**Summary:**

We thank the reviewer for providing a number of helpful suggestions on where clarification or additional information is needed. Below, we have restated each review comment in bold text followed by our reply.

**1: L31: The 2018 Brickman et al. paper documents and reveals the mechanism for these intrusions and should be referenced here:**

**Brickman, D., Hebert, D. and Wang, Z., 2018. Mechanism for the recent ocean warming events on the Scotian Shelf of eastern Canada. Continental Shelf Research, 156, pp.11-22. https://doi.org/10.1016/j.csr.2018.01.001**

We have added the suggested reference.

**2: L73: Regarding NAO effects on fisheries. The Fisher et al. 2008 paper provides an excellent example of the NAO effect on fish distributions in the model area. The authors may want to check the paper out and include a reference to it.**

**Fisher, J.A., Frank, K.T., Petrie, B., Leggett, W.C. and Shackell, N.L., 2008. Temporal dynamics within a contemporary latitudinal diversity gradient. Ecology letters, 11(9), pp.883-897.**

We appreciate this suggestion and have added the reference.

**3: L158 …163: Does the z* coord system have partial bottom cells? Authors should clarify the vertical resolution of the bottom cell.**

Yes, we edited the text to read "Vertically, the model uses a z∗ coordinate (a height coordinate that is rescaled with the free surface; Adcroft and Campin, 2004) with 75 layers and partial bottom cells"

**4: S2.3 Spinup and Hindcast: L332-342: I am not sure that I followed the spinup procedure, or perhaps I do not understand the logic. The authors describe a 10y spinup for the BGCM component (using a perpetual 1993? -- clarify) which is then used as the initial BGCM field for the main model run, which starts from rest in 1993 using the Glorys TS field. This confuses me. Because the BGCM is part of "main model" then the physics model must also be spun up. Why not use the 10y spinup to start the physics model as well? It is rare to not spinup a model, even if it is initialized from a 3D "spun up" TS field. Please clarify this procedure.**

We acknowledge that temperature and salinity are commonly spun up beforehand in regional simulations. In this case, however, we are initializing using a reliable ocean reanalysis with the same horizontal resolution as the regional model, and our experience during model development was that initializing directly from this reanalysis produced more accurate simulations than trying to initialize from a spun up run. We did try several spin up methods, including repeating the first year of the simulation or repeating the first 10 years, and found that the initial conditions produced by these runs caused large and persistent errors immediately when beginning the main run.

We did use a short spin up run for the biogeochemistry, however, because the data we have to initialize the BGC from are primarily coarse-resolution climatologies and running a spin up allows the BGC component time to adjust from these less accurate and inconsistent initial conditions, at least at the surface.

To help clarify, we added a sentence to the text: “During model development, we found that spinning up the physics first, by either repeating the first year or the first 10 years, produced less reliable initial conditions than the high resolution GLORYS12 reanalysis and led to substantial errors in the beginning of the simulation.”

The text also now states that the BGC spin up simulation “for 10 years using 1993--2002 time series of forcings described previously” to clarify that it is using the time series, not a repeating year.

**5: L364: conservatively interpolated; some details on this would be helpful**

We updated this to read “For spatial comparisons where the observed product had a resolution of 1/4° or coarser, the model data was interpolated onto the observed product grid with a first-order conservative method.” For chlorophyll, the text now says it was “interpolated onto the NWA12 model grid with a method that preserved the geometric mean chlorophyll”. The first-order conservative method conserves area averages, which is useful when regridding data from fine to coarse resolution. If a log transform is applied before using the method, and the results are transformed back afterwards, the geometric mean is conserved, which is useful for log-normal fields like chlorophyll.

**6: L384: “introduce a small bias”; further explanation needed.**

We revised this text to read: “On average this will introduce a small shallow bias in the model MLD relative to the de Boyer Montégut (2004) MLD when the model mixed layer threshold is near or above 10 m or the diurnal cycle of 0–2 m temperature is large; however, as we examine only mixed layers during winter when mixing is deeper and the diurnal cycle is weaker, the difference should be negligible.”

Although not detailed in the text, to confirm this we calculated the two different definitions of the mixed layer depth for each cast with data above 2 m and below 40 m in the CTD database that we also used for the Northeast Channel metric. Averaged over all profiles (7,109) from January–March, the average difference is 3.6 m shallower for the model definition of MLD. The

10th percentile of the difference is -11 m (i.e., in only 10% of profiles is the model definition more than 11 m shallower than the dBM definition).

**7: L385 … (and Fig 6), re GS position: There are a number of recent papers discussing changes in the GS position. For a slightly different analysis the authors should have a look at Wang et al. (2022) [Wang, Z., Yang, J., Johnson, C. and DeTracey, B., 2022. Changes in Deep Ocean Contribute to a "See‐Sawing" Gulf Stream Path. Geophysical Research Letters, 49(21), p.e2022GL100937. https://doi.org/10.1029/2022GL100937]**

We agree that there are quite a few papers on this topic, probably more than we can succinctly mention in this paragraph. We tried to cite papers that mentioned a connection to other changes in the region, such as warming or increased salinity. However, we agree the suggested citation is also useful and so we have modified the sentence in the introduction to read: "Northward shifts of the Gulf Stream have cut off the cool, southward Labrador Current and amplified warming in the region (Brickman et al., 2018; Gonçalves Neto et al., 2021; Seidov et al., 2021), although some studies have found contrasting long-term trends in the latitudinal position of the Gulf Stream (Wang et al., 2022)."

**8: L453, re EPUs (Figures 1b, 18): For researchers working on the SS/GSL/NL region, the SS EPU would not be considered part of the SS. This should be changed to eastern GoM (EGOM) as in Pontavice et al.**

We agree that this region is unfortunately named. However, for clarity and reproducibility we want to be consistent with the dataset used by NOAA Fisheries in their State of the Ecosystem reports (which we also use for observations; (https://noaa-edab.github.io/tech-doc/epu.html#epu), which calls this the Scotian Shelf-Eastern Gulf of Maine EPU. We have thus revised the text and Figure 18 to use the full name for the Scotian Shelf-Eastern Gulf of Maine EPU. We still abbreviate it as "SS" in Figure 1b, however, to be consistent with the abbreviation used in the EPU bottom temperature dataset.

**9: Results: 3.1**

**GS position (F6): F6a,b: lon/lat on these panels please**

Fixed in revised version.

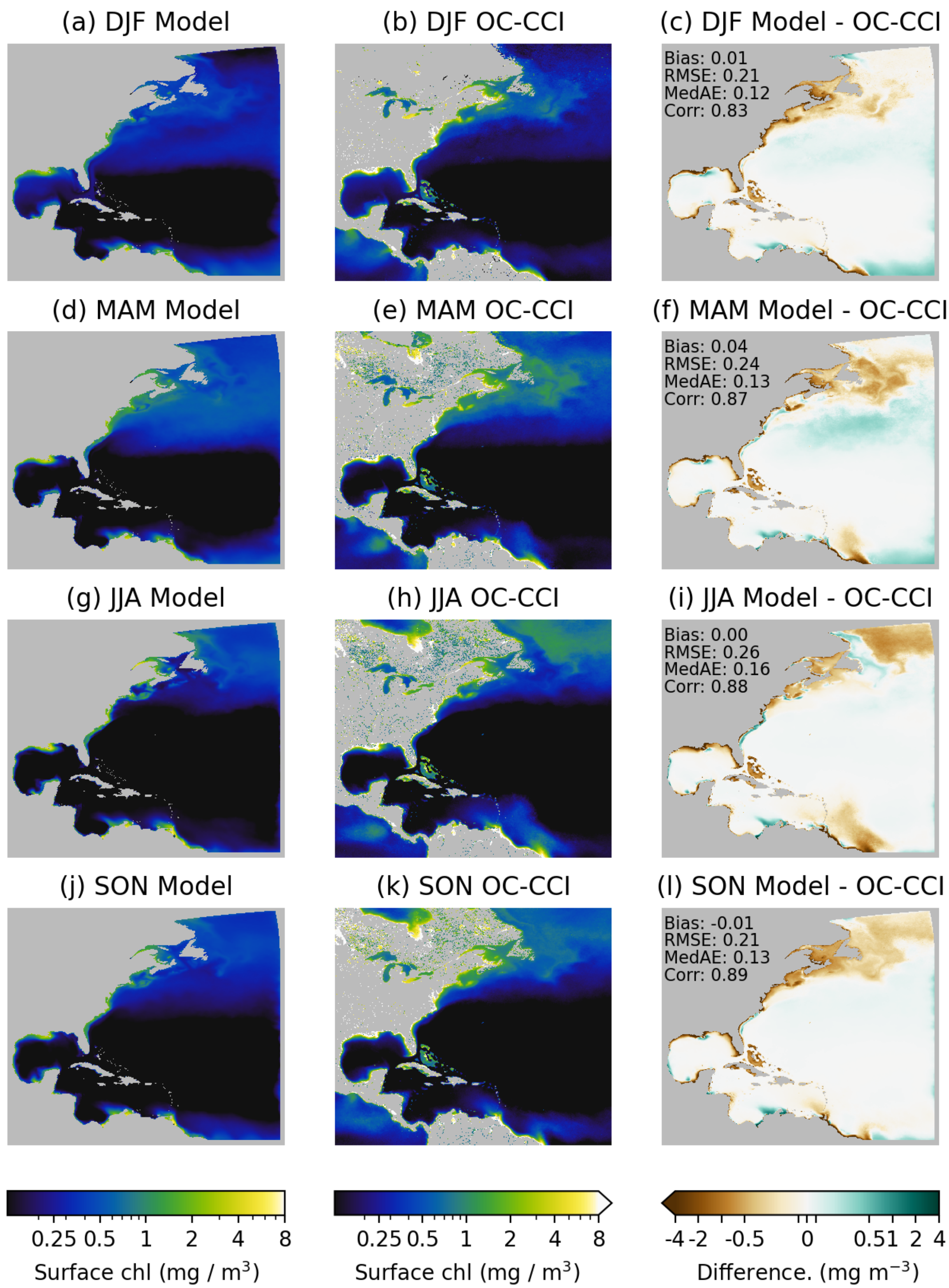
**L593: "cross-shore" is a bit confusing; consider north-south or meridional**

Fixed in revised version.

**No need for Fig12. Fig13 shows the results better**

We have kept Figure 12 because it shows some other aspects of the chlorophyll (e.g. the subtropical gyre minimum and along the South American coast). Also, in response to Reviewer

1, we have added panels to this plot showing the difference between the model and OC-CCI, which make this figure more useful.

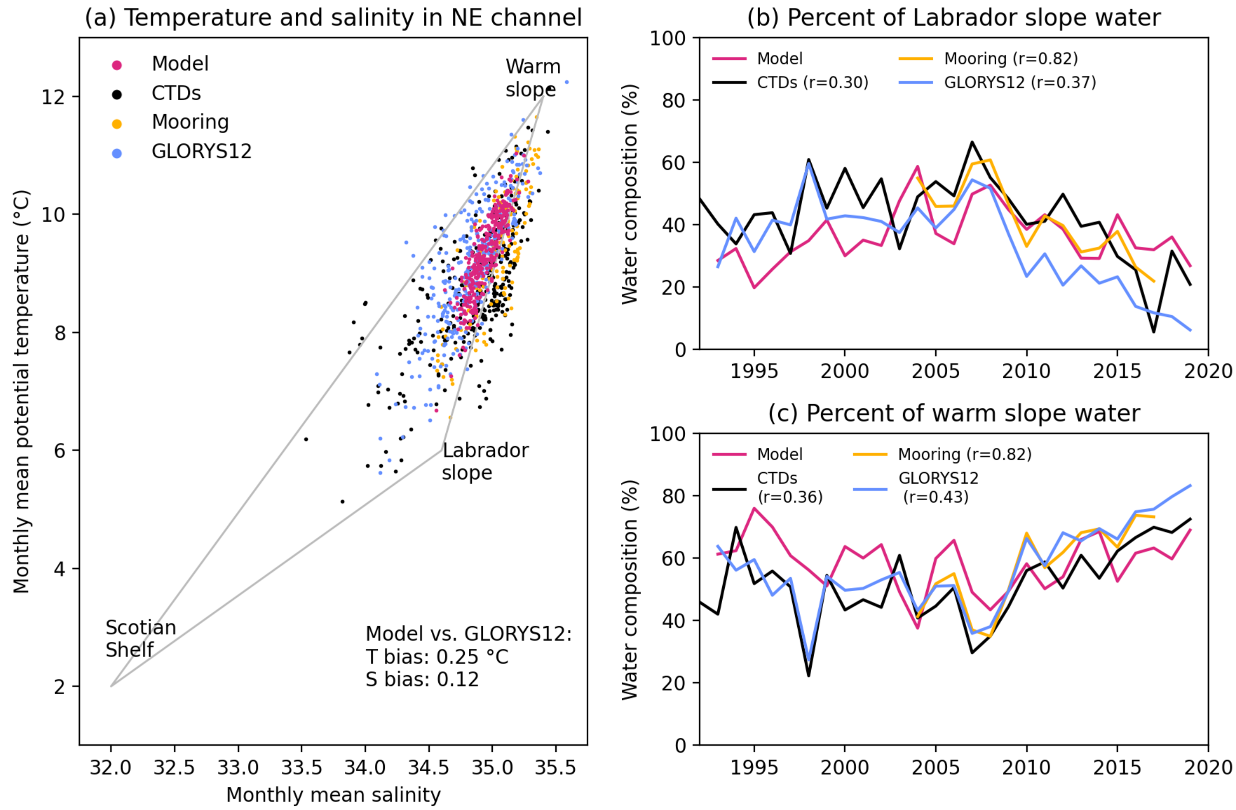



**L646: for clarity add the word "model" before "mesozooplankton".**

Fixed in revised version.

**10: S3.2**
**Deep Salinity: Fig19:**
**(a) TS fig. Please make colored symbols in the legend bigger so they can be seen.**

Fixed in revised version.




**(b,c) plots of %water masses (LSW & GSW): No mention of corrs in caption and it is not clear that the Glorys12 values are the model vs Glorys12 or Glorys12 vs one of the data series**

We added to the caption "Each correlation in parentheses gives the correlation between the model time series and the given observation dataset."

**Ice: F21,22: many models have problems with the advance, extent, and retreat of sea ice in the GSL (and on the NL shelf). This model does very well in this regard.**

We appreciate the comment and are also very pleased with how well the model simulates the sea ice.

**11. S3.3 Computational performance: Authors are commended on including this section. A couple of questions:**

**L771 (last sentence): Re scaling efficiency: The reason is not clear to me as wouldn't the (15x15 point / PE) restriction from the 2D BT solver still apply even when the BGCM component is implemented because the physics module still uses the BT solver? Please clarify.**

Yes, the BT solver would remain a bottleneck for the physics component of the model. However, with 40 additional tracers, the model with BGC spends a lot more time in the tracer routines. If these tracer routines continue to become faster as the number of PEs is increased, this would allow the model to continue to run faster with more PEs even though the barotropic solver is no longer becoming faster.

As a specific example, we compared runs from the 50x50 and 70x70 cases with biogeochemistry and looked at the average run times of individual steps and subroutines that MOM6 prints out at the end of the model run. The "Ocean barotropic mode stepping" step took a total of 2453 seconds in the 50x50 case and 2487 seconds for the 70x70 case; in other words, increasing the number of PEs from 50x50 had no impact on the time spent in the BT step. On the other hand, the "Ocean thermodynamics and tracers" step, which includes all of the biogeochemistry routines, took 15211 seconds for the 50x50 case and 10022 seconds for the 70x70 case, a speedup of about 1.5x and a meaningful reduction in the total run time. This speedup would not be nearly as prominent in a physics-only model with just temperature and salinity tracers.

We tried to clear this up in the text by changing it to:
"Furthermore, the scaling efficiency of this tracer-heavy model is better than seen in MOM6 models without coupled biogeochemistry where the computationally expensive 2-dimensional barotropic solver generally ceases to scale with fewer than 15x15 points per PE. The 40 additional tracers added by the BGC component result in more time spent in the tracer routines, which scale fairly well as the number of points per PE is reduced (and total PEs is increased) and produce a meaningful reduction in runtime even though the barotropic solver does not."

**L783: I do not understand this sentence, in particular the statement about reduced number of PEs needed by the longer thermodynamics timestep. I would have thought that the longer timestep requires fewer clock cycles, not fewer PEs. Perhaps my confusion is related to my comment above(?). In any case please clarify this**

We have deleted this sentence to reduce confusion. The sentence was referring to the computational resources needed to complete a run in a fixed amount of wall clock time. For example, if it was desired to complete a year-long simulation in exactly 12 hours, increasing the thermodynamics time step would allow the number of PEs to be decreased while maintaining the 12-hour run time. However, we realize this could be confusing and is not particularly important, so we have deleted this sentence.

**12: Discussion: L810: The Brickman et al. (2018) reference is relevant here as well.**

We have added the suggested reference.

**13: Conclusions: L866: I note that the authors mention a 0.5 corr as "commonly considered lower bound for useful prediction skill" although they do not provide a reference**

We added a citation to Murphy and Epstein 1989 ("Skill Scores and Correlation Coefficients in Model Verification"). We also changed "commonly considered" to "often considered" which in hindsight we think is more accurate since this is generally treated as a rule of thumb.