Review for PCI Ecology of the Article entitled "Predicting species distributions in the open oceans with convolutional neural networks" by Gaétan Morand et al.
https://doi.org/10.1101/2023.08.11.551418

General comments:

This article aims to use a deep learning method to predict the distribution of marine species in open oceans. To do so, a convolutional neural network (CNN) is trained using the occurrences of 38 marine taxa (mostly pelagic megafauna, including mammals, birds, turtle, fish, coral) collected from the Global Biodiversity Information Facility (GBIF) and 29 environmental variables characterizing the surface ocean, such as sea surface temperature, sea surface salinity, chlorophyll concentration, and finite-size Lyapunov exponents (FSLEs). Classical data splitting for deep learning is used (60% for training, 20% for validation, and 20% for test). Accuracy and confusion matrix were used to evaluate the performance of the CNN. The predictions of the model are then used to provide distribution maps at global scale (for 4 dates in 2021) and in the Southwestern Indian Ocean (53 weeks in 2021). Five environmental variables were then discarded without decreasing accuracy and the most determining variables among the 20 remaining ones were calculated using the integrated gradients method. Global maps are presented for 3 species, while weekly regional maps are provided for an other species. All the distribution maps produced during the study are openly accessible online on Zenodo. For 3 other species, global maps are visually compared with global biogeography from the literature. The Analysis of determining variables revealed that the top-5 influential variables were the strength of the FSLEs, the sea surface temperature, the (sea surface) pH, the bathymetry, and the (sea surface) salinity. Finally the effect of a 2°C increase in sea surface temperature was tested and 3 additional maps of distribution anomalies are provided for 3 species (2 were partly shown previously). A short discussion (less than two pages) mentions the benefits and limitations of using CNN for marine species distribution modelling and provides some suggestions to further improve the methods.

The manuscript is very well written. The description of the rationale and methods are clear and comprehensive. All along the manuscript, the given explanations are clear and most of them are sound. However, strong ecological background is cruelly missing to convince that the method is of interest for marine ecologists (compared to classical SDMs that are not based on CNN).

The title correctly reflects the content of the article and the abstract clearly presents the findings of the study. The introduction clearly explains the motivation for the study and the research question is clearly presented. The introduction build on relevant recent and past research performed in the field, although some choices of citations are arguable (see details below).

My main concerns are on the methods, the results, and the discussion of the article.

While the methods and analysis are in general described in sufficient detail (although I have not evaluated the statistical scripts and program codes), critical elements are missing regarding the environmental variables that have been used. They are also some flaws in the methodological choices (e.g.the choice of these 38 taxa, the choice of adding +2°C to the SST to provide "tentative" and "theoretical" future projections of species distributions).

Regarding the results, I have not checked the raw data and their associated description and I have not run the data transformations and statistical analyses and checked that I get the same results. Yet, to the best of my ability, I have not detected any obvious manipulation of data. The authors performed many

predictions, but retain only some of the results to present in the manuscript. However, all the predicted maps are openly available on Zenodo. While I fully understand that all the maps predicted for the 38 taxa could not be described exhaustively, the choice of showing 3 species for global maps at one date (*Caretta caretta*, *Mobula alfredi*, *Puffinus pacificus*), 1 species at regional scale for 18 dates (*Prionace glauca*), 3 species for comparison with distribution maps from the literature (*Puffinus pacificus*, *Eubalaena australis*, *Thunnus thynnus*), and 3 species for the "STT+2°C" scenario (*Caretta caretta*, *Eubalaena australis*, *Katsuwonus pelamis*) are not justified nor explained. Why these choices of species among the 38 considered taxa?

Finally, the discussion is relatively superficial (while the identified topics of discussion are relevant) and does not rely on the literature.

To conclude, while the methodology is sound regarding the use of the CNN, it is not always the case from an ecological and oceanographic point of view. The authors should better defend why the predictions obtained from their CNN framework are reliable, robust, and trustworthy. Due to the various methodological weaknesses (see details below), as well as the relatively superficial discussion, the study has very limited ecological relevance in its present form. Therefore I would recommend major revisions to address the identified flaws ans weaknesses.

Main comments:

1) Main concerns on the description of species occurrences and environmental variables in the Method section:

Occurrences:

- Please provide additional quantitative information on the datasets, e.g. number of occurrences for each taxa and the period of years covered.
-Lines 170-172: "This is obviously wrong, but as we work with a limited number of species in an extensive area and period of time, chances are slim that the model receives contradictory information": have you checked, for your grid points, how many time this may have occurred (same location, same time, of more than 1 species/taxa)? Please quantify this (and potentially remove the point (t,x,y) with more than 1 species/taxa?)
-Line 181- 183: "It shows that some taxa were easily identified by the model (the top two being Aptenodytes forsteri and Mobula alfredi). Others were harder to predict, the worst two being Istiompax indica and Carcharhinus longimanus. " It would be useful to have the number of occurrences considered for these taxa.

Environmental layers:

- From reading the introduction, it is not clear if the vertical dimension is considered in this work, or if the ocean is considered as 2D. Please clearly state from the introduction that you are not considering the vertical dimension of the ocean.
- It seems that only one value of environmental variable/layer is considered for each (lon, lat) grid point. Is it correct? If so, is it an annual mean? For which period? Or are you considering the time stamp of the specie occurrence? Usually, seasonal means and or seasonal stdev can be considered (see for instance Benedetti, F., Guilhaumon, F., Adloff, F. and Ayata, S.-D. (2018) Investigating uncertainties in zooplankton composition shifts under climate change scenarios in the Mediterranean Sea. Ecography, 41: 345-360. https://doi.org/10.1111/ecog.02434)

- In Section 2.2, many information is missing on the environmental variable: are they surface values? what are the units (e.g. diatoms: is is concentration in carbon? number of cells?). It is also unclear how the time is handled: for a given occurrence recorded at time t in coordinates (x,y), which value of SST is used? Annual value? for which period? Too many information is missing here to be able to reproduce this work.
- Similarly, more information should be given on the environmental data in Table 2: are they surface values? Depth-integrated values? What are their units? Are you considering annual mean, and if yes for which period? See for instance table 1 of Reygondeau et al. (2017, Biogeochemical regions of the Mediterranean Sea: An objective multidimensional and multivariate environmental approach. Progress in Oceanography 151, 138-148. https://doi.org/10.1016/j.pocean.2016.11.001) for the description of environmental variables.
- Space and time association is not clear in the Method section.

2) Problems that have been identified in the predictions:
- L 183: the confusion matrix also indicates a high confusion between Istiophorus and *Carcharhinus falciformis*. A quick verification on obis.org confirms that this genus and this species have a relatively similar distribution. Given the fact that for each occurrence, you consider that the other taxa are absent, this leads here to high confusion.
- There is also a problem with Acropora predictions as this coral genus is present in coastal areas, mostly in less than 10 m depth (https://obis.org/taxon/205469), and the simulated distribution provided as supplementary material though reference 42 seems offshore. Given that the bathymetry has been taken into account, this is most surprising.
- I would suggest that, for each taxa, you check if there is not obvious problems with what is know in obis. In the present form, it is not convincing enough that your results have some ecological relevance.
- L 218: Results for the Southwestern Indian Ocean: usually, in niche modelling, climatological data are used to train the model and describe the habitat probability of a given species. Therefore, weekly predictions may not be relevant. Given the methods you are using (e.g. no possible co-existence of several species), I am not sure that these predictions are reliable. Please explain.

3) Limited relevance of the "SST+2°C" scenario:
I see a flaw regarding the effect of a 2°C increase in sea surface temperature: "Predictions were computed after adding 2°C to sea surface temperature, leaving all other variables unchanged." (line 250) Yet, this is not realistic at all, since temperature increase is not expected to be homogeneous, cf the different IPPC reports and regional variations that have been reported. Although the authors acknowledge that " In the context of climate change, this is a tentative projection but it is theoretical, as there are significant and complex correlations between future changes in various environmental 252 variables", I would recommend to remove all this part of the study, or to redo it using a SST field predicted by any Earth System model from IPCC for a given scenario. In that case, changes in SST should be considered (rather than new values), see for instance how SST scenarios are handled in Benedetti et al. (2017, Investigating uncertainties in zooplankton composition shifts under climate change scenarios in the Mediterranean Sea, Ecography 40: 001–015, doi: 10.1111/ecog.02434)

4) Choices of 38 taxa:
- I would also suggest to use a few zooplankton species, as many work has been done previously to describe their distribution at global scale (e.g. Benedetti, F., Vogt, M., Elizondo, U.H. et al. Major restructuring of marine plankton assemblages under global warming. Nat Commun 12, 5226 (2021). https://doi.org/10.1038/s41467-021-25385-x) in order to compare the results you obtain with your CNN approach and more classical SDM approaches using an ensemble of models. Besides, the consideration

of planktonic species has the advantage of considering organisms that are not able to swim and move over large distances due to foraging or mating behaviours. This would also make your discussion (section 4.1 Ecological interpretation of the results, implications for offshore species distributions) more relevant.
- Also clearly state in the introduction and in the discussion how you deal with movement of species, as movement and migrations are expressly mentioned in several parts of the manuscript (e.g. first and last sentences of the abstract, line 129, line 268, line 315). It seems that sometime you are considering only marine megafauna, which is is not the case.

5) Relatively superficial discussion
Discussion of the results:
-Line 272: "This highlights the need for distribution models of fast-moving species to consider these [temporal] variations, instead of relying only on averaged values.." (extra final point to be removed) This statement is somehow obvious. Why not compare your results with previous studies using SDM for a few fast-moving species?
- Line 275: "A possible explanation is that the model may have used other variables as a proxy for low depths." Again, this statement sounds obvious.

Limits of the study:
- L 283: "We noted three main limitations of our method, namely performance metrics, biases in the input data, and some undetected patterns." The fact that your model can only predict the presence of one taxa is also a major limitations. Although this is mentioned in subsection 4.2.1 on Accuracy, this should be more clearly underlined and discussed.
- L298: "Most observation data in the open ocean come from fishing vessels, which target certain species": this is not true for plankton species, there the relevance of considering planktonic species as well.
- Line 302: "The strength of deep learning in this context is that it makes no assumption when there is no data: it replicates the results from similar well-known areas." I disagree since other SDM methods using pseudo-absences (rebuilt from the available data) also do this.
- Line 308: "some data come from scientific tracking of individual animals": which taxa? How many observations (compared to the total number of observations for these taxa)? Please be more specific.
- Section 4.2 of the discussion: no references are given. Please discuss you results in the context of state-of-the-art and relevant literature.
- The justification of considering monthly habitat mapping should be clearly explained for species exhibiting migrations.

6) The conclusion section also lacks strong scientific background.

In my opinion, these weaknesses should be addressed in a revised version of the manuscript to make it relevant for future ecological studies and better convince that the proposed CNN-based SDM provides reliable results with ecological relevance.

Minor comments :

Line 16: I would suggest replacing "Earth's climate, nutrient cycles, and biogeochemical cycles (including carbon sequestration)" by "Earth's climate and biogeochemical cycles (including nutrient cycles and carbon sequestration)".
Lines 15 and 17: Beware that references #1 and #2 are on the deep sea.

Line 22: " the most pressing challenges" => which ones? Consider replacing by "these most pressing challenges"

Line 41: "Usually, SDMs use environmental data at the exact location where the prediction is computed": it is not true as usually climatologies are used, e.g. using mean seasonal values of SST, rather that the SST value recorded when the species has been observed. See for instance the work of Benedetti and colleagues :

- Benedetti F, Vogt M, Elizondo UH, Righetti D, Zimmermann NE, Gruber N (2021) Major restructuring of marine plankton assemblages under global warming. Nature communications 12 (1), 5226. https://doi.org/10.1038/s41467-021-25385-x
- Benedetti, F, Vogt, M, Righetti, D, Guilhaumon, F, Ayata, S-D (2018) Do functional groups of planktonic copepods differ in their ecological niches?. J Biogeogr. 45: 604–616. https://doi.org/10.1111/jbi.13166
- Benedetti, F., Guilhaumon, F., Adloff, F. and Ayata, S.-D. (2018) Investigating uncertainties in zooplankton composition shifts under climate change scenarios in the Mediterranean Sea. Ecography, 41: 345-360. https://doi.org/10.1111/ecog.02434
- G Reygondeau, C Guieu, F Benedetti, JO Irisson, SD Ayata, S Gasparini, Koubbi P (2017) Biogeochemical regions of the Mediterranean Sea: An objective multidimensional and multivariate environmental approach. Progress in oceanography 151, 138-148. https://doi.org/10.1016/j.pocean.2016.11.001

Line 47: citation #12 refers to a model. I suggest to cite another citation referring to observed patterns, even if not at the species level, such as : Baudena, A., Ser-Giacomi, E., D'Onofrio, D. et al. (2021) Fine-scale structures as spots of increased fish concentration in the open ocean. Sci Rep 11, 15805. https://doi.org/10.1038/s41598-021-94368-1

Line 55: not sure why not taking "into account the high temporal variability of environmental data [...] seriously hinders the prediction of highly mobile species distributions." I would argue that is it also true for non mobile species like plankton.

Line 105: essential to ?

Line 141: " Very few of the occurrences were located in the Arctic Ocean: they were assigned the closest of these ocean basins." why not remove them ?

Line 101: Why are you using genera here? SDM are based on the niche theory that applies at the species level. Please justify.

Line 112: "When there were more than 10,000 occurrences of a taxon, a random sample of 10,000 occurrences was selected."" Why? Which taxa are concerned?? taxon => taxa

L 224-223: "Yet the maps that we produce are highly dependent on time, see Figure 6 for instance." See my comment above regarding line 218.

L 226: "Comparison of predicted distribution maps to establised maps". Here you are showing 3 examples. How have you chosen these 3 species? Indeed, it could be seen as cherry picking among your 38 taxa. establised => established

Figure 7: some information is missing in the caption. Which date in 2021 are you showing in the right panels? Could you please also cite the references (44, 45, and 46) in the caption?

Table 4: please order your variables by mean or max and provide more information in the caption (see my comment previously)

Lines 244 and 245: FSLEs could be replaced by "finite-size Lyapunov exponents"

Line 247: Please describe Figure 8 in a few sentence. What is the main message from this Figure?

Figure 9: Again, if picking 3 examples among the 38 taxa, the choices of these 3 examples should be clearly explained.

The captions of the Figures and Table should be more informative.