Dear Editor,

Thank you for taking the time to review our revised manuscript and our responses to the reviewers' comments. We greatly appreciate your careful reading and the additional suggestions you provided. We have carefully considered all of your comments and have revised the manuscript accordingly. Please find our point-by-point responses below.

We appreciate your time and consideration and hope that the revised manuscript satisfactorily addresses all of the concerns.

Sincerely,
Yiming Guo
On behalf of all co-authors

Detailed comments
Lines 87-89. This might read better if "through comprehensive comparisons with available observations" were moved forward to follow ". . shown".
Revised.
Line 95. Better "its" –> "GLORYS'"?'
Revised.
Lines 109-112 ought to reflect the response to Ref 2 comment on lines 106-109 of the previous version.
We have revised our response to Reviewer 2's comment by indicating the relevant sections of the manuscript where this issue is addressed.
Line 113. Omit "simulations" or ". . ensemble of simulations . ."?
Revised.
Line 150. "which provides" –> "providing"?
Revised.
Line 153. Regarding Ref 2 comment on line 146 of the previous version. I agree, not "forecast" but maybe "simulation" –> "predictions"?
Revised.
Lines 320-326. Ref 1 asked for a "more quantitative comment on the representation of GS displacement in ROMS-DOWN." You have added descriptive text but no values. Could some quantification be added without "additional analyses"?
Thank you for this suggestion. We agree that a quantitative assessment of the GS displacement would be valuable. However, a rigorous quantification of the GS position would require additional analyses that are beyond the scope of the present manuscript. As shown in Fig. 2f–j, the excessive northward overshooting of the GS is evident in CESM-DPLE, whereas both ROMS simulations exhibit a substantially more realistic representation of the GS path. We have therefore retained the qualitative comparison in the manuscript.
Figure 25. The caption is consistent with the terminology elsewhere but the labels along the horizontal axis are not.
Figure 15 has been revised.
Line 562. "provide" –> "provides"

Revised.
Ref. 2 commented on line 545 of the previous version (line 578 in the revised manuscript). That line refers to SSH but your response discusses SSS. I think the referee really does mean SSH which is absent from Table 1 and hardly discussed in section 3 (only lines 249-251, 264-267 and inclusion in figures 2 and 3).
Thank you for pointing this out. Our previous response mistakenly referred to SSS instead of SSH. We have corrected this oversight and revised both our response and the manuscript accordingly. Specifically, we have revised both Section 3.3 (Results) and the Discussion. In the Results section, we now describe the relatively weak SSH prediction skill and note that, although ROMS-DOWN substantially improves the mean-state representation and temporal variability of SSH relative to CESM-DPLE, these improvements do not translate into enhanced SSH prediction skill. We also briefly note that this limited improvement likely reflects the difficulty of accurately predicting Gulf Stream path variability and associated mesoscale activity, while the detailed physical interpretation is provided in the Discussion. Please see our revised response to Reviewer 2's comment below.
Subject to the OS formatting guidance, I think that all figures A.n (with their captions) should be in sequence without intermediate "Acknowledgements" etc.
Revised.

REVIEWER COMMENTS:
RC1:
The authors dynamically downscale output from the CESM Decadal Prediction Large Ensemble (CESM-DPLE) using a high-resolution regional ocean model, ROMS-DOWN, with a horizontal resolution of ~5 km. The study focuses on the multiyear prediction skill of ROMS-DOWN on the U.S. Northeast Shelf, which complements previous work on seasonal and decadal prediction in this region. Using a set of metrics including the mean state and interannual variability of key physical variables, as well as the representation of important regional phenomena such as the Mid-Atlantic Bight cold pool and slope-water mixing, the authors demonstrate substantial improvements in ROMS-DOWN relative to the CESM-DPLE simulations.

Overall, the manuscript is well written, and the analyses are clear and comprehensive. I believe this study makes a valuable contribution to the literature on regional downscaling and ocean predictability. I have only a few minor comments as listed below.

Response: We thank the reviewer for the thorough review and the positive assessment of our manuscript. We have carefully considered all of the comments and revised the manuscript accordingly. Our point-by-point responses are provided below. Our responses are shown in red, and the corresponding revisions to the manuscript are shown in blue.

Minor comments:

1. Figure 2-3: Why does ROMS-DOWN sometimes perform better than ROMS-HIND? I expected the opposite since ROMS-HIND is forced by reanalysis atmospheric anomalies that are constrained by data. Could the authors explain?

Response: We agree that ROMS-HIND would be expected to provide the most accurate representation of the historical ocean state. In our results, however, the differences between ROMS-HIND and ROMS-DOWN are generally small, and both configurations perform substantially better than CESM-DPLE. The slight advantages of ROMS-DOWN for some variables or regions should not be interpreted as indicating that it systematically outperforms the hindcast. Rather, they likely reflect the combined effects of internal variability over the relatively short evaluation period (2004–2010), model uncertainties, and differences in the forcing strategy. In particular, Figure 2 compares only the lead-year-1 ROMS-DOWN forecasts, which are reinitialized each year, with a continuously integrated ROMS-HIND simulation. The annual reinitialization may reduce the accumulation of model errors and contribute to the small differences between the two simulations. Overall, the two ROMS simulations exhibit comparable performance, and the differences between them are much smaller than the improvement of either ROMS configuration relative to CESM-DPLE.

We have added a discussion of this point in Lines 280–285: "Although ROMS-HIND is expected to provide the most accurate reconstruction, the differences between ROMS-HIND and ROMS-DOWN are generally small. Occasional lower RMSE values in ROMS-DOWN likely result from its anomaly-downscaling framework, which used the GLORYS climatological mean state, together with model uncertainties and internal variability over the relatively short evaluation period. In addition, the comparison is based on only the lead-year-1 ROMS-DOWN forecasts, which are reinitialized annually, whereas ROMS-HIND is a continuously integrated simulation."

2. Line 236: How is NES defined exactly? Is it a combination of GOM, GB, and MAB?
Response: Thank you for pointing this out. In this study, the NES refers to the entire U.S. Northeast Shelf analysis region, which consists of GOM, GB, and MAB. We agree that this definition should be stated more explicitly and have clarified it in the revised manuscript. In Lines 182-184: "For regional skill assessment (Section 2.5), metrics are aggregated over the entire NES, defined here as the combined shelf region encompassing the Gulf of Maine (GOM), Georges Bank (GB), and the Middle Atlantic Bight (MAB), as well as over each subregion individually." And we also revised the caption of Figure 1 to clarify this further: "Figure 1. Study area. (a) ROMS model domain showing bathymetry. The white box denotes the analysis region shown in panel (b). (b) Bathymetry within the analysis region. The U.S. Northeast Shelf (NES) is defined as the shelf region shallower than 200 m and consists of the Middle Atlantic Bight (MAB), Georges Bank (GB), and the Gulf of Maine (GOM), outlined in red. (c) Same as panel (b), but for the CESM bathymetry."

3. Lines 395 and 494: Have the authors looked at the variability of the Gulf Stream position (e.g., the GS index; Ross et al. 2023) or its proxy in ROMS-DOWN? Previous studies have shown that GOM temperature is strongly correlated with the meridional shifts of the GS (Pershing et al., 2015). Therefore, the high prediction skill for NES SST may partly reflect a realistic simulation of GS displacement, as the authors note. I do not think additional analyses are necessary, but a more quantitative comment on the representation of GS displacement in ROMS-DOWN would be helpful.

Response: We thank the reviewer for this insightful comment. We agree that the improved representation of Gulf Stream variability likely contributes to the enhanced SST prediction skill in ROMS-DOWN. Although we did not explicitly calculate a Gulf Stream position index in this study, the improved representation of the Gulf Stream mean position in ROMS-DOWN relative to CESM-DPLE is evident in Figure 2. In addition, previous studies have demonstrated that meridional shifts of the Gulf Stream are strongly linked to temperature variability over the U.S. Northeast Shelf, particularly in the Gulf of Maine (e.g., Pershing et al., 2015; Ross et al., 2023). We have expanded the discussion to explicitly acknowledge this connection and note that the improved representation of the Gulf Stream position and shelf–slope exchange in ROMS-DOWN likely contributes to the enhanced SST prediction skill. A detailed investigation of Gulf Stream path predictability in the regional prediction system is beyond the scope of the present study and will be the subject of future work. We have expanded the discussion in the revised manuscript (Lines 320–326): "However, the SST time series (Figure 4) also reveals that, even when the overall correlation or variance agreement with GLORYS is comparable, the higher resolution and improved representation of the Gulf Stream and fine-scale processes in ROMS-DOWN enable it to track warm and cool extremes more skillfully. Previous studies have shown that meridional shifts of the Gulf Stream strongly influence temperature variability over the NES, particularly in the Gulf of Maine (Pershing et al., 2015; Ross et al., 2023). Therefore, the improved representation of the Gulf Stream mean position and associated shelf–slope exchange in ROMS-DOWN likely contributes to the enhanced SST prediction skill."

Technical corrections:

Figure 12: label (h) is repeated. Please revise the plot.

Revised.

Caption of Figure 14: "Shelf Water (SSW)" -> "Scotian Shelf Water (SSW)"

Revised.

Figure 14: "label (e) Lead Year 8" -> "label (h) …"

Revised.

RC2:
This paper provides a detailed analysis of the performance of a downscaled predicition system for the American Northeast Shelf (NES) region, over the challenging sub-decadal timescale. The paper is well written, and the results will be useful for a range of reader – both model developers and potential users of model output.

However, the paper would benefit from some revisions that would help improve clarity of results for the reader. Firstly, there are a large number of figures that can be hard to compare at times, so some thought on use of tables for key statistics could be useful. I also feel more could be done to clarify whether the model is intended to benefit on-shelf or off-shelf predictions. Some

discussion on this for both the motivation and conclusion could help target the analysis and conclusions.

Response: We thank the reviewer for the careful and thoughtful review of our manuscript. We also thank the reviewer for the constructive suggestions to improve the clarity and presentation of the manuscript. We have carefully considered all comments and revised the manuscript accordingly. Our detailed responses are provided below in red, and the corresponding revisions to the manuscript are shown in blue.

General comments:

1. The paper has a very large number of figures and subpanels. It could be useful to summarise some statistics into a table, to aid comparison between different models and metrics used for analysis. For example, providing tables for each region (or NES), that summarise the different statistics (e.g., ACC, MSSS, etc), at different lead times, for the different models (perhaps replacing figures 6-8)? This could help avoiding comparison between different figures and subpanels, and make the results clearer.

   Response: We thank the reviewer for this helpful suggestion. We agree that a concise summary of the prediction skill is useful. Rather than replacing Figures 6–8, which illustrate the evolution of skill with lead time, we have added a summary table (Table 1) reporting the mean skill over the two forecast periods (LY2–5 and LY6–8) for the entire NES. This table complements the figures by providing a compact quantitative comparison while retaining the temporal evolution shown in the figures. We have also cited Table 1 in the Results and Summary sections to emphasize the key comparisons between ROMS-DOWN and CESM-DPLE.

   Lines 346-350: "Building on the SST prediction skill assessment, we next examine whether dynamical downscaling improves prediction skill for other key NES water properties, including SSS, T_{bottom}, and HC_{200m}. Figure 8 shows the differences in ACC, MSSS, and BrS between ROMS-DOWN and CESM-DPLE for these variables, summarized across the entire shelf and its subregions as a function of lead year. To facilitate comparison among variables and skill metrics, Table 1 summarizes the mean prediction skill over the NES during the early (LY2–5) and late (LY6–8) forecast periods."

   Lines 383-384: "As summarized in Table 1, ROMS-DOWN generally provides higher deterministic (ACC and MSSS) and probabilistic (BrS) skill during LY2-5, whereas the differences become smaller for several metrics during LY6-8."

   Lines 629-631: "Both deterministic and probabilistic metrics demonstrate that ROMS-DOWN improves forecast skill, particularly during the first five forecast years (Table 1 and Figure 8). At longer leads (LY6–LY8), CESM-DPLE retains comparable skill due to the dominance of low-frequency, externally forced variability."

In addition, following the reviewer's suggestions, we have improved the presentation of the figures by adding more descriptive panel titles, expanding the figure captions, using consistent colormaps, and adding clearer labels where appropriate to make the figures easier to interpret. We have also revised the text to better summarize the key differences among models and metrics. Please also refer to our responses to the specific comments below for details of these revisions.

2. Do not use jet or rainbow for any figures. Use perceptually uniform colormaps, e.g., sticking to Red-Blue for all anomalies, and other options for sequential data (e.g., mean temperature).

   Response: We thank the reviewer for this suggestion. Accordingly, we have replaced the rainbow/jet colormaps with perceptually uniform alternatives throughout the manuscript (i.e. Figures 2, 5, and A8).

3. Through the text, please clarify whether you use NES to refer to whole model domain, or just the shelf region. If you are more interested in model skill for former or latter, suggest clarifying this in both the introduction and summary.

   Response: We thank the reviewer for pointing out this ambiguity. Throughout the revised manuscript, we have clarified that the ROMS model domain extends beyond the U.S. Northeast Shelf to include the adjacent slope sea and Gulf Stream. However, unless otherwise stated, NES refers specifically to the U.S. Northeast Shelf analysis region, defined as the combination of Gulf of Maine (GOM), Georges Bank (GB), and Middle Atlantic Bight (MAB). We have clarified this in the Introduction, Methods, and Figure 1 caption.

   Introduction (Lines 65-68): "In this study, we evaluate the improvement in multi-year prediction skill over the NES using a high-resolution Regional Ocean Modeling System (ROMS) configuration (hereafter ROMS-DOWN) forced by large-scale anomaly fields from CESM-DPLE (Yeager et al., 2018). Although the ROMS model domain extends beyond the shelf to include the adjacent Slope Sea and Gulf Stream, our evaluation focuses primarily on prediction skill over the NES".

   Method (Lines 182-184): "For regional skill assessment (Section 2.5), metrics are aggregated over the entire NES, defined here as the combined shelf region encompassing the Gulf of Maine (GOM), Georges Bank (GB), and the Middle Atlantic Bight (MAB), as well as over each subregion individually."

   Figure 1 Caption: "Figure 1. Study area. (a) ROMS model domain showing bathymetry. The white box denotes the analysis region shown in panel (b). (b) Bathymetry within the analysis region. The U.S. Northeast Shelf (NES) is defined as the shelf region shallower than 200 m and consists of the Middle Atlantic Bight (MAB), Georges Bank (GB), and the Gulf of Maine (GOM), outlined in red. (c) Same as panel (b), but for the CESM bathymetry."

Technical comments:

Section 2: Make sure you provide sufficient references to all CMEMS products – include reference dataset ID, DOI, and date of access. (Following recommended citation format). This is provided for some datasets, but not all.

Response: We thank the reviewer for pointing this out. We have carefully reviewed all dataset citations in Section 2 and updated them, where necessary, to ensure they follow the recommended CMEMS citation format.

L83: Clarify whether model levels, or post-processed product from CMEMS? (Or are these the same?)

Response: We thank the reviewer for pointing this out. We have clarified that the GLORYS data used in this study are obtained from the CMEMS final product, which is distributed on a regular 1/12° grid with 50 standard vertical levels. This clarification has been revised to the manuscript in lines 82-85 "To evaluate multi-year prediction skill, we employ the GLORYS ocean reanalysis (Global Ocean Reanalysis and Simulation) version 12V1, produced by Mercator Ocean and distributed through the CMEMS (Jean-Michel et al., 2021). The CMEMS product is provided on a regular 1/12°(~8 km in the study area) grid with 50 standard vertical levels and spans the satellite altimetry era from 1993 to the present (Lellouche et al., 2021).".

L92: Why did you choose annual-mean rather than seasonal analysis? Please include further discussion on what impact this might have on results.

Response: We thank the reviewer for this comment. We chose to evaluate annual-mean anomalies because the primary objective of this study is to assess prediction skill on multi-year (interannual-to-decadal) timescales. Annual averaging emphasizes the lower-frequency variability targeted by the prediction system while reducing the influence of seasonal fluctuations. In addition, CESM-DPLE is initialized only once per year (each November) and is designed primarily for annual-to-decadal prediction rather than seasonal forecasting. We agree that prediction skill may vary by season; however, a seasonal analysis is beyond both the design of the prediction experiments and the scope of the present study. We have revised the manuscript to clarify the rationale for using annual means and discuss the implications of this choice in Lines 189-193: "To evaluate prediction skill on multi-year timescales, we focus on annual-mean anomalies which emphasizes interannual-to-decadal variability by reducing the influence of seasonal fluctuations. This choice is consistent with the intended application of CESM-DPLE as an annual-to-decadal prediction system. While seasonal analyses may provide additional insight into season-dependent sources of predictability, such analyses are beyond the scope of the present study and will be the subject of future work."

L92-93: The removal of climatology should be explained in more detail at the first point of introduction.

Response: We have expanded the description in Section 2.5 to explain that a lead-year-dependent climatology is removed from both the model predictions and observations to account for forecast

drift and to isolate the predictable interannual-to-decadal variability used in the skill assessment. In Lines 202-205: "To evaluate prediction skill, we remove a lead-year-dependent climatology from both the model predictions and the observations. This climatology is computed separately for each prediction lead year by averaging across all initialization years. Removing a lead-year-dependent climatology accounts for systematic forecast drift that develops with increasing lead time and isolates the interannual-to-decadal anomalies used to calculate the skill metrics (Yeager et al., 2018, 2023)."

L106-109: Why the difference between use of products here?

Response: The different initialization datasets are part of the original CESM-DPLE experimental design rather than a choice made in this study, which is introduced in Yeager et al. (2018). In CESM-DPLE, the ocean and sea ice are initialized from a CORE-forced ocean–sea ice simulation to provide an observationally constrained ocean state, while the atmosphere and land are initialized from CESM-LE because their initial conditions have relatively short memory compared with the ocean. This is indicated in Lines 109-112: "Initialization of the ocean and sea ice is based on an ocean–sea ice-only simulation forced with Coordinated Ocean Research Experiments (CORE) bulk fluxes (Helmholtz Ctr110 Ocean Res, 2016), while the atmosphere and land components are initialized from the CESM Large Ensemble (CESM-LE; Kay et al. (2015)). The external forcing is prescribed to be identical to CESM-LE."

L119-121: What about frequency of ocean lateral boundary forcing?

Response: The global CESM-DPLE does not have ocean lateral boundaries. The ROMS-DOWN uses a monthly lateral boundary from a combination of CESM-DPLE and GLORYS as indicated in lines 159-161: "Similarly, the ocean initial and lateral boundary conditions are constructed from the 1993–2021 GLORYS monthly climatology plus monthly CESM-DPLE ocean anomaly fields computed relative to the 1993–2010 monthly climatology."

L126: Curvilinear rather than variable?

Revised. Lines 131-132: "The model has a curvilinear grid with a ~5 km spatial resolution and 40 terrain-following vertical levels that are unevenly spaced with a higher resolution near the surface."

L128: For setting minimum depth, is this modifying the bathymetry or the free surface code?

Response: We thank the reviewer for this comment. The minimum depth was implemented by modifying the bathymetry. We have clarified this in the revised manuscript in lines 132-133: "A minimum depth of 10 m was imposed by modifying the bathymetry prior to the simulation to avoid wetting and drying of grid cells."

L131: Remove first sentence ("realistic" is subjective) – just confirm choice of bathymetry product and its resolution.

Revised.

L144: River data is from what data product – e.g., model or gauges?

Response: We thank the reviewer for pointing this out. We have clarified that the river forcing is derived from the global river discharge dataset developed by Dai (2021), which combines gauge observations with a hydrological modeling approach. This has been clarified in the revised manuscript in lines 148-149: "The model is forced with monthly climatological river discharge and temperature at 21 river mouths along the NES, derived from the global river reanalysis dataset of Dai (2021)."

L145: Why only M2 and S2?

Response: We thank the reviewer for this question. We included only the M2 and S2 tidal constituents because semidiurnal tides dominate the tidal energy over the U.S. Northeast Shelf, and previous studies have shown that including the dominant semidiurnal tides provides the tidal mixing necessary to realistically represent the regional circulation and hydrography (Chen and He, 2015). Since our analysis focuses on annual-mean anomalies and multi-year prediction skill, the contribution of the smaller tidal constituents is expected to have only a limited impact on the results. We have clarified these points in the revised manuscript in lines 149-151: "Tidal forcing is included in the model using the dominant semidiurnal tidal constituents (M2 and S2) from the TPXO09 global tidal model (Egbert and Erofeeva, 2002), which provides the primary tidal mixing necessary to realistically represent the circulation over the Shelf (Chen and He, 2015)."

L146: Clarify – ROMS-DOWN *forecast* simulations.

Response: ROMS-DOWN refers to the forecast (prediction) simulations, in contrast to the ROMS-HIND hindcast simulation.

Section 2.4: Please clarify why different time periods used for each climatology. Also, why are monthly anomalies used rather than daily for ocean boundaries?

Response: We thank the reviewer for this comment. The climatology periods differ because they are determined by the temporal coverage of the respective datasets. Specifically, the ERA5 and GLORYS climatologies are computed using all available years over the common analysis period (1993 onward), whereas the CESM-DPLE anomalies are referenced to the common 1993–2010 baseline to ensure consistency in the anomaly-downscaling framework. Monthly ocean anomalies are used because the CESM-DPLE ocean variables required to force the regional model are archived only as monthly means. This temporal resolution is also appropriate for the present study because the large-scale ocean variability relevant to multi-year prediction evolves primarily on monthly to longer timescales. We have clarified these points in the revised manuscript. We have clarified these points in the revised manuscript in lines 161-165: "The different climatology periods reflect the temporal coverage of the respective datasets. Monthly ocean anomalies are used because the CESM-DPLE ocean variables required to force the regional model are available only as monthly means. This temporal resolution is also appropriate because the large-scale ocean variability relevant to multi-year prediction evolves primarily on monthly to longer timescales, while also reducing computational and storage requirements."

L154: Is the “mean” referring to monthly, or otherwise?

Response: We have revised the manuscript to clarify how the CESM-DPLE anomalies are constructed. Specifically, the daily atmospheric anomalies are computed relative to a daily climatology obtained by interpolating the 1993–2010 monthly climatology, whereas the ocean anomalies are computed relative to the 1993–2010 monthly climatology because the CESM-DPLE ocean output is available only as monthly means. In the revised manuscript, we made changes to lines 157-163: “The atmospheric forcing is constructed as the 1993–2024 ERA5 monthly climatology plus daily CESM-DPLE anomaly fields, where the daily anomalies are computed relative to a daily climatology obtained by interpolating the 1993–2010 monthly climatology. Similarly, the ocean initial and lateral boundary conditions are constructed from the 1993–2021 GLORYS monthly climatology plus monthly CESM-DPLE ocean anomaly fields computed relative to the 1993–2010 monthly climatology. The different climatology periods reflect the temporal coverage of the respective datasets. Monthly ocean anomalies are used because the CESM-DPLE ocean variables required to force the regional model are available only as monthly means.”

L159-161: Move this note on northern boundary location to support decision on domain size in previous section.

Revised.

L173-175: By NES, clarify whether mean just shelf, or whole model domain.

Response: Thank you for the comment. Please refer to General Comment 3 for our response and revision.

L179-180: Are there no observations you could use for further comparison of surface or subsurface, e.g., SST, SSH, moorings, tide gauges? Please provide more justification and discussion on your choice.

Response: We thank the reviewer for this helpful comment. While direct observations are available for SST and SSH, comparable domain-wide observational datasets do not exist for subsurface variables over the full study period. We therefore use GLORYS as a common reference dataset because it assimilates a broad range of observations, including satellite SST and SSH as well as in situ temperature and salinity profiles, and has been shown to provide one of the most accurate representations of the Northeast Shelf hydrography among currently available global ocean reanalyses (Castillo-Trujillo et al., 2023). This has been justified in lines 87-90: “Independent evaluations have shown that GLORYS is one of the best global reanalysis products in representing circulation features and water mass properties in the NES region (Castillo-Trujillo et al., 2023). These characteristics make GLORYS an appropriate benchmark for assessing the predictive skill of both the regional and global prediction systems analyzed in this study.” Its limitation is also discussed in lines 602-611: “An additional source of uncertainty relates to the use of GLORYS as the verification benchmark. GLORYS does not explicitly include tidal forcing, and its representation of stratification and mixing over strongly tidal regions—such as GB and parts of the GoM—may therefore differ from ROMS simulations that

include tidal harmonics. These structural differences could influence the magnitude of skill metrics, particularly for subsurface properties. However, our evaluation focuses on annual-mean anomalies with lead-time-dependent climatology removed, which reduces sensitivity to mean-state biases and high-frequency tidal variability. While the absence of tidal forcing in GLORYS is unlikely to fundamentally alter the conclusions regarding the added value of dynamical downscaling for multi-year anomaly predictability, it may contribute to differences in T_{bottom} and HC_{200m}, particularly in strongly tidal regions such as GB and parts of GOM, through its influence on vertical mixing and heat exchange. Consequently, part of the skill differences for these variables may reflect differences in the representation of tidal mixing between GLORYS and ROMS."

L185: Clarify what LY1-LY8 refer to, e.g., for LY1, what start date used to predict 1997?

Response: We thank the reviewer for pointing this out. We have revised the manuscript to clarify the relationship between initialization year, lead year, and calendar year, and included an example for illustration in lines 198-201: "Here, LY1–LY8 denote the first through eighth forecast years following initialization. For example, a forecast initialized on 1 January 1996 has LY1 = 1996, LY2 = 1997, ..., and LY8 = 2003. Skill at each lead year is evaluated by pooling all forecasts corresponding to the same lead year across the 20 initialization years."

L219-221: It would be more beneficial to include the anomaly plots within the paper rather than mean climatology. Also, for SST and SSH, why only comparing with GLORYS, and not direct to obs? NB. GLORYS will use data assimilation – clarify whether same products used for comparison? Where are the anomaly maps for GLORYS vs Obs?

Response: We thank the reviewer for these helpful suggestions. We have chosen to retain the mean climatological fields in Figure 2 because the primary purpose of this figure is to evaluate the mean-state fidelity of the regional and global models. The prediction of annual anomalies is assessed separately throughout the manuscript. Specifically, Figure 3 evaluates the temporal RMSE of annual anomalies, while the subsequent analyses use anomaly-based skill metrics (ACC, MSSS, and Brier Score) to quantify prediction skill. For SST and SSH, GLORYS is first evaluated against the observational products (OISST and AVISO), while ROMS-HIND, ROMS-DOWN, and CESM-DPLE are evaluated relative to GLORYS to provide a consistent benchmark across all variables, including SSS, bottom temperature, and HC_{200m}, for which comparable basin-wide observations are unavailable. A comprehensive evaluation of GLORYS against observations over the U.S. Northeast Shelf has been presented by Castillo-Trujillo et al. (2023), and we refer interested readers to that study for a more detailed assessment of the reanalysis performance in Lines 87-89: "Independent evaluations have shown that GLORYS is one of the best global reanalysis products in representing the climatological hydrography, circulation, and their temporal and spatial variability over the NES region through comprehensive comparisons with available observations (Castillo-Trujillo et al., 2023)". Therefore, we believe that including additional anomaly maps would largely duplicate the information already provided by Figure 3, the anomaly-based skill metrics, and the existing evaluation of GLORYS against observations.

L230-231: Again, is the purpose of this model to reproduce shelf or off-shelf conditions? If shelf, should you focus on performance here, rather than whole domain?

Response: We thank the reviewer for this comment. The primary objective of this study is to evaluate prediction skill over the NES (shelf region). The spatial maps over the full model domain are included to provide regional context and illustrate the distribution of model errors, while the quantitative evaluation and discussion focus on the shelf region. Specifically, the regional metrics presented in Figures 3, 6–8 and throughout the manuscript are computed over the NES and its three subregions (GOM, GB, and MAB). We have clarified this distinction in the revised manuscript at the following locations:

Lines 67-68: "Although the ROMS model domain extends beyond the shelf to include the adjacent Slope Sea and Gulf Stream, our evaluation focuses primarily on prediction skill over the NES."

Lines 182-184: "For regional skill assessment (Section 2.5), metrics are aggregated over the entire NES, defined here as the combined shelf region encompassing the Gulf of Maine (GOM), Georges Bank (GB), and the Middle Atlantic Bight (MAB), as well as over each subregion individually."

Figure 1 caption: "Figure 1. Study area. (a) ROMS model domain showing bathymetry. The white box denotes the analysis region shown in panel (b). (b) Bathymetry within the analysis region. The U.S. Northeast Shelf (NES) is defined as the shelf region shallower than 200 m and consists of the Middle Atlantic Bight (MAB), Georges Bank (GB), and the Gulf of Maine (GOM), outlined in red. (c) Same as panel (b), but for the CESM bathymetry."

L236-237 & Figure 3: I don't understand what this figure shows - is it intergrated or mean RMSE?

Response: The values shown in Figure 3 are spatially averaged temporal RMSEs, rather than integrated RMSEs. Specifically, the temporal RMSE is first calculated at each grid point using the annual anomalies relative to GLORYS, and the resulting RMSE values are then spatially averaged over the NES and each subregion. We have clarified this in the revised manuscript in lines 259-262: "To further quantify errors in temporal variability, the temporal RMSE is first calculated at each grid point using annual anomalies relative to GLORYS and then spatially averaged over the NES and each subregion. The resulting regional-mean RMSE values are shown in Figure 3, while the corresponding spatial distributions of grid-point RMSE are presented in Figure A4." and in the caption of Figure 3: "Spatial average of the temporal RMSE (2004–2010; relative to GLORYS) calculated at each grid point within each shelf subregion for…".

L263: but does have a comparable or higher r for some lead times (as discussed later)?

Response: We thank the reviewer for pointing this out. We have revised the text to clarify that the advantage of ROMS-DOWN is most evident at the shorter lead times, whereas CESM-DPLE exhibits comparable or even higher correlation at some longer lead times because of its representation of externally forced low-frequency variability. Lines 291-293: "ROMS-DOWN generally tracks the GLORYS anomalies more closely across all lead years, whereas CESM-

DPLE tends to exhibit weaker amplitude and less agreement in the timing of extremes, particularly at the shorter lead times."

L299-303: Could some of these summary scores be shown in a table instead, to help comparison between the different configurations and statistics?

Response: We thank the reviewer for this suggestion. As discussed in our response to General Comment 1, we considered summarizing these statistics in tables. However, because the figures are intended to illustrate the evolution of skill across multiple lead years, variables, regions, and skill metrics, we believe that graphical presentation more effectively conveys these patterns than large tables. Instead, we have improved the readability of the figures by adding more descriptive subpanel titles, revising the figure captions, and adopting a consistent colormap. We have also added a summary table of the mean prediction skill over LY2–5 and LY6–8 for the NES, which provides a concise quantitative comparison while complementing the figures.

L312: "Improved" reveals results, suggest delete or rephrase: Performance of ...

Revised. Now the section title is "Predictability of NES Water Properties"

L328: Suggests improved, but comparing with GLORYS, which doesn't have tidal mixing?

Response: It is true that GLORYS does not explicitly include tidal forcing, which may influence the comparison, particularly in strongly tidal regions such as Georges Bank. We have therefore revised the wording to avoid implying an unequivocal improvement and instead describe the results more objectively. We also refer readers to the discussion in Section 4, where the potential impact of the absence of tides in GLORYS is discussed in more detail. Lines 602-611: "An additional source of uncertainty relates to the use of GLORYS as the verification benchmark. GLORYS does not explicitly include tidal forcing, and its representation of stratification and mixing over strongly tidal regions—such as GB and parts of the GOM—may therefore differ from ROMS simulations that include tidal harmonics. These structural differences could influence the magnitude of skill metrics, particularly for subsurface properties. However, our evaluation focuses on annual-mean anomalies with lead-time-dependent climatology removed, which reduces sensitivity to mean-state biases and high-frequency tidal variability. While the absence of tidal forcing in GLORYS is unlikely to fundamentally alter the conclusions regarding the added value of dynamical downscaling for multi-year anomaly predictability, it may contribute to differences in T_{bottom} and HC_{200m}, particularly in strongly tidal regions such as GB and parts of GOM, through its influence on vertical mixing and heat exchange. Consequently, part of the skill differences for these variables may reflect differences in the representation of tidal mixing between GLORYS and ROMS."

Sections 4.1-4.2: These are results, suggest add separate results section to focus on regional applications/case studies?

Response: Revised. These two sections have been moved under Results section 3.5-3.6.

L541: Rephrase: how basin-scale anomalies impact the NES. (*shelf* is within the acronym)

Revised.

L545: This result appears quite stark and surprising that only introduced in detail here?

Response: Thank you for this helpful comment. We agree that the relatively weak SSH prediction skill should be introduced in the Results section before being discussed in the Discussion. To address this concern, we have revised Section 3.3 by adding a paragraph describing the SSH prediction results in lines 350-355 "Unlike the other variables, SSH exhibits generally weak prediction skill in both ROMS-DOWN and CESM-DPLE throughout the forecast period (Figure A8). Although ROMS-DOWN substantially improves the mean-state representation and temporal variability of SSH relative to CESM-DPLE (Figures 2 and 3), these improvements do not translate into enhanced SSH prediction skill. This limited improvement likely reflects the difficulty of accurately predicting offshore Gulf Stream path variability and associated mesoscale activities, which strongly influence SSH variability over the shelf (Gawarkiewicz et al., 2012; Guo et al., 2023)." The Discussion section has been revised accordingly to focus on the physical interpretation of the low SSH predictability in lines 583-590 "First, as shown in Figure A8 and described in Section 3.3, SSH exhibits generally weak prediction skill in both ROMS-DOWN and CESM-DPLE. On the shelf, SSH variability is strongly influenced by offshore Gulf Stream fluctuations and associated mesoscale activity (Böhm et al., 2006; Ezer, 2016; Guo et al., 2023), which are inherently chaotic and difficult to predict deterministically, particularly in the absence of data assimilation, even with a dynamically downscaled regional model. In addition, uncertainties in atmospheric forcing–particularly wind stress variability–further limit predictability on multiyear time scales. The limited ability of ROMS-DOWN to fully reproduce the variability of complicated shelf circulation patterns and offshore Gulf Stream meandering may contribute to the low SSH skill and highlights the persistence of internal noise in 590 the regional system."

L556-557: More clarity of this would be useful in the methods section.

Response: We have added an explanation in the Methods section clarifying this. Lines 210-214: "Because the forecasts are initialized over a finite 20-year period, the verification time windows differ among lead years. For example, LY1 spans 1992–2011, whereas LY8 spans 1999–2018. Consequently, differences in prediction skill among lead years may reflect both dependence on the forecast lead time and differences in the underlying verification periods. The implications of this sampling are discussed further in Section 4."

L558: Not sure I follow this reasoning, please clarify.

Response: We have revised the text to explicitly explain that the verification periods differ among lead years and that the stronger long-term warming trend in the later verification periods may enhance prediction skill at longer lead times. This clarification makes it clear that part of the lead-time dependence of the skill may reflect differences in the verification periods rather than forecast lead time alone. Lines 596-598: "For example, the verification periods for longer lead years contain a larger contribution from the long-term warming trend than those for shorter lead years. Because this trend is relatively predictable, differences in skill among lead years may partly reflect differences in the verification periods rather than the lead time alone (e.g., Figure

4a vs. 4h).” In addition, as noted in our response to the previous comment, we have introduced this sampling issue earlier in the Methods (Lines 212–214) to clarify that the verification periods differ among lead years because of the finite number of initialization years.

L567-568: I agree it may not affect the broad conclusions. However, tides could alter the variability from mean state on annual timescales, especially at depth, if they affect the transfer of heat from surface and depth of mixed layers. This is worthy of more discussion in context of heat content and bottom temperature results.

Response: We thank the reviewer for this insightful comment. We agree that tides may also affect annual-mean bottom temperature and upper-ocean heat content through enhanced vertical mixing and vertical heat exchange. We have expanded the discussion in the revised manuscript in lines 607-611: “While the absence of tidal forcing in GLORYS is unlikely to fundamentally alter the conclusions regarding the added value of dynamical downscaling for multi-year anomaly predictability, it may contribute to differences in T_{bottom} and HC_{200m}, particularly in strongly tidal regions such as GB and parts of GOM, through its influence on vertical mixing and heat exchange. Consequently, part of the skill differences for these variables may reflect differences in the representation of tidal mixing between GLORYS and ROMS.”

L573: Need to work with partners to ensure required outputs are stored for downscaling – for both atmospheric forcing and ocean boundaries?

Response: We thank the reviewer for this suggestion. We have revised the manuscript in lines 614-618: “Future efforts should focus on extending hindcast periods, increasing ensemble size, ensuring that the atmospheric and oceanic output required for regional downscaling is routinely archived. In addition, improving the representation of circulation variability—for example through submesoscale-resolving configurations and the incorporation of atmosphere–ocean coupling in regional downscaled prediction systems—may yield additional gains in multi-year prediction skill on the NES.”

L583: Refer to variability rather than bias, to avoid confusion with performance stats?

Response: In this sentence, we intentionally use the term “mean-state biases” because we are referring to the systematic differences between the model climatology and the GLORYS reanalysis, rather than temporal variability.

L595: transferred rather than projected?

Revised.

L620: Comment on how your skill scores <8y address this? Is the 2-5y timescale more useful to stakeholders?

Response: We thank the reviewer for this helpful suggestion. We have revised this statement in lines 662-665: “Our work on multi-year prediction fills a critical gap between seasonal and decadal forecasts. In particular, the largest improvements in prediction skill occur during the 2–5

forecast years, a timescale that is especially relevant for living marine resource management because it provides outlooks beyond seasonal forecasts while remaining more actionable than decadal projections."

Figure 1: Edit caption to clarify whether NES domain refers to white box or whole figure. Define MAB, GB and GOM, and also label clearly on figure.

Response: We have revised the figure caption to clearly distinguish the ROMS model domain, the analysis region, and the NES. The revised caption now reads: "Figure 1. Study area. (a) ROMS model domain showing bathymetry. The white box denotes the analysis region shown in panel (b). (b) Bathymetry within the analysis region. The U.S. Northeast Shelf (NES) is defined as the shelf region shallower than 200 m and consists of the Middle Atlantic Bight (MAB), Georges Bank (GB), and the Gulf of Maine (GOM), outlined in red. (c) Same as panel (b), but for the CESM bathymetry." We also added labels on revised Figure 1.

Figure 2: Suggest showing anomalies rather than mean here, as easier to assess model performance. Also, please add bathymetry contours to show the location of the shelf break for reference. (This may be obvious for some variables, but not others.)

Response: We thank the reviewer for these suggestions. We have chosen to retain the climatological mean fields in Figure 2 because the primary purpose of this figure is to evaluate the mean-state fidelity of the regional and global models and to illustrate the mean hydrographic and circulation structure of the study region, which provides important context for the subsequent prediction skill analyses. The representation and prediction of annual anomalies are evaluated separately using the temporal RMSE (Figure 3) and the anomaly-based skill metrics (ACC, MSSS, and Brier Score) presented in the following sections. In addition, a comprehensive evaluation of GLORYS against available observations over the NES has been presented by Castillo-Trujillo et al. (2023). Therefore, we believe that including additional anomaly maps would largely duplicate the information already provided by Figure 3, the anomaly-based skill metrics, and the existing evaluation of GLORYS. We have, however, added the 200-m isobath and boundaries of shelf subregions to Figure 2 to indicate the shelf break and provide additional spatial reference.

Figures 6-8: Suggest convert to table(s). If retain any, please label subpanels with titles to help reader. Also, why different colormap for ACC in Figure 8?

Response: As discussed in our response to General Comment 1, we have retained Figures 6–8 because they more effectively illustrate the evolution of prediction skill across lead years, regions, variables, and skill metrics than large summary tables. To complement these figures, we have added a summary table (Table 1) of the mean prediction skill over LY2–5 and LY6–8 for the NES. We have also improved the readability of Figures 6–8 by adding more descriptive subpanel titles, revising the figure captions, and adopting a consistent colormap for the ACC panels in Figure 8.

Figure 11: Please expand caption.

Response: We've expanded caption for Figure 11.

Figures 13 & 15: Again, consider conversion of ACC stats to table, to reduce number of figures.

Response: We thank the reviewer for this suggestion. As discussed in our response to General Comment 1, we have chosen to retain these figures because they more effectively illustrate the prediction skill and its lead-time dependence than summary tables.

Figure A6: Are all markers significant?
Response: Thanks for pointing this out. We have updated Figure A6 by including the statistical significance.