Response to second review OS-2020-100 comments

Report #1

The example shown in Fig. 5 is very helpful but could be expanded easily to be of even more use. To better understand the effect of the MODE parameters that were chosen, it would be nice to show the matched objects, for example by using the same color for objects that match, by connecting them with lines, or by adding numbered markers that also show the order in which they are matched. As a reader, I'd be interested, for example, if AMM7v11 objects 7 and 12 (off the coasts of Denmark and Germany) remain unmatched or are matched with the big L4 object 10. Adding information about the matching will help readers understand the results in the following sections 4.3 which focuses on the fraction of matched objects and problems when matching objects.

This is now Fig 6 and the colours have been adjusted so the matches are evident. Anything unmatched is coloured grey.

Another example of where an example could be expanded is Fig. 11: It would be nice to use the same colors as in Fig. 10 here, so that the bloom events can be related to each other. Lines could connect dots of the same colors, to further emphasize that they are associate with the same bloom event. Going one step further, Fig. 10 and 11 can be merged. And once colored, Fig. 11 appears to contain all the information contained in Fig. 10a, and 10a could be removed.

Now Figs 11 and 12, with Fig 12 merged to become part of Fig 11. Fig 11a very specifically refers to the duration each object existed for given the threshold to define it. Fig 11b now shows the areas (with a bug fix!).

Several smaller implementation aspects of the study remain vague or are not mentioned. I have pointed out several instances in my specific comments below, one example is the log-transformation of the chl-a field. The manuscript includes a somewhat lengthy description of the choice of thresholds and that they are equally spaced in log-space, just to translate them back into linear space and discard all but 3 of them for this study. Meanwhile, an important implementation aspect concerning the log-transformation -- if smoothing is performed in log-space (I presume so but I am not sure) -- is not mentioned at all.

Your queries about the thresholds has been addressed. Hopefully the discussion is now sufficiently clear. The data sets were not transformed but thresholds were chosen to be equally spaced in log-space, thus bisecting the data appropriately (given the underlying distribution).

specific comments

l 19: "By contrast the AMM7v11 analyses produces more bloom objects in deeper Atlantic waters, which are not detected by the satellite product.": Not if sure if this is intended but this formulation makes it sound like the model is correctly producing a chlorophyll bloom while the satellite (incorrectly) fails to detect it.

This does read ambiguously and has been reworded. In addition, the abstract has been shortened somewhat to conform to the 250 word limit.

l 72: "that is": It's not clear if "that" is referring to blooms or "oceanic primary production". I assume

it is the latter and would suggest to remove the "that is" and replace it with a ",".
This has been clarified, we have replaced with “,” as suggested.

l 77: While there is definitely a link between Chl-a and phytoplankton biomass, many of the biogeochemical models often have separate variables for both phytoplankton biomass and Chl-a. So using model Chl-a instead of model phytoplankton biomass to infer phytoplankton biomass is not a common use case. But why mention biomass here at all, I would suggest to link Chl-a directly to blooms.
We have rephrased this sentence to say:

“Biogeochemical models coupled to physical models of the ocean provide simulations for the various parameters that characterise the evolution of a spring bloom, such as Chl-a concentration which can also be estimated from spaceborne ocean colour sensors (Antoine et al., 1996).”

l 84: Maybe add Mattern et al (2010) to the list here, which compares several neighbourhood-based methods for model-satellite Chl-a comparison.
We have added a reference to Mattern et al. (2010) as suggested.

l 117: Maybe add that the model grid is used for this purpose here.

This has been added.

l 125: "chlorophyll": Shouldn't this also be Chl-a?

This has been changed.

l 158: I would suggest to add one sentence about the difference between this product and the L4 product used in the comparison. Even just to mention that this is not the same dataset later used in the comparison.
We have added the following sentence:

“The L3 product is based on two of the same three ocean colours sensors used in the L4 product described in Section 2.1, but with different processing and no gap-filling.”

Fig. 1: While "L4 observations" is no longer used in the text, it is still present in the figure and its caption. I am not against using the term but for consistency, the authors may want to change it.
This was changed in figure and caption

Fig. 1: Negative exponents on the color bar are not fully in the superscript.
Color bar corrected in a revised figure.

l 212: Use "Fig. 2a" instead of "a", I was at first confused what it was referring to.

This has been changed. Fig 2 has been added to the (b) as well for clarity.

l 216: "Only the pairing with the highest score is analysed further": In this particular example, or in general? What would happen if the 3-2 pairing scored 0.75 in Fig. 2b? I assume it would also become a match but the text suggests it would not.

No, only the pairing with highest score is analysed as a matched pair, though the results are stored for all the matches that are performed. Each possible combination of objects is scored. Whether a particular match is then retained depends on whether the score is above the threshold (0.7).

Fig. 2: Nice example but it would be useful to more explicitly mention that 1 and 1 are matched. Currently only the bold text suggests it.

Good point. We have added this to the text.

Fig. 2: The 1km distance bar should be moved closer to the objects in 2a.

Done.

l 236: In the answer to my previous comment, the authors mention that the observations are smoothed as well. This should be made explicit here.

This has been clarified.

l 278: When mentioning the threshold of 1000, it would also be good to mention the time resolution that is used here. Is a distance of one unit in time treated equally to one unit in space?

In this instance the volume is the sum of the grid squares identified in each time slice of the identified space-time object. This has been added to the text.

l 290: Maybe write "to _only_ capture events of interest" ("only" added by me) to emphasize why that would mean increasing the threshold.

“only” has been added.

l 293: "though higher values are present": In the data, in the L4 product, or both?

This is somewhat ambiguous and has been clarified.

l 296: "In the paper": Here? I would suggest to rephrase to "here", "in this study" or "in this paper".

The phrase has been replace with “Here….”

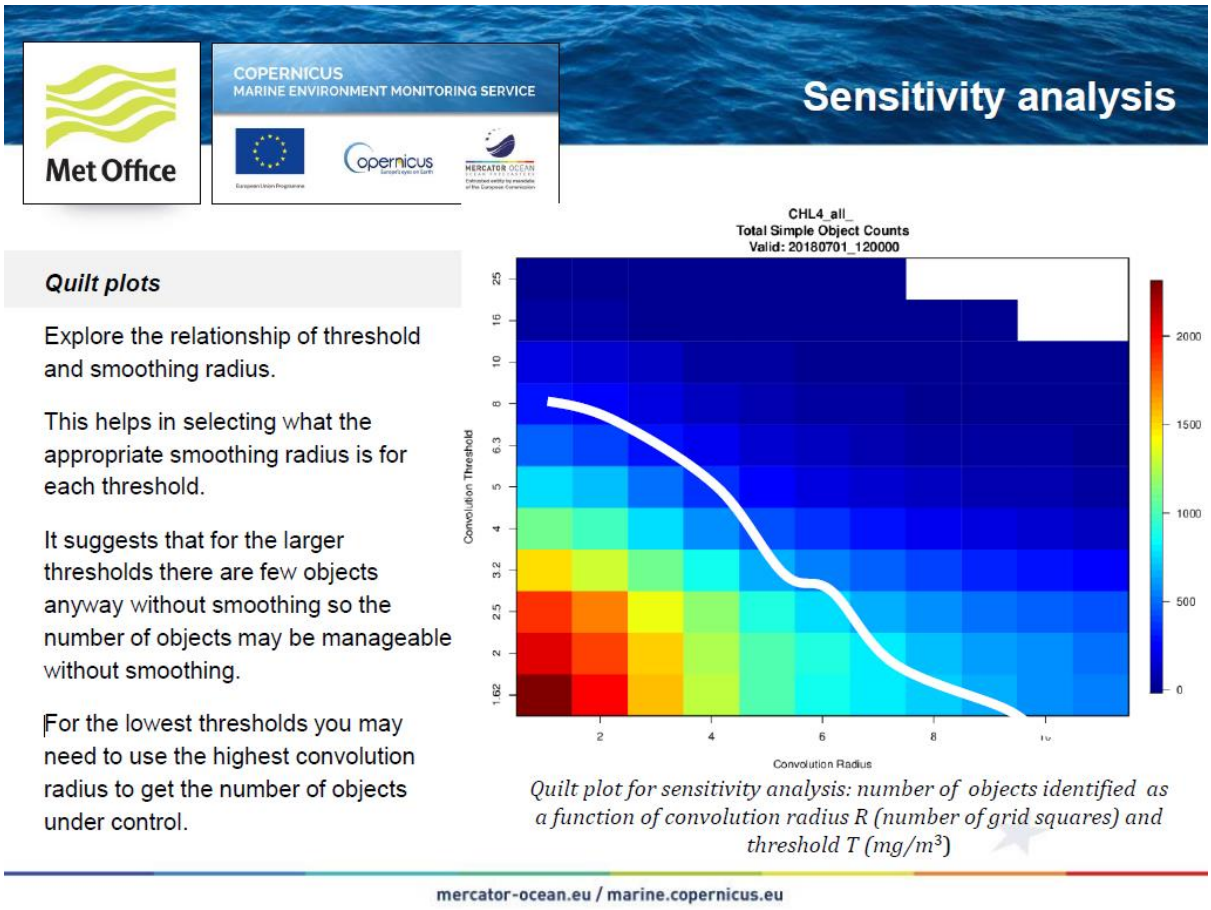
l 304: Why is a count of 30 an important limit?

These methods are intended to be inherently heuristic in that they attempt to mimic what a human would do. This number was arrived at through some pragmatic eyeballing. Beyond that number my brain could not have managed to make much sense of what to do with the identified objects!

l 304 "Furthermore, the smoothing applied needs to be reduced with increasing concentration thresholds because objects become smaller and are less frequent.": Here the logic appears to be

backwards: smoothing happens before applying the threshold. So, increased smoothing decreases the concentration, requiring a lower threshold.

Indeed, you want to apply less smoothing for the higher thresholds to avoid destroying the peak values before identifying objects (noting that the raw field information is put back into the objects after object identification). A figure similar to the one below (I think) was originally included in the manuscript, which showed the inverse relationship between the number of objects as a function of smoothing radius and concentration. See the figure below… these were results for the 2018 season sensitivity analysis but highlight the principle. It also makes it very clear that setting the threshold too low isn't a very good use of this methodology. You want to identify distinct features or objects.



Quilt plot for sensitivity analysis: number of objects identified as a function of convolution radius R (number of grid squares) and threshold T (mg/m^3)

l 308: Remove "too".

Done.

l 308: "objects may be spurious": It is not really clear what this mean in this context. Wouldn't increasing the threshold remove objects that are less relevant?

Not using a smoothing radius along with an inappropriate area threshold for object identification could result in objects being identified which are actually not meant to be objects. This is especially true for gridded observation fields, such as rainfall estimates from radar where the presence of ground clutter can still remain, even with reasonably good QC. You do not want to include objects like this in the analysis. The wording has been strengthened in the text.

l 309: "AMM7v11 analyses are on a ~7 km grid." This has been mentioned before.

This is a duplicate and has been removed.

Fig. 3, 4, 6, 7, 11: It would be useful to show grid lines in the panels

In many of these plots there are lot of lines. What is now Fig 12 has been removed and integrated with Fig 11. We have opted to include some (feint) reference lines where necessary, e.g. Fig 5.

l 332: "centile": This is the short form and less commonly used word for "percentile". Especially since percentages are never used here, I would recommend to use the more general and likely better known term "quantile" instead.

Quantile has been applied throughout.

l 332: "less than equal": Shouldn't this be "equal"?

We believe this is correct.

l 333: "corresponding AMM7v11 centile": This needs to be described a bit better, is the "corresponding" centile/quantile the quantile of the same value or the quantile that is associated with the same chl-a concentration?

The explanation has been further strengthened with specific reference to Fig 4.

l 350: Maybe a better example (rather than a constant value) would be the presence of high chl-a features in the right place, i.e. "A lack of objects suggests the presence of a model bias but it does not provide any sense of whether the model is producing enhanced Chl-a (albeit below the chosen threshold) at the right location, or not."

The fundamental premise of this method is to take away the need for the forecast (or analysis) to produce the feature in the right place at the right time, as this is a specific issue with high-resolution models. The method of equalising quantiles (or quantile mapping) works on the premise that one distribution is fixed and the other is not, and in doing so the bias is removed. This is why we had included the work on quantile mapping in earlier revisions, but this seemed to confuse rather than help in explaining the method and so the fixed threshold results have been shown here instead. Indeed, the presence of a bias, could lead to the scenario you describe. Hence, why understanding what impact the bias *might* have, is really fundamental to any threshold-based methodology. Here, we are relatively safe as, though there is a bias, it isn't prohibitive in the way you describe.

l 360: "In this instance, though there is bias, it did not prevent the identification of objects in either fields to the extent where the results did not reflect the potential for the analyses to provide features which could be matched, paired and compared.": This sentence needs to be rephrased and simplified.

We have attempted to do so.

l 415: "From an interpretation perspective": It's not clear what this means, I would recommend not using ambiguous phrases like this.

We have tried to clarify the text.

Fig. 7: Because the description explicitly mentions the total number of objects and because the lines often overlap, it may make for a better plot to show the total number of objects and the number of matched objects, instead of matched and unmatched.

The figure has been replaced to reflect the total vs matched number of objects and the text amended.

l 441: Where are the "South West Approaches"? I'd suggest to either describe the location in different terms or mark it on the map.

This has been rectified with the inclusion of new Fig. 2, and is referred to in the text.

l 441: "There are areas, for example in the South West Approaches, where there appears to be a good level of consistency.": But this is for identified and not matched objects, correct? How about similar maps that show the probability of the grid cell being part of a match given that is is part of an identified object?

Indeed, this would be very interesting to do, and is a great idea. The gridded output is not currently sophisticated enough to allow that to be done with any ease. Also, for this dataset the probabilities would be very low. The maps here provide more of a climatological view of the season.

Fig. 10: Does the same color for AMM7v11 and L4 indicate that the objects matched?

This is now Fig 11. No, the colours here denote the progression of time… approximately the same hues give a sense of the temporal proximity of the objects, but as (c) shows even if objects exist at the same time, they may not be geographically close.

l 566: Are these 26 days solely based on the first bloom object (l 501)?

This is the difference in dates between the first identified bloom objects in both datasets. This has been clarified more in the text. Now near line 543.

l 587: "MODE results suggest that the AMM7v11 bloom objects are larger than those in the L4 product.": That appears a bit surprising given the results in Fig. 5 showing large L4 objects, the higher chl-a concentration in L4 (see distribution), and the result that AMM7v11 objects are numerous but small.

This is now Fig 6, which is only one daily snapshot. The area panel in Fig 11(b) shows the area differences more clearly.

Report #2

My main concern still remains: to illustrate advantages of the presented verification methods (newly applied for Chl-a forecast evaluation), it would be nice to compare the proposed method with other metrics like, for instance, bias (classical), MAE(MAD), bloom phenological indices (Siegel et al. 2002,

Soppa et al.2016), or any methods previously used for AMM NWS Chl-a valuation with respect to complication/simplicity, possible diagnostic (meaning) and conclusions drawn by the analysis given the particular (addressed in the paper) application case/data. The study would benefit from such a comparison that should not not take lots of time.

While the focus of the paper is on introducing and exploring new methods for evaluating Chl-a, we agree that some comparison to classical metrics may be helpful for the reader. We have therefore included median bias and median absolute difference as per your suggestion, and additionally Pearson correlation coefficient. These are now shown in Tables 2 and 3 and Fig. 2, with accompanying discussion in Section 4.1. These metrics are the ones used in the CMEMS Quality Information Document for the model product (McEwan et al., 2021, https://catalogue.marine.copernicus.eu/documents/QUID/CMEMS-NWS-QUID-004-002.pdf), and so represent the routine metrics which the novel method introduced in this study would aim to complement.

We have not included bloom phenological indices as these are not commonly used for model validation. Siegel et al. (2002) and Soppa et al. (2016) are observation-based process studies for instance rather than model validation papers, and we are only aware of phenological indices being used to validate 1D models at in situ time series stations (e.g. Anugerahanti et al., 2018, https://doi.org/10.5194/bg-15-6685-2018), rather than 3D forecast models. Appropriate application of such indices to validating daily model outputs would require optimising e.g. smoothing and time-averaging procedures, which is outside the scope of the present study.

We have added the following text to the discussion section:

"Using MODE and MTD clearly gives extra information not obtained from traditional verification metrics that are more routinely used (McEwan et al., 2021). An alternative approach to assessing the representation of phytoplankton blooms might be to use phenological indices (Siegel et al., 2002; Soppa et al., 2016), which measure the day of the year on which Chl-*a* concentration first crosses a threshold based on the median concentration. Phenological indices have been used in observation process studies (Racault et al., 2012), but very rarely for model verification, and then only in 1D (Anugerahanti et al., 2018). One reason for this is that daily model Chl-*a* will frequently cross such a threshold throughout the bloom season, meaning temporal smoothing and other processing (Cole et al., 2012) would be required, which is not straightforward to apply consistently. Objective methods such as MODE and MTD, which consider individual bloom objects throughout the season, rather than assuming a single spring bloom will occur at each location, bypass these difficulties."