Detailed responses to reviewer 2 (reviewer comments are included in black, responses in blue font)

General comments

Comment:
1. In this paper, the authors compare the output of CMIP5 and CMIP6 Earth System Models (ESMs) to observations in order to determine which models are suitable to build boundary conditions for projections. A ranking analysis was performed on a large array of ESMs. However, they are only looking at surface values of 3 variables and far away from the regional model boundaries, even though they mention on lines 44-46 that it is important to look at the information imposed at the boundaries. I think the objective stated on line 67 "Our objective is to assess the performance of a number of available ESMs in reproducing present conditions on the NWA shelf in contrast to a high-resolution regional model" is more in line with what is presented in the manuscript since there is no analysis at the boundaries.

Response: We agree with the reviewer that ESMs performance offshore, where the regional model boundary is located, 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 that was also raised by reviewer 1, we will add an analysis of ESM performances along the ACM boundaries and compare those with the results from the shelf.

As mentioned in the response to reviewer 1's general comments, we will also clarify the objectives and findings in the revised manuscript and provide further discussion about the regional use of ESM data. 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.

Comment:
2. They are not discussing the processes that lead to the observed values in the region under study and they are not analysing if the models do represent these processes correctly. I believe salinity should be included, as surface temperature depends strongly on atmospheric forcing while salinity is more representative of the different water masses in that region.

Response: The study is meant to provide ESM users with information about model performance for either direct use or regional downscaling. We do discuss to some extent the potential sources of mismatch between models and observations but 1) this is not the objective of the analysis and 2) we can only speculate on the sources of errors.

We included temperature in the comparison because it is an important variable for higher trophic level studies and climate change impacts. The fact that surface temperature is available at high spatial and temporal resolution on the shelf, similar to chlorophyll, is also important. Despite the tight control by atmospheric forcing, we did find significant

differences in surface temperature across the ESMs. We believe these differences are of interest and relevant to many users.

Large scale salinity patterns in the historical simulations are likely related to those of temperature and therefore it is not clear if adding salinity to the comparison would provide additional information. However, since salinity is available from the WOA dataset at the same resolution as NO_3, we will compare simulated and observed salinity and add the results to the manuscript.

Comment:
3. Moreover, it is very surprising that a similar study (Lavoie et al. 2019) in the exact same area, with the same purpose, and using some of the same ESMs is hardly mentioned at all. No comparison of the results of this study with the 2019 study is made.

Response: We agree with the reviewer that we should discuss our results with respect to the findings of Lavoie et al. (2019). We did mention the conclusions of an earlier report (Lavoie et al., 2015) but will expand this discussion and also add the more recent study to the revised manuscript.

Comment:
4. Also there is not enough details on the comparison with the data, they appear to be comparing different time periods (see detailed comments) or on how the ESMs were brought to a single grid.

Response: We provide the information about time range, averaging and spatial mapping in the Methods. Additional information will be added for completeness, as detailed in the responses to detailed comments 18–20, 27 and 30 below.

Comment:
5. There is only a vague mention of what the improvements are between the CMIP5 and CMIP6 models. What was improved should be stated (not only biogeochemistry of physics) so that the reader can judge on the potential impact on the ranking.

Response: See response to comment 2 above and comment 25 below. Model changes from CMIP5 to CMIP6 can be significant, including in the atmospheric and terrestrial realms, and the study is not meant to find out what are the sources of improvement in performances. Give the limited output available from these models, we can only speculate on the sources of improvement based on our results. The reader is referred to the specific papers listed in Table 1 for the list of changes in the models. To clarify this point we will add the following statement L317:

"*For specific changes in the CMIP6 model versions, the reader is referred to the references listed in Table 1*."

Comment:
6. Increasing the model resolution in order to improve the representation of the circulation in the NWA has been mentioned by many authors (e.g. Loder, Brickman, Yool). Here it is stated that the resolution does not have an impact. This is a big

statement, considering the general agreement, and it should be demonstrated. The authors could show the changes in circulation of a few models they are giving in example for this.

Response: Our findings are in line with previous work, included the ones cited above, see responses to detailed comments 9, 16, and 34–35.

Comment:
7. All these points should be addressed in order for the conclusions to be more convincing (ranking based on analysis of shelf surface conditions representative of boundary conditions).

Response: These points are addressed in the detailed comments below.

Comment:
8. Also, Lavoie et al. (2019) estimated that the boundary conditions obtained with the ESMs were not as reliable for the simulation of the conditions on the Scotian Shelf and in the Gulf of Maine. It would good to know if there was an improvement in this regard with the CMIP6 ESMs.

Response: In the revised manuscript we will provide an analysis of model performance along the ACM boundaries. The change in ranking from CMIP5 to CMIP6 along the western, southern and northern boundary will show if there was an improvement in the CMIP6 models. We will discuss these new results with respect to the findings of Lavoie et al. (2019).

Specific comments

Comment:
9. Line 11: Here you say that the coarse resolution is not appropriate to represent the circulation and elemental flux but later on you say that increasing the resolution does not matter. Is it important or not?

Response: Our two statements are in agreement. It is well known that the coarse resolution of ESMs is an issue to resolve shelf-scale processes and that high resolution is necessary in these areas, as mentioned in comment 6 above. However, even the highest resolution ESM from our ensemble is too coarse to resolve shelf-scale processes and therefore it is not surprising that we do not see better performance with increasing ESM resolution. We will clarify this point by adding the following sentence 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:
10. Line 14: ability to reproduce surface observations...

Response: Will be corrected.

Comment:
11. Line 15: why is it particularly sensitive?

Response: We refer to the effect of climate change on the location and strength of the Gulf Stream and Labrador Sea currents. We will clarify the sentence in the revised manuscript.

Comment:
12. Line 16: The spatial mismatch in large-scale circulation was not demonstrated. There are references for CMIP5 but what about CMIP6. Changes, or not, in circulation should be shown/mentioned after an inspection of the ESMs results.

Response: We mentioned a warm bias in the Gulf of Maine that is in line with the results of Loder et al. (2015) and Saba et al. (2016) (Lines 365-266). Although smaller, a cold bias appears on Grand Banks in most models (Figures 4c and 5c). The biases suggest a mismatch in the large-scale currents. However, since we did not compare the position of the currents across the models, we will rephrase the sentence L15–17 as follows:

"*Most ESMs compare relatively poorly to observed nitrate and chlorophyll and show differences with observed temperature that suggest a spatial mismatch in their large-scale circulation*."

Since we will add salinity and look at offshore conditions along the ACM boundaries we will have more support for this statement.

Comment:
13. Line22: How can we say just by looking at the surface temperature, nitrate and chl a that the top three models are appropriate for boundary forcing? The model boundaries are hundreds of meters deep (and more) and are not located in the regions analysed. It should be mentioned what are the tracers that will be downscaled at the boundaries? Salinity is certainly one of them, why was is not included in the analysis?

Response: The revised manuscript will include both salinity and a comparison along the offshore boundaries of the ACM, see responses to comments 2 and 8 above, and response to comment 1 by reviewer 1.

Comment:
14. Main text Line 71: why look only at three variables? What about salinity?

Response: Salinity will be included, see response to comment 2 above. Not all variables were available for all models (ESMs and ACM) and could be compared to observations so we restricted the comparison to 3 variables, plus salinity in the revised manuscript. The selected variables are, arguably, the most important to potential users.

Comment:
15. Line 78: historical simulations are not used for projections. This should be rephrased.

Response: TBA.

Comment:
16. Lines 115-116: The ESMs horizontal resolution in the region of interest should be given in Table 1. Some models have a variable resolution and it might not be that bad in the NWA.

Response: To give a sense of horizontal resolution that is easily comparable across models we provide the number of grid cells in the three zones of interest in Table 1. This value also depends on the coverage, which can be very poor for coarse grids (e.g. IPSL–CM5), and therefore provides more information to the reader. However, as suggested by the reviewer and since resolution is typically reported in degrees for ESMs, we will add a column to Table 1 with average $\Delta lon \times \Delta lat$ on the NWA shelf.

Comment:

17. Line 117: MR and HR mean medium resolution and high resolution respectively. If they share the same grid where does the change in resolution come from?

Response: MR stands for Mixed Resolution and HR for Higher Resolution in the MPI model names. MPI-ESM-MR (CMIP5) and MPI-ESM1-2-HR (CMIP6) have the same ocean circulation model but the horizontal resolution of the atmospheric component was improved from ~200 km (MPI-ESM-MR) to ~100 km (MPI-ESM1-2-HR). Thus, model names are not related to the ocean model, which can be confusing. A similar confusion can occur from the IPSL CMIP5 model names. In this case, the models share the same ocean model, but the horizontal resolution of the atmospheric model is higher in the medium resolution (MR) version compared to the low resolution (LR) version. To avoid some confusion, the following text will be added to the caption of Table 1:

"*Note that the IPSL-CM5 models share the same ocean component with higher resolution atmospheric component in the MR version. Similarly, MPI-ESM-MR and MPI-ESM1-2-HR share the same ocean component with higher resolution atmospheric component in the HR version*."

Comment:
18. Line 123: From where were the satellite data obtained. Who did the averaging?

Response: Links to the data will be added as follows:

"*1) satellite surface chlorophyll observations from the Sea-viewing Wide Field-of-view Sensor (SeaWiFS) as 8-day averaged maps at 1/12˚ resolution (1999–2010, https://doi.org/data/10.5067/ORBVIEW-2/SEAWIFS/L3M/CHL/2018), 2) surface nitrate from the World Ocean Atlas 2009 (WOA; Garcia et al., 2010) at 1˚ resolution, and 3) surface temperature from the Operational SST and Sea Ice Analysis (OSTIA) system (Donlon et al., 2012) at 1/20˚ resolution (2006–2016, https://doi.org/10.5067/GHOST-4FK01). Monthly climatologies were calculated for each of these*."

The following references will be added:

SeaWiFS. NASA Goddard Space Flight Center, Ocean Ecology Laboratory, Ocean Biology Processing Group. Sea-viewing Wide Field-of-view Sensor (SeaWiFS) Chlorophyll Data; NASA OB.DAAC, Greenbelt, MD, USA. doi:10.5067/ORBVIEW-2/SEAWIFS/L3M/CHL/2018. Accessed on 2014/03/12.

OSTIA. UK Met Office. 2005. GHRSST Level 4 OSTIA Global Foundation Sea Surface Temperature Analysis. Ver. 1.0. PO.DAAC, CA, USA. doi:10.5067/GHOST-4FK01. Accessed on 2019/12/06.

The original data were daily (OSTIA) and 8-day (SeaWiFS) maps which were converted to monthly climatologies, as mentioned Lines 126–127.

Comment:
19. Line 128: Which data from the AZMP were used? Along the Halifax line only? Why were the data averaged seasonally and not monthly like the other data?

Response: See also response to comment 30. Yes, along the Halifax Line where both high-resolution glider data and ship-based bi-monthly or seasonal data are available. The location of the data is presented in Figure 1. We will rephrase the sentence as follows for clarity:

"*In addition, the regional model was validated using high-resolution in-situ observations along the Halifax Line (Figure 1) from the Atlantic Zone Monitoring Program (AZMP, 2000–2014, http://www.meds-sdmm.dfo-mpo.gc.ca/isdm-gdsi/azmp-pmza/index-eng.html) and glider transects between 2011 and 2016 (Ross et al., 2017)*"

The glider missions and the AZMP data collection frequency along the Halifax Line were seasonal, which is why the spatially resolved dataset was averaged into seasons rather than months. At station 2 we were able to use a bi-weekly frequency for the AZMP climatology.

Comment:
20. Line 132: So the model results are brought back onto 3 different grids, one for each variable. Are the time period also adjusted? For example SST goes from 2006 to 2016. The CMIP5 historical period ends in 2005. How can the two be compared then? Also, there are probably models that have a higher resolution than 1◦ (see my comment for lines 115-116), what is the impact of decreasing the resolution (converting from higher to lower resolution) and having on the ranking analysis. And how is the conversion of one grid to the other done?

Response: We used a heterogeneous data set and for comparison we brought the data and model to the same temporal (monthly) and spatial (observation grid) scale. We used a long-term climatology for robustness. Ideally, we would use the same time range for observations and models but this was not possible. All the ESMs used the same time range (30 years climatology, 1976–2005) so their intercomparison is robust. Note that Line 193 should read "(1976–2005)", not "(1975–2005)", which will be corrected in the revised manuscript. Unfortunately, we could only run the ACM simulation for 15 years starting in 1999 so the ESMs and ACM simulations overlap for 6 years only. Since the

CMIP6 historical simulations end in 2014 it was possible to use the range 2000–2014 with the CMIP6 models. However, the ESM intercomparison would have been less robust and we decided to use the same time range for all the ESMs.

The conversion from grid to grid is simply a linear interpolation onto the observations grid. This information will be added Line 133 as follows:

"*For comparison with the observations, each model was mapped onto the SeaWiFS, WOA and OSTIA grids using a linear interpolation*"

Comment:
21. Also, how the thickness of the first grid cell compares between the different models?

Response: The ESMs have various vertical resolution. For completeness, the number of vertical levels will be added to Table 1. The thickness of the vertical layers may influence the model performance but this is inherent to the model configuration and therefore not relevant here.

Comment:
22. Line 175: What is the main difference between the CMIP5 and CMIP6 groups, why is it better? Improved BGC? Same question hold for nitrate.

Response: In the discussion (see L329–341) we speculate about the source of improvement in surface chlorophyll and nitrate fields. The suggested sources of improvement refer to the literature as we cannot substantiate the reasons for these changes from our data.

Comment:
23. Line 200: Figure 6 does not show the annual cycle.

Response: Here we refer to the data that are used to calculate the RMSD. The sentence will be modified to:

"some models are much better at representing the observed annual cycle, as indicated by the lower RMSD (Figure 6)".

Comment:
24. Line 208: Could you explain why? From local atmospheric forcing or circulation change?

Response: We can only speculate the reason why some models have poor scores for temperature. Lines 264–266 we mention the warm bias associated with a mismatch in the location of the Gulf Stream. However, we do not know why some CMIP6 models (CESM and GISS) have a large temperature bias. These models already had poor scores for temperature in their CMIP5 version. Line 323 the following sentence will be added:

"*Models with poor scores had already poor scores in their CMIP5 version and therefore the cause of their poor performance is likely the same.*"

Comment:
25. Line 209: What are the improvements in the CMIP6 models?

Response: Here "improvement" refers to the lower chlorophyll scores for the CMIP6 models 22 and 23 (CNRM-ESM2-1 and GFDL-ESM4). These models have the best chlorophyll scores after ACM. For clarity, "improvement" will be removed and the sentence will be:

"*The range of variability in chlorophyll scores did not reduce from CMIP5 to CMIP6 and given the relatively low scores of a few CMIP6 models (i.e. 22 and 23), the range is larger in the CMIP6 group (0.8–1.4, Figure 7, right panel) than in the CMIP5 group (1–1.4, Figure 7, left panel).*"

Comment:
26. Line 215: So this means that the model ranked 2 (and others) might not be ranked as high? What is the impact on the final choice?

Response: The numbers in the parenthesis correspond to the model ID. Since we excluded Nov–Jan from the WOA dataset, i.e. when nitrate is high at the surface, models with consistently low nitrate will have lower scores than they should, which will increase their rank with respect to nitrate. The overall rank is an average of the 3 variables so it should be less sensitive to this effect. Supporting Figures S1-S5g-i indicate that this might be the case for models 4, 8, 14, 19, and 26–27. Models 4, 8 and 19 have poor rankings so the underestimation of the nitrate score have no effect on the final ranking. However, for model 14, 26–27, information on the underestimation of the nitrate scores should be provided. To reflect a possible bias in the ranking of these models we will add an asterisk in Table 2 beside the overall rank of these models and will update the caption accordingly.

Comment:
27. Line 218: Could the fact that you are using different time periods and different grid resolution for the three variables explain the lack of correlation?

Response: We used climatologies to remove as much as possible the influence of time on the comparisons. For the sake of completeness, in the supplement we will provide 1:1 comparison of observed chlorophyll, nitrate and temperature on each grid and refer to the comparisons in the revised manuscript.

Comment:
28. Line 221: How do you explain that?

Response: We do not wish to speculate about the reasons for each models' individual ranking. As stated before, given the limited output that is available for each of the models we would have to speculate but this is outside the intended scope of this study. The objective is to report on the models' behaviour.

Comment:
29. Line 230: Why? Does it relate to temperature-dependant phytoplankton growth?

Response: Again, we can only speculate. Temperature-dependant phytoplankton growth is a possibility, large scale circulation is another. If poor chlorophyll scores also correspond to poor salinity scores in our new results then large scale circulation might be a better explanation. We will include this discussion in section 4.1 of the revised manuscript.

Comment:
30. Line 245: What are the years compared for the ACM and the glider data?

Response: Information on the ACM and glider data is provided in the Methods, i.e. Lines 111 and 130. The ACM data are the same as for the comparison with the ESMs, i.e. years 2000–2014 but presented as a seasonal (Figure 9) and daily (Figure 10) climatology to match the resolution of the glider data. The AZMP years are the same as ACM. The glider missions were carried out between 2011 and 2016 but were heterogeneous in time and space (see tracks on Figure 1). To enable a quantitative comparison between the glider and ACM data (Table 3), we spatially interpolated both dataset onto a transect following the Halifax Line (black line in Figure 1). The glider missions were seasonal, which is why the spatially resolved dataset was averaged into seasons (Figure 9). For each mission, data were extracted at Station 2 to produce a monthly climatology (Figure 10). ACM data were extracted at this location for comparison. We will add this information as follows in the Methods section.

Comment:
31. Line 260: Correlation coefficients are high for nitrate despite having a large bias and RMSD. This should be explained.

Response: The correlation coefficient is a complementary measure to bias and RMSD. Correlation and bias are largely unrelated. The former is a measure the similarity in spatial or temporal variations but does not account for bias. In other words, the same correlation coefficient can occur for very different values of bias. Likewise, high correlation does not imply low RMSD. In a noisy data set the RMSD will be higher than in a data set that is smooth, while both might display the same correlation.

Comment:
32. Line 270: See my previous comments about time-period and grid differences. I think that a statement about a misrepresentation of ocean physics as the cause should be backed up since later on the cause for nitrate mismatch is stated as coming from the BGC behavior (line 279). There are refs for the CMIP5 models but was there an improvement in circulation with the CMIP6 group or not?

Response: The mismatch is partly associated with ocean physics and partly due to the BGC model. To reflect this the 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:
33. Line 288: So here again, the model-data comparison was made on a different grid than for the ESMs. Shouldn't it be done on the same grid for an appropriate comparison?

Response: No, the grid for comparison depends on the dataset. Here, since we have both high (glider) and low (AZMP) spatial resolution data, we mapped the data along the Halifax Line.

Comment:
34. Line 295: Lavoie et al. (2019) also point at the misrepresentation of the remineralisation depth in those models as a likely cause. This also explain why some models having a coarse resolution still have good results with biogeochemistry. But the statement made below that improving the model resolution does not improve the representation of circulation and main features in the models, such as the representation of the Gulf Stream detachment point and flow around the Grand Banks should be demonstrated. There is a large consensus on that and it should not be stated lightly. The authors could actually show the mean currents between the two versions of a same model with improved resolution. Especially that you state that higher resolution is required to refine the projections on line 306. There is a contradiction here.

Response: See also response to comment 9 above. We agree with the reviewer that there is a large consensus on the effect of grid resolution on large scale circulation and our discussion is in line with this consensus. The resolution of the CMIP models is much coarser than the resolution of the models used to study the effect of grid resolution on the large-scale current systems of the NWA. Therefore, as pointed out by reviewer 1 (see comment 7 by reviewer 1), it is not surprising that current ESMs do not show the effect of grid resolution on model performances; much higher resolution will be necessary to see this effect. We will clarify this point as follows 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…*"

We will also add the following sentence Line 295:

"*In the NWA, Lavoie et al. (2019) suggest that the misrepresentation of remineralisation depth may lead to poor results in some models, despite their resolution*."

Comment:
35. Line 310: Here again it appears to be contradictory as you previously mentioned that BGC improvements we the cause for improvements in the CMIP6 ranking. There are likely different versions of the 4 BGC models mentioned, which should be specified in the table and considered in the analysis. Also, it could relate to the processes that control nitrate in the regions under study, they are different for your 3 regions. And how well are these processes represented by the ESMs?

Response: Here we refer to the general BGC component. There are not enough data to compare specific model version or parameterization. This paragraph is meant to point out that, in our comparison, there was no relationship between the type of model and the overall performances. But we cannot go further, and this is not the objective of the study. For clarification we will modify the paragraph as follows:

"*Although model performance is likely influenced by the biogeochemical model structure, we did not find a clear relationship between the type of biogeochemical model and performance. Here we only refer to the model type because the same model may have different parameterizations when used by different groups. While the inner and outer ensembles share only 4 biogeochemical models (PISCES, HAMOCC, TOPAZ2, NOBM) out of 13, there was no indication of consistently better performance for the biogeochemical models in the inner ensemble. For example, models using similar ocean biogeochemistry (e.g., PISCES: 5, 12–14 (CMIP5), 22 and 26 (CMIP6), and HAMOCC: 15–16, 18 (CMIP5), 28–29 (CMIP6)) had very different ranks, with no obvious relationship between overall model rank and the ocean biogeochemical model component*."

Comment:
36. Line 335: What was updated in the ocean biogeochemistry?

Response: As mentioned in the response to previous comments above, our goal is not to discuss the details of the models. The HAMOCC biogeochemistry module includes a new parameterization of detritus sinking, which may influence surface chlorophyll and nitrate, as suggested by Lavoie et al (2019). However, this explanation is highly speculative and we do not think that is should be included here.

Comment:
37. Figure 4f: In the suppl. figures, there is more chl a in the model than in the obs. The opposite is shown here.

Response: The supplemental figures S1–S5 present the individual chlorophyll time series for the 29 ESMs, whereas Figure 4f shows ECM ensembles. The "all" ensemble time series are calculated with all the individual ESM time series in Figures S1–S5. Figure 4f shows that even though individual ESMs can be close to observations during the spring bloom (e.g. HadGEM2, Figure S2f) and even significantly larger (e.g. CESM2, Figure S4f), all ESM ensembles underestimate the bloom.

Comment:
38. Figure 7: Maybe use ACM instead of ROMS.

Response: Will do.

Comment:
39. Figure 8: Could specify that ACM has the same rank for the three variables (only see one point)

Response: We will add the following sentence to the caption:

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