Enhancing Presence-Absence Prediction Models Using Presence-Only Data

Tim Chopard[1,*,†], Darren Rawlings[2,*,†]

[1] *University of Leeds, Woodhouse, Leeds, LS2 9JT, UK*
[2] *University of Groningen, Broerstraat 5, 9712 CP Groningen, The Netherlands*

Abstract

Accurate Species Distribution Modelling (SDM) is essential for biodiversity conservation, however the limited and spatially biased nature of Presence-Absence (PA) data poses a challenge. In contrast, Presence-Only (PO) datasets are abundant but lack explicit absence records. This paper examines a two step deep learning approach to combining both PO and PA data to generate an SDM. In the first step, the model was trained on a larger PO dataset, and in the second the model was then tuned on a smaller PA dataset. Results indicate that pre-training with PO data improved the performance by 7% when subsequently fine-tuned with PA data, as measured by the samples-averaged F1-score. This approach demonstrates the potential of combining diverse data types to create more reliable species distribution models for plant biodiversity conservation.

Keywords

Species Distribution Modelling, Presence-Only Data, Presence-Absence Data, Environmental Predictors, Climatic Data, Biodiversity Conservation

1. Introduction

Understanding the complex patterns of plant species distribution can help in managing and protecting species, that are rare [1], climatically sensitive [2], or economically important [3]. Species Distribution Models (SDMs) are used to predict likely locations for these plants. These models function by correlating known species occurrences with environmental factors such as climate, terrain, and land cover [4, 5]. Through understanding these relationships, it is possible to predict the full occurrence rates of a species, even in areas that have not been surveyed.

Deep learning techniques, especially Convolutional Neural Networks (CNNs), have become popular for this purpose. CNNs are particularly good at identifying complex patterns in large, high-dimensional environmental data.

These models typically rely on two kinds of data. The first is Presence-Only (PO) data, which is widely available from sources such as citizen science projects and museum collections. However, this data only tells us where a species has been found, not where it's absent, and it can be geographically skewed [6, 7]. The second is Presence-Absence (PA) data, gathered from systematic surveys. While PA data provides reliable information on both where a species is present and where it is truly absent, it is much less common and covers smaller areas. The geographical distributions of the PA and PO data used in this paper are shown in Figure 1.

Through the combination of both PA and PO data more powerful models may be developed, namely Integrated Distribution Models. Development of these models is, however, both technically and computationally complex, and often requires notable amounts of time and resources [7, 8].

CNNs are a natural fit for SDMs because they excel at analyzing high-resolution data from remote sensing. Architectures such as ResNet have been used effectively in modeling species distribution, and served as key benchmarks in previous GeoLifeCLEF labs [9, 10].

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

[*] Kaggle team BernGron https://github.com/startung/geolife2025

[†] These authors contributed equally.

timchopard@cloudberries.io (T. Chopard); d.rawlings@student.rug.nl (D. Rawlings)
https://cloudberries.io (T. Chopard); https://github.com/startung (D. Rawlings)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

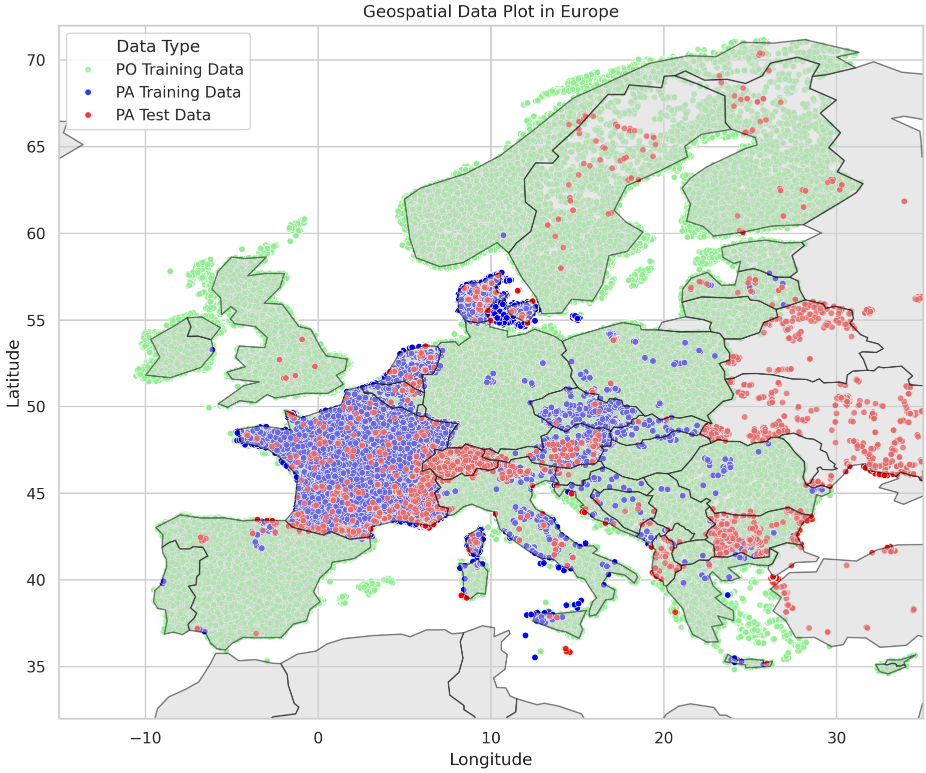

Figure 1: An overlay of the three different datasets on a map of Europe, showing the differences in geographical coverage of each dataset. Each survey is represented by a single point overlaid on a map of Europe.

In previous GeoLifeCLEF labs multi-modal models, which include more than one predictor per model, have been shown to perform well on similar tasks [11, 12]. This paper, however, focuses on the utilization PO and PA data, rather than producing the absolute highest performing solution. In this research, single modal models are used and tested against each other.

1.1. Integrating PO Pretraining with PA Fine-Tuning

To take advantage of both Presence-Only (PO) and Presence-Absence (PA) data, a two-step deep learning approach was used, with the goal of improving on previous models that exclusively used PA data [12]. First, a ResNet18 model was trained on a large subset of the PO data, allowing it to learn general features about the environment. In the second step, the model is fine-tuned using the more complete PA data, ensuring more accurate identification of places in which a species is truly absent. The goal of this transfer learning strategy is to create a more reliable model that benefits from the larger availability of PO data as well as the accuracy of PA data.

This approach was tested with the GeoLifeCLEF 2025 dataset [13], which provides over five million PO records and ninety thousand PA survey sites in Europe, along with environmental data such as satellite imagery and climate trends. The ResNet18 architecture was chosen because of its previous success in similar tasks.

Ultimately, this research investigates whether pre-training a model on PO data can improve predictions when it is later trained on PA data. This paper demonstrates that this combined method will produce more accurate results than using PA data by itself.

2. Data

The data for this study [13] was sourced from the GeoLifeCLEF 2025 competition [14]. These datasets comprised both tabular and image data.

2.1. Summary

The provided data contained two main types of species information: approximately 5 million Presence-Only (PO) sightings and 90,000 Presence-Absence (PA) survey records, all linked to specific GPS coordinates and survey ID values.

Each record is paired with a comprehensive set of environmental variables, including:

- Satellite imagery (from Sentinel-2 and Landsat)
- Climate data (CHELSA time series and bioclimatic variables)
- Land cover maps
- Human footprint indexes
- Soil data

The dataset is organized to frame the task as a multi-label classification problem, where the goal is to predict the presence of various species. This research specifically drew upon three main data sources within this collection. These three data sources were selected as they were provided as PyTorch tensors in a structure that could easily be input into a CNN such as ResNet18 without additional manipulation data:

- **Sentinel-2 (sen):** Pre-processed raster files scaled to the European continent, representing Sentinel-2 Level-2A observations. Each TIFF file corresponds to a unique observation location (surveyId).
- **Landsat (lan):** Over 20 years of Landsat satellite imagery extracted from Ecodatacube, aggregated into CSV files and data cubes representing mean spectral band values for three months preceding each observation date. Cubes are structured as (n_bands, n_quarters, n_years) where n_bands = 6, n_quarters = 4, and n_years = 21.
- **Bioclimatic Cubes (bio):** Four monthly CHELSA climatic rasters (precipitation, maximum temperature, minimum temperature, and mean temperature) with a resolution of 30 arc seconds, spanning January 2000 to June 2019. Cubes are structured as (n_year, n_month, n_bio) where n_year = 19, n_month = 12, and n_bio = 4.

2.2. Geospatial Distribution

Statistical analyses were performed to determine whether the PO and PA training and test datasets exhibit similar geographic spreads. This was done to ensure the test data's spatial distribution is an accurate representation of the training data.

The test compared the spatial density of GPS coordinates between the Presence-Absence training data, the Presence-Only training data, and the Presence-Absence test data. The Jensen-Shannon Divergence (JSD) [15] was employed, which is a method to measure how different two probability distributions are.

The following hypotheses were tested:

Null Hypothesis: The training and test datasets originate from the same geographic area, and observed spatial differences are attributable to random chance. Alternative Hypothesis: The datasets come from different geographic areas, meaning there is a significant spatial mismatch between them.

To do this, a Gaussian Kernel Density Estimate (KDE) was created for the coordinates of each dataset to visualize its spatial distribution. The JSD score was then calculated to quantify the difference between these distributions. The resulting KDE heatmaps for the PA and PO datasets are shown in Figure 2.

Second, to evaluate the statistical significance of the observed JSD, a permutation test was then conducted under the null hypothesis that the training and test samples originate from the same underlying

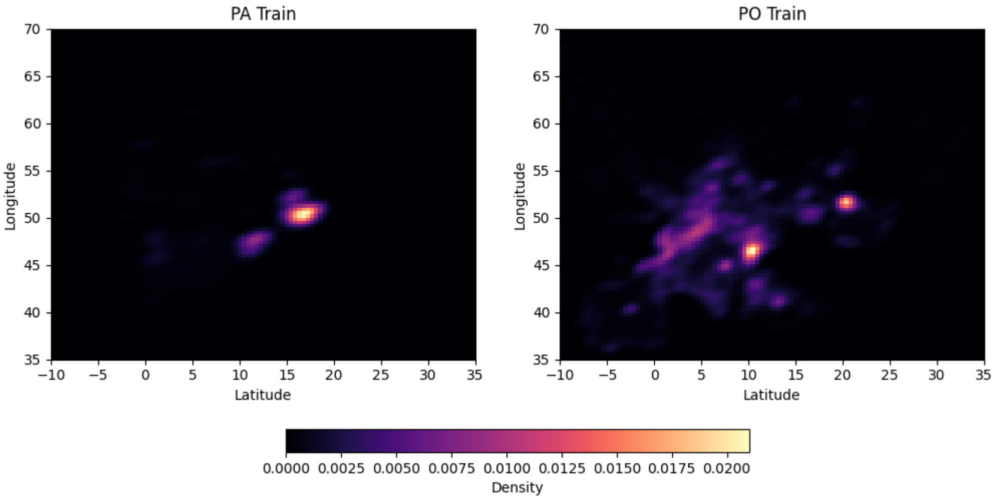

Figure 2: KDE plots of train (left) and test (right) spatial distributions. Visualizing the geographic difference of these datasets.

spatial distribution. The combined dataset was randomly permuted, and samples were reassigned into surrogate training and test groups matching the original group sizes. For each permutation, KDEs and the corresponding JSD were recomputed, generating a null distribution of divergence values.

Table 1
Jensen-Shannon Divergence values comparing the spatial distributions of different dataset combinations: Presence-Only data, Presence-Absence training data, and Presence-Absence test data. Lower values indicate greater similarity in spatial distribution between the compared datasets.

Pairing	Jensen-Shannon Divergence
PA_{train}, PA_{test}	0.5939
PA_{train}, PO	0.5515
PA_{test}, PO	0.5225

Table 1 presents values for pairwise comparisons between three datasets. All JSD values, ranging between 0.5 and 0.6, indicate moderate to substantial differences in the spatial distributions across the dataset pairs. This suggests that while the datasets share some spatial overlap, their geographic patterns exhibit meaningful divergence.

Finally, p-values were calculated as the proportion of permuted JSDs exceeding or equal to the observed JSD, thereby providing a measure of the likelihood that the observed spatial divergence occurred by chance.

In this analysis, the p-value was 0, indicating that none of the 1,000 permutations produced a divergence as large as or larger than the observed JSD. This result provides strong evidence to reject the null hypothesis at a significance level of 0.05, suggesting that the test dataset's spatial distribution is statistically distinct and not representative of the training dataset.

2.3. Species Occurrence

Value counts for each individual species were calculated for the data in order to remove low-occurrence outlier species. These counts were performed on the PA training data, as this provides a comprehensive overview of each survey site, mitigating the possibility of certain harder to identify species being overlooked, which might occur in the PO data.

The top 1,000 species were selected for use. These represent 94.12% of the PA training data and include species with occurrences of 130 or greater within this dataset.

3. Method

A CNN based on the ResNet18 architecture was trained to predict species presence from satellite imagery. The model was designed to classify species presence or absence at specified geographic locations, utilizing image tiles centered on survey coordinates as input. To optimize spatial generalization and reduce overfitting, a data strategy that combined different types of occurrence data while accounting for their respective strengths and limitations was implemented.

3.1. Incorporating Presence-Only Data to Supplement Sparse Presence-Absence Observations

Presence-Absence data were used as the primary target for model training and evaluation, as they provided both positive and negative labels necessary for supervised learning. However, the available PA dataset was relatively limited in size and spatial coverage. Many regions were underrepresented, including regions covered in the test data. To address this, the PA were supplemented with a larger and more spatially uniform Presence-Only (PO) dataset, which, although lacking explicit absence information and label balance, offered broader geographic coverage.

The PO data were used to augment the spatial diversity of the training inputs, exposing the model to a wider range of environmental and landscape contexts associated with species presence. While PO data could not be used for direct training on absence, they contributed to pretraining and representation learning stages, improving the model's ability to extract ecologically meaningful spatial features. This combined data approach allowed the model to benefit from both the accuracy of PA labels and the spatial representativeness of the PO observations.

3.2. Pipeline

This study employs a multi-label classification approach to predict species presence based on the environmental predictors described above. ResNet18 [16], a commonly used convolutional neural network architecture, was selected as the base model for all experiments. This model was chosen for its efficiency and the ability to train it on consumer hardware. The code developed for this experiment was adapted from the baselines provided as part of the Kaggle competition [13]. The following steps were undertaken:

1. **Data Preprocessing:** The training data were filtered to include only the 1,000 most common species from the PA training data. This filtering was applied consistently to both the PO and PA datasets to ensure a consistent feature space.
2. **Model Creation:** Three instances of the ResNet18 model were created, each utilizing a different input data source (Sentinel-2, Bioclimatic Cubes, Landsat). These models were adapted to suit the input format of the data sources.
3. **Pre-training Phase:** The experimental models were pre-trained for 10 epochs on the Presence-Only (PO) dataset. This initial training phase aimed to enable the models to learn generalizable features from the broader PO distribution and to address the disparity observed between the PA training and test datasets. This approach was selected for simplicity and was constrained by the limited resources available.
4. **Fine-tuning Phase:** Following pre-training, the models were fine-tuned for an additional 10 epochs on the PA training data. This fine-tuning stage adapts the pre-trained weights to the specific task of predicting species presence in the PA survey records. A set of three baseline models were also trained for 10 epochs on only the PA training data.

5. **Prediction and Evaluation:** The 25 most likely species to appear at each site in the test data were selected, to ensure consistency across models and with the baseline. The performance of the models was evaluated using the samples-averaged F1-score, which measures the overlap between the predicted and actual sets of species present at each location.

3.3. Models

All models were trained using an RTX 3090 GPU with 24GB VRAM. This use of consumer grade hardware influenced the choice of model architecture for smaller networks.

For this research, it was necessary to make some adjustments to the ResNet18 model.

Model for processing the Sentinel-2 data:

- Used pre-trained weights IMAGENET1K_V1
- The first convolutional layer was modified from 3 to 4 channels to accommodate the Infrared spectrum imagery.

Model for processing the Bioclimatic Cubes data:

- Used randomized initial weights
- Input channels were set to [4, 19, 12]
- The first convolutional layer was modified from 3 to 4 channels to accommodate cube dimensions.
- The final classification layer was changed to two linear layers: fc1 (in: 1,000 parameters, out: 2,056 parameters), and fc2 (in: 2,056 parameters, out: 1,000 parameters) in an attempt to improve classification.

Model for processing the Landsat data:

- Used randomized initial weights
- Input channels were set to [6, 4, 21]
- The first convolutional layer was modified from 3 to 6 channels to accommodate cube dimensions.
- The final classification layer was changed to two linear layers: fc1 (in: 1,000 parameters, out: 2,056 parameters), and fc2 (in: 2,056 parameters, out: 1,000 parameters) in an attempt to improve classification.

3.4. Metric

The performance was assessed using the samples-averaged F_1-score. This metric quantifies the degree of overlap between the species predicted to be present at each location and the actual species observed (ground truth). Submissions were then created consisting of a list of predicted species IDs for each test survey. These were compared against the known set of species present at that location and time, which was stored on the Kaggle platform. The F_1-score was calculated to provide an evaluation of the model's predictive accuracy.

$$F_1 = \frac{1}{N} \sum_{i=1}^{N} \frac{TP_i}{TP_i + (FP_i + FN_i)/2} \tag{1}$$

Where:

$$\begin{aligned} TP_i &= \text{ Number of predicted species truly present} \\ FP_i &= \text{ Number of predicted species that are absent} \\ FN_i &= \text{ Number of present species not predicted} \end{aligned}$$

Table 2
Hyperparameters used in the ResNet18 models trained on the Bioclimatic, Sentinel and Landsat data.

Hyperparameter	Value
learning rate Bioclimatic, Landsat	0.0006
learning rate Sentinel	0.0004
batch size	128
Sentinel input normalization (R, G, B, IR) mean	(0.485, 0.456, 0.406, 0.5)
Sentinel input normalization (R, G, B, IR) std	(0.229, 0.224, 0.225, 0.225)
optimizer	AdamW
scheduler	CosineAnnealingLR

3.5. Hyperparameters

The hyperparameters used in training the models are shown in Table 2. These were adopted from previous research [12] on similar data, performed in GeoLifeCLEF 2024. The learning rates were increased from the original value of 0.00025 to accommodate the decision to reduce the initial epochs from 20 down to 10. The learning rate of the Sentinel model was kept lower than that of the Bioclimatic and Landsat models as pre-trained weights were used.

4. Results

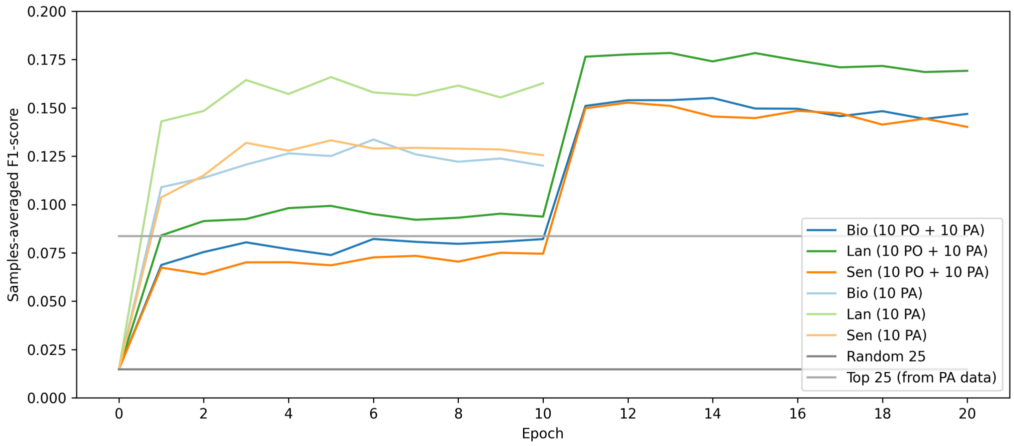

Figure 3: Comparison of models trained on a PO data (for 10 epochs) then trained on PA data (for 10 epochs), with those trained only on PA data (for 10 epochs). These are compared to two simple baselines. Random 25: where a random sample of 25 species were provided per survey. Top 25: where the 25 most common species were provided per survey.

As shown in Figure 3 all three models pre-trained on PO data outperformed their equivalent models trained on just PA training data with random weight initialization. Training solely on PO data (the first 10 epochs of the *10 PO + 10 PA* results in Figure 3) proved insufficient to produce a useful model, as the PO-trained Bioclimatic Cubes, Sentinel-2 models did not outperform a simple top-25 most frequent species prediction and the Landsat outperformed this baseline (0.08372) with a samples-averaged F_1-score of 0.09374. However, when subsequently fine-tuned with the PA training data, the three models demonstrated an approximate 7% increase in performance, as measured by the F_1-score.

Looking at the individual models, the Landsat model has the highest F_1-score in both PO+PA and PA-only training. Further, the Landsat PA-only model outperforms both the Bioclimatic and Sentinel PO+PA models. Between the two versions of the Landsat model, an improvement is also seen in the PO+PA model, which outperforms the PA-only model, resulting in an F_1-score of 0.1784.

Table 3
Training time in seconds per epoch for the different ResNet18 models on PO and PA data.

Model	PO Time Per Epoch	PA Time Per Epoch
Bioclimatic Cubes	394 s	24 s
Landsat	401 s	21 s
Sentinel-2	1,406 s	73 s

Table 3 presents the training times for the two different datasets PO and PA, segmented by the three different models, trained on the different data types. The PO data required approximately 18 times longer per epoch due to the significant difference in the number of records.

5. Conclusion

Pre-training our three baseline ResNet18 models with PO data resulted in improved performance for all three models. This improvement was observed despite the demonstrated differences between the PO and PA datasets. Notably, substantial geographical areas existed where the PA training and test data exhibited no overlap. The different data coverage for Switzerland, a country with very distinctive terrain, is shown in Figure 4. It is largely covered in the PA test data, but is only sparsely covered in the PA training data. It is, however, well covered in the PO data.

Due to the disparity in data size, training the model on PO data took longer per epoch than training on PA data. This cost in time would be a factor in choosing to use PO data, however, due to the efficiency of ResNet18, this did not push the hardware requirements beyond what can be delivered by modern consumer hardware. Furthermore, as seen in Figure 3 the models did not exhibit improvement in performance over extended training and the number of epochs could be reduced without sacrificing performance.

The PO data could, in theory, allow these trained models to generalize better to areas not adequately covered by PA training data, but this would require further analysis of the data, and access to the ground truth labels for the test set.

Due to resource limitations, limited hyperparameter tuning was performed. Further tuning is likely to result in improved performance, and when combined with more complex pre-training/fine-tuning methods, could result in a more accurate model.

Future research would be assisted by the addition of more data related to the plant species, such as family or genus information. This would allow for a deeper analysis of plant groups where the models underperform.

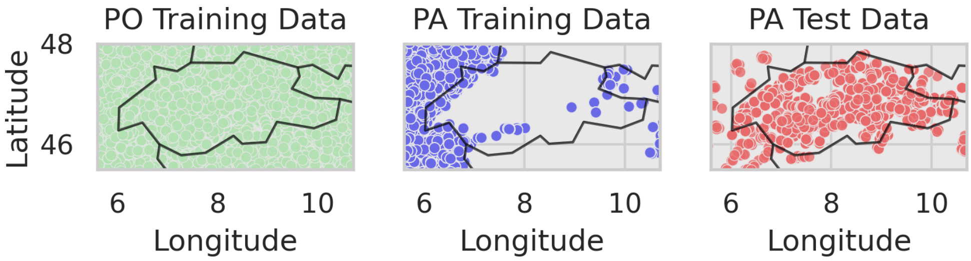

Figure 4: Variations in the survey data distributions centered on Switzerland. This shows both the PO and PA test data have a broad coverage of Switzerland, unlike the PA training data.

This research has shown that Presence-Only data can still improve predictive multi-label classification models, when supported by accompanying Presence-Absence data. Improved predictions could allow for better planning in protecting biodiversity.

Declaration on Generative AI

During the preparation of this work, the authors used Gemini 2.5 in order to: Grammar and Spelling Check; and Improve writing style. Further, the initial reviews used ChatGPT-4o for: Peer review simulation, the results of which were taken into account in later drafts. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] S. L. Pimm, C. N. Jenkins, R. Abell, T. M. Brooks, J. L. Gittleman, L. N. Joppa, P. H. Raven, C. M. Roberts, J. O. Sexton, The biodiversity of species and their rates of extinction, distribution, and protection, science 344 (2014) 1246752.

[2] C. Parmesan, M. E. Hanley, Plants and climate change: complexities and surprises, Annals of botany 116 (2015) 849–864.

[3] H. Seebens, F. Essl, W. Dawson, N. Fuentes, D. Moser, J. Pergl, P. Pyšek, M. van Kleunen, E. Weber, M. Winter, et al., Global trade will accelerate plant invasions in emerging economies under climate change, Global change biology 21 (2015) 4128–4140.

[4] J. Elith, J. R. Leathwick, The art of modelling range-shifting species, Methods in Ecology and Evolution 1 (2009) 330–342.

[5] A. Guisan, W. Thuiller, Predicting species distribution: offering more than simple habitat models, Ecology letters 8 (2005) 993–1009.

[6] S. J. Phillips, J. Elith, Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data, Ecological applications 19 (2009) 181–197.

[7] J. A. Royle, R. B. Chandler, C. Yackulic, J. D. Nichols, Integrating presence–absence and presence-only data to model species distribution, Methods in Ecology and Evolution 3 (2012) 349–359.

[8] W. Fithian, T. Hastie, et al., Bias correction in species distribution models: pooling survey and collection data for multiple species, Methods in Ecology and Evolution 6 (2015) 424–438.

[9] L. Picek, C. Botella, M. Servajean, B. Deneu, D. Marcos Gonzalez, R. Palard, T. Larcher, C. Leblanc, J. Estopinan, P. Bonnet, A. Joly, Overview of GeoLifeCLEF 2024: Species presence prediction based on occurrence data and high-resolution remote sensing images, in: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, 2024.

[10] R. Chen, et al., Transfer learning with cnns for predicting plant distributions in novel environments, Remote Sensing of Environment (2024). In press.

[11] C. Leblanc, A. Joly, T. Lorieul, M. Servajean, P. Bonnet, Species distribution modeling based on aerial images and environmental features with convolutional neural networks, CEUR-WS, 2022.

[12] D. Rawlings, T. Chopard, Exploring biodiversity: A multi-model approach to multi-label plant species prediction, in: 25th Working Notes of the Conference and Labs of the Evaluation Forum, CLEF 2024, CEUR Workshop Proceedings, 2024, pp. 2188–2200.

[13] L. Picek, C. Botella, M. Servajean, C. Leblanc, R. Palard, T. Larcher, B. Deneu, D. Marcos, P. Bonnet, A. Joly, Geoplant: Spatial plant species prediction dataset, in: A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, C. Zhang (Eds.), Advances in Neural Information Processing Systems, volume 37, Curran Associates, Inc., 2024, pp. 126653–126676. URL: https://proceedings.neurips.cc/paper_files/paper/2024/file/e4e7de47202bda8133dd3e8b46205cf2-Paper-Datasets_and_Benchmarks_Track.pdf.

[14] 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.

[15] B. Fuglede, F. Topsoe, Jensen-shannon divergence and hilbert space embedding, in: International symposium onInformation theory, 2004. ISIT 2004. Proceedings., IEEE, 2004, p. 31.

[16] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, 2015. URL: https://arxiv.org/abs/1512.03385. arXiv:1512.03385.