evaluation species are part of its training set i.e., not zero-shot. We can also add data from these species to the training set of our approach which unsurprisingly boosts performance (i.e., row 3 vs. 9), though unlike SINR, FS-SINR does not have weights associated with individual species and so the impact of seeing evaluation species during training is fairly small. As a trivial baseline, we also report performance of FS-SINR (row 4) when no location or text metadata is provided, i.e., this is simply the output of the class token. As expected,

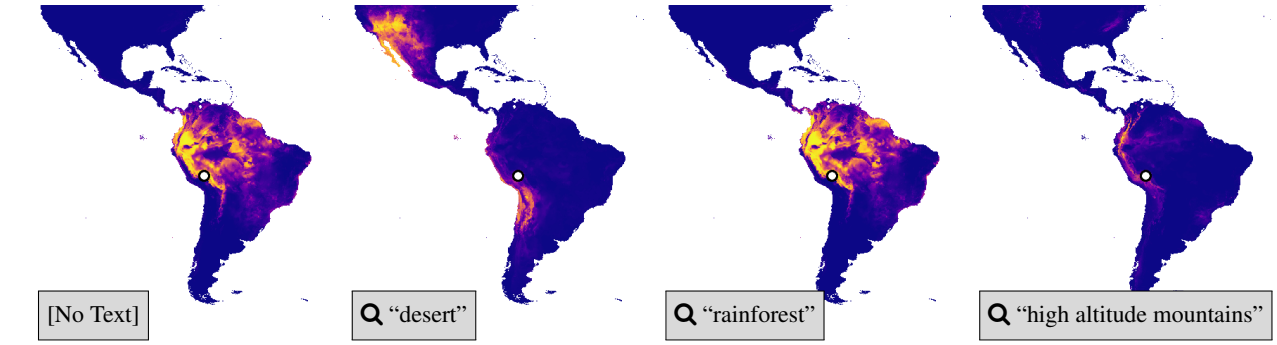


Figure 4. **Controlling range predictions using a single context location with different text**. Given the same single context location, denoted as ‘○’, FS-SINR can generate significantly different range predictions depending on the text provided. This example illustrates a use case where a user may have limited observations but some additional knowledge that can be encoded via text regarding the type of habitat a species of interest could be found in. Note, while ‘no text’ and ‘rainforest’ look similar, they are actually subtly different.

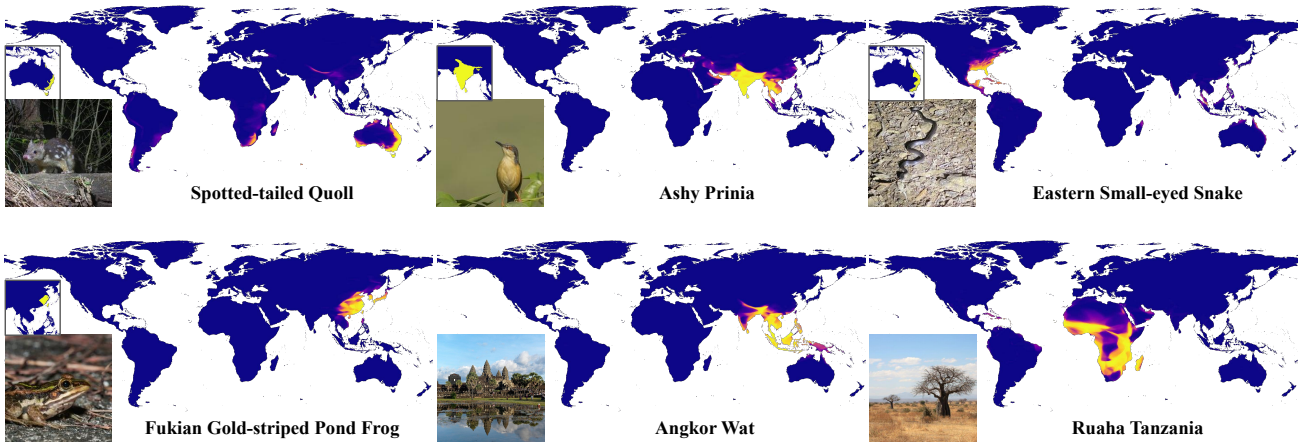


Figure 5. **Range predictions with a single context image as input**. We can condition FS-SINR on an arbitrary input image with no context locations or text, e.g., a held-out species (top row and bottom left), famous landmarks (bottom middle), or landscape images (bottom right).

this model performs poorly, but interestingly it seems to have learned some spatial prior that results in non-trivial predictions on S&T which contains bird species mostly concentrated in North America. We also compare to a version of FS-SINR (row 5) where we use taxonomic text (TRT) as in LD-SDM (Sastry et al., 2023) (see Appendix A.2 for further details).

In all instances, our FS-SINR approach outperforms LE-SINR, even when both models are provided with the same information at inference and training time (i.e., row 6 vs. 7 or row 8 vs. 9). Confirming observations from LE-SINR, we see that range text (RT) is more informative than habitat text (HT) (i.e., row 7 vs. 9). Additionally, image information provides some non-trivial signal (i.e., row 4 vs. 10), but it is not as informative as text (i.e., row 9 vs. 10), and can negatively impact performance when more informative sources are provided (i.e., row 9 vs. 11 for the harder IUCN dataset), as the model may overfit to incorrect spurious features in the image. As we can see in Figure 5 (with additional examples in Figure A24), zero-shot image predictions can be sensible, but predicting an unobserved species’ range from a single input image is ill posed. Text descriptions of range or habitat preferences are simply much more informative than a single image.

4.4. Additional Results and Ablations

In Appendix A we provide additional experimental results for FS-SINR. There we investigate uncertainty quantification to see how well calibrated the model predictions are. We also report results using a ‘distance-weighted MAP’ metric which penalizes errors more the further they are in distance away from the actual range. This metric more closely aligns human judgment of predicted range quality.

We provide a more ecologically relevant breakdown of results in Appendix B, where we find multiple potential sources of bias in our training data toward North America and Europe, and report higher evaluation performance in

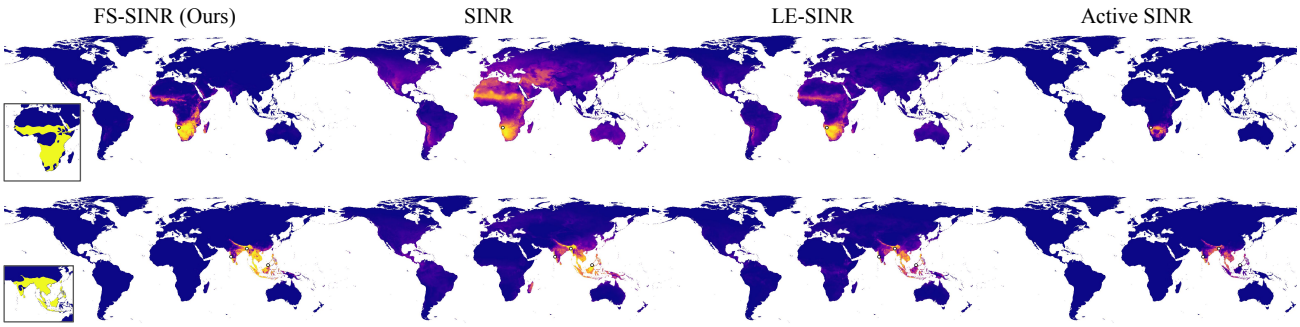


Figure 6. **Qualitative comparison of range predictions for different methods**. (Top) Predicted ranges from a single context location denoted as '○' and no additional metadata for the `Gabar Goshawk`. (Bottom) Predicted ranges from three context locations for the `Black-naped Monarch`. From left to right, FS-SINR (ours) with expert range inset, SINR, LE-SINR, and Active SINR. Please zoom in to see details.

Table 1. **Zero-shot results**. We report zero-shot performance where no location information is provided to each model, only additional metadata, comparing to SINR (Cole et al., 2023) and LE-SINR (Hamilton et al., 2024). We denote additional metadata used by models as RT for 'Range Text', HT for 'Habitat Text', and 'I' for 'Image'. TST represents 'Test Species in Train', indicating that a model uses location observations for the evaluation species at training time (e.g., SINR which provides an upper bound on performance), unlike other models where these species are excluded. TRT models are trained using 'Taxonomic Rank Text' as in Sastry et al. (2023), which are also provided with the full taxonomic description from 'class' to 'species' during evaluation. Results are presented as MAP, where higher is better.

ID	Method	Variant	IUCN	S&T
1	SINR	TST	0.67	0.77
2	FS-SINR	HT, TST	0.38	0.59
3	FS-SINR	RT, TST	0.55	0.67
4	FS-SINR		0.05	0.18
5	FS-SINR	TRT	0.21	0.34
6	LE-SINR	HT	0.28	0.52
7	FS-SINR	HT	0.33	0.53
8	LE-SINR	RT	0.48	0.60
9	FS-SINR	RT	0.52	0.64
10	FS-SINR	I	0.19	0.38
11	FS-SINR	I + RT	0.46	0.64

these regions for FS-SINR and LE-SINR. Similarly, we see that biases in the text data potentially leads to increased performance for charismatic and well-studied mammals compared to other taxonomic groups, though providing more context locations reduces this gap. FS-SINR is somewhat robust to these biases and outperforms other approaches across almost all categories. We also find that for all approaches tested, estimating very small ranges is difficult and performance varies strongly with range sizes, though FS-SINR again shows comparatively good performance.

We provide additional ablation experiments for FS-SINR in Appendix D, where we evaluate the impact of different input features, location encoders, and the amount of training data used, and explore architectural modifications such as removing the final species decoder that operates on the output of the Transformer. We observe that FS-SINR is robust to these changes, justifying its design decisions.

5. Limitations

While FS-SINR outperforms other zero and few-shot approaches for species range estimation, there are some limitations. First, given a set of input context locations FS-SINR is deterministic in that it will always generate the same output range map. In practice, in the few-shot regime, the same set of points could actually be representative of many different possible range maps. An obvious extension of our work is to introduce stochasticity into the model outputs, e.g., by treating class token output from the Transformer as a latent embedding for an additional sampling step. In Figure A25 we observe that initializing FS-SINR with different random seeds during training results in diverse range predictions across the different models, and in Appendix A we show that this can be exploited to give estimates of the uncertainty of predictions when using an ensemble of FS-SINR models. We leave further exploration of this for future work. Second, at inference time, users may wish to provide example locations indicating where a specific species has *not* been found, i.e., confirmed absences. Currently, our model is trained using presence-only data but could be adapted to use absence information, if available, which could be denoted via a different embedding type vector, to be learned during training alongside our existing token type embeddings. However, obtaining reliable large-scale absence data for tens of thousands of species is a challenging problem.

Finally, biodiversity data and particularly global-scale citizen science datasets like the one we use to train FS-SINR can contain large biases (Geldmann et al., 2016; Hughes et al., 2021), e.g., location, temporal, and taxonomic biases, among others. We do not explicitly account for these biases during training, though we make some attempt to

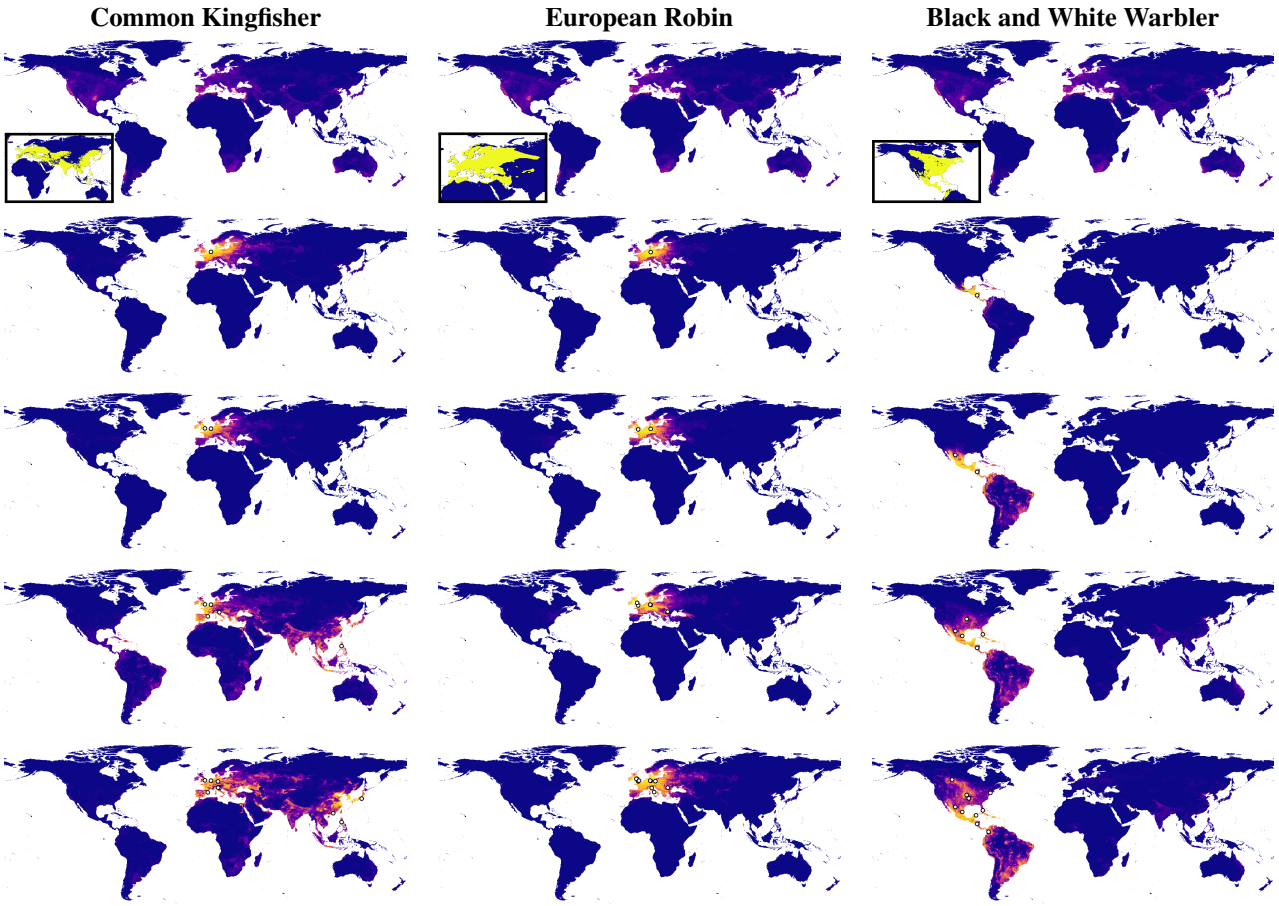


Figure 7. **Few-shot range estimation with increasing context locations**. Here we illustrate few-shot range predictions from FS-SINR given an increasing number of context locations $\{0, 1, 2, 5, 10\}$ and no other context information for the `Common Kingfisher` (left), `European Robin` (center), and the `Black and White Warbler` (right). In the first row, we show the expert-derived range inset and the prediction for the model when no context locations are provided (which is the same for all species). Then, in the remaining rows we increase the number of context locations, denoted as '$\circ$'. Please zoom in to see the context locations. As we increase the number of context locations, the predictions become closer to the expert ranges.

evaluate the impact of them in Appendix B, and thus we would caution the use of the predictions of our model in any applications that would use our range predictions in the context of biodiversity assessments. However, we note that we outperform existing and recent state-of-the-art range estimation methods, especially in the low observation data setting, and do not require any retraining at inference time.

6. Conclusion

The scientific community has limited knowledge on the geographical distributions of the majority of species on Earth. This lack of understanding is further hampered by the fact that we also have insufficient data to train models to estimate their ranges. To address this problem, we introduced FS-SINR, a new approach for few-shot species range estimation. We demonstrated that FS-SINR is able to fuse data from different modalities at inference time in a feedforward manner to efficiently make plausible range predictions for previously unobserved species. Our quantitative analysis, using expert-derived range maps, shows a 5-10% performance improvement compared to current approaches in the few-shot setting, i.e., when the number of observations equals ten, for previously unseen species. In addition, we also outperform existing methods in the zero-shot setting. While our results are promising, they also indicate that there are many open challenges in this important task.

Acknowledgements. We thank the iNaturalist community for making the species observation data available. OMA was supported by a Royal Society Research Grant. MH and SM were supported by NSF grants 2329927 and 2406687.

Impact Statement

Given the limited observations available for most species, there is a great need for reliable machine-learning based solutions for estimating their ranges. Such methods would provide us with unprecedented insight into how biodiversity is distributed on our planet and how it is changing over time. However, there are potential negative consequences associated with inaccurate range predictions generated by automated methods, e.g., a downstream conservation decision could be made based on an erroneous range map, resulting in wasted resources. Thus, it is important for practitioners to scrutinize the outputs of models such as ours.

Another issue associated with training models on species observation data is that there is a risk that sensitive information (e.g., the locations of protected species) could be leaked or extracted from the models. To respect this concern, the models in this work were trained using only publicly available information which does not include any sensitive observations. Finally, our approach integrates predictions from pre-trained large language models. These models are known to be biased and capable of hallucinating and fabricating outputs. Spatially localizing the outputs of such models runs the risk of amplifying such biases if used inappropriately.

References

Albert, C., Luque, G. M., and Courchamp, F. The twenty most charismatic species. *PloS one*, 2018.

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

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

Breiman, L. Random forests. *Machine learning*, 2001.

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

Chichorro, F., Juslén, A., and Cardoso, P. A review of the relation between species traits and extinction risk. *Biological Conservation*, 2019.

Cole, E., Van Horn, G., Lange, C., Shepard, A., Leary, P., Perona, P., Loarie, S., and Mac Aodha, O. Spatial implicit neural representations for global-scale species mapping. In *ICML*, 2023.

Darcet, T., Oquab, M., Mairal, J., and Bojanowski, P. Vision transformers need registers. In *ICLR*, 2024.

Dietterich, T. G. Ensemble methods in machine learning. In *Multiple Classifier Systems*, 2000.

Dollinger, J., Brun, P., Sainte Fare Garnot, V., and Wegner, J. D. Sat-sinr: High-resolution species distribution models through satellite imagery. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2024.

Dorm, F., Lange, C., Loarie, S., and Mac Aodha, O. Generating Binary Species Range Maps. In *Computer Vision for Ecology Workshop at ECCV*, 2024.

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.

Fang, Y., Sun, Q., Wang, X., Huang, T., Wang, X., and Cao, Y. Eva-02: A visual representation for neon genesis. *Image and Vision Computing*, 2024.

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.

Geldmann, J., Heilmann-Clausen, J., Holm, T. E., Levinsky, I., Markussen, B., Olsen, K., Rahbek, C., and Tøttrup, A. P. What determines spatial bias in citizen science? exploring four recording schemes with different proficiency requirements. *Diversity and Distributions*, 2016.

Golding, N. and Purse, B. V. Fast and flexible bayesian species distribution modelling using gaussian processes. *Methods in Ecology and Evolution*, 2016.

Hamilton, M., Lange, C., Cole, E., Samuel, H., Shepard, A., Mac Aodha, O., Maji, S., and Van Horn, G. Combining observational data and language for species range estimation. In *NeurIPS*, 2024.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. *CVPR*, 2015.

Hughes, A. C., Orr, M. C., Ma, K., Costello, M. J., Waller, J., Provoost, P., Yang, Q., Zhu, C., and Qiao, H. Sampling biases shape our view of the natural world. *Ecography*, 2021.

iNaturalist, 2025. https://www.inaturalist.org.

Jetz, W., McGeoch, M. A., Guralnick, R., Ferrier, S., Beck, J., Costello, M. J., Fernandez, M., Geller, G. N., Keil, P., Merow, C., et al. Essential biodiversity variables for mapping and monitoring species populations. *Nature ecology and evolution*, 2019.

Kellenberger, B. A., Winner, K., and Jetz, W. The performance and potential of deep learning for predicting species distributions. *bioRxiv*, 2024.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *ICLR*, 2015.

Klemmer, K., Rolf, E., Robinson, C., Mackey, L., and Rußwurm, M. Satclip: Global, general-purpose location embeddings with satellite imagery. In *AAAI*, 2025.

Lange, C., Cole, E., Horn, G., and Mac Aodha, O. Active learning-based species range estimation. In *NeurIPS*, 2023.

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

Mantyka-pringle, C. S., Martin, T. G., and Rhodes, J. R. Interactions between climate and habitat loss effects on biodiversity: a systematic review and meta-analysis. *Global Change Biology*, 2012.

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. *Communications of the ACM*, 2021.

Mora, C., Tittensor, D. P., Adl, S., Simpson, A. G., and Worm, B. How many species are there on earth and in the ocean? *PLoS Biology*, 2011.

Muennighoff, N., Su, H., Wang, L., Yang, N., Wei, F., Yu, T., Singh, A., and Kiela, D. Generative representational instruction tuning. In *ICLR*, 2025.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al. Dinov2: Learning robust visual features without supervision. *TMLR*, 2024.

Parnami, A. and Lee, M. Learning from few examples: A summary of approaches to few-shot learning. *arXiv:2203.04291*, 2022.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: an imperative style, high-performance deep learning library. In *NeurIPS*, 2019.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. Scikit-learn: Machine learning in python. *JMLR*, 2011.

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.

Picek, L., Botella, C., Servajean, M., Leblanc, C., Palard, R., Larcher, T., Deneu, B., Marcos, D., Bonnet, P., and Joly, A. Geoplant: Spatial plant species prediction dataset. In *NeurIPS Datasets and Benchmarks Track*, 2024.

Poggi, M., Aleotti, F., Tosi, F., and Mattoccia, S. On the uncertainty of self-supervised monocular depth estimation. In *CVPR*, 2020.

Rußwurm, M., Klemmer, K., Rolf, E., Zbinden, R., and Tuia, D. Geographic location encoding with spherical harmonics and sinusoidal representation networks. In *ICLR*, 2024.

Sastry, S., Xing, X., Dhakal, A., Khanal, S., Ahmad, A., and Jacobs, N. Ld-sdm: Language-driven hierarchical species distribution modeling. *arXiv:2312.08334*, 2023.

Sastry, S., Khanal, S., Dhakal, A., Ahmad, A., and Jacobs, N. Taxabind: A unified embedding space for ecological applications. In *WACV*, 2025.

Snell, J., Swersky, K., and Zemel, R. Prototypical networks for few-shot learning. In *NeurIPS*, 2017.

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.

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

Trimble, M. J. and Van Aarde, R. J. Species inequality in scientific study. *Conservation biology*, 2010.

Van Horn, G., Cole, E., Beery, S., Wilber, K., Belongie, S., and Mac Aodha, O. Benchmarking representation learning for natural world image collections. In *CVPR*, 2021.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. In *NeurIPS*, 2017.

Wang, Y., Yao, Q., Kwok, J. T., and Ni, L. M. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys*, 2020.

Wikipedia, 2025. https://www.wikipedia.org.

Zbinden, R., Van Tiel, N., Kellenberger, B., Hughes, L., and Tuia, D. On the selection and effectiveness of pseudo-absences for species distribution modeling with deep learning. *Ecological Informatics*, 2024a.

Zbinden, R., van Tiel, N., Rußwurm, M., and Tuia, D. Imbalance-aware presence-only loss function for species distribution modeling. In *ICLR Workshop on Tackling Climate Change with Machine Learning*, 2024b.

Appendix

In this appendix we provide additional quantitative and qualitative results, analysis, implementation details, and ablations.

In Appendix A, we provide additional results for uncertainty quantification, use of taxonomic rank text, a 'distance weighted' MAP metric, and show expanded results from Figure 3 for the non-low-shot setting. In Appendix B, we perform an ecologically relevant analysis of our results, showing how performance varies with region, range size, and taxonomic group. In Appendix C, we provide details on the implementation of FS-SINR and the baseline approaches, and of the training and evaluation procedure. In Appendix D, we provide additional ablations of FS-SINR, investigating the impact of training data, different input features, and modifications to the architecture. Finally, in Appendix E, we show additional qualitative results, including visualizing zero-shot and few-shot ranges for species and non-species concepts, and comparisons to ranges produced by LE-SINR and SINR approaches.

A. Additional Quantitative Results

A.1. Uncertainty Quantification

Here, we report results for an ensemble based on FS-SINR and quantify the uncertainty in the ensemble's predictions using methods adapted from Poggi et al. (2020). We create an ensemble by averaging the predictions of three FS-SINR models trained with different random seeds. Figure A25 shows range estimates from three such models, where we can see that each model can produce significantly different outputs. We take the average of these models as the ensemble prediction and treat the variance between individual model predictions as an estimate of the uncertainty of the ensemble. If all models agree that a species is either present or absent at a location, then the uncertainty will be low, while if models have different predictions, then the uncertainty will be high. in Table A1, comparing 'Ensemble' MAP to 'Model' MAP (repeated from FS-SINR 'Text' MAP in Table A4 for convenience), we can see that creating an ensemble increases MAP by 0.01 to 0.02, agreeing with typical findings that ensembling can improve performance relative to individual models (Dietterich, 2000).

In order to quantify uncertainty of our ensemble we follow an approach used in Poggi et al. (2020). We iteratively remove locations from the evaluation dataset that have the highest estimated uncertainty and recalculate the MAP without these locations. In our case we remove 2% of the data at each step. If the ensemble uncertainty aligns with how likely it is to be incorrect, then the MAP at each step will increase as the data with higher estimated uncertainty is removed from the evaluation. We can then plot the MAP against the fraction of data used for the evaluation. Taking the area under the curve generated doing this gives us the estimated 'Sparsification Error AUC' (SEAUC). If, instead, we remove 2% of locations randomly at each step then the MAP for each evaluation will remain approximately equal to the MAP when using the entire evaluation dataset. The 'random' SEAUC in this case is effectively equal numerically to the MAP using all evaluation data and can be estimated as such. Taking the difference between the estimated SEARC and the MAP then gives us the 'Area Under the Random Gain' (AURG). A positive AURG shows that the ensemble's estimate of how uncertain its predictions are is better than random guessing. In Table A1 we see that the AURG is positive for all number of context locations and increases as more context locations are provided, showing that FS-SINR ensembles can provide a useful estimate of how certain they are about a prediction and that providing more context locations allows the ensemble to be more accurate in this uncertainty estimate.

In Figure A1 we visualize the ranges (means) and uncertainties (variances) for our FS-SINR ensemble for the `Yellow-footed Green Pigeon`, using 'Range' text, 'Habitat' text, or a single context location. We observe that the mean is high in the region of the expert-derived range, and lower in areas far from the range where a single model has erroneously predicted presence. The variance is generally lower in the region of the expert-derived range where all models agree, higher at the edges of this region where different models have different estimates of the extent of the range, and high in areas far from the true range where single models have incorrectly predicted presence, such as South America when 'Habitat' text is provided, or parts of Africa when 'Range' text is provided.

A.2. Taxonomic Understanding

Here, we investigate the impact of providing FS-SINR with an understanding of the species taxonomy. For this we provide 'Taxonomic Rank Text' (TRT) instead of the Wikipedia-based free-form descriptions of a species that are used for our standard FS-SINR approach. This text gives the taxonomy of the species in decreasing taxonomic rank, in the form

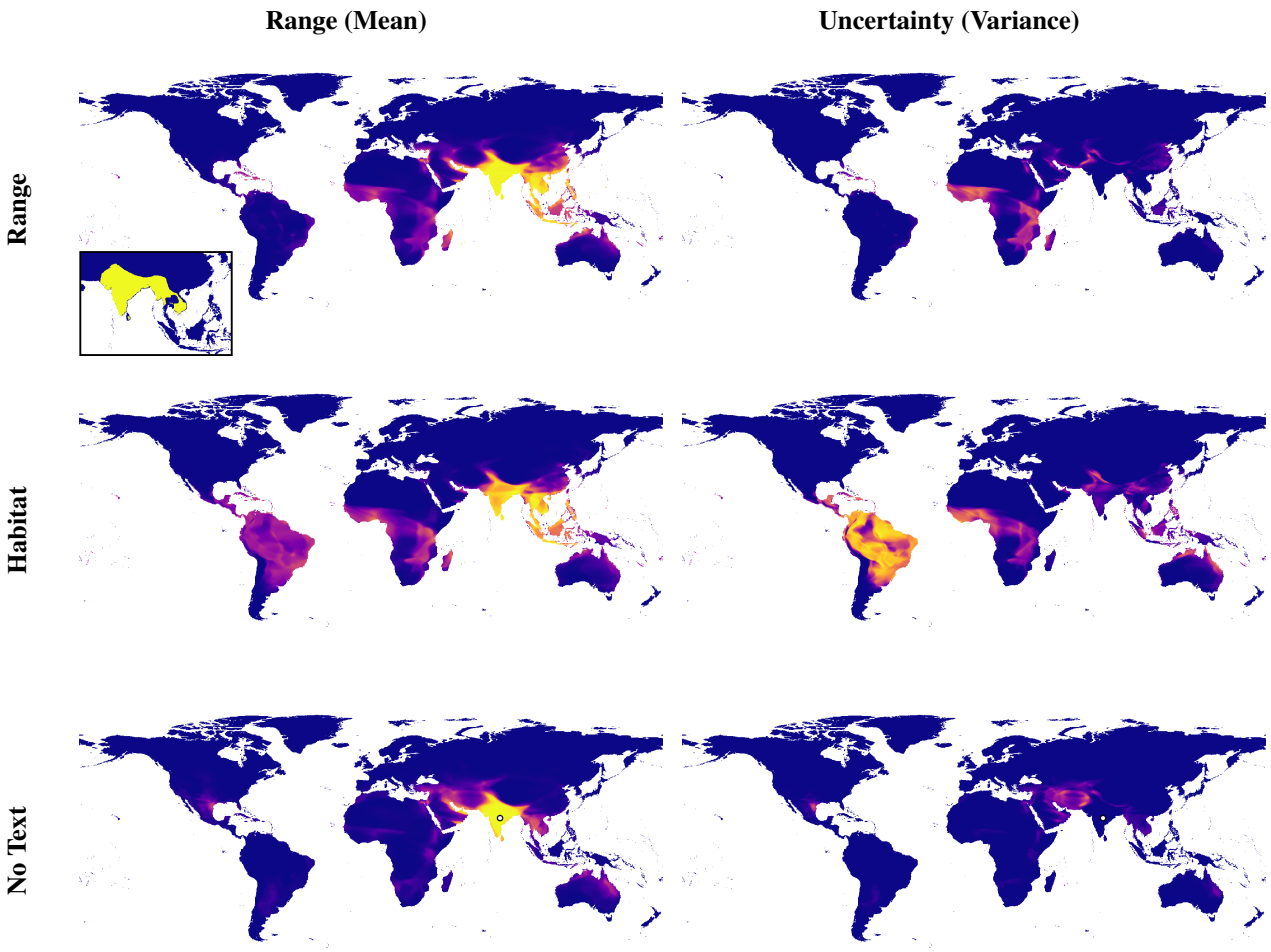


Figure A1. **Range estimates and visualized uncertainty estimates for a FS-SINR ensemble.** We display range estimates and uncertainties for the `Yellow-footed Green Pigeon` from an ensemble of three FS-SINR models. Zero-shot estimates are based on 'Range' text (top) and 'Habitat' text (middle). A few-shot estimate using no text and a single context location (bottom) is also shown. Range estimates for the ensemble (left) are a mean average of individual model predictions, while uncertainties (right) are estimated using the variances of the model predictions. The uncertainty is lower in the region of the expert-derived range where all models agree, higher at the edges of this region where different models have different estimates of the extent of the range, and high in areas far from the true range where single models have incorrectly predicted presence.
Range Text: "The yellow-footed green pigeon is found in the Indian subcontinent and parts of Southeast Asia. It is the state bird of Maharashtra."
Habitat Text: "The species is a habitat generalist, preferring dense forest areas with emergent trees, especially Banyan trees, but can also be spotted in natural remnants in urban areas. They forage in flocks and are often seen sunning on the tops of trees in the early morning."

'class order family genus species', so for a dog we would give the text '`Mammalia Carnivora Canidae Canis Familiaris`'. During training, we select a rank uniformly at random and remove all ranks beneath that. We hope that this process will force the model to learn an understanding of the distributions of not only individual species, but also genera, families, etc.. This may be helpful when facing unseen species as knowledge of the genus or family may provide clues about where this species may be found. This is similar to the approach used by LD-SDM (Sastry et al., 2023).

In Table A2 we show zero-shot performance for FS-SINR models trained on TRT on the IUCN and S&T evaluation tasks. We see that as we provide additional taxonomic information zero-shot performance improves, though it is still much worse than using habitat or range text. This implies that the model has managed to develop some understanding of the distributions of genera etc. and can use this to help map a novel species that shares higher order taxonomy with species in the training set.

In Figure A2 we provide some qualitative zero-shot and few-shot results showing the impact of training on taxonomic text. We see that the model appears to narrow down on the correct range as more specific taxonomy is revealed to it,

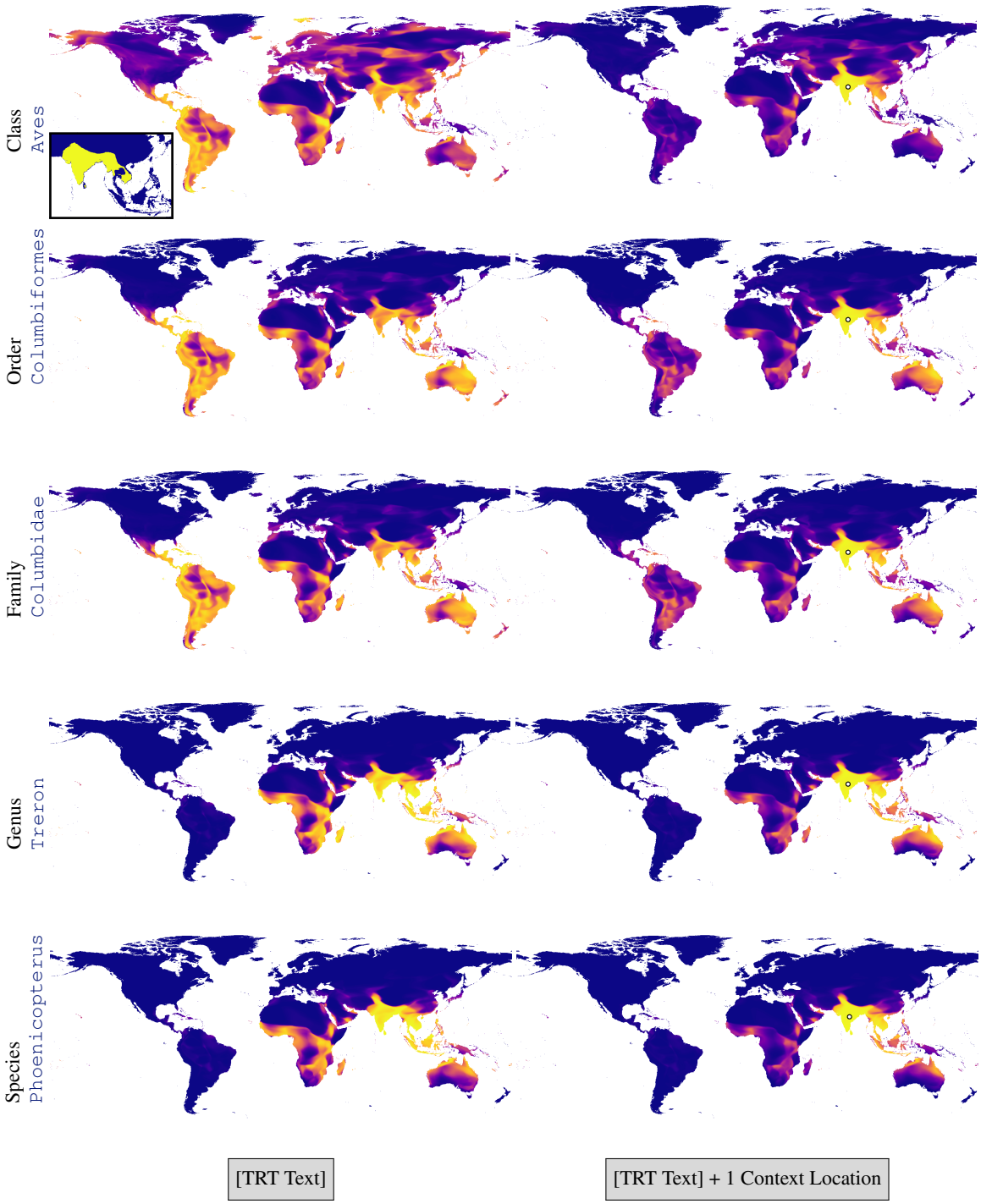


Figure A2. **Zero-shot and one-shot range estimation using Taxonomic Rank Text (TRT).** Range predictions for the species `Yellow-footed Green Pigeon` from FS-SINR model trained on taxonomic rank text as in LD-SDM (Sastry et al., 2023), with expert-derived range inset. As seen in Figure A3 and Table A2, the text-based zero-shot predictions seem to more closely match the expert-derived range as more of the taxonomic rank text of the species is provided. Taxonomic rank text allows the model to somewhat localize predictions to areas where species sharing the provided taxonomy ranks are present in the training set. For example, `Birds` are globally distributed and we see the model attempt to output this in the zero-shot 'Class' visualization. `Pigeons and Doves` are not found in the extreme north and providing these ranks reduces predictions in these areas (and much of the northern hemisphere). The model mostly manages to identify that `Green Pigeons` are found only in Africa and parts of Asia. A single observation significantly contracts the predicted ranges, particularly when less taxonomic information is provided. Click on the taxonomic rank names to visit the iNaturalist page for that taxa where the geographic distribution of observations for it can be observed.

Table A1. **Uncertainty quantification with ensembles.** We show MAP on the S&T dataset for an ensemble of three FS-SINR models using 'Range' text and differing numbers of context locations (Ensemble MAP), with metrics for uncertainty quantification adapted from Poggi et al. (2020). We report 'Sparsification Error AUC' (SEAUC) and 'Area Under the Random Gain' (AURG) for the ensemble. Positive AURG shows the ensemble is performing better than random chance at estimating its uncertainty. We also present results for the same text and context locations for individual FS-SINR models for comparison (Model MAP). Ensembling slightly improves performance for all numbers of context locations.

# Context	Model MAP	Ensemble MAP	SEAUC	AURG
0	0.64	0.66	0.68	0.03
1	0.66	0.68	0.71	0.03
2	0.67	0.69	0.73	0.03
3	0.68	0.70	0.74	0.04
4	0.69	0.71	0.75	0.04
5	0.70	0.71	0.75	0.04
8	0.71	0.72	0.76	0.04
10	0.72	0.73	0.77	0.04
15	0.72	0.73	0.78	0.05
20	0.72	0.74	0.79	0.05
50	0.73	0.74	0.78	0.05

from predicting across the entire globe when just the class `Aves` is provided, to removing northern latitudes as the family `Columbidae` is added, and finally removing the new world when the genus is provided. This broadly matches the actual distribution of these taxonomic ranks. Note that the relationship between taxonomic hierarchy and species range is likely complex as many speciation events (i.e., when a species splits into two or more new ones) can be the result of physical geographic barriers separating populations over time.

In Figure A3 we show few-shot results for FS-SINR models trained on TRT on the IUCN and SNT evaluation datasets. Zero-shot performance improvement with increasing taxonomic information is evident, but after very few provided locations this effect seems to disappear.

Table A2. **Zero-shot results with taxonomy rank text**. We denote additional metadata used by models as RT for 'Range Text' and HT for 'Habitat Text'. 'Species', 'Genus', 'Family', 'Order', 'Class' refer to models trained and evaluated using taxonomic rank text. Taxonomic information up to and including the specified rank is provided during evaluation.

Method	**Variant**	**IUCN**	**S&T**
FS-SINR		0.05	0.18
FS-SINR	HT	0.33	0.53
FS-SINR	RT	0.52	0.64
FS-SINR	Class	0.05	0.19
FS-SINR	Order	0.06	0.20
FS-SINR	Family	0.12	0.25
FS-SINR	Genus	0.18	0.30
FS-SINR	Species	0.21	0.34

A.3. Alternative Evaluation Metric

Here we provide additional results for the main models from Figure 3 using a 'distance weighted' MAP evaluation metric. This is inspired by the evaluation conducted in LD-SDM (Sastry et al., 2023). This metric is based on mean average precision (MAP), however we now weight predictions by distance from the true range, i.e., predicting the presence of a species far from where it is said to be found is penalized more than predicting the presence of a species in a location that is very close to existing observations, but is still actually outside the range. We intend that this metric more closely aligns with a human's judgment on how 'correct' a range is, compared to standard MAP. By considering both metrics we can be more confident that the improvement in range mapping performance that FS-SINR provides is not just a consequence of how we are quantifying it. We determine the weight for location $\boldsymbol{x}$ as:

$$w_{\boldsymbol{x}} = 1 + \frac{d_{range}(\boldsymbol{x})}{d_{antipodal}} h, \tag{3}$$

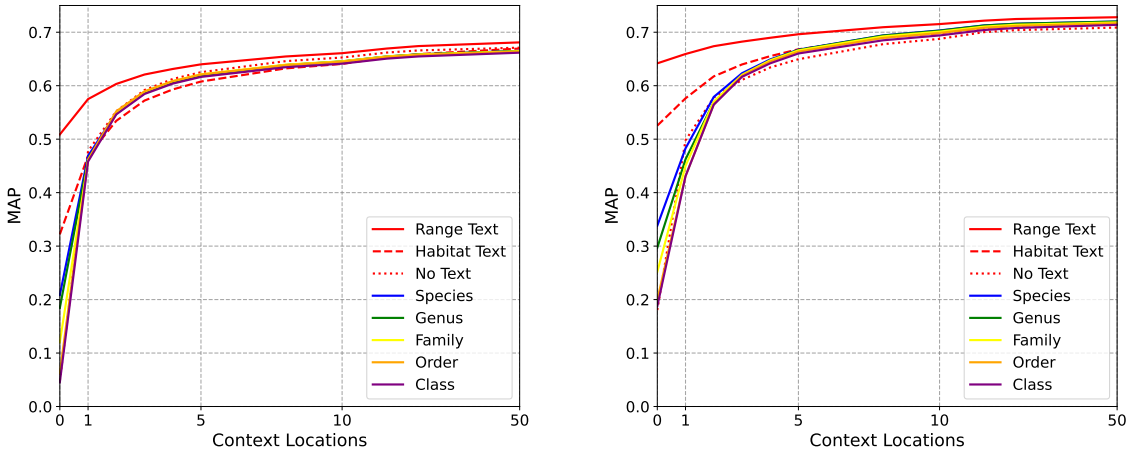


Figure A3. **Impact of training and evaluating with Taxonomic Rank Text.** Here we evaluate FS-SINR models trained using different context information on the IUCN dataset (left), and the S&T dataset (right). 'Class' indicates that only the taxonomic class of the species is provided as text during evaluation. 'Order' indicates that the taxonomic class followed by the order is provided as a text string during evaluation, and so on, such that 'Species' indicates that a text string in the format 'class order family genus species' is provided during evaluation. Providing more specific taxonomic text increases zero-shot performance. This is also presented Table A2. However we see that even the full taxonomy does not provide as much signal as habitat and range text for zero-shot range mapping. These more detailed texts provide more useful information for zero-shot range mapping - either actually mentioning geographic locations in the case of range text, or allowing the model to narrow predictions down to areas with specific features such as mountains and forests in the case of habitat text. When a single context location is provided, the choice of taxonomy text no longer seems to impact performance at all. It is possible that training on these less informative tokens means the model learns to pay less 'attention' to these text tokens compared to the Wikipedia-based text tokens usually used during training. This could explain why different rank taxonomy text tokens seemingly provide no benefit when any context locations are provided to the model.

where $d_{range}(\boldsymbol{x})$ is the distance along the earth's surface from point $\boldsymbol{x}$ to the nearest point of the expert-derived range using for evaluation, and $d_{antipodal}$ is the distance along the earth's surface between two points on opposite sides of the earth. While this distance does vary very slightly in different locations as the earth is not a perfect sphere, for this experiment we have set $d_{antipodal}$ to 20,037.5 km. h is the 'distance weight hyperparameter' and determines how much this metric penalizes incorrect predictions far from the range relative to close to the range. The metric is implemented equivalent to scikit-learn's `average_precision_score` `sample_weight` parameter (Pedregosa et al., 2011). We evaluate performance using the standard 'unweighted MAP', i.e., where $h = 0$ and so we are calculating MAP as usual, and 'distance weighted MAP' with $h = 9$ and $h = 99$. We selected these settings so that errors on the opposite side of earth from the true range are penalized 10 and 100 times more than errors close to the true range.

Results on the IUCN evaluation dataset can be found in Figure A4. We do not present results for the S&T dataset as we require knowledge of the range of each species globally to fairly apply the distance-weighted MAP, while the S&T dataset only provides range estimates for portions of the globe for each species. As the weight is increased, we observe a general reduction in overall performance. While there is no change in the relative ordering of different models, and FS-SINR outperforms LE-SINR and SINR across all settings of h, we do observe that FS-SINR and LE-SINR models that use habitat text during evaluation seem to decrease in performance more with larger h compared to other approaches. They are likely most effected by the larger weight, as habitat text can cause the model to predict presence in locations around the world with similar habitat features such as mountains, forest, or desert, despite these locations being far from the true range. This appears to be true of both FS-SINR and LE-SINR. For LE-SINR, using habitat text outperforms not using text for the *unweighted* MAP, but using habitat text performs worse than not using text for the weighted MAP. In Figure A5, we display zero-shot results for two species where there is a large difference in performance based on the two metrics. In both cases FS-SINR using only text incorrectly predicts the species to be present far from the expert-derived range.

A.4. Expanded Results

In Tables A3 and A4 we present and expanded set of results from Figure 3, for the IUCN and S&T datasets. We also report additional results for the non-few-shot setting using 500 and 1,000 context locations. Performance for most approaches including FS-SINR seems to plateau around 50 context locations, with some approaches gaining a minor boost

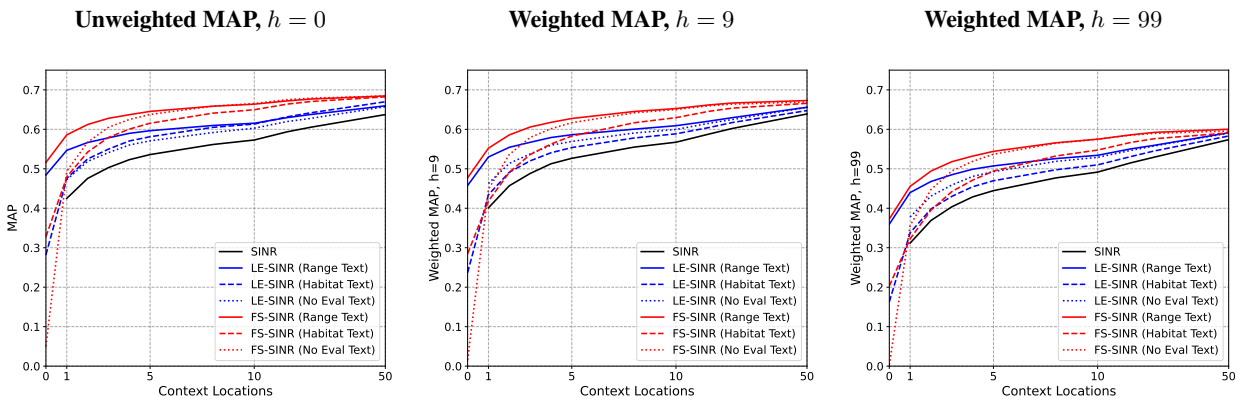
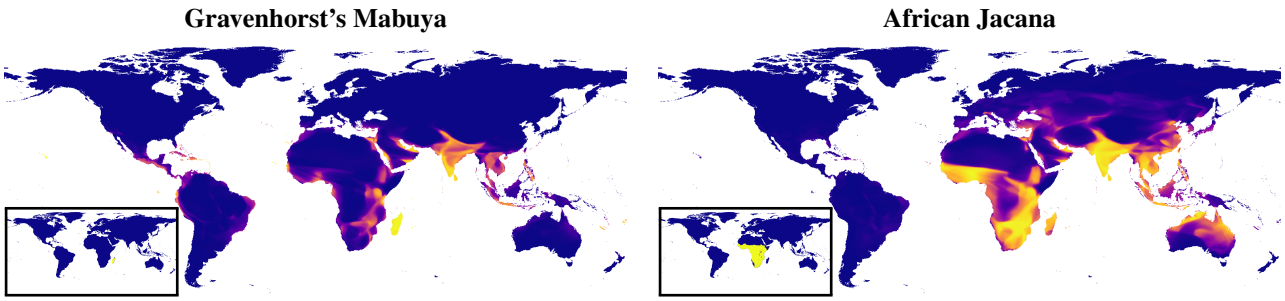


Figure A4. **Zero-shot and few-shot performance using our distance weighted MAP metric on the IUCN evaluation dataset.** We find that increasing the distance weight hyperparameter, h, reduces performance across the board without significantly changing the order of different models i.e., FS-SINR continues to outperform LE-SINR and SINR. We do see that approaches using habitat text decrease in performance more as h increases, relative to approaches not using text or using range text.



Habitat Text: "T. gravenhorstii prefers rocky areas, grassland, shrubland, and forest, but is also found in coffee plantations and ylang-ylang plantations. The species is adapted to a variety of environmental conditions, including different types of ecosystems."

Habitat Text: "The African jacana prefers shallow lakes and its preferred habitat is floating vegetation. It has a specific environmental condition, which is not mentioned in the article, but it can be inferred that it requires a certain level of water depth and vegetation cover."

Figure A5. **Examples of two species with poor distanced weighted MAP performance.** Here we visualize FS-SINR's zero-shot predictions using habitat text for two species where there is a large difference between the evaluation scores using the standard MAP metric compared to the distance weighted one (here using $h = 9$). For the `Gravenhorst's Mabuya` (left), which is endemic to Madagascar, we obtain an MAP of 0.419 but a lower distance weighted MAP of 0.175. For the `African Jacana` (right), found in most of sub-Saharan Africa, we obtain an MAP of 0.457 and a distance weighted MAP of 0.226. The distance weighted metric more heavily penalizes mistakes for these species that are very far from their true range.

in performance in the non-few-shot setting (SINR and Prototype-SINR), while others perform much worse (Active SINR). We also report results for a model trained and evaluated using context locations and features from a visual encoder with a DINOv2 backbone. This performs worse and has significantly more variable performance across runs compared to the pre-trained EVA-02 ViT, which was fine-tuned using species images during its original pre-training, used in the main paper.

A.5. Comparison to Traditional Machine Learning Methods

In Figure A6 we compare FS-SINR with two traditional machine learning approaches for classifying whether a species is present or absent at a given location. Both approaches were implemented using scikit-learn (Pedregosa et al., 2011). We compare with a Gaussian Process approach adapted from Golding & Purse (2016). While this method is designed for presence-absence data, we adapt it for the presence-only setting by providing a number of pseudo-negatives equal to the provided number of context locations, which performed best out of several strategies we investigated. Using a large number of pseudo-negatives as we do for FS-SINR both performed poorly and took excessively long to run as Gaussian Process computations scale cubically with the amount of data. For this approach we use a logit link function and a squared-exponential kernel. We also compare to a random forest classifier (Breiman, 2001). We investigated providing the same number of pseudo-negatives as for FS-SINR with appropriate class weights, however we found better performance by

Table A3. **IUCN zero-shot and few-shot results.** Here we present expanded IUCN evaluation results for the models shown in Figure 3 in tabular form. We also show results using a DINOv2 based image encoder. SINR and LE-SINR without text cannot produce a range map without at least one context location. Results are presented as MAP, where higher is better.

	FS-SINR					LE-SINR		SINR	Prototype SINR	Active SINR
# Context	Text	Image	Text + Image	No Text \Image	DINOv2	Text	No Text	No Text	No Text	No Text
0	0.52	0.19	0.46	0.05	0.13	0.48	-	-	-	-
1	0.57	0.45	0.55	0.48	0.40	0.55	0.47	0.42	0.48	0.48
2	0.60	0.54	0.59	0.56	0.49	0.57	0.52	0.47	0.53	0.55
3	0.62	0.58	0.61	0.60	0.53	0.58	0.54	0.50	0.56	0.58
4	0.63	0.60	0.62	0.62	0.55	0.59	0.56	0.52	0.57	0.59
5	0.64	0.62	0.63	0.63	0.56	0.60	0.57	0.54	0.58	0.60
8	0.65	0.64	0.64	0.65	0.59	0.61	0.59	0.56	0.60	0.62
10	0.66	0.65	0.65	0.66	0.60	0.62	0.60	0.57	0.61	0.62
15	0.67	0.66	0.66	0.67	0.61	0.63	0.62	0.59	0.61	0.62
20	0.67	0.66	0.66	0.67	0.61	0.64	0.63	0.61	0.62	0.62
50	0.68	0.67	0.67	0.67	0.61	0.66	0.66	0.64	0.62	0.60
500	0.68	0.66	0.66	0.67	0.57	0.67	0.67	0.65	0.63	0.37
1000	0.68	0.66	0.66	0.67	0.57	0.67	0.67	0.65	0.63	0.36

Table A4. **S&T zero-shot and few-shot results.** Here we present expanded S&T evaluation results for the models shown in Figure 3 in tabular form. We also show results using a DINOv2 based image encoder. SINR and LE-SINR without text cannot produce a range map without at least one context location. Results are presented as MAP, where higher is better.

	FS-SINR					LE-SINR		SINR	Prototype SINR	Active SINR
# Context	Text	Image	Text + Image	No Text \Image	DINOv2	Text	No Text	No Text	No Text	No Text
0	0.64	0.38	0.64	0.18	0.28	0.60	-	-	-	-
1	0.66	0.49	0.66	0.50	0.44	0.64	0.52	0.49	0.54	0.53
2	0.67	0.57	0.67	0.58	0.53	0.66	0.57	0.55	0.59	0.59
3	0.68	0.61	0.68	0.61	0.57	0.67	0.60	0.58	0.61	0.62
4	0.69	0.64	0.69	0.64	0.61	0.67	0.61	0.59	0.62	0.65
5	0.70	0.66	0.70	0.65	0.63	0.68	0.62	0.60	0.63	0.65
8	0.71	0.69	0.71	0.68	0.65	0.69	0.65	0.63	0.64	0.67
10	0.72	0.70	0.71	0.69	0.67	0.69	0.66	0.64	0.65	0.67
15	0.72	0.71	0.72	0.70	0.68	0.70	0.68	0.67	0.65	0.67
20	0.72	0.71	0.72	0.71	0.68	0.71	0.69	0.68	0.65	0.66
50	0.73	0.72	0.73	0.71	0.69	0.73	0.72	0.72	0.66	0.64
500	0.73	0.71	0.72	0.71	0.62	0.73	0.72	0.73	0.67	0.24
1000	0.72	0.71	0.72	0.71	0.62	0.73	0.72	0.73	0.67	0.24

providing the same number of pseudo-negatives as context locations. In both cases, performance for this task is significantly worse than FS-SINR, LE-SINR and SINR.

B. Ecologically Relevant Analysis of Results

To give a more ecologically relevant analysis of our results we present them here by continent, species range size, and taxonomic class. In these cases we display results for the IUCN evaluation dataset only, as it provides expert-derived presence-absence information globally and includes a range of taxonomic groups. In comparison the S&T dataset only includes `Aves`, i.e., bird species, and evaluates the presence or absence of each species over a portion of the globe, preventing us from calculating global range sizes and performance by continent.

B.1. Results by Region

In Figure A7 we show performance of FS-SINR, LE-SINR, and SINR models by continent for few-shot and zero-shot text-only predictions. FS-SINR outperforms other approaches on all continents except South America and Oceania, where at larger numbers of context locations LE-SINR becomes comparable. Biodiversity data and particularly global-scale citizen science datasets such as the iNaturalist-derived data we use to train FS-SINR can contain large biases (Geldmann et al., 2016; Hughes et al., 2021), and here we can see the impact of this bias within our training data.

We have more species observation training data (taken from Cole et al. (2023), which also visualizes the data distribution)

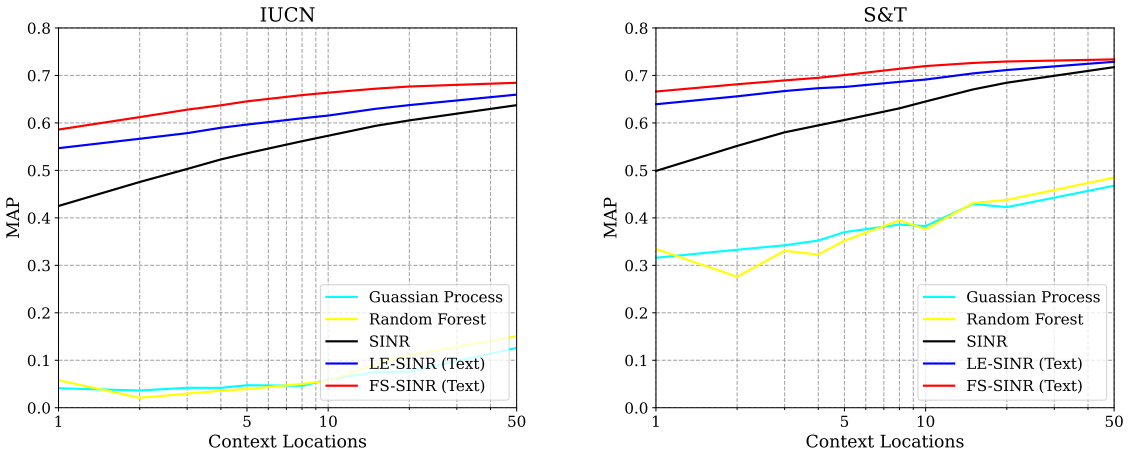


Figure A6. **Comparison to traditional machine-learning approaches.** Here we evaluate a Gaussian Process approach and a Random Forest on the IUCN (left) and S&T (right) datasets. We find that both approaches perform far worse than FS-SINR, LE-SINR or SINR for species range estimation using presence-only data. 'Range' text is used here for FS-SINR and LE-SINR, while other approaches take only context locations as input.

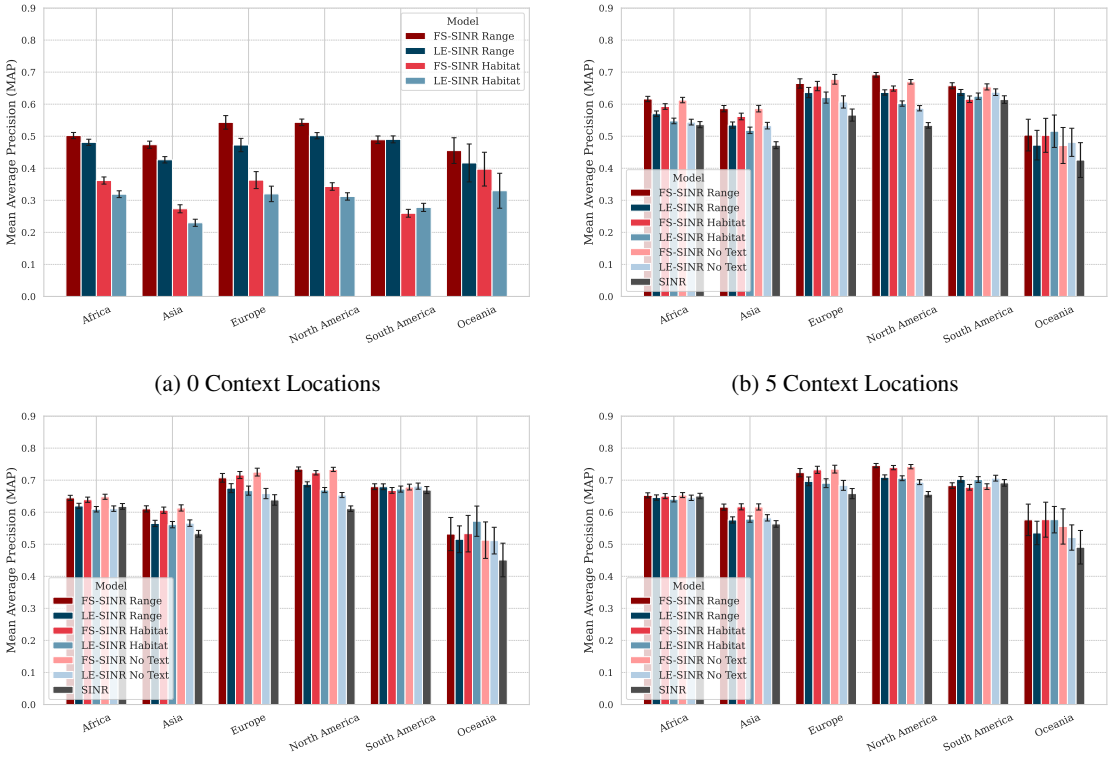
from Europe and North America than from other areas, and the observations in these regions are relatively well distributed spatially. In other continents, there are large areas with very few or no training observations for models to learn from, along with small areas that are highly observed. On top of this, our text descriptions are taken from English language Wikipedia and may be more descriptive for species found in areas where English is widely spoken, and our pre-trained large language model may have more knowledge of North American and European geography and ecology due to biases in the text data used for training. Combined, these factors lead to higher performance in North America and Europe compared to other regions. In Figure A10 we show the average false positive error for few-shot range estimation for FS-SINR on the IUCN evaluation dataset. This indicates greater error in regions where we have less training data.

B.2. Results by Species Range Size

Here, we display results showing the average MAP for species in our IUCN evaluation dataset, grouped by range size, where range size is computed from the expert-derived range maps. In Figure A9 we break down performance of zero-shot approaches by range size for FS-SINR, LE-SINR, and SINR. We find that for all models and settings, performance varies very strongly with range size. This is most significant in the zero-shot setting. FS-SINR performs well compared to the baselines, though all models struggle with very small ranges. As small-ranged species are especially vulnerable to extinction (Chichorro et al., 2019), current methods performing poorly for these species when evaluated globally may be of concern and improving the modeling of these species may be a priority from a conservation perspective. We also see that performance worsens for the very largest ranges.

B.3. Results by Taxonomic Class

Here we break down performance on the IUCN evaluation dataset by taxonomic class. Four taxonomic classes are present in our training data, namely `Amphibia`, `Aves`, `Mammalia`, and `Reptilia`. In Figure A8 (a) we display zero-shot performance for FS-SINR and LE-SINR using range and habitat text. We observe that `Aves` and especially `Mammalia` outperform the other classes, particularly when habitat text is provided. Albert et al. (2018) suggest that of the 20 most 'charismatic' species in the western world, all but the `Great White Shark` and `Crocodile` are mammals, and Trimble & Van Aarde (2010) show that scientific research is heavily focused on mammals. We may be seeing the impact of this, where mammals are more likely to have detailed Wikipedia pages which we draw our textual training and evaluation data from. In Figure A8 (b), (c), and (d), we investigate how these differences in performance between taxonomic classes change as more location data is provided. We see that for both FS-SINR and LE-SINR, providing context locations significantly reduces the differences in performance between taxonomic class, though mammals do continue to very slightly outperform other taxonomic groups for a given model and setting.



(a) 0 Context Locations

(b) 5 Context Locations

(c) 20 Context Locations

(d) 50 Context Locations

Figure A7. **IUCN Performance by continent.** Error bars show standard error of the mean.

C. Implementation Details

C.1. FS-SINR

C.1.1. Model Architecture

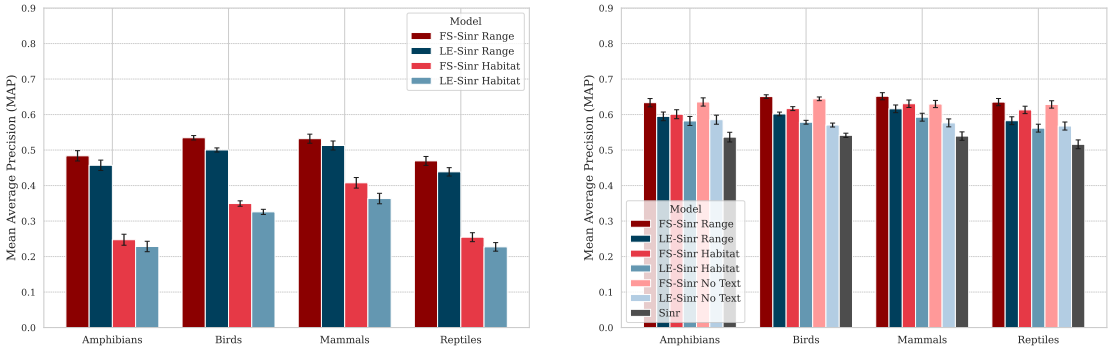
Code for FS-SINR is available at https://github.com/Chris-lange/fs-sinr. The architecture for FS-SINR consists of four components: the location encoder, f; the text encoder, t; the image encoder, a; the transformer encoder, m; and the species decoder, s. These components comprise 8,154,368 learnable parameters in total. All non-linearities in FS-SINR are ReLUs.

The location encoder, f, is identical to the one used in Hamilton et al. (2024), which is taken from Cole et al. (2023). It is composed of an initial linear layer and ReLU non-linearity followed by four residual layers, where each is a two-layer fully connected network with residual connections (He et al., 2015) between the input and output of each residual layer. Each layer contains 256 neurons, and there are 527,616 learnable parameters in total.

The text encoder, t, follows the structure of the text-based species encoder from Hamilton et al. (2024). In t, a pre-trained and frozen large language model, GritLM (Muennighoff et al., 2025), is used to produce a fixed 4,096 length embedding from input text. This is then passed through a smaller network to reduce the dimensionality to 256. This smaller network consists of two residual layers with a hidden layer size of 512. In total, the text encoder contains 3,410,432 learnable parameters.

The image encoder, a, has a structure similar to t. In a, a pre-trained and frozen vision transformer, EVA-02 (Fang et al., 2024), pre-trained on images from 10,000 species from the iNaturalist species classification dataset (Van Horn et al., 2021), is used to produce a fixed 1,024 length embedding from an input image, by extracting the `CLS` token from the final layer of the model. This is then passed through a smaller network to reduce the dimensionality to 256. This smaller network consists of two residual layers with a hidden layer size of 512. In total, the image encoder contains 1,837,568 learnable parameters. Tables A3 to A5 include results using a DINOv2-large image encoder (Oquab et al., 2024) instead of the EVA-02 ViT where all other architecture choices remain the same.

The transformer encoder, m, takes in an arbitrary length set of unordered 256 dimensional tokens produced by f, t, and a,



(a) 0 Context Locations

(b) 5 Context Locations

(c) 20 Context Locations

(d) 50 Context Locations

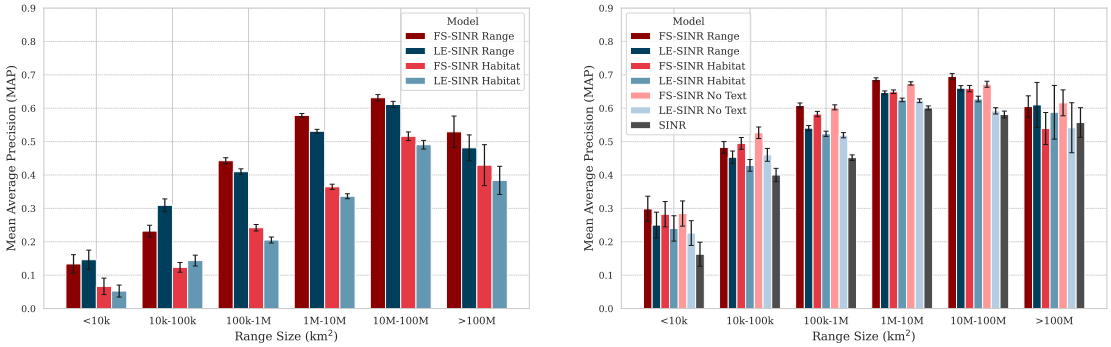
Figure A8. **IUCN Performance by taxonomic group.** Error bars show standard error of the mean.

as well as two learned tokens that are added to each set of inputs. The `CLS`, class, token produces the species range, and a 'Register' token, inspired by Darcet et al. (2024), which acts as an additional repository of global information during encoding. Element-wise addition between each token and one of five learned 256 dimensional 'token type embeddings' is performed to allow the model to differentiate between tokens from different sources. The transformer itself is composed of four transformer encoder layers, implemented using PyTorch's `nn.TransformerEncoderLayer` (Paszke et al., 2019), based on Vaswani et al. (2017). Key-query-value multi-head attention is used with two 'heads.' The feed forward components contain 512 neurons per layer, while the token dimensionality is 256. Layer norm is used in each layer, using a default epsilon value of 1e-5 for enhanced numerical stability. In total, m contains 2,176,256 learnable parameters. Finally the species decoder, s, is a simple fully connected network with two hidden layers. Each layer contains 256 neurons, and in total the decoder contains 197,376 learnable parameters.

C.1.2. Training

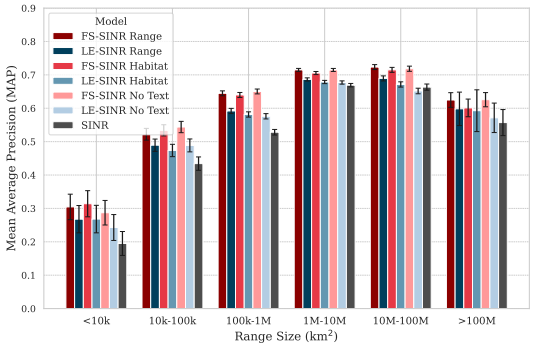
For all training we use the Adam optimizer (Kingma & Ba, 2015) with a learning rate of 0.0005, and an exponential learning rate scheduler with a learning rate decay of 0.98 per epoch, and we use a batch size of 2048. Our training data comes from Cole et al. (2023), comprising 35.5 million species observations with locations, covering 47,375 species observed prior to 2022 on the iNaturalist platform. However, we remove all species found in our evaluation datasets, leaving us with 44,181 species in our training set.

Training comprises of two steps. First, the location encoder, f, is trained. This follows the training procedure of Cole et al. (2023) using the $\mathcal{L}_{\text{AN-full}}$ loss function with positive weighting, λ, set to 2,048, training for 20 epochs with a dropout of 0.5. To reduce training time without significantly impacting performance we only train on a maximum of 1,000 examples per-species, as done in Cole et al. (2023). Thus, our training dataset for this step contains 13.8 million location observations. Secondly, we train all components of FS-SINR, except the pre-trained large language model and the pretrained vision transformer, using our $\mathcal{L}_{\text{AN-full-b}}$ loss with λ set to 2,048. We train the location encoder, f, again as this improves performance compared to freezing it,as seen in Figure A15. For this part of training, we use a dropout of 0.2. We further reduce the training data used to a maximum of 100 examples per-species, leaving 4.0 million training examples, which again increases

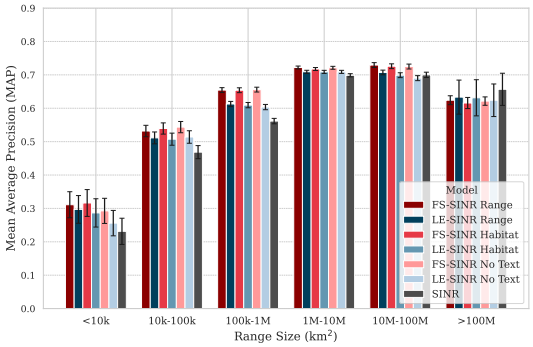


(a) 0 Context Locations

(b) 5 Context Locations



(c) 20 Context Locations



(d) 50 Context Locations

Figure A9. **IUCN Performance by range size.** Error bars show standard error of the mean.

training speed without a significant impact on performance, as seen in Figure A16. For this step we additionally train with images and text descriptions of the training species. Each instance in the training set is used once per epoch as a training example to compute the loss. The training example is not passed through the transformer encoder, m, and so does not contribute to the species embedding vector produced by this part of the model. Instead, additional context information is provided to produce the species embedding. By default, this context information consists of 20 context locations, a section of text describing the target species, and an image of the species. With 0.2 probability the context locations are dropped from the context information, and with 0.5 probability each the text or image is dropped. These context locations are taken from the training data for the target species. As such, a single instance from the training set can be used multiple times per epoch, once as a training example, and potentially many times as a context location. The impact of different distributions of locations and text provided during training is shown in Figure A13.

For the text inputs required during this stage of training, we use the text dataset from Hamilton et al. (2024) consisting of multiple sections of Wikipedia (Wikipedia, 2025) articles for each species in the train set where these are available. This dataset contains 127,484 sections from 37,889 species' articles. The evaluation text either describes the habitat or range a species, where habitat text tends to describe the local environment and range text is typically more informative as it often lists specific countries or geographic regions where the species can be found. Note that not all 44,181 train species have text data available. The images used are taken from iNaturalist (2025), and this dataset comprises 204,064 images of our train species. When an image or piece of text is not available for a species during training, and we are trying to provide these modalities and context locations to the model, we simply ignore the additional modality and only provide the context locations. When we are attempting to provide just an image or text as context, we instead skip that training instance.

In practice, during training, we pass all text sections through the frozen large language model once and then store the embeddings produced to use in the current training run and all future runs, and similarly extract and store all image embeddings after passing the images through the frozen vision transformer. This prevents us having to repeatedly query these frozen, but resource-intensive, models during training. Training takes approximately ten hours on a single NVIDIA A6000 GPU, requiring approximately six gigabytes of RAM.

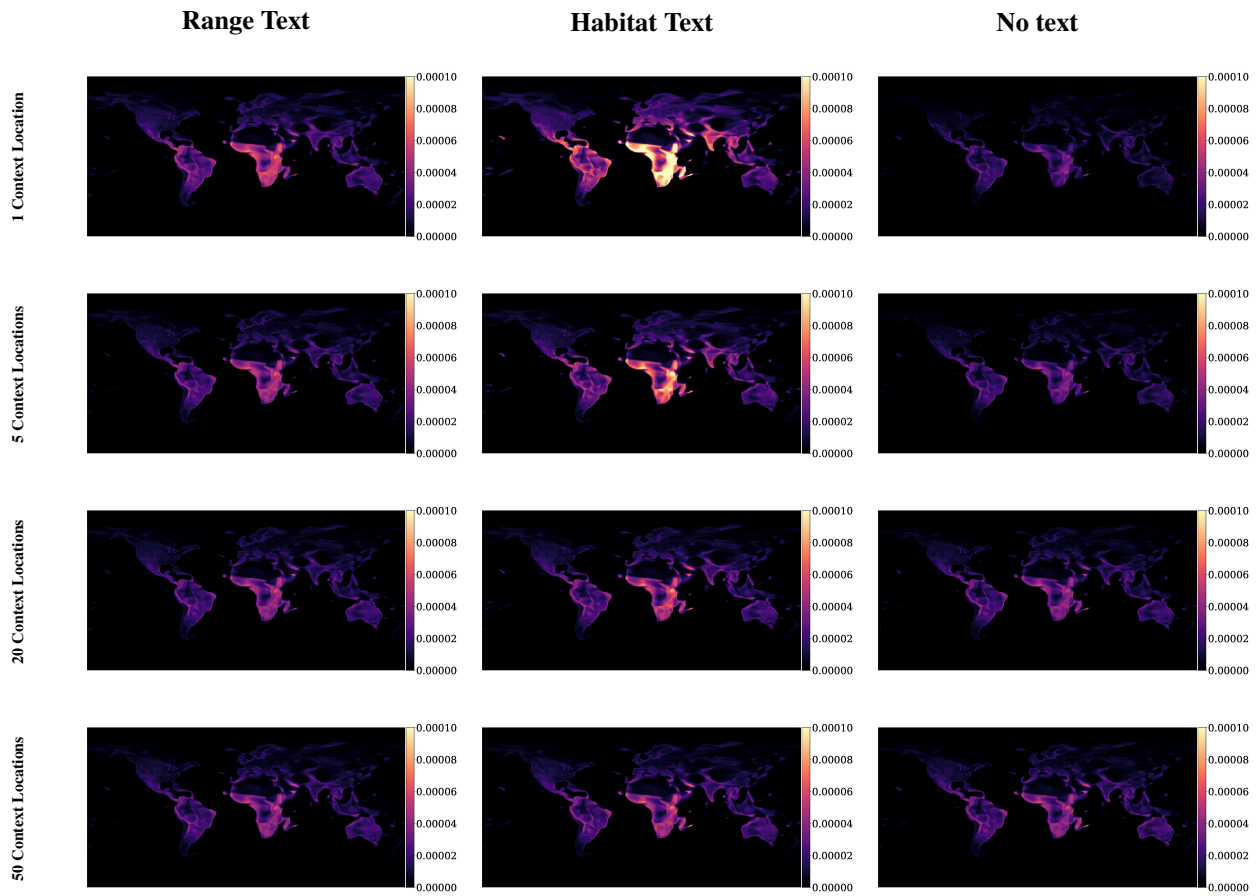


Figure A10. **Average false positive error by location for few-shot approaches.** Here we see average false positive error of FS-SINR on the IUCN dataset. Providing any text leads to an increase in the false positive error, although Figure 3 suggests that this text still helps with range estimation performance. As the number of provided context locations increases, the impact of the text is reduced and the distribution of errors appear similar.

C.2. Baselines

C.2.1. LE-SINR

We compare our approach to the recently introduced species range model LE-SINR (Hamilton et al., 2024) that can incorporate text information. We follow the original architecture and training procedure for LE-SINR and SINR, with the exception that we enforce that SINR, like LE-SINR and our approach, is trained on our reduced set of 44,181 species which do not include any of the evaluation species.

We also follow the original evaluation procedure for LE-SINR. For few-shot evaluation without text, logistic regression with L2 regularization is performed with location features as input using the few positive examples provided alongside a set of pseudo-negatives drawn half from a uniform random distribution and half from the training data distribution. The regularization weight is set to 20. For text-based zero-shot evaluation, we directly make use of the output of the text encoder with the dot product between this and location features giving us a probability of species presence. For few-shot evaluation, when text is provided, we again perform logistic regression, but the output of the text encoder is used as the 'target' that the weights are drawn towards in a modified L2 regularization term,see Hamilton et al. (2024) for more details. The regularization weight is again set to 20. In total, this model comprises 25,715,202 learnable parameters.

C.2.2. SINR

We also compare to SINR (Cole et al., 2023). The original SINR implementation requires all evaluation species to be part of the training set. We match the adaptations from Hamilton et al. (2024) to allow evaluation on unseen species. After training

we remove the learned species heads and keep only the location encoder. During evaluation, we perform logistic regression with L2 regularization using location features as input. The regularization weight is again set to 20, and the same method of selecting pseudo-negatives as above is used. In total, this model comprises 11,941,120 learnable parameters.

C.2.3. PROTOTYPE SINR

Here we describe our few-shot baseline based on Prototypical Networks (Snell et al., 2017), which we refer to as Prototype SINR. Our approach is very similar to Snell et al. (2017) although we use the SINR location encoder of our models as the 'embedding function', allowing us to generate few-shot results for a novel species without any retraining. This SINR location encoder is trained only on species found in the training set and not those used for evaluation. Using this method, SINR and LE-SINR models can be used to estimate the range of a novel species without requiring training to learn a new species embedding vector.

In order to do this, we first encode our known 'presence' locations using the location encoder of our chosen model and then take an average of these points to generate a 'prototype' for the presence class. We select pseudo-negatives in the same manner as in Hamilton et al. (2024) and similarly encode and average these in order to generate a prototype for the 'absent' class. We represent these prototypes as:

$$\boldsymbol{r}_k = \frac{1}{|S_k|} \sum_{\boldsymbol{x}_i \in S_k} f_{\boldsymbol{\theta}}(\boldsymbol{x}_i), \tag{4}$$

where $k \in \{\text{present}, \text{absent}\}$ indicates the class of the prototype, and S is the 'support set', i.e., the set of locations $\boldsymbol{x}$ that we use to create our prototypes. In our case, $S_{present}$ is the small set of available context locations for our target species, i.e., $\mathcal{C}^t$, while S_{absent} is the set of pseudo-negative locations that we have selected according to Hamilton et al. (2024). $f_{\boldsymbol{\theta}}()$ denotes the location encoder of our model.

To generate a probability of presence or absence for any location $\boldsymbol{x}$, we encode $\boldsymbol{x}$ using our location encoder and calculate the cosine distance in 'location encoder space' between $\boldsymbol{x}$ and each prototype. We then use these values as the 'logits' in a softmax function to generate our probabilities. The parameters of the location encoder are not changed. Putting this together, we can calculate the probability of presence as:

$$p_{\text{present}}(\boldsymbol{x}) = \frac{e^{d(f_{\boldsymbol{\theta}}(\boldsymbol{x}), \boldsymbol{r}_{\text{present}})}}{e^{d(f_{\boldsymbol{\theta}}(\boldsymbol{x}), \boldsymbol{r}_{\text{present}})} + e^{d(f_{\boldsymbol{\theta}}(\boldsymbol{x}), \boldsymbol{r}_{\text{absent}})}}, \tag{5}$$

where $d(\boldsymbol{a}, \boldsymbol{b})$ represents a distance metric between $\boldsymbol{a}$ and $\boldsymbol{b}$, in this case, cosine distance.

While the original implementation in Snell et al. (2017) uses the squared Euclidean distance instead of cosine distance we find that this performs significantly worse and actually results in decreased MAP as the number of context locations increases. We suggest the SINR location encoder is more suited to using cosine distance, as during training presence predictions are generated by taking the dot product of the location and species embeddings. However, when the location encoder is trained from scratch for Prototype SINR as in Snell et al. (2017) we find that using the squared Euclidean distance performs better than cosine distance, although performance is still lower than cosine distance with a SINR location encoder.

In Figure 3 we see that the performance of Prototype SINR is worse than FS-SINR and the SINR and LE-SINR baselines. In Table A5 we provide zero-shot results where the positive prototype is a species embedding produced from text, in the same manner as LE-SINR zero-shot predictions. In both cases, Prototype SINR underperforms compared to our approach. In Figure A11 we present qualitative results visualizing the few-shot estimated range for the `Kalahari Scrub-Robin` produced by FS-SINR and by the Prototype SINR baseline.

C.2.4. ACTIVE SINR

We also compare to the model introduced for active learning in Lange et al. (2023), which we call Active SINR, although in our setting there is no active learning component. This approach begins with a SINR model trained on our reduced 44,181 species which do not include the evaluation species. The weights $\boldsymbol{W}$ of the multi-label classifier of this model can be viewed as a set of species embeddings where each column vector $\boldsymbol{w_j}$ of $\boldsymbol{W}$ represents an individual species j. We can combine these species embeddings with a location embedding $f_{\boldsymbol{\theta}}(\boldsymbol{x})$ via an inner product to compute the probability that the species j is present at $\boldsymbol{x}$. At inference time, we compute the presence probabilities for all species in the training set, for all locations c in the set of available context locations $\mathcal{C}^t$ for our target species t. We then produce a new species embedding $\boldsymbol{w}_t$ by taking

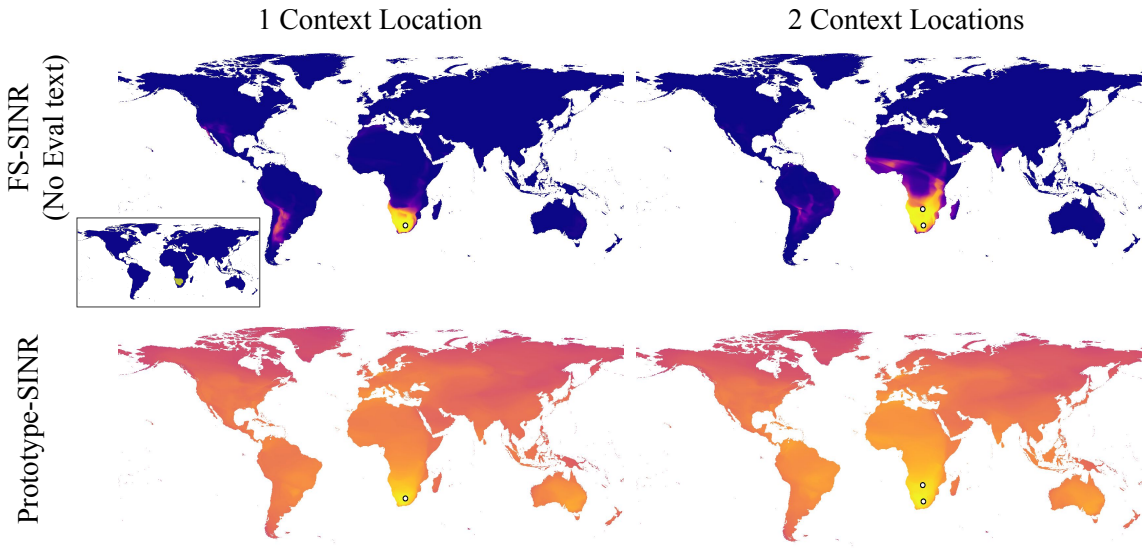


Figure A11. **Qualitative comparison of the Prototype SINR baseline.** Here we compare the predictions of our FS-SINR approach (without any text) and our Prototype SINR baseline on the `Kalahari Scrub-Robin` species that is found in Southern Africa. The Prototype SINR approach obtains an MAP of 0.54 and 0.79, for one and two context locations respectively, while FS-SINR obtains 0.79 and 0.85. As MAP is tied to the ranking of predicted probabilities rather than their absolute values, it can remain high even if the model is somewhat overconfident across the board. As long as the highest probabilities consistently align with areas where the species is truly present, the model will achieve a strong MAP, which we can see with the predictions from Prototype SINR.

a weighted average of the existing $\boldsymbol{w}_j$'s where the weight for each is the product of the probabilities of presence for that species:

$$\boldsymbol{w}_t = \sum_{j=1}^{s} P(\boldsymbol{w}_j | \mathcal{C}^t) \boldsymbol{w}_j. \tag{6}$$

We can then use this new species embedding for our target species to produce a probability of presence for any location x as in SINR (Cole et al., 2023). We present few-shot results using this method in Figure 3. We see that the performance of the Active SINR approach is competitive with FS-SINR when provided with no text, though worse than FS-SINR when provided with this additional context. However, increasing the number of provided context locations beyond a small number actually hurts performance, as it is unable to accurately represent the range of a previously unseen species via the weighted combination of those from the training set.

C.3. Evaluation

We perform three runs for each experiment using different initial random seeds and report the mean. We display the standard deviation as error bars in our figures. For all evaluations across SINR, LE-SINR, Prototype SINR, Active SINR, and FS-SINR, the same set of context locations are used for a given species, and these context locations are accessed in the same order, so all evaluations using five context locations are performed with the same five points, and four of those points are those used for evaluations using four context locations, etc. In our few-shot setting, we use at most 50 context locations during both training and evaluation.

D. Additional Ablations

Here we present additional results to investigate the impact of a variety of design choices and training procedures for FS-SINR. We present plots on a "Symlog" scale, where a linear scale is used between 0 and 10, in order to allow us to show zero-shot results alongside few-shot results. We display the mean of three runs with standard deviations shown as error bars and also show just the mean values alongside for easier interpretation.

D.1. Ablating Training Context Locations

In Figure A12 we show 'Range Text' evaluation performance on the IUCN dataset for FS-SINR models trained using different amounts of context information at training time. We see that generally increasing the context used during training improves performance, and that having a fixed number of context locations is also beneficial.

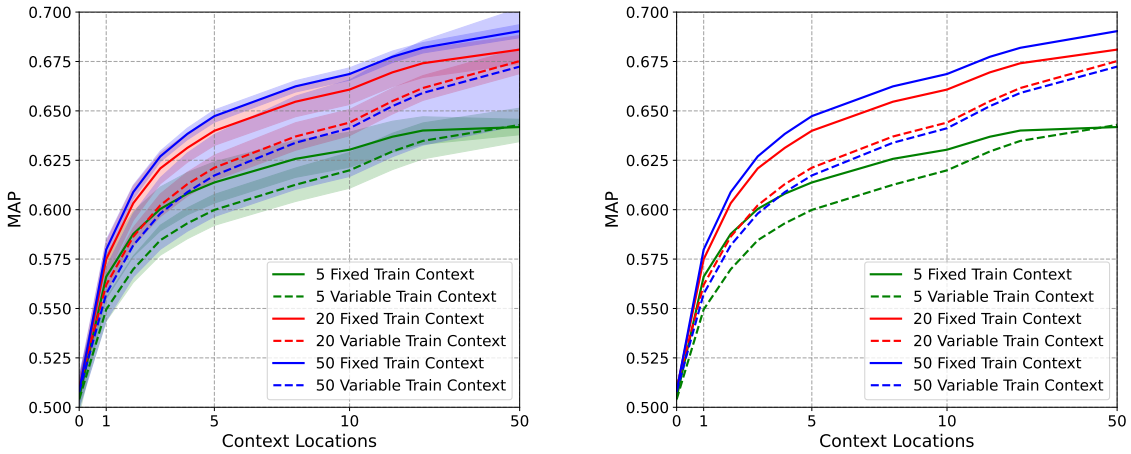


Figure A12. **Impact of number of training context locations.** Here we evaluate FS-SINR models trained using different numbers of context locations. Results are shown with standard deviations from three runs (left), and without (right) for clarity. Evaluation is performed with 'Range Text' on the IUCN dataset. 'Fixed' indicates the same number of context locations were provided for every training example. 'Variable' indicates that a uniform random distribution of context locations up to the specified number were provided with each training example. We see that 'Variable' generally under-performs compared to 'Fixed' and that increasing the train context length tends to increase evaluation performance.

D.2. Ablating Context Information

In Figure A13 we display 'Range Text' evaluation performance on the IUCN dataset for FS-SINR models trained using different combinations of text and context locations during training. We observe that good text-only zero-shot performance requires sometimes providing just text as context information during training. This forces the model to learn to produce ranges from only text information. Models that are sometimes provided with both text and locations for the same training examples perform best as the number of provided context locations increases. We also see that models trained without text can perform on par with those that see text during training when enough context locations are provided (5 - 10). As we might expect, models that are provided with token types they have not seen during training perform poorly.

D.3. Ablating Input Features

In Table A5 we provide additional zero-shot results expanding on those in Table 1 from the main paper. Specifically, we add comparisons to using a different location encoder (i.e., SATCLIP (Klemmer et al., 2025) instead of SINR), comparisons to using a DINOv2 pre-trained image encoder (DINOv2-large) (Oquab et al., 2024), comparisons to using the environmental covariates as in SINR (Cole et al., 2023) that contain information about a locations' climate in addition to the spatial coordinates.

D.4. Ablating Location Encoder

In Figure A14, we vary the number of datapoints used to pre-train the SINR encoder used in FS-SINR. For both FS-SINR and the SINR baseline, we generally observe that more data is better, and for SINR approaches we see that pretraining the encoder is much better than randomly initializing it. We also show results for a SINR model trained on evaluation species in addition to train species. As we saw in Table 1 for FS-SINR, the impact of training the location encoder with evaluation species is small.

In Figure A15, we also investigate the impact of changing the location encoder entirely. We see that replacing our SINR location encoder with a pre-trained and frozen 'SATCLIP' location encoder (Klemmer et al., 2025) significantly harms performance. This may be due to this model being frozen and trained on tasks that do not completely match ours. In

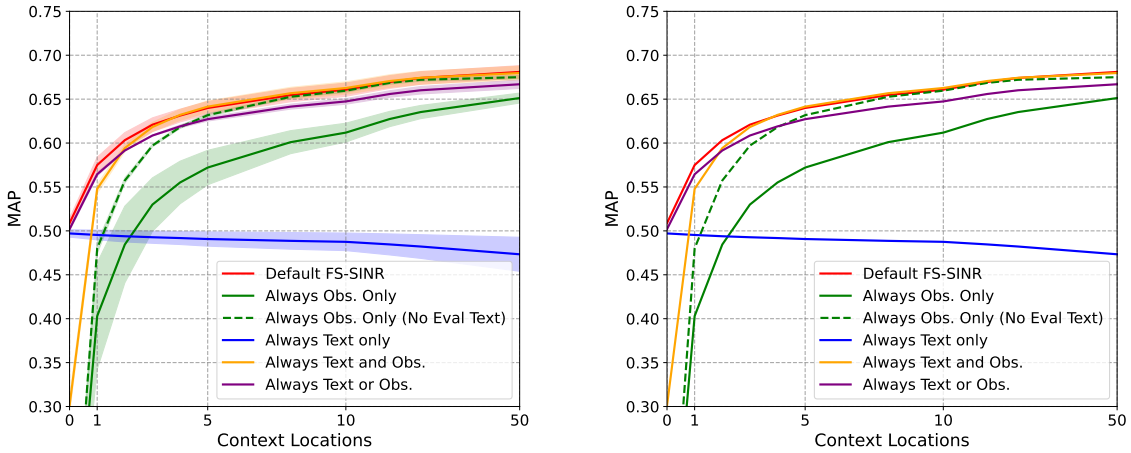


Figure A13. **Impact of train context information.** Here we evaluate FS-SINR models trained using different context information on the IUCN dataset. Results are shown with standard deviations from three runs (left), and without (right) for clarity. Evaluation is performed with 'Range Text' unless 'No Eval Text' is specified, in which case just locations are provided during eval. 70% of training examples for 'Default FS-SINR' provide both location and text context, 20% provide just locations 10% and provide just text. 'Always Obs. Only' has only seen locations during training. 'Always Text Only' has only seen Text during training. 'Always Text and Obs' is always provided with both locations and text during training. 'Always Text or Obs.' is provided with just locations for 90% of training examples, and just text for the remaining 10%.

comparison, a randomly initialized and untrained SINR backbone performs almost identically well as one that has seen a small amount of training data (10 examples per-species in the train set). We also investigate replacing the learned location encoder $f()$ with a simple form of Fourier feature encoding (Tancik et al., 2020) to encode location inputs to the transformer $m_{\psi}()$. In this setting, a pre-trained and fine-tuned SINR type location encoder $f()$ is still used to encode evaluation locations x to determine the probability of presence of species j via the inner product between the species embedding vector $\boldsymbol{w}_j$ and $f(x)$. However, $f()$ is not used to encode the context locations $\mathcal{C}^t$ before they are passed to $m_{\psi}()$. Using these two different encoders performs increasingly poorly as the amount of context information increases.

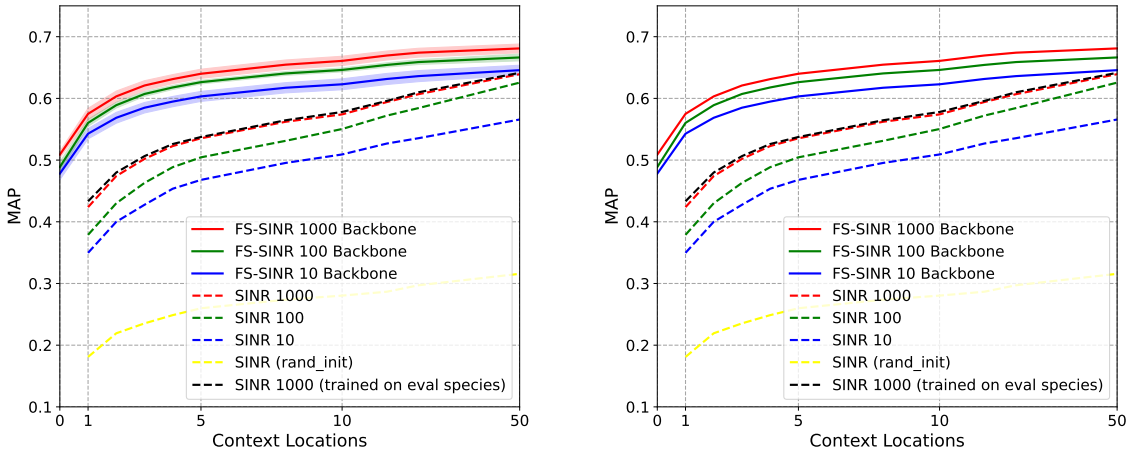


Figure A14. **Impact of Location Encoder Training.** Here we evaluate the performance of SINR and FS-SINR models when the size of the training dataset for the SINR backbone is varied. Results for FS-SINR models are shown with standard deviations from three runs (left), and without (right) for clarity. Evaluation on FS-SINR is performed with 'Range Text', while SINR can only make use of location data. '1000', '100', and '10' represent the maximum number of examples per class the SINR backbone was trained on. 'SINR (rand_init)' is initialized with random weights and is not trained. '(trained on eval species)' indicates training on all training and evaluation species.

D.5. Ablating Training Data

In Figure A16 we vary the number of examples per-species that are provided during training. The impact of this is fairly small, with models trained on an intermediate amount of data performing best. It is worth noting, that not all species in the

Table A5. **Additional zero-shot results**. We report zero-shot performance where no location information is provided to each model, comparing SINR (Cole et al., 2023), LE-SINR (Hamilton et al., 2024), and variants. We denote additional metadata as: **EN** for additional environmental covariates (Fick & Hijmans, 2017) used in Cole et al. (2023), **HT** for 'Habitat Text', **RT** for 'Range Text', **I** for 'Image' using our default EVA-02 image encoder, **I (DINOV2)** for Image using a DINOV2 based image encoder (Oquab et al., 2024), **TST** for 'Test Species in Train', **TRT** for using full taxonomic rank text, **SATCLIP** for where the SINR encoders are replaced with the image derived location encoders from Klemmer et al. (2025), and **P-LE-SINR** for 'Prototype LE-SINR'. Results are reported as MAP, where higher is better.

(a) Methods without additional environmental covariates

Method	Variant	IUCN	S&T
TST (test species in train)			
SINR	TST	0.67	0.77
FS-SINR	HT, TST	0.38	0.59
FS-SINR	RT, TST	0.55	0.67
With SATCLIP encoder			
FS-SINR	HT, SATCLIP	0.20	0.43
FS-SINR	RT, SATCLIP	0.33	0.55
Prototype SINR			
P-LE-SINR	HT	0.23	0.48
P-LE-SINR	RT	0.40	0.55
LE-SINR			
LE-SINR	HT	0.28	0.52
LE-SINR	RT	0.48	0.60
FS-SINR			
FS-SINR		0.05	0.18
FS-SINR	TRT	0.21	0.34
FS-SINR	HT	0.33	0.53
FS-SINR	RT	0.52	0.64
FS-SINR	I	0.19	0.38
FS-SINR	I (DINOv2)	0.13	0.28
FS-SINR	I + RT	0.46	0.64
FS-SINR	I (DINOv2) + RT	0.46	0.62

(b) Methods with additional environmental covariates

Method	Variant	IUCN	S&T
TST (test species in train)			
SINR	EN, TST	0.76	0.81
FS-SINR	HT, EN, TST	0.38	0.61
FS-SINR	RT, EN, TST	0.57	0.67
LE-SINR			
LE-SINR	HT, EN	0.31	0.52
LE-SINR	RT, EN	0.51	0.61
FS-SINR			
FS-SINR	EN	0.07	0.64
FS-SINR	HT, EN	0.32	0.53
FS-SINR	RT, EN	0.51	0.65

training dataset have as many as 1000 observations. We find that a model trained on only 10 examples per-species performs significantly worse.

D.6. Ablating FS-SINR Architecture

In Figure A17 we vary the underlying FS-SINR architecture. Removing different components has a small effect on model performance, with the removal of the species decoder actually improving results when range text is provided. However, as several ablations perform very similarly, it is difficult to tease out the how much of this effect is due to variance. It is clear however that removing the learnable token type embeddings causes the model to completely fail to learn during training. In Figure A18 we show further ablations based around removing the learned location encoder for inputs to the transformer and replacing it with the simple Fourier feature encoding also seen in Figure A15. When this is removed, other ablations seem to further harm performance, although results for these ablations vary significantly between runs.

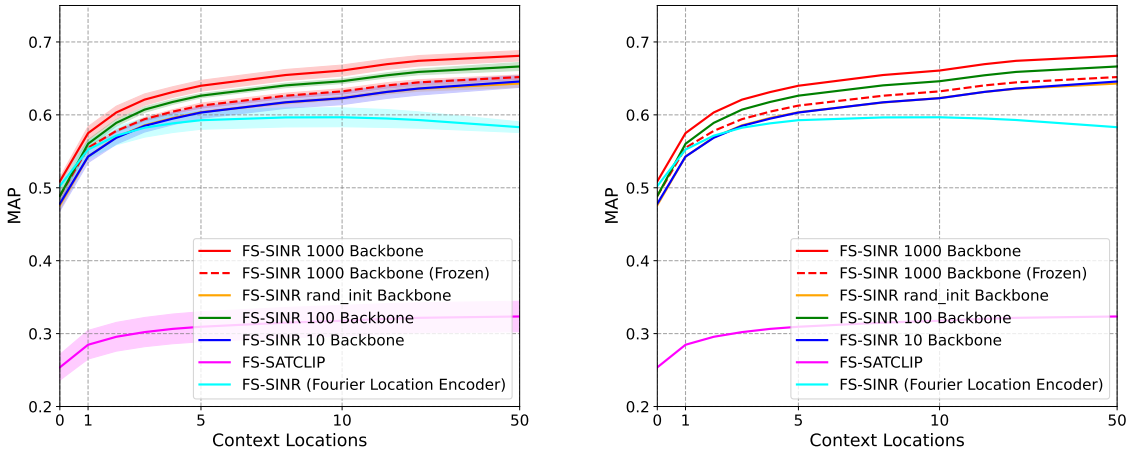


Figure A15. **Impact of location encoder.** Here we evaluate the performance of FS-SINR style models with different location encoders. Results are shown with standard deviations from three runs (left), and without (right) for clarity. Evaluation is performed with 'Range Text' on the IUCN dataset. '1000', '100', '10' represent the maximum number of examples per class the SINR backbone was trained on. '(Frozen)' indicates that the location encoder parameters were not updated during FS-SINR training. 'FS-SATCLIP' replaces the SINR location encoder with a pretrained, frozen location encoder from Klemmer et al. (2025). 'FS-SINR (Fourier Location Encoder)' uses the simple Fourier feature encoding (Tancik et al., 2020) used in Mildenhall et al. (2021) to match the 256 dimensional outputs of the SINR location encoders. These outputs are used directly as inputs to the transformer encoder. After a species token is produced in this way, it is attached to a pre-trained and fine-tuned SINR backbone to produce a range.

E. Additional Qualitative Results

E.1. Visualizing Non-Species Concepts

By jointly training on text and locations, FS-SINR is able to spatially ground abstract non-species concepts in a zero-shot manner, as is done with LE-SINR in Hamilton et al. (2024). In Figure A19 we provide another example similar to Figure 4 in the main paper. Here, we again fix the context location and show the impact of changing the text. We can see that different text prompts can result in quite different predicted ranges. In Figure A20 we see examples where different text concepts, which are very different from the species-based text provided during training, are grounded in sensible locations on the map. In Figure A21 we compare predictions made with increasing numbers of context locations in desert regions, with or without the accompanying text prompt "Desert". As we increase the number of context locations, the two different models converge to more similar range predictions.

E.2. Visualizing Estimated Species Ranges

Here, we provide additional examples of the ranges produced by FS-SINR using context locations, text, and images. In Figures A22 and A23 we visualize FS-SINR range estimates for two different species when habitat or range text is provided. We observe that the combination of text and context locations seem to result in better estimates of the range. In Figure A24 we show additional zero-shot image-only examples, where FS-SINR is provided a single image from a held-out test species at inference time. Again, we observe some plausible range predictions even with such limited input data.

In Figure A26 we show range estimates for the `Brown-banded Watersnake`, using 'range' text for FS-SINR and LE-SINR approaches. In Figure A27 we show range estimates for the `Brown-headed Honeyeater`, using 'habitat' text for FS-SINR and LE-SINR approaches. Finally in Figure A28 we show range estimates for the `Crevice Swift`, without providing text. Overall, SINR produces more diffuse ranges and requires more locations to narrow down the range. LE-SINR and FS-SINR appear to have very different zero-shot behaviors, with LE-SINR frequently seeming to predict presence in almost no locations at all, while FS-SINR tends to produce a zero-shot range that is too large.

In Figure A25 we visualize FS-SINR range predictions for the `Yellow-footed Green Pigeon` for models that have had different random initializations (i.e., different random seeds). We observe that there is a relatively large amount of variance in the outputs produced given the same input data. The same set of input context locations could represent many different possible output ranges, and thus being able to represent this variety is advantageous. We also utilize these different predictions from different seeds for ensembling and uncertainty quantification in Appendix A.

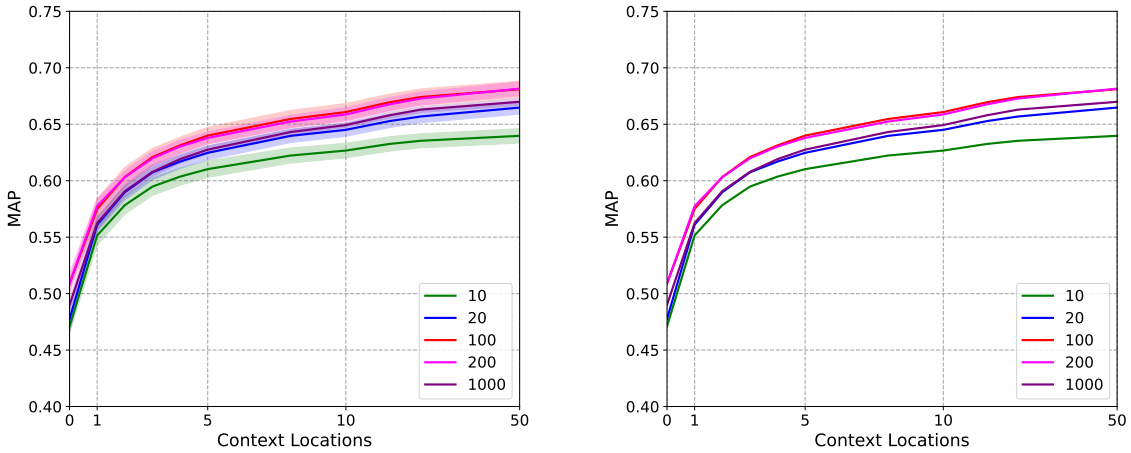


Figure A16. **Impact of training data.** Here we evaluate FS-SINR models trained with different amounts of data. Results are shown with standard deviations from three runs (left), and without (right) for clarity. Evaluation is performed with 'Range Text' on the IUCN dataset. The labels show the maximum number of examples per-species that FS-SINR is trained on. We see that training on an intermediate amount of training data leads to best performance.

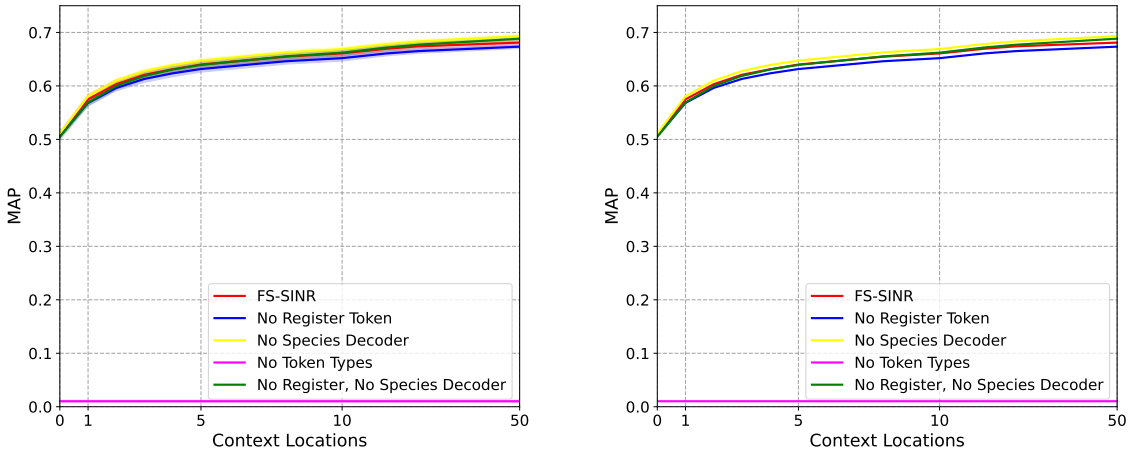


Figure A17. **Ablating model architecture components.** Here we evaluate the performance of FS-SINR style models as we ablate various design choices. Results are shown with standard deviations from three runs (left), and without (right) for clarity. Evaluation is performed with 'Range Text' on the IUCN dataset. We see small changes in performance when removing the register token and the species decoder. However removing the learned token type embeddings has a large impact.

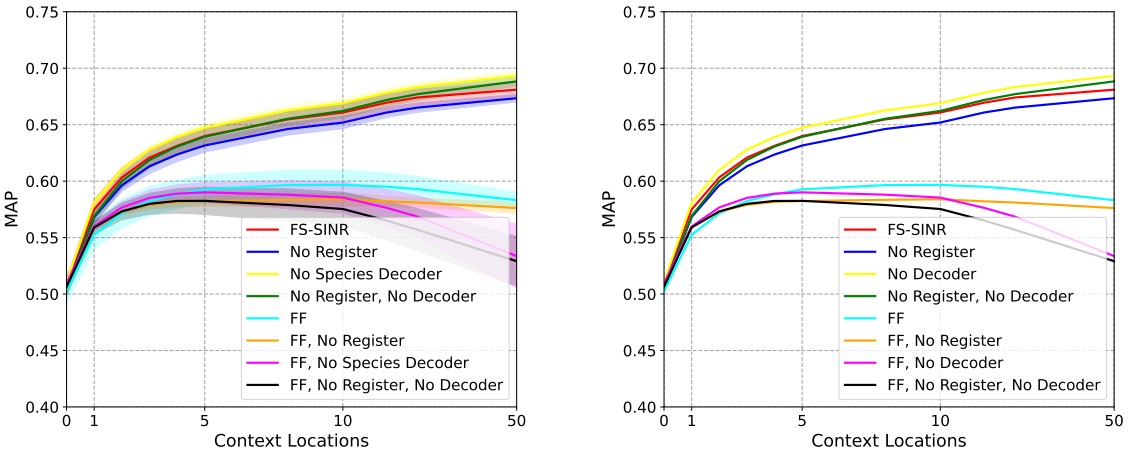


Figure A18. **Further ablating model components.** Here we evaluate the performance of FS-SINR style models as we ablate more components. Results are shown with standard deviations from three runs (left), and without (right) for clarity. Evaluation is performed with 'Range Text' on the IUCN dataset. 'FF' indicates that the model does not use a SINR backbone to encode location inputs to the transformer encoder. Instead, a simple Fourier feature encoding (Tancik et al., 2020) used in Mildenhall et al. (2021) is used to increase the dimensionality of location data to match the token dimension of the transformer encoder. These are used directly as inputs to the transformer encoder. After a species token is produced in this way, it is attached to a standard SINR backbone to produce a range. Removing the SINR backbone for encoding inputs to the transformer has a large impact on performance, especially when more context locations are supplied, and makes the model more sensitive to the impact of other ablations.

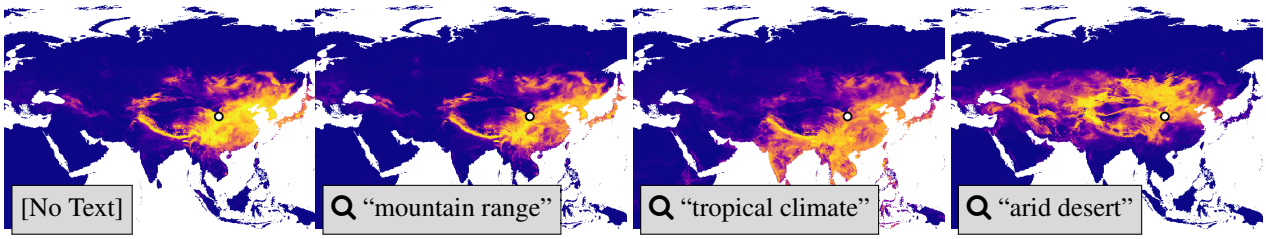


Figure A19. **Controlling range predictions using a single context location and text**. Here we show another example similar to Figure 4 in the main paper. Given the same context location, denoted as '○', FS-SINR can produce significantly different range predictions depending on the text provided. This example illustrates a use case where a user may have limited observations but some additional knowledge regarding what type of habitat a species of interest could be found in.

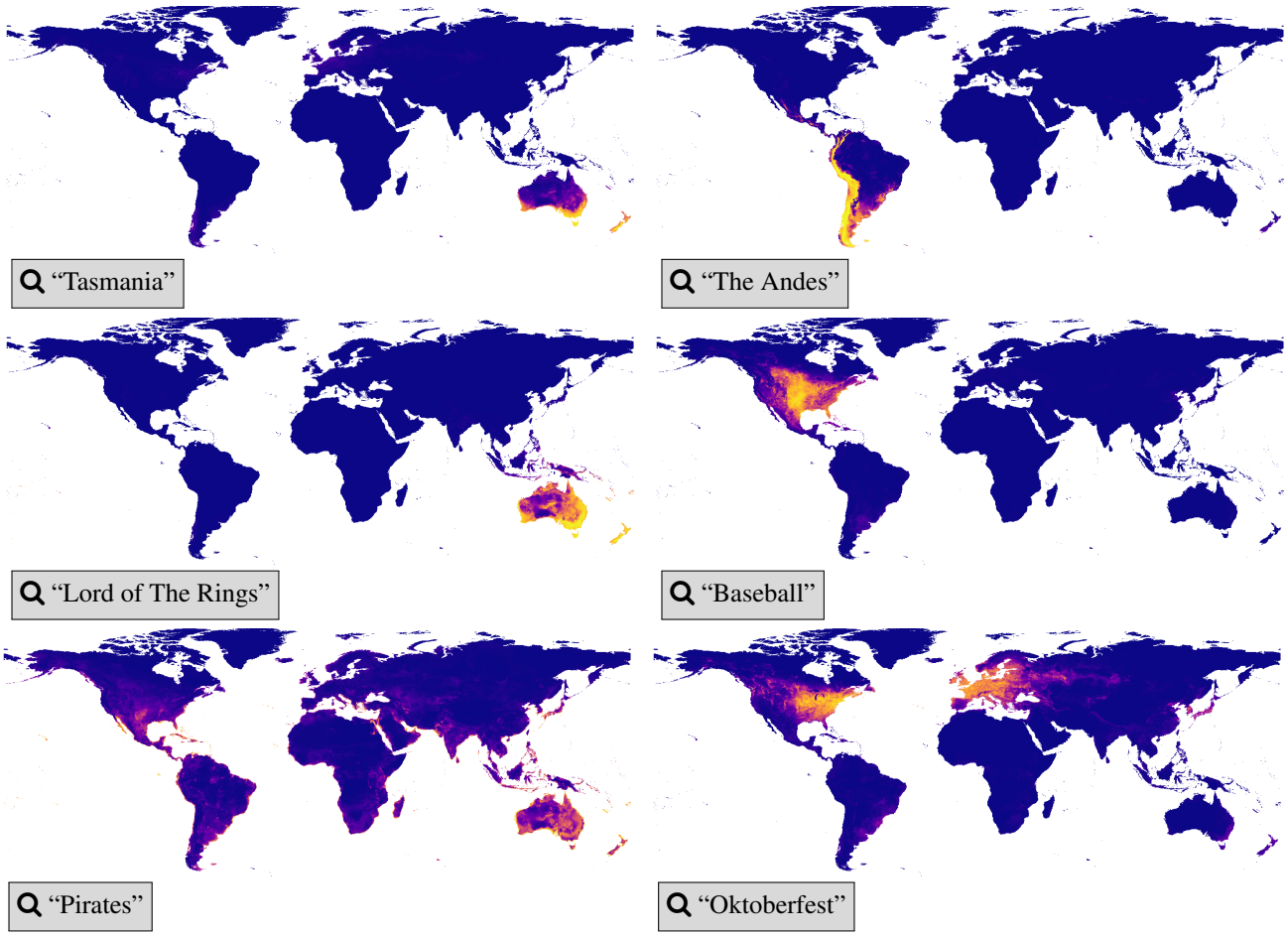


Figure A20. **Zero-shot non-species concepts**. We can evaluate FS-SINR in a zero-shot manner using only text information, i.e., without any locations. Here, we observe that FS-SINR, like LE-SINR (Hamilton et al., 2024), can localize abstract (i.e., non-species) concepts in geographic space, despite never being trained to explicitly do so. The model achieves this as it learns to make connections between species text and information already contained in the pretrained language encoder we use. However, we do note failure/ambiguous cases such as the “Pirate” example in the bottom row.

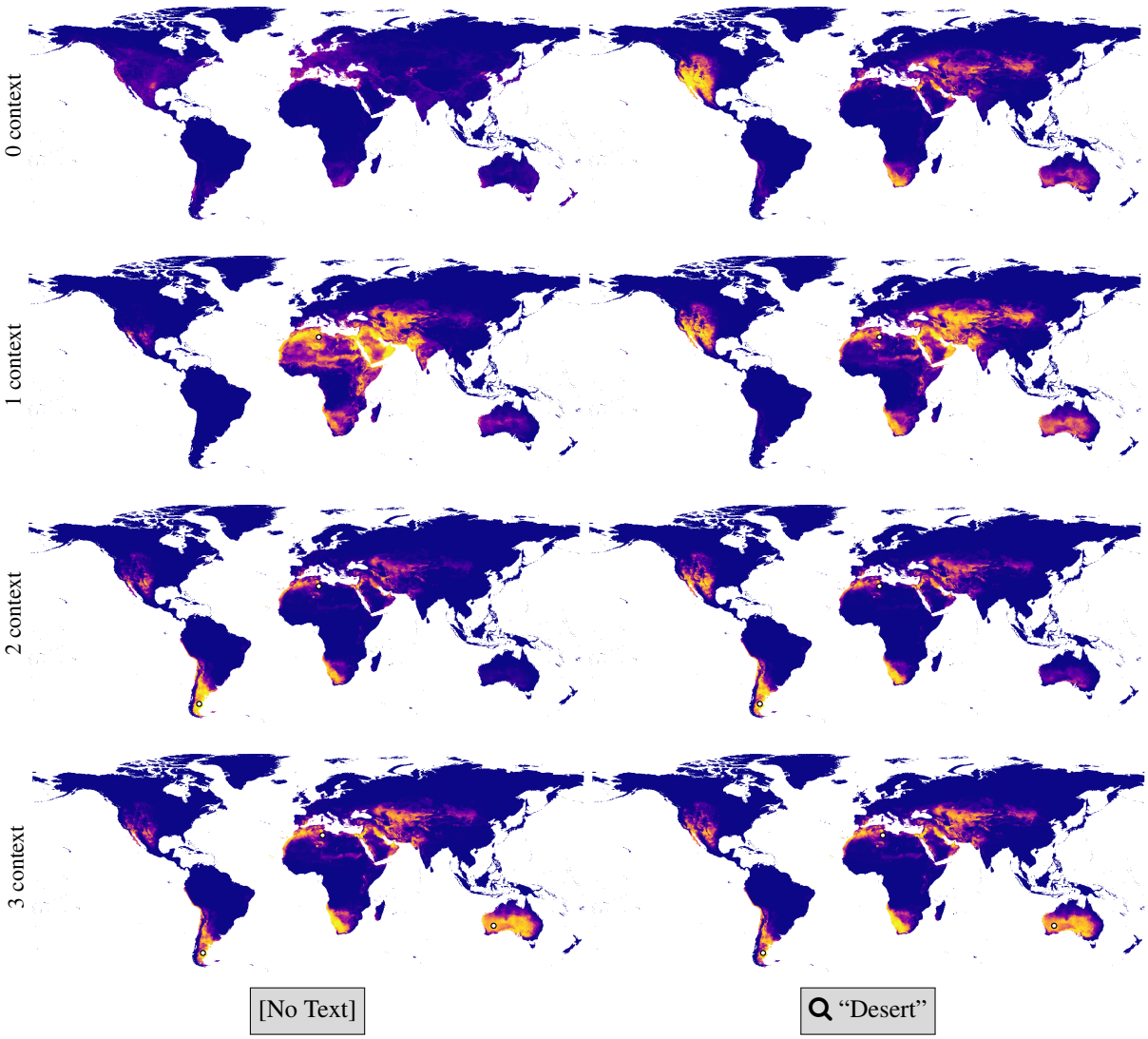
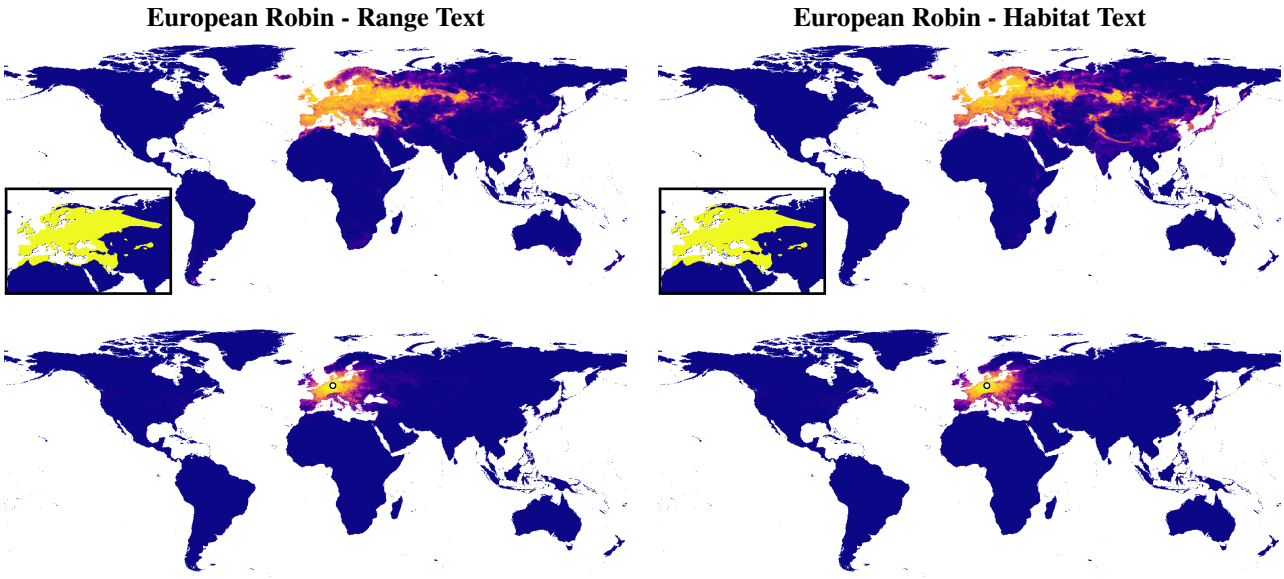


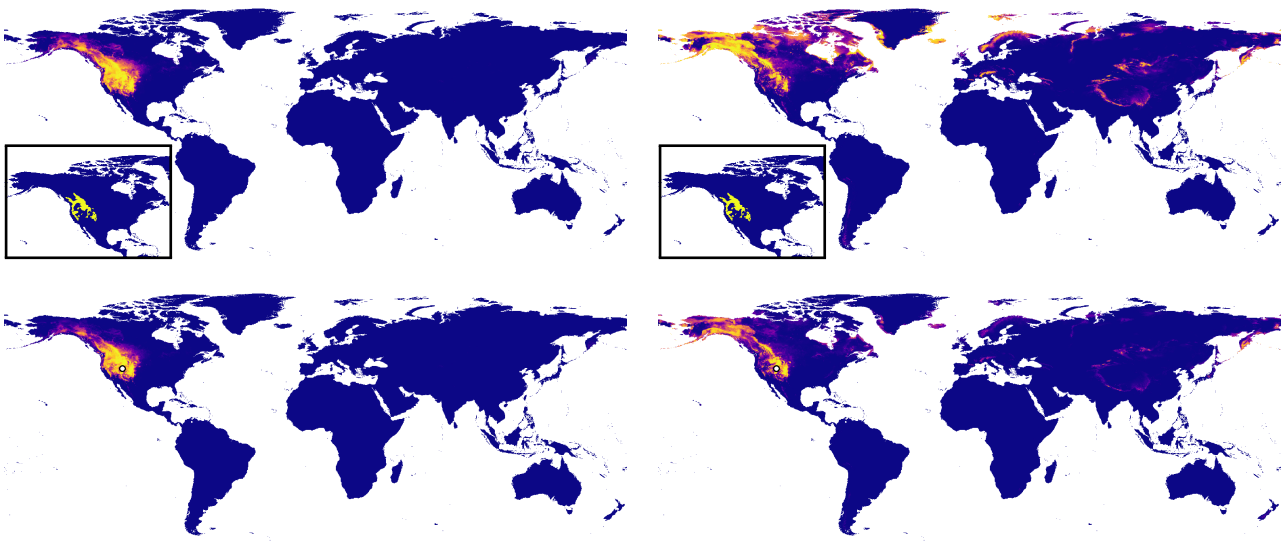
Figure A21. **Varying the context information provided**. Here we change the context information provided to FS-SINR. The model on the left column receives no text input, but the one on the right gets the text "Desert". Additionally, in each row we increase the number of context locations provided, from zero to three, denoted as '○'. We observe that the model on the right that uses text already has a strong prior about the species being present at desert-like locations, e.g., see first row where no context locations are provided. As soon as one context location is added in North Africa (second row), the model generates a new prediction with an increased probability that the species is present there.



Range Text: The European robin is found across Europe, east to Western Siberia and south to North Africa; it is sedentary in most of its range except the far north. It also occurs in the Atlantic islands as far west as the Central Group of the Azores and Madeira. It is a vagrant in Iceland and has been introduced to other regions, including North America and Australia, but these introductions were unsuccessful.

Habitat Text: The European robin inhabits a variety of habitats, including gardens, parks, woodlands, and forests. It prefers areas with dense vegetation and is often found near human settlements. It is also found in mountainous regions and can be seen in urban areas, such as cities and towns.

Figure A22. **Using text descriptions.** Here we illustrate the zero-shot (top row) and one-shot (bottom row) FS-SINR range estimations based on text descriptions for the `European Robin`, using 'Range' (left), and 'Habitat' (right) text, shown below the range estimates. Expert-derived range maps are shown inset in the top row.



Range Text: The American pika (Ochotona princeps) is found in the mountains of western North America, usually in boulder fields at or above the tree line, from central British Columbia and Alberta in Canada to the US states of Oregon, Washington, Idaho, Montana, Wyoming, Colorado, Utah, Nevada, California, and New Mexico.

Habitat Text: Pikas inhabit talus fields that are fringed by suitable vegetation in alpine areas. They also live in piles of broken rock. Sometimes, they live in man-made substrate such as mine tailings and piles of scrap lumber. Pikas usually have their den and nest sites below rock, around 20-100 cm (8-39 in) in diameter, but often sit on larger and more prominent rocks.

Figure A23. **Using text descriptions.** Here we illustrate the zero-shot (top row) and one-shot (bottom row) FS-SINR range estimations based on text descriptions for the `American Pika`, using 'Range' (left), and 'Habitat' (right) text, shown below the range estimates. Expert-derived range maps are shown inset.

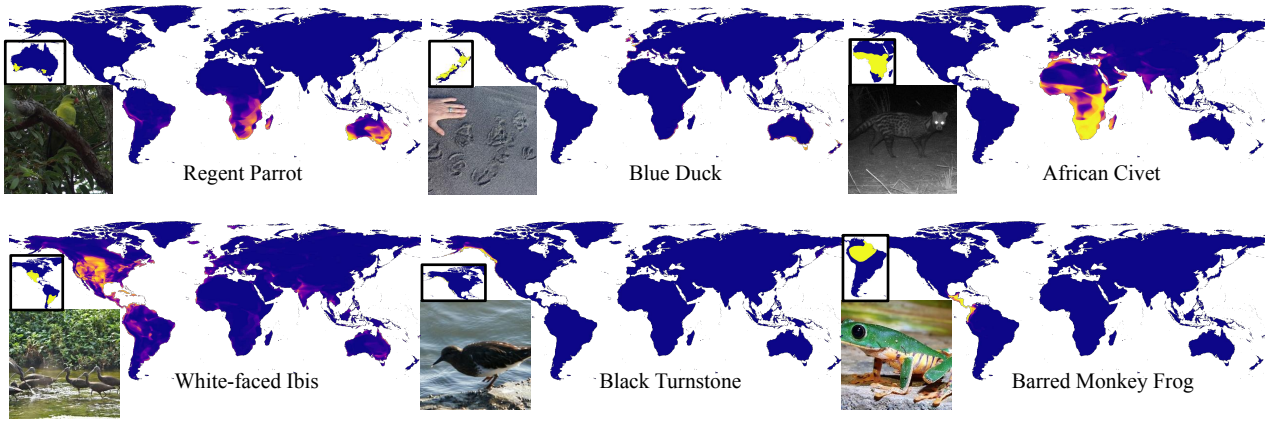


Figure A24. **Image zero-shot range estimation.** Here we see zero-shot range estimates for six species in the IUCN evaluation dataset, with expert-derived range and image inset. The `blue duck` image taken from iNaturalist (2025) only shows evidence of the species from footprints in wet sand. We see that this image generates predictions in coastal areas in various locations around the globe. The coastal background for the `Black Turnstone` could have helped the model to generate a relatively accurate prediction on the northwest coast of North America.

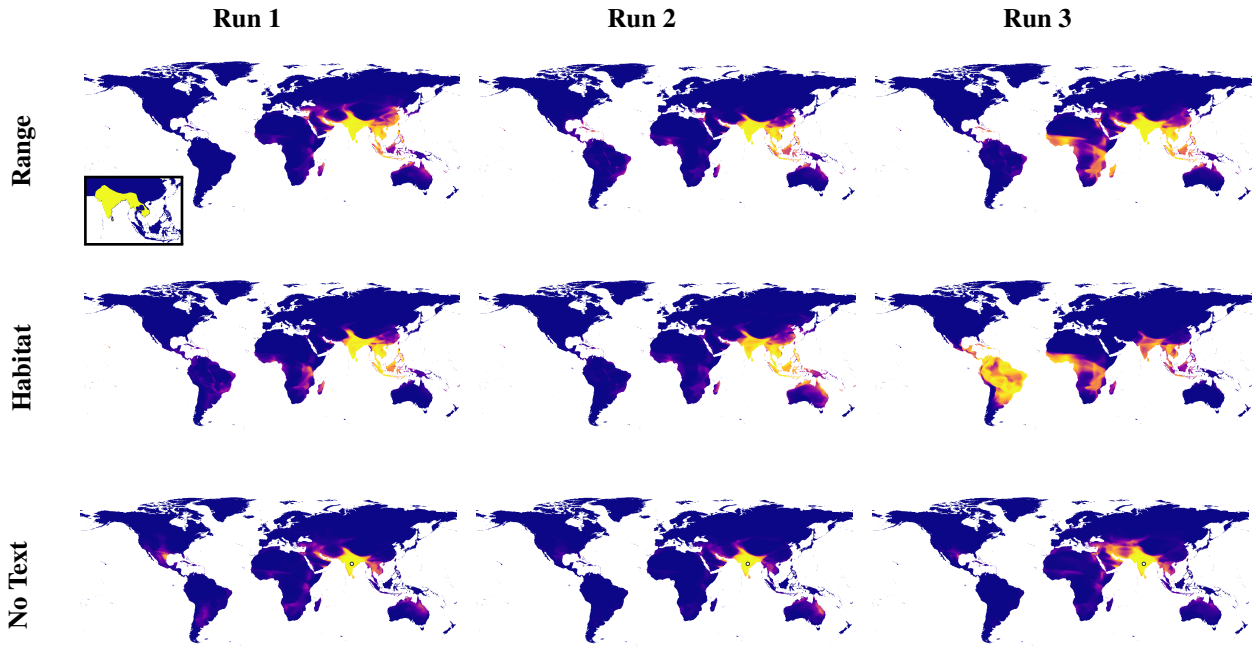


Figure A25. **Impact of random initialization on FS-SINR.** Here we display range estimates for the `Yellow-footed Green Pigeon` from three different FS-SINR models where different random seeds were used to initialize each model during training. We show zero-shot results using 'range text' (top) and 'habitat text' (middle), and also few-shot results using one context location with no text (bottom). The IUCN expert-derived range is shown inset. We see that even when provided with the same inputs, different models can perform very differently when this input is very sparse (e.g., just text or one context location). While most of the Indian part of the actual range is included for all input types and runs, there is significant variability across the runs in other geographic areas.
Range Text: "The yellow-footed green pigeon is found in the Indian subcontinent and parts of Southeast Asia. It is the state bird of Maharashtra."
Habitat Text: "The species is a habitat generalist, preferring dense forest areas with emergent trees, especially Banyan trees, but can also be spotted in natural remnants in urban areas. They forage in flocks and are often seen sunning on the tops of trees in the early morning."

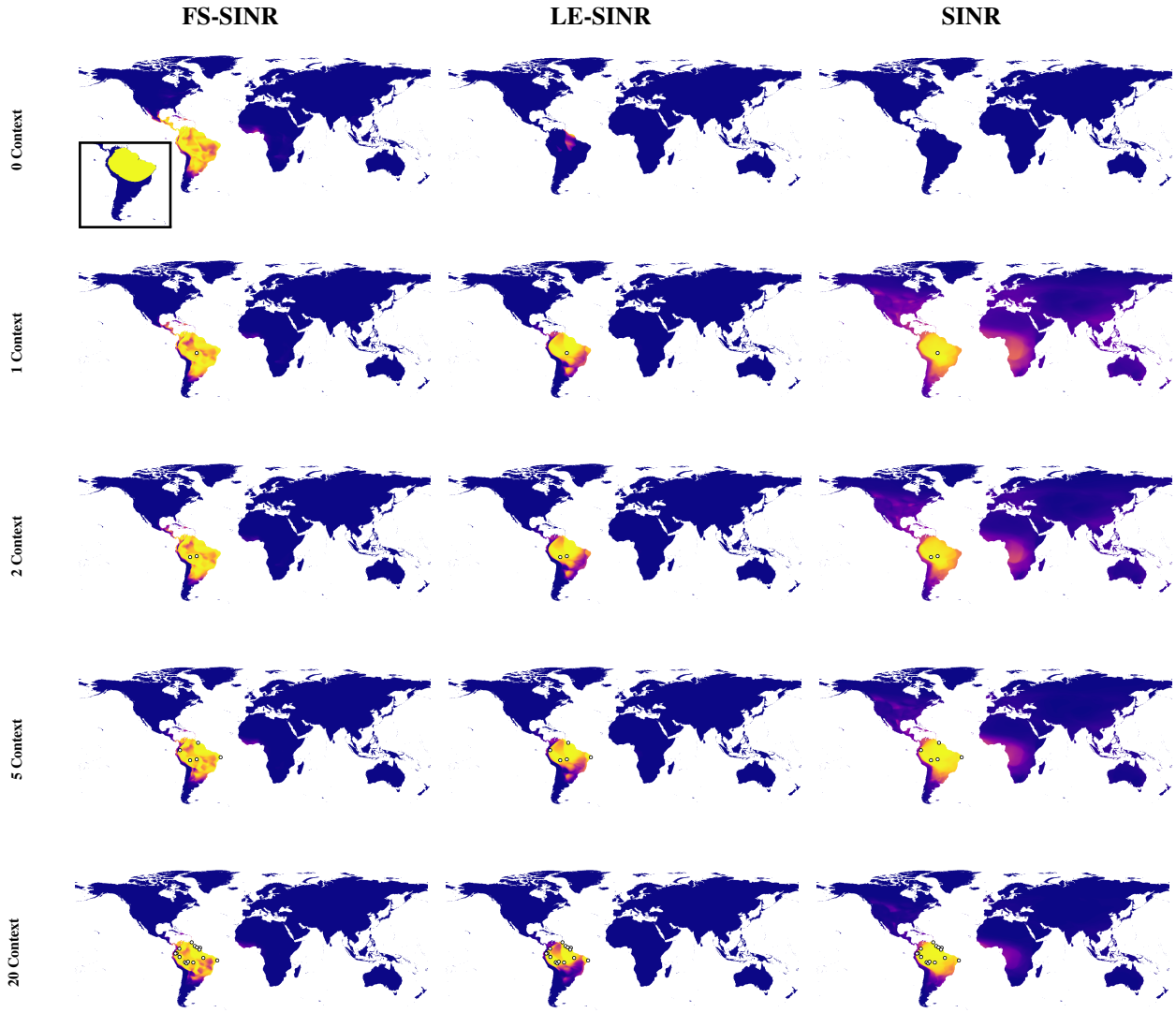


Figure A26. **Comparing estimated ranges across models with context lcoations and range text.** Here we see zero-shot and few-shot range estimates produced by FS-SINR, LE-SINR, and SINR for the `Brown-banded Watersnake`, with expert-derived range inset. We provide range text to FS-SINR and LE-SINR as well as context locations, but SINR is not capable of accepting text and so we show a blank map for the zero-shot range estimate. We see that LE-SINR underestimates the range using only text, while FS-SINR overestimates it. SINR requires more location data than the other approaches to localize the range to South America.
Range Text: "The Brown-banded water snake (Helicops angulatus) is found in tropical South America and Trinidad and Tobago."

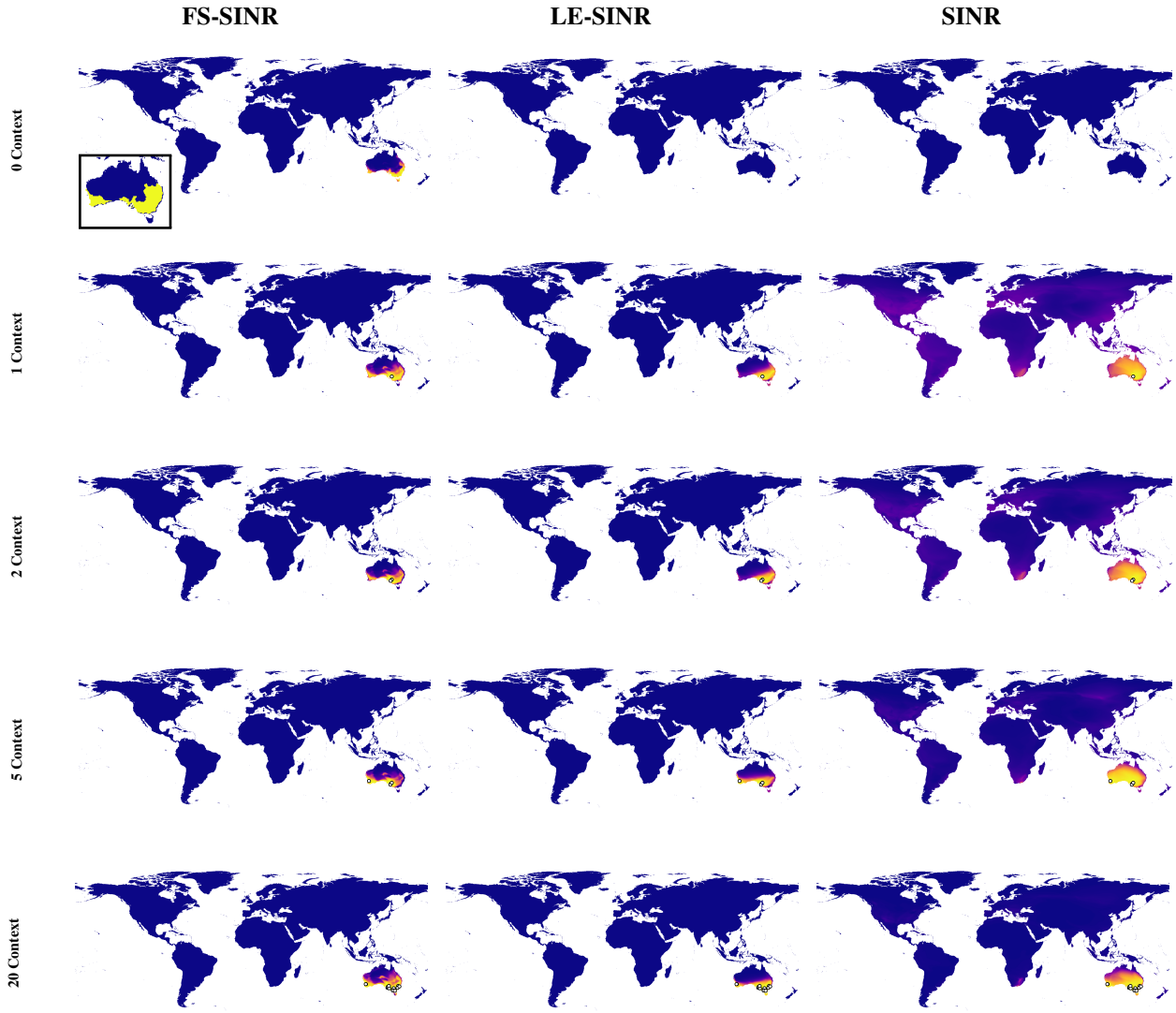


Figure A27. **Comparing estimated ranges across models with context locations and habitat text.** Here we see zero-shot and few-shot range estimates produced by FS-SINR, LE-SINR, and SINR for the `Brown-headed Honeyeater`, with expert-derived range inset. We provide habitat text to FS-SINR and LE-SINR as well as context locations, but SINR is not capable of accepting text and so we show a blank map for the zero-shot range estimate. We again see LE-SINR underestimate the range using only text, while FS-SINR has very good zero-shot performance for this species. We see that SINR again requires more location data to narrow down the range and even after 20 locations the range is still significantly larger than the other models, extending into South Africa.
Habitat Text: "The brown-headed honeyeater inhabits temperate forests and Mediterranean-type shrubby vegetation. It is typically found in tall trees, where it forages by probing in the bark of trunks and branches."

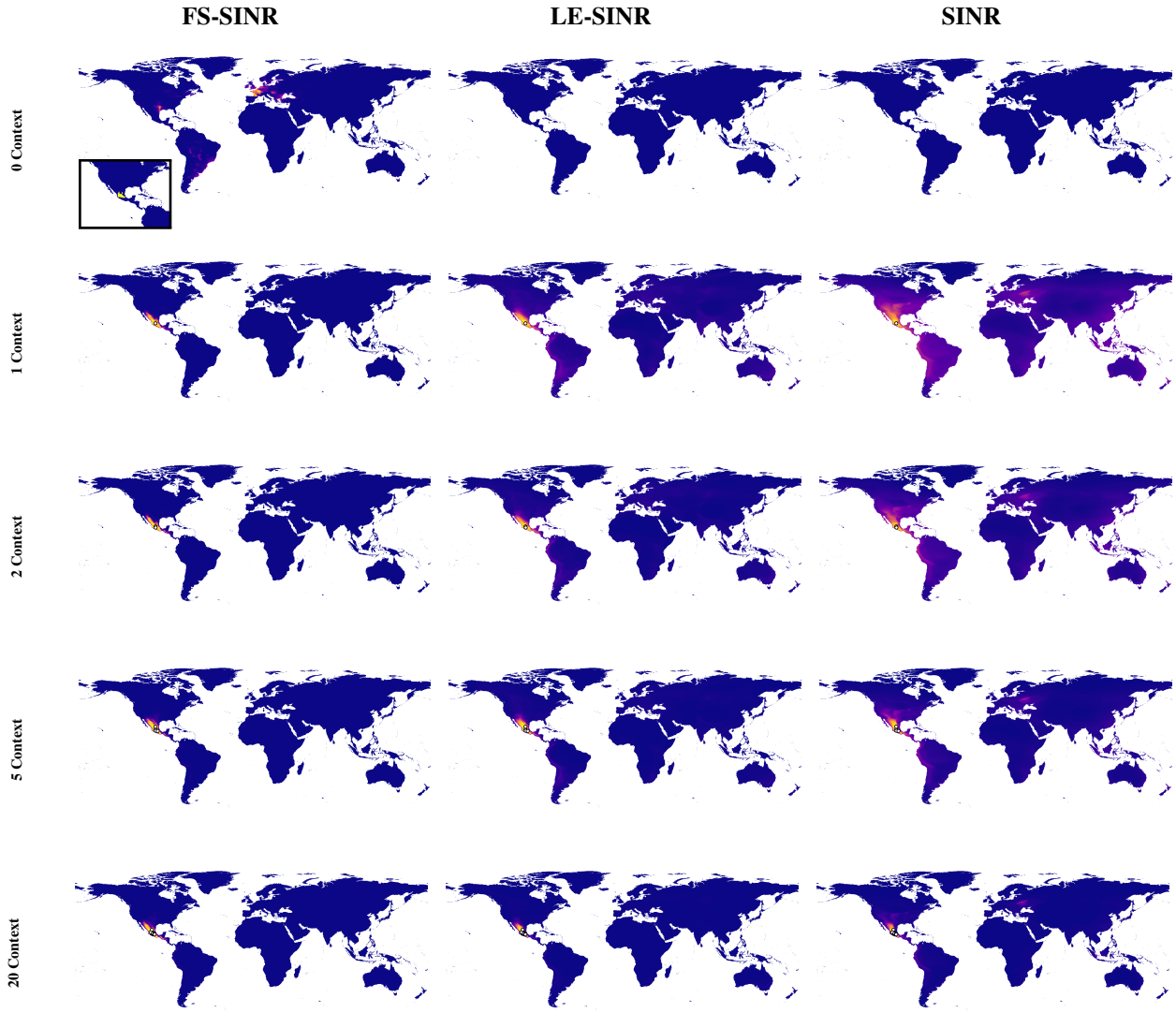


Figure A28. **Comparing estimated ranges across models.** Here we see few-shot range estimates produced by FS-SINR, LE-SINR, and SINR for the `Crevice Swift` lizard, with expert-derived range in Mexico inset. No text is provided and so no sensible zero-shot prediction can be made for any model. However while LE-SINR and SINR cannot produce an output for this and so we show a blank map, FS-SINR can generate a predicted range just from feeding the learned CLS and register tokens with no other information into the transformer encoder. The range that is produced is contained within the model or the learned tokens itself rather than from any further inputs. Absent additional information, the model seems to guide predictions towards areas where it as seen many species during training e.g., Europe and North America. This may be an unhelpful bias when attempting to model novel species. SINR again produces more diffuse ranges than the other methods, though all approaches struggle to model these small ranges, as seen in Appendix B.2.