OS-2020-100: Response to reviewer comments

We would like to thank the two reviewers for their comprehensive and thoughtful comments.

In terms of the major comments we have taken what RC1 wrote about Section 3 on board and have restructured and in places reworded the section with a new figure to help give a better sense of the method and how it works.

In terms of the analysis in Section 4, we agree that the AMM7v8 *model* analysis and forecasts are not doing a very good job at present of forecasting chlorophyll concentrations, though the spatial method does extract more useful information in terms of where the major issues of poor performance lie, e.g. the timing error in the onset of the bloom. The AMM7v11 *model* analysis was not originally part of the study, and there were no AMM7v11 forecasts produced at the time, so these cannot be included.

Upon careful consideration, given the poor performance of AMM7v8 and the confusion of comparing three datasets, we agree with RC1, and have taken the recommendation of removing all aspects of the AMM7v8 analysis and forecasts (i.e. the three-way comparison) from the paper. This has meant a change of title and fairly substantive rewrite of Section 4, with many changes also needed in other sections to match. There have been other benefits. The original paper was excessively long. The removal of AMM7v8 has allowed for a shortening of the text and rationalising of the figures, providing a paper of more conventional length. This makes the paper much less confusing with a clearer message. It also means that many of the minor comments made by RC1 and RC2 are no longer relevant. Any comments related to deleted text are shown below with a ~~strikethrough~~.

We appreciate that the word “analysis” is often used to describe anything that is on a grid; therefore, the L4 ocean colour product is also an “analysis” in that sense. The text has been made a lot more generic, removing all mention of forecasts. The L4 product is no longer referred to as an analysis. This term is reserved for referring to AMM7v11 only.

RC2 requested some further background to the new method:

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 (mentioned in lines 83 - 85) with respect to complication/simplicity, possible diagnostic (meaning) and conclusions drawn by the analysis.

Such metrics have been used elsewhere in the literature to validate the AMM NWS Chl-a, and we have added further AMM NWS references on evaluation work to the introduction, and a summary of relevant findings, which serve to put the findings of this study in context as suggested.

RC1: minor comments

l 21: "whilst several forecast blooms did not materialise in the observations": This sounds like the data is to blame for not showing the bloom, maybe rephrase to something like: "while the model forecasts also showed blooms that do not appear in the observations".
The abstract has required a substantive rewrite given the change in focus of the paper

l 22: "Whilst the model...": Most of this has been said in the previous sentence, would

be more useful to move it one sentence up as an intro.
The abstract has required a substantive rewrite given the change in emphasis of the paper.
l 34: "double penalty effect": For readers not familiar with this term, it would be beneficial to describe this term a little better. Currently it sounds like not in the right place would be a single penalty and additionally not at the right time would be the double penalty.

The explanation has been strengthened to help the reader unfamiliar with the concept understand it better.

l 114: "the models at 7 km resolution cannot resolve the coasts": It is not entirely clear what exactly is meant here: are you referring to the coastal chlorophyll a dynamics?
It is a combination of the coastal dynamics, and the grid resolution which makes many of the processes sub-grid scale. This has been clarified in the text.

~~l 126: "is" -> "are"~~
Reworded.

l 133: It would be good to introduce the "Atlantic Margin Model" the first time "AMM" is used in l 117 or l 98.
Done. It has been added around line 87.

~~l 135: I have a little trouble understanding this: is "Day 4 for the period of 1 March-31 July 2019" simply March 4th or are multiple 4-day forecasts created?~~

l 136: Mention that both the analysis and the forecast are used in this study. The previous sentence is confusing to the reader, it appears to state that the forecast is hereafter referred to as AMM7v8, when later on AMM7v8 forecast and analysis are used.
All reference to AMM7v8 has been removed from the paper.

Fig 1: I would prefer to have the log scale included in the colorbar ticks ("10ˆ{-3}" instead of "-3") rather than written out in the title. This would also eliminate potential confusion, as the values in (d) are not exponents, although the title may indicate that.
This has been changed and the colour bar of the Fig 1 top subpanels are now log scale, with actual values in the tick labels.

l 150: Use (a) and (b) instead of "left" and "right". Same for the next sentence.
This has been amended.

l 158: "can be comparable": That is a very imprecise statement, as a reader I am not sure what this is telling me. Are the observation errors of the same magnitude as the model bias?
This is not the case anymore when considering only AMM7v11.

l 166: The instrument to measure ocean colour is satellite-borne, the ocean colour is not. Maybe use "remotely-sensed" or "satellite-derived"? Also, I would not refer to colours as concentrations.
The vocabulary related to the satellite-derived chlorophyll concentration has been changed and clarified.

l 167: Out of curiosity, why is a different satellite product used for the comparison and is there a significant difference between the two satellite products?

The product used for the assimilation is the L3 NRT product, whereas the product used for the comparison is the gridded (L4) REP product (OCEANCOLOUR_ATL_CHL_L4_REP_OBSERVATIONS_009_091), it is available 6 months after NRT and benefits from the maximum level of post processing.

l 172: Is this referring to the coupling between physical and biological model components?

Yes. This has been clarified in the text.

l 173: Is the AMM7v11 data assimilation also based on 3Dvar?
Yes. This has been clarified in the text.

l 181: The "sequence of forecasts" is a bit confusing here. The next sentence refers to "a forecast". I know that data assimilation can create a sequence of forecasts but here I think it would be much easier for the reader to stick with "a forecast" throughout the description of MODE.
This section has been rewritten. We have removed the use of the word forecast and observed. The use of the word “sequence” here is justified because MTD cannot be used without having temporal sequence of some description.

l 184: The mention of a model-based analysis is not helpful to the reader here, and it also not used again in this paragraph. I would suggest to rephrase to something like: "in this context, one is typically a model forecast, the other an observed field, i.e. observations regridded to match the locations of the model grid." Later on it can be mentioned that it can also be used to compare two model fields.
This section has been rewritten as suggested.

l 189: "observed objects" -> "observed field"
We have removed all references to “observed” and “forecast” from the revised paper, so this sentence has been reworded in a more generic way.

l 198: "and is based on a disk": Is the convolution kernel really a (flat) disk? The "based on" is confusing here.
This section has been made simpler and clearer, omitting words and phrases which could confuse.

l 201: Maybe use "objects" again, as above, instead of "areas".
This has been done. See around line 222.

l 204: Are the observed fields not smoothed? In step 2 it sounds like both fields are smoothed.
This section has been rewritten substantially. Yes, both fields were smoothed.

l 212: This is unclear: the first sentence in (6) "which together are expressed as the socalled “interest” score" makes it seem like one interest score is computed summarizing the fit of all objects. In this sentence, "interest scores are computed for all" objects. Please rephrase.
This section has been significantly rewritten. Hopefully this aspect has been made clearer. See lines around line 228.

l 252: "minimum volume of 1000 grid squares": This should be "area" or is this applied

below the ocean surface? Why is it 1000 grid squares here and 10 above? Reading it a few more times, it seems like this describes a "volume" in space-time, which needs to be made more explicit. "1000 grid squares" is not a volume, is this 10 grid squares times 100 time steps?
The thresholds are independent of each other in MODE and MTD. Here the 1000 grid squares does refer to the accumulated number of grid squares over consecutive time slices. The 10 grid squares refers to the MODE area threshold. This has been clarified in the text, in and around lines 263.

l 255: I am guessing here that paired objects are those for which an equivalent was found in the other field while all others are called single. And clusters are two or more objects in one field, classified as belonging together. Please add a bit of description here, to define these 4 terms.
This forms part of the rewrite of Section 3.1, which also now includes a schematic in a new Fig 2 to provide visual aids to what these terms mean.

l 262: Using the log-transformation seems like a sensible choice for chlorophyll. It would be good to know if the transformation was applied before smoothing the fields.
The fields were not transformed but the thresholds were derived from an equally spaced progression in logarithmic space to ensure that the thresholds we applied followed the underlying distribution of the Chl-*a* values. Apologies if this seemed confusing.

l 266: "mg." -> "mg"
We believe that you would like us to remove the full stops in the units. Which we have done.

l 267: The units here would be "log_10(mg mˆ-3)". In my opinion, it would be clearest to just provide the values in linear space, e.g.: "For this study, a range of thresholds were applied to the log10-transformed chl-a fields, corresponding to chl-a concentrations between 1.62 and 25 mg mˆ-3."
We have reworded the section to make it clearer what was done.

l 268: Why use 25 mg mˆ-3 here when the values of interest are in the range 3 - 5 (previous sentence)?
We were guided by the values that have been typically used in other studies. Over the NWS the values do appear to be mostly in the range 3-5 mg.m^{-3} but larger values are present, though in too small numbers. This has been reflected in the text on line 276.

l 274: "This radius" -> "The 5 grid point radius"
This paragraph has been rewritten to reflect changes to the focus and emphasis of the paper to only include AMM7v11.

l 276: Why so much discussion of different thresholds above when only a single one will be used here? Or is "here" only referring to the sensitivity analysis? Please be more specific.
A series of thresholds were considered as per line 267 in the original manuscript. We would never have been able to discuss all of them, and in fact on the lowest thresholds provide sufficient sample size for analysis. Line 276 was there to say that most of the results presented in the paper would relate to the 2.5 mg.m^{-3} threshold, which was the second lowest we considered. This has been amended slightly near line 276 in the revised paper.

~~l 304: What is "this", maybe use "The effect of bias"~~

l 310: "would yield no useful information": One could argue that having no matched pairs contains the useful information that the model solution cannot be very good.
True. It provides a very stark outcome but in terms of trying to demonstrate what a feature-based method can provide, would not have been very helpful. It also doesn't totally capture the ability of the model to capture enhanced levels of Chl-a, which could still provide some useful information, with the caveat of knowing about the bias. We have added a sentence to reflect your comment, which is a valid one. Text has been added around line 329 in the revised paper.

l 315: Why is the L4 product referred to as an analysis here?
It is an analysis of sorts but the use of the term here, we agree, is sloppy. We have ensured that the text only refers to the AMM7v11 as an analysis and L4 is called a product.

l 324: "the two that clearly differ more dramatically from Fig 2": What is meant here is probably "the two fields that show the largest discrepancies in Fig 2".
Correct. This has been reworded, though the revised Fig 5 no longer shows quite the same discrepancies as AMM7v8 has been removed (panel (a)).

l 336: What about using different thresholds for the different fields?

This is essentially what quantile mapping does or the use of a fixed, but different, threshold derived from the dataset… which is what was referred to as the "seasonal" threshold. This is no longer relevant in the context of the revised paper.

~~339: "forecasts": Which forecast, only analysis solutions were considered thus far.~~

Fig 3: Please include the AMM7v8 distribution after the quantile mapping. And why not include the AMM7v11 distribution as well?
As the paper has been reduced in scope to only include the AMM7v11 analysis this figure now reflects only the L4 and AMM7v11 distributions. For completeness, it could have been interesting to plot the AMM7v8 quantile mapped distribution, the method did not use the season long distribution, but the daily mapped distributions.

l 356: "the observed threshold": Do you mean the threshold applied to the observations?
Yes, that is what was intended. We have refrained from using the word "observed" in the revised paper, choosing to refer to L4 as a "product" instead.

l 358: After a lot of reading the next paragraphs over and over to figure out what the thresholds are, this is the position where I lost track. Here, it needs to be stated that the "value that has the equivalent rank in the forecast distribution" is the "forecast threshold". They are currently not linked and initially I thought that the forecast threshold was the same as the (slightly misnamed) observed threshold.
This section has changed substantially to reflect the removal of AMMv8 forecasts from the paper, along with the use of variable thresholds. In doing so, the paper has become a lot more focused with less opportunity for confusion.

l 362: Based on this description it is not clear what this threshold is. It is also called "seasonal" here when Fig. 4 shows daily variations in a threshold. Are these the same thresholds, why is it called seasonal? If this is the threshold used to identify objects, please mention this explicitly.
We agree that this section was very confusing. With the decision to remove AMM7v8 this section is no longer relevant and has been removed. We have kept a revised version of Fig 4 to show the

distributions vary on a day-to-day basis and how this affects the AMM7v11 concentration value that is in the same centile of the distribution as the 2.5 mg,m^{-3} L4 distribution.

l 365: Is this procedure applied to both v8 and v11?
No, it wasn't. But this is no longer relevant with the focus on v11.

l 374: "the latter": is this the L4 product? "the latter" is not really referencing anything at this point, please just use the product name.
This whole section has been reworded to reflect the change in storyline, hopefully making it clearer too.

l 375: Please explain better what this threshold means. It is impossible to understand this paragraph without knowing what the threshold signifies.
The discussion here has been reworded in light of the fact that the concept of quantile mapping has been removed. We felt it was useful to keep the concept of distributions on the daily time scale to provide as rounded a picture of the variations of the bias on a day-to-day basis, to explain the impact of using a fixed threshold for both datasets.

Fig. 4: It would be good to include a grid and reduce white space, the threshold never exceeds 6. I would further suggest to merge this figure with Fig 5.
Fig 4 has been recreated and Fig 5 has been removed as AMM7v8 is no longer discussed anywhere. Instead of a grid we have added a horizontal dotted line at 2.5 mg.m^{-3} as a visual guide. We hope this is satisfactory.

l 390: "Error! Reference source not found." Something appeared to have gone wrong with the automated submission system.
Indeed! Not sure what happened. Sometimes the pdf conversion can also do this. Sorry we didn't spot this.

l 391: "using the built-in functionality in MODE as for Fig 4.": This is unclear, please rephrase.

As the quantile mapping has been removed from the paper, we have refrained from referring to the method in those terms. We have retained the figure in the text but referred to it in statistical terms rather than using any reference to MODE capability.

~~l 394: What is the forecast lead time?~~

~~l 405: "the forecasts are very active": What exactly does this mean, phytoplankton blooms?~~

l 406: "The latter object is not identified in the L4 ocean colour product." It is not clear what "object" this is referring to. I do not think this paragraph is very helpful to the reader at this point. So far the reader has only seen an example of what object identification should not look like (Fig. 2) and now here is a lengthy explanation of chlorophyll-a blooms leading to peaks in a threshold without showing an example of that these variable thresholds improve object identification.
This paragraph has been removed from the revised paper, and the preceding explanation of the methodology rewritten to be clearer

l 434: "this work was done without accounting for the concentration differences": Does this imply that the variable threshold from the previous section was not used? But it

seemed to be very import for successful identification.
Correct. We agree that, in hindsight, some of the elements in the submitted paper can appear contradictory. In the rewrite these sections, which refer to AMM7v8 have been removed as this part of the analysis was not done for AMM7v11, as it wasn’t available at the time. We did use the smoothing radius findings for AMM7v11.

~~l 444: Are "quilt plots" the same as "quilt “difference” plots"? I would suggest to stick with one term.~~

~~l 444: Here are two nearly identical sentences two (short) paragraphs apart: "Figure 6 provides a selection of quilt plots derived from using the L4 ocean colour products and AMM7v8 analyses during July 2018, using one of the merging options which was tested." (l 435) and "In Figure 6 some quilt “difference” plots are shown to focus on the individual characteristics of the AMM7v8 analysis and the L4 ocean colour product based on a set of initial data that was available for July 2018."~~

l 451: How useful is this analysis if it is already clear from the previous section that a comparison at the same threshold is not sensible? The 2.5 mg mˆ-3 threshold can be established without considering the model output.
True. In the rewrite of the paper this section has been removed as we are no longer considering any aspects of AMM7v8.

l 475: Again, how valid are these results if different thresholds would be used for model and data? Would the conclusions about the smoothing radius hold?
This section has been removed as this analysis was not done using the AMM7v11 results, as they weren’t available at the time.

~~Fig 6: The font is much too small.~~

This figure has been removed as it relates to AMM7v8.

l 489: Is this analysis done at the same threshold for data and model?
Yes, it is.

l 491: Is this "total area" of identified objects or ocean grid cells?
This is the total of the combined area of identified objects and has been clarified in the text.

~~l 534: The lead time has still not been properly explained.~~

l 545: So these percentiles are characterizing the chl-a distributions among different areas. How can they contain values below the 2.5 threshold applied to the observations? The "(in this case 2.5 mg mˆ-3)" makes it sound like the same threshold was applied to the two products but this seems to contradict earlier sentences. Assuming now that the threshold are different: From Fig. 3, we already know the distribution of chl values, what new information does Fig 9 give the reader? We know of the bias and have a rough idea of the distribution, and we also know that a higher threshold will likely be used for AMM7v8, which appears to be the main contributor to the differences between the distributions in Fig 9.

This figure has been removed as it relates to AMM7v8.

l 574: It would be nice for comparison to see good results for comparison. How would

Fig 10 look for a AMM7v11 to L4 comparison?

Fig 10 does show the AMM7v11 and L4 results, but not against each other. In the revised figure the AMM7v11 vs L4 comparison is the only comparison that is shown, which hopefully makes things clearer. Yes, it would be nice to show good results and these certainly are much better than for AMM7v8!

l 623: "AMM7v8 day 0 forecast to L4 and AMM7v11 (labeled AMM7v8 vs AMM7v11, and AMM7v8 vs L4)": The labels are either switched or cleverly selected to confuse the reader.

The three-way comparison has been removed so this figure has also been removed.

l 653: Is the spatial centroid the average location of the center of a series of identified objects? This needs to be explained.

Yes, it is. This has been clarified in the text.

RC2 specific comments:

Lines 15 – 16: I like that the discussed verification method allows not only identify bias but correct (mitigate its impact when carried other further analysis). How would it compare with regular bias correction?

The aspects of quantile mapping have been removed from the paper since it was not deemed necessary to evaluate AMM7v11 alone as the bias isn't as severe. How would quantile mapping compare to other bias correction methods? It depends on the data sets. Depends also on what you mean by regular. I bulk bias correction would affect all parts of the distribution in the same way and could therefore do some strange things in the tails of the distribution. Other forms of calibration-based method such as Kalman filtering would probably provide similar results but have other advantages in that they can extrapolate to values outside the distribution, if that is an issue.

Lines 526 – 530: Reads as this paragraph should not be at this place. Which figure is discussed? Figure 12?

This paragraph has been moved as it did appear to be in the wrong place. It has been to the discussion of the revised figure, though still Fig 9. The original Fig 9 has been removed entirely.

Lines 553 – 557: How would it compare with analysis of the frequency distribution (histogram)? Part 4.5 “Onset and evolution”, I would suggest change the title to “Bloom onset and evolution”. I am also curious how would the material and results presented in this subsection compare with simple bloom phenology analysis based a threshold method (Siegel et al. 2002, Racualt et al. 2012, Brody et al. 2013, Soppa et al. 2016).

Following the substantial rewrite this material has now been included in a new Section 4.4 titled “Incorporating the time dimension”. Bloom phenology analysis based on threshold methods is not yet routinely used as a model validation metric, as applying a method consistently to both satellite and model data is, in practice, not straightforward. However, we have now referenced other studies which have assessed bloom timing in the NWS system (e.g. Skákala et al., 2020, https://doi.org/10.1029/2020JC016122), to provide some context for our findings.

~~Lines 772 – 774: "Blooms were said to occur when the observed concentration threshold exceeded 2.5 mg.m-3. Forecast thresholds for MODE were then relative to this value and varied from day-to-day. For MTD the seasonal equivalent threshold for the AMM7v8 forecasts was 6 mg.m-3". What can be in general concluded from this about AMM7v8 performance?~~

Typos

~~Line 99: delete on of two words "product".~~
~~Line 187: "*merged*", "*matched*" – is italic font urgently required. Or it is a format error?~~
~~Line 197: "*further*" - a format error?~~
~~Line 206: "*not*" - a format error?~~
~~Line 257: "*single simple*" - a format error?~~
~~Lines 217, 218, 220, 232, 233, 237, 238, 242: please check/confirm if italic font used for a number of words is required (?since refer to specific options? right?).~~
Italics was used for emphasis or to highlight specific key words relevant to MODE or MTD. This whole section has been rewritten and these are no longer relevant.

~~Lines 546 – 548: again, a format error – used square brackets, "*exceeding the threshold*", "*within-object*"~~
~~Lines 550 – 552: "*within-object*" - a format error?~~
This text was a description for Fig 9 which has been removed.

~~Line 641: "*and*" - a format error?~~

Line 652: "*time*" - a format error?
Italics removed.

Figure quality

Figure 2: please improve quality of the subplots titles.
The figure has been recreated with fewer panels to remove AMM7v8. This has improved the overall sizing etc. This is now Fig 5.

~~F~~igures 3, 6 and 7: please enlarge the font used in the subplots
Figure 3 has been recreated with larger axis labels and using the AMM7v11 dataset. Figs 6 and 7 have been removed with the rewrite.

~~Figure 15: please enlarge the font used in the subplots and introduce units for Chl-a;~~
~~Figure 17: Upper and low left panels, xlabel(subtitle) brackets are missing at the end.~~