Synthesizing Joint and Deep Species Distribution Modeling to Enhance Spatial Prediction of Plant Communities at Continental Scale

Notebook for the LifeCLEF Lab at CLEF 2025

Gleb Tikhonov[1,*], Dmitry Tikhonov[2]

[1]University of Helsinki, Viikinkaari 1, P.O. Box 65, 00014 Helsinki, Finland

[2]Center for Forest Ecology and Productivity, Russian Academy of Science, Profsoyuznaya st. 84/32 building 14, 117997 Moscow, Russian Federation

Abstract

Understanding the complex mechanisms that shape biological communities and the ability to accurately predict them are central research objectives in statistical community ecology, which provides the methodological foundation for data-driven biodiversity monitoring and responding to ongoing Global Change. Historically, this predictive task has been primarily approached through species distribution modeling (SDM), which treats species in a community independently. However, the growing recognition that a biological community is more than the sum of its individual species led to the emergence of joint species distribution modeling (JSDM), which has been increasingly used in community analysis over the past decade. At the same time, the increased availability of remote sensing data and advancements in geospatial AI tools led to the development of SDMs that build on deep learning techniques to leverage raw input data, such as satellite imagery. While JSDM and deep SDM are not mutually exclusive, their integration has been limited to date. In this work, we aim to fill this methodological gap by developing and evaluating a unified prototype framework that synthesizes the advances from both approaches. Specifically, we combine a deep learning feature extractor, capable of processing the satellite imagery, with a structured JSDM output layer represented by the Hierarchical Model of Species Communities, which enables to account for inter-species relationships and the spatial study design structure. We demonstrate the predictive utility of our approach using a large dataset of plant species communities across Europe, as part of the GeoLifeClef2025 data science challenge. Our solution was ranked second by predictive F_1-score on the hidden test partition. Our findings highlight that further integration of the joint and deep SDM may reveal previously unattainable opportunities for accurate, continuous spatial predictions at regional and global scales.

Keywords

joint species distribution modeling, deep learning, spatial statistics, community ecology, community modeling, hierarchical modeling of species communities, geospatial AI, European vegetation, multivariate data

1. Introduction

The intricate interplay of life, characterized by the distribution and abundance of diverse species, forms the fundamental topics of ecological research [1]. Understanding the complex mechanisms that shape biological communities is a central research topic in community ecology, which seeks to decipher how combinations of environmental filtering, species interactions, as well as stochastic spatial and temporal processes jointly structure local species assemblages [2]. To navigate this complexity, ecologists increasingly rely on Species Distribution Models (SDM), powerful numerical tools that correlate species observations with environmental variables to predict species ranges across diverse geographical, temporal and environmental ranges [3, 4]. The predictive utility of these models is the numerical foundation of many biodiversity-focused applications that are crucial in an era of Global Change, such as biodiversity monitoring, enabling the tracking of changes in species populations and community compositions over time, and conservation planning, where model outputs inform the design

CLEF 2025 Working Notes, 9–12 September 2025, Madrid, Spain

*Corresponding author.

gleb.tikhonov@helsinki.fi (G. Tikhonov); dtikhonov66@gmail.com (D. Tikhonov)

0000-0003-3040-0307 (G. Tikhonov); 0000-0002-9517-0728 (D. Tikhonov)

of protected areas and management strategies for threatened species [5]. Moreover, the continuous integration of novel data observation techniques, improved models for environmental forecasting, and refinement of community-level modeling through more sophisticated analytical techniques, are paving the path towards the ambitious goal of creating a "biodiversity Digital Twin" — a dynamic virtual representation of ecosystems that can simulate responses to various environmental changes and management interventions, thereby revolutionizing our approach to ecological forecasting and management [6].

While early SDM approaches often focused on modeling species individually, in reality species do not exist in isolation but are embedded within complex networks of interactions [7, 8]. This limitation initiated the development of Joint Species Distribution Models (JSDM), which represent a significant conceptual and analytical advancement [9, 10, 11]. In contrast to stacked SDMs that merely aggregate predictions from single-species models, JSDMs account for the multivariate nature of ecological communities. They achieve this by simultaneously modeling the responses of multiple species to their environment while also accounting for potential co-occurrence patterns that deviate from what would be expected based on environmental responses alone [7, 12]. These residual co-occurrence patterns can offer insights into biotic interactions (e.g., competition, facilitation) or reflect shared responses to unmeasured environmental factors [8]. A key strength of JSDMs is their capacity to integrate various data types, such as species traits and phylogenetic relationships, to explore how these species-level characteristics mediate species' environmental niches and influence community assembly [8, 13]. The Hierarchical Modeling of Species Communities (HMSC) framework exemplifies this integrative approach, offering a robust statistical platform to link species occurrences, environmental drivers, species-specific traits, and evolutionary histories to the underlying processes of community organization [10].

Despite their profound conceptual advantages and growing adoption, the application of JSDMs has faced multiple computational challenges, particularly to the increasingly large community datasets that emerge as the sampling techniques evolve. Initial JSDM formulations, such as those based on multivariate probit (MVP) models, encountered significant scalability hurdles, primarily because the number of parameters in the species-to-species covariance matrix grows quadratically with the number of species [12]. Generalized Linear Latent Variable Models (GLLVMs) were widely adopted as a more scalable alternative, using a smaller number of unobserved latent variables to represent the structure of species associations and reduce model dimensionality [8, 9]. However, as datasets continued to expand, encompassing hundreds or thousands of species and vast numbers of spatial locations, GLLVMs also started facing limitations, especially when incorporating spatially explicit structures. Addressing these persistent challenges became a critical research focus. For instance, Tikhonov et al. [14] achieved significant advances by integrating modern spatial statistical scalability techniques such as Gaussian predictive processes and nearest-neighbor Gaussian processes into the HMSC framework, enabling the analysis of community datasets with hundreds of species distributed over hundreds of thousands of spatial units. Concurrently, algorithmic innovations have been adopted : Pichler and Hartig [11] developed sjSDM, a JSDM approach that circumvents latent variables by employing approximate Monte Carlo integration for the joint likelihood, coupled with elastic net regularization, achieving substantial speed-ups especially with GPU execution and demonstrating scalability to thousands of species. More recently, efforts like the Hmsc-HPC extension package have focused on accelerating existing, well-established JSDM software by porting its computations to GPUs, resulting in potentially over 1000-fold performance speedup for large datasets without altering the core model structure [15]. These developments are crucial for leveraging the full potential of modern ecological datasets.

Another recent and rapidly evolving paradigm in species distribution modeling involves the utilization of deep learning techniques, giving rise to the deep SDMs [16]. These models, frequently employing computer vision architectures such as Convolutional Neural Networks (CNN) or Visual Transformers (ViT), are designed to learn intricate non-linear patterns directly from the raw data without the need to conduct careful feature extraction [17]. The development of robust deep SDMs largely benefits from emergence of dedicated, pre-trained foundation models for Earth observations. Such foundation models, trained on massive volumes of unlabeled satellite and/or environmental data, can learn generalizable

representations of the Earth's surface, which can then be fine-tuned for specific downstream tasks [18]. However, modern deep SDM limitedly reflect the well-established practices from the JSDM literature. First of all, as far as we are aware, published deep SDMs approaches predominantly can be classified neither as stacked SDM nor as JSDMs: they resemble JSDM in the sense that the training is done for all species simultaneously, but in their final network layer they treat all species as effectively independent outcomes like stacked SDMs. Further, while traditional JSDMs place a strong emphasis on parameterizing and interpreting ecological processes, such as species interactions or trait-mediated environmental responses, deep SDMs typically prioritize predictive accuracy at the expense of direct ecological interpretability. On the other hand, recent studies emphasise that both DNN-specific interpretability tools and generic classification assessment techniques can be employed to facilitate the transition from the fitted "black-box" deep SDM model towards human-intelligible insights desired by ecologists [17].

The GeoLifeCLEF initiative is an annual data science competition with a long history of innovation in biodiversity analysis, which pushes the boundaries of species distribution modeling [19]. The expert-validated presence-absence data, derived from European Vegetation Archive (EVA), is unique in its sheer scale compared to typical ecological datasets available for model development. This extensive dataset, while invaluable, is aggregated from multiple distinct surveys and sources, which introduce heterogeneity and potential biases that modeling approaches must robustly address. Furthermore, the GeoLifeCLEF contest offers a vast dataset of presence-only records, primarily sourced from GBIF. Incorporating this GBIF-extracted data to the analysis presents both significant challenges due to inherent sampling biases and lack of systematic absence information, but also great scientific interest, as presence-only data constitutes the most abundant and widely available form of biodiversity information globally [20].

This working note summarizes our analysis undertaken within the GeoLifeCLEF 2025 competition, where our best solution was ranked second by predictive F_1 score on the hidden test partition. Our core research objective was to assess the comparative performance of JSDM and deep SDM approaches, and to partially fill the gap between these methodologies by developing and evaluating a prototype method that will leverage the features of both frameworks. We additionally seek to compare the performance of our focal methods against the solutions produced by other competitors of the GeoLifeCLEF 2025 challenge. By testing these distinct modeling approaches with the massive data and rigorous evaluation criteria of GeoLifeCLEF 2025, our aim is to provide insight into their relative advantages and limitations for subsequent applied use cases of biodiversity assessment and high-resolution predictive mapping.

2. Data

The GeoLifeCLEF 2025 dataset contains species observation data, including Presence-Only (PO) occurrences and Presence-Absence (PA) surveys, paired with numerous environmental predictors. The predictors provided by the competition organizers include a rich set of environmental rasters, Sentinel-2 satellite images, 20 years of monthly climatic time series, and 21 years of quarterly Landsat time-series point values. There are around 5 million PO occurrences, 89 thousand PA survey records and 14.8 thousand of test locations. PA data features 5016 unique species, PO data contains 9709 species, and there are 11255 species totally in both PA and PO. All species names are anonymized and replaced with numerical indices. The dataset includes survey records in 43 European countries, covering nine biogeographic regions: Alpine, Anatolian, Atlantic, Black Sea, Boreal, Continental, Mediterranean, Pannonian, and Steppic. The data were collected between 2017 and 2021.

Notably, the spatial coverage of the data is highly heterogeneous in the study area. In particular, many regions contain only PO observations or PA test records, but totally lack training data from the PA training set. This spatial imbalance introduces limitations for model development, especially in areas without labeled PA training data for supervised learning. A visualization of the spatial distribution of the PA train, PO train is provided in Figure 1.

PA train data include information on the survey index, the list of species recorded in each survey, the country and biogeographic region, the geographic coordinates of the site along with their uncertainty, the

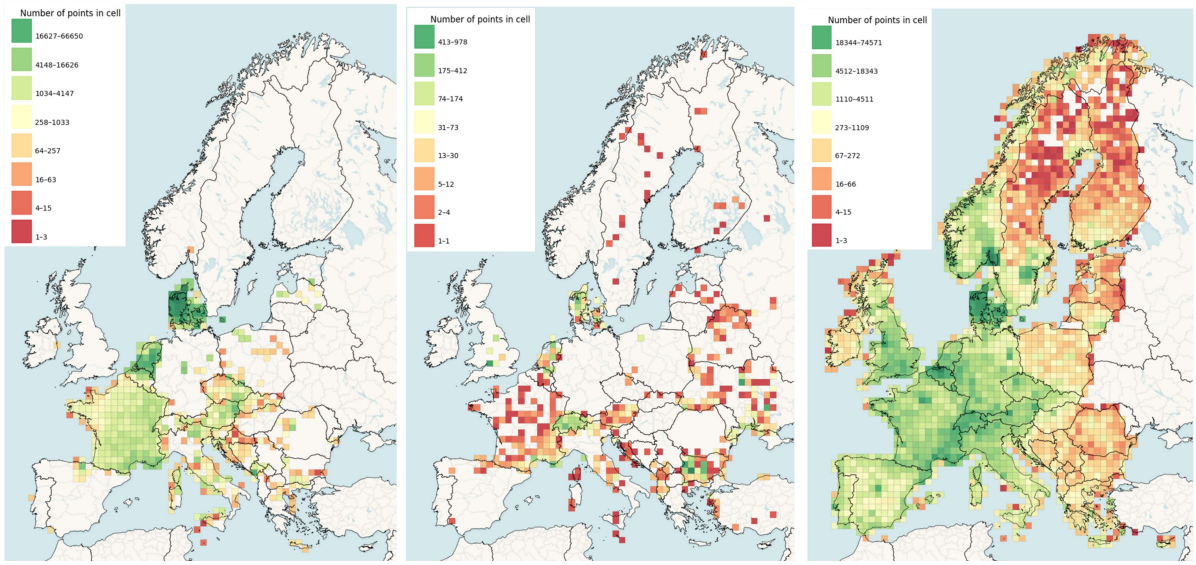

Figure 1: Spatial distribution of the PA train surveys (left), test surveys (center) and PO train observations (right) in GeoLifeCLEF 2025 competition. 90% of the test surveys are located in 8 countries: Bulgaria (22%), Ukraine (20%), Switzerland (13%), France (10%), Denmark (10%), Netherlands (6%), United Kingdom (6%) and Italy (4%). At the same time 90% of PA train surveys are located in 4 countries: Denmark (55%), Netherlands (17%), France (15%) and Italy (3%). Some test surveys on the eastern fringe of the study area are over 500 km away from the closest PA train or PO train data.

area of the surveyed plot, and the year of the survey. The test data are similar to PA train, with the obvious difference that the species lists are not available. PO occurrence data include the observation index, the recorded species, the date of the observation, the geographic coordinates with their uncertainty, and the publisher source from which the data were obtained.

In our final modeling approaches we used multiple environmental data sources provided within the GeoLifeCLEF dataset. Specifically, we relied on:

- **Sentinel-2 satellite imagery**, including RGB and near-infrared bands, capturing data over a 640 meter × 640 meter area at a 10-meter resolution, formatted into 64 × 64 pixel patches.
- **Landsat time series** (2000–2017), providing quarterly composites of six spectral bands at 30-meter resolution. We omitted the data from the years 2018-2020 as we noticed that for many PA train surveys some of that data was corrupt.
- **Monthly climatic time series** from CHELSA. We used only mean temperature and total precipitation. We aggregated the monthly data to quarterly data and truncated to years 2000-2017 to match the temporal support of the Landsat time series.
- **Soil variables**, describing physical and chemical soil properties.
- **Elevation data**.
- **Land cover information**, derived from global classification schemes. We reduced the originally provided data to 7 classes so that we avoid inclusion of highly correlated predictors.

In addition to the datasets provided by the organizers, we incorporated two auxiliary environmental characteristics to enrich the analysis. Specifically, we derived land cover information from the Sentinel-2 WorldCover product, which offers a global land cover classification at 10-meter spatial resolution for the year 2020. This dataset enables a detailed characterization of land use and vegetation structure across the study area. Furthermore, we utilized the CHELSA Snow Cover Duration dataset, which provides high-resolution data on the average duration of seasonal snow cover.

During the data exploration phase, we identified a high level of regional heterogeneity in the properties related to the vegetation sampling effort. For instance, plot area in the PA train data ranged from 0.01 m^2 to 8,000 m^2. Such high variation most likely also imply substantial variation in the applied surveying

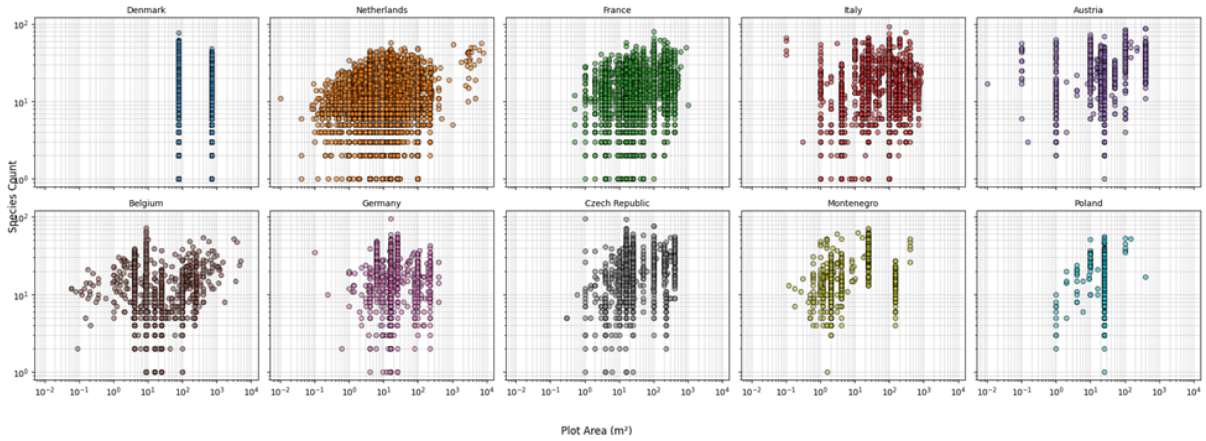

Figure 2: Log-log plots displaying the survey plot area vs number of species recorded in the survey. The patterns are visualized for top-10 countries with most PA train surveys. For majority of presented countries the expected pattern of positive relationship between survey area size and the species richness is visually recognizable. However, the PA training data contains both very large surveys (e.g. 707 m² in Denmark) with very low number of species and very small surveys with relatively large number of species (e.g. 0.01 m² in Austria). According to our domain knowledge, this is more expected to be an artifact of multiple data sources aggregation, where data was collected with different survey protocols, rather than the true ecological result. Unfortunately, such variation is notoriously difficult to account for in the predictive modeling.

protocol. Furthermore, some countries have highly standardized plot sizes (e.g. in Denmark all plots are either 79 m² or 707 m²), while other countries exhibit a broad variation of sampling plot areas (for instance, the Netherlands has 258 different plot sizes). Though our initial exploratory analysis showed only a very obscured relationship between recorded survey plot area and the number of species in that plot once considering all PA train data at once, the patterns became much more clear once we split the data according to the country from which it was collected (Figure 2). Once we recognized this major variation between different countries in the study, we made an attempt to assist our models to account for that. For this purpose we introduced 5 indicator variables, coding whether the survey was made in Denmark, Netherlands, France, Italy or any other country. In addition to that, we used the log of survey area.

Some predictors that we used in our modeling contain missing values. Whenever we had to make a choice with the missing values, we imputed them with the mean value of the corresponding predictor.

3. Methods

In this section we summarize the collection of modeling methods that we have tried in the GeoLifeCLEF 2025 competition. We order them in the chronological order in which we developed their corresponding analytical pipelines. We focus on the high-level descriptions of the final version of the methods that we developed and refer the interested reader to our publicly available GitHub repository that contains all the analytical code developed in scope of the competition. The lighter computational tasks and prototyping were done on a local desktop, equipped with NVIDIA 4070s GPU, while the more heavy computational jobs were executed in the GPU partition of Mahti HPC cluster with NVIDIA A100 GPUs.

3.1. Deep SDM

We started our participation in GeoLifeCLEF 2025 by expanding the "Single Modality baseline with Landsat data" example code based on ResNet-18 architecture that was generously provided by the organizers. We incorporated the climatic time series by aligning their temporal support to the Landsat data and adding them as extra channels for the Landsat+Climate cube. For the Sentinel rasters we employed the foundation ViT-type Prithvi-EO-2.0-300M model [18], for which we froze the encoder

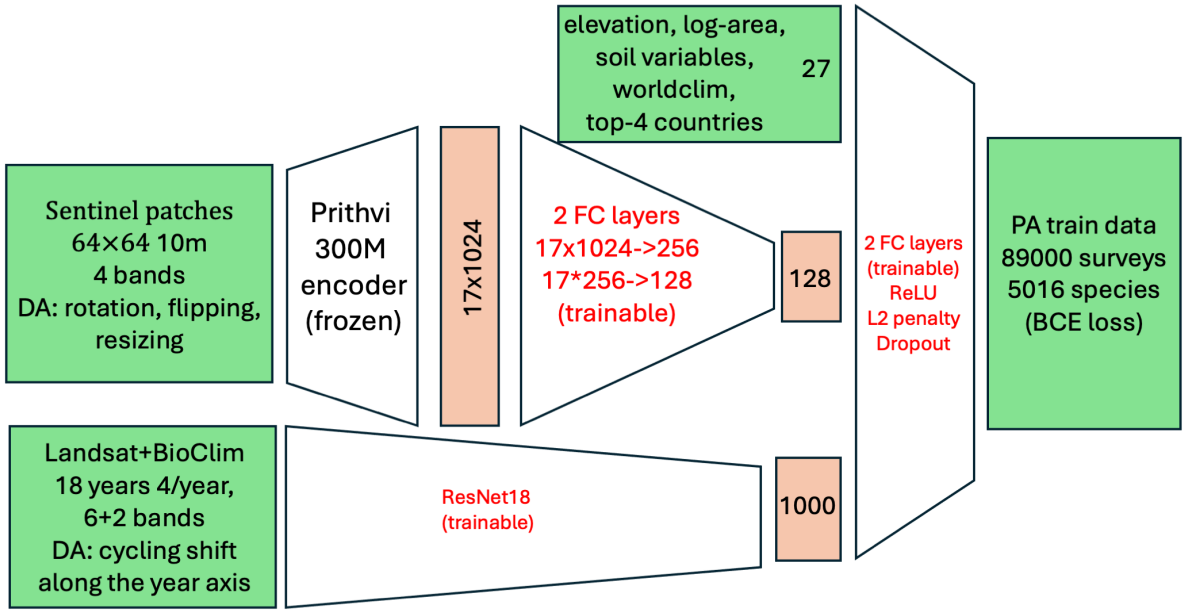

Figure 3: Final architecture of the DNN in our deep SDM approach. The same model without last FC layer was used as the deep feature extractor for the DNN-1 + HMSC approach.

and used a two-layer decoder motivated with examples distributed by this foundation model developers. We stacked the outputs of these networks with the rest of the non-raster covariates and added fully connected (FC) layer, ReLU activation and final FC layer with output corresponding to unique species in PA train data. Figure 3 visualizes the resulted DNN architecture. We used binary cross entropy loss and AdamW optimizer with cosine learning rate scheduler.

We normalized each channel in Landsat+Climate cubes and Sentinel images, and each covariate to zero mean and unit standard deviation on the PA train data. We used flip, rotation and rescaling data augmentation for the Sentinel images. For the Landsat+Climate cubes we used circular shift augmentation along the years axis of the cube. We used dropout in the two last hidden layers (except the covariates nodes) and L2 normalization in the last two FC layers of the resulted DNN.

The resulted DNN was trained with 90% random subset of the training PA data and the rest 10% of data used as validation for early stopping. We did a minor manual search for the more suitable included covariates, hyperparameters of the model (number of layers and nodes in hidden layers) and hyperparameters of the fitting algorithm (batch/layer normalization, dropout rate, L_2 penalty, learning rate, scheduler strategies).

3.2. HMSC

While there are plenty of JSDM software available, we focused solely on the HMSC approach [8, 10]. This choice was largely motivated by the fact that one of the authors is among the HMSC framework developers, therefore having a particular research interest in the testing the boundaries and comparing specifically this framework. We used the Hmsc-HPC extension to accelerate the time-consuming MCMC model fitting through its placement to GPU device [15].

We modified the data feeding pipeline developed for the deep SDM training to compile a single species presence-absence matrix Y and single covariate dataset X, which are required by the HMSC. We extracted the mean value for each channel of Sentinel images, and the mean values for each channel $\times$ season pair from the Landsat+Climate cube, and combined these with other included covariates, resulting in 68 predictors. We dropped the species that were observed less than 15 times in the PA train data.

We tested HMSC models both without and with spatial random effects. For the latter we reduced the computational burden by using spatial random surfaces, approximated with piecewise-constant

Gaussian Process based on k-mean centroids of the PA train survey locations. We conducted the analysis with 10-40 latent factors and 100-400 centroids.

3.3. DNN-1 + HMSC

In our focal approach we combined the methodological components from deep SDM and HMSC model. Namely, we selected the trained deep SDM variant, which was performing best on the public leaderboard (PLB), and removed the last FC layer from it. We compiled the single covariate dataset $\hat{X}$ consisting of 1128 features extracted by the truncated DNN layers of the selected deep SDM. We proceeded with HMSC modeling using Y and $\hat{X}$ similar to the pure HMSC approach. Unfortunately, the Hmsc-HPC execution of the model variants with spatial random effects failed with out-of-memory (OOM) error for that many covariates. To mitigate this issue we used a principal component (PC) dimension reduction either down to 228 PCs that encapsulated 90% of variance in $\hat{X}$ or to 408 PCs that captured 95% of variance.

3.4. Predicted species list

All the modeling approaches described above produce in predictions of the probabilities that each species in the community is present at the set of test surveys. However, the GeoLifeCLEF 2025 evaluation criteria is based on F_1 score, which requires predicting species lists. Therefore, converting vector of species presence probabilities to the list of predicted species is a pivotal task. A simplistic solution, exemplified by the organizers in their example code is to predict a constant number of species, which presence probabilities are the highest. Alternatively, some highly-ranked participants of the GeoLifeCLEF competitions in previous years reported that they achieved superior performance by building a regression model for the length of predicted species list.

In our approach we decided to avoid introducing an extra predictive model, but to rely on the evaluated and predicted species presence probabilities. Thus, given a predicted vector of species-specific presence probabilities at a test site $[p_1, p_2, \cdots, p_{n_s}]$, we locate the optimal threshold p_* for this site so that each species j satisfying $p_j > p_*$ are included in the prediction list ($\hat{y}_j = 1$) and the rest species are excluded ($\hat{y}_j = 0$). We denote the predicted vector of species presences as $\hat{\mathbf{y}}) = [\hat{\mathbf{y}}_\mathbf{1}, \hat{\mathbf{y}}_\mathbf{2}, \cdots, \hat{\mathbf{y}}_{\mathbf{n_s}}]$. The threshold is selected so that the expected F_1 score is maximized under the predicted presence probabilities of species in this site.

$$p_* = \underset{p \in [0,1]}{\operatorname{argmax}} E_{y_j \sim \text{Bern}(p_j)} \left[F_1(\mathbf{y}, \hat{\mathbf{y}}) \right]$$

Such approach enables to produce the species lists of variable length that are justified by the predictive model's belief about estimated species presence probabilities. On the downside, this method is prone to the miscalibration of the probability estimates and ignores the potential species co-occurrence patterns. The expectation term probably does not have any simple analytical expression and was approximated with Monte-Carlo method.

3.5. PO data aggregation

Combining PO data with PA data is an ever-going challenge in statistical community ecology. While the PO data is the most commonly available type of ecological data nowadays, it is generally subject to multiple flaws that discourage its usage. The fact that most PO data does not bear any information on the sampling effort is a principal concern, which complicate the distinction between true missingness and missingness due to lack of sampling effort. Furthermore, many PO datasets exhibit high preferential sampling — both in terms of spatial distribution of sampling effort and in terms of what species could be recorded.

Yet, as the spatial coverage of PO data largely exceeds the coverage of much more structured PA train data (Figure 1), we made an attempt to aggregate it towards pseudo PA samples and use for model training. Specifically, we a) divided the study region with a square grid splitting it into 2244 cells, b) for

each grid cell we listed what WorldCover classes appear at the locations of PO data, c) for each grid cell $\times$ WorldCover class pair we aggregated the occurrences into a vector of species presence-absences, which we added as extra rows to the species matrix Y. We also calculated the mean coordinates and covariate or deep features values, which were added as extra rows of the X and $\hat{X}$ matrices. We added an indicator covariate, which coded whether the data row originated from the PA train data or from aggregated PO data, and recoded the logarithm of the total number of occurrences that were attributed to the grid cell $\times$ WorldCover class pair as a proxy of the sampling effort. This procedure resulted in approximately 13 thousands of presence-absence pseudo surveys being added to the training PA data. We reran our analytical pipelines of HMSC and integrated DNN-1 + HMSC using this combined data instead of the original PA train.

3.6. Ensemble

Ensembling multiple different predictions is a very common post-modeling strategy in data science competitions. Not surprisingly, many top competitors of GeoLifeCLEF from previous years reported that it boosted the performance of their predictive models. On the other hand, ensemble solutions generally decreases the interpretability, which greatly reduce their scientific value. Therefore, we put only very limited effort in an ensemble solution with a single final submission in the last minutes of the GeoLifeCLEF 2025 challenge. We used 15 predicted species lists from previous submissions that we designated as sufficiently different from each other. For each test survey we weighted the species from the considered solutions at this survey proportionally to their corresponding PLB scores and summed them up. Our ensemble solution consisted of the species with top summed weights, where the number of reported species was taken as the PLB-weighted average of the length of individual solutions.

4. Results

Altogether our team made 164 technically successful submissions to the Kaggle web platform hosting GeoLifeCLEF 2025 competition. However, due to the Kaggle's design of limited submissions per day, a good portion of these were merely minor variations of the previous prominent solutions. We used their PLB score to obtain a slightly better comprehension of the test set, as otherwise these submission attempts would be simply lost. We summarize the results of notable distinct model variants in the Table1, generally presenting only few top results per model class.

Due to the chronological order in which we conducted our method developments, our submissions with deep SDM framework exhibit huge variation. Our mature variants with this method achieved private LB scores of around 0.21. Notably, both our early stopping criteria and public LB score were not sufficient to robustly identify the model variants and fits that performed best on the private LB. Thus, the model with best private LB score actually fared mediocre on the public LB.

Rather surprisingly, our solutions with non-deep HMSC approach scored relatively well in the range of 0.186 to 0.198 on the private LB. The non-spatial HMSC were consistently inferior compared to the spatial counterpart, but the increase of the spatial random field approximation quality beyond the simplest 100-centroid approximation did not yield any predictive improvement. Also HMSC model slightly benefited from incorporation of pseudo PA training data that was aggregated from PO occurrences. According to the published results, our best solutions from this class would secure a Top-6 place.

Our solutions of DNN-1 + HMSC class achieved the best performance among the non-ensemble models, though the exact score varied from 0.209 to 0.217, depending on the chosen HMSC features. Similar to non-deep HMSC case, inclusion of spatial random effects improved the predictive performance, but this time the margin was smaller. The non-spatial models with PC dimension reduction tended to produce better results than their counterparts that were using the original extracted deep features, though both results were very close to the deep SDM that was used as deep feature extractor. In contrast to non-deep HMSC, extending the training dataset with pseudo PA data decreased both the public and private LB scores.

Table 1
Public and private score results for selected subset of model classes and variants that we submitted during the GeoLifeCLEF 2025 competition. Our ensemble solution (bottom line) achieved the best result among all our submissions both on the public and private LB. Apart of that, the best and second best result of individual models are marked with bold and underscore font respectively.

	model description	key hyperparameters	**public**	**private**
PA only	Multimodal DNN, used in DNN-1 + HMSC	ResNet18(1000), MLP(1128), 75 epoch	0.24136	0.21146
	Multimodal DNN	ResNet18(1000), MLP(1128), 70 epoch	0.24120	0.21258
	Multimodal DNN	ResNet14(500), MLP(1128), 75 epoch	0.24002	0.20862
	Multimodal DNN	ResNet14(500), MLP(628), 67 epoch	0.23864	0.21294
	HMSC without spatial random effects	$n_s = 2519$, $n_c = 69$	0.20526	0.18562
	HMSC with spatial random effects	$n_s = 2519$, $n_c = 69$, $n_f = 20$, $n_p = 100$	0.21734	0.19686
	DNN-1 + HMSC without spatial random effects	$n_s = 2519$, $n_c = 1129$	0.23740	0.21036
	DNN-1 + PC(408) + HMSC without spatial r.e.	$n_s = 2519$, $n_c = 409$	0.24183	0.21461
	DNN-1 + PC(408) + HMSC with spatial r.e.	$n_s = 2519$, $n_c = 409$, $n_f = 10$, $n_p = 100$	<u>0.24168</u>	0.21481
	DNN-1 + PC(228) + HMSC with spatial r.e.	$n_s = 2519$, $n_c = 229$, $n_f = 20$, $n_p = 100$	**0.24343**	**0.21727**
PA+PO	HMSC without spatial random effects	$n_s = 2269$, $n_c = 71$	0.21071	0.19112
	HMSC with spatial random effects	$n_s = 2269$, $n_c = 71$, $n_f = 20$, $n_p = 100$	0.22009	0.19849
	DNN-1 + HMSC without spatial random effects	$n_s = 2269$, $n_c = 1131$	0.23541	0.20883
	DNN-1 + PC(408) + HMSC without spatial r.e.	$n_s = 2269$, $n_c = 411$	0.23786	0.21221
	DNN-1 + PC(408) + HMSC with spatial r.e.	$n_s = 2269$, $n_c = 411$, $n_f = 20$, $n_p = 100$	0.23997	0.21402
	DNN-1 + PC(231) + HMSC without spatial r.e.	$n_s = 2269$, $n_c = 231$	0.24111	0.21487
	DNN-1 + PC(231) + HMSC with spatial r.e.	$n_s = 2269$, $n_c = 231$, $n_f = 20$, $n_p = 200$	0.24158	<u>0.21588</u>
Ensemble of multiple predicted species lists, weighted by PLB-score			0.24784	0.22153

Our final ensemble solution resulted in a significant predictive performance boost, both in terms of the public (0.248) and private (0.222) LB scores.

5. Discussion

Our team entered the GeoLifeCLEF 2025 competition with several clear goals in mind. The first objective was to improve our knowledge and skills in how deep learning can be applied to the SDM tasks, and specifically how to incorporate large pre-trained foundation models into the multimodal DNN architectures. Overall, we evaluate our experience and results in achieving this task positively, even though we cannot claim that we properly mitigated the negative overtraining effects or explored the models' and training strategies' hyperparameter space comprehensively. We identified several prospective directions for further educating ourselves in this context, such as deepening our practical skills in rigorous tracking of numerous training instances of DNN variants and automated hyperparameter tuning.

Our second and primary goal was to prototype an integrative framework that would combine the benefits of both deep SDM and JSDM. Due to the tight competition schedule, we made a strategic decision to avoid the development of an all-rounded end-to-end training approach, and opted for a simpler two-stage model fitting scheme relying on existing software solutions. Given that our best non-ensemble model results were achieved with this approach, we obtained solid evidence that such an approach is justified, operational and possible to implement with existing software.

Yet, we would like to point out that generalization of our specific GeoLifeCLEF 2025 experience to a typical applied spatial community analysis and prediction task should be conducted with caution.

First of all, we largely relied on the public LB scoring for model assessment and decision-making concerning further development. This approach is clearly viable only for the particular design of Kaggle-hosted competition but not suited for general community analysis setups. Nevertheless, in our opinion, the dramatic mismatch in spatial coverage between training and test data in GeoLifeCLEF 2025

very likely causes non-negligible degree of variation in sampling methodologies, making performance assessment using classical validation schemes inferior, and even impossible for those countries that are very poorly represented in the training data.

Another key applied limitation of our study is that is that we did not assess the relative contribution of individual predictor variables to the final model predictions. This decision was again primarily dictated by the evaluation design of the GeoLifeCLEF 2025 competition, which focused solely on predictive performance on a hidden test set. As a result, our modeling and submission pipeline prioritized maximizing predictive accuracy rather than interpretability or variable importance analysis.

Finally, the amount of compromises that we had to make in our two-step DNN-1 + HMSC approach clearly highlights its limitations. First, we relied on binary cross entropy loss in our DNN stage that statistically corresponds to the assumption of no co-associations between species, and may limit the quality of extracted deep features that we used in the HMSC stage. Next, we have to admit that the extracted deep features were pushing the Hmsc-HPC package to its current computational limits and even slightly beyond them, largely eliminating the benefits of its rigid MCMC-based Bayesian inference. Furthermore, out of the multiple beneficial JSDM features of HMSC, we managed to exploit only spatial random effects. The key reason is that anonymized species identities in the GeoLifeCLEF 2025 dataset do not allow inclusion of species traits or phylogenies, which would potentially improve the inference for the rare species through borrowing statistical signal from their common relatives.

Nevertheless, we are strongly convinced that our experience and results represent a lower bound of what a combined deep + joint SDM framework is potentially capable of. The competitive predictive performance of our prototyped DNN-1 + HMSC solution in GeoLifeCLEF 2025 serves as a solid proof of concept, but uncovering its full potential requires a considerable amount of further research, development and testing that we leave for subsequent studies.

Acknowledgments

The authors thank CSC—IT Center for Science of Finland for providing access to HPC infrastructure and high-end GPU devices. Prof. Otso Ovaskainen significantly contributed to the design of presence-only data integration, motivating us to overcome the challenges of adding this challenging modality to our analytical pipelines. We also greatly appreciate discussions with Prof. Graham Tailor on modern best practices in deep learning. We would like to thank the whole GeoLifeCLEF 2025 organising team and particularly Dr. Lukáš Picek for providing valuable support to competition participants on the Kaggle platform.

Declaration on Generative AI

During the preparation of this work, the authors used Gemini 2.5 in order to: paraphrase and reword, grammar and spelling check. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] M. Begon, J. L. Harper, C. R. Townsend, Ecology: individuals, populations and communities., Blackwell Science Ltd, Oxford, 1996.

[2] M. A. Leibold, M. Holyoak, N. Mouquet, P. Amarasekare, J. M. Chase, M. F. Hoopes, R. D. Holt, J. B. Shurin, R. Law, D. Tilman, M. Loreau, A. Gonzalez, The metacommunity concept: a framework for multi-scale community ecology, Ecology Letters 7 (2004) 601–613. doi:https://doi.org/10.1111/j.1461-0248.2004.00608.x.

[3] J. Elith, J. R. Leathwick, Species distribution models: Ecological explanation and prediction across space and time, Annual Review of Ecology, Evolution, and Systematics 40 (2009) 677–697. doi:https://doi.org/10.1146/annurev.ecolsys.110308.120159.

[4] A. Norberg, N. Abrego, F. G. Blanchet, F. R. Adler, B. J. Anderson, J. Anttila, M. B. Araújo, T. Dallas, D. Dunson, J. Elith, S. D. Foster, R. Fox, J. Franklin, W. Godsoe, A. Guisan, B. O'Hara, N. A. Hill, R. D. Holt, F. K. C. Hui, M. Husby, J. A. Kålås, A. Lehikoinen, M. Luoto, H. K. Mod, G. Newell, I. Renner, T. Roslin, J. Soininen, W. Thuiller, J. Vanhatalo, D. Warton, M. White, N. E. Zimmermann, D. Gravel, O. Ovaskainen, A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels, Ecological Monographs 89 (2019) e01370. doi:https://doi.org/10.1002/ecm.1370.

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

[6] K. de Koning, J. Broekhuijsen, I. Kühn, O. Ovaskainen, F. Taubert, D. Endresen, D. Schigel, V. Grimm, Digital twins: dynamic model-data fusion for ecology, Trends in Ecology & Evolution 38 (2023) 916–926. doi:https://doi.org/10.1016/j.tree.2023.04.010.

[7] L. J. Pollock, R. Tingley, W. K. Morris, N. Golding, R. B. O'Hara, K. M. Parris, P. A. Vesk, M. A. McCarthy, Understanding co-occurrence by modelling species simultaneously with a joint species distribution model (jsdm), Methods in Ecology and Evolution 5 (2014) 397–406. doi:https://doi.org/10.1111/2041-210X.12180.

[8] O. Ovaskainen, G. Tikhonov, A. Norberg, F. Guillaume Blanchet, L. Duan, D. Dunson, T. Roslin, N. Abrego, How to make more out of community data? a conceptual framework and its implementation as models and software, Ecology letters 20 (2017) 561–576.

[9] J. Niku, F. K. C. Hui, S. Taskinen, D. I. Warton, gllvm: Fast analysis of multivariate abundance data with generalized linear latent variable models in r, Methods in Ecology and Evolution 10 (2019) 2173–2182. doi:https://doi.org/10.1111/2041-210X.13303.

[10] G. Tikhonov, Ø. H. Opedal, N. Abrego, A. Lehikoinen, M. M. de Jonge, J. Oksanen, O. Ovaskainen, Joint species distribution modelling with the r-package hmsc, Methods in ecology and evolution 11 (2020) 442–447.

[11] M. Pichler, 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 (2021) 2159–2173. doi:https://doi.org/10.1111/2041-210X.13687.

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

[13] J. Niku, F. K. C. Hui, S. Taskinen, D. I. Warton, Analyzing environmental-trait interactions in ecological communities with fourth-corner latent variable models, Environmetrics 32 (2021) e2683. doi:https://doi.org/10.1002/env.2683.

[14] G. Tikhonov, L. Duan, N. Abrego, G. Newell, M. White, D. Dunson, O. Ovaskainen, Computationally efficient joint species distribution modeling of big spatial data, Ecology 101 (2020) e02929.

[15] A. U. Rahman, G. Tikhonov, J. Oksanen, T. Rossi, O. Ovaskainen, Accelerating joint species distribution modelling with hmsc-hpc by gpu porting, PLOS Computational Biology 20 (2024) e1011914.

[16] T. Larcher, L. Picek, B. Deneu, T. Lorieul, M. Servajean, A. Joly, Malpolon: A framework for deep species distribution modeling, 2024. URL: https://arxiv.org/abs/2409.18102. arXiv:2409.18102.

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

[18] D. Szwarcman, S. Roy, P. Fraccaro, Þorsteinn Elí Gíslason, B. Blumenstiel, R. Ghosal, P. H. de Oliveira, J. L. de Sousa Almeida, R. Sedona, Y. Kang, S. Chakraborty, S. Wang, C. Gomes, A. Kumar, M. Truong, D. Godwin, H. Lee, C.-Y. Hsu, A. A. Asanjan, B. Mujeci, D. Shidham, T. Keenan, P. Arevalo, W. Li, H. Alemohammad, P. Olofsson, C. Hain, R. Kennedy, B. Zadrozny, D. Bell, G. Cavallaro, C. Watson, M. Maskey, R. Ramachandran, J. B. Moreno, Prithvi-eo-2.0: A versatile multi-temporal foundation model for earth observation applications, 2025. URL: https://arxiv.org/abs/2412.02732. arXiv:2412.02732.

[19] L. Picek, C. Leblanc, T. Larcher, M. Servajean, P. Bonnet, A. Joly, Overview of GeoLifeCLEF 2025: Plant species presence prediction with environmental and high-resolution remote sensing data, in: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, 2025.
[20] R. Valavi, G. Guillera-Arroita, J. J. Lahoz-Monfort, J. Elith, Predictive performance of presence-only species distribution models: a benchmark study with reproducible code, Ecological Monographs 92 (2022) e01486. doi:https://doi.org/10.1002/ecm.1486.