Overview of LifeCLEF location-based species prediction task 2020 (GeoLifeCLEF)

Benjamin Deneu[1,2], Titouan Lorieul[1], Elijah Cole[3], Maximilien Servajean[4], Christophe Botella[5], Pierre Bonnet[6], Alexis Joly[1]

[1] INRIA, UMR LIRMM, Univ Montpellier, France
[2] AMAP, Univ Montpellier, CIRAD, CNRS, INRAE, IRD, Montpellier, France
[3] Caltech, Pasadena, US
[4] LIRMM, Université Paul Valéry, University of Montpellier, CNRS, Montpellier, France
[5] CNRS, LECA, France
[6] CIRAD, UMR AMAP, F-34398 Montpellier, France

Abstract. Understanding the geographic distribution of species is a key concern in conservation. By pairing species occurrences with environmental features, researchers can model the relationship between an environment and the species which may be found there. To advance the state-of-the-art in this area, a large-scale machine learning competition called *GeoLifeCLEF 2020* was organized. It relied on a dataset of 1.9 million species observations paired with high-resolution remote sensing imagery, land cover data, and altitude, in addition to traditional low-resolution climate and soil variables. This paper presents an overview of the competition, synthesizes the approaches used by the participating groups, and analyzes the main results. In particular, we highlight the ability of remote sensing imagery and convolutional neural networks to improve predictive performance, complementary to traditional approaches.

Keywords: LifeCLEF, biodiversity, environmental data, species distribution, evaluation, benchmark, species distribution models, methods comparison, presence-only data, model performance, prediction, predictive power

1 Introduction

In order to make informed conservation decisions it is essential to understand where different species live. Citizen science projects now generate millions of geo-located species observations every year, covering tens of thousands of species. But how can these point observations be used to predict what species might be found at a new location?

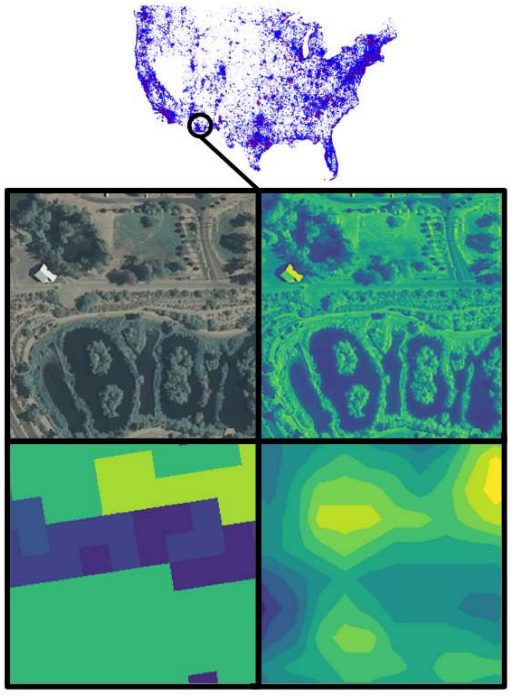


Fig. 1: Each species observation is paired with high-resolution covariates (clockwise from top left: RGB imagery, IR imagery, altitude, land cover).

A common approach is to build a *species distribution model* (SDM) [3], which uses a location's *environmental covariates* (e.g. temperature, elevation, land cover) to predict which species may be found there. Once trained, the model can be used to make predictions for any location where those covariates are available.

Developing an SDM requires a dataset where each species occurrence is paired with a collection of environmental covariates. However, many existing SDM datasets are both highly specialized and not readily accessible, having been assembled by scientists studying particular species or regions. In addition, the provided environmental covariates are typically coarse, with resolutions ranging from hundreds of meters to kilometers per pixel.

In this work, we present the results of the GeoLifeCLEF 2020 competition. This competition aimed at bridging these gaps by (i) sharing a large-scale dataset of observations paired with high-resolution covariates and (ii) defining a common evaluation methodology to measure the predictive performance of models trained on this dataset. The dataset is based on over 1.9 million observations of plant and animal species. Each observation is paired with high-resolution satellite imagery as well as traditional environmental covariates (e.g. climate, altitude and soil variables). To the best of our knowledge, this is the first publicly available dataset

to pair remote sensing imagery with species observations. Our hope is that this analysis-ready dataset and associated evaluation methodology will (i) make the SDM problem more accessible to machine learning researchers and (ii) facilitate novel research in large-scale, high-resolution, and remote-sensing-based species distribution modeling.

2 Dataset and Evaluation Protocol

Data collection: A detailed description of the GeoLifeCLEF 2020 dataset is provided in [1]. For completeness, we give a brief description here. The GeoLifeCLEF 2020 dataset consists of 1,921,123 observations from the US (1,097,640) and France (823,483) covering 31, 435 plant and animal species. Each species observation is paired with high-resolution covariates (RGB-IR imagery, land cover and altitude) as illustrated in Figure 1. These high-resolution covariates are resampled to a spatial resolution of 1 meter per pixel and provided as 256×256 images covering a 256m $\times$ 256m square centered on each observation. RGB-IR imagery come from the 2009-2011 cycle of the National Agriculture Imagery Program (NAIP) for the U.S.[1], and from the BD-ORTHO® 2.0 and ORTHO-HR® 1.0 databases from the IGN for France[2]. Land cover data originates from the National Land Cover Database (NLCD) [7] for the U.S. and from CESBIO[3] for France. All elevation data comes from the NASA Shuttle Radar Topography Mission (SRTM)[4]. In addition, the dataset also includes traditional coarser resolution covariates: 19 bio-climatic rasters (30arcsec^2/pixel (above 1km^2/pixel) from WorldClim [6]) and 8 pedologic rasters (250m^2/pixel, from SoilGrids [5]). The details on these rasters are given in Table 1.

Train-test split: The full set of occurrences was split in a training and testing set using a spatial block holdout procedure (see Figure 2). This limits the effect of *spatial auto-correlation* in the data as explained in [8]. This means that a model cannot achieve a high performance by simply interpolating between training samples. The split was based on a global grid of 5 km $\times$ 5 km quadrats. 2.5% of the quadrats were randomly sampled for the test set, and the remaining quadrats were assigned to the training set.

Evaluation metric: For each occurrence in the test set, the goal of the task was to return a candidate set of species with associated confidence scores. The main evaluation criterion is an adaptive variant of the top-K accuracy. Contrary to a classical top-K accuracy, this metric accounts for the fact that the number of species K may not be the same at each location. It is computed by thresholding the confidence score of the predictions and keeping only the species above that threshold. The threshold is determined automatically so as to have $K = 30$

[1] National Agriculture Image Program, https://www.fsa.usda.gov
[2] https://geoservices.ign.fr
[3] `http://osr-cesbio.ups-tlse.fr/~oso/posts/2017-03-30-carte-s2-2016/`
[4] `https://lpdaac.usgs.gov/products/srtmgl1v003/`

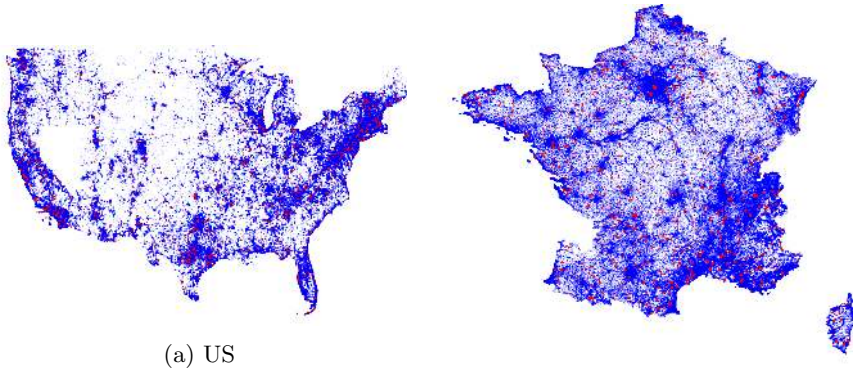
Name	Description	Resolution
bio_1	Annual Mean Temperature	30 arcsec
bio_2	Mean Diurnal Range (Mean of monthly (max temp - min temp))	30 arcsec
bio_3	Isothermality (bio_2/bio_7) (* 100)	30 arcsec
bio_4	Temperature Seasonality (standard deviation *100)	30 arcsec
bio_5	Max Temperature of Warmest Month	30 arcsec
bio_6	Min Temperature of Coldest Month	30 arcsec
bio_7	Temperature Annual Range (bio_5-bio_6)	30 arcsec
bio_8	Mean Temperature of Wettest Quarter	30 arcsec
bio_9	Mean Temperature of Driest Quarter	30 arcsec
bio_10	Mean Temperature of Warmest Quarter	30 arcsec
bio_11	Mean Temperature of Coldest Quarter	30 arcsec
bio_12	Annual Precipitation	30 arcsec
bio_13	Precipitation of Wettest Month	30 arcsec
bio_14	Precipitation of Driest Month	30 arcsec
bio_15	Precipitation Seasonality (Coefficient of Variation)	30 arcsec
bio_16	Precipitation of Wettest Quarter	30 arcsec
bio_17	Precipitation of Driest Quarter	30 arcsec
bio_18	Precipitation of Warmest Quarter	30 arcsec
bio_19	Precipitation of Coldest Quarter	30 arcsec
orcdrc	Soil organic carbon content (g/kg at 15cm depth)	250 m
phihox	Ph x 10 in H20 (at 15cm depth)	250 m
cecsol	cation exchange capacity of soil in cmolc/kg 15cm depth	250 m
bdticm	Absolute depth to bedrock in cm	250 m
clyppt	Clay (0-2 micro meter) mass fraction at 15cm depth	250 m
sltppt	Silt mass fraction at 15cm depth	250 m
sndppt	Sand mass fraction at 15cm depth	250 m
bldfie	Bulk density in kg/m3 at 15cm depth	250 m

Table 1: Summary of environmental variable rasters provided.

results per occurrence on average on the test set. Traditional top-K accuracy with $K = 30$ is used as secondary evaluation metric. See [1] for full details and justification.

Course of the challenge: The training data was publicly shared in early April 2020 through the AICrowd platform[5]. Any research team wishing to participate in the evaluation could register on the platform and download the data. The test data was shared a few weeks later but without the species labels, which were kept secret. Each team could then submit up to 10 submissions corresponding to different methods or different settings of the same method. A submission (also called a *run*) takes the form of a CSV file containing the predictions of the method being evaluated for all observations in the test set. For each submission, the evaluation metrics are computed and made visible to the participant. Once the submission phase was closed (mid-June), the participants could also see the evaluation metric values of the other participants. As a last important step, each

[5] `https://www.aicrowd.com/`



(a) US

(b) France

Fig. 2: Occurrences distribution over the US and France. Blue dots represent training data, red dots represent test data.

participant was asked to provide a *working note*, i.e. a detailed report containing all technical information required to reproduce the results of the submissions. All LifeCLEF *working notes* are reviewed by at least two members of LifeCLEF organizing committee to ensure a sufficient level of quality and reproducibility.

3 Participants and methods

40 participants registered for the GeoLifeCLEF 2020 challenge and downloaded the dataset. Only two teams succeeded in submitting results in the end: **Stanford** and **LIRMM**. A major obstacle to participation was the volume of data as well as the computing power needed to train a model. For instance, it took the LIRMM team almost two weeks to train a convolutional neural network on the full dataset using 8 GPUs. The details of the methods and systems used by the two participants are summarized below.

- **LIRMM**: This research team uploaded four submissions to the AICrowd platform but they reported problems for two of them afterwards, so we only report the correct ones here:
  - Submission 1 - Random forest trained on environmental feature vectors only (*i.e.* on the 27 climatic and soil variables).
  - Submission 3 - Convolution neural network trained on high-resolution image covariates (*i.e.* on 6-channel tensors composed of RGB-IR images, land cover image and altitude image).

  More information about the used methods can be found in the individual working note of this team [2].
- **Stanford**: This research team uploaded five submissions to the AICrowd platform. Most of the submissions were based on deep neural networks, but the authors informed us that they encountered convergence issues resulting

in performance poorer than expected. Thus, it was mutually agreed that they would not provide a working note describing their method in detail. Only one of their submissions (referred as Submission 3) was valid. This was a baseline method that always predicted the list of the most frequent species in the training set.

4 Global results of the evaluation

Participant	Submission #	Adaptive top-30 acc.	Top-30 acc.
LIRMM	Submission 3	23.3%	23.5%
LIRMM	Submission 1	21.3%	20.4%
Stanford	Submission 3	4.8%	4.8%

Table 2: GeoLifeCLEF 2020 main results, adaptive top-30 accuracy and top-30 accuracy per submission (sorted by decreasing adaptive top-30 accuracy).

In Table 2, we report the performance measured for each of the 8 submissions. The main outcome is that the method achieving the best results (LIRMM Submission 3) was based solely on a convolutional neural network (CNN) trained on the high-resolution covariates (RGB-IR imagery, land cover, and altitude). It did not make use of any bioclimatic or soil variables, which are often considered to be the most informative in the ecological literature. On the contrary, LIRMM Submission 1 was a machine learning method classically used for species distribution models [4] trained solely on the climatic and soil variables. This shows two things: (i) important information explaining the species composition is contained in the high-resolution covariates, and, (ii) convolutional neural networks are able to capture this information. The performance achieved by the baseline predictor of Stanford shows that the other methods are consistently better than just returning the most common species everywhere.

5 Complementary analysis

In this section we provide complementary analyses of the submitted results focusing on certain aspects of the dataset. In particular, we will consider the two main methods submitted which we denote:

- **RF (env.)**: the model from LIRMM Submission 1 consisting of a random forest trained solely on environmental variables;
- **CNN (high res.)**: the model from LIRMM Submission 3 consisting of a CNN trained on the high-resolution covariates.

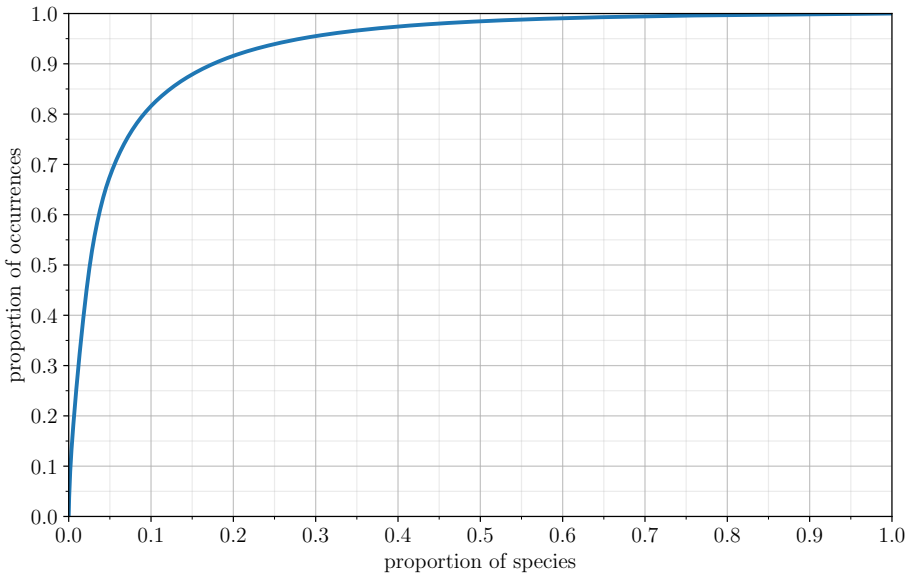


Fig. 3: Cumulative distribution of number of occurrences per species (ordered from most to least represented species in the training set) highlighting the long tail characteristic of the dataset.

Species-wise performance: First, we study the performance of the models depending on species frequencies. The dataset contains $31,435$ species distributed according to a long tail distribution which can be seen in Figure 3. For example, the curve shows that the most common 10% of species represent more than 80% of the occurrences. On the other hand, the least common 70% of species account for fewer than 5% of the occurrences. In such long-tailed contexts, metrics that are averaged over all observations can seem satisfactory while the model only performs well on a few common species. To evaluate model performance across all species, we computed the top-30 accuracy for each species independently and then averaged the resulting scores.

The results, presented in Figure 4, show that the top-30 accuracy averaged over the species is much lower than the one averaged over the occurrences. This is expected because the average over species gives much more weight to less represented species. However, it is interesting to compare how this affects each of the two models. The CNN drops from an accuracy of 23.5% to an accuracy of 13.2% with a 44% relative loss. The random forest falls from 20.4% to 6.9% with a 66% relative loss. RF (env.) is thus more affected by the average over species. To better understand this, Figure 5 shows the performance of the two models on each species of the test set ranked by their frequency in the training set (with an adaptive sliding average for display purposes). It shows that the

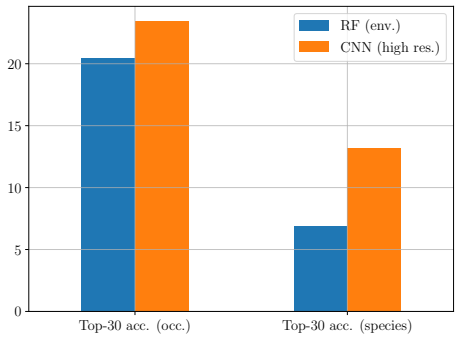


Fig. 4: Comparison of top-30 accuracy averaged over the occurrences with top-30 accuracy averaged over the species.

CNN outperforms the random forest on nearly every species. The random forest is slightly better for only the most frequent species.

Figure 6 shows a similar graph but instead of ranking the species by their frequency, species are ranked by the average top-30 accuracy achieved by the model. On this figure we can see how many species are predicted by the model with a top-30 accuracy over a given threshold. We can see in particular that for the CNN (high res.) model, $2,300$ species over the $7,794$ species are predicted with a top-30 accuracy greater than 0. In comparison, only $1,100$ have a top-30 accuracy greater than 0 when using the RF (env.) model. Moreover, we can observe that the CNN (high res.) model is able to perfectly predict (with a score of 1) more species than the RF (env.) model. The CNN (high res.) model perfectly predicts near 400 species of the test set whereas as the RF (env.) model only predicts 150 species perfectly.

Analysis by kingdom and geographic area: As the dataset covers France and US and contains both plants and animals species, it is interesting to compare the accuracy obtained by the models over these criteria. The resulting top-30 accuracy values are provided in Figure 7. Concerning the prediction of plants vs. animals, both models have similar results as shown in Figure 7a. Both have a slightly better prediction on plants, which can in part be explained by the greater number of occurrences per plant species than animals. Concerning the geographical area, Figure 7b shows that the performance is globally lower in France, in particular for animals where the performance is dramatically low. The Table 3 gives the average number of occurrences per species by kingdoms and regions. It can be noted that the very poor prediction performance on animals in France may be due to the low number of occurrences per animal species, on average 12.6. The lower performance on plants in France, however, is more difficult to explain since the average number of occurrences per species is pretty high (238). A possible interpretation could be that the lack of animal occurrences globally degrades the performance of the model, for instance because animals

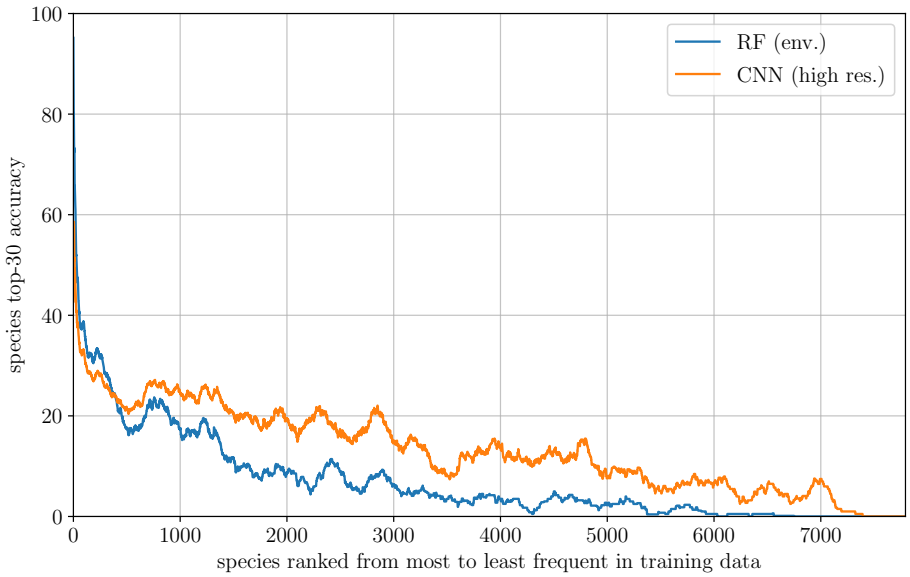


Fig. 5: Top-30 accuracy of the models over species with an adaptive sliding average. Species are ranked on the x-axis by their frequency in the training set.

could be stronger markers of the habitat. But this could be due to other reasons such as distribution of occurrences over species or to macro-ecological differences between US and France.

	Nb. of occurrences		**Nb. of species**		**Nb. of occ. per sp.**	
Kingdom	**US**	**France**	**US**	**France**	**US**	**France**
Plants	524,280	741,010	11,369	3,114	46.1	238.0
Animals	551,563	61,865	13,882	4,899	39.7	12.6
All	1,622,120	802,875	25,251	8,013	42.6	100.2

Table 3: Statistics of the number of occurrences, of species and of occurrences per species grouped by kingdoms and regions computed on the training set.

Fusion of the predictions of the high-resolution CNN and classical punctual environmental model: As discussed in previous sections, an important outcome of this evaluation is that the CNN (high res.) model performs better than the more classical RF (env.) model which means that the CNN (high res.) model is able to capture important information explaining the species distribution from the high-resolution covariates. Now, an important question is to what extent this information is complementary to the information extracted by

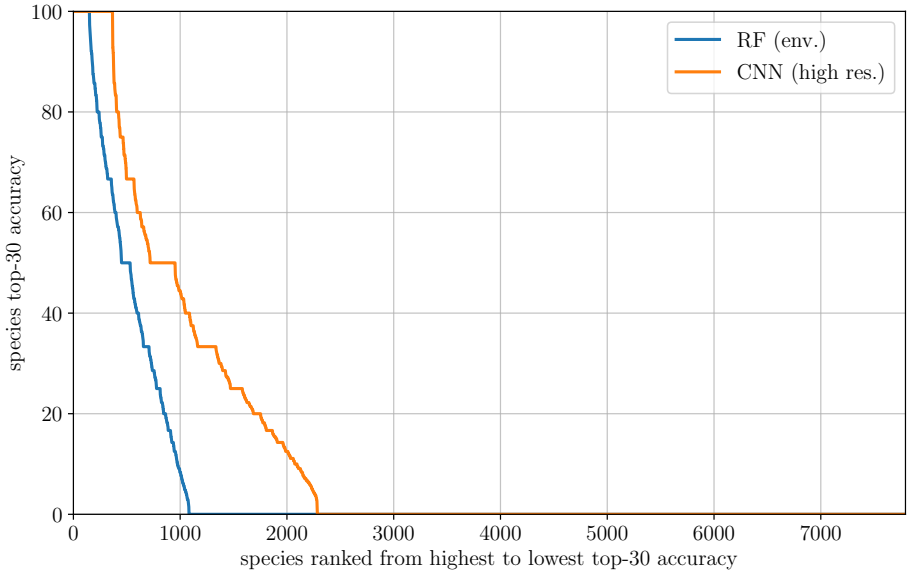
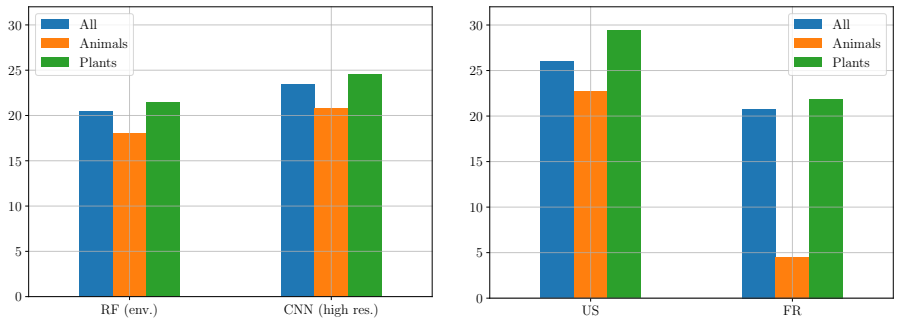


Fig. 6: Top-30 accuracy of the models over species. Species are ranked on the x-axis by the top-30 accuracy of the models on that species. First species is the best predicted by the model, last species is the worst predicted by the model. The two curves are represented on the same plot but the ranking is different for each model.



(a) Per model.

(b) Per country for CNN (high res.).

Fig. 7: Top-30 accuracy per kingdom and geographical area

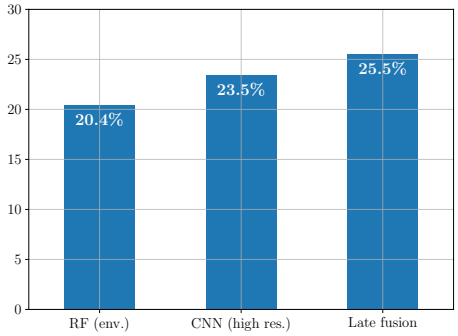


Fig. 8: Late fusion of RF (env.) and CNN (high res.) improves the global top-30 accuracy.

the RF (env.) model. The RF (env.) model is based solely on climatic and soil variables which are classically used in ecology to model the niche of the species, i.e. their environmental preferences. It is likely that these ecological preferences may also be partially inferred by the CNN (high res.) from the high-resolution covariates. For instance, it can recognize a particular habitat from the topology, landscape, or the forest's canopy. But it may also miss some important bioclimatic factors that cannot be inferred from the chosen high-resolution covariates. To assess how much complementary information is captured by the CNN (high res.) and RF (env.) models, we computed the accuracy of a simple fusion approach consisting of averaging the predicted probabilities of each model. In practice, we first had to renormalize the probability values provided in the the submitted CSV files because the number of predictions was limited to 150 per observation because of storage issues. The result of the fusion approach compared to each model alone is presented in Figure 8. It shows that the top-30 accuracy of the fusion approach is consistently better (25.5%). This result suggests that the models capture some distinct information and that more advanced methods for combining the high-resolution variables with the environmental rasters should be explored.

6 Discussion and Conclusion

The main outcomes of the evaluation conducted in this paper are related to the comparison of two radically different approaches: one approach based on high-resolution satellite imagery and convolutional networks, CNN (high res.), and one more classical approach based on bioclimatic and soil variables processed with a random forest model, RF (env.). Beyond the slightly better performance of the CNN (high res.) model, a more in-depth study of the species predictions reveals important differences between the models. The top-30 accuracy per species, in particular, reveals a much larger performance gap. Moreover, by

comparing the predictions species by species according to their frequency, we observed that the CNN is actually much better for the large majority species, and particularly for the less represented ones. This raises several points questions related to model evaluation. In the context of species distribution studies, it seems particularly important to be able to predict the distribution of rare species, especially for protection and conservation purposes. However, a raw evaluation of the performance of the models on occurrences is biased by the long-tailed distribution of occurrences per species. The CNN (high res.) model trained on high-resolution covariates is both better on less represented species and able to predict more species perfectly than the RF learned on environmental variables, while it is slightly less efficient on more frequent species. Random forest tends to predict mostly the most frequent species. Even if the random forest's prediction is good on average on the test occurrences, it is clearly less relevant than the predictions of the CNN. Unfortunately, as both models were trained on different data, it is difficult to determine whether the origin of this difference lies in the model structure or the input data. It is important to note, however, that the high spatial resolution data has made it possible to learn a model capable of rivaling and even surpassing a model derived from a more classical approach learned on environmental variables. In addition, combining two approaches with late fusion produces a gain in performance, indicating that the models have captured complementary information. The use of high-resolution data seems to be an interesting way to learn models with high predictive power. None of the models submitted by participants were able to use the high and low resolution data together. However, if even simple late fusion improves performance, it is likely that a model trained on both data sources simultaneously will provide even better performance.

Acknowledgement

This project has received funding from the French National Research Agency under the Investments for the Future Program, referred as ANR-16-CONV-0004 and from the European Union's Horizon 2020 research and innovation program under grant agreement No 863463 (Cos4Cloud project).

References

1. Cole, E., Deneu, B., Lorieul, T., Servajean, M., Botella, C., Morris, D., Jojic, N., Bonnet, P., Joly, A.: The GeoLifeCLEF 2020 dataset. arXiv preprint arXiv:2004.04192 (2020)
2. Deneu, B., Servajean, M., Joly, A.: Participation of LIRMM / Inria to the GeoLifeCLEF 2020 challenge. In: CLEF working notes 2020, CLEF: Conference and Labs of the Evaluation Forum, Sep. 2020, Thessaloniki, Greece. (2020)
3. Elith, J., Leathwick, J.R.: Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. Annual Review of Ecology, Evolution, and Systematics (2009)

4. Evans, J.S., Murphy, M.A., Holden, Z.A., Cushman, S.A.: Modeling species distribution and change using random forest. In: Predictive species and habitat modeling in landscape ecology, pp. 139–159. Springer (2011)
5. Hengl, T., de Jesus, J.M., Heuvelink, G.B., Gonzalez, M.R., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M.N., Geng, X., Bauer-Marschallinger, B., et al.: Soilgrids250m: Global gridded soil information based on machine learning. PLoS one **12**(2) (2017)
6. Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G., Jarvis, A.: Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology: A Journal of the Royal Meteorological Society **25**(15), 1965–1978 (2005)
7. Homer, C., Dewitz, J., Yang, L., Jin, S., Danielson, P., Xian, G., Coulston, J., Herold, N., Wickham, J., Megown, K.: Completion of the 2011 national land cover database for the conterminous united states – representing a decade of land cover change information. Photogrammetric Engineering & Remote Sensing **81**(5), 345–354 (2015)
8. Roberts, D.R., Bahn, V., Ciuti, S., Boyce, M.S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J.J., Schröder, B., Thuiller, W., et al.: Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography **40**(8), 913–929 (2017)