Detailed responses to reviewer 1 (reviewer comments are included in black, responses in blue font)

Overview

The study addresses the important question of whether the Earth system models can accurately simulate biogeochemical processes/fields on the Northwest North Atlantic shelf and if they are suitable for projecting the regional impact of climate change. The study ranks the CMIP5 and CMIP6 models in terms of capturing the observed seasonal variability in surface temperature, chlorophyll and nitrate in the Northwest North Atlantic shelf, highlighting that ESMs are not adequate to project accurately future changes in the NWA shelf. Based on the comparison with these ocean surface observations the study clearly demonstrates how a regional model outperforms the Earth system models in terms of capturing the seasonal variability in the region during the historical period. The study then provides plausible reasons for the better performance of some of the Earth system models and the improvement from the CMIP5 to CMIP6.The study is interesting, timely and I believe it is a substantial contribution towards improving our understanding and ability to provide accurate regional projections of climate change on the Northwest North Atlantic shelf, particularly in the context of biogeochemical fields. The methods and results are reasonable, justified and well presented. Hence, in my opinion the study fits well the Biogeosciences scope and should be consider for publication after addressing some very minor comments/revisions as I explain below.

Response: We appreciate the positive and constructive feedback. We provide detailed responses to the comments below.

General comments

Comment:
1. I agree that not all the ESMs are suitable to force regional model projections and should be selected carefully as not to force the regional model with unrealistic large-scale patterns. The choice of the 'parent model' should be based on its performance on capturing the large scale patterns/features that are introduced in the regional models such as: large scale circulation patterns, and biogeochemical and physical fluxes in-out of the regional model domain (the study highlights well this point in the introduction, lines 48-59 and in discussion, lines 343-351). Misrepresentation of the ESMs surface fields in the Gulf of Maine, Scotian Shelf and Grand Banks may be due to the models not being able to accurately represent on-shelf fine-scale biogeochemical and physical processes due to lack of resolution (which is inadequate to resolve the physical/biogeochemical processes on the shelves in all the ESMs, as mentioned in the manuscript), lack of accurate river discharge, biogeochemical model parameter optimisation for global use rather than the region of interest, or lack/misrepresentation of other local processes (e.g., enhance mixing associated with tides). The regional model is meant to address the above on-shelf problems, and as long as the large-scale patterns are introduced correctly at the open boundaries it should be able to perform well. Hence, I am a little reserved about the conclusion that 3 specific ESMs are most appropriate to force regional model projections in NWA (lines 21-23 and 77-79) based on the ESMs' performance on the shelf at the surface. In my understanding, there is no explicit analysis in the study to link the ESMs

ranking on the shelf with the large-scale physics/circulation patterns off-shelf, or across-shelf properties exchange beyond a moderate correlation of surface temperature and surface chlorophyll, and no explicit link between the ESMs performance on the shelf and their performance off the slope in the vicinity of the regional model boundaries. So in my opinion, I would suggest to further discuss and clarify this issue: does low ranking on the shelf translate directly to low ranking of the larger scale patterns in the region and specifically to low ranking in the quality of the ESMs exchange/input of physical and biogeochemical properties at the location of the open boundaries offshore the Scotian shelf, Gulf of Maine and Grand Banks? Why are the ESMs with high ranking on the shelf the ones that can provide the best open boundary conditions coming off the shelf?

Response: We agree with the reviewer that ESMs performance offshore, along a regional model boundary such as the ACM, is most important for regional downscaling and may differ from those on shelf, although we suspect that performances in/out of the shelf are related. To address this specific point in the revised manuscript we will provide an additional analysis of ESM performances along the ACM boundaries and compare those with the results from the shelf.

We will also clarify the objectives and findings of our analysis and provide further discussion about ESM data, which are not only used to provide boundary conditions for regional downscaling. For example, ESM projections can be used to drive higher trophic level models and to assess societal impacts of climate change, such as fish catch, that affect mainly Exclusive Economic Zones (EEZ), i.e. coastal ecosystems. Our results indicate that the choice of ESM for these projections is very important and that 3 specific ESMs are most appropriate in the northwest North Atlantic region.

Specific comments

Comment:
2. Line 75 (typo): I think an ‘on’ is missing such that: ‘based on the mismatch...’

Response: Corrected.

Comment:
3. Lines 147-148 (just a suggestion): The authors can refer and link R and Rbio symbols in Table 2 directly with the rank for temperature, chlorophyll and nitrate, and the rank of chlorophyll and nitrate only, respectively, to help the readers follow more easily the ranking metric.

Response: Thank you for the suggestion, we will refer to $\overline{R}$ and $\overline{R}^{bio}$ here.

Comment:
4. Line 179: Please can you clarify which months you consider for the winter bias for nitrate? In my understanding winter is defined as December-February (based on Table3) but November-January are excluded from the observations for nitrate, so is it only February?

Response: Nov–Jan were excluded from the World Ocean Atlas (open/close circles in Fig 5g-i) and therefore in the Gulf of Maine there are only February data for the winter. However, for the Scotian Shelf and Grand Banks AZMP data were also available (black squares in Fig 5h-i) and are included in the winter (December-February) bias calculation. This will be clarified in Figures 4 and 5 captions.

Comment:
5. Line 233 (just a suggestion): Maybe if you could clarify that here you mean the best overall ESM in terms of nitrate and chlorophyll (the Rbio in Table 2) by adding something along the lines: 'ACM and model 22 (the best overall ESM for combined nitrate and chlorophyll) indicates...'.

Response: The sentence will be changed to:

"*ACM and model 22 (the best overall ESM for combined nitrate and chlorophyll, Table 2) indicates that…*"

Comment:
6. Lines 268-270: I agree that some of the errors in chlorophyll are linked with the temperature bias and the misrepresentation in general circulation. However, in my opinion i) the relative moderate correlation between chlorophyll and temperature r=0.51 and ii) the improvements in chlorophyll in CMIP6 relative to CMIP5 models that are not associated with any improvement in temperature (as discussed in lines 321-328) indicate that the errors in surface chlorophyll are also driven to a significant degree by a poor biogeochemical model component rather than by the ocean physics only. Hence, I suggest that the authors could modify this sentence to reflect this.

Response: This sentence will be modified as follows:

"*The correlation between temperature and chlorophyll scores indicated that errors in surface chlorophyll concentration were partly driven by the misrepresentation of the general circulation and, more generally, of ocean physics. The improvement in chlorophyll from CMIP5 to CMIP6 without an associated improvement in temperature suggest that the errors in surface chlorophyll were also driven to some extent by a poor biogeochemical model component*"

Comment:
7. Section 4.2: Although the authors mention it, in my opinion, they could highlight even more that the lack of correlation between resolution and rank is not surprising as none of the ESMs have the resolution to explicitly resolve the processes in shelf-scales rather than parameterise them. Maybe the authors could add a sentence after line 303 along the lines: 'This lack of correlation between model resolution and accuracy on the NAW shelf and the primary control of performance by the model set-up is not surprising as all ESMs are coarse and do not explicitly resolve the shelf-scale processes but rather rely on their parameterisation.

Response: We will add the following sentence after Line 305:

"*The lack of correlation between model resolution and performance on the NWA shelf is not surprising as all ESMs are coarse and do not explicitly resolve shelf-scale processes but rather rely on their parameterisation. Much higher resolution will be necessary…*"

Comment:
8. Figure 8 caption (just a suggestion): The authors mention that the temperature rank is hidden for model 6, however, in my understanding ranks for other models are also hidden as they coincide (model 1, 30 and 18). Maybe just for clarity and to make the link between figure 8 and table 2 explicit, you could mention in the caption something along the lines 'Coinciding ranks as shown in table 2 are hidden'.

Response: The following sentence will be added to Figure 8 caption:

"*Hidden coinciding ranks (models 1, 6, 30 and 18) are provided in Table 2.*"