General Comments

This article provides an interesting insight into the regional dynamics of phytoplankton Chlorophyll at a regional scale and makes good use of a multi-sensor time series to provide new information on phenology. The justification of the dataset and algorithm used appears valid and the data processing is clearly explained. The linking of the chlorophyll dynamics to sea-surface temperature would benefit from a cross correlation analysis between the two time series to allow a numerical support to the statements made. A couple of minor corrections are required to the text (mostly for ease of reading) and a couple of the plots (such as missing legends or the addition of regression lines).

We appreciate that the reviewer liked the article and thank her/him for the suggestions that certainly improved the quality of the paper. The current version of the manuscript takes account of all the comments from the Reviewer. Moreover, the comments from both Reviewers stimulated a thorough and careful re-reading of the entire manuscript. We have more strongly stressed the focus of the paper, which is worth reminding develops in the operational context of CMEMS. Main limitations of this work are now mentioned and discussed in the manuscript.

The entire manuscript went through a general re-editing:

- *The abstract and Introduction have been nearly totally rewritten.*
- *Section Data (now Data and methods) now includes a new section with the description of the statistical analysis (formerly present at the beginning of the results).*
- *Similarly, the structure of section Results (now Results and discussion) has been improved by including a forth subsection (Algorithm regional calibration) to better ease the reading.*
- *Conclusions should now read less cumbersomely as more general statements did replace lots of the many summarizing and purely technical and repetitive sentences.*

All figures have been edited.

- *Figure 1 (formerly showing the matchup spatial distribution only) now includes the space-temporal matchup distribution along with the in situ CHL frequency distribution.*
- *Figures 2 and 3 about the scatterplot of satellite versus in-situ CHL now include a qualitative colour legend and regression lines for each plot.*
- *Figure 4 now includes a panel to show more clearly the variability and robustness of the regression lines obtained within the bootstrapping exercise.*
- *In Figure 5 the zero is now part of one bin of the histogram, while formerly was the boundary between two bins.*
- *Former figure 6 on the full time series of CHL has been removed because the discussion was and still is only focused on the climatological values (current figure 6) and the supplementary material already contained detailed figures of each CHL yearly time series. On the other hand former figure 8 on the full time series of SST have been moved to the supplementary material. Both figures 6 and 8 were not adding any specific information while potentially distracting the reader.*
- *A brand new figure 7 has been added to show that the adopted algorithms can also be used to monitor the space-time variability of the cyanobacteria in the Central Baltic, consistent with previous findings (Kahru and Elmgren, 2014).*
- *In figure 8 only the years with corresponding SST data are shown.*

The specific issues raised by the Reviewer are addressed below (in italic).

Specific Comments

page 2289: line 15: The authors state that there are “many more” invalid data in the MLP product. It might be of use to give a mean % valid pixels difference for s sample of representative images so that readers can quantitatively consider the difference.

On average, for all the time series, MLP had around 15 % more flagged pixels that OC4 and OC5. This has been specified in the manuscript.

pg2290 lines 5-20: There is no information provided on the methodology of the in situ chlorophyll a measurements. Were the data used from the database measured using HPLC, extracted fluorometric methods, both?

Relevant info has been added to the section, including the non-fully traceability of the methods used by all data providers to data repositories: data are included into the repository regardless whether measurement protocols are specified.

pg2291 line 20: The authors state that they discard outliers, where outliers are defined as possessing an error greater than a set criteria (ratio). I understand the need to remove outliers, especially in log normally distributed data, but are the points discarded evenly distributed across the range of the dataset? The distribution of the 'very poor' data is of interest in itself. Also, the authors state that 3.5% of the values are discarded using the ratio threshold. One assumes that the approximately normal distribution of satellite and in situ measurements also leads to a somewhat normal distribution of errors. Would it be possible to state that points where errors were outside the range 2nd-98th percentile of all errors were removed as outliers? If not then could the authors justify the error ratio chosen to dictate outliers?

As explicitly mentioned in the main text, "The discarded data were evenly distributed over the entire range of CHL variability and without any specific temporal or spatial patterns". Moreover, the criterion implemented in this work discards (3.5%) as many outliers as if the other criterion (2nd < X < 98th) was chosen (4% by definition). However, our method has the advantage of ensuring the analysis to be conducted over exactly the same datasets, thus allowing proper statistical inter-comparison among algorithms.

pg 2295 line 12: The authors state that the horizontal-averaged CHL for OC4v6corr were computed for images with a minimum of 1000 valid pixels. I assume that this is referring to the mean CHL across the entire basin? If so then it might be worth stating the total number of pixels that make up the region of study. This would then allow a statement along the lines of "Horizontal-averaged (whole Baltic) CHL for OC4v6corr were computed only for images with at least 1000 valid pixels (x% of basin observed)." Obviously observing the whole basin in a single day is near extremely unlikely but it would be good to know if 1000 pixels corresponds to 10%, 20%, 30% etc when considering the following phenological discussion.

Relevant info has now been added to the main text to incorporate the Reviewer comment. In this respect, the whole Baltic has 21424 pixels. Gulf of Bothnia has 5750, Skagerrak and Kattegat have 2625 and the Central Baltic has 13049. 1000 pixels correspond to 5 %, 17 %, 38 % and 7 % of their respective surfaces. It is a statistical property that the expected error of the horizontal average depends on the absolute number of pixels, not on the relative. For this reason, we decided to establish the minimum 1000 pixels equally for all occasions when horizontal averages are computed.

pg 2297: lines 2-8: The authors state that there is a relationship between SST anomaly and CHL anomaly time series. It would be worth performing a brief cross correlation analysis of the two time series such that the link could be quantified beyond "high-amplitude temperature anomalies induce similar growth and decay in CHL.". This may also provide information of the lag of a system response to forcing factors, which would be of interest to those cited in the discussion of lines 9-29.

The coefficient of determination R2 of such cross correlation has been added. Temporally, the range has been restricted between the days 150 and 250 (May 30th and September 7th) to focus onto the summer cyanobacterial bloom that is supposed to have a tighter link to temperature. As it is now clearly mentioned, the scope of this work is to assess the best performing algorithm among the many available for operational use. The lagged correlation analysis would surely be of interest to better highlight the link between SST and CHL, and in this respect we added some text and reference on both the complexity of the Baltic ecosystem dynamics and on the fact that "In the specific context of this cross-correlation analysis, we are implicitly *assuming that both SST and CHL respond to the calm weather conditions with the same time lag". Moreover, according to Kahru et al. (1993), cyanobacteria can heat the water column from their absorbed light, but of course cyanobacteria also respond to temperature variations, so the final observed cause-effect relationship might be not obvious.*

Technical Comments
pg2283 Title of article: The authors should hyphenate 'multi-sensor' as both terms apply to the following noun.

Corrected.

pg 2284 line 2: "Fifteen-year" should be "A fifteen-year"

Corrected.

line 8: Authors should expand the sentence "Statistics showed low linearity". Is this for all the provinces, individually and combined? Is this for all algorithms?

We changed "Statistics showed low linearity" with "In general, statistics showed low linearity" to better refer to the general outcome of the whole matchup analysis.

pg 2285 lines 2-6: A very long sentence. Perhaps best to break after Øresund? Change "Great Belt and Øresund, thus leaving the Skagerrak..." to "Great Belt and Øresund. This leaves the Skagerrak..."

Introduction significantly changed and this sentence has been removed.

line 18: It would be more clear if "Thus statistics should be consequently calculated." were changed to "Therefore the statistical assessment of algorithm performance should be performed on the area as a whole."

Introduction significantly changed and the concept expressed by this sentence has been rephrased and moved later into the introduction to better match the new version.

pg2286 line 4: "to cite a few" seems superfluous.

Removed.

line 27: perhaps include chl in the units for "1 mg m-3" otherwise the value could technically refer to concentrations of some other optical component.

Within the new version of the Introduction this bit was totally rephrased.

pg2287 line 4: "Case 2" to "Case II" for consistency with cited literature.

Corrected.

pg 2288 line 4: change "remains up to date" to "is currently"

Removed.

pg 2292 lines 3-4: change "that it can be derived" to "that can be derived"

Corrected.

line 9: change "BIAS respect to the..." to "BIAS with respect to the..."

Corrected.

line 27-28: It might be more clear to state that “In each region, OC4v6 overestimates CHL by > 40% on average.” as “In all cases, OC4v6 overestimates CHL more than 40 %.” could be interpreted to mean that for OC4v6 overestimates CHL by > 40% for all data points.

Corrected.

pg 2295 line 19: Change “reflect this phenomenon (Fig. 3a). Causes could be” to “reflect this phenomenon (Fig. 3a), possibly due to”

This entire sentence has been removed.

line 22: Change “are excessively risen” to “are excessively raised”

This entire sentence has been removed.

line 25: Possibly change “eventual coccolithophore” to something like “occasional coccolithophore” or “annual coccolithophore” to reflect the actual frequency of coccolithophore blooms in this region.

This sentence has been moved to the conclusions.

line 28: Might be clearer to change “However, few spikes in the time series...” to “Additionally, a few spikes inthe whole-Baltic time series...”.

This sentence has been removed.

pg 2298 line 5: Change “Fifteen years-long merged multi sensor daily CHL data” to “A fifteen-year merged-multi-sensor-daily dataset of CHL”

Corrected.

line 14: Change “the dynamics was similar as in the Central Baltic” to “the dynamics were similar to the Central Baltic”

This sentence has been rephrased.

lines 16-20: These would be supported by metrics from a cross correlation analysis.

Done.

line 24: It is not just a higher R2 that would be preferable but a smaller RMS and data collection over a full range of regional conditions to avoid the situation seen at high CHL in the Skagerrak and Kattegat region.

This sentence has been rephrased.

pg 2306 Figure 2: This figure has a colour scale for density of points but there is no legend to show what the colour scale range. Also, it may be of help to the reader to show the linear regression on the plots in addition to the 1:1 line (as is done in figure 3) so that one can get a better understanding of the regression parameters at a glance.

A qualitative colour legend indicating this low-to-high density scale has been added to both Fig. 2 and Fig. 3. Here, what really matters is the density of the matchup population within the plot, considering that a quantitative estimation of the point distribution would not add any particular crucial information.

Regression lines have been added in all scatterplots of the Matchups and Validation sections.

pg2307 Figure 3: The caption states that the line of equal value is dashed and black, but in the plot it is solid and black. Either the text or the plot require correction. This plot also requires the addition of a colour scale legend.

Done.

General comments

The manuscript provides the first long time remote sensing chlorophyll data for the Baltic Sea. There is a good reason why such analysis have not been undertaken before– simply the chlorophyll-a algorithms do not provide accurate enough concentrations to make such analysis reasonable. It has been shown by several authors (cited in the manuscript) that the blue-green ratios (like the OC4) do not work in the Baltic Sea. Mainly because the reflectance signal in the blue band is not determined by chlorophyll-a, but CDOM. The results of this study confirm these findings as the highest correlation coefficient found for all tested algorithms is R2=0.44. Therefore, the whole following analysis seems a bit artificial to me.

An important element that characterizes the new version of the manuscript is the clear focus on the operational CMEMS context in which the paper builds and develops. One of the main issues with previous CHL validation works over the Baltic (not many and all cited within the manuscript) is linked to the used datasets that were limited in both the geographic extent and in their seasonal representativeness, casting doubts on their overall relevance: most of these only use data from a few cruises. As it is clearly mentioned in the revised manuscript this constitutes one of the main motivations for this work, which uses the largest in situ dataset ever used for CHL algorithm calibration and validation over the Baltic region. Moreover, this is also the first work providing statistics for the entire Baltic, using the merged multi-sensor CCI and GlobColour products both covering a timeframe of fifteen years.

As for the blue-to-green ratio algorithms, it is true that this approach has demonstrated to perform poorly into the Baltic, though tested with limited datasets. It is also true that the ocean colour Baltic product currently provided in the context of CMEMS is far from being optimal (it is the worst of the proposed algorithms in all areas), as clearly shown by this analysis for the first time. All the Reviewer's concerns about the reliability of blue-to-green ratio algorithms (like the OC4) over the Baltic Sea are now included into the manuscript to increase the readers' awareness on this topic.

Moreover, the results presented here about the phytoplankton phenology in the Baltic are in line with what is expected from the literature with the Central Baltic blooming twice a year, in spring and summer (Reissmann et al., 2009), the Gulf of Bothnia only during spring (Carstensen et al., 2015) and the Skagerrak and Kattegat showing a minimum in Summer (Edelvang et al., 2005).

The current version of the manuscript takes account of all the comments from the Reviewer. Moreover, the comments from both Reviewers stimulated a thorough and careful re-reading of the entire manuscript. We have more strongly stressed the focus of the paper, which is worth reminding develops in the operational context of CMEMS. Main limitations of this work are now mentioned and discussed in the manuscript.

The entire manuscript went through a general re-editing:

- *The abstract and Introduction have been nearly totally rewritten.*
- *Section Data (now Data and methods) now includes a new section with the description of the statistical analysis (formerly present at the beginning of the results).*
- *Similarly, the structure of section Results (now Results and discussion) has been improved by including a forth subsection (Algorithm regional calibration) to better ease the reading.*
- *Conclusions should now read less cumbersomely as more general statements did replace lots of the many summarizing and purely technical and repetitive sentences.*

All figures have been edited.

- *Figure 1 (formerly showing the matchup spatial distribution only) now includes the space-temporal matchup distribution along with the in situ CHL frequency distribution.*

- *Figures 2 and 3 about the scatterplot of satellite versus in-situ CHL now include a qualitative colour legend and regression lines for each plot.*
- *Figure 4 now includes a panel to show more clearly the variability and robustness of the regression lines obtained within the bootstrapping exercise.*
- *In Figure 5 the zero is now part of one bin of the histogram, while formerly was the boundary between two bins.*
- *Former figure 6 on the full time series of CHL has been removed because the discussion was and still is only focused on the climatological values (current figure 6) and the supplementary material already contained detailed figures of each CHL yearly time series. On the other hand former figure 8 on the full time series of SST have been moved to the supplementary material. Both figures 6 and 8 were not adding any specific information while potentially distracting the reader.*
- *A brand new figure 7 has been added to show that the adopted algorithms can also be used to monitor the space-time variability of the cyanobacteria in the Central Baltic, consistent with previous findings (Kahru and Elmgren, 2014).*
- *In figure 8 only the years with corresponding SST data are shown.*

The specific issues raised by the Reviewer are addressed below (in italic).

Detailed comments

Being familiar with unpublished (yet) results from different countries around the Baltic Sea it is hard to say whether universal chlorophyll algorithm for the Baltic Sea is feasible. Studies on the specific inherent optical water properties show that optical properties of the spring bloom assemblages are very different compared to cyanobacteria in summer. Therefore, two sets of chlorophyll retrieval algorithms may be needed. These are results for the open parts of the Baltic Sea. Other studies show that absorption and backscattering coefficients (determining the reflectance) differ by order of magnitude between rocky granite shores (Sweden, Finland) and sandy shores (Russia, Estonia, Latvia, Lithuania, Lithuania, Poland, Denmark). Third study (also unpublished) shows that the correlation between the OC4v6 and chlorophyll is close to zero even if reflectances created by HydroLight model (i.e. free from atmospheric correction problems, glint, sensor noise, etc.) are used. So, it is really hard to believe that the algorithm used in this study will ever produce reasonable results for the Baltic Sea. In the conclusions the Authors state themselves (like tens of authors before) that green to red bands have to be used in order to get reasonable chlorophyll-a estimates.

It is surely true that one single algorithm for the entire domain (this in theory applies to the global ocean as well) might provide unrealistic observations given the phenological heterogeneity of the area. It is also true, however, that different algorithms meant to capture both the space and time variability (as the Reviewer is suggesting) would be extremely difficult to merge operationally. Here, as it should now be more clear, we aim at evaluating algorithms for the whole Baltic area, while still trying to provide region-specific metrics for the Skagerrak-Kattegat, the Central Baltic and for the gulf of Bothnia. Obviously, this fragmentation could be continued ad infinitum, but in the context of the operational oceanography, it is very unlikely that provided CHL products will reach such level of geographic and seasonal distinction in the near future.

On page 8 the Authors discuss problems related with vertical distribution of phytoplankton biomass. During most of the year this should not be an issue as top 10-20 m is mixed. However, vertical distribution becomes a huge issue during the period of cyanobacterial dominance. Unlike other phytoplankton cyanobacteria can regulate their buoyancy and (in calm weather) tend to be at the depth most optimal for them. It has been shown before (Kutser et al. 2008) that vertical distribution of cyanobacteria has significant impact on the reflectance i.e. the same biomass distributed differently in the water column produces very different reflectance.

The discussion on the phytoplankton vertical distribution is meant to provide details on the method used to convert a vertical profile into a "surface" observation. And the purpose of the method is exactly aimed to address the issue raised by the Reviewer. We now added a sentence to better stress the purpose of this particular bit of work.

Findings in the page 15 are contradictory to what was proposed by Kahru et al. (1993). Not the elevated temperature causes blooms (how it can be elevated?) but bloom absorbs solar radiation and heats the water.

We expanded a bit more on the motivations for this part of the analysis. The Reviewer's comment has been included into the main text and helped us expanding a little more on the motivations for this analysis. Moreover, as explicitly asked by the other Reviewer, we added a cross-correlation analysis between CHL and SST anomalies over the Central Baltic during summer to better support what was in the previous version a more qualitative discussion.

I obviously cannot agree with the last conclusion that the analysis provides a good confidence level about ocean colour retrieval over the Baltic Sea.

The sentence has been removed.

I recommend the Authors to read biological literature about the phenology of the Baltic Sea phytoplankton, what kind of concentrations of chlorophyll-a have been actually observed in the Baltic Sea during different bloom periods and how this matches/contradicts with their findings. There is plenty of literature available for the Baltic Sea.

We did follow the Reviewer's suggestion and incorporated findings from the literature into the manuscript, and found that the ranges of CHL variability described in this work are fully consistent with those from the literature. Analogously, as mentioned above, the tentative description of the phytoplankton phenology sketched in this work is coherent with that available from the literature for the three regions.

Cyanobacterial blooms in the Baltic Sea have been known for large mats (scum) floating on the water surface. These mats may be several centimetres thick and cover areas of 200 000 km2 (Kahru and Elmgren 2014). These issues have not been discussed at all. If standard processing chains are used then the scum pixels are masked out as errors whereas the algorithms cannot cope with "terrestrial" reflectance (high NIR) in the middle of the sea. If these pixels are masked out then what kind of chlorophyll dynamics we discuss here at all?

This comment has been included into the main text and motivated the new figure 7, which confirms (after Kahru and Elmgren, 2014) that even standard algorithms can be used for the monitoring of phytoplankton biomass into the Baltic. Moreover, Kahru and Elmgren (2014) in their in depth analysis of the cyanobacteria blooms explicitly mention that only a small (<5%) fraction of the bloom pixels is affected by atmospheric correction failure, due to surface scum.