Reviewer #1

The paper provides an extensive study on downscaled retrospective forecast in the Northwest Atlantic Ocean from GFDL global model using a 1/12 configuration based on MOM6, previously designed and assessed by the Authors in another dedicated paper.

The methodology used for assessing the forecast is very interesting and quite comprehensive as well as the process-oriented analysis, supported by discussed results.

We appreciate the reviewer for reading the manuscript and providing encouraging comments.

Reviewer #2

I very much appreciate the opportunity to review this manuscript. This paper focuses on evaluating seasonal predictability of surface temperature and salinity and bottom temperature over the North America East Coast by using a dynamically downscaled model forecast system (MOM6-NWA12) and compared with the parent SPEAR model forecasts. Detailed discussions about sources of improved prediction skill from downscaling are included for the Northeast U.S. region, as well as discussions on the effect of long-term warming trends over this area. Besides, this paper also contributes a useful discussion on the ensemble size for reasonable prediction skill when predicting SST. This is a very important work with high quality contributing to the research field, and I only have a few minor comments on this work:

We thank the reviewer for taking the time to read the manuscript and provide a helpful review.

1. Some description about seasons in the Result section are confusing, not sure if authors are talking about initialization seasons or forecast seasons. For example, on lines 237-239, "the downscaled model has skill greater than persistence and SPEAR across a wide range of times, except in the winter…". It is not clear if "winter" here refers to the initialization month of December or forecast months in winter.

We see how this could be confusing. For lines 237-239, we have reworded it to "except for forecasts verifying in December". We have also clarified in this and a few of the following paragraphs whether the months or seasons we mention are the initial or verification months.

2. Lines 267-268, the Southeast U.S. LME, as shown in Figure 1, is not narrow compared to most other LMEs. I also question on its dominance by the Gulf Stream, as Gulf Stream is usually referred to the western boundary current north of Cape Hatteras (so north of the Southeast U.S. LME).

We reworded this to say that the Southeast US LME "has a narrow shelf and is dominated by the western boundary current".

3. Description of the forecast-observation mean bias (for Figs 5-7) could be more focused on those forecasts that have significant forecast-observation correlation coefficient (Figs 2-4).

In the revised manuscript we now mention that the SST biases in the SS and NEUS LMEs are improved in the downscaled model for forecasts verifying in autumn and early winter when the downscaled forecasts have the most skill.

4. Lines 278-284: authors could just write out the season name, instead of "first season", "last season", and "seasons 0 and 2".

In the revised version we have replaced "seasons 0 and 2" with "the first and third seasons".

5. Lines 294-295:
(1) "remaining three regions" -> "remaining four regions"?

Yes, we have fixed this to now read "remaining four regions".

(2) "aside from the increased spread in the Southeast U.S." not sure why it is "increased" when comparing with SS and NEUS based on Figure 9, please consider rephrasing this sentence.

We agree that this was worded confusingly. It now reads "Differences between the two models are smaller in the remaining four regions, aside from NWA12 having higher spread than SPEAR in the Southeast U.S. and lower spread and RMSE than SPEAR in the Floridian region".

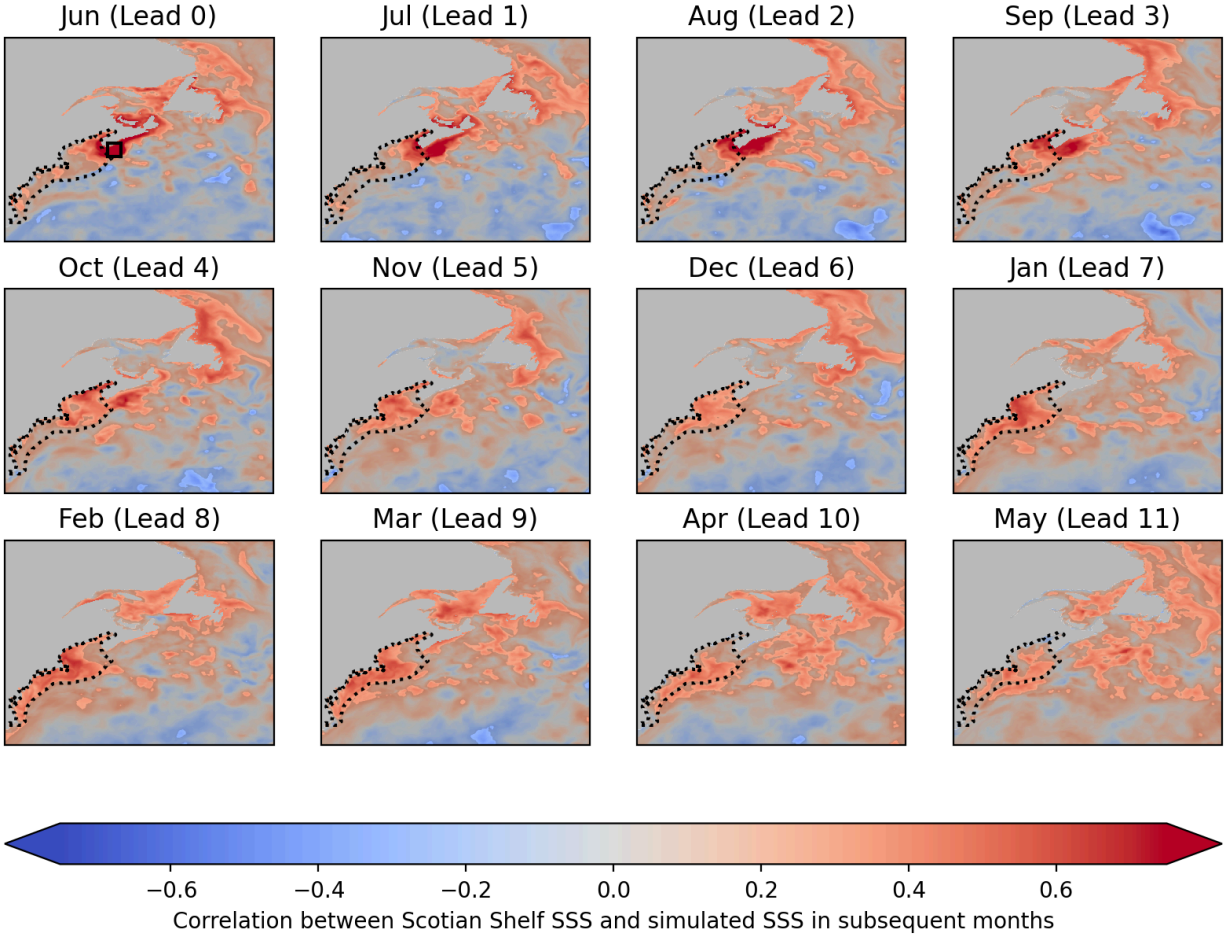
6. Line 340: "mid-Atlantic Bight" -> MAB

We appreciate this suggestion and have also replaced a few other instances of "Mid-Atlantic Bight" with MAB.

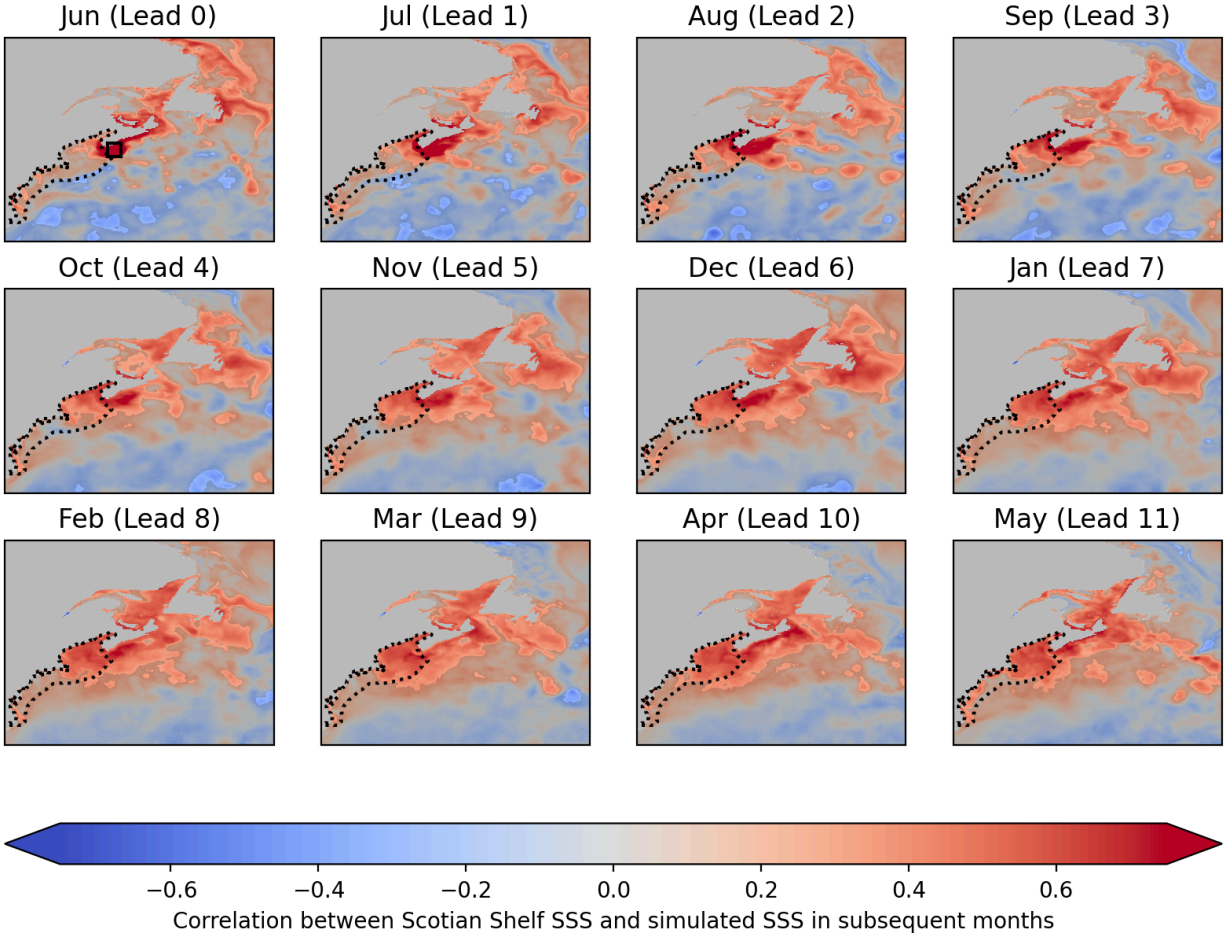
7. Figure 11: Please indicate correlation significance in the figure for each panel. Figure 11 shows the correlation in the GOM is minimum at Lead 6 but increases at Lead 7. Could you please explain it? Please consider adding the location of the Scotian Shelf box in Figure 1.
-and-
8. Figure 12-13: please consider adding correlation significance in each correlation map.

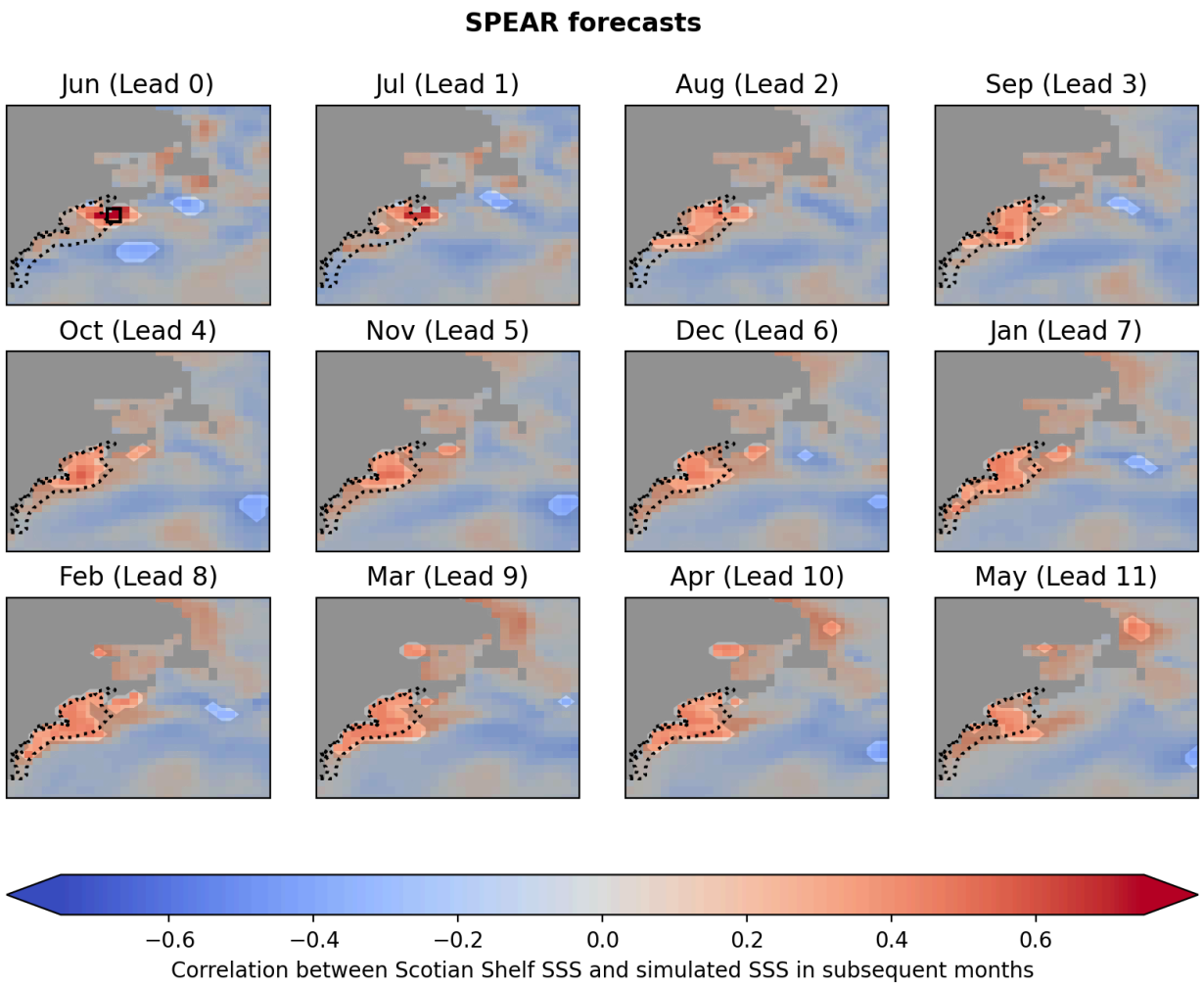
We have a semi transparent gray shading to the regions where the correlation is not significant at alpha = 0.1 in Figures 11, 12, and 13. A copy of these figures is included below. We have also added an outline of the Scotian Shelf box to the first panel of each figure.

Nudged analysis simulation



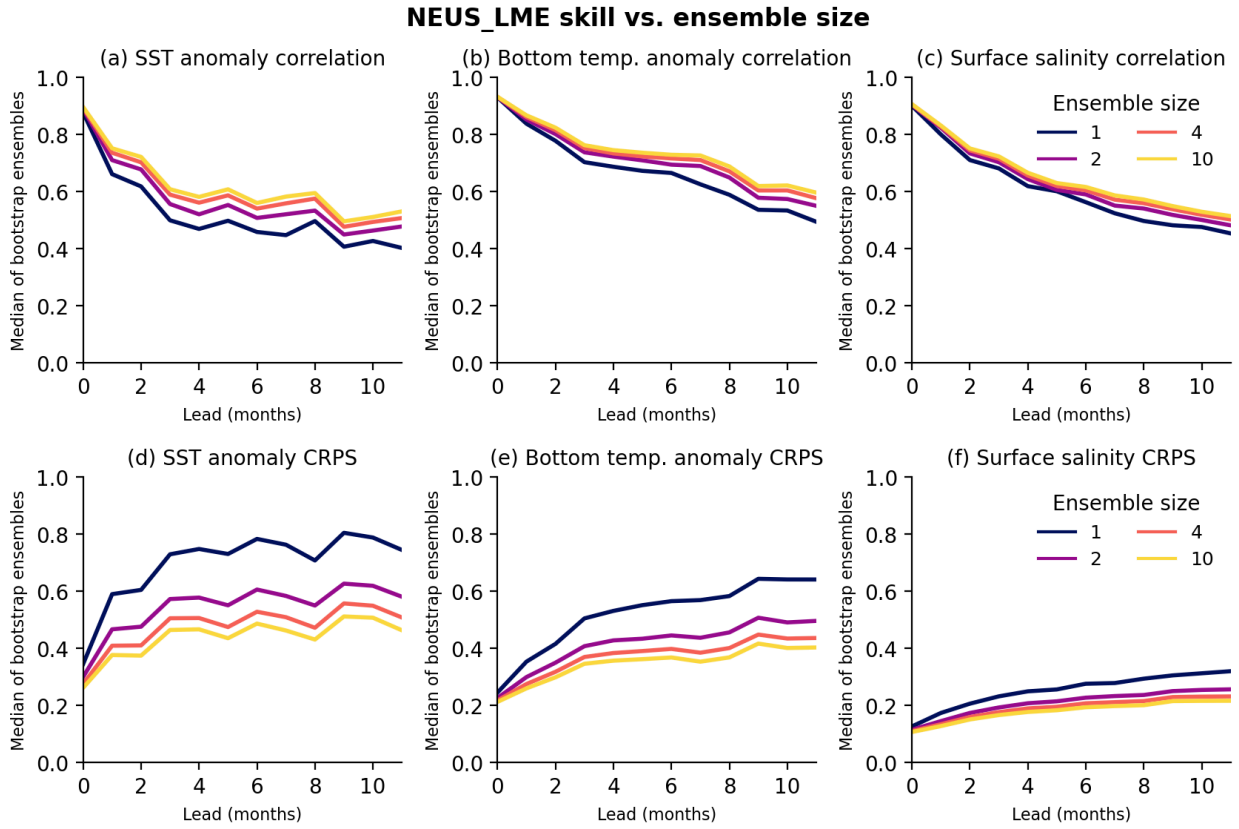
NWA12 downscaled forecasts





9. Figure 20: Do predictions of bottom temperature also require approximately 4 ensemble members to provide a reasonable compromise between computational costs and prediction skill?

Yes, bottom temperature and surface salinity have similar patterns of skill vs ensemble size. To show this, we have revised figure 20 to include panels for bottom temperature and surface salinity (figure also included below). We have also revised the text to mention that the effect of ensemble size is similar for all three variables.



Reviewer #3

This article discusses the skill of historical seasonal forecasts utilizing a regional North Atlantic ocean model initialized with the GLORYS12 ocean analysis (approximately the same resolution as the ocean model) and forced by atmospheric forecast conditions from the SPEAR seasonal forecast. Results are then compared with the global, low resolution, coupled forecasts of SPEAR.

I have always been an advocate for the usage of ocean reanalysis as a tool for both initialization of forecasts – and for their use as a diagnostic tool to assess the ocean in regions where observations are sparse, or non-existent. However, the usage here, to use the GLORYS12 product as initialization for the regional seasonal forecasts – and then to assess skill against the GLORYS12 reanalysis seems somewhat incestuous to me – especially since the SPEAR ocean analysis likely differs substantially from the GLORYS12 analysis The study then really becomes one of assessing the initialization of a high resolution ocean model with GLORYS12 versus initializing with the ocean model component of SPEAR, and not particularly a "downscaling" of a seasonal forecast.

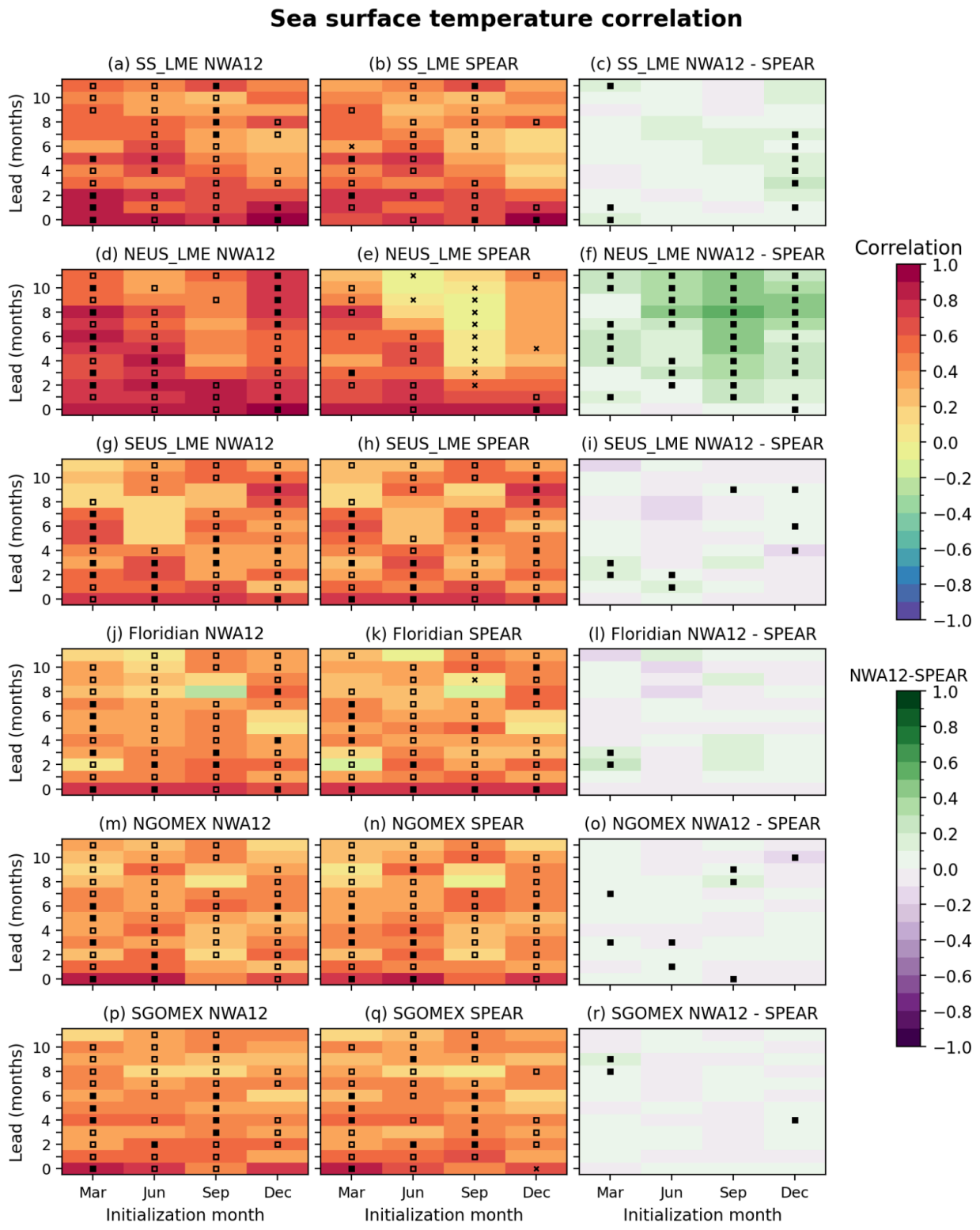
My question to the authors: If this system was to become an operational forecast system for the U.S. East Coast, would the goal be to initialize such forecasts with the real time

GLORYS12 analysis (Skill assessed in this manuscript), or initialized by downscaling the SPEAR ocean analysis (Skill not assessed in this manuscript). If it is the former, then perhaps utilizing the full multi-model ensemble of NMME, as opposed to only SPEAR atmospheric component forcing, would be a more prudent approach, as that likely would increase the underdispersiveness of only using SPEAR atmospheric forecast forcing.

Recommendation: Despite my trepidation with regards to the skill assessment primarily against the system initialization product, I would recommend publication after the authors answer my question and comments.

We appreciate the reviewer for raising these concerns with the methods and how the results are discussed. Ultimately, we believe that the methodology we employed is the best for examining the question at hand—whether a high resolution model can produce skillful seasonal predictions in the study region—and we plan to employ the same methodology (initializing with a simulation nudged towards the GLORYS12 reanalysis) for a quasi-operational product as part of NOAA's Climate, Ecosystems, and Fisheries Initiative. If the GLORYS reanalysis was a poor quality product with severe biases, we agree that using it for initialization and verification would give a biased assessment of prediction skill. However, if the GLORYS reanalysis was a perfect match for actual observations, aside from observation error, then using it for initialization and verification would be ideal. Since, as we cite in the manuscript, the GLORYS reanalysis has been found to match observations more closely than many other reanalysis, we argue that it is appropriate to use GLORYS for both initialization and verification.

To quantitatively explore this question, we compared the sea surface temperature forecasts from the MOM6-NWA12 and SPEAR models with the OISST v2 dataset instead of the GLORYS reanalysis. As we note in an answer below, the OISST dataset was assimilated by the SPEAR model, so this would potentially bias the skill assessment in favor of SPEAR. However, we see no meaningful difference whether OISST is used as the observations (figure below) or GLORYS is used as in the manuscript. In fact, at most lead times the skill of MOM6-NWA12 actually increases if OISST is used as the observations.



We also note that Jacox et al. 2023 (“Downscaled seasonal forecasts for the California Current System”, PLOS Climate) similarly compared their seasonal forecasts with the same reanalysis

that was used for initialization, and they likewise found it made little difference if the skill assessment was performed using the GLORYS or OISST products instead.

In terms of whether or not this is truly a study of downscaling because the initial conditions are not derived from the model being downscaled, we note Jacox et al. 2023 and Kearney et al. 2021 also initialized from a different source yet still referred to their forecasts as “downscaled”.

We have added two paragraphs to the Discussion in reply to this comment:

The analysis showed that the high resolution regional model had significantly higher forecast skill than the global model in many cases, and that this skill comes from several sources including better representation of re-emergence, advection of water masses, and Gulf Stream variability and trends. Given the experimental design used in this study, however, where the high resolution model uses initial conditions from a different, higher-resolution source, it is difficult to determine how much of the increased forecast skill comes from the higher resolution initial conditions and how much comes from evolving the initial conditions forward in time with higher resolution. In an analysis for the U.S. West Coast, Jacox et al. (2023) examined two sets of downscaled retrospective forecasts, one initialized from a high resolution reanalysis (similar to the present study) and the other initialized from the coarse resolution parent model. Initializing from the high resolution reanalysis yielded generally negligible improvements in forecast skill for surface and bottom temperature, aside from in the first month. The high resolution reanalysis did significantly improve the skill of sea surface heights in their analysis. However, it is worth noting that improved (bias-corrected) atmospheric forcing was also included in their model runs initialized from the high resolution product. Overall, additional experiments are needed to conclusively determine the role of the resolution of the initial conditions in downscaled seasonal forecast skill.

As we noted in Section 2.4, the skill assessment could have been biased in favor of the high resolution model, which was initialized with the same GLORYS reanalysis used as the observations in the assessment. On the other hand, the GLORYS reanalysis has been repeatedly found to closely match in situ observations (Amaya et al., 2023; Carolina Castillo-Trujillo et al., 2023), which would suggest that comparing against the GLORYS reanalysis should be similar to comparing against in situ observations. To determine whether any bias could be an issue, we repeated the assessment of forecast SST anomaly correlation using the OISST dataset (Reynolds et al., 2007) instead of the GLORYS reanalysis (Figure S1). Even though the SPEAR model derived its initial conditions by assimilating data from OISST, there is no meaningful difference between the forecast skill relative to GLORYS or OISST. In fact, in many cases the downscaled model has slightly higher prediction skill if OISST is used as the observations. A lack of sensitivity to the dataset used for verification was also found by Jacox et al. (2023) who downscaled seasonal forecasts for the U.S. West Coast.

Itemized Comments:

1. I believe other studies (not seasonal forecasts, however) have been undertaken with 1/12th degree North Atlantic systems, although admittedly I could not find a particularly relevant study in my quick search. Perhaps the authors could [be] more explicit with regards to the definition of their 1/12th grid: Is the grid identical to a North Atlantic

subset of the GLORYS12 grid, or how does it differ from the ORCA12 grid utilized by GLORYS?

We refer to Ross et al. 2023 at the beginning of the methods section for details about the model grid and configuration. To address the reviewer's comment, however, we have added a note in the revised version that the NWA12 model grid is a subset of the North and Equatorial Atlantic model grid from Chassignet and Xu (2017). The original model by Chassignet and coauthors is a HYCOM-based model, and the grid has no relation to the GLORYS/ORCA12 grid.

2. It is not the responsibility of the authors to discuss the ocean initialization of SPEAR, but nonetheless, how it is initialized, and in particular, how its ocean state estimation approach differs from GLORYS12 is an important component of this study. More information is required to assess this, preferably with some explicit text in the manuscripts, but minimally by explicit citations of the SPEAR ocean initialization procedure. The manuscripts does show "0 lead" (actually 0.5 lead I believe) results that can be used to assess these differences somewhat, but some more explicit comparisons, particularly for the reemergence discussion would be useful – for instance, the manuscript shows the reemergence in the GLORYS12 reanalysis – is it also present in the SPEAR ocean analysis (or concatenation of 0 leads).

We have added a sentence to the methods section: "Ocean initial conditions for the SPEAR retrospective forecasts were obtained by assimilating the OISSTv2 sea surface temperature product, vertical profiles from Argo floats, and several other sources using an Ensemble Adjustment Kalman Filter; see Lu et al. (2020) for details."

3. The statement in the conclusion, " Finally, full data assimilation, rather than nudging towards a reanalysis, could improve prediction skill through better initial conditions · · ·," might be true – but not necessarily when basing that skill on the reanalysis being nudged towards. This may be particularly true if not many observations are going into the ocean analysis in the areas of skill assessment, which unfortunately may be true for the coastal region, strong ocean current (short ARGO float retention) regions under study in this manuscript.

We believe we have addressed this comment with our reply to the beginning general comment.

4. The spread error discussion was interesting, and the skill versus ensemble size (including CRPS results) did expand on this. But I am always interested in expanding on the probabilistic nature of the ensemble – and it would seem the reemergence diagnostics utilized here might be a natural way to expound on this. I assume the reemergence diagnostics are perform on the ensemble mean? Could an member by member diagnostics be performed that might lead to a "probability" of re-emergence that could be accessed for skill (Brier Score)?

This is an interesting suggestion, but we believe that developing a probabilistic measure of re-emergence would be best suited as a topic for a future manuscript. The reviewer is correct that in the present manuscript we are using the ensemble mean for the reemergence diagnostic. We have not revised the text in response to this comment.

5. I remind the authors that Atlantic Overturning Circulation variability can be driven by atmospheric variability as well (Jackson et al, 2016; https://doi.org/10.1038/ngeo2715), with particular implications to density anomalies along western Atlantic.

We appreciate the reminder. We added a citation to this paper where we think it is most relevant: to support our remark that anomalies entering our model domain from the northern boundary typically take more than 1 year to reach the LMEs of interest in our study.

6. The authors should highlight their skill assessment of ocean currents is performed using an independent "observation" source, and therefore is not as sensitive to the initial conditions as their temperature and salinity skill assessment. Although it then may be instructive to give some evidence of current skill in the GLORYS12 analysis (Aijaz et al, 2023; https://doi.org/10.1016/j.ocemod.2023.102241)

Although technically the assessment of ocean currents is based on a satellite altimetry dataset rather than the GLORYS12 reanalysis, GLORYS assimilates the altimetry data and in practice the sea surface heights of the two datasets are similar. Furthermore, based on our comparison with the OISST dataset and the other sources we cited, we do not believe that in this study it makes a difference whether the dataset being used for evaluation is also used for initialization. We have not revised the text in response to this comment.

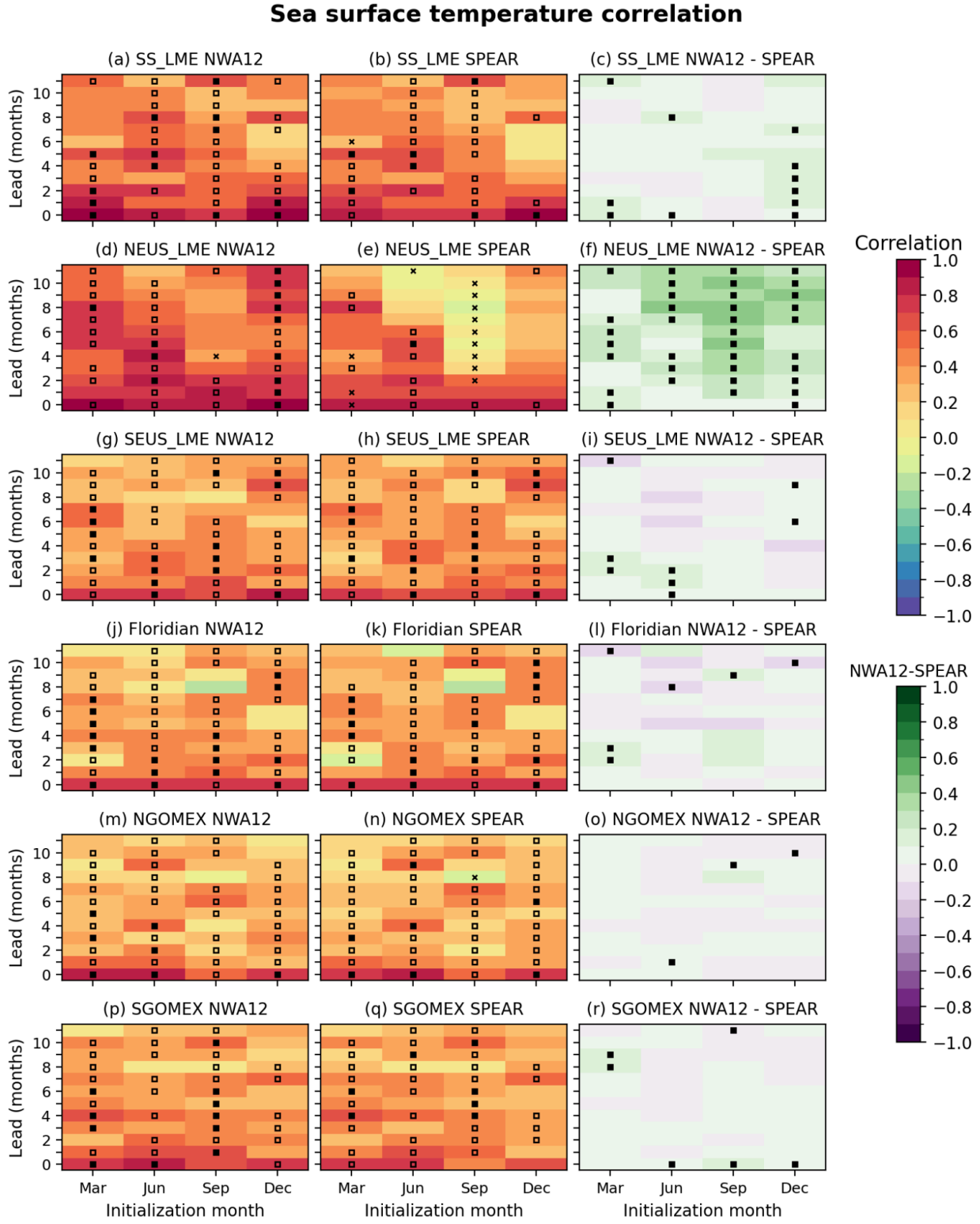
7. In light of the previous two points, I wonder why authors did not present a more detailed skill assessment of currents beyond just a trend analysis?

The present study is primarily focused on metrics that are readily translatable to fisheries management, such as surface and bottom temperature, and that also provide an overall picture of the model skill (for example, accurately predicting surface salinity requires accurately simulating advection and mixing). Ocean currents are not currently used by the marine resource managers we are working with. Comparison with the 1° global model is also potentially less interesting and relevant for ocean currents; for example, small coastal currents like the Eastern Maine Coastal Current *are* potentially relevant to fisheries and water quality, but they are obviously not represented in the 1° model. We believe an analysis of prediction skill for coastal currents in the high resolution model would be an interesting and useful subject for a separate study.

8. I found the divergenging color schemes used in plots to not be particularly easy to distinguish null results, particulary with the red/green scheme used in figures 1-4 (plus I believe it is not particularly colour blind friendly). The purple/green scheme of figures 5-7

seems somewhat better – but either an explicitly white color marker for 0 difference, or 3 colour scheme (i.e. yellow as zero difference) might be preferable.

We appreciate this suggestion to improve the accessibility of the paper. In the revised version, we have replaced the red/green color scheme with the same purple/green scheme used in figures 5–7. At the reviewer's suggestion, we also experimented with adding a pure white color to the middle of the colormap, but we did not think that this enhanced the readability of the figure. A sample revised figure is included below.


Sea surface temperature correlation
(a) SS_LME NWA12
(b) SS_LME SPEAR
(c) SS_LME NWA12 - SPEAR
(d) NEUS_LME NWA12
(e) NEUS_LME SPEAR
(f) NEUS_LME NWA12 - SPEAR
(g) SEUS_LME NWA12
(h) SEUS_LME SPEAR
(i) SEUS_LME NWA12 - SPEAR
(j) Floridian NWA12
(k) Floridian SPEAR
(l) Floridian NWA12 - SPEAR
(m) NGOMEX NWA12
(n) NGOMEX SPEAR
(o) NGOMEX NWA12 - SPEAR
(p) SGOMEX NWA12
(q) SGOMEX SPEAR
(r) SGOMEX NWA12 - SPEAR
Lead (months)
0
2
4
6
8
10
Mar
Jun
Sep
Dec
Initialization month
Correlation
NWA12-SPEAR
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
−1.0