We thank the reviewer for her/his positive *comments*. We report here our proposed replies to the specific points. If accepted, these notes will be properly included in the revised version.

1. (P8 L1-10) Authors discuss further in the manuscript that direct comparison of model results with the sensor data should be done with caution as the applied corrections may not reflect true value. Thus, application of the introduced metrics which is consistent within itself (e.g. normalising the data to its own surface value) is a good approach. As these are new metrics, for their applicability to other regions, the choice of the criteria deserves a discussion. Do 10% of the surface chl value, or 2mmol/m3 nitrate have a significant meaning? Are these values applicable enough throughout the Mediterranean or have the authors seen regional inconsistencies? Are the strict choice of seasons (jan-mar and apr-oct) valid in practice (can't tell much from Figure 9, but wouldn't using MLDs from the ARGOs yield better estimation of which criteria to use? ARGO profiles show deep chl formation in late Jan and late March 2016 hence much deeper mwb than then model)

REPLY 1 - The metrics concerning chlorophyll and the relative two periods for the computation have been derived from the outcomes of Lavigne et al. (2015), who identified some standard shapes for chlorophyll profiles from the analysis of a large number of chlorophyll (fluorescence data) profiles in the Mediterranean Sea (see their Fig. 2). In particular, our summer period is based on the consideration that the DCM shape profile (useful to define our DCM index) is typically observed from April to October (Fig. 5 by Lavigne et al., 2015). Then, three other shape profiles (i.e., "homogeneous", "high surface chlorophyll-HSC" and "complex" in Fig. 2 by Lavigne et al., 2015) are characterized by decreasing values with depth, and typically occur between January and March in different Mediterranean regions (Fig. 5 by Lavigne et al., 2015).
The choice of the 10% criterion for the MWB index was made after a sensitivity analysis varying the limit between 1 to 10% (not shown). The value of 10% gave results qualitatively consistent with those reported by Lavigne et al. (2015). Further, lower percentage values gave more unclear patterns because the depth increases substantially and the thickness of model layers has also an impact (i.e. thickness of model layers is 5 m at 70 m depth, 7 m at 100 m and 10 m at 140 m). Thus, the selection of the period might depend on a preliminary investigation of the typical behavior of the ecosystem. Alternatively, an analysis employing machine learning technique and additional data such as DCM values can provide a better tool to identify shape profiles and their distribution in time, however this is out the scope the present work and might require a considerable amount of data.
The use of the MLD for the choice of seasons' limits represents a promising alternative option, since it would account for the specific conditions at each float profile. However, it must be noted that there are several possible definitions of MLD which can give slightly different results, thus a sensitivity analysis of the biogeochemical metrics to the MLD definition would be necessary to tune the choice of the criteria based on MLD. Therefore, we decided to adopt an a priori, while rigid, temporal subdivision based on literature to test the feasibility of the computation of the chlorophyll criteria.

Regarding the criteria for nutricline metrics, we are aware that selecting an unique criterion to detect the nutricline might be sensitive and controversial. Further, we think that the important aspect is to rather track the time evolution of the nutricline. For these reasons, we tested two different approaches: the depth of the 2 mmol/m^3 concentration isopleth (NITRCL1) and the depth of the maximum nitrate vertical gradient (NITRCL2). According to Manca et al. (2004), the values of nitrate concentration at depth higher than 400 m are around 4-5 in the eastern basin and 6-7 mmol/m^3 in the western, therefore the isopleth of 2 mmol/m^3 can be considered a safe value to detect the rapid change between the very low concentration typically measured at the surface and the high concentration at depth in all areas of the Mediterranean.

2. (P11 L8-10) Does this partly explain the lower modelled NO3 concentrations (e.g. nwm) due to the lack of N river load time-series? How does model perform in terms of N/P ratios? Does it represent the high N/P ratio character of the Med Sea and its regional differences?

REPLY 2 - We agree: the effect of the lack of high frequency data of nutrient discharges is one of the most important sources of uncertainty at the daily/weekly time scale (not at seasonal/annual scale) and at very local coastal scales, as discussed by Teruzzi et al. (2018). Indeed we highlighted this potential issue at (P17 L13-17).
Uncertainty in NWM is partly related to a possible underestimation of the river input forcing and possibly to the effect of lateral circulation from ALB and SWM1 surface waters (see Fig. 5 in the manuscript). A sensitivity analysis of the impacts of the different factors would help in elucidating the most relevant factor. However, the relevant point is that the fine sub-division of the Mediterranean Sea in 16 sub-basins allows to detect the relevant spatial gradients and, thus, to highlight possible issues. The choice of 16 sub-basins was a trade-off between having a number of areas as larger as possible, the need for having robust in situ statistics and the known characteristic on dynamics derived from literature. Thus, our proposed sub-division is a relevant result in itself. We propose to include this point in the part concerning the discussion on the sub-basin subdivision at (P16 L10-17), adding a comment on the detection of the surface nitrate underestimation of Fig. 5 at (P16 L17):
"In fact, the comparison of nutrients profiles of Fig. 5 allows to highlights the satisfactory model performance in reproducing the mean spatial gradients and possible anomalies such as the underestimation of Nitrate in NWM sub-basin, which is related partly to possible underestimation of the terrestrial input and partly to the impact of the incoming Atlantic waters".

Concerning N/P ratios, the performance of OGSTM-BFM model system was assessed in Lazzari et al. (2016) for nitrate and phosphate separately, showing a general higher-than-Redfield ratio in the Mediterranean Sea (closer to a N:P=22 ratio, with exception of the Alboran Sea area, characterized by a lower-than-Redfield ratio) and significant spatial variability. The present operational configuration of the MedBFM incorporates the results presented by Lazzari et al. (2016; see their Fig. 7 here reported as Fig.R1, for further details please refer to the paper) and provides consistent results on N:P ratio (see also N and P values in Fig. 5).

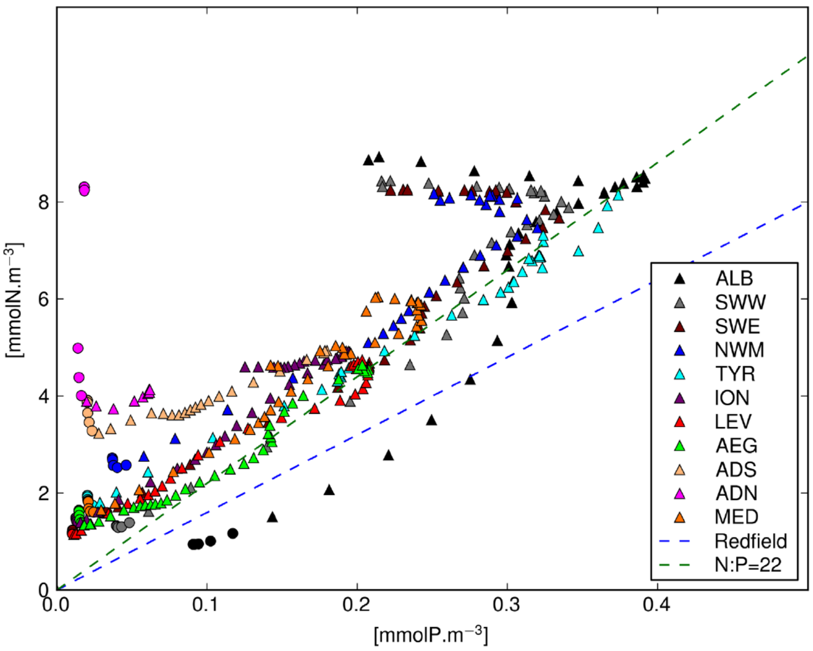

Fig.R1 - Results from Lazzari et al. (2016), model-derived vertical profiles of phosphate (x-axis) and nitrate (y-axis) averaged by sub-basins (reported in Fig. 1 of Lazzari et al., 2016) for the period 1999–2004. Each entry represents the spatial-temporal average at a certain depth: 0–50 m layer (circles) and 50-bottom layer (triangles).

3. (Fig9) How does BGC-Argo surface chl compare with the satellites? P11 L4 suggest the model has a higher (0.015) bias for the winter (model vs satellite), supported by Fig8 with more pronounced bias for the west/northwest Med, while paragraph of P11 L32 suggest the model has lower values when compared to BGC-Argo data. Is there a consistent ratio between satellite and float data, and how applicable is it to use global correction of division by 2 as suggested by Roesler et al. (2017) taking into account the regionally different ratios shown in the same article. As the Mignot et al. (2018) manuscript is in review, I cannot comment about their results but can the application of their method suggest different correction factors with a better regional fit?

REPLY 3 – We thank the reviewer for having raised this point, since the inconsistency in surface chlorophyll observations between satellite and BGC-Argo floats has been already observed in our investigations (see Fig.R2) and represents a potential issue, not only for validation purposes as discussed here, but also for multi-platform data assimilation (e.g., combined assimilation of chlorophyll from BGC-Argo floats and satellite). Regional corrections of BGC-Argo float data can be advisable, such as regional algorithms for the ocean color exist for satellite, to provide chlorophyll in different regions of the global ocean. Methods such as that of Mignot et al. (2018) can be helpful proven the availability of a sufficient amount of in situ independent data.
Such investigation is off-topic with respect manuscript, we can comment that operational systems are optimal tools to test the consistency of many different sources of information as already reported in the submitted manuscript citing She et al. (2016) at P2-L23 and P17-L2.

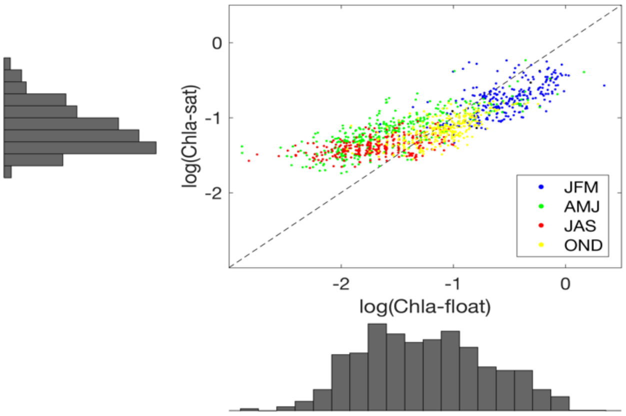

Fig. R2 - Chlorophyll concentration comparison between surface BGC-Argo floats data and satellite data (CMEMS ESA-CCI Mediterranean data). Data of all floats available between 2015-2017 are here included. The matchup between data is performed on a daily basis and using a bilinear spatial interpolation. Positive deviation of BGC-Argo floats w.r.t. satellite data is evident in winter while negative deviation of BGC-Argo data is present in summer and spring months.

4. (P14 L20-23) Authors point out that the introduced BGC-ARGO related metrics are already being implemented for data assimilation purposes with consequent improvements in model solutions. Before the assimilation phase (e.g. pre-opetaional runs), does the skill assessment documented here (of BGC-Argo metrics) reveal any prior messages for model parameter adjustment such as for light attenuation or nutrient assimilation rates, or errors of physical model origin? I can see the use of this dataset not only for forecasting purposes and skill assessment purposes, but its high resolution coverage including ocean interior is of high value. A short comment on that would be good scientific addition to the manuscript.

REPLY 4 – We thank the review for this comment. The integration of BGC-Argo floats within an analysis and forecasting system (in terms of data assimilation) paved the way to an in-deep study of the interior of the sea and its dynamics. Given the fact that the BGC-Argo provide also profiles of PAR, salinity and temperature, which are forcing mechanisms of the biogeochemical processes (or proxies for the forcing impacting the biogeochemistry), a global analysis of the uncertainty can be made comprehensively using multivariate statistical analysis (e.g. PCA, neural network methods), with the aim of disentangling the sources of error on profiles.
However, it must be noted that this analysis would need not simple measures of distance between observations and model values (such as BIAS and RMSD), but indexes that can put in relation the shape and intensity of the profiles with the underlying processes. We think that our work provides a first step to identify and quantify several functional indexes. Another critical point is the availability of a sufficient amount of profiles for variables like nitrate and oxygen, which may allow for statistically significant analysis. A short comment will be further added on this point in the discussion section.

5. I see that the manuscript is designed as a document for the overall skill assessment of MedBFM, but both the abstract and the manuscript throughout have stressed the importance and usefulness of their new metrics (GODAE Class 1 and 4, and especially the use of BGC-Argo), and I agree with them, and these sections of the manuscript stand out as the novel scientific content. The title fails *to give this message and won't promote this novel scientific content of the manuscript. I leave it to the authors consideration.*

REPLY 5 – We agree: the title will be changed highlighting the novel metrics and the usefulness of BGC-Argo float data. Possible title may become: “Skill performance of the CMEMS Mediterranean marine ecosystem forecasts: improving model uncertainty assessment using BGC-Argo floats data-based metrics”.

Minor/technical comments will be also thoroughly addressed in the review, correcting in particular Figs. 10 and 13.

Best Regards