Ecology Letters

WILEY

ECOLOGY LETTERS

METHOD

From Presence-Only to Abundance Species Distribution Models Using Transfer Learning

Benjamin Bourel[1] | Alexis Joly[1] | Maximilien Servajean[2,3] | Simon Bettinger[4] | José Antonio Sanabria-Fernández[5] | David Mouillot[4]

[1]Inria, University of Montpellier, LIRMM, CNRS, Montpellier, France | [2]LIRMM, University of Montpellier, CNRS, Montpellier, France | [3]AMIS, Paule Valery University, Montpellier, France | [4]MARBEC, University of Montpellier, CNRS, IFREMER, IRD, Montpellier, France | [5]Department of Ecology and Evolution, Doñana Biological Station (EBD-CSIC), Sevilla, Spain

Correspondence: Benjamin Bourel (benjamin.bourel@inria.fr)

Received: 8 November 2024 | **Revised:** 20 June 2025 | **Accepted:** 26 June 2025

Editor: Jonathan Lenoir

Funding: This work was supported by European Union's Horizon research and innovation programme—MAMBO project (101060639), European Union's Horizon research and innovation programme—GUARDEN project (101060693), IA-Biodiv ANR—FISH-PREDICT project (ANR-21-AAFI-0001-01), Biodiversa+—BioBoost+ project (EU grant agreement 101052342), Junta de Andalucía Postdoctoral Fellowship (DGP_POST_2024_00757).

Keywords: abundance | deep learning | fish | Mediterranean Sea | random forest | rare species | remote sensing | satellite imagery

ABSTRACT

Species Distribution Models based on Convolutional Neural Networks (CNN-SDMs) have recently emerged, demonstrating greater effectiveness than traditional SDMs in several contexts. A limited number of studies, however, have focused on species abundance patterns, as the datasets available for this purpose are generally too small to effectively learn a deep learning model with millions of parameters. Our study demonstrated that CNN-SDMs can circumvent the small sample size of species abundance datasets through the combined use of a large presence-only species dataset and transfer learning to significantly improve the performance of abundance-based CNN-SDMs. Applied to Mediterranean coastal fishes, our approach significantly improves the abundance prediction performance of CNN-SDMs, with average gains of 35% (D-squared regression score). This allows CNN-SDMs to perform better than classical SDMs in abundance prediction, with average gains of 10%. These gains are stemming from enhanced abundance predictions for rare species and where widespread species are locally rare.

1 | Introduction

The local population abundances of many taxa are declining worldwide, even within protected areas, under the combined pressure of climate change and human exploitation (Chaikin et al. 2024; Edgar et al. 2023; van Klink et al. 2024; Nowakowski et al. 2023; Pollock et al. 2022; Rigal et al. 2023). To identify the most at-risk populations (Ceballos et al. 2020), it is therefore crucial to accurately predict species abundances across space and time. Species Distribution Models (SDMs) are key tools that establish relationships between species distributions (presence-only, presence-pseudo-absences, presence-absence, and abundances) and the covariates that inform about habitat, environmental conditions, and anthropogenic conditions affecting their distribution (Srivastava et al. 2019; Waldock et al. 2022). However, predicting species abundances remains a challenge, especially for rare species or those with aggregated distributions (non-uniform distribution throughout the space, but rather concentrated in certain areas) (Chardon et al. 2022; Finn et al. 2023).

In recent years, new types of SDMs based on deep learning, including Convolutional Neural Networks (CNN-SDMs) have emerged (Botella et al. 2018). Basically, a CNN-SDM is an SDM that uses a CNN to predict the presence or abundance of species with spatial data like environmental rasters or satellite images centred on the location of observations (Botella et al. 2018; Deneu, Joly, et al. 2021). CNN-SDMs potentially possess greater predictive ability than classical SDMs (e.g., Random Forest) for three main reasons. Firstly, their architecture allows complex and non-linear combinations of covariates without over-constraining their functional form. This is an important advantage over classical SDMs, which generally use classification or regression techniques with a limited dimensionality of habitat, environmental, and anthropogenic covariates (Phillips et al. 2004, 2006; Waldock et al. 2022). Secondly, the use of spatio-temporal data representing these covariates and remote sensing images as input data enables these models to capture a broader spatial context to predict species distributions (Botella et al. 2018) which remains difficult to achieve with classical SDMs. Thirdly, in the context of SDMs that simultaneously predict the distribution of multiple species (multi-class SDM), unlike conventional SDMs, CNN-SDMs can indirectly capture and use biotic associations (species co-occurrences, correlations in species abundances, etc.) via their convolutional layers (Chen et al. 2017) to predict assemblages of multiple species (Brun et al. 2024; Hu et al. 2025).

CNN-SDMs have been successfully applied to better predict site occupancy from presence-only species data than machine learning methods like random forest (RF) (Deneu, Joly, et al. 2021; Estopinan et al. 2022), but their ability to predict species abundances is still poorly known, whereas this information is crucial for providing indications of habitat quality, conservation priorities, and ecosystem management. The limited application of CNN-SDMs for predicting species abundances is primarily due to the lack of large datasets because CNN-SDMs need more data than classical models to be trained. As shown by the main database aggregators (GBIF 2023; OBIS 2024), the presence-only species datasets are much larger and more numerous than the species abundance datasets, which remain generally far below what is typically required to train CNN-SDMs (Botella et al. 2018). Although theoretically more efficient, CNN-SDMs can perform less well than classical SDMs when the training dataset is small. For this reason, most current CNN-SDMs utilise occurrence datasets to predict species occurrence probabilities (Wang and Tabeta 2023), excluding abundance predictions.

The issue posed by the relatively small size of species abundance datasets for training CNN-SDMs could be circumvented by a transfer learning procedure. Transfer learning allows the transfer of layer weights, i.e., the adjustable coefficients that modulate the transmission of information between artificial neurons, from a first CNN-SDM trained on species occurrence predictions towards a second CNN-SDM performing a new task like species abundance predictions (Gupta et al. 2022). The underlying idea is that the second model will be able to draw on the knowledge acquired by the first model to learn the new task more effectively. As explained by Taylor and Stone (2009) and then Weiss et al. (2016), since large training datasets can be difficult to acquire for some tasks or research fields, creating high-performance learners trained with other massive data can represent an alternative and relevant option. Yet, in general, there is a close connection between the data used for transfer learning and the final data to be predicted (Weiss et al. 2016; Yeh et al. 2020). So, using CNN-SDMs trained on species occurrences to better predict species abundances is not intuitive and cannot be taken for granted in ecology, even if some studies have already attempted to estimate species abundances using presence-only data (e.g., Pearce and Boyce 2006). However, it has been pointed out that these various methods have many limitations and are proxies at best (Bradley 2016; Pearce and Boyce 2006).

In light of this, our article aims to address two primary objectives. Firstly, we evaluate the extent to which CNN-SDMs can predict fish species abundances along the Mediterranean coast by leveraging transfer learning from occurrence-based CNN-SDMs to overcome the small sample size of fish abundances. To achieve this, we hypothesise that the layers of the neural network previously trained with massive presence-only data could capture general information and patterns that could be reused for a different but related task of predicting species abundances (Gupta et al. 2022). Secondly, we assess whether this procedure enables CNN-SDMs to outperform classical SDMs (here RF) in predicting fish species abundances with small datasets. We show that our two-step approach has the potential to revolutionise the way we train and utilise CNN-SDMs, even when working with limited datasets across various tasks.

2 | Materials and Methods

2.1 | Fish Occurrence and Abundance Datasets

The fish occurrence dataset was retrieved from GBIF by extracting species presence-only data (taxa of the Chordata, excluding Tunicata and Tetrapoda) recorded in the Mediterranean Sea through observations carried out between 2011 and 2022, with a spatial resolution accuracy ranging from 0 to 100 m ("GBIF dataset" 2022; "GBIF dataset" 2023). We selected fish occurrences located up to 20 km from the coast in order to include as many areas as possible where abundances were collected (see next paragraph). Species observed at less than 5 km from areas with depth lower than 50 m were also selected (Figure 1A) to take into account small islets that are not georeferenced, as well as shallow reefs located on continental shelves. Thereafter, occurrences not identified at the species level and those where environmental data were not available were removed. Finally, species with fewer than 10 occurrences were removed, resulting in the final dataset of 62,240 occurrences for 181 fish species (Figure 1B and Table S1). This threshold of 10 occurrences is relatively low but is useful for taking rare species into account (Breiner et al. 2015).

The fish abundance datasets were provided by underwater fish counts (species abundances per site at a given time for all observed fish individuals). This survey was performed in the Mediterranean Sea between 2011 and 2020 by the Reef Life Survey programme (Galaxy for Ecology 2022). This Reef Life Survey dataset is also available in GBIF (key: 38f06820-08c5-42b2-94f6-47cc3e83a54a) but has not been included in our fish occurrence dataset (see column 2 of Table S1). To construct this abundance dataset, divers counted and identified all the fish

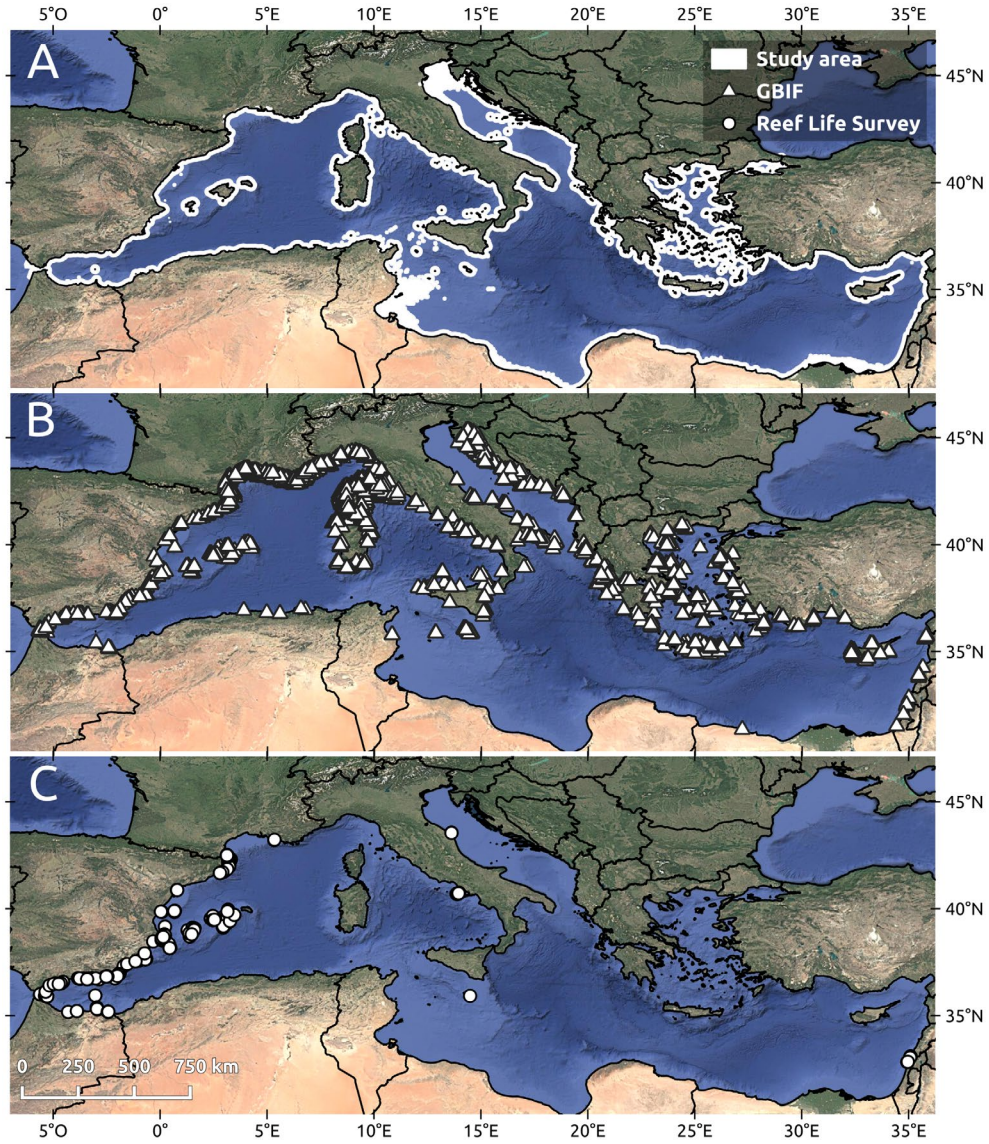

FIGURE 1 | Maps of the Mediterranean Sea showing (A) the study area, (B) the location of the 62,240 GBIF occurrences, and (C) the 217 sites grouping the 406 fish abundance counts of Reef Life Surveys.

observed in two 5 m wide by 5 m deep blocks along 50 m transect (Edgar et al. 2020). Here, diver counts were aggregated based on their SurveyID in order to combine the counts of the different blocks of the same transect. After that, our abundance dataset consists of 406 fish abundance counts spread over 217 sites with some counts made at the same site on different dates (Figure 1C). Taxa not identified at the species level and species present in fewer than 10 fish abundance counts were removed. The final fish abundance dataset used in this article (Table S2) consists of 47 species, including 3 species absent from the presence-only dataset.

2.2 | Environmental Dataset

For each fish occurrence or abundance count, the corresponding environmental dataset was composed of 15 rasters representing 14 environmental covariates and one satellite image (Table 1). Each raster resulted from on-the-fly extraction (the source data used for extraction are indicated in Table 1) and was spatially centred on the GPS position of the associated observation. Three sizes of raster windows (size of the raster centred on the GPS position) were arbitrarily defined in order to have enough cells inside each raster window to obtain spatial patterns of 3×3, 30×30, or 63×63 km. Eight out of 15 rasters were multi-channel. The different channels in the same raster represent either different colours (e.g., RGB rasters), different depths (e.g., salinity raster) or different spatial axes (e.g., coordinate raster). Details regarding these channels are available in Table 1. Data in the rasters are quantitative, except for the substrate, which had 9 classes, one per raster (Table S3).

2.3 | Data Pre-Processing

We used two datasets, one for fish species occurrences and another for fish species abundances, which were both split into training, validation, and test sets. The validation set was

TABLE 1 | Description of rasters and associated environmental data. The rasters are spatially centred on the GPS position of the associated fish occurrences or abundances. For mean and standard deviation, the values are ordered from the first to the last channel.

Raster	Description	Raster size (in cells)	Cell resolution (in km)	Size of the raster window (in km)	No. of bands	Nature	Mean (M) and standard deviation (SD) by band for our GBIF data	Temporality	
rgb	True colour satelite image in RGB (band 0 = red, band 1 = green, and band 2 = blue)	300 × 300	0.010	3 × 3	3	q.uantit	M = [36.7, 47.9, 65.4] SD = [35.7, 28.1, 22.8]	The most recent day with the lowest cloud cover at the event coordinates between August 20 and September 29, 2022	Band TCI of Copernicus Sentinel-2 Data (2023)
nir	Satelite image in near-infrared (central wavelength = 842 nm and bandwidth = 115 nm)	300 × 300	0.010	3 × 3	1	quantit.	M = [2061.9] SD = [1181.9]		Band B08 of "Copernicus Sentinel-2 data" (2023)
coord	GPS coordinates of the event in uniform raster (band 0 = latitude and band 1 = long.)	256 × 256	NA	NA	2	quantit.	Not used in this article	No temporality	"GBIF dataset" (2022), "GBIF dataset" (2023), and "Galaxy for Ecology" (2022)
bathy	Bathymetry (m)	31 × 31	0.095	3 × 3	1	quantit.	M = [−110.7] SD = [169.8]	No temporality	"EMODnet Bathymetry Consortium" (2020)
chloro	Average daily chlorophyll concentration (mg/m^3) in sea water at surface	30 × 30	1.000	30 × 30	1	quantit.	M = [0.15] SD = [0.36]	The event day	Variable "CHL" in field "[...]-gapfree-multi-1km_P1D" of E.U. Copernicus Marine Service Information (2023b)
sst	Average daily sea surface temperature (°K)	32 × 32	0.950	30 × 30	1	quantit.	M = [295.0] SD = [3.9]	The event day	Variable "analysed_sst" in field "[...]_010_004_c_V2" of E.U. Copernicus Marine Service Information (2023a)
east_vow_day	Average daily eastward sea water velocity (m/s) at 1.01 m (band 0), 26.2 m (band 1), and 51.4 m (band 2)	15 × 15	4.200	63 × 63	3	quantit.	M = [−0.06, −0.06, −0.05] SD = [0.14, 0.12, 0.11]	The event day	Variable "uo" in field "med-cmcc-cur-rean-d" of E.U. Copernicus Marine Service Information (2023c)
east_vow_month	Average monthly eastward sea water velocity (m/s) at 1.01 m (band 0), 26.2 m (band 1), and 51.4 m (band 2)	15 × 15	4.200	63 × 63	3	quantit.	M = [−0.06, −0.06, −0.05] SD = [0.12, 0.11, 0.09]	Calendar month of the event	Variable "uo" in field "med-cmcc-cur-rean-m" of E.U. Copernicus Marine Service Information (2023c)

(Continues)

TABLE 1 | (Continued)

Raster	Description	Raster size (in cells)	Cell resolution (in km)	Size of the raster window (in km)	No. of bands	Nature	Mean (M) and standard deviation (SD) by band for our GBIF data	Temporality	
north_vow_day	Mean day northward sea water velocity (m/s) at 1.01 m (band 0), 26.2 m (band 1), and 51.4 m (band 2)	15×15	4.200	63×63	3	quantit.	M = [−0.06, −0.03, −0.03] SD = [0.14, 0.12, 0.10]	The event day	Variable "vo" in field "med-cmcc-cur-rean-d" of E.U. Copernicus Marine Service Information (2023c)
north_vow_month	Average monthly northward sea water velocity (m/s) at 1.01 m (band 0), 26.2 m (band 1), and 51.4 m (band 2)	15×15	4.200	63×63	3	quantit.	M = [−0.06, −0.04, −0.03] SD = [0.11, 0.10, 0.09]	Calendar month of the event	Variable "vo" in field "med-cmcc-cur-rean-m" of E.U. Copernicus Marine Service Information (2023c)
salinity_day	Average daily salinity (psu) at 1.01 m (band 0), 26.2 m (band 1), and 51.4 m (band 2)	15×15	4.200	63×63	3	quantit.	M = [38.15, 38.17, 38.02] SD = [0.48, 0.38, 0.40]	The event day	Variable "so" in field "med-cmcc-sal-rean-d" of E.U. Copernicus Marine Service Information (2023c)
salinity_month	Average monthly salinity (psu) at 1.01 m (band 0), 26.2 m (band 1), and 51.4 m (band 2)	15×15	4.200	63×63	3	quantit.	M = [38.15, 38.18, 38.02] SD = [0.47, 0.38, 0.40]	Calendar month of the event	Variable "so" in field "med-cmcc-sal-rean-m" of E.U. Copernicus Marine Service Information (2023c)
sea_floor_temp_day	Average daily of sea water potential temperature (°C) at sea floor	15×15	4.200	63×63	1	quantit.	M = [14.3] SD = [1.9]	The event day	Variable "bottomT" in field "med-cmcc-tem-rean-d" of E.U. Copernicus Marine Service Information (2023c)
sea_floor_temp_month	Average monthly sea water potential temperature (°C) at sea floor	15×15	4.200	63×63	1	quantit.	M = [14.3] SD = [1.9]	Calendar month of the event	Variable "bottomT" in field "med-cmcc-tem-rean-m" of E.U. Copernicus Marine Service Information (2023c)
substrate	9 seabed substrate classes based on 14 classes in the file Geodatabase of Vasquez et al. (2021)	256×256	0.012	3×3	1	catego.	Not used in this article	No temporality	Field "substrate" of Vasquez et al. (2021)

used during the learning phase to provide independent evaluation of the model's performance and to refine the model's hyperparameters. The test set was used after the training and validation phase to evaluate the performance of the model on completely new data. The data split was carried out with spatial blocks in each marine ecoregion defined by Spalding et al. (2007) to avoid the underestimation of the predictive error (Roberts et al. 2017). These blocks were defined as 10 × 10 km for occurrences and 5 × 5 km for abundance counts. In each ecoregion, the blocks are randomly divided into training, validation, and test sets, with 70%, 15%, and 15% respectively for the occurrence dataset and 60%, 20%, and 20% respectively for the abundance dataset since we had fewer observations. When the number of blocks was not sufficient to meet these exact proportions, the surplus was automatically allocated to the training set (Data S1).

For environmental data, the rasters were resized to 256 × 256 pixels using nearest neighbour interpolation to ensure that all rasters had the same pixel size, thus enabling consistent inputs into the model. This transformation changed the pixel size of the rasters, but not the size of raster windows (spatial extents), which remained fixed (Table 1). Each undefined value ("NaN") was replaced by the mean value of the associated channel. The values of rasters were all standardised (Data S1), except for the rasters representing the marine substrate and coordinates (Table 1). Additionally, for substrate raster, the single channel with 9 classes (8 marine substrate classes and the land class) was replaced by 8 new channels with 3 classes per channel: class 1 for one of the 8 substrate classes, class 0 for the other substrates and class-1 for land. This modification increased the number of channels in the 15 rasters from 30 (Table 1) to 37 (30–1 + 8).

Finally, the channels of the 15 rasters were concatenated into a 3D tensor of dimensions 256 × 256 × 37, where 37 was the total number of raster channels. Each tensor was associated with the corresponding species occurrences and species abundances in each dataset (Table S1 and Table S2). For the abundance dataset, we obtained vectors, one per fish abundance count, consisting of 47 values representing the 47 fish species. The values in these vectors were organised according to the alphabetical order of the species (Table S2).

2.4 | CNN-SDMs

The CNN-SDMs were built using a ResNet-50 from scratch (He et al. 2016) in PyTorch (Paszke et al. 2019). ResNet-50 is a deep learning model, and more specifically, a CNN designed to take the 3 channels of RGB images as inputs. Here, we modified it to take the 37 channels of our 3D tensor (environmental data) as inputs. In addition, the last fully connected layer of the ResNet-50 was replaced by a sequence that begins with a dropout layer to prevent overfitting, followed by the original last fully connected layer. For fish abundance models, the PyTorch's relu activation function was added at the end of this sequence to replace negative outputs with 0. Finally, a data augmentation step was incorporated. Details of data augmentation, optimiser, hyperparameters, and regularisation methods are presented in Data S1.

CNN-SDMs trained with occurrence data used the Cross Entropy Loss (Equation 1) of PyTorch to perform a multi-class classification with 181 classes corresponding to the 181 fish species. These occurrence models predicted species occurrence probabilities using presence-only data, so they did not require the use of absences or pseudo-absences. This choice avoided the methodological issues related to pseudo-absence generation (Raiter and Hawlena 2024). CNN-SDMs trained with fish abundance data employed the L1 Log Loss of PyTorch (Equation 2).

$$\text{Cross Entropy Loss} = -\sum_{i=1}^{n} p_i \times \log\left(\hat{p}_i\right) \tag{1}$$

where n = number of classes (here species), p_i = true probability distribution for the i class, and $\hat{p}_i$ = predicted probability distribution for the i class.

$$L1\ \text{Log Loss} = \frac{1}{n}\sum_{i=1}^{n}\left|\log\left(y_i + 1\right) - \log\left(\hat{y}_i + 1\right)\right| \tag{2}$$

where n = number of samples (here species abundance counts), y_i = true value of i sample, and $\hat{y}_i$ = predicted value of i sample.

The training of a model consists of several epochs. An epoch corresponds to a complete passage of all the learning data. The training of the model stops automatically when the performance of the model on the validation data has not improved after 6 epochs. At the end of training, the layer weights saved are those corresponding to the epoch with the best performance on the validation set before overfitting, if any (Data S1). The results of CNN-SDMs predicting fish occurrences were based on a single cross-validation (hold-out validation) using a single split from presence-only data (Table S1). This choice of the single split is explained in Data S1. The results of CNN-SDMs for abundances were based on a k-fold cross-validation, using k = 20 random splits of spatial blocks from the dataset (Table S2). These 20 random folds were the same for models with and without transfer learning.

2.5 | Transfer Learning From Occurrence to Abundance CNN-SDMs

The transfer learning procedure consisted of transferring the weights of the layers obtained from the model trained to predict species occurrences to the model predicting species abundances, with the exception of the last fully connected layer that was initialised randomly. In view of the preliminary tests carried out during the training phase of the abundance model, all layer weights were updated.

2.6 | Benchmarking With Random Forest Model

To evaluate the performance of the CNN-SDMs, we employed an RF, a more conventional yet highly effective machine learning model (e.g., Waldock et al. 2022). We used an RF classifier on occurrence data and an RF regressor on abundance data to compare them, respectively, to the CNN-SDM models predicting species occurrences and to the CNN-SDM models predicting abundances. For that, we trained and tested the RF models with the same splits

and folds between the training, validation, and test sets as those used with the corresponding CNN-SDMs (Tables S1 and S2).

To train and test these different RFs, we used scikit-learn (Pedregosa et al. 2011). As for the CNN-SDMs, we used the Cross Entropy Loss criterion for the RF predicting species occurrences to perform a multi-class classification and predict species occurrence probabilities. As for the CNN-SDMs, this RF did not require species absences or pseudo-absences. For the RF predicting abundances, we used the Mean Absolute Error criterion. The settings of the RFs used for predicting both occurrences and abundances were optimised using GridSearchCV from scikit-learn (Pedregosa et al. 2011). Basically, GridSearchCV optimised the hyperparameters of RFs by testing all the combinations of values and selecting the best one. Only our test and validation sets were used for this. All the details of the optimisation and the selected hyperparameter values can be found in Data S1. Finally, unlike CNNs which process raster data as spatially structured inputs (i.e., 2D arrays), random RFs require vector-format inputs. To solve this problem and perform a fair comparison, we used the mean and the standard deviation of the pixels for each channel in RFs (Data S1) (Tables S4 and S5).

3 | Metrics

For the evaluation of the models predicting species occurrences, we used both the micro average and macro average accuracy, which are complementary (Equations 3 and 4). Micro average accuracy measured overall accuracy by considering all fish observations together, which favours the major or dominant species. Macro average accuracy instead calculates the average accuracy per species, giving equal weight to each species, whatever the number of fish observations. For both average accuracies, we reported the Top 1, Top 5, Top 10, and Top 20 values. The Top-k accuracy is the proportion of cases where the right species are within the first k species predicted by the model (the k species with the highest prediction probabilities).

$$\text{Micro average accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FT} + \text{FN}} \tag{3}$$

$$\text{Macro average accuracy} = \frac{1}{n}\sum_{i=1}^{\text{n}} \frac{\text{TP}_i + \text{TN}_i}{\text{TP}_i + \text{TN}_i + \text{FP}_i + \text{FN}_i} \tag{4}$$

where TP = True Positives, TN = True Negatives, FP = False Positives, FN = False Negatives, n = number of classes (here species), and i = for class i.

For the evaluation of the fish abundance models, we use the D-squared score function on the log-transformed data ($D2\log$) and the R-squared regression score function on the log-transformed data ($R2\log$) (Equations 5 and 6), estimating the error in abundance predictions. We used the $R2\log$ because it is a standard metric in ecology, but it is more sensitive to outliers and more prone to instability caused by extreme values (which is the case here) than the $D2\log$ (Willmott and Matsuura 2005). These two metrics have values ranging from -∞ to 1; the closer the value is to 1, the fewer errors in the predictions (the perfect model is at 1). We also evaluated the extent to which the ranking of species abundances (the order in which species are ranked on a site according to their abundance) is well predicted by the model for each site (independently of abundance prediction errors) using the Spearman rank-order coefficient (Spearman coefficient) (Equation 7). This metric has values ranging from −1 to 1; the closer the value is to 1, the better species ranking is respected (the perfect model is at 1).

$$D2\log = 1 - \frac{\sum_{i=1}^{n} \left|\log(y_i + 1) - \log(\hat{y}_i + 1)\right|}{\sum_{i=1}^{n} \left|\log(y_i + 1) - \log(\tilde{y} + 1)\right|} \tag{5}$$

$$R2\log = 1 - \frac{\sum_{i=1}^{n} \left(\log(y_i + 1) - \log(\hat{y}_i + 1)\right)^2}{\sum_{i=1}^{n} \left(\log(y_i + 1) - \log(\tilde{y} + 1)\right)^2} \tag{6}$$

where n = number of species abundance counts, y_i = true abundance of species i in a given site, $\hat{y}_i$ = predicted abundance of species i in a given site, $\tilde{y}$ = median true abundance of species, and $\bar{y}$ = mean of true abundance of species.

$$\text{Spearman coefficient} = 1 - \frac{6\sum_{i=1}^{n} \left(u_i - \hat{u}_i\right)^2}{n\left(n^2 - 1\right)} \tag{7}$$

where n = number of species abundance counts, u_i = rank from smallest to largest of the i^{th} true abundance of species (in all true species abundances) and $\hat{u}_i$ = rank from smallest to largest of the i^{th} predicted abundance of species (in all predicted abundance of species). Equal species abundances are assigned to the mean rank for their positions.

3.1 | Statistical Tests

For each metric, we first performed a Welch's ANOVA via Scipy (Virtanen et al. 2020) to check whether significant differences were present between the abundance models. We use a Welch's ANOVA because the data are heteroscedastic but are approximately normally distributed (Table S6) (Liu 2015). For metrics with a $p < 0.05$ in Welch's ANOVA, we then performed a Dunn's test with Scikit Posthocs (Terpilowski 2019) to assess in which models these differences were significant ($p < 0.05$). In addition to Dunn's test, we performed a pairwise permutation test, which does the same thing but is more reliable and flexible on small datasets.

4 | Results

4.1 | CNN-SDMs Versus RFs for Fish Occurrences

For the CNN-SDM predicting fish occurrences, the training was chaotic during the first epochs, but became more regular as the number of epochs grew and the learning rate decreased (Figure S1). We saved the layer weights corresponding to the best model on the validation set before overfitting, in this case, epoch 9. These layer weights were used for the final results (Table 2) and transfer learning.

Regarding macro average accuracy, the CNN-SDM predicting fish occurrences clearly outperformed RF predictions. Although similar for Top 1, the macro average accuracy was at least 1.5

times higher for Top 5, Top 10, and Top 20 species (Table 2). In terms of micro average accuracy, the CNN-SDM predicting fish occurrences performed slightly better (+0.00 to +0.02) than the RF for Top 1, Top 5, Top 10, and Top 20 species (Table 2).

4.2 | CNN-SDMs With and Without Transfer Learning for Fish Abundances

The CNN-SDM predicting fish abundances with transfer learning consistently outperformed those of the same model without transfer learning for the three performance metrics, whatever the fold (Figure 2 and Table S7). The average gain obtained with transfer learning was 35% for the *D*2log with a maximum gain of 82%, 15% for the Spearman coefficient with a maximum gain of 41%, and 72% for the *R*2log with a maximum gain of 171%. In terms of performance homogeneity between folds, the standard deviation of the three metrics was divided between 2 and 3 with transfer learning (Figure 2). The differences between these two models for each of the three metrics were significant. For each metric, this is supported by Welch's ANOVA, followed by Dunn's test and by the pairwise permutation test (Table S6).

TABLE 2 | Performance of the CNN-SDM and of the Random Forest model on the presence-only fish occurrence test set.

	Random forest	CNN-SDM
Micro average accuracy		
Top 1	0.06	0.06
Top 5	0.25	0.26
Top 10	0.43	0.45
Top 20	0.65	0.67
Macro average accuracy		
Top 1	0.02	0.02
Top 5	0.08	0.12
Top 10	0.13	0.22
Top 20	0.25	0.42

4.3 | CNN-SDMs With Transfer Learning Versus RFs for Fish Abundances

The CNN-SDM predicting fish abundances with transfer learning outperformed RF predictions for both the *D*2log and the Spearman coefficient whatever the fold, except for fold 17 where it was a little less efficient (Figure 2 and Table S8). More precisely, the average gain of the CNN-SDM predicting fish abundances with transfer learning compared to RF was 10% for the *D*2log with a maximum gain of 23%, and 4% for the Spearman coefficient with a maximum gain of 8%. On the contrary, *R*2log values were always better for the RF model predicting fish abundances than for the CNN-SDM with transfer learning (Figure 2 and Table S8). The differences between these two models for each of the three metrics were significant. For each metric, this is supported by Welch's ANOVA, followed by Dunn's test and by the pairwise permutation test (Table S6).

The mean *D*2log values per species over the 20 folds showed that the CNN-SDM predicting fish abundances with transfer learning performed better than the RF when species were present on less than 30% of the total fish abundance counts (Figure 3). Using predicted fish abundance maps, we also find that CNN-SDM with transfer learning was better where widespread species are locally rare (Figure 4). The abundance prediction maps for both models for each species are available in Data S2.

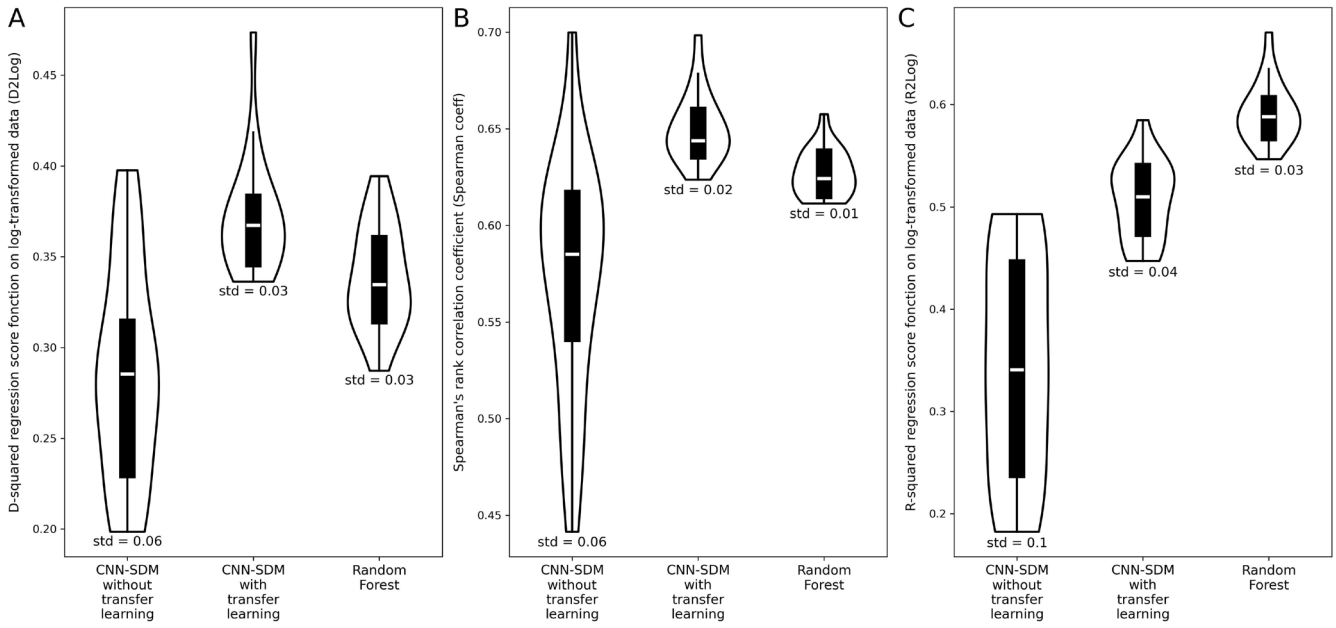

FIGURE 2 | Violin plots showing models' performances on the fish abundance counts for test sets over the 20 folds for (A) the D-squared regression score function on the log-transformed data (*D*2log), (B) the Spearman rank-order coefficient (Spearman coefficient) and (C) the R^2 regression score function on log-transformed data (*R*2log). For these three metrics, the closer the value is to 1, the better the model. Std = Standard deviation.

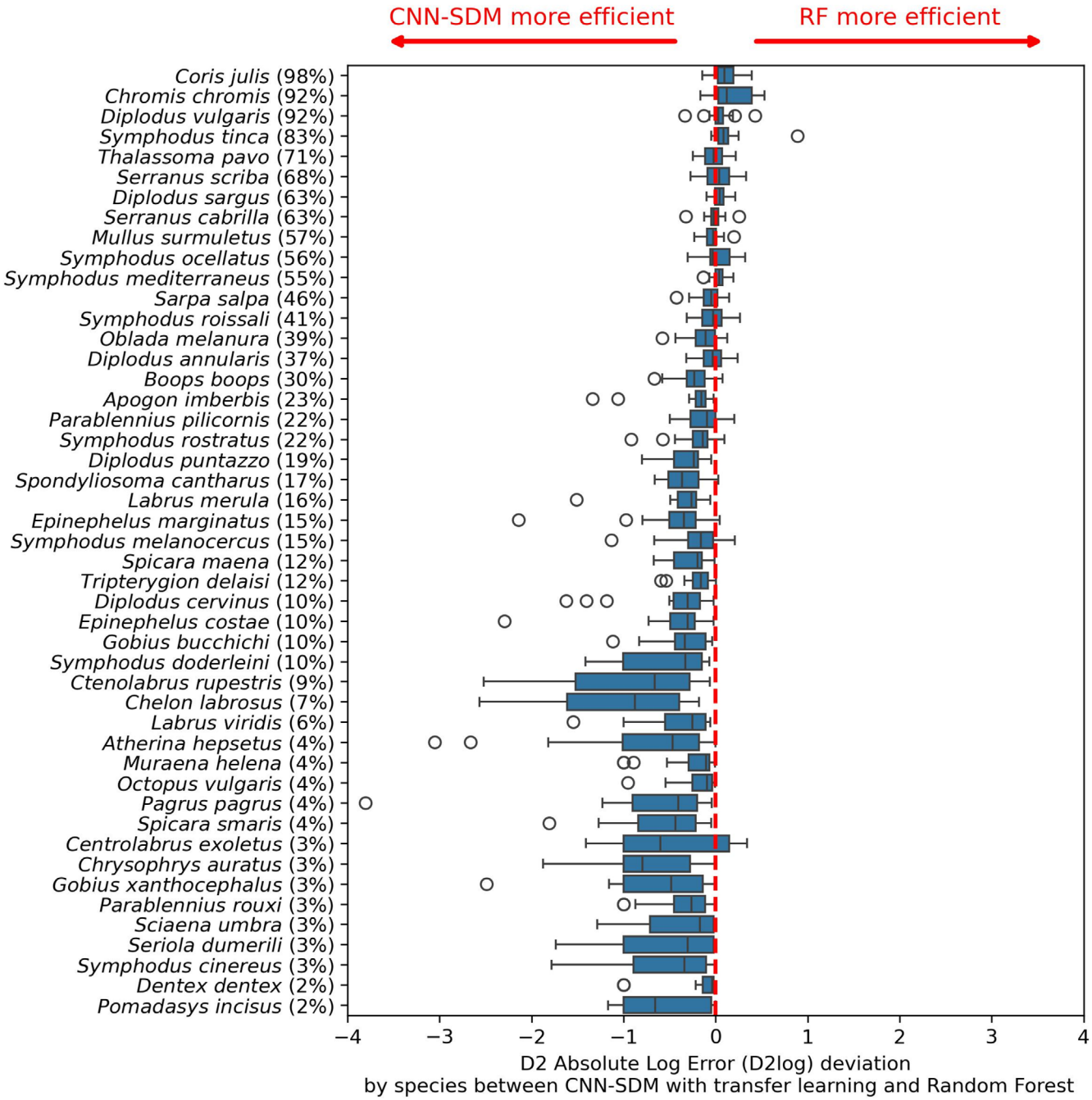

FIGURE 3 | *D*-squared regression score function on the log-transformed data (*D*2Log) deviation by species between the CNN-SDM with transfer learning and the Random Forest calculated for each of the 20 folds. Here, $D2\log\ \text{deviation} = \left(D2\log_A - 1\right) - \left(D2\log_B - 1\right)$ because $D2\log \in]-\infty;1]$. The percentage in brackets next to the name of each species indicates the percentage of fish abundance counts on which the species is present for the 406 fish abundance counts of Reef Life Surveys.

5 | Discussion

5.1 | Benefits of CNN-SDMs Over RFs to Predict Rare Species Occurrences

The relative performances of the CNN-SDM and RF predicting species occurrences are similar in terms of micro average accuracy, but the CNN-SDM is more efficient than the RF in terms of macro average accuracy (Table 2). To interpret these results, it is necessary to bear in mind that the fish occurrence dataset used in this study is highly imbalanced. It contains occurrences of 181 species, but the 13 most frequent species account for 50% of total occurrences (Table S9). This challenge of class imbalance is almost always present in all biodiversity datasets due to the natural hyper-dominance of some species (Callaghan et al. 2023; ter Steege et al. 2013) and the rarity of many others with long-tailed distributions (Beery et al. 2020; Enquist et al. 2019) but also sampling biases linked to socioeconomic factors (Troudet et al. 2017) or fishing gears (Mbaru et al. 2020). In the context of imbalanced datasets, the most common species (over-represented classes) have a disproportional influence on the micro average accuracy, while the influence of rare species

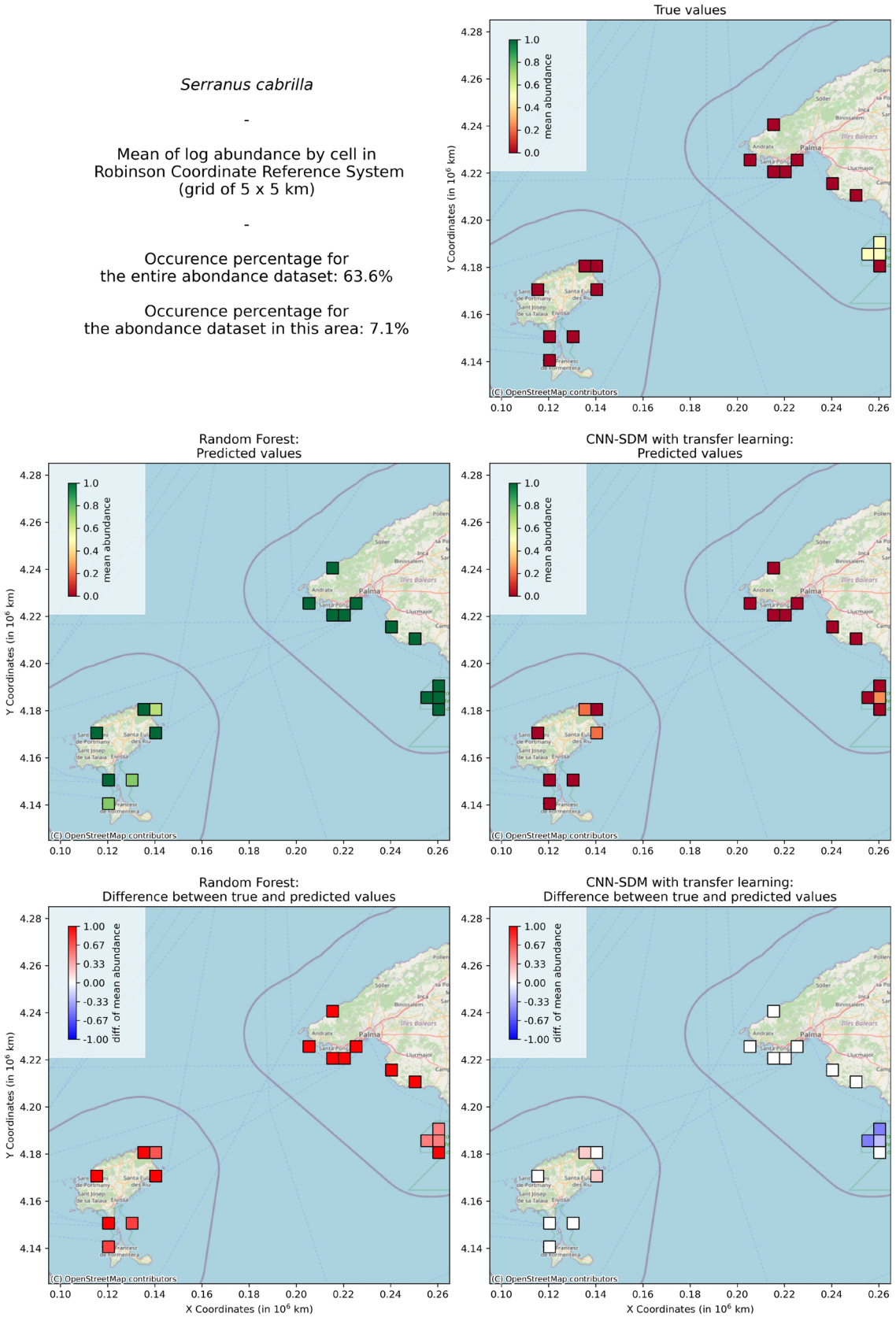

FIGURE 4 | Maps of Formentera, Ibiza, and Mallorca showing the average abundance (number of fish) of *Serranus cabrilla* over 5 km^2 areas between 2011 and 2022 for the true abundances, the abundances predicted by the CNN-SDM with transfer learning, the abundances predicted by the RF, and the difference between predictions and true values for the two models. The predicted abundances for each fish abundance count are based on the average of the predicted values across the 20 test folds, with the predicted abundance values in each fold first rounded up to the nearest value with a threshold of 0.05. Occurrence percentage: The percentage of fish abundance counts on which the species is present.

is negligible (underrepresented classes). In contrast, the rare and common species have an equivalent impact on the macro average accuracy (Estopinan et al. 2022). Our results show that CNN-SDMs perform as well as the RF for predicting occurrences of common species, but that CNN-SDMs perform much better for predicting rare species occurrences. This aspect is critical since rare species are the main targets of conservation strategies like protected areas (Sanchez et al. 2024) and can be essential for ecosystem functioning (Soliveres et al. 2016) and contributions to human societies (Dee et al. 2019).

These results are novel in the marine domain and are in line with the few studies carried out in the continental domain. For example, in predicting orchid species occurrences worldwide, Estopinan et al. (2022) highlight the better performance of CNN-SDMs for rare species. Complementarily, Deneu, Servajean, et al. (2021) conclude that, when comparing four models (punctual Deep Neural Networks, Boosted Trees, RF, and CNN-SDM) for predicting plant occurrences, the CNN-SDM performs the best for rare species, up to twice better than the RF. In our study, the performance gain in Top 5, 10, and 20 macro average accuracy provided by the CNN-SDM over the RF is of comparable magnitude (Table 2). Deneu, Servajean, et al. (2021) demonstrate that the ability of CNN-SDMs to associate species occurrences with the spatial structuring of their local landscape is a significant advantage over other SDMs for predicting the distribution of rare species. Similarly, the ability of CNN-SDMs to capture and reuse biotic associations in the context of multi-class SDM is also probably an advantage over classical SDMs for this task (Chen et al. 2017).

5.2 | Effectiveness of Transfer Learning for CNN-SDMs Predicting Species Abundances

The important variations in the metrics of CNN-SDM predicting species abundance without transfer learning between the folds (Table S7) are linked to the small amount of data to be split. Transfer learning induces a significant reduction in this phenomenon for the three performance metrics (Figure 2). In addition to this regularisation of performance, compared to the same model without transfer learning, the performance gains of the CNN-SDM predicting species abundances with transfer learning are significant for both absolute abundance values and the ranking of abundances among species. This improvement in the prediction of species abundances can be seen with the significant increase in $D2\log$ and $R2\log$, and the improvement in the ranking of species abundances through the increase in the Spearman coefficient.

A large field of deep learning applications successfully uses transfer learning (Do et al. 2022; Karaman et al. 2021; Nawaz et al. 2023; Qasim et al. 2022; Schwessinger et al. 2020; Yeh et al. 2020), but, to our knowledge, this is the first time that this approach is applied to SDMs predicting species abundances. With our CNN-SDMs, we show that learning patterns from presence-only species data can deeply improve predictions of species abundances. This paves the way for a new generation of SDMs pre-trained with large species occurrence datasets and transferred to tasks with smaller available datasets, like species abundances in the marine realm.

5.3 | Advantages of CNN-SDMs With Learning Transfer Over Traditional Approaches for Abundance Prediction

The $D2\log$ indicates that the CNN-SDM predicting species abundances with transfer learning performs better than the RF on the same task with the same data split and folds, but the $R2\log$ indicates the opposite. This can be easily explained by the fact that (i) the $R2\log$ is much more sensitive to outliers and extreme values in terms of species abundances than the $D2\log$ and (ii) the RF predicting species abundances is slightly better on the common species, although weaker for rare species (Figure 3). It is also important to note that in areas where a widespread species is locally rare, the CNN-SDM with transfer learning performs better (Figure 4). Yet, some of these species are present in shoals up to tens of thousands of individuals, so their counts are extremely approximate and the resulting errors are generally very large (Edgar et al. 2020). Moreover, some sites surveyed twice on the same day host, for example at the site CAT11, either 395 *Chromis chromis* individuals or 2838 individuals (*cf.* Survey ID n°912349329 and n°912349330 in Table S2) so one diver has surely seen a shoal and the other has not. The Spearman coefficient is also consistent with this and $D2\log$, showing that the CNN-SDM predicting species abundances with transfer learning more closely respects the general ranking in species abundances than the RF does (Figure 2). This is a remarkable result in terms of ecological interpretation and potential applications in conservation.

However, it should be noted that these advantages come at a cost in terms of computing time and energy. It is an approximation, but if all the calculations had been carried out on the same computer, the whole process of training our CNN-SDM with transfer learning would have required 30 times more time than the RF. The quality of the source data used for transfer learning (here, presence-only) is also a limitation. A model trained on poor-quality data may transfer or amplify errors to the target model (here, abundances), resulting in 'negative transfer' that can degrade performance. It is therefore recommended to compare performance with and without transfer, and to consider specific methods to limit this risk when the quality of data sources are known to be low (De et al. 2020). Finally, our approach does not correct the imbalance between classes. Models may therefore exhibit biases in favour of underrepresented species. This is a problem that affects both CNNs and RFs and therefore should not significantly impact the model comparison.

6 | Conclusion

This study produces several novel results that extend the range of applications and the predictive capacity of CNN-SDMs in ecology towards a new generation of SDMs based on deep learning (deep-SDMs). We demonstrate that these deep-SDMs can significantly improve species abundance predictions even on limited abundance datasets by applying a transfer learning from CNN-SDMs trained on large presence-only species datasets. This transfer learning compensates for the small size of abundance datasets, which is a classical limitation in ecology when using deep learning methods (Hu et al. 2025), by taking advantage of massive occurrence datasets in presence-only. This

ability of deep-SDMs to extract relevant information predicting species abundances from presence-only data is a new result. This finding paves the way towards a more general use of deep learning to better predict local population sizes through space and time, which remains challenging (Lertzman-Lepofsky et al. 2025), and to feed indicators like the Red List Index (Dulvy et al. 2024) or the Living Planet Index (Dove et al. 2023). Yet, additional work is needed at the interface between computer science and ecology to better comprehend how the information extracted from presence-only models can be successfully used by species abundance models.

Author Contributions

B.B., A.J., M.S., and D.M. designed the study. J.A.S.F. organised the collection of some of the data used in this article (Spanish data from the Reef Life Survey programme). B.B. and S.B. searched, extracted, and formatted the data from online databases for the analyses. B.B. and M.S. performed the analyses. B.B. wrote the first draft of the manuscript, and all the authors contributed substantially to the revisions.

Acknowledgements

This work has been mainly funded by the IA-Biodiv ANR programme—FISH-PREDICT project (ANR-21-AAFI-0001-01) and by the Biodiversa+programme—BioBoost+ project (EU grant agreement 101052342). It was also partially funded by the European Union's Horizon research and innovation programme under grant agreement n° 101060639 (MAMBO project) and n° 101060693 (GUARDEN project). This work was granted access to the HPC resources of IDRIS under the allocation 2022-AD011013891 made by GENCI. JAS-F was supported by a Junta de Andalucía Postdoctoral Fellowship (DGP_POST_2024_00757).

Data Availability Statement

The raw data on the occurrences of fish in presence-only are available in GBIF here: https://doi.org/10.15468/dl.pc3csa and https://doi.org/10.15468/dl.srq7cd. Due to the evolution of the "GBIF Backbone Taxonomy", it is possible that the GBIF web page indicates an inconsistent number of included occurrences in the download. If this is the case, please ignore it and download the data normally because the DarwinCore archive is permanent. The raw data on the fish abundance counts are available here: https://ecology.usegalaxy.eu/published/history?id=fa3c3b177fc40300. The cleaned raw data used in the study are presented in Table S1 for fish occurrences and Table S2 for fish abundance counts. The sources of each environmental data used in this study are specified in Table 1 in the "Source of data" column. The environmental data cited above were used to create rasters centred on the GPS positions of the associated occurrence or count sites for training our species distribution models. These rasters, the codes, and all other data needed to reproduce the results of species distribution models are available on GitHub in this repository: https://github.com/Beniofh/CNN_SDM_and_RF_for_Fish_2024.

Peer Review

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/ele.70177.

References

Beery, S., Y. Liu, D. Morris, et al. 2020. Synthetic Examples Improve Generalization for Rare Classes 2020 IEEE Winter Conference on Applications of Computer Vision (WACV). Presented at the 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, Snowmass Village, CO, USA, pp. 852–862.

Botella, C., A. Joly, P. Bonnet, P. Monestiez, and F. Munoz. 2018. "A Deep Learning Approach to Species Distribution Modelling." In *Multimedia Tools and Applications for Environmental & Biodiversity Informatics, Multimedia Systems and Applications*, edited by A. Joly, S. Vrochidis, K. Karatzas, A. Karppinen, and P. Bonnet, 169–199. Springer International Publishing.

Bradley, B. A. 2016. "Predicting Abundance With Presence-Only Models." *Landscape Ecology* 31: 19–30.

Breiner, F. T., A. Guisan, A. Bergamini, and M. P. Nobis. 2015. "Overcoming Limitations of Modelling Rare Species by Using Ensembles of Small Models." *Methods in Ecology and Evolution* 6: 1210–1218.

Brun, P., D. N. Karger, D. Zurell, et al. 2024. "Multispecies Deep Learning Using Citizen Science Data Produces More Informative Plant Community Models." *Nature Communications* 15: 4421.

Callaghan, C. T., L. Borda-de-Água, R. van Klink, R. Rozzi, and H. M. Pereira. 2023. "Unveiling Global Species Abundance Distributions." *Nature Ecology & Evolution* 7: 1600–1609.

Ceballos, G., P. R. Ehrlich, and P. H. Raven. 2020. "Vertebrates on the Brink as Indicators of Biological Annihilation and the Sixth Mass Extinction." *Proceedings of the National Academy of Sciences of the United States of America* 117: 13596–13602.

Chaikin, S., F. Riva, K. E. Marshall, J.-P. Lessard, and J. Belmaker. 2024. "Marine Fishes Experiencing High-Velocity Range Shifts May Not Be Climate Change Winners." *Nature Ecology & Evolution* 8: 936–946.

Chardon, N. I., J. Nabe-Nielsen, J. J. Assmann, et al. 2022. "High Resolution Species Distribution and Abundance Models Cannot Predict Separate Shrub Datasets in Adjacent Arctic Fjords." *Diversity and Distributions* 28: 956–975.

Chen, D., Y. Xue, S. Chen, D. Fink, and C. Gomes. 2017. "Deep Multi-Species Embedding."

Copernicus Sentinel-2 Data. 2023. "Copernic. Sentin.-2 Data Process. ESA Modif. Sentin.-Hub." https://www.sentinel-hub.com. Accessed 4 April 2023.

De, S., J. Britton, M. Reynolds, R. Skinner, K. Jansen, and A. Doostan. 2020. "On Transfer Learning of Neural Networks Using Bi-Fidelity Data for Uncertainty Propagation."

Dee, L. E., J. Cowles, F. Isbell, S. Pau, S. D. Gaines, and P. B. Reich. 2019. "When Do Ecosystem Services Depend on Rare Species?" *Trends in Ecology & Evolution* 34: 746–758.

Deneu, B., A. Joly, P. Bonnet, M. Servajean, and F. Munoz. 2021. "How Do Deep Convolutional SDM Trained on Satellite Images Unravel Vegetation Ecology?" In *Pattern Recognition. ICPR International Workshops and Challenges*, Lecture Notes in Computer Science, edited by A. Del Bimbo, R. Cucchiara, S. Sclaroff, et al., 148–158. Springer International Publishing.

Deneu, B., M. Servajean, P. Bonnet, C. Botella, F. Munoz, and A. Joly. 2021. "Convolutional Neural Networks Improve Species Distribution Modelling by Capturing the Spatial Structure of the Environment." *PLoS Computational Biology* 17: e1008856.

Do, S., É. Ollion, and R. Shen. 2022. "The Augmented Social Scientist: Using Sequential Transfer Learning to Annotate Millions of Texts With Human-Level Accuracy." *Sociological Methods & Research* 53: 1167–1200.

Dove, S., M. Böhm, R. Freeman, L. McRae, and D. J. Murrell. 2023. "Quantifying Reliability and Data Deficiency in Global Vertebrate Population Trends Using the Living Planet Index." *Global Change Biology* 29: 4966–4982.

Dulvy, N. K., N. Pacoureau, J. H. Matsushiba, et al. 2024. "Ecological Erosion and Expanding Extinction Risk of Sharks and Rays." *Science* 386: eadn1477.

Edgar, G. J., A. Cooper, S. C. Baker, et al. 2020. "Establishing the Ecological Basis for Conservation of Shallow Marine Life Using Reef Life Survey." *Biological Conservation* 252: 108855.

Edgar, G. J., R. D. Stuart-Smith, F. J. Heather, et al. 2023. "Continent-Wide Declines in Shallow Reef Life Over a Decade of Ocean Warming." *Nature* 615: 858–865.

EMODnet Bathymetry Consortium. 2020. "EMODnet Digit. Bathymetry DTM 2020." https://doi.org/10.12770/bb6a87dd-e579-4036-abe1-e649cea9881a.

Enquist, B. J., X. Feng, B. Boyle, et al. 2019. "The Commonness of Rarity: Global and Future Distribution of Rarity Across Land Plants." *Science Advances* 5: eaaz0414.

Estopinan, J., M. Servajean, P. Bonnet, F. Munoz, and A. Joly. 2022. "Deep Species Distribution Modeling From Sentinel-2 Image Time-Series: A Global Scale Analysis on the Orchid Family." *Frontiers in Plant Science* 13: 839327.

E.U. Copernicus Marine Service Information. 2023a. "Mediterr. Sea Ocean Colour Plankton MY L4 Dly." *Gapfree Obs. Climatol. Mon. Obs.* https://doi.org/10.48670/moi-00300.

E.U. Copernicus Marine Service Information. 2023b. "Mediterr. Sea High Resolut." Ultra High Resolut. Sea Surf. Temp. Anal. https://doi.org/10.48670/moi-00172.

E.U. Copernicus Marine Service Information. 2023c. "Mediterrian Sea Physics Reanalysis." https://doi.org/10.25423/CMCC/MEDSEA_MULTIYEAR_PHY_006_004_E3R1.

Finn, C., F. Grattarola, and D. Pincheira-Donoso. 2023. "More Losers Than Winners: Investigating Anthropocene Defaunation Through the Diversity of Population Trends." *Biological Reviews* 98: 1732–1748.

Galaxy for Ecology. 2022. "Reef Life Survey Mediterrian Sea." Accessed 29 September 2023. https://ecology.usegalaxy.eu/published/history?id=fa3c3b177fc40300.

GBIF. 2023. *Glob. Biodivers. Inf. Facil. Intergov. Oceanogr. Comm.* UNESCO. Accessed 20 October 2023. https://www.gbif.org.

GBIF Dataset. 2022. "GBIF Occur. Download." https://doi.org/10.15468/dl.pc3csa.

GBIF Dataset. 2023. "GBIF Occur. Download." https://doi.org/10.15468/dl.srq7cd.

Gupta, J., S. Pathak, and G. Kumar. 2022. "Deep Learning (CNN) and Transfer Learning: A Review." *Journal of Physics Conference Series* 2273: 12029.

He, K., X. Zhang, S. Ren, and J. Sun. 2016. Deep Residual Learning for Image Recognition Presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.

Hu, Y., S. Si-Moussi, and W. Thuiller. 2025. "Introduction to Deep Learning Methods for Multi-Species Predictions." *Methods in Ecology and Evolution* 16: 228–246.

Karaman, O., H. Çakın, A. Alhudhaif, and K. Polat. 2021. "Robust Automated Parkinson Disease Detection Based on Voice Signals With Transfer Learning." *Expert Systems with Applications* 178: 115013.

Lertzman-Lepofsky, G., A. J. Dolezal, M. T. Waters, et al. 2025. "Temporal Changes in Taxon Abundances Are Positively Correlated but Poorly Predicted at the Global Scale." *Ecography* 2025: e07195.

Liu, H. 2015. "Comparing Welch's ANOVA, a Kruskal-Wallis Test and Traditional ANOVA in Case of Heterogeneity of Variance." Theses Diss.

Mbaru, E. K., N. A. J. Graham, T. R. McClanahan, and J. E. Cinner. 2020. "Functional Traits Illuminate the Selective Impacts of Different Fishing Gears on Coral Reefs." *Journal of Applied Ecology* 57: 241–252.

Nawaz, U., U. A. Raza, A. Farooq, M. J. Iqbal, and A. Tariq. 2023. "Voice Cloning Using Transfer Learning With Audio Samples." *UMT Artificial Intelligence Review* 3: 65–77.

Nowakowski, J. A., J. I. Watling, A. Murray, et al. 2023. "Protected Areas Slow Declines Unevenly Across the Tetrapod Tree of Life." *Nature* 622: 101–106.

OBIS. 2024. "Ocean Biodiversity Information System." Accessed 28 March 2024. www.obis.org.

Paszke, A., S. Gross, F. Massa, et al. 2019. "PyTorch: An Imperative Style, High-Performance Deep Learning Library." In *Advances in Neural Information Processing Systems*. Curran Associates, Inc.

Pearce, J. L., and M. S. Boyce. 2006. "Modelling Distribution and Abundance With Presence-Only Data." *Journal of Applied Ecology* 43: 405–412.

Pedregosa, F., G. Varoquaux, A. Gramfort, et al. 2011. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research* 12: 2825–2830.

Phillips, S. J., R. P. Anderson, and R. E. Schapire. 2006. "Maximum Entropy Modeling of Species Geographic Distributions." *Ecological Modelling* 190: 231–259.

Phillips, S. J., M. Dudík, and R. E. Schapire. 2004. "A Maximum Entropy Approach to Species Distribution Modeling." In *Proceedings of the Twenty-First International Conference on Machine Learning, ICML '04*, 83. Association for Computing Machinery.

Pollock, H. S., J. D. Toms, C. E. Tarwater, T. J. Benson, J. R. Karr, and J. D. Brawn. 2022. "Long-Term Monitoring Reveals Widespread and Severe Declines of Understory Birds in a Protected Neotropical Forest." *Proceedings of the National Academy of Sciences of the United States of America* 119: e2108731119.

Qasim, R., W. H. Bangyal, M. A. Alqarni, and A. Ali Almazroi. 2022. "A Fine-Tuned BERT-Based Transfer Learning Approach for Text Classification." *Journal of Healthcare Engineering* 2022: e3498123.

Raiter, K. G., and D. Hawlena. 2024. "Managing Multiple Uncertainties in Species Distribution Modelling." *Diversity and Distributions* 30: e13857.

Rigal, S., V. Dakos, H. Alonso, et al. 2023. "Farmland Practices Are Driving Bird Population Decline Across Europe." *Proceedings of the National Academy of Sciences of the United States of America* 120: e2216573120.

Roberts, D. R., V. Bahn, S. Ciuti, et al. 2017. "Cross-Validation Strategies for Data With Temporal, Spatial, Hierarchical, or Phylogenetic Structure." *Ecography* 40: 913–929.

Sanchez, L., N. Loiseau, G. J. Edgar, et al. 2024. "Rarity Mediates Species-Specific Responses of Tropical Reef Fishes to Protection." *Ecology Letters* 27: e14418.

Schwessinger, R., M. Gosden, D. Downes, et al. 2020. "DeepC: Predicting 3D Genome Folding Using Megabase-Scale Transfer Learning." *Nature Methods* 17: 1118–1124.

Soliveres, S., P. Manning, D. Prati, et al. 2016. "Locally Rare Species Influence Grassland Ecosystem Multifunctionality." *Philosophical Transactions of the Royal Society, B: Biological Sciences* 371: 20150269.

Spalding, M. D., H. E. Fox, G. R. Allen, et al. 2007. "Marine Ecoregions of the World: A Bioregionalization of Coastal and Shelf Areas." *Bioscience* 57: 573–583.

Srivastava, V., V. Lafond, and V. C. Griess. 2019. "Species Distribution Models (SDM): Applications, Benefits and Challenges in Invasive Species Management." *CABI Reviews* 2019: 1–13.

Taylor, M. E., and P. Stone. 2009. "Transfer Learning for Reinforcement Learning Domains: A Survey." *Journal of Machine Learning Research* 10: 1633–1685.

ter Steege, H., N. C. A. Pitman, D. Sabatier, et al. 2013. "Hyperdominance in the Amazonian Tree Flora." *Science* 342: 1243092.

Terpilowski, M. A. 2019. "Scikit-Posthocs: Pairwise Multiple Comparison Tests in Python." *Journal of Open Source Software* 4: 1169.

Troudet, J., P. Grandcolas, A. Blin, R. Vignes-Lebbe, and F. Legendre. 2017. "Taxonomic Bias in Biodiversity Data and Societal Preferences." *Scientific Reports* 7: 9132.

van Klink, R., D. E. Bowler, K. B. Gongalsky, M. Shen, S. R. Swengel, and J. M. Chase. 2024. "Disproportionate Declines of Formerly Abundant Species Underlie Insect Loss." *Nature* 628: 359–364.

Vasquez, M., H. Allen, E. Manca, et al. 2021. "EUSeaMap 2021. A European Broad-Scale Seabed Habitat Map (No. DOI 10.13155/83528). EMODnet."

Virtanen, P., R. Gommers, T. E. Oliphant, et al. 2020. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python." *Nature Methods* 17: 261–272.

Waldock, C., R. D. Stuart-Smith, C. Albouy, et al. 2022. "A Quantitative Review of Abundance-Based Species Distribution Models." *Ecography*: 2022.

Wang, J., and S. Tabeta. 2023. "Four-Channel Generative Adversarial Networks Can Predict the Distribution of Reef-Associated Fish in the South and East China Seas." *Ecological Informatics* 78: 102321.

Weiss, K., T. M. Khoshgoftaar, and D. Wang. 2016. "A Survey of Transfer Learning." *Journal of Big Data* 3: 9.

Willmott, C., and K. Matsuura. 2005. "Advantages of the Mean Absolute Error (MAE) Over the Root Mean Square Error (RMSE) in Assessing Average Model Performance." *Climate Research* 30: 79–82.

Yeh, C., A. Perez, A. Driscoll, et al. 2020. "Using Publicly Available Satellite Imagery and Deep Learning to Understand Economic Well-Being in Africa." *Nature Communications* 11: 2583.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.