Response to Reviewer 2 comments

Comments for Basher et al. (2018) - ESSD Discussion

The present manuscript aims to present a « novel » digital atlas of environmental (meaning physical, chemical, biogeochemical) climatologies, from which scientists may download numerous environmental layers that are typically used for developing spatial statistical models, such as species distribution models (SDMs). The authors did a fine job in compiling many published datasets, and gathering all of them in a homogeneous and central atlas. Consequently, the Global Marine Environment Dataset (GMED) is the online platform with the widest range of environmental layers. The GMED proposes environmental data at a finer spatial resolution compared to previous comparable atlases (mainly AQUAMAPS, MARSPEC and Bio-ORACLE). The GMED also supplies past and future fields for some of the environmental layers (temperature, salinity, ice cover), thus allowing the community to quickly test « long-term » changes in species distributions and diversity. Consequently, it might attract marine ecologists aiming to easily model the niches and distributions of marine taxa, whether those are benthic, pelagic, coastal or inhabiting offshore conditions. But that might also be an issue as I will develop below.

In spite of the added value of the dataset might present, I have identified some major points that may help improve the completeness, the quality of the atlas and the manuscript. I will now give my step-by-step review of the manuscript and data access based on the ESSD review guidelines. Then I will detail my major concerns and comments regarding the GMED itself.

We thank the reviewer for the positive and important comments. We appreciate the inputs on how to improve the overall quality of the manuscript/ data repository with additional datasets by the reviewer. We will implement some of these steps immediately and some in near future to improve the overall quality of the GMED data layers.

1. The data presented consist of a compilation of pre-existing datasets so the data themselves are not « new », but the atlas is clearly more exhaustive than previous and comparable ones, even though some of the data used are clearly outdated (but see major comments below). Also, the data presented here have been interpolated to follow a higher resolution grid, so they represent an improvement for end users (SDM users). I do like that the authors added variables such as distance to land, or to closest port, because these often have to be calculated separately and can be very useful to account for sampling biases in marine species distributions. Therefore, I agree that this dataset could be useful for future studies. Although I consider the methods description to be thorough, I do miss proper uncertainty estimates in the layers provided online, and especially for the past and future environmental layers (but see major comments below). For

controlling the quality of the data, the authors completely rely on the controls undergone y other authors for the first publication of the data compiled. Furthermore, the perform some sort of completely circular cross- validation to control the output of their interpolation. Proper quality control would at least require some independent data. Therefore, they do not really perform « quality control » in my humble opinion. As a result, key information are missing for the reader about the way the environmental layers were developed initially. Here, it is not sufficient to simply state that « *All of the primary datasets used in the GMED compilation had undergone quality control checks by the primary data collectors and processors* ».

Thank you for the understanding and comments about the value for the compilation. The principal objective of GMED is to make marine datasets readily available to ecologists for their modeling work, to reduce the lengthy process of data compilation and standardization. We understand the issue with circulation validation, we would very much like to provide an uncertainty estimate to end users. However, as no other extensive independent global dataset was available at the time when GMED was created, for that reason we only provided a validation method to ensure users we did not introduce any new errors to the new dataset with our interpolation process. We will use the independent data layers you mentioned and by another reviewer to create uncertainty estimates for all GMED data layers in future.

2. I have identified some gaps in the niche modeling literature (lines 53-59) that I would like the authors to address carefully because I think it may have lead them to forget important predictors in their atlas (but see major comments below). I would also like the authors to mention the update of the Bio-ORACLE v2 dataset (Assis et al., 2018 - DOI: 10.1111/geb.12693) in their manuscript. Yet I acknowledge it might have been published after the authors finished their atlas.

Thank you for mentioning about the new BioOracle dataset. We are aware of the recent this recent update, and will add the citation in the revised manuscript.

3. The dataset is easily accessible via the online portal. I had no problem downloading, unarchiving and then reading the data with R. The files are encoded in ASCII, which is easy to read with the « *raster* » R package (Hijmans (2017). *raster: Geographic Data Analysis and Modeling*. R package version 2.6-7. https://CRAN.R-project.org/ package=raster). I honestly have no experience with reading ASCII tables with other commonly-used languages, such as Matlab or python, but I am convinced the people concerned will not have too much trouble with that. The online dataset seems complete regarding to what is described in the manuscript. However, I am missing error estimates and/or quality flags in the data tables provided. For now, the only « quality flag » consists in the existence of a cropped version of the layers (at 70°N and S

because of the satellite data). I would like to know whether it would be possible to add uncertainty estimates (linked to initial observations density biases, or model uncertainties when models are part of the process) to the data tables so one can identify where the less reliable data are geographically located (especially for biogeochemical and future fields)? Nevertheless, I would say the data cleaning, treatment and comparison to previous similar datasets are adequate: what the authors did is clear and well described. One of the authors' main claims is that the finer resolution of their layers should lead to more reliable SDMs compared to previous products. Although I agree this should be the case, a formal test of this assertion is needed (like developing a few standard SDMs for a virtual species based from the present data and then compare them to SDMs built from the previous atlases). But I am not sure this is within the scope of an ESSD paper, which focuses on the data itself.

Thank you for the complement. Several studies have conducted SDM with empirical data and using GMED and we will now cite these in the MS, namely: Jayathilake et al. 2018, Asad et al. 2017, Basher et al. 2016, Saeedi et al. 2016

4. Overall, I find the dataset to be usable in its current format. Maybe others would prefer to be able to download it as a text file (.txt or .csv). The metadata are provided in the appendices and can easily be found online (http://gmed.auckland.ac.nz/layersd.html). The issue when compiling pre-existing datasets is that one may simply refer to the original publication of the data for the full metadata. These are not provided in the present manuscript but the authors do refer to the original publications and website (like in any other publication of this nature, see Tyberghein et al., 2012). Overall, the language and the figures are of good-quality in my opinion. However, I strongly recommend that the authors provide a quantitative scale and the appropriate units with the maps, instead of the rather arbitrary « low » and « high ». I think there is a typo in the caption of Figure A53: *Temperature 1AB Scenario*? This really looks like a salinity map. Also, in Figure 1, the authors need to clearly indicate which steps they performed themselves. Indeed, they did not compile all the satellite/ model/ Therefore, from what I understood, the data processing steps actually performed by the authors are those indicated by the 4th and 5th arrows (after « Raster Grid »). To summarize: the authors interpolate the older layers on a new and finer grid, and then evaluate the interpolation's output by computing variation coefficient and standard error between those and the initial layers. By doing so, the authors do not claim to actually control the quality of the data, but rather the « *interpolation quality* ». I think the authors are right in stating so, but I do find the process a bit circular…

To conclude, I do think the data presented here are complete and could be useful to quickly run and test some SDMs. It does comprise a very comprehensive compilation of different environmental variables that are commonly used as

predictors in species distribution modeling. However, I cannot conclude that the data presented here are « unique ».

Thank you for pointing out about the misplaced figure A53. Correct figure of Temperature will be included in the revised manuscript. We would like to iterate we performed all the steps in Figure 1, from Raster Grid onwards. Data was sourced from both existing data layers as well as satellite/modelled data from other sources (i.e., two of the temperature and salinity data layers were sourced from NASA PSD reanalysis datasets via Giovanni interface). We would very much like to have a data quality check step added to the current process but as mentioned earlier due to the lack of comparable global dataset we had to settle with the interpolation quality assurance only.

Major Comments

1) Mixed-layer depth and variables averaged over the mixed-layer?

In the introduction (l. 53-59), the authors rightfully state that SDMs have been relatively less used for studying marine taxa compared to their terrestrial counterparts. Then, they mention the marine groups that have been studied through SDMs with the associated literature. Here they fail to mention the recent (and less recent) studies that performed niche modeling for the marine plankton (both phytoplankton and zooplankton), apart from Bentlage et al. (2013) whom quickly performed SDMs using climatologies from before 2005…Haphazardly, you should mention some of the following studies:

Beaugrand, G. & Helaouët, P. (2008) Simple procedures to assess and compare the ecological niche of species. *Marine Ecology Progress Series*, **363**, 29-37.

Beaugrand, G., Edwards, M., Brander, K., Luczak, C. & Ibanez, F. (2008) Causes and projections of abrupt climate-driven ecosystem shifts in the North Atlantic. *Ecology Letters*, **11**, 1157-1168.

Beaugrand, G., Lenoir, S., Ibañez, F. & Manté, C. (2011) A new model to assess the probability of occurrence of a species based on presence-only data. *Mar. Ecol. Prog. Ser*, **424**, 175-190.

Reygondeau, G. & Beaugrand, G. (2011) Future climate-driven shifts in distribution of *Calanus finmarchicus*. *Global Change Biology*, **17**, 756-766.

Irwin, A.J., Nelles, A.M. & Finkel, Z.V. (2012) Phytoplankton niches estimated from field data. *Limnology and Oceanography*, **57**, 787-797.

Beaugrand, G., Mackas, D. & Goberville, E. (2013) Applying the concept of the ecological niche and a macroecological approach to understand how climate influences zooplankton: advantages, assumptions, limitations and requirements. *Progress in*

Oceanography, **111**,
75-90.

Chust, G., Castellani, C., Licandro, P., Ibaibarriaga, L., Sagarminaga, Y. & Irigoien, X. (2014) Are *Calanus* spp. shifting poleward in the North Atlantic? A habitat modelling approach. *ICES Journal of Marine Science: Journal du Conseil*, **71**, 241-253.

Pinkernell, S. & Beszteri, B. (2014) Potential effects of climate change on the distribution range of the main silicate sinker of the Southern Ocean. *Ecology and Evolution*, **4**, 3147-3161.

Villarino, E., Chust, G., Licandro, P., Butenschön, M., Ibaibarriaga, L., Kreus, M., Larrañaga, A.
& Irigoien, X. (2015) Modelling the future biogeography of North Atlantic zooplankton communities in response to climate change. *Marine Ecology Progress Series*, **531**,
121-142.

Brun, P., Vogt, M., Payne, M.R., Gruber, N., O'Brien, C.J., Buitenhuis, E.T., Le Quéré, C., Leblanc, K. & Luo, Y.W. (2015) Ecological niches of open ocean phytoplankton taxa. *Limnology and Oceanography*, **60**, 1020-1038.

Barton, A.D., Irwin, A.J., Finkel, Z.V. & Stock, C.A. (2016) Anthropogenic climate change drives shift and shuffle in North Atlantic phytoplankton communities. *Proceedings of the National Academy of Sciences*, **113**, 2964-2969.

Brun, P., Kiørboe, T., Licandro, P. & Payne, M.R. (2016) The predictive skill of species distribution models for plankton in a changing climate. *Global Change Biology*, **22**,
3170-3181.

Benedetti, F., Vogt, M., Righetti, D., Guilhaumon, F. & Ayata, S.-D. (2018) Do functional groups of planktonic copepods differ in their ecological niches? *Journal of Biogeography*, **45**,
604-616.

But more importantly: in several of these publications, SDMs were developed using mixed-layer depth (MLD) as a predictor, or other variables (PAR, SST, Chlorophyll concentration) integrated over the mixed layer. The noteworthy paper of Brun et al. (2015) even identified MLD as the most important variable for modeling the niches of phytoplankton species. MLD greatly contributes to the temperature, light conditions and nutrients dynamics perceived by the plankton, the basis of evert marine food-web. Its role in controlling Ocean-Amosphere heat fluxes and in shaping bloom dynamics has been studied for decades now. It should always be considered as a potential predictor even for fishes and/or top predators because of its probable effect through bottom-up processes. Overall, MLD is arguably one of the most important oceanographic variable so I was extremely surprised not to see it among the variables compiled. Why is that?

I highly recommend that the authors add at least one MLD product to their atlas. The most recent one I can think of would be: Holte, J., Talley, L.D., Gilson, J. & Roemmich, D. (2017) An Argo mixed layer climatology and database. *Geophysical*

Research Letters, **44**, 5618-5626. Which can be found here: http://mixedlayer.ucsd.edu/

I also encourage the authors to compute mixed-layer averages for several other variables such as temperature, irradiance, salinity, nutrients concentrations, Chlorophyll-a concentration etc. It seems like the authors do not benefit from the recent wave of Argo floats data. Which brings me to my second major point.

Thank you for the suggestions about SDM references and MLD dataset. We consider GMED as a living repository which will be updated with data layers as they became available over time. As suggested, we will add the references to the revised manuscript and add MLD product in next update of GMED dataset.

2) Outdated data sources.

While reviewing the sources of the data compiled, I was surprised that many of the layers still rely on data from the World Ocean Atlas of 2009, or from the Sea-WiFS satellite era. Since then, the World Ocean Atlas has undergone not one but two updates (it is currently at the WOA 2013v2 stage: https://www.nodc.noaa.gov/OC5/woa13/) and the MODIS- Aqua sensor has been operational since 2002. The WOA 2013v2 provides monthly/ seasonal/ annual climatologies at a 5°, 1° and sometimes 1/4° resolution, with standard depth levels, and with detailed and proper quality controls. I am very surprised the authors did not take the time to assimilate these layers since they are widely known in the oceanographic community.

For other chemical and biogeochemical variables, way more recent and valuable datasets can be found in ESSD:

https://www.earth-syst-sci-data.net/7/261/2015/essd-7-261-2015.pdf https://www.earth-syst-sci-data.net/8/325/2016/essd-8-325-2016.pdf
https://www.earth-syst-sci-data.net/8/297/2016/essd-8-297-2016.pdf https://www.earth-syst-sci-data.net/8/383/2016/essd-8-383-2016.pdf
And, of course, updated and controlled observations and re-analyses can be found on the Copernicus data portal: http://marine.copernicus.eu/services-portfolio/access-to- products/

I am a bit uncomfortable as I do not want to dismiss all of the work carried out by the authors, but I must *strongly* encourage them to go through all these data products and update their data sources. Otherwise the community is just given recycled and outdated data products that do not not reflect the state of the art, nor the efforts of the climate and ocean scientists. This point is also valid for the past and future environmental layers provided in the GMED, which brings me to my third major comment.

Thank you for the suggestions. We are aware of the WOA 2013 and even the most

recent WOA 2018. As GMED was initially created in 2013, released on 2014, we did not have opportunity to incorporate these recent data layers into the repository. We will be incorporating all these new data layers progressively in future updates of the repository, and most importantly, noting to users that new versions of the primary data are available which when processed would improve the quality of the maps..

3) Fields of future environmental conditions.

One of the reasons why SDMs got so popular in the last 20 years is because they allow to handily explore temporal changes in species distribution, and therefore diversity, following climate change (greenhouse gas emissions actually) scenarios. Knowing that, the authors added some predictions of SST, SSS, seabed temperature and salinity, primary productivity and ice concentration. This could have been interesting had the scenarios not been completely outdated. Indeed, the data compiled here were issued for the 4th AR of the IPCC (CMIP3 exercice). I do not believe the authors are unaware of the existence of the IPCC's 5th AR which presents Representative Concentration Pathways (RCPs) that are now the standard when it comes to model climate change impacts. I know, from my own experience, that RCPs data are not always available for regional models, but this is definitely not the case for the global ocean. Proof is that even the latest version of Bio- ORACLE (Assis et al., 2018 - DOI: 10.1111/geb.12693) provides RCPs outputs, with some uncertainty estimate across AOGCMs. Why did the authors not consider the latest standards?

The authors fail to provide crucial information about model set-up, calibration, configuration, validation, bias correction…The two links provided in the references below Table 1 are not functional. Do the layers presented correspond to the absolute fields obtained for the 2090-2100 period? Or to model biases between the contemporary period and the end-of-century period that were added to the observation-based climatologies? What are the uncertainties within each projection? And then between projections? Why did the authors rely on just two models (IPSL and HadCM3) among all the existing ones? Why are future surface temperature and salinity given for two emission scenarios but not seabed temperature? What is the model configuration that generated the future PP product? There are *tremendous* uncertainties across the suite of coupled ecosystem models that can provide biogeochemical projections (just have a look at Laufkötter et al., 2015 - doi:10.5194/bg-12-6955-2015), and this is well known in the community. I am sorry but these future layers cannot be used as of now. The choice of the climate model can make up a significant part of the uncertainties in SDM-based climate change predictions. Please see:

Diniz-Filho, J.A.F., Bini, L.M., Rangel, T.F.L., Loyola, R.D., Hof, C., Nogués-Bravo, D. & Araújo, M.B. (2009) Partitioning and mapping uncertainties in ensembles of forecasts of species turnover under climate change. *Ecography*, **32**, 897-906.

Buisson, L., Thuiller, W., Casajus, N., Lek, S. & Grenouillet, G. (2010) Uncertainty in ensemble forecasting of species distribution. *Global Change Biology*, **16**, 1145-1157.

Garcia, R.A., Burgess, N.D., Cabeza, M., Rahbek, C. & Araújo, M.B. (2012) Exploring consensus in 21st century projections of climatically suitable areas for African vertebrates. *Global Change Biology*, **18**, 1253-1269.

The risk is that young scientists might implement SDMs based on the contemporary layers provided the GMED, with the default settings user-friendly modeling platforms, and simply project those in the future conditions, without any prior knowledge about the way the data were produced…Knowing the data you use (meaning understanding where it comes from, its limitations and quality, the uncertainties associated) is a crucial part of any modeling experiment, and species distribution modeling is not an exception. And this brings me to my fourth and final major comment.

Thank you for pointing out about the invalid link below Table 1. We will update the correct URL in the revised manuscript. Regarding the quality assurance for projected data layers, all the future data layers were originally published either by Bio-Oracle v1 or AquaMaps (mentioned in the appendix and http://gmed.auckland.ac.nz/layersd.html webpage), and we compiled the LGM data layer ourselves from the primary data source. As discussed in the manuscript all details about the model setting, calibration, and other information should be available from the meta-data from AquaMaps and Bio-Oracle website. We will include these references and link to metadata next to the future data layers in the revised version of the Manuscript. We will also include a warning to the website stating the some maps will be updated using more recent and improved primary data sources (i.e., new RCP's for future data layers).

4) Compilation of environmental predictors takes time…and maybe it should do so.

The manuscript's abstract stipulates the following: « *Marine environmental datasets available for species distribution modelling (SDM) have different spatial resolutions and are frequently provided in assorted file formats. This makes data assembly one of the most time-consuming parts of any study using multiple environmental layers for biogeography visualization or SDM applications* ». I assume this motivated the authors (but others also) to implement user-friendly and publicly available compilations of environmental data to facilitate (and accelerate) the process. This could make sense when the quality of the data used and therefore the whole procedure is not affected. But I am confident this cannot be the case when using the present GMED (because of all the previously mentioned reasons).

Instead, I argue it is crucial that students and young scientists take the time that is required to: (i) review the environmental datasets available; (ii) thoroughly examine their origins (metadata), advantages and limitations; and (iii) investigate how alternative choices in the environmental data impact final SDMs outputs. How are they supposed to perform state of the art modeling if they do not even understand the ins and outs of the data they use? Data assembly is time-consuming because it

is the process that will determine data quality and thus the quality of any SDM projection. The identity and resolution of the environmental predictors available and suitable for a niche modeling exercice depend on its goals (testing for niche overlap, species distribution visualization, climate change impacts projections), and the type of biological data available (abundance, presence only, presence-absence etc.). I totally get that our community is experiencing increasing pressure because of competition for fundings, pressure to publish, and demands from stakeholders to provide climate change predictions, and therefore tries to gain time when possible. But simplicity and easy-to-use products should not take over the quality that any scientific experiment is entitled to.

To conclude, I would like the authors to know that I am truly sorry that I could not provide more positive comments. I hope they will take it as an encouragement to deeply re- organize and actualize their data so they comply with the quality required for any ESSD dataset. I encourage them to work more closely with oceanographers and climate scientists to help them find better and updated marine environment data.

We completely agree with the comment and objective of making quality data available for rapid SDM applications. With GMED we want to approach towards that direction. As suggested we will work to include some validation statistics as suggested by the reviewer and incorporate inputs from oceanographers and climate scientists whenever possible to improve the overall quality of the distributed data layers over time. We will also include warnings where appropriate in the website to notify users about datasets which might be available with newer model outputs (i.e., IPCC AR5 RCP's) from primary data providers.

Dr. Fabio Benedetti
ETH Ziirich, 0-USYS, IBP, UP Group.
On the 03/10/2018.

Cited References

Jayathilake, D. R. M., & Costello, M. J. (2018). A modelled global distribution of the seagrass biome. Biological Conservation, 226, 120–126. doi:10.1016/j.biocon.2018.07.009

Saeedi, H., Dennis, T.E., Costello, M.J., 2016. Bimodal latitudinal species richness and high endemicity of razor clams (Mollusca). J. Biogeogr. 44, 592–604. https://doi.org/10.1111/jbi.12903.

Asaad, I., Lundquist, C. J., Erdmann, M. V., & Costello, M. J. (2017). Ecological criteria to identify areas for biodiversity conservation. Biological Conservation, 213, 309-316.

Basher, Z., & Costello, M. J. (2016). The past, present and future distribution of a deep-sea shrimp in the Southern Ocean. PeerJ, 4, e1713.