We thank both reviewers for their careful, constructive reviews. Their comments have substantially improved the manuscript. Below, reviewers' comments are given in black and our response in blue. The reviewers' comments retain the original submitted figure numbers; our responses use the revised numbering.

Summary of changes made:

1. Incorporated most of the reviewers' suggestions regarding changes to the manuscript title, interpretation of prediction skill, and organization of the main text and methods. In the few cases where our responses departed slightly from the reviewers' suggestions, we provide clear rationale to support our choices.
2. Revised all figures for improved visibility, quality and interpretability.
3. Added detailed explanations, suitable to a broad readership, to the Methods sections on various model simulations carried out in this work.
4. Added a table that clearly contrasts the distinct attributes and configurations of each simulation.
5. Added a new figure that shows the main drivers of bottom O_2, pH and Ω_{ar} during the study period.

RC1:

Review for Koul et al. – Decadal biogeochemical predictions for the bottom marine environment of the Northeast U.S. Continental Shelf

This manuscript presents a decadal scale forecast for the Gulf of Maine, describing the skill of hindcast models that can be used to parse mechanisms useful for predicting changes to temperature and other water quality conditions that are important for marine biota. The baseline model performs favorably against multiple sources of regridded observational data and is a step towards operationalizing longer-term forecasts that would be important for fishery and ecosystem management purposes.

Overall, the paper is well written and does a good job describing skill assessments and limitations of multiple approaches. It would also benefit the paper to further describe the dataset error bars presented, specifically what they correspond to (spatio-temporal variability?) and how they were computed/propagated after regridding. After reading the paper, it is my understanding that substantial degradation in predictive skill is unavoidable after ~2 years of simulations into future conditions (Figure 4), inherently limiting the forecast window of this approach since it will necessarily be forced to guess at realistic boundary conditions that influence the projected results. However, this was a bit difficult to discern from my reading of the paper, and ties together with some general confusion about differences between the NWA12_HIND and NWA12_HIST model simulations, particularly with respect to boundary conditions and how they are being used for forecast analyses. As it reads in the methods section, I am unsure if NWA12_HIST simulations are exactly equivalent to NWA12_HIND but are simply missing an initialization period? If so, I remain confused about how the authors are considering the impacts of unavailable future boundary conditions to influence their model results, particularly with some of the lags found in Figure 5 when attempting to simulate accurate probabilistic predictions of multiple variables. I believe that many of these points, some of which are addressed in the detailed comments below, are not insurmountable and could be more thoroughly addressed by the authors before publication.

Thank you for your detailed comments and suggestions. We recognize that our submitted draft used nomenclatures that, while taken in the fields of climate prediction, can be confusing and ambiguous. We have extensively revised the descriptions of all model simulations in this study to provide greater clarity and detail. Additionally, we have included a new Table 1 that clearly contrasts the distinct attributes and configurations of each simulation. We believe with these revisions, the interpretation and comparison of prediction skill is unambiguous.

Detailed Comments

Line 105: Suggest striking "however" to make sentence flow better.

Removed.

Figure 1: Suggest modifying aspect ratio (widen) panel a to improve readability, a little difficult to see any detail for bathymetry like that present in the Northeast Channel. May also be beneficial to include arrows showing sources of Warm Slope, Labrador Slope, and Gulf Stream water masses. What do the error bars correspond to, just standard deviations over the regridded datasets?

We have now added a new stand-alone Figure 1 which shows the region of intertest, the 1000m isobath and other features through a larger map.

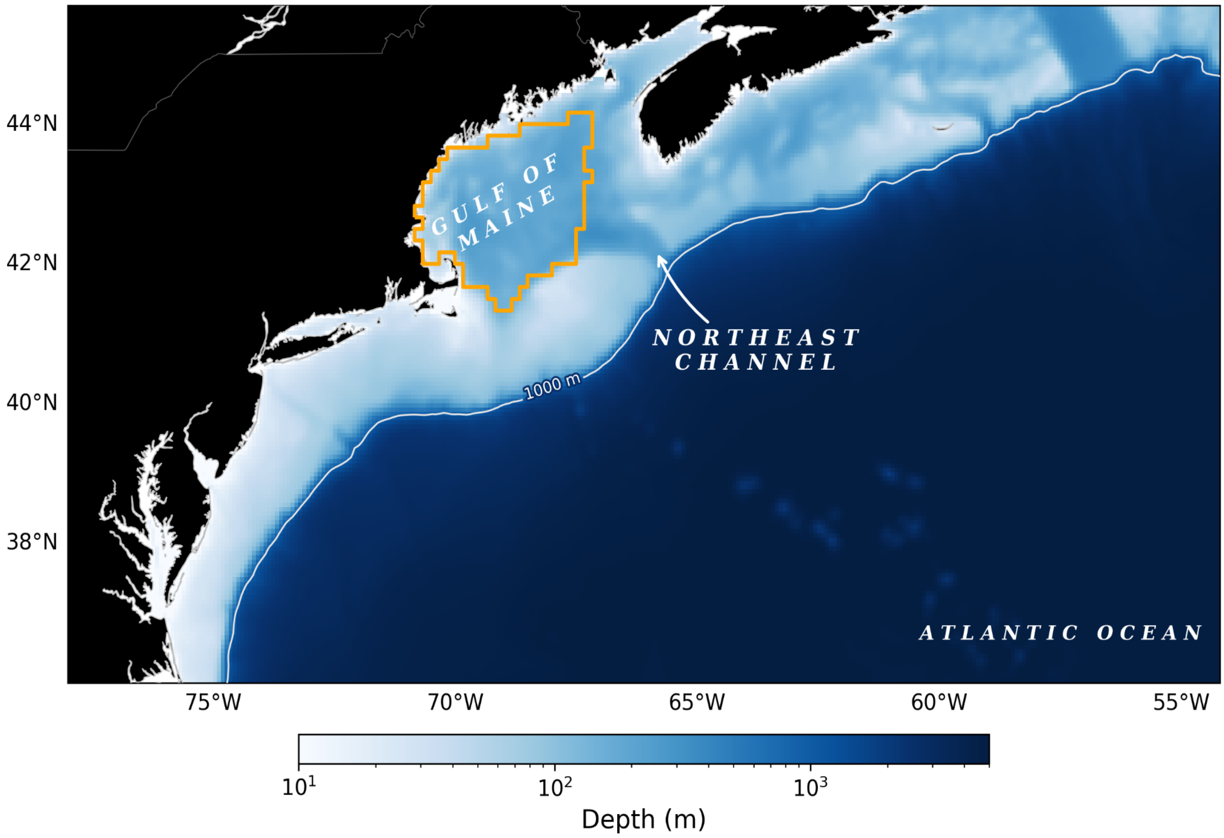


Fig 1: The study region and its characteristics. *The locations of the Gulf of Maine Ecological Production Unit (GOM_EPU) and the Northeast Channel along the Northeast US continental shelf are shown along with the model bathymetry, The names and approximate locations of the Warm Slope Water and Labrador Slope Water are also shown. The 1000-meter isobaths is also shown as a black and white line respectively.*

The error bars in Figure 2 characterize the uncertainty (standard error of the mean) in the annual mean anomaly, which arises because it is estimated from the mean of a limited set of imperfect monthly anomaly estimates. The error bars are computed separately for each observational product.

The steps in the calculation are:

1. The observations are area-weighted-averaged over each EPU (point products (Fishbot and the vessel/glider data) are first interpolated to a 1/4° grid each month, then area-averaged)
2. Monthly anomalies are formed relative to the 1994–2023 monthly climatology
3. The annual mean anomaly is the average of that year's monthly anomalies, accounting for differences in the number of days in each month. Annual mean anomalies are retained only when at least 4 valid months are present otherwise not computed;
4. The plotted uncertainty for that year is $\sigma\_monthly / \sqrt{n}$, i.e. the sample standard deviation of that year's monthly anomalies divided by the square root of the number of valid months to get the uncertainty in the mean value.

We have clarified the calculation of the error bars in the revised Methods section (Lines 293-300).

I'm unsure what the implications may be for normalizing temperature to the variables in d-f, as it may imply that a stronger direct relationship between unit changes in O2, pH, and omega for changes in temperature when the mechanisms are more complex. I also am a bit confused by the choice to invert temperature on panel d, wouldn't it be more beneficial to show that increases in oxygen anomalies are correlated with decreases in temperature and vice versa? As plotted now, it requires a much more careful reading of the plot to confirm this.

We agree that in the original figure the bottom-temperature curve in (d)–(f) was linearly rescaled, which could indeed be read as implying a fixed unit-for-unit relationship between temperature and O_2 / pH / Ω_{ar}. That was not our intent. The curve was only meant to indicate temporal co-variability. In the revised figure (now Figure 2), bottom temperature is drawn on its own dedicated secondary (right-hand) axis in true units (°C). The temperature curve now conveys only the sign and timing of co-variability, not an implied unit conversion. For example, the anti-correlation between bottom oxygen and bottom temperature is now read directly: when the temperature anomaly falls, the oxygen anomaly rises and vice versa. We have also revised the figure caption to make this explicit.

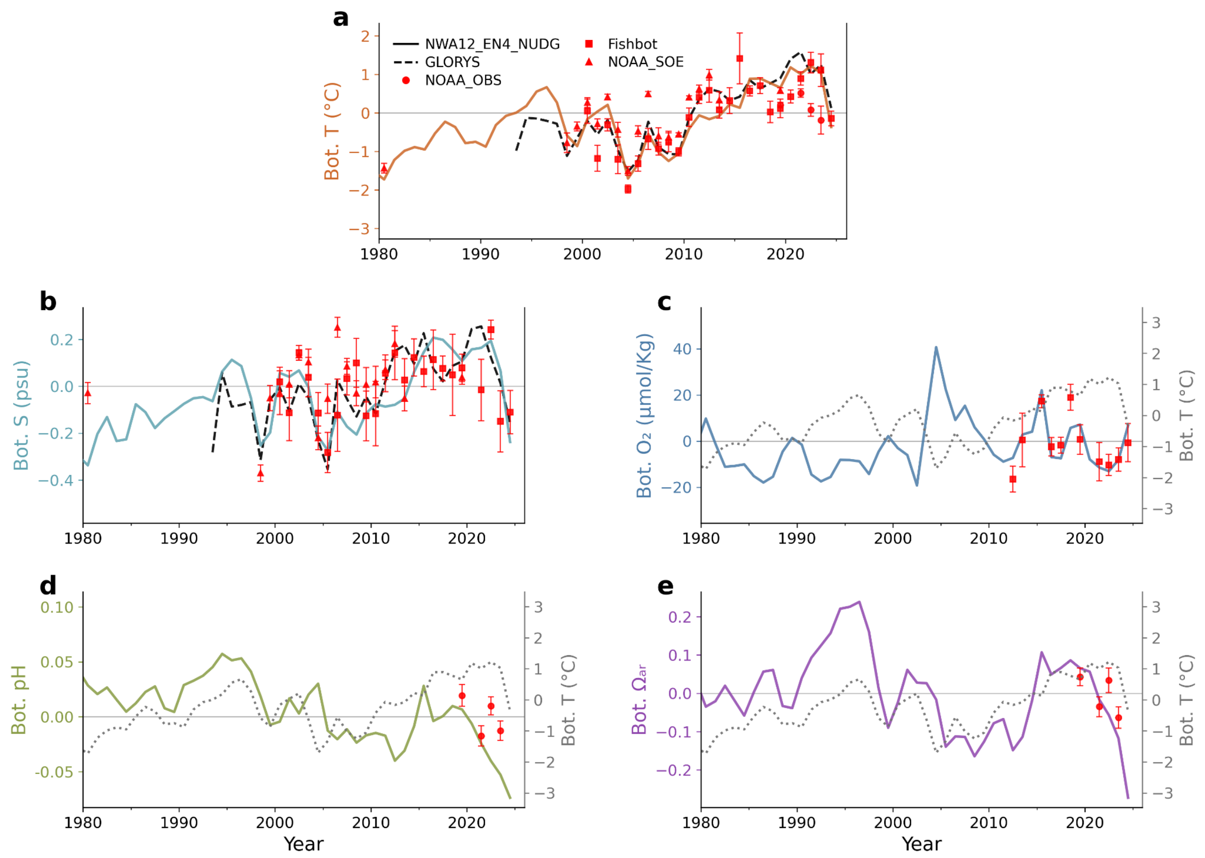


Fig 2: Physical and biogeochemical variability in the GOM. *Annual mean anomalies of (a) bottom temperature (°C) (b) bottom salinity (psu) (c) bottom oxygen (μmol/kg), (d) bottom pH and (e) bottom aragonite saturation state in the GOM from the ensemble mean of the regional downscaled simulations (NWA12_EN4_NUDG, colored solid lines) and observations (markers with error bars). Observations from the NOAA SOE reports (NOAA_SOE, triangles), the Fishbot bottom data (squares) and the aggregated point observations from vessels and gliders (NOAA_OBS, circles) are also shown for available years. The error bars depict the uncertainty of the annual mean anomaly (see Methods). The dashed black lines in (a) and (b) are from the GLORYS12 dataset. All anomalies are calculated with respect to the 1994-2023 mean. To highlight physical and biogeochemical co-variability, bottom temperature lines are shown as dotted grey lines in (c)-(e).*

Line 243-249: Are you still referring to Figure 3 in this paragraph? Or Figure 1?

We are still referring to Figure 3 (now Figure 4 in the revised manuscript), so we have added the figure reference at the end of the sentence.

Line 254-259: This explanation seems plausible, but as you note the large increase in bottom oxygen concentrations in 2015 seem to also be driven by a large positive winter MLD anomaly (Fig. S5) and that does seem to match well with Fishbot data. Are there other plausible explanations for the rapid declines in prediction skill?

The episodic bottom-O_2 maxima (2004, 2015, 2019) are driven by deep winter mixing that is itself largely meteorologically forced and therefore not predictable at multi-year leads. This is the primary reason for the low skill in bottom-O_2. The rapid drop in the skill beyond 2 years indicates that the signal in the advective water-mass variability through the Northeast Channel is weaker and not the principal contributor to the skill. The absence of any anthropogenic long-term trend in the bottom-O_2 time series also limits the prediction skill.

Line 292-293: Missing end quote, but also suggest rewording to "… ask about forecast reliability of anomalously warm or cool conditions."

Added the missing closing quotation mark.

Line 303-323: Perhaps this is also addressed in the discussion, but it seems from looking at Figure 5 that the model is most limited by transition periods and its own internal variability that increases the persistence of ongoing trends. For all variables shown, there are numerous examples of high confidence projected by the model for the persistence of a trend, followed by an eventual transition 1-2 years later (or sometimes missed altogether as seen in the 2010s for pH and omega). I don't think that this necessarily indicates a critical lack of model skill, but it does warrant some further elaboration and discussion here.

We have added a new skill metric to Figure 7 to help quantitatively demonstrate the model's skill at predicting transitions by comparing with a persistence forecast. This metric is a skill score based on the Ranked Probability Score (RPS).

The Ranked Probability Score (RPS) for a single forecast is defined as

$$RPS = \frac{1}{K-1}\sum_{k=1}^{K}\left[\sum_{i=1}^{k} P_i - \sum_{i=1}^{k} O_i\right]^2$$

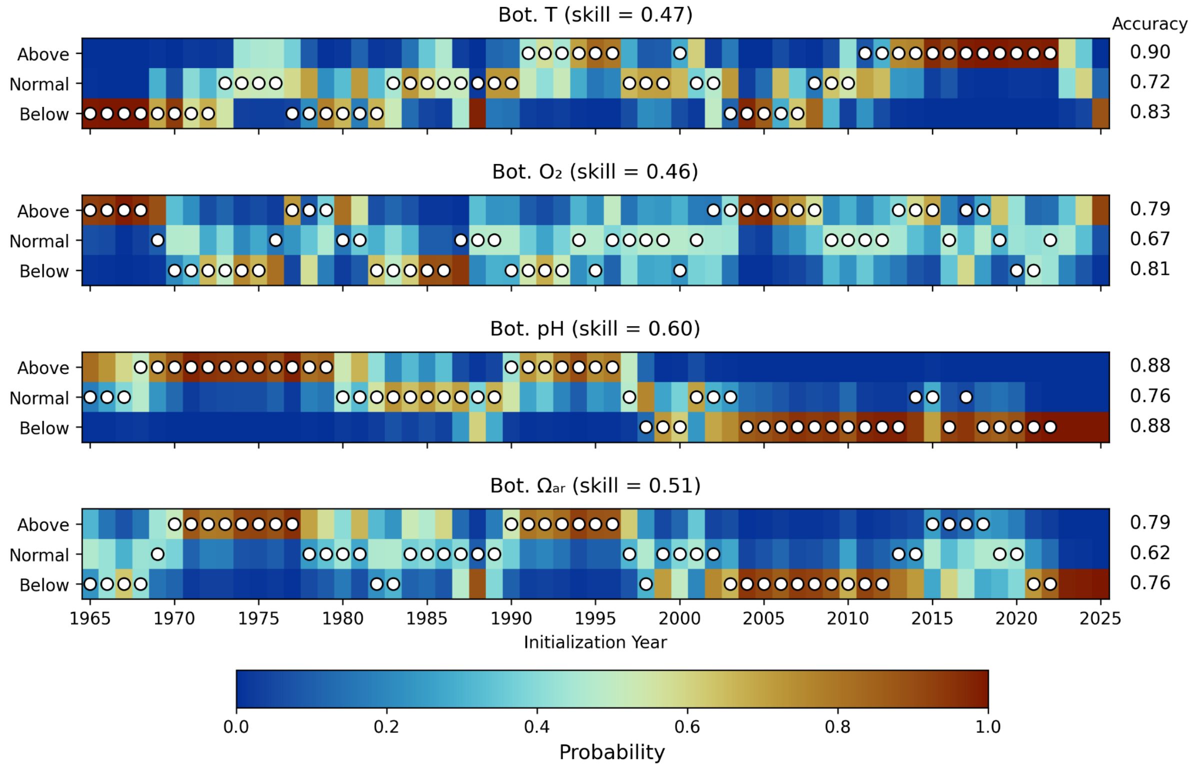
where K is the number of categories (3 in our case: below, normal, or above), P_i is the forecast probability for category i and O_i is the observed probability (0 or 1) for category i, such that the inner sums calculate the cumulative probability up to outcome k. A perfect forecast, which assigns 100% probability to the observed category, will have an RPS = 0. Predicting a 100% chance of the opposite outcome (i.e., below when the outcome is above) gives an RPS of 1, while predicting normal when an outcome is above gives an RPS of 0.5. The RPS is notable in that it evaluates the distribution of the forecast probability around the outcome, penalizing outcomes that are further away. This makes RPS a more sensitive metric than accuracy.

To compare the model forecast accuracy against a baseline forecast of no transition (i.e., persistence), we calculated the ranked probability skill score (RPSS):

RPSS = 1 - RPS (forecast) / RPS (persistence)

where RPS (forecast) is the sum of the ranked probability scores comparing the forecast probabilities with the observed categories, and RPS (persistence) is the sum of the ranked probability scores comparing the observed categories with the categories observed in the previous year (in other words, we assign 100% probability to the category observed in the year before the forecast was initialized, and compare that with the actual category for the next 3 years). By the definition of the skill score, if RPSS = 1, the forecasts had perfect accuracy at probabilistically predicting the observations. If RPSS = 0, the forecasts had the same overall score as the persistence forecasts of the previous year's state; in other words, the forecasts had no skill at predicting transitions.

Our simulations had RPSS values between 0.46-0.6 (see the new Figure 7). The RPSS is highest for pH and lowest for oxygen which is broadly consistent with the results in Figure 3. Since all the skill scores are greater than 0, we can conclude that the model forecasts are better than persistence for probabilistically predicting the tercile for the next 3 years, which means that the forecasts do have skill at predicting transitions (since persistence would never forecast a transition). However, since the skill scores are less than perfect (1.0), there is room for improvement. We have added a short discussion of these new methods and results to the text (Lines 332-360 and Lines 606-615).



Line 377-379: I would agree that the model does largely predict past variability well for temperature and salinity, excluding a few outliers in the 2000s, and would be curious to know if the bottom temperature skill is comparable to the surface temperature skill referenced in Koul et al. (2024). It seems like a bit of a stretch to say that it captured past variations in pH and omega (Figure 1) given the relatively limited dataset.

The SST skill in Koul et al 2024 is higher than the bottom temperature skill presented in this study. The primary reason is that the highly predictable impact of GHG warming on the ocean temperature diminishes as one moves from the surface to deeper in the water column. For the past variations in pH and omega, we acknowledge that the available observations are limited. In lieu of observations, we are relying on the fairly well-established covariance between these variables and temperature and salinity, which the model is able to reproduce.

We revised this section in the manuscript as follows (Lines 713-719):

However, the model successfully simulated observed variability in bottom temperature and salinity and captured the expected covariance between these two variables and the bottom biogeochemistry. Koul et al. (2024) also demonstrated forecast skill for surface temperature when compared against actual observations. Furthermore, the model is likely accurate in predicting much of the forced component of skill, which is driven by steadily increasing greenhouse gas concentrations and resulting impacts on ocean warming and chemistry.

Line 391: suggest replacing “however” with “therefore”

Replaced

Line 393-398: This sentence is a bit of a weak conclusion to some impressive work in the paper. Suggest striking or finding other references that you can use to stress the importance of probabilistic forecasting on longer timescales.

We agree and appreciate this suggestion. We’ve removed the last sentence and re-inserted the references that note the benefits of probabilistic forecasts for decisionmaking in the section where the probabilistic predictions are first introduced.

Line 466: Please define or provide citation for SSP585.

We now define SSP585 as the high-emission Shared Socioeconomic Pathway (SSP5-8.5, Riahi et al. (2017))

Figure S5: Typo of Fishbot in caption and would be helpful to define what errorbars shown correspond to.

We corrected the typo and added the meaning of error bars in the figure caption (Figure 2) and in the revised methods.

General figure comments: Some parts of the figures can be a little difficult to read (e.g., axis labels in Fig 1-2, 4, legend and skill metrics in Fig. 2), recommend doing a quick reformat to make text more visible and consistent among them

We revised all figures to ease their visibility and improve their quality in suggested aspects.

RC2:

Overview

This paper presents and evaluates a prediction system for physical and biogeochemical variables in the Gulf of Maine, largely motivated by applications to fishery management. The authors carefully evaluate a sophisticated set of models and simulations to show that there is some predictive skill in the modeling framework, which is an important new accomplishment.

I struggled to review this paper because it was difficult to understand for two main reasons: (1) the modeling system is complex and not described in sufficient detail for all but the most sophisticated readers to understand and (2) many of the figures are difficult to read.

We agree and have taken several steps to address these issues. First, we have expanded and reorganized our methods section according to the reviewers’ recommendations and now include a concise table summarizing the three model configurations used in this study and the purpose served by each. We have also edited the figures for clarity in response to the detailed comments of both reviewers.

I believe the paper is an important contribution to ocean forecasting. The paper makes a clear distinction between forced and natural variability, which is valuable. I am very pleased to see inclusion of important biogeochemical variables (oxygen, pH, and aragonite saturation state) in the forecasting system. The explanations of the causes of variability are good but fall short in several cases for the biogeochemical variables.

Decadal predictions are not new with this system, as they were done in Ross et al. (2024) and Koul et al. (2024), but I think what is new in this paper is the inclusion of biogeochemical predictions. That point could be brought out a little more clearly; I was having a hard time figuring out exactly what was new in this study.

We agree and now state explicitly, in the introduction, that the novel contribution of this study is coupled biogeochemical decadal prediction.

I think the problems I bring up are fairly straightforward to address and I have tried to be clear and constructive with my criticisms and suggestions.

Thank you for your efforts in reviewing this manuscript and for providing helpful suggestions.

One last point is that the title is a little misleading by including "Northeast U.S. Continental Shelf." The paper is almost exclusively about the Gulf of Maine, so if a particular region is going to be mentioned in the title then it should be the Gulf of Maine. If the authors want the paper to reach a more general audience, they could end the title with " … bottom marine environment of a coastal ecosystem."

This is a fair point—although many of the changes discussed in the paper took place over the full Shelf and surrounding ocean (shifting Gulf Stream and changing currents, for example), the forecast skill evaluation was focused only on the Gulf of Maine subregion. We have changed the title to "Decadal biogeochemical predictions for the changing bottom marine environment of the Gulf of Maine". The relevance of results to other coastal ecosystems is noted in the abstract, keywords, introduction and discussion (Title and Lines 23-26, 130-138).

Methods

Most of my time was spent trying to figure out what the authors actually did. The authors present a fairly sophisticated modeling system, with two models (a global model and a regional model), each used in very different ways. Ensembles add another layer of complexity. If one is an expert in this sort of system and has read all of the papers by this research group, then reviewing this paper might be straightforward. Because I was not familiar with the prediction system, reviewing the paper took me about 15 hours. The authors talk about the real-world applications of this research, so it seems that they would like to target a more general audience. If that is the case, then I would suggest that the authors provide more clarity and motivation in their description of the methods. I do not know if it was the authors' choice to place the methods at the end of the paper or if this is required by the journal, but doing so makes the paper more difficult to read. There should be clear links between motivation, methods, and results; placing the methods at the end breaks such linkages.

We want to thank the reviewer again for their effort and dedication in reading the paper and providing a helpful review.

We have revised the paper to place the methods section in the more typical order, between the introduction and results, and to provide additional information and context to help a wider audience readily interpret the paper. Details of these revisions are provided in the responses to your more specific points below.

I struggled to figure out all of the simulations that were done in this study. The simulations go by many names, including historical, retrospective, forecasts, hindcasts, nudging, downscaled, and predictions. A difficulty is that words that seem to have similar meanings (like retrospective and historical) are used to describe very different simulations. I think it would be helpful if the authors used one descriptor for each type of simulation and stuck to that descriptor throughout the text.

We absolutely agree that methodology needs to be clarified and each model simulation needs to be explicitly described. We now complement the improved and relocated Methods section with a Table summarizing each component of the prediction system and the role that it plays.

I will summarize what I think the authors did. There are three types of simulations using the high-resolution (regional) model: (1) retrospective simulations, (2) retrospective forecasts (hindcasts), and (3) historical simulations. Detail on each follows.

(1) Retrospective simulations. These are simulations from 1960 to 2024, which use nudging at a time scale of 90 days towards the full vertical profiles of T and S in the EN4 1-degree gridded data set wherever the ocean bottom is deeper than 1000 m. (As an aside, it would be helpful to get an idea of where the 1000 m isobath is. I suggest that the authors place a map larger than Figure 1a in the paper and clearly indicate where the 1000 m isobath is, as well as the various regions described in the paper: GOM, MAB, SS, and GB.) These simulations are forced at the surface and open boundaries by the global atmosphere–ocean–sea-ice SPEAR model, which itself is nudged towards observed SST and reanalysis winds. This regional model simulation is called NWA12_EN4_NUDG in the authors' nomenclature. It is confusing to me that these retrospective simulations are presented in terms of their ensemble mean. I think there are 10 members, so that would be a total of 65 years x 10 members = 650 years of simulation. I do not understand how the ensembles were created for retrospective simulations. If these simulations are forced by SPEAR, which has also been nudged towards observations, then how is a true ensemble being created? Also, it would seem difficult to generate a true ensemble if T and S are being restored towards observations. SPEAR must be free-running at some point in order to create the ensemble, but I cannot tell when the authors let go of the restoring of SPEAR towards observed SST and winds. The fact that the authors do not discuss the ensemble spread in Figures 1 and 2 adds to the confusion. I do not understand the point of creating an ensemble if only the ensemble mean is presented. Finally, I was surprised at the river input being climatological (lines 443–444). Is it well established that interannual variability in these shelf waters is independent of river input?

Your summary of the retrospective simulations (NWA12_EN4_NUDG) is accurate. We have now expanded our description to clarify the points of confusion regarding ensemble generation and river input that you have raised (Lines 144-222, 226-233). We have also added a new table (Table 1) to help clarify some of the similarities and differences between the three types of downscaled simulations.

The use of an ensemble in NWA12_EN4_NUDG is motivated by the desire to have multiple plausible initial states to initialize the ensemble forecasts highlighted in Figs. 5, 6 and analyzed probabilistically in Fig. 7. The 10 ensemble members of the NWA12_EN4_NUDG ensemble (called "Retrospective simulations") are derived by running the regional model with member specific ocean initial conditions, atmospheric forcing, and open ocean boundary conditions all derived from 10 different ensemble members of the SPEAR_LO global coupled reanalysis, covering the period 1960-2024. The ensemble members of the SPEAR_LO reanalysis were initialized from the SPEAR_LO large ensemble simulation, which is free running, so each member starts with a different realization of the internal ocean state. From 1958 on, the SPEAR_LO reanalysis was nudged towards JRA55 atmosphere data and ERSSTv4 sea surface temperature. Because the initial conditions for each member are different, and the nudging is relatively weak and applied only at the ocean surface, enough divergence remains in the SPEAR_LO reanalysis ensemble to provide a suitable ensemble of initial conditions for forecast simulations. Additional details are available in Yang et al., 2021, On the Development of GFDL's Decadal Prediction System: Initialization Approaches and Retrospective Forecast Assessment. JAMES. https://doi.org/10.1029/2021MS002529.

In NWA12_EN4_NUDG, we simply carry forward the member-wise information present in the ocean initial conditions, atmospheric forcing, and open ocean boundary conditions from the 10-member SPEAR_LO. In addition, we relax the model T, S solution towards EN4 T, S wherever the ocean bottom depth exceeds 1000 m. The 1000 m isobath is now shown in Fig 1. Again, because this is nudging and not a complete override of the data, and is only applied off shelf regions, there are enough differences in the ensemble to generate significant ensemble spread for the forecast simulations, as illustrated explicitly in Fig. 5 and 6.

We focus on the ensemble mean in Figs. 2 and 4 to efficiently assess the consistency of the historical variability in the model reanalysis with the observed variability. While the ensemble members vary around this mean pattern, the primary variations are common across them. This is illustrated by the consistency of the initialized anomalies in Fig. 6. The ensemble spread in the reanalysis simulation is mainly relevant for initializing the ensemble of forecasts (Fig. 5, 6).

Regarding the rivers, a climatology is used because we do not have coastwide observations of river runoff dating back to the 1960s that would be needed to force the model. The GloFAS reanalysis product that we use does extend back to 1979, but rather than having an inconsistency in the simulation before/after 1979 we chose to use a climatology for the full-time period. We also think this is justified because direct river runoff is a minor source of freshwater to the Gulf of Maine (external sources are more than half the freshwater (Smith, P.C., Houghton, R.W., Fairbanks, R.G. and Mountain, D.G., 2001. Interannual variability of boundary fluxes and water mass properties in the Gulf of Maine and on Georges Bank: 1993–1997. *Deep Sea Research Part II: Topical Studies in Oceanography*, *48*(1-3), pp.37-70.) and almost all of the nitrogen loading (Townsend, 1998)) and most of the interannual variability is driven by variability in whether the inflowing water consists of Warm Slope, Labrador Slope, or Scotian Shelf Water (see Smith et al., 2001 and others). We have added a sentence explaining this to the manuscript (Lines 226-236).

(2) Retrospective forecasts (hindcasts). These are a series of 10-year simulations that vary depending on the start year and the ensemble member. Every year starting in 1960 has a set of 10 ensemble members. I am uncertain as to what the last start year is. Is it 2024? Or is 2024 the last end year, in which case the last start year would be 2015? Looking at Koul et al. (2024), I think the last start year might be 2024. Either way, this set of simulations represents a lot of simulation time: 10 members x 10 years per simulation x 55 (or 65) start years = 5500 (or 6500 years). The authors call these simulations NWA12_HIND. The 10 ensemble members seem to be tied to the same 10 ensemble members in the retrospective simulations above, both for the regional and global models, but because I do not understand the ensemble design for the retrospective simulations, it is difficult for me to understand the ensemble design for the hindcasts. One last puzzle is related to the relaxation time

scale north of 45 deg N, which "decreases sinusoidally to 0 at 45 deg N." Since sinusoids are not monotonic, I cannot see how one can have a sinusoidal decrease.

The NWA12_HIND (called "retrospective forecasts" or simply "hindcasts" in our paper and the broader prediction literature) are indeed 10-yearlong simulations started every year from 1965 (first start year of the 10-yearlong simulation) to 2025 (last start year 10-yearlong simulation). So, the last 10-yearlong simulation ("forecast") is from 01-01-2025 to 31-12-2034.

The 10-member NWA12_HIND ensemble is indeed started from ocean initial conditions taken from the corresponding 10 members of NWA12_EN4_NUDG. The 10 member NWA12_HIND hindcast ensembles are then generated using atmospheric forcing and open ocean boundary conditions from the corresponding members of the global SPEAR_HIND forecasts used to generate the NWA12_EN4_NUDG ensemble member. This ensures a consistency between the initialization and subsequent forecast.

These aspects of the hindcast are now clarified in the methods (Lines 237-258 and in Table 1).

North of 45°N, we relax the NWA12_HIND T, S solution towards the SPEAR_HIND T, S with 90-day timescale at the northernmost boundary and no relaxation at 45°N. In between the stretch of the relaxation is tapered off using a cosine curve. We agree the use of the term "sinusoidal" is inappropriate. We now provide the explicit function used:

The latitudinal damping rate (inverse of relaxation time scale) was prescribed as:

$$r(\text{latitude}) = (r\_max / 2) \times [1 + \cos(\pi \times (\max(\text{latitude}) - \text{latitude}) / (\max(\text{latitude}) - 45))]$$

for latitudes from 45°N to max(latitude) in the regional model, and zero south of 45°N and

$r\_max = 1/90\ \text{day}^{-1}$ at max(latitude), where max(latitude) in the regional model is ≈ 58°N

This aspect of our approach in now clarified in Lines 248-256.

(3) Historical simulations. The authors also call these "uninitialized," which I do not understand because every simulation must be initialized. These simulations are forced by SPEAR, but apparently not the same configuration of SPEAR used in the first two sets of simulations (NWA12_NUDG and NWA12_HIND). This configuration of SPEAR is called SPEAR_HIST. In this case SPEAR does not seem to be nudged towards observations, though it is difficult for me to tell. Very little information is provided about SPEAR_HIST. The description is given in the Methods from lines 457 to 468 and seems incomplete to me. All that is stated is that "greenhouse gas concentrations and other radiative forcings" are included. No references are provided. 10 ensemble members of SPEAR are used to force the regional model, I think from 1850 to 2100. So that would be 250 years x 10 ensemble members = 2500 years of simulation. Is that correct? But only output from the time period 1965 to 2024 is used, so a total of 600 years? The corresponding regional model simulations are called NWA12_HIST.

We apologize for the confusion. The "uninitialized" and "historical" descriptors used in the paper are drawn from the terminology used for global climate change projections conducted for the Coupled Model Intercomparison Project (CMIP) in association with the Assessment Reports of the Intergovernmental Panel on Climate Change. In the CMIP context, a climate change projection is broken into "pre-industrial" (pre-1850), "historical" (1850-2012 in CMIP6) and "projected" (usually

2012-2100) segments. These segments are characterized by different concentrations of greenhouse gases, land use patterns and aerosols that impact climate.

Since the main purpose of the historical and projected segments is to assess climate change - the change in the statistics of weather and ocean conditions - there is no attempt to initialize the historical portion of the run with the actual state of ocean variability (e.g., ENSO) in 1850. This differs from seasonal to decadal climate predictions, where the ocean state is carefully initialized, usually via data assimilation, to capture patterns prior to the forecast, and is the origin of the "uninitialized" terminology used in the paper and elsewhere. Instead, arbitrary years from the pre-industrial segment are used to spawn different historical ensemble members. These historical runs have ocean variability (e.g., ENSO), but the variations will not match the timing of those in the historical record. The mean ocean state and the statistics of climate variations, however, do respond to historical greenhouse gases, land use and aerosol changes. The mean of an ensemble of historical climate change simulations averages out the variability, and isolates climate change-forced shifts in mean conditions. This is generally referred to as the "forced" signal.

Accessible reviews and further details of this terminology can be found in Stock et al., (2011) (Stock, C.A., Alexander, M.A., Bond, N.A., Brander, K.M., Cheung, W.W., Curchitser, E.N., Delworth, T.L., Dunne, J.P., Griffies, S.M., Haltuch, M.A. and Hare, J.A., 2011. On the use of IPCC-class models to assess the impact of climate on living marine resources. *Progress in Oceanography*, *88*(1-4), pp.1-27.) and Tommasi et al., (2017). While we cannot cover these topics in similar detail in this paper, we have expanded our methods considerably and added a summary table to provide the reader with sufficient information to read and understand the paper without extensive review of additional documents.

Thus, "SPEAR_HIST" is an ensemble of simulations running from pre-industrial times to 2012 that is forced by historical changes in greenhouse gases, aerosols, land use and insolation. As is customary, we enlisted corresponding ensemble members from the CMIP6 projection period to extend these to 2034. We chose SSP8.5 scenario for this purpose, as differences between scenarios are relatively small early in the 21st century. The HIST runs were spawned from an arbitrary pre-industrial year. No attempt was made to initialize the run with the precise ocean state in 1850. The ensemble mean of these runs can thus be used to approximate the signals forced by changing greenhouse gases, aerosols, land use and solar forcing (i.e., the "forced signal").

The "NWA12_HIST" ensemble is simply the downscaled version of the "SPEAR_HIST" ensemble for the years 1960-2034. The mean of the NWA12_HIST ensemble thus estimates the forced signal associated with climate change.

These points are now clarified in Lines 256-266 of the Methods and in Table 1.

Figures

Reviewing the paper was made challenging because many of the figures are difficult to read. The main problem is that the font sizes are too small. Though it is possible to use a computer to zoom in to the figures and see everything, one should be able to print out the paper (which is what I did) and have it be readable. Take Figure 1 as an example. The font sizes on all of the tick mark labels and legends are very small. I looked to see if EGU has any guidelines on font size in figures but could find none. I saw that AGU has a lower limit of 8 point, with 6 point for subscripts:

https://onlinelibrary.wiley.com/page/journal/21699402/homepage/graphics.htm

I doubt Figure 1 satisfies these reasonable criteria. All other figures are also difficult to read, with the exception of Figures 3 and S3–S6, which are very legible.

Also the map in Figure 1a is extremely small, maybe just slightly bigger than a postage stamp. It is also difficult to clearly see on panels 1b, 1d, and 1e when various events described in the text occur because these panels do not have horizontal axes just below them. A solution to this would be to include horizontal axes or, maybe better, to have thin vertical dashed lines on the years 1990, 2000, 2010, and 2020, which would have the advantage of more easily relating changes in one variable to another.

We have introduced the region and other aspects in a new bigger Figure 1. We have also added horizontal axes labels to the other panels in what is now Fig. 2.

There is too much information on Figure 4. Here are some suggestions to make the figure more understandable (in addition to the font size recommendation made earlier): (1) In the caption, note that the solid black lines (NWA12_EN4_NUDG) are the same as the lines in Figure 1b, d, e, and f; (2) do not show individual ensemble members but rather some measure of the ensemble spread (e.g., maximum and minimum values or interquartile range); (3) remove overlapping hindcasts (for example, remove hindcasts starting in 2004 and 2022).

Thank you for these suggestions, we have revised Figure 6 which now shows the min-max envelope around the ensemble mean rather than individual members.

Results

It is encouraging that the retrospective simulations capture bottom temperature and salinity variability (lines 139–174, Figure 1b and c), but it is not clear how much of this signal comes from the nudging the full water column to T and S observations where the bottom depth is greater than 1000 m. Some discussion is needed.

This is a good question. The need to nudge the interior model solution towards EN4 T, S in NWA12_EN4_NUDG was borne out from our preliminary tests which showed that open ocean boundary conditions (particularly at the northern boundary) from SPEAR_LO introduced some unrealistic variability in the subsurface ocean for a few years in the middle of the simulation time period which was propagated onto the shelf in the regional model. This resulted in a lower quality of regional simulations and thus necessitated nudging. While not central to the analysis, we have revised this section to include more discussion on the use of nudging (Lines 208-225). We also note that the bottom waters of the Gulf of Maine analysed here (the deep basins and the Northeast Channel) lie well inshore of the 1000 m isobath and are therefore not directly nudged towards EN4. The relaxation constrains only the off-shelf slope source waters. Their signal then enters the Gulf through the Northeast Channel and is carried onto the shelf by the model’s own dynamics. The evaluated bottom temperature and salinity variability in the Gulf is thus an emergent model response to better-constrained boundary source waters rather than a direct imprint of the nudging.

For bottom oxygen in the retrospective simulations, the authors point out that the variations result from different water masses (advection, effectively), solubility, and biological oxygen demand (lines 179–182). Instead of this very general statement, I recommend that the authors attempt to distinguish causes of these variations by computing the saturation concentration ([O2]sat, a function of potential temperature and salinity) and the difference [O2] – [O2]sat. So [O2] = [O2]sat + [O2]other, where [O2]other represents any processes that drive [O2] away from saturation, which one expects to be dominated by respiration for bottom waters because surface waters are generally close to saturation. [O2]other is basically the negative of the apparent oxygen utilization (AOU = [O2]sat – [O2]).

We have now added a new Figure 4 which shows the $[O_2]$, $[O_2]$ sat and AOU. The AOU accounts for approximately 70% of the interannual variance of the bottom-O_2 anomaly ($r = 0.89$ with total O_2), compared with about 21% for the solubility term. The dominant O_2 variations (including the episodic maxima of 2004, 2015 and 2019) are therefore driven by ventilation of the deep basins through deep winter mixing (Fig. S5) and by differences in the AOU of the inflowing slope waters, rather than by temperature and salinity-dependent solubility alone; solubility instead imparts a smaller, smoother decline as the bottom water warms.

Similarly, the authors should attempt to diagnose causes of variability in pH and CaCO3 saturation state (omega). The authors claim that the overall declining trend in these variables is consistent with the increase in atmospheric CO2 (lines 197–199) but they cannot say if that is the case quantitatively without further analysis. To understand pH and omega variations, it is best to first think about changes in the master variables DIC and alkalinity (Alk) because (unlike pH and omega), they will not change with temperature and salinity, will mix conservatively, and will also respond in straightforward, quantitatively consistent ways to processes such as gas exchange and respiration. pH and omega will decrease as DIC increases and as Alk decreases, which can be seen using equations of the full carbonate system or with some approximations.

For pH, it is best to think in terms of the hydrogen ion concentration, which, from the second dissociation step of CO2, is given by K2[HCO3]/[CO3], where HCO3 and CO3 are shorthand for the bicarbonate and carbonate ions. Then the approximations HCO3 = 2DIC – Alk and CO3 = Alk – DIC, so you can see how H+ increases (pH decreases) as DIC increases and Alk decreases. You can also see how H+ might vary with T and S due to variations in K2. So if the decline in pH is really due to anthropogenic CO2, then DIC should be going up roughly in line with expectations from the Revelle factor, and that DIC increase should be of a magnitude to cause the pH decline.

The difference between Alk and DIC can be particularly illuminating for understanding acidification metrics, such as pH and omega. I suggest the authors look at the following paper and use its ideas for understanding pH and omega variability: Xue, L., & Cai, W.-J. (2020). Total alkalinity minus dissolved inorganic carbon as a proxy for deciphering ocean acidification mechanisms. Marine Chemistry, 222, 103791.

A similar approach should be used for omega, which is equal to [CO3][Ca]/Ksp, where [Ca] is the calcium ion concentration, which is roughly proportional to salinity except in very low salinity (S < 5 psu) water, and Ksp is the solubility product, which depends on temperature (mainly) but also salinity and pressure. The authors are arguing that omega is decreasing due to an increase in anthropogenic DIC (impacting CO3), but omega could also be changing due to changes in Ca and Ksp, or changes in CO3 resulting from other processes influencing DIC and Alk (advection, respiration, CaCO3 dissolution, etc.).

Some effort is made in Figure 2 to quantify the impact of natural variability on omega, which is well done. However, the causes of the long-term declines in pH and omega remain unclear. I see that the historical simulations in Figure 4 are used to investigate causes of long-term change. This seems backwards to me. Before removing the long-term trend in Figure 2, the reader should first see the long-term trend itself.

Thank you for the substantive feedback. We have taken several steps to address your concerns. First, we note that our decision to focus on pH and Ω_{ar} was driven by their high relevance to living marine resources. As such, there is high interest in understanding the predictability of these quantities directly, including all their non-linear dependencies on ocean conditions.

Second, the statement on lines 197-199 was only intended to qualitatively establish that the downward trends in pH and Ω_{ar} are consistent with expectations from increasing CO_2 before further analysis of the contributions of the forced change and variability. We have now made this clear in the text.

Third, given the length of our simulations, we were forced to be very selective with the quantities saved for our analysis and did not save DIC and Alkalinity at the bottom. However, we had sufficient bottom information to reconstruct the contribution and estimate the roles of DIC, Alk, temperature and salinity on inter-annual variations and trends (see methods). In the revised manuscript, the causes of the long-term declines and variations are now diagnosed quantitatively in the new Fig. 4, which decomposes the pH and Ω_{ar} anomalies into contributions from DIC, alkalinity, temperature and salinity using the retrospective (NWA12_EN4_NUDG) simulation. The long-term trend and its attribution are therefore established before the impact of the trend on prediction skill is interpreted. These changes are described in more detail in response to your specific comments below.

The authors state in line 199–201 that omega varies more than pH but I do not see how one can make a statement like that when comparing two different variables. If there were two time series of the same variable, then one could simply compare their standard deviations. Such a comparison for pH and omega seems flawed.

The statement is based on the size of the variations around the mean trend in Fig. 2 relative to the trend. We have made this statement more precise (Lines 488-491). Our statement that omega has more interannual variability than pH in the Gulf of Maine is consistent with Salisbury and Jonsson, 2018, who wrote:

"[compared to pH,] The decadal variability is more pronounced for omega_AR, which is dominated by the variation in TA and combined SST and salinity, with less dependence on the OA signal."

We have added a citation to Salisbury and Jonsson, 2018 after our statement (Lines 488-491). The sentences following it in the manuscript elaborate on how the results are consistent with Salisbury and Jonsson.

The result that omega and temperature are correlated on interannual time scales is a nice finding (Figure 2). The explanation of the relationship (lines 203–208), however, needs to be fleshed out. The authors talk about the response of the carbonate chemistry to warming and the increased alkalinity associated with WSW. These effects should be separated. Further, WSW probably has high DIC as well, which would work in the opposite direction on omega. This is why I would emphasize discussion of DIC and Alk in order to explain omega. The authors go on to make connections to pH (lines 207–209), but there is no solid basis for the explanations.

We have now added a new Figure 4 in which we attributed the bottom pH and Ω_{ar} anomalies to DIC, Alk, temperature and salinity (see Methods in the revised manuscript). Following your suggestion, we adopt the alkalinity-minus-DIC (Alk − DIC) framework of Xue and Cai (2020) to interpret the acidification metrics; Alk − DIC is shown in Fig. 4b. In the model the Alk − DIC anomaly is nearly identical to the bottom Ω_{ar} anomaly (r = 0.996; Fig. 4b, d) but tracks pH relatively loosely (r = 0.82, because pH also responds directly to warming). On interannual timescales Alk − DIC (and hence Ω_{ar}) is set by water-mass alkalinity changes, while its long-term decline (−2.2 μmol kg^{-1} $decade^{-1}$), rising DIC outpacing a smaller alkalinity increase, is the anthropogenic acidification trend. We have added a sentence to the results section stating this (Lines 518-524).

To further address the causes of the long-term declines (comments above) and your specific concern about DIC, we now report the total-change attribution for both variables. The pH decline (−0.013 $decade^{-1}$) is driven by rising DIC (−0.010) and warming (−0.007), partially buffered by increasing alkalinity (+0.005). The Ω_{ar} decline (−0.015 $decade^{-1}$) is likewise led by DIC (−0.035 $decade^{-1}$), but increasing alkalinity (+0.016) and warming (+0.004) together offset about 58% of that DIC-driven decline. On interannual timescales alkalinity dominates the Ω_{ar} variability (about 57% of the detrended variance) while the direct thermodynamic temperature effect is minor, so the Ω_{ar}–temperature correlation primarily reflects the covariance of temperature and alkalinity through the WSW/LSW variations.

Finally, on your specific point that WSW might carry high DIC and oppose the alkalinity effect: in the model it does not. The detrended DIC and alkalinity anomalies are weakly anti-correlated (r = −0.28) and the overall variations in DIC are small. The small interannual DIC anomaly reinforces rather than opposes the alkalinity-driven variability. These results are now in the Results text and Fig. 4c,d. These small DIC variations are consistent with published work suggesting that WSW and LSW have similar DIC concentrations and differ primarily in their alkalinity (e.g., Stewart et al., 2025).

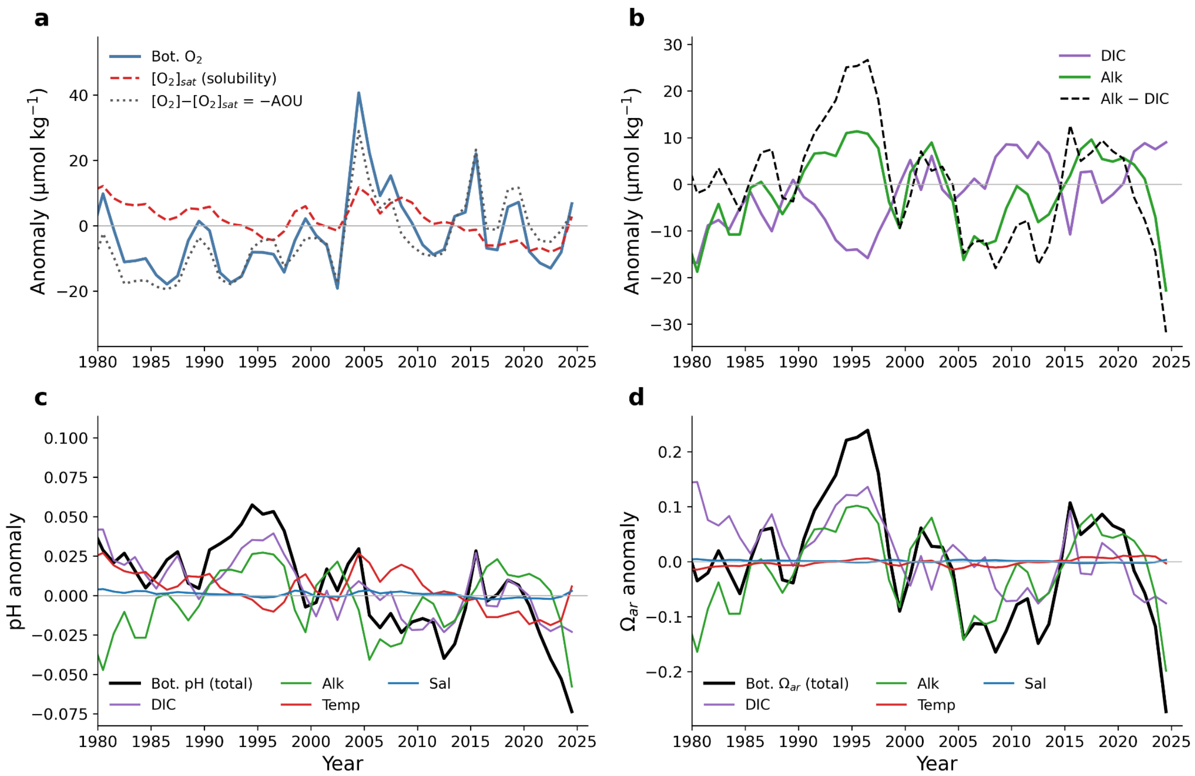


Figure 4: Decomposition of bottom-water biogeochemical variability in the Gulf of Maine (NWA12_EN4_NUDG retrospective simulation, ensemble mean; anomalies relative to 1994–2023). (a) Bottom-O_2 anomaly partitioned into the solubility term $[O_2]$ sat and the non-solubility residual $[O_2] - [O_2]$ sat (≈ −AOU). (b) Bottom dissolved inorganic carbon (DIC), total alkalinity (Alk) and Alk − DIC. (c, d) Attribution of the bottom (c) pH and (d) Ωar anomalies to DIC, Alk, temperature and salinity, obtained by perturbing each driver about the 1994–2023 mean state; the black line is the total anomaly (see Methods).

Figure 3 is a helpful figure. What's missing, however, is the salinity plot. Showing the zero-correlation line would also be helpful. In general, salinity seems to be dropped as a variable to discuss after Figure 1. Maybe the authors have a good reason for this, but I could not find one in the paper.

We did not add a separate salinity panel to Figure 3 because bottom salinity co-varies strongly with bottom temperature ($r = 0.84$; Fig. 2a,b, Fig. S3), so such a panel would largely duplicate the temperature panel. Salinity is not dropped from the analysis: it is retained as an explicit driver in the new Fig. 4 decomposition of the pH and Ω_{ar} anomalies (Fig. 4c,d), where its direct contribution is found to be small, and we have added a sentence to the results section making the role of salinity explicit.

Figure 4 is interesting but very difficult to read, as noted above. The authors might consider adding a salinity panel. I am curious to know if there is some long-term trend due to climate change; I would expect freshening, given enhanced precipitation, runoff, and ice melt at high latitudes, the signal of which might be advected southward in the Labrador current.

We have now improved how Figure 6 shows the retrospective predictions. Regarding a salinity panel on the hindcast figure (now Fig. 6): bottom salinity closely tracks bottom temperature (Fig. S3, $r = 0.84$), so we did not add a redundant panel, but its variability is shown in Fig. 2b and its contribution to the carbonate anomalies in Fig. 4c, d. On the long-term trend: in the deep Gulf of Maine, bottom salinity is dominated by the Warm Slope Water / Labrador Slope Water balance rather than by a monotonic high-latitude freshening signal. During the 2004–2020 warming the bottom water became warmer and saltier (increasing WSW influence), whereas the recent LSW resurgence brings cooler, fresher water. Any high-latitude freshening advected by the Labrador Current is therefore superimposed on, and over recent decades largely masked by, this water-mass variability.

Figure 5 is helpful to show the skill of short-term predictions.

Minor comments

Typo: Should be "skillful" in line 24

Corrected.

Line 112: Koul et al. (2024) is cited as predicting biogeochemical variability, but this paper does not include biogeochemistry.

Corrected.

Figures S1 and S2. I recommend that the titles of these panels read "Mid-Atlantic Bight," "Gulf of Maine," etc. so that readers can easily understand what is being shown.

We have edited figures to improve visibility.

Line 434: It seems that Figure S1 should also be cited here.

Cited.