Answers to editor comments and reviews

Dear authors,

Thank you for considering PCI Ecology for examining and recommending your work.

We have received 3 detailed reviews that you will find on the PCI website.

I concur with the 3 reviewers that the topic is interesting and timely, and that all the data and methods are clearly presented so that the analyses can be reproduced.

However the reviewers also raised some critical points that would require careful consideration in a revision.

There are two categories of major comments and recommended improvements:

- to better explain the ecological context, the biological properties of the species considered in the study, and to better discuss the results in light of the ecological context.
- to better explain and discuss why and how using a classifier of all taxa allows properly assessing the possible co-occurrence of different taxa at a given location. One of the reviewers did insightful comments and suggestions on this point. I don't mean that the methodology should be completely rethought, but more discussion will certainly be useful for the readership.

I hope you will find all the comments and suggestions of the reviewers useful for your revision.

We look forward to receiving a new version of this manuscript.

Sincerely,

François

Dear editor and reviewers,

We would like to thank you for your high quality reviews that certainly helped improve the manuscript significantly. We answered to each comment, and indicated the line numbers of the updates. Both versions can be displayed side by side with the changes highlighted at this link: https://draftable.com/compare/tlNMENiuxztZ

Regarding your first point, we added information, especially in the discussion. Our detailed answers below also expand on the difficulty we had to validate our predictions ecologically. Regarding the second point, we clarified the goal of the model in the introduction, and also the way we deal with co-occurrences, as it seems that our phrasing was misleading and let the reader think that co-occurrences were not possible due to the model being a classifier.

Unfortunately, we are not in capacity of altering the training process (including input data), as the funding for several of the authors' employment has expired, and it would require several months of additional work. When reviewers suggested such changes, we added them into the discussion so that they may be considered for future research. We hope you will find that the method is worth publishing even if not perfect in its current form.

Finally, we would like to point out that there is a slight change in the order of influent variables, as we recalculated their influence using the whole data set rather than a random sample (which was a flaw that Reviewer 1 pointed out).

Sincerely,

The authors

Review #1

# Review for "Predicting species distributions in the open oceans with convolutional neural networks"

This article uses the method defined by Botella et al (2018) and Deneu et al (2021) to model the occurrences of 38 marine taxa based on a new set of environmental variables appropriate for this environment. This is a very good idea since this method seems extremely promising and should be relevant to capture important mesoscale features in the ocean, such as fronts.

Furthermore, the authors should be commended for the effort they made to make their research reproducible and reusable by others. They shared the input data, the results but also the code used to fit the model and the code to extract relevant environmental data around a point of interest, in the form of a Python package. This is excellent.

While I find the premises of the study interesting and its technical setup exemplary, I have several remarks about the content, which I detail below. I hope that they will be perceived as useful by the authors.

## Major remarks

### Baseline model

In quite a few ways, this model is different from that of Deneu et al (2021): the species and environment are different of course, but also the way the missing values are handled, possibly the loss function, the evaluation metrics etc. For all these reasons, it would be necessary to evaluate the results against a baseline model. It would help highlight the advantage of using a CNN and substantiate such claims in the discussion: "This method holds promise in helping researchers uncover new correlations between the oceanic conditions and species distributions".

The comparison that would best show the potential advantage of adding the convolutional part would be with the "Punctual deep neural network (DNN-SDM)" of Deneu et al, keeping the dense part exactly identical to the one used in the current model.

An alternative that would allow comparing with a "classic" approach would be to predict each species independently with Gradient Boosted Trees (or a Random Forest) and pseudo absences at the points of other species.

Yet another alternative that would keep the multivariate response but use classic tools would be to use Gradient Boosted Trees in a classification setting, with categorical cross entropy loss (`logloss` for xgboost).

We trained a Punctual DNN model, and added a paragraph on this in the Methods section (line 226). We also mention this result in the discussion (line 338)

### Is this model a Species Distribution Model?

55-56: "Furthermore, SDMs rarely take into account the high temporal variability of environmental data [14], which seriously hinders the prediction of highly mobile species distributions."
This, and other aspects of the paper pose the question "what is a niche" and whether the model trained here is a niche model. The ecological niche of a species is the set of conditions in which this species can survive and reproduce; it is shaped by natural selection. The purpose of a Species Distribution Model, or a Niche Model, is to capture that niche and project it spatially, as the region in which the species should be found, i.e. the distribution *range* of the species. A reason why most SDMs use climatological summaries of the conditions in a given place (mean, but also min and max, etc.) is because the presence of a species in a given place on Earth is not determined by the immediate conditions in this place but by the long-term range of conditions experienced there. This is easily understandable for non-mobile species: the persistence of an oak forest depends on the range of conditions experienced over the last decades, not on the temperature at any specific date. Even for mobile species, the purpose of a SDM would still be to project the complete range of existence of that species; its extent would be determined by natural selection, again, not by the individual movement ability of the organisms.

I would argue that a model that relates the presence of organisms to the immediate, dynamic environmental conditions of a given place, as done here, is not a "niche" model anymore. A broader term in which this could fit would be "habitat models". It uses the same general reasoning (organisms are constrained by their environment) and often the same statistical tooling, but the underlying hypothesis is different: while in niche models, the presence of a species in given conditions is to be related to its ability to persist in those conditions over generation, here, the presence is determined by the organisms ability to move towards favourable conditions at a specific time of the year (or rapidly multiply in these conditions, but all animals here are long-lived so this does not apply). In passing, this makes the non-negligible hypothesis that the animals in question have the ability to detect those regions of favourable conditions from far away (through direct mechanisms or proxies) or at least to sense when they are in bad conditions and move away from them.

I would like to this this question discussed in the paper and, while you can refer to niche models as inspiration and regarding their general principle, I think that the mechanisms represented here do not pertain to the ecological niche of the species and this should be made explicit.

We reorganized the “Existing methods for predicting species distributions” subsection to discuss and clarify this. We acknowledge that our model may have different goals from most traditional SDMs, but according to the definitions in Melo-Merino et al., 2020, dynamic distribution modeling is part of species distribution modeling, although different from ecological niche modelling.

This would have implications for section 3.2, where you compare a snapshot distribution over one date (March 2021) to the map of the distribution range of the species. It is unsurprising that they do not match and you cannot even hypothesise from them that "the established geographical range is not fully used by the species". I suggest that you can:
(1) either use those maps to check that the distribution you predict is *within* the range of the

species, as a broad check of the method, and not more;
(2) or predict those maps for several seasons of several years (probably 20 to capture a full climatic period), produce an average, and then compare this average with pre-existing species distribution maps.
I assume solution 2 is computationally very demanding.

Indeed, solution 2 would require days of data download; we settle for solution 1. We removed the hypothesis that the geographical range is not fully used.

The discussion should also be reviewed in that light. Part 4.1 compares this to SDM studies and conclusions such as "This highlights the need for distribution models of fast-moving species to consider these variations, instead of relying only on averaged values" are misleading. It is just that habitat models using snapshot of dynamic fields vs. average fields do not answer the same ecological question at all. If one wants to model the distributional range of a species, then climatological averages *are* the correct answer (or possibly averages of snapshot predictions over climatological scales).

We removed the second part of the sentence and supported the first part with two references (line 328). In the new paragraphs of the introduction we clarified that we work on dynamic SDMs, and that including spatial features as predictors forces us to drop climatological summaries (lines 54-63).

### Choice of species

Related to the points above regarding the nature (SDM or not) of the model, I find the inclusion of the Acropora coral very strange in this list. All other species are large, quickly moving organisms, which could indeed follow water masses of suitable properties; therefore what is modelled is the set of favourable conditions throughout the year (rather than the niche). This relates to "movement ecology" to which you refer a couple of time. In the case of Acropora, the relationship between presence and immediate conditions in the vicinity (in time and space) of the observation make much less sense; for this taxon, the long term conditions are what determines species presence. While your model may still capture the fact that Acropora lives in rather warm waters, for example (because no occurrences will be recorded beyond tropical regions), a model based on climatologies would make much more sense. I would suggest to remove it. I realise that it means re-running all model fitting and predictions, since the model fits all species together...

Furthermore, it would be relevant to justify the choice of species in the text. Again you seem to have made a deliberate choice of large, wide ranging/moving species, state it, explain why, and draw the conclusions for what your model is about (see above).

Unfortunately we cannot remove a species from our dataset without redoing the whole study. We added a few sentences to justify our choice of species (line 118).

### The model is a classifier

While the model is presented as predicting species distribution, which is usually seen as a regression problem, it is actually a *classifier* that predicts which species, among those modelled, is the most likely to occur in any given pixel of the ocean; and there can only be one. As the authors briefly discuss (286-290), this is problematic when two species occupy the same environmental space and should therefore be equally likely to occur in a given pixel. The probelm appears for the accuracy metric (which is what the authors discuss in section 4.2.1) but is also, and more profoundly, affecting the model itself.

If two species occur in the *exact* same conditions (i.e. in the exact same places and times, in the context of this paper), then the model has no way of sorting out these contradictory informations and will come to a 50-50% answer, which is the right one.
Now, in the more realistic case where two species occur in very similar but not exactly the same conditions (for example, close to each other geographically), they should still be predicted 50-50%. However, the softmax and cross-entropy loss will instead push the model towards extreme answers in terms of "probability": to minimise the loss, not only should the model predict the correct one of those species in a given pixel/set of conditions, it should also predict it with high probability, and therefore all others will have low probabilities. To achieve this, the model will pick up minute differences in conditions, that are probably irrelevant biologically, and output predictions that are far from 50-50%. I would venture that this is one reason for the very spotty aspect of the predicted maps, in the Indian ocean for _C caretta_ or the west Pacific for _P pacificus_ for example (one other reason may be mesoscale structures that would be visible in the FSLE, temperature, etc. but it is difficult to say without a map of those). Actually, this could even be the reason why FSLE is one of the most predictive variables: it has very strong local structure and therefore is one variable that can be exploited by the network to come up with (artificial) differences between nearby occurrences.

Unfortunately, I do not have an easy solution for this. Deneu et al, who use the same approach, get (probably partially) around this problem by predicting a very large number of species (>4000), which likely helps smooth things out during training (especially is sufficient regularisation is used), and use metrics that consider the top k predictions only (not top 1), which diminishes this issue during evaluation. This is not applicable for you. Maybe another loss function would be more appropriate in your case?

Thank you for your input. This is a very interesting point and we added a paragraph on this in the Discussion (lines 345-350) and a paragraph in the Methods (lines 206-219). We believe that the use of averaged binary cross entropy (cf. below) reduces the issue you mention about pushing other species probabilities towards 0.

A common solution to reconcile disagreeing inputs in SDMs (e.g. presence and absence of the same species in similar conditions) is binning. Because the models operates in environmental space, such binning should ideally be done in the n-dimensional environmental space. However, it is commonly performed in the 2-d geographical space because it is easier and has the added benefit of correcting some of the observation bias (reduce the number of inputs in regions frequently observed; as noted in your section 4.2.2). In your case, the environmental space is only defined after the inputs pass in the feature extractor and it has likely >1000 dimensions (the output size of Resnet50); so using this is out of the question. You consider geography *and* time to fetch the environmental data, so binning should be 3-d (lat,lon,time), not 2-d as usual. The idea is then that, if you get observations for two species relatively close in lat,lon,time then you consider them as a single input with two 1s on this row, hence capturing the fact that these two species co-occur. For this to happen, your bins probably need to be quite large and one possibility is to consider only the week or month of the year for time. This should also be done on the full GBIF output, before the subsampling of 10,000 per species. From then on, the loss function should cope with the fact that there can be several 1s on the same input; I do not know enough about cross entropy to know if it does. MultiR2 would but has other drawbacks.

Overall, (i) the fact that this model is a classifier and the meaning of predictions should be made clearer in the text, right from the abstract and introduction, (ii) the effect that this fact has on the predicted maps should be discussed more, unless, of course, (iii) an alternative solution is found (through a different encoding or loss function).

We clarified in the abstract and the introduction that the model is a classifier (line 89). We discussed this further in the Accuracy part of the Discussion (line 345).

The problem with binning is that we would lose the potential correlation between the spatial structures and the occurrences.

156-171: as a side note, you mention using binary cross entropy although your response has length 38, not 2; you are probably using categorical cross entropy, like Deneu et al, right?

We used element-wise Binary Cross Entropy which is then averaged over the training batch. This is quite odd but we found that in our case it worked best to obtain convergence. We've added this into the *Training* subsection of the Methods (line 204), and mentioned it in the discussion (line 424).

### Geographical predictors

136-140: I am ambivalent regarding the inclusion of the geographical variables justified by the fact that they constitute barriers. If this is considered a SDM, the barriers should have been present for a sufficient amount of time in the evolution of the species to let subpopulations evolve different preferences in the different basins and this is not mentioned/justified. If it is not a SDM but a "preferred habitat" model (which I think it is), the reader is still lacking justifications that most species do not regularly cross these barriers. But most species considered (tunas, marlins, sharks, whales, dolphins, etc.) have global distributions and, while they are often managed/considered as different stocks, they likely could move from one ocean basin to the next. For the model, the question becomes whether these subpopulations have different environmental preferenda in those various basins, warranting the inclusion of those variables to capture interactions with other variables (e.g. different temperature range in the Atlantic vs Pacific). I do not know enough about the biology of all species to be conclusive but I think the authors need to justify that choice further.

As you say, this issue is made worse by the confusion between varying definitions of SDMs: these barriers come from the long-term distributions, but we use them in a context of dynamic SDMs. We added a few sentences to justify that choice in the methods (line 171). Hopefully the clarification of the type of SDMs we use in the introduction will also help.

The most problematic choice is the inclusion of hemisphere, which creates a completely artificial boundary in the distribution of _Caretta caretta_ in the Indian Ocean (Fig 5a) for example. I understand that was included to avoid predicting arctic species in the antarctic but the artefact above is reason enough to remove it, in my opinion. Maybe the arctic/antarctic separation can be ensured by masking on the prediction: setting the proba of arctic species to 0 for all points in the southern hemisphere, or south of a given latitude, and rescaling the others to sum to 1?

We do not want to add custom mask for specific species, as it would make it much more difficult to scale the method to a large number of species. We added a potential solution to these artefacts in the discussion (line 426).

Sidenote (related to line 144): the polar front is a pretty strong barrier for many species, even though it is not "physical" in the sense of a continent. If the warm waters of the equator are considered barrier to movement, then this could be also, which makes including hemisphere even more questionable.

We define the equator as a barrier because the conditions on both sides are similar, so it is the only way to explain that some species are not present in both hemispheres. When species move towards the poles, the temperatures get colder and colder but never go back up (except in the opposite ocean for species that can cross the poles, but that is already covered by our Ocean variable), so there is no need for an “artificial” barrier to explain the absence of species.

Finally, if I understand correctly, these binary variables were included by giving a completely uniform 32×32 input tensor (filled with either 1 or 0, I suppose). First, this wastes a bit of computational resources since performing convolutions/pooling on a constant input just outputs the same number.

This doesn't necessarily consume more GPU, as processing this data in vector rather than raster form would have required an additional branch in the network architecture, as well as a fusion layer to aggregate the extracted features with those of the other modalities. Generally speaking, in a GPU processor, the more compact the input data (=in the form of a single tensor) and the simpler the computation graph (=a sequence of homogeneous tensor operations), the more efficient the computations (operations being optimally parallelized for this type of tensorial computation graph). We added this in the methods (line 143).

Second, I am not completely sure how those would behave during the convolutions with the other tensors (i.e. other variables) but it is likely that tensors of 1s just have no effect while tensors of 0s mask the other variables.

The different channels of a tensor are not multiplied together during a convolution operation, so there is no such masking effect. Locally (=for a given output pixel of a given output channel), convolution is simply a weighted sum of the input variables (=a linear model). When one of the input channels is locally constant, it's just that the same input value will appear several times in the linear model (with different weights due to their random initialization). The important information extracted by the network for such locally constant variables is their average weight that determines the strength and direction of the influence of this variable on the predicted output (exactly as a classical linear or GLM model does). The main advantages of exploiting these locally constant variables as tensors rather than vectors are: (i) simplicity of implementation (managing tensors of different sizes in the same network would lead to a very complicated and potentially less efficient architecture) (ii) homogenizing the contribution of the different input variables in order to avoid convergence problems (e.g.

over- or under-representation of one variable compared to the others induced by the network architecture itself).

In both cases I think this is not what you want: you want the effect of the patterns in the other tensors to be *conditional* to the geographical variable, e.g. you want the temperature patterns to always show but have a potentially different effect in the Pacific vs. Atlantic ocean.

This is precisely one type of information the model can learn. Each convolution layer is indeed composed of a battery of locally linear models (one for each output channel) followed by RELU activation functions (that allows detecting patterns such as the one you mention). The network is then composed of a sequence of such layers which allows capturing even more complex compositions of patterns.

Furthermore, in the first layers of the network, the convolutions of those constant tensors will be limited with the tensors that happen to be placed next to them in the 32×32×29 stack, which, again, is probably not what you intended.

I'm not sure to well understand this remark but if it is related to long-distance statistical relationships it is true that CNN can not capture them (as most existing models anyway). The interest of CNNs is to capture local patterns (e.g. the topology of the site and its relative importance to other environmental variables such as temperature, etc.).

For all these reasons I suggest that, if some geographical info is to be kept, it should be included as additional scalar values (of 0 and 1) concatenated to the output of the feature extractor, i.e. in the first layer of the Multi Layer Perceptron. This way, they are not used in the convolutions and they can interact with all other variables through the fully connected layers.

### 2ºC increase simulation

I do not find this section useful. You acknowledge yourselves that more than temperature will change in the future. So just adding 2ºC everywhere, without considering the associated change in stratification, circulation, primary production, etc. is just unrealistic.

The minimal relevant way to do this would be to consider full earth system models outputs, for at least one model and one scenario, to fit and project the model in the present and project it on the future time (both taken from the same run!) and compute the difference between present and future. More appropriate studies would transform the climate model to fit in the space of observations, using something like CDF-t, and to fit+predict the occurrences in this new space, in both the present and future, for different earth system models and scenarii. This is clearly out of the scope of this study. So this part should just be cut.

We removed this section.

## Minor remarks

Abstract: "for prospective modelling of the impacts of future ocean conditions on oceanic species" : nowhere in the Abstract is the +2ºC experiment mentioned so the reader cannot know what this sentence refers to. But see above regarding the relevance of this +2ºC experiment.

This sentence was removed

15: Ref [1] applies to the deep ocean, not to the surface; it is not be the most appropriate to justify this sentence.

We replaced this reference with more relevant ones (line 15).

25: the influence of temporal variations: what does it mean exactly? At this point the reader has not read about the time resolution of the study and similar models frequently use climatologies, hence erasing time. It would be necessary to explain a bit more.

This part of the sentence was removed

26: identify areas that are most at risk: risk of what? SDMs can predict species distribution but then, from that, how to identify regions at risk or not? Rephrase.

This was rephrased (line 25).

79: (about Sentinel2 data) "we cannot rely on this information in the open ocean": why? Sentinel-2 data is readily available and can be used to describe, in great detail, phytoplankton blooms for example. The product derived from Sentinel's raw data make most sense on land (to capture vegetation cover etc.) but the raw RGB data can be used at least.

We meant that little information can be used from RGB data directly. Ocean colour derived products (chl, phytoplankton, etc.) are available so we used these directly. We rephrased the sentence (line 91).

94-95: at this point a few sentences should be added about the fact that the response is multivariate, otherwise the reader cannot understand why the output of the model is "a vector of observation probabilities". The most common approach in a species-environment model would have been to make one univariate model per species.

We added this at the beginning of the methods section (just before we mention the vector of probabilities) (line 110)

108: GBIF data being what is it (invaluable but sometimes worsened by records of lower quality), I am very surprised that you did not have to further clean the data. If you did inspect it then please mention what you did. If you did not, I think it should be checked for:
- duplicate records (including with a buffer around each point since the duplicate may have slightly different coordinates)
- records that are far out of the geographical distribution of the rest of the points and that are likely to be mis-identification or wrong geographical coordinates; this can be done with density-based methods.

Duplicates could also be multiple sightings. As we cannot inspect the numerous sources one by one, we chose not to intervene in the input data, We acknowledge that this choice is arbitrary and in future work it may be interesting to inspect input data more thoroughly. We added this into the manuscript (line 133, line 398).

131-134: the change in lon/lat ratio when moving from the equator to the poles is acknowledged but how was it taken into account? Was is considered when performing the interpolation from 241×241 to 32×32?

The change in lon/lat ratio is considered when downloading data with geoenrich, as the downloaded data is defined by a buffer in kilometers (line 156). So there is no need to consider it further when interpolating.

Table 2: some choices in environmental variables are strange.
Why take Chlorophyll from a source (33) different from that of the phytoplankton groups (35)? Chlorophyll is actually available from the same Copernicus product and using this would ensure a more consistent signal among the this and phytoplankton functional types. This would avoid having a signal for phytoplankton groups but missing data for Chl --which then appears replaced by the median-- as visible in Fig 3.

We used chlorophyll from OCCI as we were told it was the best quality source, despite the missing data. Indeed it is provided daily instead of monthly for the Copernicus product. This could be experimented with in a future version of the work. We now mention this as an example in the discussion (line 405).

Why consider u,v for wind and current but strength and orientation for FSLEs? I think the relevant choice is strength and orientation for all; this is what animals will more directly be sensitive to. Furthermore direction is an angle, meaning that, in its raw form, slight changes from 359° to 1° would appear as a major feature/ridge in the 32×32 tensor (as is visible in Fig 3 actually) and this artefact will be picked up by the convolution filters. It should be transformed to be continuous. Common transformation for angles are logit or cos and/or sin.

U/v or strength/orientation is an arbitrary choice. We agree that for most variables strength and orientation is more logical. But for wind and current we thought that the East-West and North-South components may benefit from explicit characterization due to the different ways they appear, in relation with Earth rotation.

A solution would be, as you suggest, to use strength + cos + sin which would tick all boxes. It is a great idea and we wish we had it before. We've added this to the discussion (line 411).

Regarding environmental data, what is the time resolution of the products? Was any time-wise interpolation done to define the array of values taken or it was just the closest available date? If it is the later, knowing the time resolution is particularly important.

The closest available date was used. We added this into the text (line 156) and time resolution in the Environmental Variables table.

150: what does "clipped" mean: replaced by a missing value (and therefore replaced further by the median) or replaced by the previous most extreme value?

It was replaced by the most extreme value. We've clarified this in the text (line 184).

153: you used the median to fill in missing values while Deneu et al (2021) mention "Furthermore, rasters contains sea pixels and other undefined values that should be attributed a numerical value. To avoid as much as possible potential errors related to this constraint, we chose a value sufficiently distinct from the other values, here we choose a value under the minimum of the values of valid pixels". And indeed, in your case, when pixels are land for example, they should probably get a different value which would allow the model to pick that up. Can you justify the choice of the median vs. the choice made by Deneu et al. above?

The choice of Deneu et al. is to encode all missing values as a single specific pattern with the idea that this pattern could be informative (typically the presence of a coast in their case). Theoretically, then, they don't consider these to be missing values (=unobserved values of the variable), but full-fledged values (=characteristic of a certain information). This hypothesis may be relevant for some raster but there is no guarantee it is relevant for any raster. It could even be counterproductive if the presence of missing values is correlated to some confounding factor.

In our case, we prefer to consider missing values in the classical theoretical sense of the term (=unobserved values of the variable). Taking the median values is one the most employed method in this regard. Furthermore, the model already has access to land/sea values through bathymetry.

We added some text in the *Feature scaling* subsection of the Methods to clarify (line 189).

174: how was the 80-20-20 split done? Given the origin of data, there may be autocorrelation among the data points (several occurrences reported in close proximity in space and time). Not taking this into account by doing a purely random split would
(i) push towards overfitting (if points in the validation set are close to points in the training set, fitting tightly to those points in the training set would decrease validation loss) and
(ii) inflate accuracy (if some points in the test set are close to some in the training set, they are "easy" to predict).
Common solutions to this are block cross-validation, withholding of complete regions (in space-time for you) for val/test, or choosing points that are far from others by examining the density of observations and picking points in low density regions.

We did a purely random split. We've added this into the discussion (line 420).

192: what are the resolutions of the prediction grids in both regions, in degrees? What resolution was it interpolated to afterwards? How (linear, spline, other)?

The grids have a 100km resolution, so the resolution in degrees varies. The World grid was interpolated to 3600x1800 pixels and the Western Indian Ocean grid to 800x800 pixels using cubic interpolation. This was added to the *From probability predictions to distribution maps* subsection (line 244).

194: is there any clever trick in the fetching of environmental data on a grid? Indeed, it is likely that the regions around successive grid points overlap and therefore that it is possible to fetch a large region once and then cut it into chunks rather than downloading each one separately (leading to multiple downloads of the same pixels). If such cleverness is built into geoenrich, mention it, it is worthwhile!

We added a sentence on caching in the *Enrichment* subsection of the Methods (line 154).

197: can you make it a little more explicit what you mean by "relative probabilities"? Probabilities sum to 1 for each pixel, by definition, so what does the added "relative" mean? My understanding is that, in the maps, the probabilities of occurrence are rescaled per species, so that, even if a species has a low probability of being the first predicted species everywhere, the map still goes from white to dark blue. Is that the case? If so I would call them "rescaled probabilities"; or I would just display the probability but keep the colour scale independent for each plot (no matter if the max is 0.9 or 0.01, it is dark blue).

The probabilities are not rescaled to enable comparison between species. We clarified this sentence (line 246).

208: the general principle of the "integrated gradients method" would need to be explained here. It is not well known enough to assume that readers will know what it does and it would be nice to avoid them having to read the underlying paper to understand it.

We added a general explanation of the method (line 259).

209: how representative is this 1000 random sample? Was it stratified geographically to ensure it covered various conditions?
Ideally, these 1000 points should cover enough of the environmental space of the full 36506 to be considered representative. One way to check this is to extract the feature vectors of the 36506 points (i.e. the values at the end of the feature extractor, or possibly at a further layer of the MLP), draw the density distribution of values for each feature, do the same with only the 1000 and compare the density distributions. You want those to match as best as possible. The mismatch can be quantified with the Kolmogorov-Smirnov statistic, for example, and deciding whether 1000 is enough can be done by using 10 (which will not be enough), then 100, then 500, etc. and check when the statistic saturates.
A cheaper and approximate way to do it is to extract only the centre pixel for each variable and do the same: i.e. does the distribution of temperature at the 1000 points look like the worldwide distribution.

We regenerated the corresponding table using all data rather than a random sample. We kept the same figure as before for taxon-specific data, as a sample of 1000 is more representative in these case (fewer data points in total).

IMPORTANT: this is the reason why the order of the most important variables changed.

210-211: I am not sure I understand this. The summing over geographical area would tell you if the most explanatory variables are similar between the Atlantic and Pacific for example, right? But those results are never shown. Then the sentence suggests that it is the values aggregated by area that are then summed per taxon, but Fig 8 is only per taxon; I assumed this was just a straight sum over the 1000 points.

We meant that it was aggregated over the whole world area, we rephrased (line 262).

Section 3.2: on several occasions, in that section, you explain the geographical discrepancies between the theoretical distribution and the predicted one by under-representation of occurrences where the species is not predicted (l. 232, 241). But the model operates in environmental space (except for the few geographical variables) so the representation that matters is that of the environmental conditions, not of the geographical locations. This is actually the whole point of such habitat-based models: predict probability of occurrence in data poor region. So you should be a bit more careful with the wording here (in addition to reconsidering the general point of view of this section, as explained above).
NB: This is also one more reason to avoid adding geographically constraining variables.

Indeed this was not clearly worded. We added a few sentences to clarify this in the *Puffinus* subsection (lines 297-302).

280: the fact that the effect of variables has to be studies afterwards is not a drawback of solely this method. It is the case for all other machine-learning based methods (where partial dependence plots have to be drawn after model fitting) and even multivariate linear models where the true understanding of the contribution of variables come from effects plots. It may be longer and more computationally demanding to undertake here, because of the complexity of the input data, but is not different conceptually.

This sentence was removed.

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. This partly compensates for sampling effort heterogeneity." This is true for habitat models in general (it is actually the purpose of these models); there is nothing specific about deep learning here.

This sentence was removed.

304: "But this only works when there is a homogeneous population". Actually, this model (and other machine learning-based ones) likely has enough degrees of freedom to capture multimodal responses. So if a species has two stocks, which respond differently to environmental conditions, the same model should still be able to capture both relationships and predict presence in all places that are favourable in at least one of the two stocks. But of course, one needs samples from each stocks to start with. For _T thynnus_, the issue is likely the imbalance between the West and East (only ~900 occurrences east of -35º, over a total of ~9000). So, overall, this is a data problem, which fits in this "Observer bias" section, but the wording makes it sound like a model problem.

This was rephrased (lines 373-380).

Section 4.2.3: pointing what the model missed is good and honest. But the reader is expecting explanations as to why this was missed. Is it a lack of occurrence data, the lack appropriate environmental variables relevant to capture the conditions towards which the animals migrate, a problem with the fit of the model?

Unfortunately, this information is hard to obtain. We added in the text that the causes are unclear, but that we propose several ways to improve the results in the following section (line 393).

Section 4.3.3: This is relevant and, indeed, an interesting perspective. But how would you deal with pixels on shore in that case? This is related to my enquiry above regarding the use of the median value to fill missing values, which differs from Deneu et al.

Bathymetry can be used by the model as a proxy for the land/sea divide. This means that the value of other variables on shore would not have any impact.

335: this sentence is actually true of all habitat models. You should make it a bit more specific.

We rewrote this sentence to clarify what is new in our method (lines 440-443).

## Specific corrections

Abstract: due to scarce observations -> due to the scarcity of observations

Done

Abstract: observations, and -> observations and : in general, there should never be a comma before and in an enumeration of two items. Please review throughout.

We removed all of these commas except in two sentences where a break is needed.

Abstract: 38 taxa which include -> 38 taxa comprising : "which" should introduce a sub-sentence and be separated from the main sentence by a comma.

Done

Abstract: mammals, as well as marine -> mammals, marine

Done

Abstract: this black box model -> this purely correlative model : black box is a bit of a catch-all term. The model is not so black box after all, since you can get insight into which variable is most explanatory. I would avoid the term

Done

Abstract: insight for species-specific movement ecology studies : why "movement" in particular?

Removed “movement”

16: climate, nutrient cycles, and biogeochemical cycles : remove nutrient cycles, it is redundant with biogeochemical cycles.

We put it within brackets in accordance with another reviewer’s suggestion

22: To focus on solving -> To solve

Done

22: it is essential to -> a necessary first step is to

Done

23: the open oceans -> the open ocean. "Open ocean" is a general term here, not designating any single ocean in particular. Change everywhere.

Done

37: There is a wide -> A wide variety of Species Distribution Models (SDMs) have been discussed

Done

38: environmental niche: the area where : The niche is defined in environmental space; it does not designate a region, it designates a set of conditions in which the species thrives.

Removed

40: with specific environmental conditions : I don't understand what this part of the sentence means.

Removed

41: where the prediction is computed -> where the observation is recorded

Done

43: area does not -> area would not

Removed

44: seamounts or trenches. -> trenches for example.

Done

46: these spatial structures are essential to understanding species distributions -> these spatial structures represent processes essential for determining species distributions

Done

48: include the values of these environmental data -> include the environmental data

Done

49: variables -> predictors

Done

51: summarize input data as fewer significant variables -> summarize input data into fewer relevant variables

Done

51: made manually -> carried out manually

Done

60: invented -> designed

Done

62: image classification -> image processing

Done

66: objects is probably enough

We left animals and plants to provide explicit examples.

71: at the point of occurrence -> at a given point : this is relevant also for prediction

Done

72: temperature fronts -> fronts : there are salinity-driven fronts which are equally important

Done

81: relies only on environmental data : isn't Sentinel-2 data environmental data? Rephrase.

Done

83: we explore and report the possibilities -> we explore the possibilities (or we explore and report *on* the possibilities; but explore is probably enough)

Done

91: to build a model to link environmental data to species presence. -> to build a model to relate species presence to environmental data.

Done

92: we used freely available occurrence data -> we used occurrence data : the env data is also free and nothing special was mentioned about it.

Done

Fig 1: Point grid -> Grid

Done

102: large pelagics -> large pelagic fishes (or something else but pelagic is an adjective, it requires a noun).

Done

103: delete "These taxa may be replaced or complemented with others in the future"; you say it in the conclusion and it has its place there.

Done

108: move "Furthermore, convolutional neural networks are known to be 109 robust against occasional labelling mistakes [19]." before "We removed..."

Done

111: When there were more than 10,000 occurrences of a taxon, a random sample -> When more than 10,000 occurrences of a taxon were available, a random sample : overall, "there is" is to be avoided in written text; please check throughout.

Done. Removed everywhere except in one sentence.

126: and made available -> and is made available (and congrats again for packaging this and making it available!)

Done. Thank you!

Table 2: please group similar variables together (e.g. SST, SST -5d, SST -15d). Please mention from which type of source Eddy kinetic energy is computed (I assume models).

EKE is computed using geostrophic current. We added this information in the caption. We also grouped SST and Chlorophyll.

Fig 3: like in the table, keep related variables together, to ease comparison.

Done

178-180: those are results and should be placed there; possibly in the section about quality assessment of the model (3.2).

Done

178: those statistics are on the test set? If so, state it explicitly.

Yes they are, we added this information

182: easily identified -> well predicted

Done

Fig 4: formatting numbers >0 with leading zeroes (001 instead of 1) is slightly misleading visually (the number of errors "looks" similar between 001 and 999)

There are no leading zeros, those numbers are decimals (0.01 to 0.99). These numbers are here for reference but the intended indicator is the darkness of cell colour (proportional to the number). We added a sentence to the caption to clarify.

Fig 6: what do you mean by "the contrast was increased"? Does it simply means the colour scale is not the same as in Fig 5? If so, use a different colour, it will make it more natural. Is it still the same for all panels of Fig 6?

Yes, the scale is different from other figures, but it is consistent for all subfigures. We changed the colour and clarified the caption.

Table 4: Redefine the quantity displayed in the legend of the table. Also, the values are sorted by median but the mean and standard deviation are also reported. If the distribution of values is such that the median makes more sense than the mean, then report only the quantiles (not the mean) and the median absolute deviation (instead of the standard deviation).

We updated the table with more accurate results (computed over the whole dataset), and reported the statistics you suggest.

267: The variables that were identified are -> The variables that were identified as important are

Done

268: important movement predictor -> important predictor of movement

Done

269: I suggest adding the part in brackets: "Sea surface temperature was also expected to be an important predictor, [since it has important physiological consequences and is therefore] the most frequently used descriptor in." SST is not important because it is widely used; it is widely used because it is important.

Done

273: two final periods. Remove one.

Done

286: what does "depending on the ecological context" mean?

We added a few sentences to clarify (lines 351-354).

322: I think what you mean here is sourcing from different datasets within GBIF instead of randomly. But this is not obvious and should be rephrased.
But, again, binning would be a better way than resampling and would correct some the sampling bias.

Indeed, we rephrased to clarify (line 400).

335: species distribution at all and all areas -> species presence at all dates and all areas

Done

## Conclusion

Overall, while this work is a valuable contribution, has the potential to be a very interesting one, and could prove seminal for the future of such approaches, I cannot advise for recommending it at this point. At the very least some points need to be discussed more in depth and it is likely that some computation needs to be added/redone.

I would be happy to interact with the authors if some of my remarks are not clear enough.

Thank you for this very thorough review, we hope that we answered most of your concerns.

Review #2

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 |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 |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 |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?

These species were selected to appear in the article because the distribution maps had interesting features to comment (positive or negative). As you mentioned, all other maps are available: we are not hiding anything.

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

The discussion is now much more in-depth and includes numerous references.

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 |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.

We’ve added the number of occurrences for each taxa. We prefer not to add the period of years to avoid overloading the table, but this information is easily available in the csv file we provide

-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 1species/taxa?)

We rephrased this paragraph to clarify: co-occurrences are not a problem, the model converges towards a middle-ground (lines 217-220).

-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.

We added the number of occurrences in Table 1, and a comment in the text (line 277).

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.

We considered the ocean as 2D. This was added to subsection *Description of the environmental data* (line 147).

- 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)

The environmental values are the values at the date of occurrence (line 107).

- 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.

The environmental values are the values at the date of occurrence (line 107).
We provide the source for each variable in table 2, which includes their unit and many more information. We chose not to duplicate this in the table to keep it readable.

-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 ifyes 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.

Again, data sources are available in the table. We added that we used surface values (line 147). We also added the spatial & temporal resolutions of variables to the table, as they are the most important metadata in our context.

-Space and time association isnot clear inthe Method section.

We do not understand this comment.

2) Problems that have been identified in the predictions:

-L183: the confusion matrix also indicates ahigh 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.

As you point out, these two taxa have similar distributions. Therefore it is expected of the model to have a high confusion rate between the two.

-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.

Thank you for this remark. We had not noticed this. It seems that the predictions for Acropora work well on the test data set, but not on the world grid, which might be due to auto-correlated data in the GBIF dataset. We added a paragraph on that in the discussion (lines 415-422).

- |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.

This was our first intention, but it rapidly appeared that observation data is heavily biased (which is the reason why we need this kind of models) (reference: https://doi.org/10.1890/07-2153.1). For example, according to OBIS data (https://obis.org/taxon/137092), humpback whales are all at the surface, except a small

number that are between 20 and 30 meters from the surface. Obviously this kind of data cannot be used to judge the validity of our predictions. This is why we resorted to established ranges, as in section 3.3. Even this solution is far from perfect, with maps that often contradict others of the same taxon, and for some taxa very hard to find if they even exist.
We agree that this is far from ideal but we have yet to find a better way to validate our predictions.

- 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), |am not sure that these predictions are reliable. Please explain.

Our method does not forbid co-existence of several species. We rewrote this section of the methods to clarify (lines 217-220). We also added a paragraph on this in the discussion (lines 347-350).

3) Limited relevance of the "SST+2°C" scenario:
|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 variables", |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)

This section was removed completely, as suggested by several reviewers.

4) Choices of 38 taxa:
- |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.

We cannot add or remove species at this point, as it would require months to download and process the data, as well as retraining the model and regenerating all the results. However we thank you for your suggestion and we will include plankton in a future version of this work.

-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.

We do not study movement in particular, but it is an important factor to justify that distributions should be modelled dynamically. We clarified this in the *Existing methods for predicting species distributions* subsection of the Introduction (lines 56-63).

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?

We did not find examples of dynamic SDMs that show movement. The closest we have is information on Humpback Whales migrations, which we mention in the *Undetected patterns* subsection.

- 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.

We expanded on that comment (lines 332-334).

Limits of the study:
- L283: "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.

The model can predict several taxa, this was clarified in the Methods (lines 217-220) and the Discussion (lines 347-350).

- 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.

We now mention planktonic species in the *Suggestions* subsection (line 403).

- Line 302: "The strength of deep learning in this context is that it makes no assumption when there is no data: itreplicates the results from similar well-known areas." |disagree since other SDM methods using pseudo-absences (rebuilt from the available data) also do this.

We agree. This sentence was removed.

-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.

We changed the sentence to conditional phrasing (line 381). We cannot check 315,000 occurrences.

-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 discussion now includes many references.

-The justification of considering monthly habitat mapping should be clearly explained for species exhibiting migrations.

We consider dynamic habitat: predictions can be computed at any point in time. We clarified this in the introduction (lines 55-56).

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.

We believe that, in the absence of baseline data (cf. our previous answer on checking against OBIS data), our only choice is to allow experts to criticize our results, like you did with Acropora data. This leads to investigating issues and proposing fixes for subsequent studies.

We added this into the conclusion (lines 465-468).

Minor comments :

Line 16: |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)".

Done

Lines 15 and 17: Beware that references #1 and #2 are on the deep sea.

We replaced the citations with more appropriate ones.

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

Done

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

This sentence was rephrased. Thank you for the references.

Line 47: citation #12 refers to a model. |suggest to cite another citation referring to observed patterns, even ifnot 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

Done

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." |would argue that is it also true for non mobile species like plankton.

This paragraph was rewritten.

Line 105: essential to ?

We do not understand this comment.

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 ?

Both options were possible. Because of the small number it does not make a significant difference.

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

The model we present is not based on niche theory (we added a paragraph in the introduction to clarify that, lines 56-58). Some of our taxa are genera because grouping species was need to provide enough data points.

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

This is to reduce the number of occurrences to keep calculation time reasonable. Now the table shows the number of occurrences per taxon.

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.

We clarified in the introduction that this method provides dynamic predictions rather than static distributions (cf answers to previous comments).

L226: "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.

We picked species with interesting comments to be made. As you can see in the article, we do not hide negative results, cf. our comments on Acropora, Humpback whales, Thunnus tynnus, etc.

establised => established

Done

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?

We added the dates on the right panels and the references in the main caption.

We added a comment that states using other dates yields the same comments (maps are public).

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

We chose to order this table by median as we believe it is represents better the importance of variables. We changed the statistics according to another reviewer's comments.

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

Done

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

We added the main messages in the text (lines 316-319).

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

This section was removed, as suggested.

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

We added information in almost all the captions.

Review #3

Thank you to the authors for an interesting and exciting article and to PCI for the opportunity to review it! Additionally, I'd like to thank the authors for the impressive commitment to making their code, data, and manuscript publicly available. My review is below.

### Summary of Article

The authors present a multi-species, temporally explicit SDM built using typical CNN methodologies for a broad suite of 38 marine species and genera using data from GBIF. The CNNs were built using 25 environmental variables at varying spatiotemporal resolutions and the output layer of the model was configured to output the predicted probabilities of each taxa, which were then interpreted as two primary forms of species distribution prediction. Finally, the authors conduct a variable importance analysis of the environmental variables using integrated gradients. They conclude with a discussion of potential future work and improvements for the model and their results.

### Summary of Review

The authors' article contributes to a hot topic exploring the application of deep learning to species distribution modeling. There is, I believe, a common philosophy that DL (and specifically CNNs) should be a clear winner for SDMs and a reasonable value proposition in understanding which domains are more or less suited to being modeled via CNNs and still little collective understanding of best practices when following this approach. I have some suggestions for critical details that should be described in the manuscript, and key questions about the interpretation of the results, but I support the fundamentals of what the authors have done here and believe it could be a particularly valuable contribution to the literature even if only my simpler feedback was addressed.

Overall, I'd recommend this article be revised and resubmitted.

### Major feedback and questions for the authors

* Did you use a pretrained resnet-50 model? If so, which?

No, we trained it from scratch. We added this in the « Training the model » subsection (line 202).

* How was your model extended to a 29d input layer? This is not a trivial extension of the Deneu or Botella models and the new architecture should be described more completely.

We only changed the first and last layers. We added this into the methods (line 201).

* Additional detail is needed on the alignment process for environmental variables, particularly for variables that were downscaled. It should be clearer throughout the manuscript what resolution was being modeled.

There is no alignment process because the same geographical buffer is used for each variable download. We added the final grid resolution (7km) to the *Enrichment* subsection of the method (line 163).

* I have a fundamental issue with the treatment of model predictions as a multi-class classification problem, including the one-hot encoding of training data and the interpretation of the predicted class probabilities. In particular, the "accuracy" of predicting the most likely species within each pixel follows the GeoLifeCLEF problem formulation and is fundamentally problematic, particularly when modeling with presence-only data from GBIF. Additionally, treating the predicted class label probabilities as a surrogate for suitability or RPoO and presenting spatial visualizations of these probabilities as a snapshot SDM is a misinterpretation, in my opinion.

Treating the problem as a classification task allows estimating the conditional probability of y (the observed species) given that an observation has been made in the environment x. It has the advantage to be (asymptotically) invariant to the spatial sampling effort but it is sensitive to the taxonomic reporting bias (the fact that some species are more observed than others). In the absence of taxonomic reporting bias, the estimated probabilities would converge to the relative probability of each species given the environment. Mapping those probabilities is thus equivalent to mapping the species suitability relatively to the other species suitability. It can be related to the "target-group" approach for generating pseudo-absences. The species are considered to be absent at the occurrences of the other species. In the presence of taxonomic reporting bias, it is unclear how this affects the maps. If the bias is constant over space, this affects the map of each species only by a constant factor but if the bias is more complex, it remains an open problem (as for most SDMs).

We added this into the text (lines 206-214).

This space is still being defined within the literature, so I wouldn't object to publishing this treatment, but I would encourage the authors to at least address this treatment of presence-only SDM data in their discussion of the interpretability of the distribution predictions.

We clarified that both in the introduction (line 89) and in the discussion (lines 345-350).

* The authors dismiss the need for data cleaning because of the robustness of CNNs to mislabeled or erroneous records. However, recent studies (Zizka et al., 2020) have estimated the proportion of problematic GBIF records to be as high as 41-44% of all records. I'm not aware of a paper which has investigated the robustness of CNN-SDMs to GBIF errors and so would encourage the authors to reconsider this lack of data cleaning for their model.

We removed the sentence where we dismiss the need for data cleaning. Instead, we acknowledge that we made this choice and that data cleaning may/should be conducted in future studies (line 133).

* I like the description in §1.3 of how CNNs pool feature detectors at varying levels of coarseness and how that might be paralleled in a climate/environmental SDM. However I

think the language of what is happening as the models learn different weights for the convolutional layers is imprecise and could mislead readers.

We added a sentence to clarify (line 70).

It would also make an excellent conceptual figure, space and time permitting.

Thank you for the idea. Unfortunately we do not have time for this, but hopefully in the future.

### Minor feedback and line items

- §2.7: Why were some variables removed apriori before variable importance?

We only removed variables that were highly correlated with others, in order to have a proper view of their influence. Otherwise, they would share the influence between them and falsely appear less important.

- "Figure X", "Table X", and "Section X" should generally be capitalized throughout the manuscript (e.g. on lines 103, 111, and 115)

Done

- The term "probabilities predictions" is used a few times throughout the MS (e.g. L87) and should be replaced with "predicted probabilities"

Done

- L51, "This work may be made manually": doesn't scan for me, perhaps prefer "These summary variables may be constructed manually by experts, ..."

This was already rephrased following advice from Review #1

Again, thank you to the authors for the opportunity to review their impressive project and I look forward to seeing it in print soon!