Biodiversity Profile Estimation with Earth Observation

Moritz Dieing[1][0009−0009−4862−1109], Michele Volpi[2][0000−0003−2771−0750], and Victor Boussange[3][0000−0002−4202−3599]

[1] Department of Computer Science, ETH Zurich, Zurich, Switzerland
[2] School of Computer and Communication Sciences, EPFL, Lausanne, Switzerland
[3] Swiss Data Science Center, ETH Zurich and EPFL, Switzerland

Abstract. Estimating the diversity of ecosystems requires extrapolating sparse, fine-scale field observations to spatial units orders of magnitude larger, while summarizing the multifaceted nature of biodiversity into comprehensive metrics. We introduce DIPSnet, a multiscale convolutional neural network that directly ingests dense geospatial tiles to jointly predict a full biodiversity profile, expressed as Hill diversity indices, across ecologically and policy relevant spatial scales. Extending MuScaRi [4], DIPSnet learns spatial context end to end from raw geospatial data without relying on handcrafted spatial statistics, and simultaneously predicts species richness (H_0), Shannon diversity (H_1), and Simpson diversity (H_2), the three canonical members of the Hill family, enabling comprehensive characterization of biodiversity including community evenness, a property linked to ecological resilience. Predictions are structured through a Weibull rarefaction parameterization that enforces statistically motivated saturation behavior and enables principled extrapolation from partial observations to fully sampled target areas. On simulated and empirical benchmarks, DIPSnet achieves the best performance among the evaluated baselines, reducing RMSE by up to 20% over the strongest existing baseline considered. Applied to vascular plant communities across the Netherlands, Czech Republic, and Quebec, it produces spatially coherent diversity maps at multiple scales, faithfully recovering scale-dependent biodiversity patterns and delivering actionable information for large-scale ecological monitoring.

Keywords: biodiversity monitoring · Earth observation · multi-scale learning · Hill diversity indices · rarefaction

1 Introduction

Biodiversity is undergoing unprecedented change due to global stressors, with anthropogenic activities driving significant species losses at both local and regional scales [18]. Empirical evidence consistently shows that losses in species richness reduce ecosystem multifunctionality, amplifying vulnerability and the risk of crossing ecological tipping points [25, 22]. Measuring diversity across ecologically relevant extents is therefore essential for quantifying biodiversity responses to global change.

Ecological surveys conducted by experts are labor-intensive and therefore usually provide only sparse data at fine spatial granularity. As processes that shape biodiversity (dispersal limitation, species interactions, habitat heterogeneity) can operate differently across scales [21], upscaling local observations to regional extents remains a major challenge. While species richness is generally the most widely used index for biodiversity assessment [12], it captures only the number of species present and does not account for compositional variation such as relative species abundances. Complementing species richness with abundance-weighted indices such as Shannon and Simpson diversity enables a more nuanced characterization of community structure, capturing community evenness alongside species counts [15].

Due to the limited availability of high-resolution occurrence data across large spatial extents, a range of methods has been developed to estimate species richness indirectly. Historically, these approaches have been based on species-area relationships (SARs), which infer richness across scales from observations in nested sampling units by assuming a specific functional form, most commonly the Arrhenius power law [2]. While SARs have proven useful for characterizing broad macroecological patterns, they rely on strong assumptions about the shape of the area-richness relationship and require nested sampling designs that are seldom available in practice. More recent rarefaction-based approaches characterize how observed richness increases with sampling effort (the total survey area within a study unit; see Sec. 3.1) within a spatial unit, offering greater flexibility for upscaling from scattered plots. However, these methods generally do not yield spatially explicit predictions transferable across regions, and critically, none of them can leverage the rich spatial information contained in modern dense geospatial datasets (see Sec. 2 for a review). MuScaRi [4] conditions a parametric rarefaction model on aggregated statistics of environmental covariates, enabling spatially-explicit multi-scale predictions from scattered plot data. However, its reliance on aggregated statistics discards the intra-unit spatial structure present in the geospatial context and is limited to species richness prediction.

In this work we introduce DIPSnet[4], based on a multi-scale convolutional neural network (CNN) encoder that directly ingests dense geospatial tiles, enabling it to capture ecologically meaningful spatial patterns (such as habitat fragmentation) that aggregation-based approaches cannot resolve. DIPSnet extends MuScaRi to simultaneously predict three Hill diversity indices: species richness (H_0), Shannon diversity (H_1), and Simpson diversity (H_2) [16, 8], within a single end-to-end framework. Our experiments show that DIPSnet consistently outperforms shallower baselines, with the rarefaction-curve parameterization being critical under low sampling effort. Our contributions are:

- A multi-scale CNN encoder that explicitly processes geospatial tiles to jointly predict biodiversity indices at varying spatial scales, achieving the best performance among the evaluated baselines;
- A controlled evaluation using simulated data with known ground-truth diversity, enabling direct assessment of extrapolation performance;

[4] For "Diversity Index Prediction across Scales", also named after the plant *Dipsacus*.

– A curated vegetation plot dataset collating European and Canadian vegetation survey data for training and evaluation, and associated predicted vascular plant diversity maps across multiple spatial scales.

2 Related Work

Deep learning and EO data for species distribution modelling. CNN-based species distribution models can explicitly learn spatial patterns and improve predictive performance over classical pixel-based counterparts [10]. Fine-grained EO data can further improve these models [11, 13, 28]. These studies demonstrate the value of EO data, but focus on predicting species distributions rather than community-level properties such as diversity indices, which provide aggregated, ecologically interpretable summaries of the full species pool.

Species richness and diversity estimation. SAR models have long served as a standard framework for estimating species richness, most commonly as the Arrhenius power law $S = cA^z$ [2], but rely on restrictive assumptions about the functional form of the area-richness relationship and require nested-sampling designs that are often unavailable in practice [14]. Rarefaction and asymptotic estimators [7] offer greater flexibility: these methods characterize how observed richness increases with sampling effort, enabling upscaling from scattered small-scale plots without contiguous nested samples. Parametric approaches assume an explicit accumulation curve, while non-parametric estimators such as Chao1 [6] infer asymptotic richness from the frequency of rare species. Neither class yields spatially explicit predictions transferable across environmentally similar but sparsely sampled regions.

With the availability of large opportunistic databases collating field surveys [9], machine learning alternatives become attractive. Andermann et al. [1] showed that neural networks can predict biodiversity measures at different scales directly from environmental variables, but without modeling the accumulation process, making them inappropriate for extrapolating beyond observed sampling effort. MuScaRi [4] addresses this by conditioning a parametric rarefaction curve (a 4-parameter Weibull function) on aggregated statistics of CHELSA bioclimatic covariates (mean and standard deviation per spatial unit) to predict species richness across scales and sampling efforts. While effective, this aggregation discards intra-unit spatial structure and is limited to species richness.

Multiscale EO architectures. Recent approaches move toward scale-aware representation learning. Scale-MAE learns multiscale geospatial representations through scale-aware masked autoencoding [24]. AnySat provides a unified EO model operating across resolutions and modalities using scale-adaptive encoders [3]. RAMEN addresses heterogeneous EO inputs with a resolution-adjustable encoder whose branch weights are conditioned on target resolution [17], a design principle adopted by DIPSnet.

No existing method exploits intra-unit spatial structure to predict diversity indices from geospatial data, conditioning on both scale and sampling effort. DIPSnet fills this gap by combining a multi-scale CNN encoder for geospatial

tiles with a rarefaction-curve parameterization that jointly predicts all three Hill diversity indices.

3 Methods

3.1 Definitions

Spatial unit of observation. A *spatial unit of observation* is a fixed geographic area A_u, represented by a geospatial tile. It may contain one or more biodiversity plots, and is associated with a sampling effort a defined as the total area covered by those plots. DIPSnet predicts diversity for a target subarea of extent $A_t \leq A_u$, specified through the zoom factor $\sigma_z = \sqrt{A_t/A_u}$, which is supplied to the model together with the geospatial tile.

Hill diversity indices. Hill diversity indices H_q [16], also known as effective numbers of species, are a family of generalized diversity measures parameterized by order q, which controls the sensitivity to rare versus common species. Given a community of S species with relative abundances $p_1, \ldots, p_S$, they are defined as $H_q = \left(\sum_{i=1}^{S} p_i^q\right)^{\frac{1}{1-q}}$. Increasing q progressively down-weights rare species. DIPSnet jointly predicts Hill indices of orders 0, 1, and 2, corresponding to species richness, Shannon diversity [26], and Simpson diversity [27]. Hill diversity across multiple orders provides information on community evenness, a structural property often linked to ecological stability and resilience.

Rarefaction curve. For a spatial unit, the rarefaction curve $\mathcal{D}_q(a)$ describes the expected Hill diversity of order q as a function of sampling effort a. $\mathcal{D}_q$ is monotonically increasing and converges to an asymptotic value as $a \to \infty$; extrapolating to asymptotic effort yields an estimate of diversity at full observation. In practice, field surveys only partially sample the spatial unit ($a \ll A_u$), motivating the use of a parametric rarefaction model.

Problem statement. Let $\mathcal{D} = \{(M_i, a_i, \sigma_{z,i}, y_i)\}_{i=1}^{N}$ be a dataset of N samples, where each sample consists of a d-dimensional geospatial tile of square side s, $M_i \in \mathbb{R}^{s \times s \times d}$, a sampling effort a_i, a zoom factor $\sigma_{z,i}$, and ground-truth diversity indices $y_i = (H_{0,i}, H_{1,i}, H_{2,i})$. Sampling effort a_i is the total area of biodiversity plots within the target spatial unit for $\sigma_{z,i}$. The goal of DIPSnet is to predict Hill indices from geospatial inputs when $a_i = (\sigma_{z,i} s)^2$ spans the entire unit.

3.2 Model Architecture

The proposed model (Figure 1) encompasses two variants: DIPSnet, which constrains predictions via a parametric rarefaction model enabling principled extrapolation to full sampling effort, and DIPSnet-Unconstrained, which predicts diversity directly but lacks this inductive bias. DIPSnet-Unconstrained serves as a spatial baseline for DIPSnet.

Multi-Scale Encoder. The encoder comprises n parallel convolutional branches, each built from a sequence of blocks with 3×3 convolutions (stride 2, padding 1) followed by ReLU nonlinearities. Feature dimensionality doubles after each block

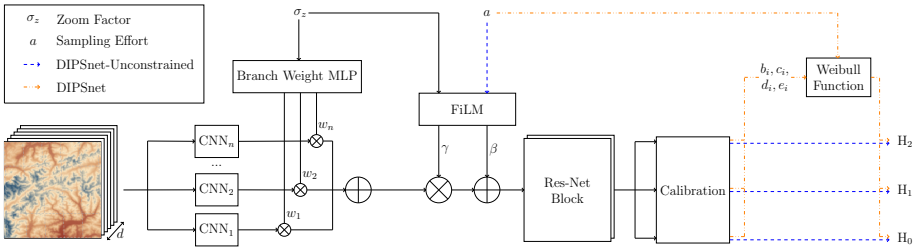

Fig. 1: DIPSnet and DIPSnet-Unconstrained share a multi-scale CNN encoder and a ResNet decoder. The encoder maps the input tile through n convolutional branches, adaptively aggregated via MLP+softmax conditioned on σ_z. A FiLM layer modulates the latent embedding. In DIPSnet, the decoder predicts the parameters $\Theta = \{c, d, b, e\}$ of the rarefaction curve (Equation 1), which are evaluated at sampling effort a to yield Hill indices. In DIPSnet-Unconstrained, the decoder predicts diversity indices directly, conditioned on both a and σ_z.

as spatial resolution decreases. A bottleneck convolution with kernel size equal to the remaining feature-map size produces a single latent vector per branch. The per-branch vectors are stacked along the branch dimension into a $(n \times c)$ matrix, jointly layer-normalized, and combined via a weighted sum following the RAMEN design [17]: a small MLP followed by softmax produces normalized coefficients conditioned on σ_z for adaptive branch aggregation, allowing each branch to specialize in features at a different scale.

FiLM Layer. Feature-wise Linear Modulation (FiLM) [23] incorporates scalar conditioning variables as $\hat{x} = x \cdot \gamma + \beta$, where γ and β are predicted by an intermediate MLP. In DIPSnet-Unconstrained, the FiLM layer receives both a and σ_z. In DIPSnet, it is conditioned only on σ_z: feeding a to the encoder would allow it to circumvent the saturation constraint of the rarefaction curve, undermining the monotonicity and saturation inductive bias that is key to reliable extrapolation. A second FiLM conditioning on σ_z is applied at the decoder input, ensuring scale information is present throughout the forward pass.

Decoder and Calibration. The decoder uses a ResNet-style architecture with pre-normalized residual blocks (Layer Normalization, GELU, dropout) and an initial input projection. In DIPSnet, following MuScaRi [4], the decoder outputs the parameters Θ of the rarefaction curve for each of the three Hill indices (12 outputs total); diversity values are then obtained by evaluating g_Θ at effort a. In DIPSnet-Unconstrained, diversity indices are predicted directly. A learnable affine transformation calibrates scale and bias across all outputs.

The parametric rarefaction function g_Θ is defined as the 4-parameter Weibull function, following [4, 30]:

$$g(a) = c + (d - c)\left(1 - \exp\left[-\ln(a/e)^b\right]\right) \tag{1}$$

where c is the intercept at minimum effort, d is the asymptotic diversity, b controls the shape of accumulation, and e is a scale parameter. The 4-parameter Weibull function provides a reliable parametric estimator of rarefaction curves

by enforcing monotonicity and saturation: g_Θ is strictly increasing and saturates at d as $a \to \infty$, encoding the prior that diversity accumulates sublinearly with sampling effort and enabling reliable extrapolation in the partial-observation regime [4].

4 Experiments

We evaluate DIPSnet against baselines in three settings: simulated data with ground-truth diversity to directly assess extrapolation; empirical data across three regions; and ablation studies examining encoder design and context.

4.1 Data

Simulated Biodiversity Data. We use synthetic vegetation plots based on a radial niche model. Conditioned on environmental covariates, the simulator generates species-specific suitability surfaces and presence-only plots via spatially biased sampling. Species niches are defined by environmental niche centers and widths; local individual counts are drawn from suitability-dependent Poisson distributions. Each plot provides spatial coordinates and detected species IDs, yielding a controlled benchmark where ground-truth diversity at full observation is exactly known.

Empirical Biodiversity Data. Following [4], we use 502,724 vegetation plots from the European Vegetation Archive (EVA, [9]; project 172; retrieved 27 February 2023), restricting to the Netherlands (69,123 plots) and the Czech Republic (60,305 plots), containing the highest densities of available plots. The dataset is publicly available at `https://huggingface.co/vboussange/MuScaRi`. We additionally use data from the Ministère des Ressources naturelles et des Forêts of Quebec (donneesquebec.ca; retrieved 23 February 2026), comprising 28,398 plots collected under a standardized sampling protocol over a shorter time interval, yielding a cleaner and more evenly distributed dataset.

Geospatial data. Model inputs are seven bioclimatic features from the CHELSA dataset [19, 5] (climatological data derived from local observations, Earth observation, and remote sensing) at 1 km resolution, capturing temperature, water availability, and energy balance.

Training Sample Generation. Samples are generated by randomly selecting spatial unit centers and sampling a square with side between 1×1 km and 25×25 km, which determines σ_z. In the simulated setting, sampling effort is modeled by individual subsampling: a value $\eta \in [0, 1)$ is drawn, the population size N is reduced to $\hat{N} = \eta N$ via multinomial resampling, and effort is set to $\hat{N}$. In the empirical setting, effort is the union of plot areas within the spatial unit [4]. CHELSA bioclimatic tiles are extracted from the largest square centered on the spatial unit midpoint.

To ensure unbiased evaluation, we use six train/validation/test splits based on spatially structured partitioning. For empirical datasets, a block-based cross-validation strategy is adopted following [4]: the study area is partitioned into 1×1 km blocks, each assigned to one of six folds, and each vegetation plot is

mapped to the fold of its block. For each split, one fold is held out for testing, one for validation, and the remaining four for training. For simulated data, a 6×4 grid is overlaid on the study region; each cell is assigned to training/validation or test, and buffer zones are enforced to ensure sampled tiles do not overlap across splits.

4.2 Model and Training Setup

Baseline models. We compare against three baselines: (1) **MLP**, using the aggregated-feature architecture of MuScaRi but predicting diversity indices directly without a rarefaction curve; (2) **MuScaRi** [4], extended to jointly predict three diversity indices with its parametric rarefaction formulation; and (3) **DIPSnet-Unconstrained**, using the same multi-scale encoder and ResNet decoder as DIPSnet while predicting diversity directly.
Implementation details. Input tiles are 25×25 grid cells (25×25 km at 1 km CHELSA resolution). The encoder comprises $n = 10$ convolutional branches; encoder and decoder hidden dimensions are 128 and 64; the decoder has two ResNet blocks. Models are trained for up to 50 epochs with batch size 128 using AdamW [20] (learning rate 10^{-3}; 5×10^{-4} for rarefaction-constrained models on empirical data to improve convergence), weight decay 10^{-4}, and MSE loss. A `ReduceLROnPlateau` scheduler (factor 0.5, patience 2, minimum 10^{-6}) and early stopping (patience 5 epochs) are applied. All experiments were conducted on an NVIDIA RTX 3090. Code for all models, training scripts, and experiments is publicly available at `https://github.com/moritzdieing/dipsnet`.

4.3 Simulated Data Experiments

We compare all models against ground-truth diversity under three sampling scenarios:

- **Full observation.** Training and test sampling effort set to the full population ($\eta = 1$).
- **Variable sampling effort.** $\eta \in [0, 1)$ for both training and test.
- **Partial observation.** $\eta \in [0, 0.1)$ at training; $\eta = 1$ at test. This is the most challenging and realistic setting, where models trained on sparse observations must extrapolate to fully observed areas.

Performance is reported as RMSE. To compare across diversity indices with different numerical ranges, we additionally report the weighted mean absolute percentage error:

$$\mathrm{WMAPE}_j = \frac{\sum_{i=1}^{N} \left| H_{j,i} - \hat{H}_{j,i} \right|}{\sum_{i=1}^{N} |H_{j,i}|}, \tag{2}$$

where $H_{j,i}$ and $\hat{H}_{j,i}$ denote observed and predicted values for index j. WMAPE is preferred over MAPE for its robustness to small target values.

Table 1: Model performance as mean RMSE (± std) on simulated data.

	Full Observation			**Variable Sampling Effort**			**Partial Observation**		
Model	H_0	H_1	H_2	H_0	H_1	H_2	H_0	H_1	H_2
MLP	5.19 ±0.33	1.59 ±0.19	1.23 ±0.18	5.48 ±0.70	2.03 ±0.28	1.61 ±0.28	1469.95 ±2011.81	97.62 ±128.35	55.68 ±66.86
MuScaRi	5.83 ±0.40	1.87 ±0.20	1.46 ±0.18	5.34 ±0.59	2.12 ±0.31	1.71 ±0.32	14.82 ±2.48	3.26 ±0.59	2.30 ±0.48
DIPSnet-Unconst.	**3.86** ±1.19	**1.00** ±0.18	**0.77** ±0.15	4.87 ±0.70	1.67 ±0.42	1.31 ±0.34	22.44 ±11.79	11.84 ±5.82	7.22 ±3.33
DIPSnet	4.10 ±1.19	1.20 ±0.31	0.96 ±0.25	**4.52** ±0.91	**1.49** ±0.32	**1.18** ±0.26	**14.40** ±3.06	**2.62** ±0.63	**1.91** ±0.55

4.4 Empirical Data Experiments

As empirical datasets do not provide ground-truth at full observation, only the variable sampling effort setting is evaluated. DIPSnet is compared against baselines on the Netherlands, Czech Republic, and Quebec datasets.

4.5 Ablation Studies

Convolutional expert contribution. We compare DIPSnet variants with 5, 10, and 15 convolutional experts (1.1M, 1.7M, and 2.6M parameters) to assess the trade-off between predictive performance and computational cost.
Multi-scale capabilities. We compare DIPSnet trained with a fixed zoom factor $\sigma_z = 0.2$ against one trained with randomly varying zoom factors, evaluating both at $\sigma_z = 0.2$ (5×5 km) on the Czech EVA dataset.
Context relevance. Because spatial context surrounding a target area may contain useful information about its species composition [10], we compare DIPSnet trained on 25×25 km tiles (including a 5 km context ring beyond the maximum 15×15 km target) against one trained on 15×15 km tiles matched to the target extent, using Czech EVA data.

5 Results

5.1 Simulated Data Experiments

Numerical Results. Table 1 presents the evaluation on simulated data across three sampling scenarios.

Across all settings, both DIPSnet variants substantially outperform MuScaRi, confirming that directly ingesting EO data with a CNN is beneficial. In the full-observation setting, DIPSnet-Unconstrained achieves the best performance, reducing RMSE relative to the best shallow baseline (MLP) by 25.6% for H_0, 37.1% for H_1, and 37.4% for H_2. The unconstrained variant's greater flexibility is

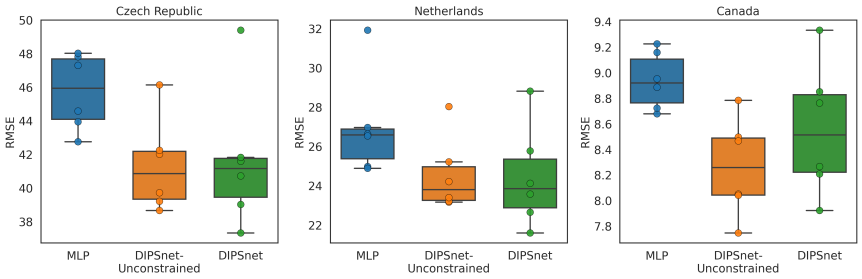

Fig. 2: Species richness (H_0) interpolation performance (RMSE) across empirical datasets and model variants.

advantageous when no extrapolation is required. In the variable-sampling-effort setting, DIPSnet performs best, improving over MLP and MuScaRi by 22.9% on average across all three indices.

In the partial-observation setting, the most practical scenario where models learn from sparse surveys but must predict full-coverage diversity, unconstrained models fail catastrophically: MLP RMSE for H_0 reaches 1469.95 due to severe overestimation under low sampling effort. In contrast, MuScaRi and DIPSnet, which impose a rarefaction-curve structure, remain stable. The saturation encoded by the rarefaction curve prevents runaway predictions when sampling effort is far below the training distribution. Within this regime, DIPSnet further improves over MuScaRi by 2.8% for H_0, 19.6% for H_1, and 17.0% for H_2, demonstrating that deeper feature extraction complements the extrapolation inductive bias. These improvements in H_1 and H_2 are particularly notable: MuScaRi relies on spatial statistics (means and standard deviations of bioclimatic variables) that can capture broad richness gradients but miss the fine intra-unit spatial structure relevant to abundance-sensitive indices. By directly encoding spatial patterns via the CNN, DIPSnet can leverage habitat heterogeneity and landscape structure as cues for evenness, which are precisely the features lost by aggregation.

Diversity Index Comparison. Prediction accuracy differs consistently across Hill indices. Across all settings, H_1 (Shannon diversity) yields the lowest relative error. In the partial-observation setting, WMAPE for DIPSnet reaches 24.4 for H_0 but only 11.0 and 11.4 for H_1 and H_2, respectively. This pattern is consistent with the ecology of sampling: rare species contribute disproportionately to H_0 but are progressively down-weighted in H_1 and H_2, so sparse sampling, which preferentially misses rare species, degrades richness estimation far more sharply than Shannon or Simpson diversity. In the first two settings, errors remain comparably small across all indices.

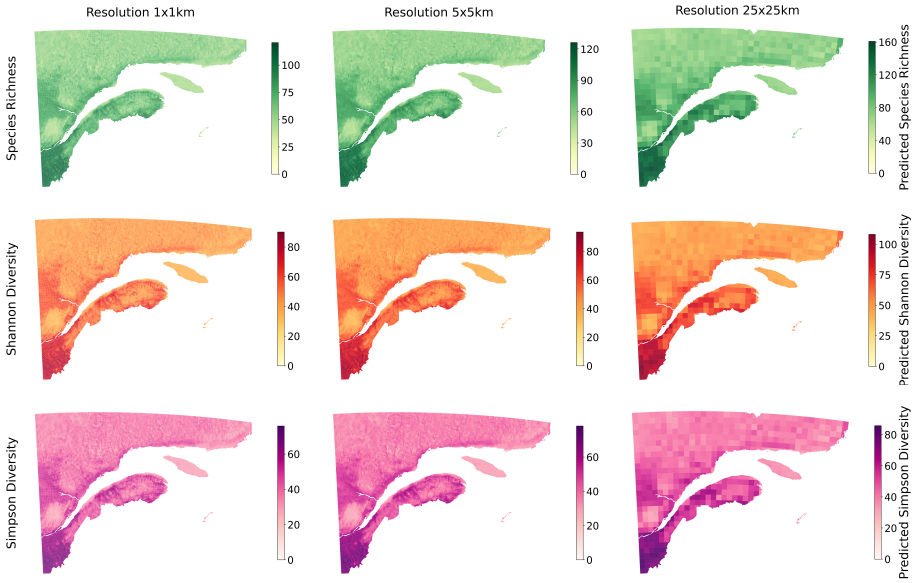

Fig. 3: **Predicted diversity across Quebec at three spatial scales.** Columns: 1×1, 5×5, and 25×25 km. Rows: Species richness H_0 (green), Shannon diversity H_1 (orange), Simpson diversity H_2 (pink).

5.2 Empirical Data Experiments

Performance on empirical datasets confirms the trends from simulation: both DIPSnet and DIPSnet-Unconstrained consistently outperform the MLP baseline (Figure 2). As empirical datasets do not provide diversity at full observation, this evaluation measures interpolation under observed sampling effort, rather than extrapolation ability under asymptotic sampling effort. Independent exhaustive-inventory datasets such as GIFT [29] would be more suited for assessing extrapolation; however, they are aggregated to administrative units whose spatial resolution is too coarse for DIPSnet at the selected input scale. On the Czech dataset, DIPSnet reduces RMSE relative to MLP by 8.9%, 11.2%, and 11.1% for H_0, H_1, and H_2. On the Dutch dataset, reductions are 9.5%, 8.9%, and 9.4%. On the Canadian dataset, absolute errors are markedly smaller, reflecting the dataset's narrower species pool (536 species vs. up to 2187 in EVA) and more uniform sampling conditions relative to the long-assembled, heterogeneous EVA plots; model differences are also narrower, with DIPSnet-Unconstrained slightly outperforming constrained DIPSnet (7.5%, 7.7%, and 8.0% vs. 4.4%, 5.2%, and 5.3% reduction relative to MLP for H_0, H_1, H_2).

In Figure 3, predicted H_0, H_1, and H_2 across Quebec at 1×1, 5×5, and 25×25 km show broadly consistent spatial patterns across all scales, confirming that the model captures stable geographic biodiversity gradients. Coarser grids progressively smooth local hotspots, producing more homogeneous maps. Predicted diversity increases toward the south, consistent with warmer climatic conditions, the influence of the Great Lakes, and mixed forest ecosystems; lower

diversity in the north reflects boreal forest dominance. Maximum predicted diversity increases systematically with coarser resolution, as larger cells aggregate more species. Notably, H_0, H_1, and H_2 maps reveal similar spatial patterns, suggesting that evenness is fairly consistent across the region and that species dominance varies only weakly in space, an ecologically informative result that species richness alone could not provide.

5.3 Ablation Studies

Convolutional expert contribution. Among the tested variants, the 10-expert model achieves the best overall performance (Czech EVA RMSE: 41.66, 29.44, 24.10 for H_0, H_1, H_2). The 5-expert variant performs markedly worse, particularly for H_1 and H_2 (57.69 and 42.25). Increasing to 15 experts does not improve performance and increases training cost. Under the current experimental setup, ten convolutional experts provide the most effective balance of capacity and performance.

Multi-scale capabilities. The multi-scale encoder generalizes across scales with negligible loss compared to single-scale training. At 5×5 km, the fixed-scale model achieves RMSE of 30.75, 23.49, 19.49 for H_0, H_1, H_2, versus 31.40, 23.55, 19.85 for the multi-scale model, a relative difference of 2.1% for H_0. Multi-scale training does not substantially compromise single-scale accuracy.

Context relevance. Extending input tiles from 15×15 to 25×25 km (adding a 5 km context ring) yields marginal changes: RMSE shifts from 37.57 to 37.51 for H_0, from 26.41 to 26.86 for H_1, and from 21.49 to 21.91 for H_2. The spatial smoothness of CHELSA bioclimatic variables at 1 km resolution limits the ecological signal in the surrounding ring relative to the target unit itself. DIPSnet therefore uses tiles matched to the maximum target extent; high-resolution satellite imagery could make additional context more informative in future work.

6 Conclusion

DIPSnet extends MuScaRi [4] by replacing aggregated-statistic inputs with a multi-scale CNN encoder capable of ingesting raw geospatial data, and by generalizing the rarefaction-based framework to the joint prediction of three Hill diversity indices. Across simulation and empirical experiments, DIPSnet consistently outperforms shallower baselines, showing that deep latent representations learned directly from EO tiles are more effective than aggregated statistics. Critically, the rarefaction-curve parameterization is essential in the partial-observation regime (the setting closest to real-world monitoring), where unconstrained models fail catastrophically while DIPSnet remains stable. By combining this extrapolation inductive bias with direct spatial EO feature extraction, DIPSnet provides a scalable framework for multi-scale biodiversity mapping,

delivering a comprehensive, spatially-explicit view of the diversity profile of a spatial unit of observation. This enables not only recovery of scale-dependent biodiversity patterns, but also characterization of spatial variation in evenness, opening the way to more informative large-scale assessments of ecological organization and resilience.

Limitations and future work. Extrapolation performance is established only on simulated data; empirical validation against exhaustive inventory datasets such as GIFT [29] remains outstanding. The model is trained per region and is not expected to generalize beyond its training geographical extent; combining multiple regions in a single model is a promising next step. The present comparison is also limited to aggregated-feature and DIPSnet variants; future benchmarks should include diversity estimates obtained from multi-species species distribution models [13, 28]. The coarse resolution of CHELSA bioclimatic features (1 km) limits the fine-grained spatial signal available to the encoder; a natural next step is to replace or complement them with higher-resolution inputs such as satellite imagery or pretrained geospatial embeddings [24, 3, 17]. Further directions include a probabilistic variant of DIPSnet for uncertainty quantification in low-observation and extrapolation regimes; multi-region training for geographic generalization; and a temporal extension to track biodiversity trends under global change.

Acknowledgements. Victor Boussange was supported by the Swiss National Science Foundation (SNSF) through the Spark grant CRSK-3_237783. We thank Dominique Gravel for helpful discussions on the manuscript.

Bibliography

[1] Andermann, T., Antonelli, A., Barrett, R., Silvestro, D.: Estimating Alpha, Beta, and Gamma Diversity Through Deep Learning. Frontiers in Plant Science **13**, 839407 (Apr 2022). https://doi.org/10.3389/fpls.2022.839407

[2] Arrhenius, O.: Species and Area. Journal of Ecology **9**(1), 95–99 (1921). https://doi.org/10.2307/2255763, https://www.jstor.org/stable/2255763

[3] Astruc, G., Gonthier, N., Mallet, C., Landrieu, L.: AnySat: One Earth Observation Model for Many Resolutions, Scales, and Modalities (May 2025). https://doi.org/10.48550/arXiv.2412.14123, http://arxiv.org/abs/2412.14123

[4] Boussange, V., Wuyts, B., Brun, P., Malle, J.T., Midolo, G., Portier, J., et al.: Multi-scale species richness estimation with deep learning (Aug 2025). https://doi.org/10.48550/arXiv.2507.06358, http://arxiv.org/abs/2507.06358, arXiv:2507.06358 [q-bio]

[5] Brun, P., Zimmermann, N.E., Hari, C., Pellissier, L., Karger, D.N.: Global climate-related predictors at kilometer resolution for the past and future. Earth System Science Data **14**(12), 5573–5603 (Dec 2022). https://doi.org/10.5194/essd-14-5573-2022, https://essd.copernicus.org/articles/14/5573/2022/

[6] Chao, A.: Non-parametric estimation of the classes in a population. Scandinavian Journal of Statistics **11**, 265–270 (Jan 1984). https://doi.org/10.2307/4615964

[7] Chao, A., Chiu, C.H.: Species Richness: Estimation and Comparison. In: Wiley StatsRef: Statistics Reference Online, pp. 1–26. John Wiley & Sons, Ltd (2016). https://doi.org/10.1002/9781118445112.stat03432.pub2, https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118445112.stat03432.pub2

[8] Chao, A., Chiu, C.H., Jost, L.: Unifying Species Diversity, Phylogenetic Diversity, Functional Diversity, and Related Similarity and Differentiation Measures Through Hill Numbers. Annual Review of Ecology, Evolution, and Systematics **45**(Volume 45, 2014), 297–324 (Nov 2014). https://doi.org/10.1146/annurev-ecolsys-120213-091540, https://www.annualreviews.org/content/journals/10.1146/annurev-ecolsys-120213-091540

[9] Chytrý, M., Hennekens, S.M., Jiménez-Alfaro, B., Knollová, I., Dengler, J., Jansen, F., et al.: European Vegetation Archive (EVA): an integrated database of European vegetation plots. Applied Vegetation Science **19**(1), 173–180 (2016). https://doi.org/10.1111/avsc.12191, https://onlinelibrary.wiley.com/doi/abs/10.1111/avsc.12191

[10] Deneu, B., Joly, A., Bonnet, P., Servajean, M., Munoz, F.: Very High Resolution Species Distribution Modeling Based on Remote Sensing

Imagery: How to Capture Fine-Grained and Large-Scale Vegetation Ecology With Convolutional Neural Networks? Frontiers in Plant Science **13** (May 2022). https://doi.org/10.3389/fpls.2022.839279, https://www.frontiersin.org/journals/plant-science/articles/10.3389/fpls.2022.839279/full

[11] Dollinger, J., Brun, P., Sainte Fare Garnot, V., Wegner, J.D.: SatSINR: High-Resolution Species Distribution Models Through Satellite Imagery. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences **X-2-2024**, 41–48 (Jun 2024). https://doi.org/10.5194/isprs-annals-X-2-2024-41-2024, https://isprs-annals.copernicus.org/articles/X-2-2024/41/2024/, conference Name: ISPRS TC II Mid-term Symposium "The Role of Photogrammetry for a Sustainable World" - 11–14 June 2024, Las Vegas, Nevada, USA

[12] Fleishman, E., Noss, R.F., Noon, B.R.: Utility and limitations of species richness metrics for conservation planning. Ecological Indicators **6**(3), 543–553 (Aug 2006). https://doi.org/10.1016/j.ecolind.2005.07.005, https://www.sciencedirect.com/science/article/pii/S1470160X05000592

[13] Gillespie, L.E., Ruffley, M., Exposito-Alonso, M.: Deep learning models map rapid plant species changes from citizen science and remote sensing data. Proceedings of the National Academy of Sciences **121**(37), e2318296121 (Sep 2024). https://doi.org/10.1073/pnas.2318296121, https://www.pnas.org/doi/10.1073/pnas.2318296121

[14] Guilhaumon, F., Gimenez, O., Gaston, K.J., Mouillot, D.: Taxonomic and regional uncertainty in species-area relationships and the identification of richness hotspots. Proceedings of the National Academy of Sciences **105**(40), 15458–15463 (Oct 2008). https://doi.org/10.1073/pnas.0803610105, https://www.pnas.org/doi/abs/10.1073/pnas.0803610105

[15] Haegeman, B., Hamelin, J., Moriarty, J., Neal, P., Dushoff, J., Weitz, J.S.: Robust estimation of microbial diversity in theory and in practice. The ISME Journal **7**(6), 1092–1101 (Jun 2013). https://doi.org/10.1038/ismej.2013.10, https://doi.org/10.1038/ismej.2013.10

[16] Hill, M.O.: Diversity and Evenness: A Unifying Notation and Its Consequences. Ecology **54**(2), 427–432 (1973). https://doi.org/10.2307/1934352, https://www.jstor.org/stable/1934352

[17] Houdré, N., Marcos, D., Turckheim, H.R.d., Ienco, D., Wendling, L., Kurtz, C., et al.: RAMEN: Resolution-Adjustable Multimodal Encoder for Earth Observation (Dec 2025). https://doi.org/10.48550/arXiv.2512.05025, http://arxiv.org/abs/2512.05025, arXiv:2512.05025 [cs]

[18] IPBES: Global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services. Tech. rep., Zenodo (May 2019). https://doi.org/10.5281/zenodo.6417333, https://zenodo.org/records/6417333

[19] Karger, D.N., Conrad, O., Böhner, J., Kawohl, T., Kreft, H., Soria-Auza, R.W., et al.: Climatologies at high resolution for the earth's land surface areas (2021). https://doi.org/http://dx.doi.org/10.16904/envidat.228, https://www.envidat.ch/dataset/chelsa-climatologies

[20] Loshchilov, I., Hutter, F.: Decoupled Weight Decay Regularization (Jan 2019). https://doi.org/10.48550/arXiv.1711.05101, https://arxiv.org/abs/1711.05101

[21] Mi, X., Bao, L., Jianhua, C., Ma, K.: Point process models, the dimensions of biodiversity and the importance of small-scale biotic interactions. Journal of Plant Ecology **7**(2), 126–133 (Apr 2014). https://doi.org/10.1093/jpe/rtt075, https://doi.org/10.1093/jpe/rtt075

[22] Mori, A.S., Isbell, F., Cadotte, M.W.: Assessing the importance of species and their assemblages for the biodiversity-ecosystem multifunctionality relationship. Ecology **104**(8), e4104 (2023). https://doi.org/10.1002/ecy.4104, https://onlinelibrary.wiley.com/doi/abs/10.1002/ecy.4104

[23] Perez, E., Strub, F., Vries, H.d., Dumoulin, V., Courville, A.: FiLM: Visual Reasoning with a General Conditioning Layer (Dec 2017). https://doi.org/10.48550/arXiv.1709.07871, http://arxiv.org/abs/1709.07871, arXiv:1709.07871 [cs]

[24] Reed, C.J., Gupta, R., Li, S., Brockman, S., Funk, C., Clipp, B., Keutzer, K., Candido, S., Uyttendaele, M., Darrell, T.: Scale-MAE: A Scale-Aware Masked Autoencoder for Multiscale Geospatial Representation Learning (Sep 2023). https://doi.org/10.48550/arXiv.2212.14532, http://arxiv.org/abs/2212.14532

[25] Rockström, J., Gupta, J., Qin, D., Lade, S.J., Abrams, J.F., Andersen, L.S., et al.: Safe and just Earth system boundaries. Nature **619**(7968), 102–111 (Jul 2023). https://doi.org/10.1038/s41586-023-06083-8, https://www.nature.com/articles/s41586-023-06083-8

[26] Shannon, C.E.: A Mathematical Theory of Communication. Bell System Technical Journal **27**(3), 379–423 (1948). https://doi.org/10.1002/j.1538-7305.1948.tb01338.x, https://onlinelibrary.wiley.com/doi/abs/10.1002/j.1538-7305.1948.tb01338.x

[27] Simpson, E.H.: Measurement of Diversity. Nature **163**(4148), 688–688 (Apr 1949). https://doi.org/10.1038/163688a0, https://www.nature.com/articles/163688a0

[28] Tikhonov, G., Tikhonov, D.: Synthesizing Joint and Deep Species Distribution Modeling to Enhance Spatial Prediction of Plant Communities at Continental Scale (2025), http://hdl.handle.net/10138/628128

[29] Weigelt, P., König, C., Kreft, H.: GIFT – A Global Inventory of Floras and Traits for macroecology and biogeography. Journal of Biogeography **47**(1), 16–43 (2020). https://doi.org/10.1111/jbi.13623, https://onlinelibrary.wiley.com/doi/abs/10.1111/jbi.13623

[30] Zou, Y., Zhao, P., Axmacher, J.C.: Estimating total species richness: Fitting rarefaction by asymptotic approximation. Ecosphere **14**(1), e4363 (Jan 2023). https://doi.org/10.1002/ecs2.4363, https://esajournals.onlinelibrary.wiley.com/doi/10.1002/ecs2.4363