Reviewer #2

I am satisfied with the revisions made by the authors, and have only one last comment on the "nudging" method that authors used to the MOM6-NWA model:

In the methodology (lines116-118), authors suggest that temperature and salinity from the MOM6-NWA12 model "were nudged towards monthly means from the GLORYS12 reanalysis with a 90 day damping time scale", and "addition of nudging helps maximize the accuracy of the initial conditions" (lines 119-120). However, it is not described at all how the nudging is performed, e.g. the details of the nudging method and references. We only know it is not that "sophisticated" (line 128).

We thank the reviewer for reviewing the manuscript again. In response to this comment, we have changed the text as follows (new text in italics):

[...] temperature and salinity throughout the model domain were nudged *(i.e., restored using Newtonian relaxation)* towards monthly means from the GLORYS12 reanalysis with a 90 day damping time scale.

This should ensure the reader understands that the nudging we used is the commonly known Newtonian relaxation.

Reviewer #3

Manuscript Synopsis 2nd Review (although 1st review was done under tight time constraints and without access to usual sources of published research). This publication introduces a regional seasonal ocean forecast system for the United States east coast forced by the global SPEAR seasonal forecast and initialized with a subset of the GLORYS12 1 12 ◦ global ocean re-analysis over the domain in question.

I feel I still need to push back on the major point of my first review, which is the treatment of the GLORYS12 [Lellouche et al., 2013] re-analysis as both the verifying truth and initial conditions. As a producer of ocean analysis, I do not want to discourage their use for either the initialization of forecast, nor as a tool for verification. Verification of both atmospheric and ocean forecasts against their corresponding atmospheric or ocean analysis is done all the time – although I should add, not without its detractors, which I do not count myself as one – indeed, I regularly engage in the process. However, dynamical ocean and atmospheric analysis, as the authors are well aware – they spend several paragraphs of their introduction explaining why they wish to produce a dynamical forecast as opposed to existing statistical forecasts, which ultimately is the same achievement – is a best fit of the observations within a dynamically balanced

system to obtain a best guess estimate of the ocean or atmosphere, at least on the grid and resolution in question. Uncertainty is always associated with this estimate, especially when observations are sparse, as they almost always are for the ocean sub-surface, or ocean surface salinity. The former is substantially improved by ARGO (effectively sometime around 2005), and the latter could be corrected by remotely sensed sea surface salinity as is currently provided by the SMOS European Space Agency satellite, and formerly by the NASA Aquarius mission – but this is not as yet a standard assimilation observation in ocean analysis like GLORYS12.

The authors have made some attempts to elicite this in the manuscript, but I think it is important (at least to me) that this gets further discussed in the manuscript. However, despite a rather long list of major comments, I am really only asking for a very minor change as suggested in item #6. My goal is to inform – not to impose major changes on the authors' manuscript. For this reason, and the fact I unfortunately either missed, or forgot to include some minor points in my original recommendation, my recommendation is still for some Minor Revisions prior to publication.

1. Thank you for addressing all my concerns, and other reviewers concerns, with the earlier version of the manuscript. The remaining items amongst the major comments are some further push back to these responses. I do not expect any major structural changes to the manuscript, and indeed in the end suggest only a minor change as requested in item #6. Most of the minor comments are unfortunately minor points I missed in doing the last review – some noted last time, but omitted from my review (access was lost to the marked-up pdf), or simply not noted previously.

We appreciate the reviewer's thoroughness and helpful comments. We have made the change requested in item #6, as well as in several other comments where a revision was suggested.

2. As I stated in the synopsis: Although I do appreciate the authors attempts to recognize that the analysis do indeed come with some uncertainty, and I do highlight these below, I do believe some further discussion, particularly when discussing the proposed skill of the downscaled system, is still warranted. Points where the authors have established this uncertainty are:

- **ll. 127-128. We acknowledge, though, that deriving the initial conditions from a data assimilation process or a more sophisticated nudging method could improve the forecast spread and skill.**
- **ll. 277-280. It should be noted that the GLORYS12 reanalysis used as the observations in this comparison does not simulate salinity as well as it does temperature, and some of the reduction in bias may be due to the use of this reanalysis in the derivation of the initial conditions for the downscaled forecasts.**

In response to comments #2, 3, and 6, we have made the revision ultimately requested in item #6.

3. However, neither of the references they use to highlight the qualities of the GLORYS12 reanalysis [Amaya et al., 2023, Carolina Castillo-Trujillo et al., 2023], which are multi-system intercomparisons of which GLORYS12 is included, investigate whether the multi-model ensemble mean provides a better quality than GLORYS12. I am not recommending that the use of a multi-system ensemble mean would be a better set of initial condition – it would not be a valid dynamical state for one – but rather wish to again highlight there is uncertainty in the GLORYS12 analysis. The authors may wish to see Toyoda et al. [2015], which is a (now old) multi-system comparison of mixed layer depths for a set of global reanalysis, of which a much older version of GLORYS (0.25◦) is contributing. Northern winter mixed layer depths (Figure 11 of the article) along the North American east coast are amongst the most uncertain (largest normalized spread), indicating any analysis in that region is likely to be uncertain – particularly before 2005, after which ARGO becomes well established, as shown in Figure 1 of Storto et al. [2019].

See above.

4. The example of validating against OISST SST given as a counter-example in the authors' response to reviewers comments is likely the least interesting for me. SST is by far the most observed variable in the ocean, with satellite remote sensing able to accurately observe the sea surface temperature to a nominal resolution of 4km. Clould cover or rain may deplete those high resolution observations on a temporal basis, but I would expect different SST analysis to not actually differ substantially. So the fact that NWA12 performs better than SPEAR when validated with OISST SST analysis assimilated by the Kalman Filter assimilation in SPEAR is not really surprising to me. 5. Sea surface salinity (SSS) in not observed well in the ocean as already commented on by the authors. Similarly subsurface T/S profiles, mostly measured through ARGO floats that do not have long retention periods in the Gulf Stream / western boundary current areas of interest in this manuscript, are relatively sparsely observed. I would very strongly suspect that GLORYS12 will perform much better than the 1◦ SPEAR Kalman Filter assimilation in correctly estimating the subsurface, especially since the additional assimilation of sea surface height, in addition to GLORYS12 higher resolution giving it the ability to constrain the (at least larger scale) mesoscale activity that is known to improve ocean water mass properties in the analysis throughout the water column in the absence of any local InSitu observations [Fujii et al., 2024]. So the results I believe are really dependent on initial conditions, especially in the light of the fact that persistence is particularly skillful at predicting this, would be the bottom temperature results.

We agree that one should not expect SST to show the largest differences between sources of data used for initialization or verification. We thought it presented a useful contrast, though, since data from GLORYS was used in the initialization of the downscaled model, whereas data from OISST was used in the initialization of the SPEAR model. These two datasets are also produced differently, with GLORYS being a full data assimilative analysis and OISST primarily deriving data from satellites. We have not changed the manuscript in response to this comment.

6. Ultimately, I believe I would be satisfied if one further statement (modified to the authors preference) is added to their discussion of ll. 233-235: Forecast correlation coefficients are higher for bottom temperature (Figure 3), which partially reflects its increased persistence. Most downscaled forecast correlations are higher than the persistence correlation, however, though the majority are not significantly higher. This is perhaps not unexpected, as the downscaled NWA12 bottom temperatures are initialized by the exact persisted values coming from the GLORYS12 analysis used as validation. The lower correlated SPEAR bottom temperature values will likely be initialized somewhat differently.

We appreciate the reviewer's suggestion and have revised the manuscript accordingly (new text in italics below):

In the Northeast U.S. (Figure 3d--f), the pattern of forecast skill and downscaling improvement is similar to that seen for surface temperature. In other regions, NWA12 bottom temperature predictions have significantly higher skill than SPEAR for some cases where they did not for surface temperature. *This is perhaps not unexpected, as the downscaled NWA12 bottom temperature forecasts were initialized with data from the same GLORYS12 reanalysis used as validation, and the lower correlated SPEAR bottom temperature values partially reflect the persistence of differences between the SPEAR initialization and the GLORYS12 reanalysis. With this caution in mind,* skill and improvement on the Scotian Shelf […]

7. I would like to respond to the authors' response to the use of the AVISO gridded current product (DOI:10.48670/moi-00148). Yes, they are both Copernicus Marine Products. Yes, they are based on the same set of satellite data. No they may not be as similar as the authors state – the GLORYS12 currents being more than just geostrophic for one. I haven't done a comparison, and do not know of any comparison off hand – although I would be surprised if it is not in the Copericus quality 2 assessment document (QUID) of one, or both the products. It certainly was an omission to not include an independent observation only source of gridded currents in the Aijaz et al. [2023] manuscript, as this certainly was a standard in most of the ORA-IP papers [Toyoda et al., 2015, Shi et al., 2015, Palmer et al., 2015, Storto et al., 2015, Uotila et al., 2019]. This, however, harkens back to the previous point: There will be uncertainty in the assessment of currents, particularly near the coast, where the altimetry data is not as reliable (has larger error). The GLORYS12 product certainly was one of the better performing products in Aijaz et al. [2023], but not universally so, and likely the difference between the analysis is more indicative of this uncertainty, than it is of the individual analysis performance – in other words, the multi-system ensemble almost always out performs the member systems (again, not particularly explored in Aijaz et al. [2023]), but well explored in the other ORA-IP papers, even when mixing higher and lower resolution systems. Ultimately, the current assessment done in the manuscript here, which is only a trend assessment, does not depend too highly on this, and I will leave it there. No action required.

We thank the reviewer for sharing their knowledge on this topic.

Minor Comments

1. ll. 39-44. The NAO was shown to be predictable in Scaife et al. [2014]. However, Smith et al. [2020] do explain why large ensemble would be necessary to pull the low NAO signal out of the noise in the application of the seasonal forecasts to real world scenarios, although this it really first developed in Eade et al. [2014]. That being said, the SPEAR journal publication [Delworth et al., 2020] does not give any details on whether any NAO predictiability is present in SPEAR, I suspect not as I don't believe they will get a significant signal with only a 1◦ ocean model (at least 1 4 ◦ is needed), but this is my own personal prejudice. Without the NAO signal available in the atmospheric forcing from the parent model, the regional model will not benefit from this possibility, no matter the number of ensemble members. If there are oceanic precursors for the NAO, the initializing GLORYS12 ocean analysis will almost certainly contain them, but they will not be properly integrated forward without the signal in the driving atmosphere. However, reemergence of existing ocean signatures could conceivably precede without too much undo intervention required by the atmosphere. The driving atmosphere is all beyond the authors' control – although it would be helpful if the authors could give some statement on SPEARS ability to forecast the NAO if that could be tracked down from existing analysis of the system.

We believe this topic is best reserved for a different manuscript.

2. ll. 68-72. I will just gently prod the authors that they could have investigated whether the 30 low resolution SPEAR members could have outperformed the 10 high resolution NWA12 members as an addition to the provided 10 to 10 member analysis of skill. But of course, that was not the purpose of the manuscript. Perhaps the authors' may with to say why they choose not to pursue this.

We agree that this is beyond the scope of the present manuscript.

3. ll. 115,118. The manuscript does not actually provide a citation for GLORYS12 [Lellouche et al., 2013]?

We thank the reviewer for catching this oversight. In the newly revised version, we have added citations for the GLORYS12, ERA5, and GloFAS reanalyses where they are first mentioned.

4. ll. 153-155. I believe the authors are trying to find ways to improve their boundary conditions without breaking the seasonal hindcast paradigm of not using future information – which would prevent them from using GLORYS as they do in the Ross et al. [2023] reanalysis-forced historical simulation. A common way of using other than climatological boundary conditions in long term forecasts is to use an anomaly persistence, or damped anomaly persistence at the boundaries. In other words, use the GLORYS12 reanalysis as initial boundary conditions, but instead of continuing to use the

GLORYS12 reanalysis, use the climatological boundary conditions plus the initial condition anomaly of GLORYS12 with respect to the climatology. You then potentially have the opportunity to damp the provided anomaly over some time period. An example of using a climatologically evolving persisted anomaly with the SST boundary condition of a (formerly) atmosphere only system can be found in Lin et al. [2016].

Yes, we were careful to not include future information in the retrospective simulations. Our argument here was that due to the large extent of the regional model domain, the timescale for anomalies to propagate from the boundaries to the interior is longer than the 1 year length of the forecasts, so it does not matter much what we use for the open boundary condition data, and we can simplify things by only using climatology. Using persistence of damped persistence of anomalies is an interesting idea, though, that we will experiment with in the future. We have not changed the current version of the manuscript in response to this comment.

5. Figures 11, 12 & 13, SubSection 4.1.2. Firstly I apologize, I realize it is late in the game, and this is one comment I know I was going to comment on in my first review, but then failed to do so, but it is really hard to follow the discussion of Section 4.1.2 by having to flip back and forth between the text and at least 2 of 3 figures – even with modern ways of doing things, like having 4 versions of 3 the manuscript lined up across my two screens. I think you would do your readers a favour – and significantly help the weight of your argument if one could have all three results (i.e. Figures 11, 12 & 13) lined up as a single figure. For instance, if you restricted yourself to a snapshot every 2nd month (starting at 0), that would likely be frequent enough to show the propagation in each system, and then one could have the 3 systems (analysis, forecast, SPEAR forecast) lined up vertically, with the 6 time snapshots arranged horizontally. There is ample white space to the left and right to do this with no loss of magnification. But if you did insist on keeping the current 12 snapshots, one could still consider having a full page figure with the 3 systems across and the 12 snapshots vertically, but this would likely result in a decrease in the individual snapshot sizes. Please consider doing one of these two options.

We appreciate this suggestion for improving the readability of the manuscript. In the revised version, we have replaced the original 3 figures with a new single figure that shows all three datasets together. To keep the figure at a readable size, the new figure only shows 1, 3, 5, and 7 month lead times; this is enough to show the features mentioned in the text, and includes the 3 and 7 month lead times that are explicitly called out.

6. The authors may wish to know – and the reader may wish to know too, that there are theoretical formula that can explain the dependence on number of ensemble members in the presence of ensemble spread versus error for the CRPS in equation 3 of Leutbecher [2019], or the number of ensemble members in the presence of average correlation between members and average correlation with observation of the individual members in the correlation of the ensemble mean with observation with equation 2 of Murphy [1990],

p. 99. Indeed, for CRPS, one can define a “fair CRPS” score that would be independent of ensemble size as in equation 4 of Leutbecher [2019].

In response to this suggestion, we revised the manuscript to read (new text in italics):

[...] the improvement from a single member to two members is substantial, while the improvement from 4 to 10 members is minor. *This is consistent with the expected effect of ensemble size on estimation of the mean and CRPS (Leutbecher, 2018)*, [...]

7. ll. 525-528. My impression from the bias shown in Figures 5-7 is that they may be caused by errors in (too much) vertical mixing, and are not necessarily linked to the external/coupled atmospheric forcing. This impression is due to the phase relation (~ 90°) between surface and bottom – and could of course be an incorrect impression. Correcting the fluxes may not be the correction required.

This is a fair point. We added a sentence to this paragraph:
Reducing biases in ocean mixing, whether by correcting the wind forcing or adjusting the model parameterizations, may be especially important for reducing bottom temperature biases.