Response to Reviewer 1 (RC1):

- *My main concern is that the use of coarse gridded data products for model evaluation is not ideal for a regional model of this scale. These products (e.g., WOA, CODAP-NA, OISSTv2.1) are coarser than the model being evaluated which can make direct comparisons misleading. They are interpolated from sparse observations which can introduce biases particularly in regions with strong gradients (upwelling zones). As a result, the differences we see in many figures may not be due to model deficiencies. Moreover, comparisons with coarse gridded products do not highlight the added value of the model. I recommend further evaluation using ship-sampled data (i.e. CTDs and bottle data) or Argo data to provide a more thorough evaluation particularly of the biogeochemistry in the model. The use of direct in situ observations will be appreciated by ecologists who wish to use these data on the shelf.*

  In response to the reviewer's comment, we have augmented these evaluations with additional comparisons against in-situ CalCOFI data, including temperature, salinity, nitrate, and oxygen at multiple depths. For temperature and salinity, NEP10k maintained similar skill levels across all data points as GLORYS despite not assimilating this data in the domain interior (Fig. S26). Similar levels of agreement (r ≥ 0.96) were achieved across all points for nitrate, oxygen, phosphate, and silicate (Fig. S27). The model was more challenged to represent the temporal variation observed across decades for individual sampling sites and depth strata (Fig. S28). Agreement was best at the surface and for temperature, but generally decreased with depth. Skill improved when values averaged across the CalCOFI sampling grid were considered (Fig. S29). We also added comparisons against individual tide gauges (Fig. S13). These were added at the request of reviewer 2, but they are also responsive to this comment.

  We agree that widely applied gridded products provide an imperfect basis for model evaluation, but these carefully constructed products do provide a useful and appropriate foundation for assessing large-scale patterns across the ecosystems and fisheries-critical shelf-scale temporal variations of central importance to the intended model use. We further note that the value of the model is not limited to resolution. It provides an internally consistent set of dynamics capable of recreating patterns across multiple datasets, and can be applied in predictive applications (e.g., Ross et al., 2024).

  Our initial submission built upon the foundation of comparisons against gridded community standards with judiciously chosen direct comparisons against ship-sampled data for fisheries-critical phenomena. This included the Bering Sea "cold pool" against Alaska Fisheries Science Center data (Figures 19-20) and oxygen trend analyses against in-situ CalCOFI data (Figure 25). Direct comparisons against mesozooplankton biomass (Fig. 11), higher-resolution satellite-based measurements (Figs. 10, 18) and high-resolution data-assimilative products (GLORYS) with demonstrated skill against in-situ observations (Amaya et al., 2023a) were also included (Fig. 17).

The key implication of this comparison is that one should not expect NEP10k to match variability observed at individual observation points sampled at approximately the same time in approximately the same place across multiple years. We are not surprised that this is the case. The NEP10k hindcast does not assimilate observations, so any biases in the mean locations of fronts and other features is compounded by stochastic mesoscale and submesoscale features whose precise locations and timing will not match those observed. Coherent patterns emerge after averaging over such features (e.g., Fig. S29, Fig. 17, Fig. 20). We have enhanced discussion of this model limitation in lines 1060-1070 of the manuscript text.

We hope the additional analyses have ameliorated the reviewer's concerns, and we recognize the value of comparisons against individual datasets. Our capacity to handle the many local data sets within the domain in a single paper, however, is limited. Once the version 1.0 foundation of the model has been established, we will steadily expand comparisons and analyses in the context of specific case studies. We feel that this is consistent with the GMD's objective, and we highlight the value of bringing additional local datasets to bear in the discussion.

- *I recommend including a single composite metric like the Kling-Gupta efficiency (see Jackson et al 2019 https://doi.org/10.1016/j.envsoft.2019.05.001) and its components. This single metric that could be compared to other models. There are other options (Willmot score), but KGE has variability as one of its components and that is something you do not assess. I like that you consider bias separately to provide a clear explicit measure of error, but the analysis could benefit from a holistic assessment of how the bias interacts with the variability and correlation.*

  As the reviewer suggests, we have calculated the Kling-Gupta metric for all of the time series presented in the paper and present those, together with a breakdown of each component, in Table S1. This approach was consistent with prior usage of this metric in hydrologic time series studies (e.g., Jackson et al., 2019). To support this addition, we added a description of the Kling-Gupta metric in the methods (lines 387-388) and discussed performance throughout our results section.

- *The clarity of the writing in the manuscript could be improved by rewriting several sentences that have unclear antecedents (examples listed):*

  - *L55 "This includes [...]" suggested rewrite-> "These ecosystems include valuable fisheres that represent [...]"*

    We have made the recommended change to the manuscript text. It now reads:

Lines 55-56: *"These ecosystems include valuable fisheries that represented roughly 42% of the $4.6 billion in commercial U.S. domestic landings in 2020 (National Marine Fisheries Service, 2022)."*

- *L170 : "This was ..." This overmixing?*

  We have made the recommended change to the manuscript text. It now reads:
  Lines 179-180: *"This overmixing was ameliorated by including a scaling factor for the turbulent decay length scale"*

- *L315: "This ..."*

  We have edited the manuscript text to now read:
  Lines 345-356: *"These tidal phases and amplitudes were compared against TPXO9 to demonstrate the ability of the model to incorporate and propagate tidal boundary forcings."*

- *L325:*

  We have edited the manuscript text to now read:
  Lines 353-354: *"These nutrient limitation distributions specifically illustrate where macronutrients nitrate and phosphate or micronutrient iron are the primary nutrient limitation of phytoplankton growth."*

- *L415: "This ..." -> "This division..."*

  We have made the recommended change to the manuscript text. It now reads:
  Lines 469-470: *"This division yields an ~10 x 10 grid (i.e., square) decomposition of model grid cells on each PE"*

- *L517 "This ..." These biases?*

  We have made the recommended change to the manuscript text. It now reads:
  Lines 580-581: *"These biases correspond with the most prominent region of overmixing (Fig. 4)."*

- *L525*

  We have edited the manuscript text to now read:
  Lines 588-589: *"These biases are consistent with shallow mixed layer biases in the Gulf of Alaska (Fig. 4)"*

- *L550 "This gradient?"*

  We have made the recommended change to the manuscript text. It now reads:

Lines 614-616: *"This distribution of dissolved iron results in large-scale patterns of phytoplankton iron limitation in the NEP10k simulation (Fig. 9, right panel) that are consistent with those observed (e.g., Moore et al., 2013; Hutchins et al., 1998)."*

- ○ *L638*

  We have edited the manuscript text to now read:
  Line 701: *"These surface alkalinity biases are aligned with positive salinity biases that penetrate to depth (Fig. 3)."*

- ○ *L913*

  We have edited the manuscript text to now read:
  Lines 998-1001: *"This weaker correlation was not necessarily surprising, given the volatile and patchy nature of coastal chlorophyll and observing challenges in such environments, but points to the need for further scrutiny of both the model and observations before predictive chlorophyll applications can be realized in most systems."*

- ○ *L935*

  We have edited the manuscript text to now read:
  Lines 1021-1023: *"This decline in bottom coverage by the coldest watermass category coincides with a dramatic monthly reduction in NEP10k's SEBS sea ice extent relative to satellite estimates (Fig. S20, May - April and June - May)."*

Technical Corrections

- *Lines 60-61: consider referencing Christian and Holmes 2016 https://doi.org/10.1111/fog.12171 and Thompson et al. 2023 https://doi.org/10.1098/rstb.2022.0191*

  We have added the suggested references to the manuscript text, which now reads:
  Lines 59-61: *"…potentially driving fluctuations in living marine resource abundance due to habitat range shifts (e.g., Pinsky et al., 2013; Christian and Holmes, 2016; Smith et al., 2021; Chasco et al., 2022; Thompson et al., 2023)"*

- *L 63 and elsewhere- Check that your citations are in chronological order*

  We have revised the reference order here and elsewhere in the manuscript text to be in chronological order

- *L100. Revise this sentence for clarity. I find the words "have contributed to" to be unclear. Climate models such as the NPGO and PDO result from a variety of different processes (e.g. Newman et al. 2016). They are associated with (correlated with) ecosystem regime* *shifts, but they are not phenomena in and of themselves and cannot, therefore, cause anything.*

Per the reviewer's recommendation, we have clarified the manuscript text. It now reads:

Lines 98-102: *"While correlation with the El-Nino Southern Oscillation (ENSO) can be found (e.g., Bailey et al., 1995; Whitney and Welch, 2002; Amaya et al., 2023), lower frequency modes of decadal climate variability tend to predominate (e.g., Di Lorenzo et al., 2008) and are associated with marked decadal-scale ecosystem regime shifts (Anderson and Piatt, 1999; Hare and Mantua, 2000) and modulations in fisheries and ecosystem risks (Hauri et al., 2021b, 2024)."*

- *L112 – there is evidence that CTW can propagate the ENSO signal to the GoA (Amaya et al 2023; https://doi.org/10.1038/s41467-023-36567-0)*

We have added the Amaya et al. 2023 reference to the earlier paragraph wherein we describe the Gulf of Alaska ecosystem (The line indicated by the reviewer occurs in a paragraph dedicated to describing the California Current Ecosystem). The manuscript text now reads:

Lines 98-102: *"While correlation with the El-Nino Southern Oscillation (ENSO) can be found (e.g., Bailey et al., 1995; Whitney and Welch, 2002; Amaya et al., 2023), lower frequency modes of decadal climate variability tend to predominate (e.g., Di Lorenzo et al., 2008) and are associated with marked decadal-scale ecosystem regime shifts (Anderson and Piatt, 1999; Hare and Mantua, 2000) and modulations in fisheries and ecosystem risks (Hauri et al., 2021b, 2024)."*

- *L149 - "time step"*

We have made the recommended change to the manuscript text. It now reads:
Lines 151-152: *"Simulations used a baroclinic time step of 400 seconds and a variable barotropic time step set to maintain stability (Hallberg, 1997; Hallberg and Adcroft, 2009)."*

- *L255 – how long did it take for the model to "converge"? how do you know?*

Our goal with the model spinup was to ensure that any drifts in the upper ocean properties (i.e., down to 500m) critical to fisheries habitat were generally small relative to interannual ocean variability critical for understanding past fisheries fluctuations. We now state this goal in the methods and include an analysis of the time-evolution of habitat-critical properties for each of the regions in Fig. 1 (Figure S3).

Lines 271-277: *"The purpose of implementing a spinup was to omit drifts in the biogeochemistry associated with the adjustment of the model from its initialized state, which was generally based on coarse-resolution observation-based products, to the model's characteristic solution. We focused on fisheries-relevant variables in the top 500m. We* *found that a spinup period of 10 years generally resolved initial model adjustments, which were strongest in the British Columbia region (Fig. S3). While 10 years removed the strongest drifts, subtle trends remain in some regions, suggesting the potential value of longer spinup periods.These spinup sensitivities are left to future NEP10k development efforts."*

- *L377: "We compared.." show me don't tell me – what did you find?*

  We report our findings on seasonal Bering season sea ice in the results (Section 3.2.1). Bering Sea-specific indicators, Lines 824-828, and in Figures 21 and Supplemental Figure 21.

- *L404: "We also assessed the long-term trends [...]" where is this? what did you find? How did the bottle data compare to the model?*

  We report our findings and describe the results of comparing NEP10k Oxygen trends against CalCOFI Section 3.2.3 California Current-specific indicators, Lines 878-883, and in Figure 25.

- *L419 – in the caption of Fig. 1 you said that the white part was not in the computational domain. But here you say that you omit grid cells that contain only land. These can't both be true; there are grid cells that contain both land and water.*

  The Figure 1 caption states that, "White coloration indicates non-ocean (i.e., masked) grid cells that are not computed in model integrations". This means the computer does not perform any calculations for these grid cells. However, these grid cells are still part of the domain and may be allocated to a computer processor, even though there are no calculations to be made for that specific grid cell. Model grid cells are designated either "land" or "ocean", there are no grid cells that contain both. We generate subsets of the NEP10k domain to distribute to computer processors; here our subsets are roughly 10 grid cells x 10 grid cells in size. Some of these 10x10 subsets contain all "ocean" grid cells, some contain a mix of "ocean" and "land" grid cells and some contain only "land" grid cells. When a 10x10 subset contains only "land" grid cells, that subset is not allocated to a computer processor because there is nothing that needs to be computed. Whereas, when a 10x10 subset contains both "land" and "ocean" grid cells, calculations are performed for the "ocean" grid cells while the "land" grid cells are ignored (i.e., skipped by the "for loops" used to perform the integrations).

  Lines 158-159: Figure 1 caption: "*White coloration indicates non-ocean (i.e., land-masked) grid cells that are not computed in model integrations, which include the Sea of Okhotsk."*

- *L501 space needed at start of paragraph*

  We have added indentations at the start of each paragraph.