Sat-SINR: High-Resolution Species Distribution Models Through Satellite Imagery

Johannes Dollinger[1], Philipp Brun[2], Vivien Sainte Fare Garnot[1], Jan Dirk Wegner[1]

[1]Department of Mathematical Modeling and Machine Learning (DM^{3}L), University of Zurich, 8057 Zürich, Switzerland
[2]Swiss Federal Research Institute WSL, 8903 Birmensdorf, Switzerland

KEY WORDS: species distribution modeling, deep learning, sentinel-2, multi-modal fusion

ABSTRACT:

We propose a deep learning approach for high-resolution species distribution modelling (SDM) at large scale combining point-wise, crowd-sourced species observation data and environmental data with Sentinel-2 satellite imagery. What makes this task challenging is the great variety of controlling factors for species distribution, such as habitat conditions, human intervention, competition, disturbances, and evolutionary history. Experts either incorporate these factors into complex mechanistic models based on *presence-absence* data collected in field campaigns or train machine learning models to learn the relationship between environmental data and *presence-only* species occurrence. We extend the latter approach here and learn deep SDMs end-to-end based on point-wise, crowd-sourced *presence-only* data in combination with satellite imagery. Our method, dubbed Sat-SINR, jointly models the spatial distributions of 5.6k plant species across Europe and increases the spatial resolution by a factor of 100 compared to the current state of the art. We exhaustively test and ablate multiple variations of combining geo-referenced point data with satellite imagery and show that our deep learning-based SDM method consistently shows an improvement of up to 3 percentage points across three metrics. We make all code publicly available at `https://github.com/ecovision-uzh/sat-sinr`.

1. INTRODUCTION

With the increase of economic and political salience of environmental issues, the need has grown for a sound understanding of the state of the biosphere and its response to global change (Navarro et al., 2017; Dıaz et al., 2019). In particular, investigating the geographical distribution of species and its evolution is a key ecological question. Such knowledge is required, for example, to quantify the ecological capital and target conservation efforts (Rockström et al., 2023). Furthermore, successful modeling of the drivers shaping species distribution is necessary to anticipate future shifts under climate and habitat change.

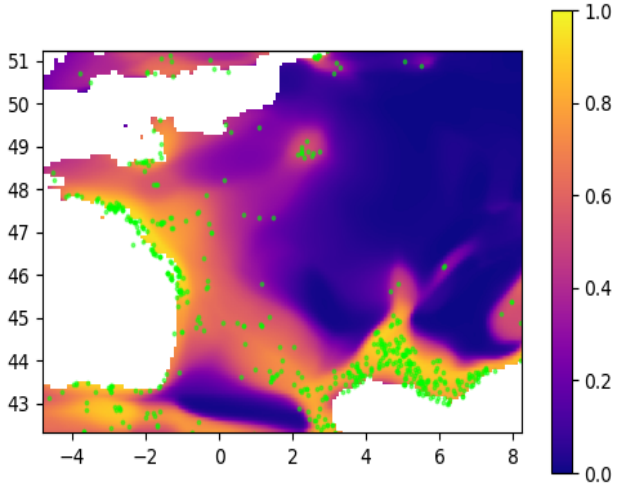
Species distribution modeling (SDM) is a task of quantitative ecology that aims at correctly predicting such distributions. Traditionally, maps of species distribution have been created by ecologists deriving theoretical models from their expert knowledge. These models are then fitted and evaluated on datasets of in-situ *presence-absence* (PA) surveys. PA surveys provide valuable information for SDM fitting as they indicate both the presence *and* absence of a given species in a given region. Yet, as they require a highly methodical survey process, they typically are carried out with limited spatial coverage. Over the past three decades, novel empirical approaches using noisy *presence-only* (PO) occurrences emerged. In this more difficult setting, the data only informs on the presence of a given species, but do not provide *negative* samples on the absence of the species. Though more challenging, these datasets such as the Global Biodiversity Information Facility[1] (GBIF) have gained popularity, as they enable a larger spatial coverage.

SDMs are fitted by relating occurrence information to co-occurring environmental conditions, using classical statistical methods or machine learning algorithms. More recently, deep-learning-based multi-species SDMs have been introduced to increase the representational capacity of the models. The authors of Spatial Implicit Neural Representations (SINR) (Cole et al., 2023), for instance, advocate for a model using solely longitude and latitude of an observation as input. They map patterns based only on confirmed occurrences instead of extrapolating to unobserved locations using local environmental information, although their work does also include experiments with bioclimatic inputs.

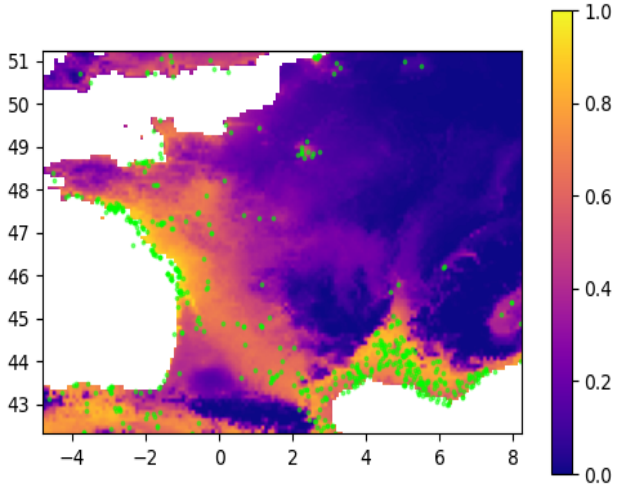
In contrast, we build upon SINR, but argue that high-resolution local information is essential to further develop SDMs. We address the problem of modeling the spatial distribution of 5.6k plant species across Europe. The growth of plants is strongly constrained by local factors such as soil, climate, human intervention, fauna, and terrain. Rasterized information on such factors is available, but generally limited to low spatial resolutions, that fail to capture sub-kilometer nuances. Furthermore, the rasters with large coverage are often based on historical data, physics-based models, or interpolation between measurements, introducing a variety of biases. Therefore, we use satellite imagery as an accurate, up-to-date, high-resolution modality, which has been shown to contain valuable information for SDMs (Deneu et al., 2022). This increases the resolution of the most high-resolved input data from 1 km of the bioclimatic variables (Fick and Hijmans, 2017) to 10 meters of Sentinel-2 (Drusch et al., 2012) satellite images. We show experimentally that fusing SINR with the additional satellite input is both beneficial to the performance on unseen PA data and the overall quality of the maps created from the SDMs predictions (Figure 1). In summary, our contributions are as follows:

- A framework for fusing satellite imagery into the SINR model.
- A quantitative comparison of multimodal fusion mechanisms.
- Quantifying the impact of various modality mixes on the resulting SDM.

[1] GBIF. Available from https://www.gbif.org. Accessed 24.01.24



(a) SINR SDM range map with location as input.



(b) Sat-SINR SDM range map with location, bioclimatic variables and Sentinel-2 images as input.

Figure 1. The resulting species distribution maps of four averaged models for *Quercus Ilex* (Evergreen Oak) in France. Compared to the location-based SINR (**Top**), Sat-SINR (**Bottom**) introduces a new level of local detail to the distribution maps. **X-axis**: longitude east, **y-axis**: latitude north, **colours**: predicted probability of 0 (blue) to 1 (yellow), **green dots**: occurrences of the evergreen oak in the training dataset. Predictions are sampled in a grid of 3 to 10 kilometers.

2. RELATED WORK

2.1 Species Distribution Models

Historically, the data used for species distribution modelling has been limited to the number of locations in a study area where ecologists would note down each observed species. While such data would reliably carry both *presence* and *absence* information, it would only cover small areas. A different approach emerged three decades ago, creating models trained on presence-only data. The advent of platforms such as iNaturalist[2] that collect citizen scientist observations further accelerated this research direction. Today, such collections provide over 50% of the data in GBIF.

Following the successes of deep learning on large datasets in fields like vision and language (Goodfellow et al., 2016), neural networks are increasingly being tested out as the next stage of progress in species distribution modeling (Beery et al., 2021). Deep learning uses large architectures with substantially higher numbers of parameters, capable of modeling intricate relationships between an input and an output distribution at the cost of explainability. As in related fields such as canopy height estimation (Lang et al., 2023; Jiang et al., 2023), cocoa plantation classification (Kalischek et al., 2023), deforestation detection (Karaman et al., 2023), and conifer cell analysis (Katzenmaier et al., 2023), deep learning approaches are showing encouraging results in species distribution modelling.

Due to the nature of deep learning, both positive and negative feedback is required for the model training. Strategies for addressing the lack of negative feedback in presence-only data rely on pseudo-absence data (Rew et al., 2021), rank-based loss functions (Brun et al., 2024), or assume full absence while up scaling the impact of presences (Aodha et al., 2019; Zbinden et al., 2024). SINR by (Cole et al., 2023) and further adapted in (Rußwurm et al., 2023) is a strong PO-trained model, using the location and, in auxiliary experiments, bioclimatic variables as inputs. This work extends SINR by additionally including satellite imagery, a data source that has recently shown promising results in species distribution modeling (Deneu et al., 2022; Botella et al., 2023a; Teng et al., 2023).

2.2 Multi-Modal Networks

Multi-modality concerns itself with the fusion of different data types into a single prediction in a neural network. Here, we focus on fusing a 1-dimensional vector with an image. In our setting, we aim at inferring an SDM based on a vector of location and bioclimatic variables, as well as a satellite image of the location. Fusion methods can typically be categorised into early, late, and middle fusion (D’mello and Kory, 2015). In early fusion, all input modalities are combined into a single tensor before being processed by the network (Lang et al., 2021; Teng et al., 2023). Late fusion derives an independent prediction or representation for each modality, only merging them at the end (Aodha et al., 2019; de Lutio et al., 2021; Sastry et al., 2023). SatCLIP (Klemmer et al., 2023) presents a special case of this, where the embeddings produced by SINR are aligned with the embeddings produced by a satellite embedder to serve as satellite image-free embeddings in a downstream task. A less common fusion approach is middle fusion that merges the information with multiple layers both above and below the merge (Damer et al., 2019) or along multiple layers across the whole network (Zhang et al., 2023). In this paper, we explore several approaches for merging satellite information into the predefined SINR architecture. In addition to early and late fusion, we implement a middle fusion scheme inspired by ControlNet (Zhang et al., 2023).

3. METHODS

3.1 Problem statement

We follow the problem statement of SINR: our dataset $\mathcal{D}$, indexed by $\mathcal{N}$, contains samples characterised by their geo-location L, a vector of bioclimatic variables (precipitation, temperature)

[2] iNaturalist. Available from https://www.inaturalist.org. Accessed 24.01.24

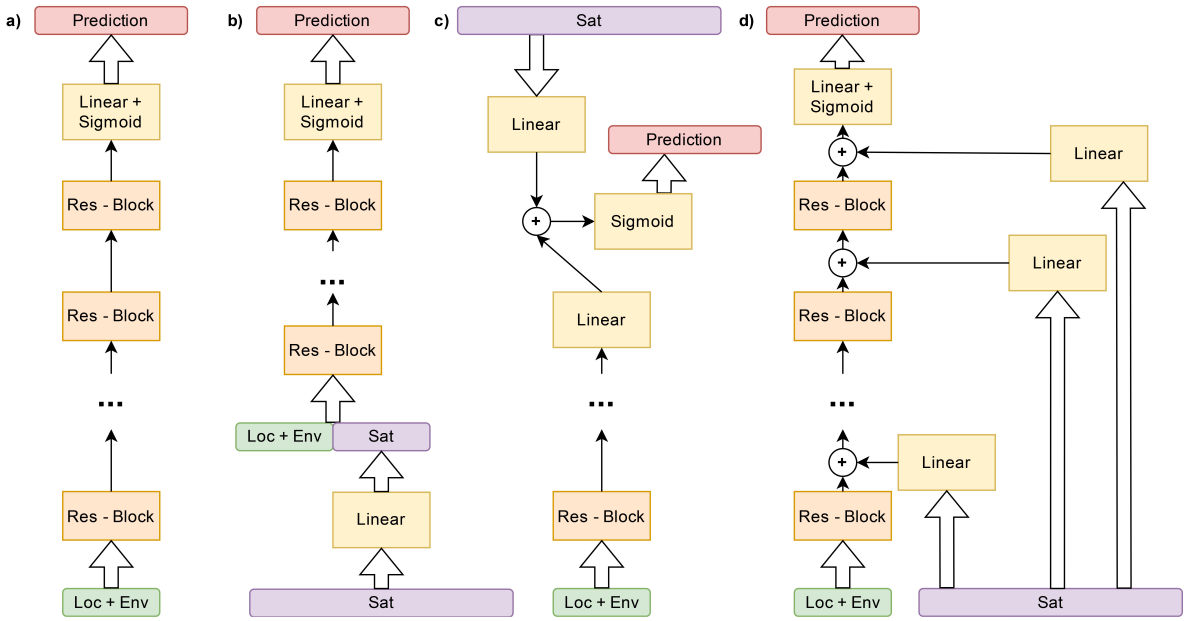


Figure 2. SINR and three proposed Sat-SINR fusion schemes for leveraging satellite data. a) **SINR:** The location L and environmental variables E are mapped to a multi-class prediction; b) **Early Fusion:** Satellite embedding c is reduced in dimensionality to fit with L and E and fed into the model at the beginning; c) **Late Fusion**: An independent embedding is created from c and added to the SINR embedding before sigmoid; d) **Middle Fusion**: c is scaled to the hidden size and added after each layer.

E, and the species identifier y, covering a limited set of S species. Since we extend this task to using remote sensing observations, we also include remote sensing images in the dataset. For each sample location L_i, we collect a satellite image I_i centered at the given latitude and longitude, and of shape $H \times W$:

$$\mathcal{D} = \{((L_i, E_i, I_i), y_i), i \in \mathcal{N}\} \quad (1)$$

The task at hand is then to predict the presence of the correct species y_i given the input context (L_i, E_i, I_i). Importantly, due to the high acquisition cost of presence-absence data, we strive to propose a method leveraging presence-only data. To this end, we choose to train on presence-only data and use our presence-absence data for evaluation only. Therefore, at train time the model needs to predict only the correct species observed in a given location. At test time, multiple species can be present in the same location, hence y_i is a one-hot vector in the training data, and a binary vector in the test data. Training a joint model for many species allows for shared representations that offer better performance than binary single-species models.

3.2 SINR

Before extending the task to the combined use of remote sensing data, we detail the main elements of the original SINR approach of Aodha et al. (2019) and Cole et al. (2023).

Feedforward architecture In SINR, a stack of residual blocks $\{\mathbf{R}_j, j \in \{1, ..., D\}\}$ processes the combined geo-location L and environmental embedding E, where the output of layer j for sample i is $h_i^j = \mathbf{R}_j(h_i^{j-1}) + h_i^{j-1}$. A linear layer then projects the last embedding to a vector of dimension $|S|$ from which per-species presence probabilities $\hat{y}$ are obtained using sigmoid activation.

Location encoding In SINR, the latitude and longitude of each data sample is encoded into a four dimensional sine-cosine embedding. For a given latitude and longitude (lat_i, lon_i), the location embedding is obtained with:

$$L_i = [sin(\pi\, lon_i), cos(\pi\, lon_i), sin(\pi\, lat_i), cos(\pi\, lat_i)] \quad (2)$$

This vector representation can be further enriched with bioclimatic variables into a vector $[L_i, E_i] \in \mathbb{R}^{24}$.

Training As explained earlier, one challenge of SDMs is the fact that presence-only data does not integrate well with conventional machine learning training losses, e.g., cross entropy for classification. Cole et al. (2023) introduce a specific loss $\mathcal{L}_{\text{AN-full}}$ to address this issue. This loss assumes that no other species than the observed one are present at the location, in addition to assuming no species being present at a location randomly sampled from the dataset scope.

$$\mathcal{L}_{\text{AN-full}}(\hat{y}, y_i) = -\frac{1}{|S|}\sum_{s=1}^{|S|}[y_{i_s} * \lambda * log(\hat{y}_s) + (1 - y_{i_s}) * log(1 - \hat{y}_s) + log(1 - \hat{y}'_s)] \quad (3)$$

The loss is based on two predictions from the model: $\hat{y}$ for the current observations location and $\hat{y}'$ for a randomly sampled location. For each class in S, their predicted probability is reinforced if their label in y_i is 1, and penalized otherwise. Their predicted probability for the randomly sampled location is also penalized regardless of their actual presence in that area as a form of negative background sampling. The loss includes a hyperparameter λ that balances the presence versus absence contribution to the final loss value.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-2-2024-41-2024 | © Author(s) 2024. CC BY 4.0 License.

3.3 Sat-SINR

We now explain how we extend this framework to incorporate satellite remote sensing observations.

Satellite image encoding In addition to the location and environment embeddings, Sat-SINR also takes the satellite image I_i as input. The satellite image is first processed by a convolutional encoder $\mathbf{C}$ into a 1-dimensional embedding c_i of dimension d_{sat} and becomes $c_i = \mathbf{C}(I_i)$. For $\mathbf{C}$, we rely on a simple CNN design based on Higgins et al. (2016), containing multiple convolutional layers followed by flattening and a series of fully connected layers. The convolutional encoder is trained end-to-end with the rest of the architecture.

Fusion strategy After encoding the satellite image into a one-dimensional vector, we explore different ways of fusing that information with the location and environmental vectors. L and E are considered the same modality. As portrayed in Figure 2, we design multiple architectures following early, late and middle fusion schemes:

- **Early fusion** concatenates all data sources into a single embedding before pushing it through the network. We choose to reduce the dimensionality of c_i to $\mathbb{R}^{24}$ in order to avoid dominating the location and environmental information. This method allows for the network to correlate information from all data sources early in the process and is the most common strategy employed in related remote sensing applications.

- **Late fusion**, on the other hand, creates a separate prediction for different modalities, merging them through addition at the last step before the application of the sigmoid function.

- **Middle Fusion** is inspired by ControlNet (Zhang et al., 2023), fusing the satellite data into the model after each layer. This enables the model to keep the predictive power of SINR while correlating the satellite information with the location information throughout the network.

In our middle fusion scheme, the information coming from the satellite image is fused with the other modalities throughout the layers of the feedforward network. For a given sample i, let h_i^j be the hidden vector after applying the j-th residual block of the network $\mathbf{R}_j$. For each layer, we use a linear layer $\mathbf{l_j}$ to project the satellite embedding c_i to a vector of the same dimension as h_i^j and add it residually before the following layer:

$$h_i^j = \mathbf{R}_j(h_i^{j-1}) + h_i^{j-1} + \mathbf{l_j}(c_i) \tag{4}$$

The trainable linear layer at each stage enables the model to control the amount of satellite information that is fused at different stages.

We amplify this control by initializing both the weight and the bias matrices to zero (Zhang et al., 2023). In that case, the first prediction contains no satellite information. Once the weights get updated, an increasing amount of satellite information flows into the predictions.

Training In each of the three fusion architectures we propose, all components are trained end-to-end. We use the same training loss as in SINR, with the important distinction of the sampling for $\hat{y}'$ in the third summand of the loss term. In the case of SINR, this location is sampled uniformly from the scope of the dataset, which in their case is the whole world. Due to the usage of satellite imagery, we cannot sample a random location, as retrieving and pre-processing the satellite image is too computationally expensive to do during training. We resolve this by employing a random training sample along with its satellite image as pseudo-absence data in place of a uniform sampling.

4. EXPERIMENTS

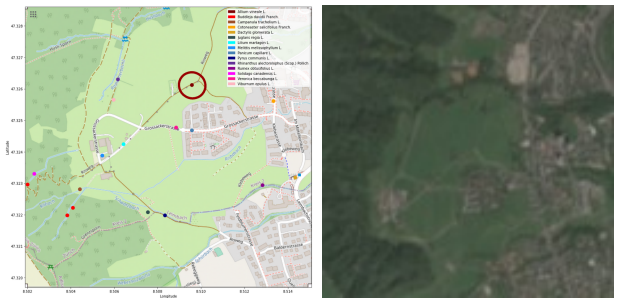


Figure 3. **Left:** A 1km by 1km square from the PO dataset in the west of Zurich. Background from OpenStreetMap. https://www.openstreetmap.org. Accessed 24.01.24. **Right:** The RGB Sentinel-2 images associated with the Allium vineale L. occurrence (dark red circle), a species of wild onion, whose stem grows up to a meter tall.

4.1 Presence-Only Training Data

In this work, we adapt the GLC23 dataset (Botella et al., 2023b), specifically the presence-only samples, using the public presence-absence surveys as test data. The original PO dataset consists of more than 5 million observations of 10k plant species covering most of Europe (38 countries) with a heavy spatial sampling bias favoring western Europe. The data is a mix of governmental and *research grade* crowd-sourced observations collected between 2017 and 2021 downloaded from GBIF.
The original dataset has a strong class imbalance, with some species occurring 4500 times and others only once. We follow the method of SINR and remove all species with less than 10 observations and cap the occurrences per class to 1000, yielding a dataset of 2 million observations of about 5.6k plant species. Figure 3 highlights the sparsity and sampling bias of the data. Due to the majority of the data being crowd-sourced, it mostly consists of observations of plants that are directly beside or visible from easily accessible streets or paths, and observations from artificially planted urban spaces. This dataset thus represents a rather anthropocentric impression of plant distributions. The spatial distribution of the data is bounded between the latitudes $[34.5686, 71.1839]$ and longitudes $[-10.4760, 34.5579]$, which we use as maximum and minimum for the cyclical values of the location embeddings.

4.2 Environmental and Satellite Data

Environmental context Similar to SINR, we use the Bioclim 2.1 and elevation data (Fick and Hijmans, 2017) at a resolution of 1km as environmental representation of our location. The bioclimatic data contains various metrics based on precipitation and temperature across the year.

Satellite imagery We use pre-processed Sentinel-2 mosaics from the Open Environmental Data Cube Europe[3] as part of the GLC23 dataset. Each image is the median Sentinel-2 value across the whole year in which the species observation was recorded. The image contains the four 10m-resolution channels RGB and NIR at a size of $H \times W = 128 \times 128$ pixels, thus covering an area of $1.28\text{km} * 1.28\text{km} \approx 1.64\text{km}^2$ centered around the observation location (Figure 3 **right**). Using satellite images thus increases the maximum predictor resolution of the model 100-fold compared to the environmental rasters.

4.3 Test Data and Metrics

The test data are 6k checklists filled out by botanical experts, sampled between 2017 and 2021 in France and Great Britain (excluding Northern Ireland), covering 2k plant species of which nearly all appear in the training data. A survey contains on average 13 observed species, with a minimum of 1 and a maximum of 73. This change between training and test data both in sampling and structure signifies a distributional shift that makes the test task quite challenging. In the GLC23 challenge no team on the leaderboard trained only on the PO data (Botella et al., 2023b), and the only baseline fitted on the PO data was the least performant. The top teams' model performance plummeted in an ablation study when removing PA data from their training pipeline, proving once again how challenging the task of training on PO data only is.
We calculate two metrics over the test data to measure our models' performance. In a first step, we define the **30** species with the highest probability in our model predictions as present, and all other species as absent. This is an arbitrary value, but yielded the best scores in preliminary experiments. We experimented with using a threshold to define presence from the probabilities, but this yields hundreds of present classes and thus does not fit with the statistics of the PA surveys. It is left for future work to apply a more ecologically sound interpretation to the output probabilities. The first metric is the micro F1 - score:

$$\text{F1} = \frac{1}{N} \sum_{i=1}^{N} \frac{\text{TP}_i}{\text{TP}_i + (\text{FP}_i + \text{FN}_i)/2} \tag{5}$$

$$\text{Where} \begin{cases} \text{TP}_i = \text{Number of predicted labels truly present} \\ \text{FP}_i = \text{Number of labels predicted but absent} \\ \text{FN}_i = \text{Number of labels not predicted but present} \\ N = \text{Total number of test surveys} \end{cases}$$

It calculates a trade-off between the precision and recall for each survey and averages that value across the whole test data. A second, more common metric in ecology is the receiver operating characteristic (ROC) curve, calculating precision and recall for a single species over all surveys with varying thresholds. The areas under the resulting precision-recall curves are averaged across all species. Macro ROC-AUC signifies an egalitarian average, with each species' ROC-AUC value contributing the same to the average. Weighted ROC-AUC, on the other hand multiplies each ROC-AUC value with the support (number of occurrences in the test data), causing strongly represented species to have significantly higher impact in the metric than little-sampled ones. Including both metrics allows to compare the models performance for both abundant and rare species.

[3] https://stac.ecodatacube.eu/. Accessed 24.01.24

4.4 Training Setup

Implementation details Training is done on a Nvidia T4 at 16GB. Runs are implemented with a batch size of 2048. The residual blocks are implemented as in Cole et al. (2023). The size of SINR is kept consistent across all models, quadrupeling in size compared to the original work with 8 layers of 512 hidden size to allow for more information to pass through the network. Dropout is reduced to 0.3. Hyperparameters are set based on a intuitively guided grid search around the original SINR parameters in multiple iterations. Not all viable combinations are exhausted due to computational constraints. The learning rate is set to 0.0007 and λ to 2048. Each model is run up to four times with different weight initializations to calculate the average and standard deviation for the three test metrics. We furthermore implement Logistic Regression (LogReg) by replacing SINR with a single linear layer.

4.5 Creating Species Range Maps

Once trained, our model makes a point-wise prediction for any queried location. To combine these into visually interpretable maps, we sample our study area with a regular 502 times 408 coordinate grid. This results in a distance between samples of about 10 kilometers along latitude, and roughly 3 to 8 kilometers along longitude. For each grid point, we calculate the location and bioclimatic embedding and pull a Sentinel-2 image from the Ecodatacube in the same manner as outlined in the GLC23 dataset to fit the structure of our training samples. The model prediction is then calculated for each grid point and mapped out. The maps thus created from the various runs for one model are averaged to reach the final image. This method of taking the average between multiple runs allows us to also create a map of the standard deviation for each location as a simple form of uncertainty quantification. The maps in Figure 1 and Figure 4 are for *Quercus Ilex* (Evergreen Oak), as it happens to be the lowest-index species present in the first presence-absence survey used as test data, thus no cherry picking took place. This oak appears 1000 times in the training data, and is native to the Mediterranean and Atlantic coast, reaching into the inner parts of France and Spain.

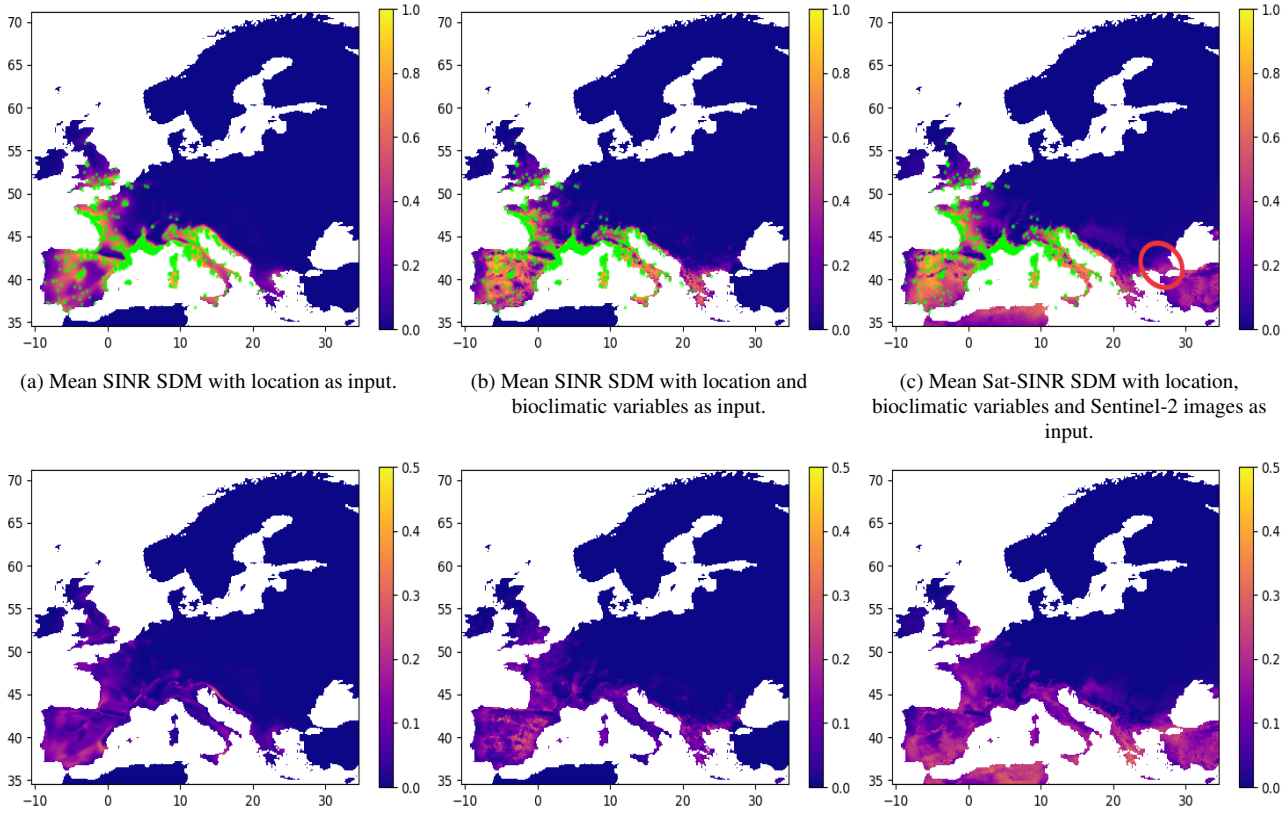
5. RESULTS

Numerical performance We show the main results of our experiments in Table 1. Adding satellite images improves the performance on all metrics, most notably the micro F1 - score. Each of the Sat-SINR fusion methods achieves the best score on a different metric, with late fusion showing the best cumulative results, although the standard deviation across runs makes the performance ranges overlap significantly. Middle fusion slightly outperforms on the weighted ROC-AUC, while being worse on the macro ROC-AUC. The trade-off for the significant performance improvement compared to SINR is a 30-fold runtime increase. However, we note that runtime is dominated by data loading, and a more efficient pipeline could help reducing it. We carry out all following visualizations and ablation studies with the late fusion version of Sat-SINR.

SDM prediction The resulting range map for Sat-SINR (Figure 4) highlights key advantages of this novel model design and input modality. It has smoother contours compared to the bioclimatic SDM while keeping a high level of detail compared to the location-only version. This is achieved partly due to confidence in areas far away from sample points, an effect

Table 1. Comparison of the baselines with the proposed architectures. All values are averaged over five runs. The macro ROC-AUC takes an average over the ROC-AUCs of all classes, while weighted ROC-AUC takes the same average while weighting each class's ROC-AUC based on its occurrence count in the validation data. We additionally include the number of parameters in million and the time for each model to converge. All metrics are listed as percentages.

Model	Predictors	Macro ROC-AUC	Weighted ROC-AUC	Micro F1	Parameters	T to convergence
LogReg	Loc + Env	70.77±0.01	66.58±0.05	3.33±0.23	251k	1.5 hrs
SINR	Loc	73.98±0.08	71.00±0.02	3.23±0.70	9.4M	1 hr
SINR	Loc + Env	76.62±0.07	73.16±0.01	4.48±0.17	9.4M	1 hr
Sat-SINR EF	Loc + Env + Sat	**78.53±0.06**	74.99±0.01	7.01±0.20	9.7M	40 hrs
Sat-SINR LF	Loc + Env + Sat	78.48±0.04	75.18±0.04	**7.22±0.28**	12.3M	30 hrs
Sat-SINR MF	Loc + Env + Sat	77.51±0.05	**75.19±0.02**	6.90±0.13	12.8M	30 hrs



(a) Mean SINR SDM with location as input.

(b) Mean SINR SDM with location and bioclimatic variables as input.

(c) Mean Sat-SINR SDM with location, bioclimatic variables and Sentinel-2 images as input.

(d) Standard deviation SINR SDM with location as input.

(e) Standard deviation SINR SDM with location and bioclimatic variables as input.

(f) Standard deviation Sat-SINR SDM with location, bioclimatic variables and Sentinel-2 images as input.

Figure 4. The resulting mean and standard deviation probability maps for *Quercus Ilex* (Evergreen Oak) when averaging four model outputs. Sat-SINR shows the ability to generalize to areas without training samples (red circle). **X-axis**: longitude east, **y-axis**: latitude north, **colours**: predicted probability of 0 (blue) to 1 (yellow), **green dots**: occurrences of the evergreen oak in the training dataset.

that can be attributed to the change in the sampling of $\hat{y}'$ in the loss. In SINR, all regions regardless of training sample density are penalized, while in Sat-SINR, regions are penalized based on sample density due to the availability of satellite images at runtime. Another effect of removing this penalty for non-sampled areas is the model placing probability mass in eastern Greece and western Anatolia, which are devoid of training samples, but do in fact constitute fitting environments for the evergreen oak. This higher confidence in non-sampled areas can also be attributed to a strong signal from the satellite image and bioclimatic variables. The map of standard deviation between different runs allows for uncertainty estimation of each prediction, exposing the higher variability between local model predictions in Sat-SINR. This increased uncertainty hints at the ability of Sat-SINR to learn multiple plausible distributions from the data, instead of arriving at the same boundaries in each run. Furthermore, this higher disagreement between local predictions does not impact the model performance on the test data, as seen in the low standard deviations of the metrics in Table 1. We leave it for future work to evaluate the calibration of these uncertainties.

5.1 Ablation Studies

Modality mixes We test each combination of the three input modalities considered in this work. The satellite-only model is the late fusion model with the SINR embedding set to zero. All other models including satellite images are late fusion Sat-SINR. The results in Table 2 highlight the value of the satellite images. The satellite-only model shows a performance similar to combinations of Loc and Env, but significantly increases

in performance when combined with either of the two. Yet it still lags behind the combination of all three, especially on the macro ROC-AUC. This experiment highlights the complementary nature of the information that the model derives from the three predictors, as neither a satellite-only model nor SINR show performance comparable to Sat-SINR.

Table 2. Modality contribution (average of five runs).

Predictors	Macro RA	Weighted RA	Micro F1
Loc	73.99	71.01	3.92
Env	75.34	72.53	4.59
Sat	75.63	73.31	4.66
Loc + Env	76.63	73.17	4.48
Loc + Sat	77.41	74.62	6.89
Env + Sat	77.08	74.86	7.28
Loc + Env + Sat	78.48	75.18	7.22

Satellite image size We reduce the size of the satellite image, keeping it centered around the observation (Figure 5 **top**). The size reduction entails a marginal drop in performance across all metrics. This shows that a large spatial context is not required for the model to extract relevant information for the task at hand. This could be because the model is only focusing on pixels close to the location of interest, or because it extracts some properties of the local terrain that remain fairly similar across a 1 kilometer stretch. We note that reducing the image size could therefore lead to a better performance/runtime trade-off.

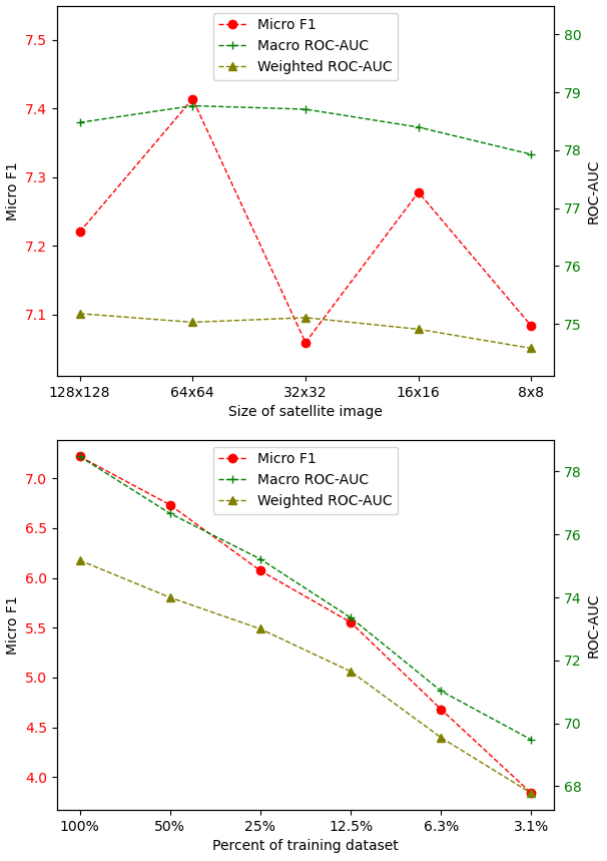


Figure 5. The metrics when reducing the image size centered around the occurrence (**Top**) and when reducing the amount of training data (**Bottom**).

Training data impact We reduce the amount of training data by randomly removing half of it up to five times (Figure 5 **bottom**). This causes a significant drop in performance, with the descent becoming slightly steeper in the low-data regime, even though less datapoints are being removed. This near-linear progression implies that with fewer training samples, each samples carries more weight, and we observe a diminishing return per new datapoint added. Furthermore, we experiment with removing all training data in France (Table 3), reducing the training dataset size by a third. France contains 73% of test surveys, thus this ablation hints at the extrapolation abilities of the models to un-sampled, but suitable areas. In this setting, Sat-SINR significantly outperforms SINR, highlighting its generalization capabilities.

Table 3. The impact of removing all training observations in France while retaining all test surveys in France.

Predictors	Macro RA	Weighted RA	Micro F1
Loc	51.0	50.88	1.49
Loc + Env	60.19	49.83	2.76
Loc + Env + Sat	76.1	72.55	5.91

6. CONCLUSION

We proposed Sat-SINR, a satellite imagery extension to the presence-only species distribution model SINR. Including satellite imagery increases the performance on presence-absence surveys in Europe and the resulting maps show novel characteristics such as plausible extrapolation to under-sampled areas and detailed yet smooth contours. Late fusion achieved the best results overall, with a significant advantage on Micro-F1 and a close second on the other two metrics. The experiments show that satellite images are a valuable source of information even under strong distributional shift between training and test data, with the added benefit of generalizing to un-sampled areas. The model can in practice be used with higher resolution satellite data, which allows to identify individual plants instead of just the surrounding environment.

It remains an advantage of non-satellite SDMs to naturally use synthetic future scenarios as input, such as the precipitation and temperature regime in a location modeled 100 years in the future. But equally, simulations of the change of vegetation and landscapes on a high resolution might be used as synthetic satellite images to bridge this gap.

In future works, we believe that interesting contributions could be made in trying to better handle the high imbalance of the data at hand, and in leveraging the temporal dimension of satellite data for better distribution modelling.

References

Aodha, O. M., Cole, E., Perona, P., 2019. Presence-Only Geographical Priors for Fine-Grained Image Classification. *CoRR*, abs/1906.05272. http://arxiv.org/abs/1906.05272.

Beery, S., Cole, E., Parker, J., Perona, P., Winner, K., 2021. Species Distribution Modeling for Machine Learning Practitioners: A Review. *Proceedings of the 4th ACM SIGCAS Conference on Computing and Sustainable Societies*, COMPASS ’21, Association for Computing Machinery, New York, NY, USA, 329–348.

Botella, C., Deneu, B., Gonzalez, D. M., Servajean, M., Larcher, T., Leblanc, C., Estopinan, J., Bonnet, P., Joly, A., 2023a. Overview of GeoLifeCLEF 2023: Species composition prediction with high spatial resolution at continental scale using remote sensing. *Working Notes of CLEF*.

Botella, C., Deneu, B., Marcos, D., Servajean, M., Estopinan, J., Larcher, T., Leblanc, C., Bonnet, P., Joly, A., 2023b. The GeoLifeCLEF 2023 Dataset to evaluate plant species distribution models at high spatial resolution across Europe. arXiv:2308.05121 [q-bio, stat].

Brun, P., Karger, D. N., Zurell, D., Descombes, P., Witte, L. C. d., Lutio, R. d., Wegner, J. D., Zimmermann, N. E., 2024. Multispecies deep learning using citizen science data produces more accurate plant community models. *Nature Communications*. accepted for publication.

Cole, E., Van Horn, G., Lange, C., Shepard, A., Leary, P., Perona, P., Loarie, S., Mac Aodha, O., 2023. Spatial Implicit Neural Representations for Global-Scale Species Mapping. arXiv:2306.02564 [cs].

Damer, N., Dimitrov, K., Braun, A., Kuijper, A., 2019. On Learning Joint Multi-biometric Representations by Deep Fusion. *2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, IEEE, Tampa, FL, USA, 1–8.

de Lutio, R., She, Y., D’Aronco, S., Russo, S., Brun, P., Wegner, J. D., Schindler, K., 2021. Digital taxonomist: Identifying plant species in community scientists’ photographs. *ISPRS Journal of Photogrammetry and Remote Sensing*, 182, 112–121. https://www.sciencedirect.com/science/article/pii/S0924271621002641.

Deneu, B., Joly, A., Bonnet, P., Servajean, M., Munoz, F., 2022. 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. https://www.frontiersin.org/articles/10.3389/fpls.2022.839279.

Dıaz, S., Settele, J., Brondızio, E., Ngo, H., Guèze, M., Agard, J., Arneth, A., Balvanera, P., Brauman, K., Butchart, S. et al., 2019. IPBES (2019): Summary for policy makers of the global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services. *IPBES secretariat, Bonn, Germany: Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services*.

D’mello, S. K., Kory, J., 2015. A Review and Meta-Analysis of Multimodal Affect Detection Systems. *ACM Computing Surveys*, 47(3), 43:1–43:36. https://dl.acm.org/doi/10.1145/2682899.

Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., Martimort, P. et al., 2012. Sentinel-2: ESA’s optical high-resolution mission for GMES operational services. *Remote sensing of Environment*, 120, 25–36.

Fick, S. E., Hijmans, R. J., 2017. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *International journal of climatology*, 37(12), 4302–4315.

Goodfellow, I., Bengio, Y., Courville, A., 2016. *Deep learning*. MIT press.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., Lerchner, A., 2016. beta-vae: Learning basic visual concepts with a constrained variational framework. *International conference on learning representations*.

Jiang, Y., Rüetschi, M., Garnot, V. S. F., Marty, M., Schindler, K., Ginzler, C., Wegner, J. D., 2023. Accuracy and consistency of space-based vegetation height maps for forest dynamics in alpine terrain. *Science of Remote Sensing*, 8, 100099.

Kalischek, N., Lang, N., Renier, C., Daudt, R. C., Addoah, T., Thompson, W., Blaser-Hart, W. J., Garrett, R., Schindler, K., Wegner, J. D., 2023. Satellite-based high-resolution maps of cocoa planted area for C\ˆote d’Ivoire and Ghana. arXiv:2206.06119 [cs].

Karaman, K., Sainte Fare Garnot, V., Wegner, J. D., 2023. Deforestation detection in the Amazon with sentinel-1 SAR image time series. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 10, 835–842.

Katzenmaier, M., Garnot, V. S. F., Björklund, J., Schneider, L., Wegner, J. D., von Arx, G., 2023. Towards ROXAS AI: Deep learning for faster and more accurate conifer cell analysis. *Dendrochronologia*, 81, 126126.

Klemmer, K., Rolf, E., Robinson, C., Mackey, L., Rußwurm, M., 2023. SatCLIP: Global, General-Purpose Location Embeddings with Satellite Imagery. *arXiv preprint arXiv:2311.17179*.

Lang, N., Jetz, W., Schindler, K., Wegner, J. D., 2023. A high-resolution canopy height model of the Earth. *Nature Ecology & Evolution*, 7(11), 1778–1789. https://www.nature.com/articles/s41559-023-02206-6. Number: 11 Publisher: Nature Publishing Group.

Lang, N., Schindler, K., Wegner, J. D., 2021. High carbon stock mapping at large scale with optical satellite imagery and spaceborne LIDAR. arXiv:2107.07431 [cs].

Navarro, L. M., Fernandez, N., Guerra, C., Guralnick, R., Kissling, W. D., Londono, M. C., Muller-Karger, F., Turak, E., Balvanera, P., Costello, M. J. et al., 2017. Monitoring biodiversity change through effective global coordination. *Current opinion in environmental sustainability*, 29, 158–169.

Rew, J., Cho, Y., Hwang, E., 2021. A robust prediction model for species distribution using bagging ensembles with deep neural networks. *Remote Sensing*, 13(8), 1495.

Rockström, J., Gupta, J., Qin, D., Lade, S. J., Abrams, J. F., Andersen, L. S., Armstrong McKay, D. I., Bai, X., Bala, G., Bunn, S. E. et al., 2023. Safe and just Earth system boundaries. *Nature*, 1–10.

Rußwurm, M., Klemmer, K., Rolf, E., Zbinden, R., Tuia, D., 2023. Geographic location encoding with spherical harmonics and sinusoidal representation networks. *arXiv preprint arXiv:2310.06743*.

Sastry, S., Xing, X., Dhakal, A., Khanal, S., Ahmad, A., Jacobs, N., 2023. LD-SDM: Language-Driven Hierarchical Species Distribution Modeling. *arXiv preprint arXiv:2312.08334*.

Teng, M., Elmustafa, A., Akera, B., Bengio, Y., Abdelwahed, H. R., Larochelle, H., Rolnick, D., 2023. SatBird: Bird Species Distribution Modeling with Remote Sensing and Citizen Science Data. *arXiv preprint arXiv:2311.00936*.

Zbinden, R., van Tiel, N., Rußwurm, M., Tuia, D., 2024. Imbalance-aware Presence-only Loss Function for Species Distribution Modeling. *arXiv preprint arXiv:2403.07472*.

Zhang, L., Rao, A., Agrawala, M., 2023. Adding Conditional Control to Text-to-Image Diffusion Models. arXiv:2302.05543 [cs].