We have read the reviews of our submitted ESSD manuscript "PlanktoShare: A large (50k+) and FAIR learning set for the Plankton Imager (Pi-10) for the Greater North Sea and NE Atlantic, based on a new flexible classification protocol". We thank the reviewers for their constructive and thorough comments. We think we can address the reviewers' comments and would like to submit a revised version. Below we present an overview of how we aim to address the reviewers' comments and suggestions.

Overall changes to structure

Both reviewers are positive about our paper, note its value as a data paper and support publication after appropriate revision. Upon reading the reviewers' comments we noticed that a substantial part of the comments is related to the example application of our learning set and classification protocol by training two ResNet 50 image classification CNN models. As we mentioned on line 85, our main aim is to present a flexible, standardise annotation scheme for plankton imaging data, provide a training set of Plankton Imager PI-10 images and describe a framework to support the long term management and development of this training set. In section 5 we added the CNN models as an example application of this framework and learning set. However, upon reading the reviewers' comments we think the CNN examples should be removed from the manuscript as the trained CNN models detract from the main focus of the paper. Removal of the CNN models is also suggested by reviewer 2 in a comment on line 85.

In the revision we will therefore remove the section and figures concerning the CNN models. Instead we will refer the readers to a Github repository where these models, including later improved versions, will be available. With this change we hope to resolve the comments of the reviewers related to the CNN models.

Please find below a detailed response to the comments by reviewer 1 and 2.

Response to reviewer 1 comments

Major comments

1. ***"The manual annotation and quality-control procedures should be described in greater detail.****The manuscript repeatedly emphasizes that the learning set is manually labelled and of high quality, but the current description of the annotation workflow, expert involvement, cross-checking, and consistency control remains limited."*
   This is a valid point. We will describe the annotation and quality-control procedures in greater detail by adding details on annotation workflow and detail in which parts of the workflow taxonomic experts were involved in quality control.
2. ***"The model evaluation relies mainly on random splitting, which may overestimate cross-cruise or cross-region generalization.*** *Images collected*

from the same cruise, time period, or nearby locations may be highly similar. A random train–validation split may therefore lead to overly optimistic performance estimates when the classifier is applied to independent cruises, different regions, different vessels, or different seasons. I recommend adding a lightweight but more informative validation, such as a blocked split by cruise, vessel, month, or region, or a leave-one-cruise-out test. If this is not feasible because of sample-size limitations or class imbalance, the authors should at least state clearly in the discussion that the random-split results primarily reflect within-dataset performance and should not be interpreted as general performance across all Pi-10 observational settings."

By removing the CNN models we consider this comment resolved. Note that each image in the learning set has coordinates so users can later decide on how to sample from the learning set.

3. ***"Class imbalance and the performance of rare taxa should be reported more transparently.*** *Appendix A shows large differences in sample size among classes: some dominant classes contain thousands to tens of thousands of images, whereas some rare taxa are represented by very few images. Under such imbalance, the macro F1-score is useful but insufficient to support the conclusion that all classes perform consistently well. I recommend providing class-wise precision, recall, F1-score, and support, preferably in a table or supplementary material. For classes with very limited support, model performance should be described cautiously to avoid implying that all classification units have equal reliability."*

   By removing the CNN models we consider this comment resolved.

4. ***"The representativeness and application boundaries of the dataset should be more clearly defined.*** *The spatial coverage of PlanktoShare is mainly concentrated along the Dutch coast, the Greater North Sea, the Celtic Sea, and selected NE Atlantic transects. This coverage is highly valuable for European marine monitoring, but the resulting classifiers may not directly generalize to other ocean regions, turbidity regimes, biogeographic provinces, or plankton community structures. I suggest adding a concise table summarizing the number of images and major class composition by cruise, vessel, year, season, and region."*

   We will add the requested table detailing the number of images by cruise, vessel, year and season and add a written description of the major class composition of the images for each cruise in the results section. In addition, we will emphasize that this information is available for each of the images in the database of the learning set.

Minor comments

1. *"In Section 3, the phrase "a specific? bespoke? learning set instance" appears to contain residual editing marks. Please remove the question marks and revise the wording."*
   Question marks will be removed and wording revised.

2. *"In the conclusion, the manuscript refers to "Fig. 7" when discussing the spatial coverage of the three vessels, but this should be Fig. 9. Please correct the figure number."*
   We will do a thorough check of figure numbering and references to figure numbers on the revised version.
3. *"The labels in the confusion matrices in Figs. 7 and 8 are very dense and difficult to read at the current page size. Please improve the resolution, enlarge the font, or move the full matrices to the supplementary material while retaining a simplified summary figure in the main text."*
   These figures will be removed.
4. *"4 shows the spatial distribution of image locations, but it would be helpful to distinguish different vessels, years, or data sources using different colours, so that readers can better understand the composition of the dataset."*
   We will improve figure 4 so that it visible where different ships collected which data.
5. *"Section 5.3 states that the OSPAR classifier uses 53,538 images, whereas the full learning set is earlier reported to contain 52,882 images. Please explain why the classifier training set exceeds the total reported size of the learning set, and clarify whether this reflects duplicate images, augmented samples, excluded/added categories, or version differences."*
   We will thoroughly check consistency of image counts throughout the revised manuscript and investigate any inconsistencies.

Response to reviewer 2 comments

Many remarks on the PDF. We will address the major remarks. Below we summarized the main concerns per section and write how we will address this.

"all figures need (major) updates"
Figures will be updated, addressing the comments provided in the PDF.

"the need for interoperability - perhaps mentions how the scheme can work for other sensors? We should indeed go beyound the mere 'taxonomic' scope."
We will expand the discussion to highlight the applicability of the scheme to other sensors. This will also contribute to improving then scientific background, commented on in a later comment?

"there is a strong need for additional tables and documentation."
We will add additional tables and documentation in response to the detailed comments in the pdf as well as as suggested by reviewer 1.

"there is a strong mismatch between a training set on on end (relatively well tackled); but on modeling, CNN, classificiatin tools on the other end which is not tackled at all and just comes in like that."
In response to this and other comments on the CNN modeling section, this section will be removed so the focus is on the training set and classification/labeling system.

"i have strong issues with the current architecture of the paper - it reads quite difficult and lacks a more conventient style (m&m, results and discussion, conclusion)."
We will evaluate on how re-structure the paper to have a more conventional structure. We think that re-ordering and adding main headers to the present content will likely already solve a lot of these issues.

"the numbers, figures, and actual images in the training set are not matching at all - or it is poorly described; or there are several versions of PlanktoShare which are not published as yet? This needs to be fixed. Similarly - a canonical record manifest, exact file formats, folder structure, identifiers, metadata files, checksums, licences and a version-specific DOI is entirely missing."
The dataset descrition and metadata will be checked thoroughly for the revision and expanded where necessary.

"i have strong issues for the lack of scientific backgorund. To elevate the paper to a worthy release, much more scientific background is needed (read and catch up on recent literature)."

Although it is a bit diffiult to judge on which aspects the Reviewer want additional scientific background, we will try to add additional recent references were possible.

"there are many typographic errors and remnant of draft versions throughout the manuscript (i stopped commenting them after a while i have to admit). I hope authors were equally involved in proof-reading as obvious typos are still there."
We apologize for this and will do a thorough check of typographic errors and remnants of draft versions for the revised version.

"consistent terminology. The manuscript alternates between database, learning set, training set, library and folder. These terms should be clearly defined and used consistently."
We will define the used terms more clearly upon their first use in the manuscipt, and check that later on in the manuscript the terminology remains consistent.

"the annotation quality controll eneds to be addressed in detail. Who, how, more then one expert? CNN only? I strongly believe the scheme proposed needs to be revised. On one end, to include such information (essential as in time we evolve to a non-human validated dataset); but also to avoid the ZERO/NA thing which is poorly explained?"
We will expand the method section to include additional details on the quality control of the dataset.

"The confusion matrices are too small and insufficient by themselves. Per-class precision, recall, F1 and support, as well as macro and weighted F1, balanced accuracy, confidence intervals and major confusion patterns..."
As mentioned above, the CNN section will be removed.

"audit the bibliography - quite some missing."
Apologies for the missing references, this should not have happened. We will do a thorough check of the bibliography for the revised version.

"explain way more thoroughly the applicability for users; and similarly (as said above) make sure to explain what is in the scheme and in what format (table!)."
In the discussion we will explain more thoroughly for which users our dataset and classification scheme is applicable.

"you speak of a database, but i see nothing hierarchical? how are thing coupled? how are images saved - are these in a noSQL server database? I have the feeling I "miss" something?"
We will use the term "dataset" instead of database to avoid confusion.

"the model software, Plankton Imager Classifier by geojoost; is not documented at all here."

The CNN model part will be removed from the manuscript as detailed above. We will include a reference to the geojoost repository as a possible application so prospective users can still find it.