STEM-LM: Spatio-Temporal Ecological Modeling via Masked Language Model for Joint Species Distribution

Jacky Kaiyuan Li[*1], Wonseop Lim[*2], Fiona Margaret Callahan[3], Levi Yoder Raskin[2], Maya Lemmon-Kishi[3], Rasmus Nielsen[2,3,4,5]

[1]Biostatistics Division, University of California, Berkeley
[2]Department of Integrative Biology, University of California, Berkeley
[3]Center for Computational Biology, University of California, Berkeley
[4]Department of Statistics, University of California, Berkeley
[5]Globe Institute, University of Copenhagen

{li_jacky,david9456,fiona_callahan,
levi_raskin,maya_lemmon-kishi,rasmus_nielsen}@berkeley.edu

[*]Equal contribution

Abstract

Joint species distribution models (JSDMs) are central to biodiversity forecasting and conservation decision-making. As ecological datasets grow in size, dimensionality, and spatio-temporal resolution, there is a need for flexible yet scalable JSDMs tailored to large-scale species observation data. Recent advances in masked language modeling for text and genomics suggest a natural alternative: by treating each species' presence or absence as a token, and a site's species assemblage together with its spatio-temporal and ecological covariates as a sentence, we can learn joint co-occurrence structure by reconstructing masked species from their neighboring sites. We propose STEM-LM[1], a Transformer-based JSDM that frames joint species distribution modeling as masked language modeling. By varying the masking rate during training, a single trained model supports both purely spatio-temporal/ecological prediction and conditioning on arbitrary subsets of observed species for joint co-occurrence inference at a given site. On a North American butterfly and a global plant distribution dataset, STEM-LM performs better or on par with other statistical and deep-learning based methods in terms of discriminative ranking, while producing substantially better rank-calibrated occurrence probabilities. Utilizing partial species observations at the same site greatly enhances prediction performance.

1 Introduction

Understanding the distribution of species and how they interact with one another and with the environment is central to biodiversity forecasting and to making conservation decisions. Species distribution models (SDMs), a class of models that relate species distribution data to environmental or ecological patterns, are useful for making such predictions [1, 2]. While early studies modeled the relationship between a single species and environmental or ecological variables (e.g., [1, 3, 4]), incorporating multiple species through co-occurrence patterns into a "joint" species distribution model (JSDM) often improves prediction [5–7]. However, existing JSDMs typically handle only a handful of species or make sparsity assumptions about species co-occurrence patterns [8–11]. Beyond co-occurrence, accounting for the spatio-temporal structure of distributions is similarly shown to be useful, particularly for migratory taxa and climate-driven range shifts, and numerous statistical

[1]Code available at `https://github.com/JackyKaiyuanL/STEM-LM`

works have developed spatio-temporal SDMs that explicitly model how species occurrence and co-occurrence evolve in space and time [12–18].

Both biotic and abiotic ecological interactions exhibit nonlinear, complex dynamics [1, 2], making machine learning an appealing choice for modeling them [2, 19–22]. Early applications, such as those based on maximum entropy [23, 24], decision tree-based models [25–27], and shallow artificial neural networks [28, 29], showed some promise for the use of machine learning in SDM, and complex models have been shown to have higher predictive accuracy [30]. More recently, application of deep neural network architectures (e.g., [2, 21, 31, 32]) to massive species observation datasets proved useful in making predictions of species distribution considering complex interaction not only between species and environment but also between species in a scalable manner. However, even with increasingly large datasets, the best models are only moderately successful at predicting species presences for spatially separated test data [33].

Existing work on deep-learning-based SDM has mostly focused on incorporating new modalities (e.g., [32, 34–42]) or using deep neural networks as a drop-in inside a parametric backbone model (e.g., [43–45]). A smaller set of fully neural models pushes instead on the learning problem itself: Brun et al. [46] scale a multispecies deep neural network ensemble to millions of citizen-science observations with ranking-based losses, MaskSDM [47, 48] masks environmental variables to obtain variable-subset flexibility, and CISO [49] conditions on a partial set of known species via state embeddings. Yet none of these conditions on the communities observed at other sites in space and time, and at most encode time or space as an input feature [19, 46, 50, 51]; existing biotic conditioning, where present, is restricted to species at the target location, leaving the spatio-temporal structure that statistical spatio-temporal SDMs capture explicitly outside the deep-learning framework.

In this work, along the lines of these studies, we present STEM-LM (Spatio-Temporal Ecological Modeling via masked Language Model), a scalable Transformer [52]-based JSDM that frames joint species distribution modeling within a masked language modeling framework [53]. STEM-LM uses associations among species and environmental covariates, and explicitly utilizes the spatio-temporal structure of species distributions to make presence–absence predictions. By varying the species masking rate during training, a single trained model can support prediction based on environmental and spatio-temporal information, and also conditioning on arbitrary subsets of observed species. To our knowledge, STEM-LM is the first deep SDM to combine masked-species prediction with explicit cross-attention over neighboring sites in space and time, allowing biotic conditioning to be drawn from other observations rather than being restricted to the target site.

2 STEM-LM model

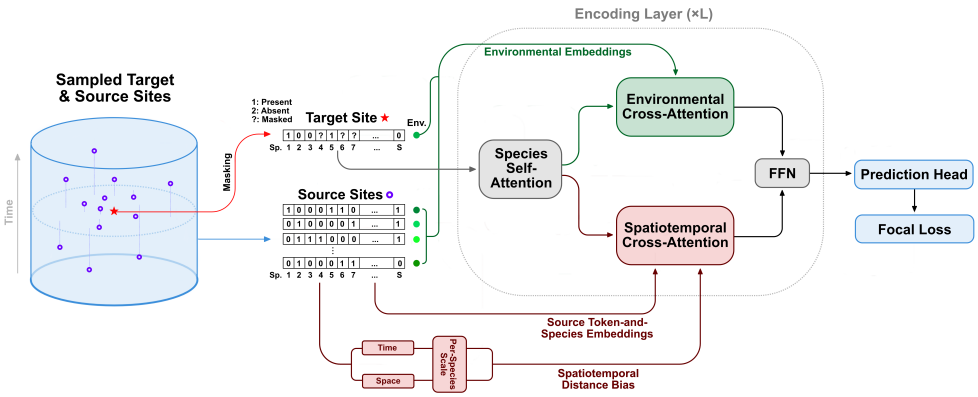

Figure 1: Overview of the architecture of STEM-LM.

STEM-LM is an encoder-only Transformer [52]-based joint species distribution model that predicts the presence or absence of species at a given location and time, conditioned on other species' states, spatio-temporal context from nearby sites, and environmental covariates. The architecture adapts the row self-attention and column cross-attention structure typical of alignment-based genomic

language models (e.g., [54–56]): each row is presence–absence data for each observation at the spatio-temporal site, and each column corresponds to a species. Species self-attention lets the target site's presence–absence observation (target row) attend within itself, and two parallel cross-attention pathways bring in context from nearby source sites: in which each species (column) of the target site attends to the source sites' species embeddings, and an environmental pathway in which each species attends to a small set of learned environmental-group embeddings pooled from the source sites' environmental covariates (Fig. 1). Unlike vision Transformers [57] or voxel-based 3D representations (e.g., 3D CNNs [58] or video Transformers [59]) that operate on a fixed grid, the permutation-invariant attention [60] over irregular source sites is suited to sparse, irregularly distributed ecological occurrence data. We use a masked language modeling [53] framework, dynamically masking a random portion of species at the target observation, and training the model to reconstruct them.

Input data and source site sampling We assume M observations, each with spatial coordinates, time, E environmental covariates, and presence–absence of S species. For each target site i, we sample a set of N source sites (default $N = 64$) whose species observations the model uses as context. Because absolute spatial and temporal ranges of nearby sites vary across regions and seasons, we rescale distances per target so that one target's notion of "nearby" is comparable to another site. Specifically, for each target i we first draw a pool of N sites with probability proportional to $1/(d_{ij}^{\text{sp}} + \epsilon)$ (ϵ a small constant for numerical stability) and take the median spatial and temporal distances from i within this pool as local scale parameters $s_{\text{sp}}(i)$ and $s_{\text{tp}}(i)$. We then form the normalized spatio-temporal distance: $d(i,j) = \sqrt{\left(d_{ij}^{\text{sp}}/s_{\text{sp}}(i)\right)^2 + \left(d_{ij}^{\text{tp}}/s_{\text{tp}}(i)\right)^2}$, and draw N source sites with probability proportional to $1/(d(i,j) + \epsilon)$. The first pool is used only for scale estimation and discarded.

Tokenization and input encoding Each species' state is tokenized as absent (0), present (1), or masked (2), and mapped via a shared state embedding $\mathbf{W}_{\text{state}} \in \mathbb{R}^{3\times H}$, where H is the model's hidden size (default $H = 256$). A learned species identity embedding $\mathbf{W}_{\text{sp}} \in \mathbb{R}^{S\times H}$ is added so that species are distinguishable regardless of state: $\mathbf{h}_s^{(0)} = \mathbf{W}_{\text{state}}[y_s] + \mathbf{W}_{\text{sp}}[s]$, where $y_s \in \{0,1,2\}$ is the tokenized state of species s. No positional encoding is used along the species axis, as column order is arbitrary. Source-site environmental covariates are projected to $\mathbb{R}^H$ and pooled into G group embeddings (default $G = 5$) via G learned group queries that cross-attend over the projected covariates [61], grouping correlated covariates. The target site's own environmental vector is separately projected through a two-layer MLP: an input layer normalization [62], a linear projection to $\mathbb{R}^H$, a SiLU [63] activation, a second linear map $\mathbb{R}^H \to \mathbb{R}^H$, and an output RMSNorm [64]. This target-environmental embedding is concatenated with the G source-pooled group embeddings to form the final environmental context $\mathbf{C}_{\text{env}} \in \mathbb{R}^{(G+1)\times H}$.

Main architecture The model consists of L identical layers (default $L = 4$), each with pre-RMSNorm [64, 65] residual sub-blocks. Each attention sub-block uses $A/2$ heads (default $A = 8$). Within each layer, three attention sub-blocks operate in sequence.

(1) Species self-attention. At the target site, the S species embeddings attend to one another; queries, keys, and values are all computed from the target row, yielding per-head $S \times S$ attention maps over species pairs. This pathway is how cross-species co-occurrence structure is captured in our model; the spatio-temporal cross-attention below operates at the per-species level.

(2) Spatiotemporal cross-attention. For each species s, the target site token attends to that same species' embeddings at the N source sites. This restriction avoids an $S \times S \times N$ attention tensor; cross-species signal from neighboring sites is recovered through species self-attention and through residual mixing across layers. Analogous to the phylogenetic distance-based inductive bias of GPN-Star [56], attention logits are augmented with a learned per-species FIRE [66] distance bias, computed by passing a fixed transformation of the pairwise distance through a small learned MLP:

$$b_{s,t,n} = e^{\alpha_s}\text{FIRE}_{\text{sp}}(D_{t,n}^{\text{sp}}) + \mathbf{1}_{\text{temporal}} \cdot e^{\beta_s}\text{FIRE}_{\text{tp}}(D_{t,n}^{\text{tp}}),$$

where α_s, β_s are learned per-species log-scales (initialized to zero), and $\text{FIRE}(d) = \text{MLP}(\phi(d))$, where each MLP is a one-hidden-layer MLP of width H_{FIRE} (default 32) with SiLU activation and scalar output. The fixed transformations $\phi_{\text{sp}}, \phi_{\text{tp}}$ are built from a log-monotone scalar $\psi_{c,d_{\max}}(d) =$

$\log(cd+1)/\log(cd_{\max}+1)$, with $d_{\max}$ the maximum pairwise distance in the dataset:

$$\phi_{\text{sp}}(d) = \psi_{c_{\text{sp}},\, d_{\max}^{\text{sp}}}(d) \in \mathbb{R},$$
$$\phi_{\text{tp}}(d) = \big(\psi_{c_{\text{tp}},\, d_{\max}^{\text{tp}}}(d),\, \cos(\omega_1 d), \sin(\omega_1 d), \ldots, \cos(\omega_K d), \sin(\omega_K d)\big) \in \mathbb{R}^{1+2K},$$

where $c_{\text{sp}}, c_{\text{tp}} > 0$, ω_k, and the MLP weights are learnable. The ω_k are initialized at ecologically meaningful periods (annual, semi-annual), and the MLP weights acting on the $\sin$ / $\cos$ components of ϕ_{tp} are initialized to zero. When the model is configured as purely spatial, the temporal term is dropped.

(3) Environmental cross-attention. In parallel, each target species attends to the $G+1$ environmental context embeddings ($\mathbf{C}_{\text{env}}$). Because queries carry species identity, different species attend to environmental embeddings with different weights, learning species-specific habitat preferences implicitly.

The two cross-attention outputs are summed and added to the residual stream after a shared pre-RMSNorm on the species-self-attention output. Each layer concludes with a SwiGLU [67] feed-forward network of intermediate size F (default $F = 2H$).

(4) Prediction head. To let the model learn an idiosyncratic linear response to the environment for each species beyond what the shared cross-attention captures, after L layers, the target-row hidden states are projected to per-species logits by a linear head augmented with a parallel per-species environmental head: $\mathbf{e}^\top \mathbf{A}\mathbf{B} + \mathbf{b}_{\text{sp}}$, where $\mathbf{e} \in \mathbb{R}^E$ is the target site's raw environmental vector, $\mathbf{A} \in \mathbb{R}^{E\times r}$ and $\mathbf{B} \in \mathbb{R}^{r\times S}$ form a low-rank factorization (default $r = 8$, $\mathbf{A}$ zero-initialized), and $\mathbf{b}_{\text{sp}} \in \mathbb{R}^S$ is a per-species bias.

Masking and data splits For each target row, a fraction p of species is selected uniformly at random and replaced with the mask token. p can be a fixed scalar, or sampled per target from Uniform$[0, 1]$ so that a single model learns to predict from any degree of partial observation. We split observations into training/validation/test sets ($80/10/10$ by default) via spatial blocking [68] to prevent spatial autocorrelation from inflating held-out performance. For geographic coordinates, entire cells of Uber's H3 hexagonal grid [69] at resolution 2 ($\sim$158 km edge hexagon) are assigned to a single split. We do not additionally apply temporal blocking: H3 spatial blocking already ensures that no site appears in both train and test, regardless of year, eliminating the dominant source of observation-level leakage.

Training The model minimizes sigmoid focal loss [70, 71] ($\gamma = 2.0$, $\alpha = 0.25$ following the suggestion of [70, 71]), to counter the heavy class imbalance between presence and absence labels. We optimize with AdamW [72, 73] (learning rate 10^{-4}, weight decay 0.01 applied to linear weights only; biases, normalization layers, and per-species log-scales are excluded from decay) and cosine annealing [74] decaying to $\eta_{\min} = 10^{-6}$, with specified batch size, gradient clipping [75] at norm 1.0, and dropout [76] of 0.1. At each epoch, we evaluate the validation split under four masking probabilities $p \in \{0.25, 0.5, 0.75, 1.0\}$ and save the model maximizing per-species AUROC averaged across species and across the four p (instead of AUPRC, following the suggestion of McDermott et al. [77]), so that selection is not biased toward any single deployment regime. The model is implemented in PyTorch [78].

Inference We report mean per-species AUROC, continuous Boyce index (CBI) [79, 80], and expected calibration error (ECE) [81] on the test split. We mainly focus on AUROC and CBI metrics: AUROC measures discrimination and is the standard ranking metric; CBI is the rank-calibration metric widely used in ecology, testing whether predicted-to-expected presence ratios increase monotonically with the predicted score. Note that "calibration" here means rank-monotonicity, not absolute-probability accuracy.

Metrics are computed under two masking schemes evaluated separately. The first matches training: a fraction $p \in \{0.25, 0.5, 0.75, 1.0\}$ of species is masked uniformly at random. The second simulates presence-only deployment: all absences are masked, plus a fraction p of presences. Predictions at each p are averaged over $K = 10$ passes that share the masking pattern but resample source sites. Because focal-loss training distorts the absolute logit scale, we optionally apply post-hoc temperature scaling [82], which divides the logits by a single positive scalar $T^\star$ fit on validation data at p=1.00 by maximizing log likelihood, and applied at every test p. We fit a single $T^\star$ at p=1.00 rather than

per-p, since at lower p, the deployment-time masking rate is usually unknown. Temperature scaling preserves the rank ordering of predictions and rescales the sigmoid output to better match empirical positive rates, substantially reducing ECE, which measures the absolute gap between predicted probabilities and observed positive rates.

3 Experiments & benchmarks

3.1 Dataset preparation

We used two publicly available species distribution datasets: the fully spatio-temporal eButterfly [83] (accessed via GBIF[2]), covering continental North America from 2011–2025 with S=173 butterfly species across 17,077 complete checklists, and the spatial-only sPlotOpen [84] (v. 2.0), covering the globe with S=1,201 plant species across 95,104 vegetation plots and treated as temporally static given the slow dynamics of broad-scale plant distributions. Both were augmented with climatic/phenological (ERA5-Land [85] daily 2 m air temperature and total precipitation, and MOD13Q1 [86] NDVI and EVI for eButterfly; WorldClim [87] v. 2.1 for sPlotOpen), pedologic (SoilGrids [88] v. 2.0) and elevation (Copernicus GLO-30 DEM [89, 90]) variables. Full preparation details are in Appendix A.

3.2 Benchmarking and ablation study

We benchmarked STEM-LM against widely used statistical SDM baseline models: logistic regression [91], Maxnet [92], and generalized additive model (GAM) [93] as well as two deep-learning-based SDMs: MaskSDM [47, 48] and CISO [49]. Logistic and GAM are each fit per species in three variants: environment-only (subscript env), spatial(-temporal) coordinates only (st on eButterfly, s on sPlotOpen), and both combined (full); Maxnet uses environmental covariates only by design (full specifications in Appendix B). The main STEM-LM run uses the default architecture, training, and inference configuration of Sec. 2: $L = 4$ layers with $A = 8$ attention heads, hidden size $H = 256$, $N = 64$ source sites, $G = 5$ environmental groups, low-rank per-species head $r = 8$, and the default training hyperparameters ($p \sim \mathrm{Uniform}[0, 1]$, focal loss $\gamma = 2.0$, $\alpha = 0.25$). On eButterfly, the FIRE temporal periods were initialized at $\{365, 182, 122, 91\}$ days, providing a flexible basis of annual and sub-annual harmonics that can capture diverse ecological cycles across species; on sPlotOpen, the temporal cross-attention pathway was disabled. Each dataset was trained for 100 epochs (eButterfly) or 50 epochs (sPlotOpen) with a batch size of 128 and bfloat16 mixed precision (per-run compute usage is summarized in Appendix C). All methods used the same training, validation, and test splits. All deep-learning methods were run on three random seeds, and the mean across seeds is reported. STEM-LM at the default configuration has $\sim$3.5 M parameters on eButterfly and $\sim$3.8 M on sPlotOpen.

We conducted an ablation study on the eButterfly dataset to assess the contribution of each cross-attention component, evaluating three variants against the full model: (a) *no spatio-temporal cross-attention* (`no_st`), in which the target site does not attend to neighboring source sites; (b) *no environmental cross-attention* (`no_env`), in which the model has no access to environmental covariates; and (c) *species self-attention only* (`no_st_env`), where both cross-attention modules are removed and the model relies solely on co-occurrence patterns among observed species at the target site. We additionally ablate the number of source sites N sampled per target, comparing $N \in \{32, 64, 128\}$ to identify a good trade-off between predictive performance and computational cost. We also explored the effect of the type of loss function used. Focal loss can be viewed as a binary cross-entropy (BCE) loss that up-weights rare positive examples through a focusing parameter γ, encouraging the model to concentrate on hard, infrequent species. To assess this design choice, we conducted an additional ablation comparing the two losses: focal and BCE loss, on the full model with otherwise identical training and evaluation settings for both eButterfly and sPlotOpen datasets. All runs use the default architecture, training, and inference configuration of Sec. 2.

3.3 Results

Ablation: cross-attention heads Following our model selection criterion, we report test AUROC across masking rates in Table 1 (left). The full model achieves the highest test AUROC across

[2] `https://www.gbif.org/dataset/cf3bdc30-370c-48d3-8fff-b587a39d72d6`; accessed 04/13/2026

all masking rates. Removing the spatio-temporal cross-attention (`no_st`) produces a particularly steep degradation as the masking rate increases, dropping from 0.880 at p=0.25 to 0.803 at p=1.00. Removing the environmental cross-attention (`no_env`) leads to a smaller but consistent drop in AUROC as masking proportion increases. When both cross-attention heads are removed (`no_st_env`), performance collapses substantially, with mean AUROC falling to 0.745 and AUROC at p=1.00 dropping to 0.500 (i.e., random) as expected, since the model has no information about the target site. Together these results show that both cross-attention components are essential to the model.

Table 1: eButterfly ablations: test AUROC by masking rate p (mean ± std). *Left:* cross-attention head ablation. *Right:* source-site count N ablation.

Cross-attention heads

	p=0.25	0.50	0.75	1.00
Full	**0.891**±0.006	**0.891**±0.002	**0.884**±0.001	**0.868**±0.002
`no_st`	0.880±0.004	0.872±0.003	0.853±0.003	0.803±0.006
`no_env`	0.886±0.003	0.888±0.003	0.880±0.002	0.861±0.004
`no_st_env`	0.855±0.003	0.839±0.006	0.786±0.008	0.500±0.000

Source-site count N

N	p=0.25	0.50	0.75	1.00
32	0.888±0.007	0.888±0.002	0.880±0.002	0.861±0.003
64 (default)	0.891±0.006	0.891±0.002	0.884±0.001	0.868±0.002
128	**0.893**±0.005	**0.892**±0.002	**0.886**±0.001	**0.870**±0.001

Ablation: source-site count Increasing N from 64 to 128 yields the highest test AUROC at every masking rate (Table 1, right), but the improvement over our default of $N = 64$ is marginal ($+0.001$ to $+0.002$ AUROC) and within seed variance, while source-site aggregation takes roughly $1.5\times$ more computation time due to the cross-attention structure. Reducing to $N = 32$ lowers AUROC by 0.003–0.007 relative to $N = 64$. We therefore set $N = 64$ as the default in our main experiments.

Table 2: Rarity-stratified per-species AUROC and CBI on eButterfly at p=1.00. Quartiles are over species sorted by training-set presence count; Q1 = rarest (counts: 21–116 / 120–211 / 212–475 / 486–4274; n=44, 43, 43, 43). Mean over species in each quartile; deep-learning methods report mean±std across three seeds.[4] **Bold** = best, underline = second-best per column.

AUROC

Method	Q1	Q2	Q3	Q4
GAM_{full}	0.900	0.880	**0.900**	**0.869**
Maxnet	0.850	0.824	0.836	0.785
CISO	0.842±0.025	0.843±0.012	0.825±0.009	0.733±0.013
STEM-LM (F)	0.898±0.006	0.905±0.004	0.874±0.005	0.803±0.002
STEM-LM (B)	**0.910**±0.006	**0.910**±0.002	0.877±0.005	0.808±0.003

CBI

Method	Q1	Q2	Q3	Q4
GAM_{full}	0.407	0.222	0.458	0.601
Maxnet	0.176	0.098	0.474	0.653
CISO	−0.085±0.112	0.183±0.073	0.395±0.105	0.579±0.079
STEM-LM (F)	**0.743**±0.032	**0.766**±0.025	**0.920**±0.025	**0.968**±0.010
STEM-LM (B)	0.397±0.056	0.485±0.033	0.783±0.038	0.861±0.036

Table 3: Test ECE and post-temperature-scaling (T-scaled) ECE for STEM-LM, by loss and masking rate p (mean±std, three seeds). $T^\star$ is fit on validation data at p=1.00.

Focal

	eButterfly		sPlotOpen	
p	ECE	T-scaled ECE	ECE	T-scaled ECE
0.25	0.070±0.001	0.018±0.001	0.027±0.001	0.006±0.000
0.50	0.068±0.001	0.015±0.001	0.027±0.001	0.005±0.000
0.75	0.067±0.001	0.014±0.001	0.029±0.001	0.005±0.000
1.00	0.065±0.003	0.013±0.000	0.038±0.002	0.006±0.000
	$T^\star$=0.557±0.012		$T^\star$=0.494±0.014	

BCE

	eButterfly		sPlotOpen	
	ECE	T-scaled ECE	ECE	T-scaled ECE
	0.017±0.000	0.019±0.000	0.005±0.000	0.005±0.000
	0.014±0.000	0.016±0.000	0.005±0.000	0.004±0.000
	0.012±0.000	0.015±0.001	0.004±0.000	0.004±0.000
	0.012±0.000	0.014±0.000	0.005±0.000	0.005±0.000
	$T^\star$=1.161±0.022		$T^\star$=0.985±0.027	

Ablation: loss type On the eButterfly dataset, BCE achieves a marginally higher test AUROC for the full model, with the gap of roughly 0.006–0.009 consistent across all masking rates (Table 4). However, focal loss yields a substantially better rank-calibrated model: across all masking levels, test CBI rises with an average improvement of 0.195, with the largest gain at p=0.25 (0.628 → 0.837; Table 4). The trade-off is therefore strongly asymmetric: focal loss accepts a small loss in discriminative ranking in exchange for a large gain in rank-calibration.

Stratifying by training-set presence count exposes how methods diverge with rarity. AUROC is consistently elevated in the rarer quartiles and lowest in the most-common quartile in every method (Table 2): the well-known class-imbalance inflation effect for rare-positive classes [94, 95], where

[4] MaskSDM is omitted because per-species predictions were not retained.

abundant easy-negative pairs inflate the rank-ordering score, so AUROC should not be over-interpreted in rarity-stratified comparisons. CBI runs the opposite direction (lowest for rare species in every method) and is where focal training substantively wins: in the rarest quartile ($\leq$116 presences), focal loss-trained STEM-LM reaches CBI 0.74, nearly double both BCE-trained STEM-LM (0.40) and the strongest baseline GAM_{full} (0.41), with Maxnet (0.18) and CISO (-0.09) far below.

The trade-off also surfaces in absolute calibration: focal loss inflates uncalibrated ECE ($\sim$0.07 vs. $\sim$0.012 for BCE; Table 3) because the same down-weighting shifts predicted probabilities away from observed positive rates without changing their rank ordering [96]. Post-hoc temperature scaling preserves rank ordering and substantially reduces ECE on focal-trained models, with the post-temperature-scaling ECE (T-scaled ECE; test ECE after dividing logits by $T^\star$) uniformly $\sim 5\times$ smaller than the uncalibrated ECE across all mask rates (Table 3). This makes focal plus temperature scaling a better operating point than BCE when both rank discrimination and absolute calibration are required.

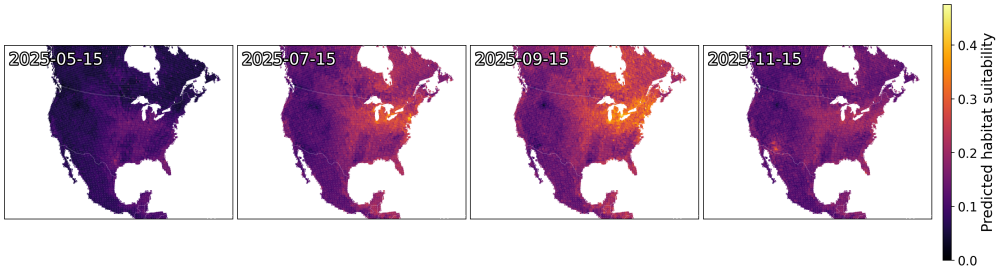

Figure 2: Predicted distribution of monarch butterfly (*Danaus plexippus* (Linnaeus 1758)) across North America in 2025 at four temporal snapshots (May, July, September, November), inferred from the full STEM-LM with focal-loss training and temperature scaling. Color encodes per-cell predicted score (habitat suitability) on a 0.5° grid.

eButterfly Table 4 reports test AUROC and CBI on the eButterfly dataset across masking rates $p \in \{0.25, 0.50, 0.75, 1.00\}$. STEM-LM matches or marginally exceeds all baselines on AUROC and substantially outperforms them on CBI. At p=1.00, where the model relies solely on spatio-temporal and environmental context (matching the setting of traditional SDMs), STEM-LM (focal) achieves AUROC 0.868 and CBI 0.860, on par with the strongest AUROC baseline GAM_{full} (AUROC 0.872, CBI 0.392) and substantially exceeding the strongest CBI baseline Maxnet (CBI 0.526). The partial-observation-aware deep-learning baseline CISO trails on both (AUROC 0.807, CBI 0.305). As the masking rate decreases, STEM-LM's advantage grows as it can exploit information regarding pre-observed species, outperforming CISO at all masking rates in both metrics. If maximizing discriminative ranking (AUROC) is the primary goal, STEM-LM trained with standard BCE loss achieves the highest AUROC across all settings, at the cost of lower rank-calibration (CBI).

Using the full model trained with STEM-LM (i.e., first row of Table 1, left), we inferred the predicted distribution of the monarch butterfly (*Danaus plexippus*) across North America in 2025 at four temporal snapshots (May, July, September, and November; Figure 2). We first prepared a 0.5° resolution grid with the same environmental variables used to train the full model, and ran inference at each site conditioned on the observations in the full eButterfly dataset (same scenario as $p = 1.0$ masking), followed by post-hoc temperature scaling. Temperature scaling was performed using the same training-validation-test split as in model training, and reduced the monarch's test-set ECE substantially from 0.0554 to 0.0063 (per-species ECE for this species; Table 3 reports the mean across all species).

In May, predicted scores are uniformly low across the continent with a faint peak in Texas, a known migratory corridor for monarchs. In July and September, the model recovers the well-known pattern of northward expansion, with peak scores concentrated over the Midwest, a well-known summer breeding ground. By November, predicted scores retreat from the northern range, leaving a faint peak in Arizona, another known migratory corridor. Notably, the model does not put strong signals to Texas or northern Mexico, which is the main migration corridor, during the spring and fall movement seasons; we attribute this to the absence or low density of observations in these regions in the training dataset and revisit the implications of this bias in Section 4.

Table 4: Test AUROC and CBI on eButterfly by masking rate p (mean $\pm$ std). STEM-LM (F) and (B) denote focal and BCE loss, respectively. **Bold** = best, underline = second-best per column.

	AUROC				CBI			
Model	$p = 0.25$	0.50	0.75	1.00	$p = 0.25$	0.50	0.75	1.00
$\text{Logistic}_{\text{full}}$	—	—	—	0.816	—	—	—	0.445
$\text{Logistic}_{\text{env}}$	—	—	—	0.784	—	—	—	0.431
$\text{Logistic}_{\text{st}}$	—	—	—	0.799	—	—	—	0.385
Maxnet	—	—	—	0.810	—	—	—	0.526
GAM_{full}	—	—	—	0.872	—	—	—	0.392
GAM_{env}	—	—	—	0.784	—	—	—	0.431
GAM_{st}	—	—	—	0.868	—	—	—	0.434
MaskSDM	—	—	—	0.821±0.001	—	—	—	0.523±0.024
CISO	0.839±0.009	0.837±0.010	0.825±0.009	0.807±0.011	0.510±0.069	0.532±0.061	0.473±0.033	0.305±0.015
STEM-LM (F)	0.891±0.006	0.891±0.002	0.884±0.001	0.868±0.002	**0.837**±0.002	**0.891**±0.002	**0.885**±0.006	**0.860**±0.013
STEM-LM (B)	**0.900**±0.003	**0.897**±0.002	**0.890**±0.003	**0.874**±0.003	0.628±0.024	0.691±0.012	0.721±0.005	0.655±0.023

Table 5: Test AUROC and CBI on sPlotOpen by masking rate p (mean $\pm$ std). STEM-LM (F) denotes focal-loss STEM-LM. **Bold** = best, underline = second-best per column.

	AUROC				CBI			
Model	$p = 0.25$	0.50	0.75	1.00	$p = 0.25$	0.50	0.75	1.00
$\text{Logistic}_{\text{full}}$	—	—	—	0.892	—	—	—	0.200
$\text{Logistic}_{\text{env}}$	—	—	—	0.901	—	—	—	0.065
$\text{Logistic}_{\text{s}}$	—	—	—	0.785	—	—	—	−0.419
Maxnet	—	—	—	0.939	—	—	—	0.592
GAM_{full}	—	—	—	0.930	—	—	—	0.328
GAM_{env}	—	—	—	0.901	—	—	—	0.064
GAM_{s}	—	—	—	0.908	—	—	—	0.468
MaskSDM	—	—	—	**0.947**±0.000	—	—	—	0.677±0.005
CISO	0.970±0.001	0.968±0.001	0.964±0.001	0.944±0.001	0.696±0.004	0.697±0.010	0.695±0.009	0.513±0.010
STEM-LM (F)	**0.978**±0.001	**0.975**±0.001	**0.970**±0.001	0.943±0.001	**0.864**±0.010	**0.899**±0.005	**0.909**±0.002	**0.843**±0.006

sPlotOpen Table 5 reports test AUROC and CBI on the sPlotOpen dataset across masking rates $p \in \{0.25, 0.50, 0.75, 1.00\}$. STEM-LM achieves the highest AUROC when partial species presence–absence information is given. The calibration advantage is even more pronounced than on eButterfly: at p=1.00, STEM-LM (F) attains CBI 0.843, compared to 0.677 for MaskSDM (the strongest baseline with respect to CBI) and below 0.5 for all logistic and GAM variants, some of which produce negatively correlated CBI scores ($\text{Logistic}_{\text{s}}$: -0.419). At lower masking rates, STEM-LM continues to improve on both AUROC and CBI, reaching AUROC 0.978 and CBI 0.864 at p=0.25. The pattern observed on eButterfly thus seems to generalize to a larger, spatial-only, geographically broad, and more taxonomically diverse dataset: STEM-LM provides both the strongest discriminative ranking and substantially better rank-calibrated occurrence probabilities than existing species distribution models.

From partial presence-only observations to complete assemblages Citizen-science records typically report only a handful of observed species per site without recorded absences; recovering the full assemblage from such partial input is a practical deployment of STEM-LM. We evaluate STEM-LM (default) under an absence-mask scheme that mimics this setting: all absences at the test site are masked, together with a fraction p of presences (Table 6). At p=1.00, this collapses to the corresponding row of Tables 4–5; at smaller p, the model conditions on the unmasked presences and recovers the rest. AUROC improves as more presences are observed on both datasets ($0.868 \rightarrow 0.886$ on eButterfly and $0.943 \rightarrow 0.976$ on sPlotOpen as p decreases from 1.00 to 0.25), confirming that even sparse presence-only context carries useful biotic signal. Because typically most species are absent at any site, p=0.75 already corresponds to only a minuscule fraction of the full species pool, yet substantially improves over the presence-free baseline.

4 Limitations and next steps

Like many SDMs, STEM-LM assumes high-quality presence–absence input and can be biased by imperfect detection, which degrades the performance of SDMs when unaccounted for [97–100]. Presence-only data, represented by crowd-sourced citizen-science data, is the extreme case of

Table 6: STEM-LM (default) under the presence-only evaluation scheme: a fraction p of presences masked and 100% absence masking (mean±std).

	AUROC				CBI			
Dataset	p=0.25	0.50	0.75	1.00	p=0.25	0.50	0.75	1.00
eButterfly	0.886±0.002	0.881±0.001	0.882±0.003	0.868±0.002	0.752±0.029	0.835±0.010	0.862±0.017	0.860±0.013
sPlotOpen	0.976±0.001	0.974±0.001	0.969±0.001	0.943±0.001	0.833±0.008	0.880±0.001	0.897±0.003	0.843±0.006

imperfect detection, and is naturally cast as a single-positive multi-label (SPML) learning problem [101–104], and Cole et al. [51] have shown the framing transfers to JSDM with presence-only data. However, SPML treats positive labels as drawn randomly, while presence-only observations are affected by observer effort, spatial sampling bias, and species-specific detectability [50, 105–107]. Even structured presence–absence surveys inevitably reflect observer effort and accessibility, leaving regions and seasons critical to a species' ecological dynamics underrepresented; the Texas–northern Mexico monarch migration corridor being largely absent from our eButterfly training data, and STEM-LM's correspondingly weak signal there during the spring and fall movements (Fig. 2), is a concrete example. A natural extension is to condition the masked-species loss on per-site, observer, and species detectability, learned from spatio-temporal context and observer-effort covariates where available, and to use this calibrated loss to fuse presence–absence with presence-only data during training, similar to observer-conditional geographical-prior loss explored in Mac Aodha et al. [50]. This would be a fundamentally different objective from applying a model trained on presence–absence survey data to partially observed, presence-only query points, as we do during our inference procedure, and would let abundant citizen-science records fill the spatio-temporal gaps that high-effort surveys might leave [108, 109].

One natural future direction is to project STEM-LM's inference into future climatic scenarios to estimate distributions decades ahead [110]. In such a scenario, no observations exist at the proximity of the target time, so the source-site cross-attention has no temporally proximate context, and the current implementation of temporal FIRE basis fits a single annual cycle shape from contemporary data that does not adapt at inference time to possible climate-driven phenological shifts. Both could be addressed by relaxing the temporal basis to allow cross-cycle shape variation, complemented by ecological process-based components that remain valid outside the training period [111, 112]. Projecting back in time shares a similar challenge, but a wider range of anchor data exists. Ancient environmental DNA (aeDNA), recovered from dated sediment samples, is a natural source of such data. Compared to macrofossils and skeletal remains, aeDNA captures a far broader range of taxa from a single sample and includes taxa that rarely enter the fossil record, making it a suitable proxy to study past ecological and environmental dynamics. aeDNA has been shown to capture species occurrences and co-occurrences across past warm intervals, glacial-interglacial transitions, and other states partially analogous to projected futures [113, 114]. A possible refinement is to take k-mer counts as input rather than reference-sequence-mapped taxon presence–absence calls, retaining information that is lost when reads cannot be confidently assigned to known taxa.

5 Conclusion

We proposed STEM-LM, a Transformer-based joint species distribution model that frames species presence–absence prediction as masked language modeling, combining species self-attention at the target site with cross-attention over neighboring sites in space and time and over environmental covariates. By tokenizing species presence–absence and assembling each site's assemblage with its spatio-temporal and environmental context as a "sentence," STEM-LM brings these signals into a single coherent framework for joint species distribution modeling. Empirically, STEM-LM matches or exceeds baselines on AUROC and substantially improves rank-calibration (CBI) on both datasets, with the CBI advantage particularly pronounced for rare species. Performance improves further when partial observations are available, relevant to the biodiversity forecasting and conservation applications motivating this work. Natural extensions include fusing presence-only citizen-science records through observer-conditional losses, projecting under future climate scenarios, and applying STEM-LM to ancient environmental DNA — directions that together span the spatio-temporal reach most relevant to biodiversity monitoring and conservation under global change.

Acknowledgments and Disclosure of Funding

This research is supported by funding from the Novo Nordisk Foundation (NNF24SA0092560).

References

[1] J. Elith and J. R. Leathwick. Species distribution models: ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics*, 40:677–697, 2009. doi: 10.1146/annurev.ecolsys.110308.120159.

[2] S. Beery, E. Cole, J. Parker, P. Perona, and K. Winner. Species distribution modeling for machine learning practitioners: A review. In *Proceedings of the 4th ACM SIGCAS Conference on Computing and Sustainable Societies*, pages 329–348, 2021. doi: 10.1145/3460112.3471966.

[3] A. Guisan and N. E. Zimmermann. Predictive habitat distribution models in ecology. *Ecological Modelling*, 135(2-3):147–186, 2000. doi: 10.1016/S0304-3800(00)00354-9.

[4] A. Guisan and W. Thuiller. Predicting species distribution: offering more than simple habitat models. *Ecology Letters*, 8(9):993–1009, 2005. doi: 10.1111/j.1461-0248.2005.00792.x.

[5] M. B. Araújo and M. Luoto. The importance of biotic interactions for modelling species distributions under climate change. *Global Ecology and Biogeography*, 16(6):743–753, 2007. doi: 10.1111/j.1466-8238.2007.00359.x.

[6] E. S. Meier, F. Kienast, P. B. Pearman, J. C. Svenning, W. Thuiller, M. B. Araújo, A. Guisan, and N. E. Zimmermann. Biotic and abiotic variables show little redundancy in explaining tree species distributions. *Ecography*, 33(6):1038–1048, 2010. doi: 10.1111/j.1600-0587.2010.06229.x.

[7] M. S. Wisz, J. Pottier, W. D. Kissling, L. Pellissier, J. Lenoir, C. F. Damgaard, C. F. Dormann, M. C. Forchhammer, J. A. Grytnes, A. Guisan, R. K. Heikkinen, T. T. Høye, I. Kühn, M. Luoto, L. Maiorano, M. C. Nilsson, S. Normand, E. Öckinger, N. M. Schmidt, M. Termansen, et al. The role of biotic interactions in shaping distributions and realised assemblages of species: implications for species distribution modelling. *Biological Reviews*, 88:15–30, 2013. doi: 10.1111/j.1469-185X.2012.00235.x.

[8] D. I. Warton, F. G. Blanchet, R. B. O'Hara, O. Ovaskainen, S. Taskinen, S. C. Walker, and F. K. C. Hui. So many variables: joint modeling in community ecology. *Trends in Ecology Evolution*, 30(12):766–779, 2015. doi: 10.1016/j.tree.2015.09.007.

[9] M. Pichler and F. Hartig. A new joint species distribution model for faster and more accurate inference of species associations from big community data. *Methods in Ecology and Evolution*, 12(11):2159–2173, 2021. doi: 10.1111/2041-210X.13687.

[10] A. U. Rahman, G. Tikhonov, J. Oksanen, T. Rossi, and O. Ovaskainen. Accelerating joint species distribution modelling with Hmsc-HPC by GPU porting. *PLoS Computational Biology*, 20(9):e1011914, 2024. doi: 10.1371/journal.pcbi.1011914.

[11] F. M. Callahan, J. K. Li, and R. Nielsen. Challenges in detecting ecological interactions using sedimentary ancient DNA data. *Environmental DNA*, 7:e70067, 2025. doi: 10.1002/edn3.70067.

[12] D. Fink, W. M. Hochachka, B. Zuckerberg, D. W. Winkler, B. Shaby, M. A. Munson, G. Hooker, M. Riedewald, D. Sheldon, and S. Kelling. Spatiotemporal exploratory models for broad-scale survey data. *Ecological Applications*, 20(8):2131–2147, 2010. doi: 10.1890/09-1340.1.

[13] D. Fink, T. Damoulas, and J. Dave. Adaptive Spatio-Temporal Exploratory Models: Hemisphere-wide species distributions from massively crowdsourced eBird data. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*, 2013. doi: 10.1609/aaai.v27i1.8484.

[14] J. T. Thorson, M. D. Scheuerell, A. O. Shelton, K. E. See, H. J. Skaug, and K. Kristensen. Spatial factor analysis: a new tool for estimating joint species distributions and correlations in species range. *Methods in Ecology and Evolution*, 6(6):627–637, 2015. doi: 10.1111/2041-210X.12359.

[15] J. T. Thorson, J. N. Ianelli, E. A. Larsen, L. Ries, M. D. Scheuerell, C. Szuwalski, and E. F. Zipkin. Joint dynamic species distribution models: a tool for community ordination and spatio-temporal monitoring. *Global Ecology and Biogeography*, 25(9):1144–1158, 2016. doi: 10.1111/geb.12464.

[16] O. Ovaskainen, G. Tikhonov, A. Norberg, F. G. Blanchet, L. Duan, D. Dunson, T. Roslin, and N. Abrego. How to make more out of community data? A conceptual framework and its implementation as models and software. *Ecology Letters*, 20(5):561–576, 2017. doi: 10.1111/ele.12757.

[17] S. Domisch, M. Friedrichs, T. Hein, F. Borgwardt, A. Wetzig, S. C. Jähnig, and S. D. Langhans. Spatially explicit species distribution models: A missed opportunity in conservation planning? *Diversity and Distributions*, 25(5):758–769, 2019. doi: 10.1111/ddi.12891.

[18] F. K. C. Hui, D. I. Warton, S. D. Foster, and C. R. Haak. Spatiotemporal joint species distribution modelling: A basis function approach. *Methods in Ecology and Evolution*, 14(8):2150–2164, 2023. doi: 10.1111/2041-210X.14184.

[19] C. Botella, A. Joly, P. Bonnet, P. Monestiez, and F. Munoz. A deep learning approach to species distribution modelling. In A. Joly, S. Vrochidis, K. Karatzas, A. Karppinen, and P. Bonnet, editors, *Multimedia Tools and Applications for Environmental & Biodiversity Informatics*, chapter 10, pages 169–199. Springer, Cham, 2018. doi: 10.1007/978-3-319-76445-0_10.

[20] D. Tuia, B. Kellenberger, S. Beery, B. R. Costelloe, S. Zuffi, B. Risse, A. Mathis, M. W. Mathis, F. van Langevelde, T. Burghardt, R. Kays, H. Klinck, M. Wikelski, I. D. Couzin, G. van Horn, M. C. Crofoot, C. V. Stewart, and T. Berger-Wolf. Perspectives in machine learning for wildlife conservation. *Nature Communications*, 13:792, 2022. doi: 10.1038/s41467-022-27980-y.

[21] R. Zbinden, N. van Tiel, B. Kellenberger, L. Hughes, and D. Tuia. Exploring the potential of neural networks for species distribution modeling. In *The Eleventh International Conference on Learning Representations*, 2023.

[22] L. J. Pollock, J. Kitzes, S. Beery, K. M. Gaynor, M. A. Jarzyna, O. Mac Aodha, B. Meyer, D. Rolnick, G. W. Taylor, D. Tuia, and T. Berger-Wolf. Harnessing artificial intelligence to fill global shortfalls in biodiversity knowledge. *Nature Reviews Biodiversity*, 1(3):166–182, 2025. doi: 10.1038/s44358-025-00022-3.

[23] S. J. Phillips, M. Dudík, and R. E. Schapire. A maximum entropy approach to species distribution modeling. In *Proceedings of the Twenty-First International Conference on Machine Learning*, 2004. doi: 10.1145/1015330.1015412.

[24] S. J. Phillips, R. P. Anderson, and R. E. Schapire. Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, 190(3-4):231–259, 2006. doi: 10.1016/j.ecolmodel.2005.03.026.

[25] A. M. Prasad, L. R. Iverson, and A. Liaw. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. *Ecosystems*, 9(2):181–199, 2006. doi: 10.1007/s10021-005-0054-1.

[26] D. R. Cutler, T. C. Edwards, K. H. Beard, A. Cutler, K. T. Hess, J. Gibson, and J. J. Lawler. Random forests for classification in ecology. *Ecology*, 88(11):2783–2792, 2007. doi: 10.1890/07-0539.1.

[27] G. De'ath. Boosted trees for ecological modeling and prediction. *Ecology*, 88(1):243–251, 2007. doi: 10.1890/0012-9658(2007)88[243:BTFEMA]2.0.CO;2.

[28] S. Lek, M. Delacoste, P. Baran, I. Dimopoulos, J. Lauga, and S. Aulagnier. Application of neural networks to modelling nonlinear relationships in ecology. *Ecological Modelling*, 90(1):39–52, 1996. doi: 10.1016/0304-3800(95)00142-5.

[29] S. L. Özesmi and U. Özesmi. An artificial neural network approach to spatial habitat modelling with interspecific interaction. *Ecological Modelling*, 116(1):15–31, 1999. doi: 10.1016/S0304-3800(98)00149-5.

[30] R. Valavi, J. Elith, J. J. Lahoz-Monfort, and G. Guillera-Arroita. Flexible species distribution modelling methods perform well on spatially separated testing data. *Global Ecology and Biogeography*, 32(3):369–383, 2023. doi: https://doi.org/10.1111/geb.13639.

[31] B. Deneu, M. Servajean, P. Bonnet, C. Botella, F. Munoz, and A. Joly. Convolutional neural networks improve species distribution modelling by capturing the spatial structure of the environment. *PLoS Computational Biology*, 17(4):e1008856, 2021. doi: 10.1371/journal.pcbi.1008856.

[32] Y. Hu, S. Si-Moussi, and W. Thuiller. Introduction to deep learning methods for multi-species predictions. *Methods in Ecology and Evolution*, 16(1):228–246, 2025. doi: 10.1111/2041-210X.14466.

[33] R. Valavi, G. Guillera-Arroita, J. J. Lahoz-Monfort, and J. Elith. Predictive performance of presence-only species distribution models: a benchmark study with reproducible code. *Ecological Monographs*, 92(1):e01486, 2022. doi: https://doi.org/10.1002/ecm.1486.

[34] L. Jeantet and E. Dufourq. Improving deep learning acoustic classifiers with contextual information for wildlife monitoring. *Ecological Informatics*, 77:102256, 2023. doi: 10.1016/j.ecoinf.2023.102256.

[35] M. Teng, A. Elmustafa, B. Akera, H. Larochelle, and D. Rolnick. Bird distribution modelling using remote sensing and citizen science data. In *The Eleventh International Conference on Learning Representations*, 2023.

[36] M. Teng, A. Elmustafa, B. Akera, Y. Bengio, H. R. Abdelwahed, H. Larochelle, and D. Rolnick. SatBird: Bird species distribution modeling with remote sensing and citizen science data. In *Advances in Neural Information Processing Systems 36*, 2023.

[37] T. Chen and Y. Y. Chiang. MiTREE: Multi-input transformer ecoregion encoder for species distribution modelling. In *Proceedings of the 7th ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, pages 110–120, 2024. doi: 10.1145/3687123.3698297.

[38] M. Hamilton, C. Lange, E. Cole, A. Shepard, S. Heinrich, O. Mac Aodha, G. Van Horn, and S. Maji. Combining observational data and language for species range estimation. In *Advances in Neural Information Processing Systems 37*, 2024.

[39] T. Larcher, L. Picek, B. Deneu, T. Lorieu, M. Servajean, and A. Joly. MALPOLON: A framework for deep species distribution modeling. *arXiv*, 2024. doi: 10.48550/arXiv.2409.18102.

[40] Y. Guo, K. Morkany, S. R. Levick, and J. Yang. Spatioformer: a geo-encoded transformer for large-scale plant species richness prediction. *IEEE Transactions on Geoscience and Remote Sensing*, 63:4403216, 2025. doi: 10.1109/TGRS.2025.3534654.

[41] C. Lange, M. Hamilton, E. Cole, A. Shepard, S. Heinrich, A. Zhu, S. Maji, G. Van Horn, and O. Mac Aodha. Feedforward few-shot species range estimation. In *Proceedings of the 42th International Conference on Machine Learning*, 2025.

[42] N. van Tiel, T. Zbinden, E. Dalsasso, B. Kellenberger, L. Pellissier, and D. Tuia. Multi-scale and multimodal species distribution modeling. In *2024 European Conference on Computer Vision*, 2025. doi: 10.1007/978-3-031-92387-6_10.

[43] L. Tang, Y. Xue, D. Chen, and C. P. Gomes. Multi-entity dependence learning with rich context via conditional variational auto-encoder. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. doi: 10.1609/aaai.v32i1.11335.

[44] E. Seo, R. A. Hutchinson, X. Fu, C. Li, T. A. Hallman, J. Kilbride, and W. D. Robinson. StatEcoNet: Statistical ecology neural networks for species distribution modeling. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence*, 2021. doi: 10.1609/aaai.v35i1.16129.

[45] C. L. Davis, Y. Bai, D. Chen, O. Robinson, V. Ruiz-Gutierrez, C. P. Gomes, and D. Fink. Deep learning with citizen science data enables estimation of species diversity and composition at continental extents. *Ecology*, 104(12):e4175, 2023. doi: 10.1002/ecy.4175.

[46] P. Brun, D. N. Karger, D. Zurell, P. Descombes, L. C. de Witte, R. de Lutio, J. D. Wegner, and N. E. Zimmermann. Multispecies deep learning using citizen science data produces more informative plant community models. *Nature Communications*, 15:4221, 2024. doi: 10.1038/s41467-024-48559-9.

[47] R. Zbinden, N. van Tiel, G. Sumbul, B. Kellenberger, and D. Tuia. MaskSDM: Adaptive species distribution modeling through data masking. In *2024 European Conference on Computer Vision*, 2025. doi: 10.1007/978-3-031-92387-6_14.

[48] R. Zbinden, N. van Tiel, G. Sumbul, C. Vanalli, B. Kellenberger, and D. Tuia. MaskSDM with Shapley values to improve flexibility, robustness and explainability in species distribution modelling. *Methods in Ecology and Evolution*, 17(1):188–206, 2025. doi: 10.1111/2041-210x.70200.

[49] H. R. Abdelwahed, M. Teng, R. Zbinden, L. Pollock, H. Larochelle, D. Tuia, and D. Rolnick. CISO: Species distribution modelling conditioned on incomplete species observations. *Methods in Ecology and Evolution*, 17(3):947–962, 2026. doi: 10.1111/2041-210x.70238.

[50] O. Mac Aodha, E. Cole, and P. Perona. Presence-only geographical priors for fine-grained image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9596–9606, 2019. doi: 10.1109/ICCV.2019.00969.

[51] E. Cole, G. Van Horn, C. Lange, A. Shepard, P. Leary, P. Perona, S. Loarie, and O. Mac Aodha. Spatial implicit neural representation for global-scale species mapping. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

[52] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, 2017. doi: 10.5555/3295222.3295349.

[53] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4171–4186, 2019. doi: 10.18653/v1/N19-1423.

[54] R. Rao, N. Bhattacharya, N. Thomas, Y. Duan, X. Chen, J. Canny, P. Abbeel, and Y. S. Song. MSA transformer. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

[55] G. Benegas, C. Albors, A. J. Aw, C. Ye, and Y. S. Song. A DNA language model based on multispecies alignment predicts the effects of genome-wide variants. *Nature Biotechnology*, 43(12):1960–1965, 2025. doi: 10.1038/s41587-024-02511-w.

[56] C. Ye, G. Benegas, C. Albors, J. C. Li, S. Prillo, P. D. Fields, B. Clarke, and Y. S. Song. Predicting functional constraints across evolutionary timescales with phylogeny-informed genomic language models. *bioRxiv*, 2025. doi: 10.1101/2025.09.21.677619.

[57] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16×16 words: Transformers for image recognition at scale. In *The Ninth International Conference on Learning Representations*, 2021.

[58] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3D convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4489–4497, 2015. doi: 10.1109/ICCV.2015.510.

[59] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid. ViViT: A video vision transformer. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 6836–6846, 2021. doi: 10.1109/ICCV48922.2021.00676.

[60] J. Lee, Y. Lee, J. Kim, A. R. Kosiorek, S. Choi, and Y. W. Teh. Set Transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

[61] A. Jaegle, F. Gimeno, A. Brock, A. Zisserman, O. Vinyals, and J. Carreira. Perceiver: General perception with iterative attention. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

[62] J. L. Ba, J. R. Kiros, and G. E. Hinton. Layer normalization. *arXiv*, 2016. doi: 10.48550/arXiv.1607.06450.

[63] S. Elfwing, E. Uchibe, and K. Doya. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Networks*, 107:3–11, 2017. doi: 10.1016/j.neunet.2017.12.012.

[64] B. Zhang and R. Sennrich. Root mean square layer normalization. In *Advances in Neural Information Processing Systems 32*, 2019. doi: 10.5555/3454287.3455397.

[65] R. Xiong, Y. Yang, D. He, K. Zheng, S. Zheng, C. Xing, H. Zhang, Y. Lan, L. Wang, and T. Y. Liu. On layer normalization in the transformer architecture. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

[66] S. Li, C. You, G. Guruganesh, J. Ainslie, S. Ontanon, M. Zaheer, S. Sanghai, Y. Yang, S. Kumar, and S. Bhojanapalli. Functional interpolation for relative positions improves long context transformers. In *The Twelfth International Conference on Learning Representations*, 2024.

[67] N. Shazeer. GLU variants improve transformer. *arXiv*, 2020. doi: 10.48550/arXiv.2002.05202.

[68] D. R. Roberts, V. Bahn, S. Ciuti, M. S. Boyce, J. Elith, G. Guillera-Arroita, S. Hauenstein, J. J. Lahoz-Monfort, B. Schröder, W. Thuiller, D. I. Warton, B. A. Wintle, F. Hartig, and C. F. Dormann. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8):913–929, 2017. doi: 10.1111/ecog.02881.

[69] Uber Technologies, Inc. H3: A hexagonal hierarchical geospatial indexing system. `https://h3geo.org`, 2018.

[70] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017. doi: 10.1109/ICCV.2017.324.

[71] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):318–327, 2020. doi: 10.1109/TPAMI.2018.2858826.

[72] D. P. Kingma and J. L. Ba. Adam: A method for stochastic optimization. In *The Third International Conference on Learning Representations*, 2015.

[73] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *The Seventh International Conference on Learning Representations*, 2019.

[74] I. Loshchilov and F. Hutter. SGDR: Stochastic gradient descent with warm restarts. In *The Fifth International Conference on Learning Representations*, 2017.

[75] R. Pascanu, T. Mikolov, and Y. Bengio. On the difficulty of training recurrent neural networks. In *Proceedings of the 30th International Conference on Machine Learning*, 2013.

[76] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[77] M. B. McDermott, H. Zhang, L. H. Hansen, G. Angelotti, and J. Gallifant. A closer look at AUROC and AUPRC under class imbalance. In *Advances in Neural Information Processing Systems 37*, 2024.

[78] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, 2019.

[79] M. S. Boyce, P. R. Vernier, S. E. Nielsen, and F. K. A. Schmiegelow. Evaluating resource selection functions. *Ecological Modelling*, 157(2-3):281–300, 2002. doi: 10.1016/S0304-3800(02)00200-4.

[80] A. H. Hirzel, G. Le Lay, V. Helfer, C. Randin, and A. Guisan. Evaluating the ability of habitat suitability models to predict species presences. *Ecological Modelling*, 199(2):142–152, 2006. doi: 10.1016/j.ecolmodel.2006.05.017.

[81] M. P. Naeini, G. F. Cooper, and M. Hauskrecht. Obtaining well calibrated probabilities using Bayesian binning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015. doi: 10.1609/aaai.v29i1.9602.

[82] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

[83] K. L. Prudic, K. P. McFarland, J. C. Oliver, R. A. Hutchinson, E. C. Long, J. T. Kerr, and M. Larrivée. eButterfly: Leveraging massive online citizen science for butterfly conservation. *Insects*, 8(2):53, 2017. doi: 10.3390/insects8020053.

[84] F. M. Sabatini, J. Lenoir, T. Hattab, E. A. Arnst, M. Chytrý, J. Dengler, P. De Ruffray, S. M. Hennekens, U. Jandt, F. Jansen, B. Jiménez-Alfaro, J. Kattge, A. Levesley, V. D. Pillar, O. Purschke, B. Sandel, F. Sultana, T. Aavik, S. Aćić, A. T. R. Acosta, et al. sPlotOpen – an environmentally balanced, open-access, global dataset of vegetation plots. *Global Ecology and Biogeography*, 30(9):1740–1764, 2021. doi: 10.1111/geb.13346.

[85] J. Muñoz-Sabater, E. Dutra, A. Agustí-Panareda, C. Albergel, G. Arduini, G. Balsamo, S. Boussetta, M. Choulga, S. Harrigan, H. Hersbach, B. Martens, D. G. Miralles, M. Piles, N. J. Rodríguez-Fernández, E. Zsoter, C. Buontempo, and J. N. Thépaut. ERA5-Land: a state-of-the-art global reanalysis dataset for land applications. *Earth System Science Data*, 13 (9):4349–4383, 2021. doi: 10.5194/essd-13-4349-2021.

[86] K. Didan. MOD13Q1 MODIS/Terra Vegetation Indices 16-Day L3 Global 250 m SIN Grid V061. `https://doi.org/10.5067/MODIS/MOD13Q1.061`, 2021.

[87] S. E. Fick and R. J. Hijmans. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, 37(12):4302–4315, 2017. doi: 10.1002/joc.5086.

[88] L. Poggio, L. M. de Sousa, N. H. Batjes, G. B. M. Heuvelink, B. Kempen, E. Ribeiro, and D. Rossiter. SoilGrids 2.0: producing soil information for the globe with quantified spatial uncertainty. *SOIL*, 7(1):217–240, 2021. doi: 10.5194/soil-7-217-2021.

[89] European Space Agency. Copernicus DEM – Global and European Digital Elevation Model. `https://doi.org/10.5270/ESA-c5d3d65`, 2021. Copernicus Digital Elevation Model, GLO-30.

[90] European Space Agency and Sinergise. Copernicus Global Digital Elevation Model. `https://doi.org/10.5069/G9028PQB`, 2021. Distributed by OpenTopography.

[91] J. Pearce and S. Ferrier. An evaluation of alternative algorithms for fitting species distribution models using logistic regression. *Ecological Modelling*, 128(2-3):127–147, 2000. doi: 10.1016/S0304-3800(99)00227-6.

[92] S. Phillips. *maxnet: Fitting 'Maxent' Species Distribution Models with 'glmnet'*, 2021. URL `https://CRAN.R-project.org/package=maxnet`. R package version 0.1.4.

[93] A. Guisan, T. C. Edwards, and T. Hastie. Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecological Modelling*, 157(2-3):89–100, 2002. doi: 10.1016/S0304-3800(02)00204-1.

[94] J. M. McPherson, W. Jetz, and D. J. Rogers. The effects of species' range sizes on the accuracy of distribution models: ecological phenomenon or statistical artefact? *Journal of Applied Ecology*, 41(5):811–823, 2004. doi: 10.1111/j.0021-8901.2004.00943.x.

[95] J. M. Lobo, A. Jiménez-Valverde, and R. Real. AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, 17(2):145–151, 2008. doi: 10.1111/j.1466-8238.2007.00358.x.

[96] J. Mukhoti, V. Kulharia, A. Sanyal, S. Golodetz, P. H. S. Torr, and P. K. Dokania. Calibrating deep neural networks using focal loss. In *Advances in Neural Information Processing Systems 33*, 2020.

[97] W. Gu and R. K. Swihart. Absent or undetected? effects of non-detection of species occurrence on wildlife–habitat models. *Biological Conservations*, 116(2):195–203, 2004. doi: 10.1016/S0006-3207(03)00190-3.

[98] J. M. Lobo, A. Jiménez-Valverde, and J. Hortal. The uncertain nature of absences and their importance in species distribution modelling. *Ecography*, 33(1):103–114, 2010. doi: 10.1111/j.1600-0587.2009.06039.x.

[99] J. Lahoz-Monfort, G. Guillera-Arroita, and B. A. Wintle. Imperfect detection impacts the performance of species distribution models. *Global Ecology and Biogeography*, 23(4):504–515, 2014. doi: 10.1111/geb.12138.

[100] A. Miller-ter-Kuile, A. Bui, A. Apigo, S. Lamm, M. Swan, J. S. Sanderlin, and K. Ogle. If you're rare, should I care? How imperfect detection changes relationships between biodiversity and global change drivers. *Global Change Biology*, 31(7):e70362, 2025. doi: 10.1111/gcb.70362.

[101] E. Cole, O. Mac Aodha, T. Lorieul, P. Perona, D. Morris, and N. Jojic. Multi-label learning from single positive labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 933–942, 2021. doi: 10.1109/CVPR46437.2021.00099.

[102] M. L. Xie, J. Xiao, and S. J. Huang. Label-aware global consistency for multi-label learning with single positive labels. In *Advances in Neural Information Processing Systems 35*, 2022.

[103] N. Xu, C. Qiao, J. Lv, X. Geng, and M. L. Zhang. One positive label is sufficient: Single-positive multi-label learning with label enhancement. In *Advances in Neural Information Processing Systems 35*, 2022.

[104] B. Liu, N. Xu, J. Lv, and X. Geng. Revisiting pseudo-label for single-positive multi-label learning. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

[105] E. H. Boakes, P. J. K. McGowan, R. A. Fuller, C. Q. Ding, N. E. Clark, K. O'Connor, and G. M. Mace. Distorted views of biodiversity: Spatial and temporal bias in species occurrence data. *PLoS Biology*, 8(6):e1000385, 2010. doi: 10.1371/journal.pbio.1000385.

[106] A. Johnston, N. Moran, A. Musgrove, Fink. D., and S. R. Baillie. Estimating species distributions from spatially biased citizen science data. *Ecological Modelling*, 422(15):108927, 2020. doi: 10.1016/j.ecolmodel.2019.108927.

[107] J. Arroyo, P. Perona, and E. Cole. Understanding label bias in single positive multi-label learning. In *The Eleventh International Conference on Learning Representations*, 2023.

[108] O. J. Robinson, V. Ruiz-Gutierrez, M. D. Reynolds, G. H. Golet, M. Strimas-Mackey, and D. Fink. Integrating citizen science data with expert surveys increases accuracy and spatial extent of species distribution models. *Diversity and Distributions*, 26(8):976–986, 2020. doi: 10.1111/ddi.13068.

[109] D. Romera-Romera and D. Nieto-Lugilde. Should we exploit opportunistic databases with joint species distribution models? Artificial and real data suggest it depends on the sampling completeness. *Ecography*, 2025(2):e07340, 2025. doi: 10.1111/ecog.07340.

[110] C. F. Dormann. Promising the future? Global change projections of species distributions. *Basic and Applied Ecology*, 8(3):387–397, 2007. doi: 10.1016/j.baae.2006.11.001.

[111] K. L. Yates, P. J. Bouchet, M. J. Caley, K. Mengersen, C. F. Randin, S. Parnell, A. H. Fielding, A. J. Bamford, S. Ban, A. M. Barbosa, C. F. Dormann, J. Elith, C. B. Embling, G. N. Ervin, R. Fisher, S. Gould, R. F. Graf, E. J. Gregr, P. N. Halpin, R. K. Heikkinen, et al. Outstanding challenges in the transferability of ecological models. *Trends in Ecology Evolution*, 33(10): 790–802, 2018. doi: 10.1016/j.tree.2018.08.001.

[112] N. J. Briscoe, J. Elith, R. Salguero-Gómez, J. J. Lahoz-Monfort, J. S. Camac, K. M. Giljohann, M. H. Holden, B. A. Hradsky, M. R. Kearney, S. M. McMahon, B. L. Phillips, T. J. Regan, J. R. Rhodes, P. A. Vesk, B. A. Wintle, J. D. L. Yen, and G. Guillera-Arroita. Forecasting species range dynamics with process-explicit models: matching methods to applications. *Ecology Letters*, 22(11):1940–1956, 2019. doi: 10.1111/ele.13348.

[113] L. Parducci, K. D. Bennett, G. F. Ficetola, I. G. Alsos, Y. Suyama, J. R. Wood, and M. W. Pedersen. Ancient plant DNA in lake sediments. *New Phytologist*, 214(3):924–942, 2017. doi: 10.1111/nph.14470.

[114] I. G. Alsos, V. Boussange, D. P. Rijal, M. Beaulieu, A. G. Brown, U. Herzschuh, J. C. Svenning, and L. Pellissier. Using ancient sedimentary DNA to forecast ecosystem trajectories under climate change. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 379 (1902):20230017, 2024. doi: 10.1098/rstb.2023.0017.

[115] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2024. URL `https://www.R-project.org/`.

[116] S. N. Wood. *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC, Boca Raton, 2 edition, 2017. ISBN 9781498728331. doi: 10.1201/9781315370279.

A Dataset preparation

eButterfly We retained only structured survey protocols that permit valid absence inference: Traveling Survey, Area Survey, Timed Count, Point Count, Atlas Square, Pollard Walk, and Pollard Transect, and excluded Incidental Observation (presence-only) and Historical (no reliable date) records. The geographic scope is continental North America extended through Mesoamerica; observations from 2011 to 2025 were retained to maintain consistent observational density. Each sampling event constitutes one row, with $S = 173$ species (those with ≥ 100 presences in the retained subset) encoded as binary presence–absence, yielding $M = 17{,}077$ final checklists. Train/validation/test splits at nominal $80/10/10$ ratios were assigned via H3 spatial blocking at resolution 2 ($\sim$158 km edge, seed 42), giving $13{,}825/950/2{,}302$ checklists, respectively.

For climate, we used ERA5-Land [85] daily 2 m air temperature (min/mean/max) and total precipitation for the observation date, lapse-rate corrected at $-6.5\,^{\circ}\mathrm{C\,km^{-1}}$ using the Copernicus GLO-30 DEM [89, 90] against the model orography of the corresponding ERA5 grid cell: $T_{\mathrm{site}} = T_{\mathrm{ERA5}} + 0.0065\,(\mathrm{orog}_{\mathrm{ERA5}} - \mathrm{DEM}_{\mathrm{site}})$. For vegetation dynamics, we used MOD13Q1 [86] NDVI and EVI (16-day composites, 250 m) from the composite nearest the observation date. Static covariates are Copernicus GLO-30 DEM elevation and eight SoilGrids [88] v. 2.0 properties at 0–5 cm depth (Table 7). All raster sources are sampled at the observation point and standardized (z-scored) per dimension on the training split.

sPlotOpen sPlotOpen [84] (v. 2.0) is a global plant species occurrence dataset optimized for macroecological analysis. We retained all records with valid coordinates and did not apply the climate-balanced subsample filter (which downsamples the database to a climate-stratified subset) to maximize geographic and floristic coverage. Plant distributions were treated as approximately time-invariant; this is more defensible for broad-scale plant patterns than for animal occurrence data, where short-term movement and seasonal dynamics dominate. Each vegetation plot constitutes one row, with $S = 1{,}201$ species (≥ 300 presences in the retained subset) encoded as binary presence–absence, yielding $M = 95{,}104$ final plots spanning all inhabited continents. The same H3 resolution-2 spatial blocking with $80/10/10$ split ratios (seed 42) was used, giving $76{,}699/8{,}636/9{,}769$ plots, respectively.

For environmental covariates, we used the 19 standard WorldClim [87] v. 2.1 bioclimatic variables, together with the same Copernicus GLO-30 DEM elevation and eight SoilGrids v. 2.0 properties used in eButterfly (Table 7). Covariates are standardized (z-scored) per dimension on the training split.

Table 7: Summary of environmental covariates used.

Source	Variables	Resolution	Dataset
ERA5-Land [85]	2 m air temperature (`t2m`; min/mean/max), total precipitation (`tp`)	0.1° ($\sim$9 km), daily	eButterfly
MOD13Q1 [86]	NDVI (Normalized Difference Vegetation Index), EVI (Enhanced Vegetation Index)	250 m, 16-day composite	eButterfly
WorldClim 2.1 [87]	19 bioclimatic variables: 11 temperature-derived (`bio1`–`bio11`; annual / seasonal / extreme means) and 8 precipitation-derived (`bio12`–`bio19`; annual sum, seasonality, quarterly extremes)	30 arc-second ($\sim$1 km)	sPlotOpen
SoilGrids 2.0 [88]	`bdod` (bulk density), `cec` (cation exchange capacity), `cfvo` (coarse-fragment volume), `clay` / `sand` / `silt` (texture fractions), `nitrogen` (total N), `phh2o` (pH in H_2O); all at 0–5 cm depth	250 m	both
Copernicus GLO-30 DEM [89, 90]	Elevation (`dem`)	30 m	both

B Benchmark model specifications

Logistic regression A generalized linear model (GLM) with a logistic link [91] was used as a basic single-species benchmark with either (a) only environmental covariates ($\mathrm{Logistic}_{\mathrm{env}}$), (b) latitude, longitude, an interaction between latitude and longitude, and (if applicable) a Fourier basis on day of year ($\mathrm{Logistic}_{\mathrm{st}}$), or (c) all of the above ($\mathrm{Logistic}_{\mathrm{full}}$). The Fourier basis comprised sine and cosine encodings at periods of 365, 182, 122, and 91 days, matching the temporal Fourier basis used

by STEM-LM. Models were fit in R [115] via `stats::glm` with `family = binomial(link = "logit")` and a maximum of 200 IRLS iterations; no regularization was applied. Models were run separately for each species, and benchmark metrics were averaged across species.

Maxnet Maxnet [92] (an R reimplementation of the popular niche modelling software Maxent [23, 24] as a regularized GLM) was used to fit flexible models with environmental covariates only. The regularization multiplier (`regmult`) was tuned over the grid $\{1, 2, 4, 6, 8, 10, 12, 16, 20, 24, 32\}$ by maximizing the mean validation AUROC across a random subset of 20 species (seed 42); the selected value was applied uniformly to all species. Feature transformations (linear, quadratic, hinge, product) are auto-selected by Maxnet based on each species' presence count, following the package defaults. Models were run separately for each species, and benchmark metrics were averaged across species.

GAM A generalized additive model (GAM) [93] was fit with `mgcv::bam` [116] (v. 1.9.1, `discrete = TRUE` for fast big-data fitting, REML smoothness selection) as a stronger non-linear single-species baseline with either (a) only environmental covariates entered linearly (GAM_{env}),[5] (b) latitude and longitude entered through a two-dimensional thin-plate smooth $s(\text{lat}, \text{lon}, k{=}50)$ together with (if applicable) day of year through a cyclic cubic regression smooth $s(\text{doy}, bs{=}\text{cc}, k{=}12)$ (GAM_{st}), or (c) all of the above combined: linear environmental covariates plus the spatial and temporal smooths (GAM_{full}). All variants use the binomial logit link. Models were run separately for each species, and benchmark metrics were averaged across species.

MaskSDM MaskSDM [47, 48] is a deep-learning SDM that tokenizes each environmental covariate and applies a Transformer encoder over the resulting feature tokens, with a learned mask token replacing a random subset of features during training so that the model learns to predict from any subset of available predictors. Per-species presence probabilities are output by a single shared linear head on the pooled token representation. We trained MaskSDM on the same train split with the upstream-default FT-Transformer configuration ($d_{\text{hidden}}{=}192$, 7 blocks, 8 attention heads, dropout 0.1) for 1,000 epochs at batch size 256 with the upstream default optimizer and learning rate, and a weighted-BCE loss using species inverse-frequency weights. Inference was performed with all environmental features available (analogous to STEM-LM's $p{=}1.0$ evaluation). The best checkpoint was selected by validation AUROC, and benchmark metrics were averaged across species over three random seeds (1337, 1338, 1339).

CISO CISO [49] is a deep-learning SDM that conditions species predictions on partial observations of other species at the same site: each species token carries a learned state embedding (presence, absence, or unknown), and a Transformer over species tokens is fed the environmental covariates through a SimpleMLP backbone. We trained CISO on the same train split with the upstream-default architecture ($d_{\text{hidden}}{=}256$, partial-label quantization disabled so labels remain binary) for 100 epochs (eButterfly) or 50 epochs (sPlotOpen) at batch size 64 and learning rate 10^{-3}, then benchmarked it at evaluation masking levels $p \in \{0.25, 0.5, 0.75, 1.0\}$ to mirror STEM-LM's random-masking evaluation, where p is the fraction of species masked. The best checkpoint was selected by validation AUROC, and benchmark metrics were averaged across species at each p over three random seeds (1337, 1338, 1339).

[5] Since the env-only variant has no continuous coordinate to smooth over, it is fit with `stats::glm` rather than `bam` and is structurally identical to $\text{Logistic}_{\text{env}}$.

C Compute usage

Table 8: GPU compute usage

Run	Seed	GPU	Platform	Time (h:m)
eButterfly BCE loss				
ebut_bce	41	1× NVIDIA A40 (46 GB)	UC Berkeley Savio	1:48
ebut_bce	42,43	1× NVIDIA A5000 (24 GB) each	UC Berkeley Savio	∼1:44 each
eButterfly ablation				
full	42	1× NVIDIA A5000 (24 GB)	UC Berkeley Savio	1:41
no_st	42	1× NVIDIA A5000 (24 GB)	UC Berkeley Savio	1:03
no_env	42	1× NVIDIA A40 (46 GB)	UC Berkeley Savio	1:53
no_st_env	42	1× NVIDIA A40 (46 GB)	UC Berkeley Savio	1:08
full	41,43	1× NVIDIA A100 (40 GB) each	Google Colab	∼1:48 each
no_st	41,43	1× NVIDIA A100 (40 GB) each	Google Colab	∼1:23 each
no_env	41,43	1× NVIDIA A100 (40 GB) each	Google Colab	∼1:46 each
no_st_env	41,43	1× NVIDIA A100 (40 GB) each	Google Colab	∼1:20 each
32 sites	41,42,43	1× NVIDIA A100 (40 GB) each	Google Colab	∼1:41 each
128 sites	41,42,43	1× NVIDIA A100 (40 GB) each	Google Colab	∼2:31 each
sPlotOpen				
splot_focal	41,42,43	1× NVIDIA A100 (80 GB) each	Google Colab	∼27:36 each