Multi-Year Predictability of Hydrography and Circulation on the U.S. Northeast Shelf: A Dynamical Downscaling Perspective

Yiming Guo[1], Alma Carolina Castillo-Trujillo[2,3], Ke Chen[2], Young-Oh Kwon[2], Sydney Perkins[4], Hyodae Seo[5], Paula Fratantoni[6], Michael Alexander[7], and Vincent Saba[8]

[1]Department of Geography, Geology and the Environment, Illinois State University, Normal, Illinois
[2]Woods Hole Oceanographic Institution, Woods Hole, Massachusetts
[3]atDepth Inc.
[4]University of California, Berkeley, California
[5]University of Hawai'i, Honolulu, HI, USA
[6]Northeast Fisheries Science Center, NOAA NMFS, Woods Hole, MA, USA
[7]Department of Atmospheric and Oceanic Sciences, University of Colorado, Boulder, CO, USA
[8]NOAA Northeast Fisheries Science Center, Geophysical Fluid Dynamics Laboratory, Princeton University, Princeton, NJ, USA

Correspondence: Yiming Guo (yguo10@ilstu.edu)

Abstract. The U.S. Northeast Shelf (NES) is a dynamic and economically important marine ecosystem where temperature and salinity variability are shaped by interactions among large-scale climate variability, Gulf Stream shifts, mesoscale eddies, and local shelf processes. Predicting these variations on multi-year timescales remains a major challenge for current climate systems, as global models at typically 1-2° resolution exhibit poor skills for the NES. Here, we evaluate a high-resolution regional prediction of NES based on the downscaling of global Community Earth System Model Decadal Prediction Large Ensemble (CESM-DPLE) using the Regional Ocean Modeling System (ROMS-DOWN) to assess its potential for improving interannual-to-decadal prediction skill on the NES. Compared to CESM-DPLE, ROMS-DOWN substantially reduces mean-state biases in temperature, salinity, sea surface height, and upper-ocean heat content across the shelf and slope, where bathymetry effect and shelf-slope exchange are critical but poorly resolved in global models. Both deterministic and probabilistic metrics indicate improved forecast performance with lead time up to 5 years. The predictive skill reflects contributions from externally forced trends and interannual-to-decadal internal variability, with dominant timescales of predictability differing among variables. ROMS-DOWN also skillfully reproduces key shelf features such as the Middle Atlantic Bight cold pool and slope-water mixing characteristics in the Gulf of Maine, though their predictability remains moderate likely due to internal variability, boundary condition biases, and model uncertainties. Overall, these results demonstrate that dynamical downscaling can effectively bridge large-scale climate predictability and regional coastal processes, providing a foundation for improved multi-year prediction and understanding of ocean variability on the NES.

1 Introduction

The U.S. Northeast Shelf (NES) Large Marine Ecosystem is one of the most dynamic and socioeconomically important marine systems in the world, supporting valuable fisheries and coastal communities (Link et al., 2002; Lucey and Nye, 2010; Gartland et al., 2023). Over recent decades, the NES has experienced substantial changes in its hydrography, circulation, and ecosystem structure in response to natural climate variability and anthropogenic forcing (Claret et al., 2018; Chen et al., 2020; Gonçalves Neto et al., 2021; Guo et al., 2022a; Friedland et al., 2024). These physical changes strongly influence biological productivity and fish distributions, creating challenges for resource management and climate adaptation (Pershing et al., 2015; McHenry et al., 2019; Friedland et al., 2025). Developing reliable predictions of the physical environment on seasonal to multi-year timescales is therefore a critical step toward supporting climate-informed fisheries management and improving ecosystem forecasts for the NES.

Skillful predictions for the NES is challenging due to the region's complex dynamics and multi-scale processes. The NES lies at the confluence zone of contrasting subpolar and subtropical water masses, where cold, fresh waters advected southward by the Labrador Current meet warm, saline waters carried northward by the Gulf Stream and slope currents (Loder et al., 1998; Chapman and Beardsley, 1989; Fratantoni and Pickart, 2007; Kwon et al., 2010). This intersection, together with strong interactions between the shelf and open ocean through multiscale dynamic processes and frequent atmospheric disturbances along the mid-latitude storm track, drives large variability on timescales from subseasonal to decadal (Kelly et al., 2010; Chen et al., 2014a; Guo et al., 2023). Standard-resolution global prediction models with 1° horizontal resolution struggle to capture these variabilities because they poorly resolve Gulf Stream mean position and variability, shelf–slope exchange, coastal currents, and fine-scale bathymetry, all of which are critical to reproducing NES water characteristics and their variability (e.g., Saba et al., 2016; Jacox et al., 2020; Chen et al., 2022).

At the same time, observational and modeling studies suggest that variations in water properties on the NES may be influenced by predictable large-scale signals. Variability in the North Atlantic Oscillation (NAO), Atlantic Meridional Overturning Circulation (AMOC), and Gulf Stream path have all been linked to temperature and salinity changes on the NES with time lags of 1–5 years (Mountain, 2012; Xu et al., 2015; Saba et al., 2016; Gonçalves Neto et al., 2021; Karmalkar and Horton, 2021). For instance, Labrador Slope Water intrusions into the Gulf of Maine are correlated with AMOC variability 1 to 2 years earlier (Saba et al., 2016), while Gulf Stream path shifts have been shown to modulate shelf temperature and fish distributions (Nye et al., 2011; Davis et al., 2017; Gonçalves Neto et al., 2021). This implies that multi-year predictability over the NES may be achievable if large-scale oceanic and/or atmospheric signals are effectively transmitted to the coastal environment through enhanced model resolution and better representation of local dynamics.

Recent advances in global climate prediction systems have demonstrated that skillful forecasts on seasonal to decadal timescales are increasingly achievable, particularly in regions where large-scale ocean dynamics provide memory to the climate system (Meehl et al., 2009; Yeager et al., 2018; Christensen et al., 2020; Yeager et al., 2022). In the North Atlantic, initialized decadal prediction systems—including the CESM Decadal Prediction Large Ensemble (CESM-DPLE)—have shown notable skill in predicting variability in the subpolar gyre, AMOC, Labrador Sea convection, and basin-scale SST patterns several

years in advance (Yeager et al., 2018; Athanasiadis et al., 2020). These forecasts capture coherent low-frequency ocean signals linked to heat content anomalies, buoyancy forcing, and large-scale circulation changes, underscoring the potential for predictable pathways that can influence downstream regions such as the NES. The demonstrated skill of CESM-DPLE in the broader North Atlantic provides a strong foundation for regional downscaling: if large-scale predictable signals are reproduced in the global system, a high-resolution regional model may translate that information into improved predictability of coastal ocean properties.

Building on these advances in large-scale prediction, dynamical downscaling provides a pathway to bridge large-scale climate signals with regional-scale ocean processes by embedding a high-resolution regional model within global forecast systems. Previous downscaling studies have demonstrated substantial improvement in the representation of NES hydrography using dynamical downscaling at both seasonal (Ross et al., 2024) and decadal timescales (Koul et al., 2024). However, these approaches target distinct prediction horizons, and the extent to which dynamical downscaling enhances forecast skill on intermediate (multi-year) timescales remains less well understood. This timescale is particularly relevant for living marine resource management, where decisions often require outlooks beyond seasonal forecasts and are more actionable than decadal projections.

In this study, we evaluate the improvement in multi-year prediction skill ~~achieved through dynamical downscaling on~~ 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. By quantifying the improvements and limitations of dynamical downscaling, we aim to achieve a better understanding of multi-year predictability on the NES and to provide guidance for advancing climate-informed ocean forecasting and ecosystem management in the region.

2 Data and Methods

2.1 Observations and reanalysis

Two satellite-based observational datasets are used to evaluate mean-state fidelity of both the reanalysis dataset and the model outputs on the NES: one for sea surface temperature (SST) and the other for sea surface height (SSH). For SST, we use the NOAA Optimum Interpolation Sea Surface Temperature (OISST; Reynolds et al. (2007); Huang et al. (2021)), which provides high-resolution daily fields based on blended satellite and *in situ* observations. For SSH, we use the AVISO (Archiving, Validation, and Interpretation of Satellite Oceanographic) gridded altimetry product archived from the data portal Copernicus Marine Environment Monitoring Service (CMEMS), which merges measurements from multiple satellite missions to generate consistent maps of sea level anomalies relative to a long-term mean. Both datasets are provided at a horizontal resolution of approximately 25 km and are used exclusively to validate the climatological mean distributions across the model domain, thereby providing a reference for the mean pattern of ocean circulation and surface water properties on the NES.

To evaluate multi-year prediction skill, we employ the GLORYS ocean reanalysis (Global Ocean Reanalysis and Simulation) version 12V1, produced by Mercator Ocean ~~. GLORYS is a global, eddy-permitting reanalysis available at~~ and distributed through the CMEMS (Jean-Michel et al., 2021). The CMEMS product is provided on a regular $1/12^\circ$ ($\sim$8 ~~-~~km in the study area) ~~horizontal resolution~~ grid with 50 ~~vertical levels, spanning~~ standard vertical levels and spans the satellite altimetry era from 1993 to ~~present (Jean-Michel et al., 2021)~~the present (Lellouche et al., 2021). It assimilates a wide range of observations, including satellite-derived SST and SSH, sea-ice concentration, and in situ profiles of temperature and salinity from ARGO floats, and ship-based measurements. Independent evaluations have shown, through comprehensive comparisons with available observations, that GLORYS is one of the best global reanalysis products in representing ~~circulation features and water mass properties in~~ the climatological hydrography, circulation, and their temporal and spatial variability over 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. One limitation of GLORYS is that it does not explicitly include tidal forcing, which can contribute to residual temperature and salinity biases in strongly tidal regions such as Georges Bank and parts of the Gulf of Maine (Castillo-Trujillo et al., 2023). However, in this study prediction skill is evaluated using annual-mean anomalies with a lead-time-dependent climatology removed, which reduces sensitivity to mean-state biases and high-frequency variability associated with tidal processes. In addition, ~~its~~ GLORYS's representation of shallow coastal waters is limited, which may introduce additional uncertainties in nearshore regions.

2.2 CESM Decadal Prediction Large Ensemble (CESM-DPLE)

The global simulations used in this study are based on the Community Earth System Model version 1.1 (CESM1.1), a fully coupled global climate model that integrates atmosphere, ocean, sea ice, and land components through a central flux coupler (Hurrell et al., 2013). The atmospheric component employs the Community Atmosphere Model version 5 (CAM5) at approximately $1^\circ \times 1^\circ$ horizontal resolution with 30 vertical levels (Park et al., 2014), while the ocean component uses the Parallel Ocean Program version 2 (POP2) at a comparable $1^\circ \times 1^\circ$ resolution with 60 vertical levels (Smith et al., 2010). Sea ice is simulated with the Community Ice Code version 4 (CICE4) on the ocean grid (Hunke et al., 2010), and land processes are represented by Community Land Model version 4 (CLM4) on the atmospheric grid (Lawrence et al., 2011).

To provide large-scale lateral boundary condition and surface forcing fields as well as the initial condition for the dynamic downscaling experiments, and to compare skill directly with a global prediction system, we use the CESM Decadal Prediction Large Ensemble (CESM-DPLE; Yeager et al. (2018)). CESM-DPLE is a global prediction system, which consists of 40 ensemble members, each initialized every November 1st from 1954 to 2015, and integrated for 10 years. The system uses CESM1.1 with nominal 1° horizontal resolution in both the ocean and atmosphere. 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 Ctr 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.

CESM-LE is a fully coupled companion ensemble of simulations designed to represent the response of the climate system to external forcing (e.g., greenhouse gases, aerosols, volcanic eruptions) in the absence of initialization from observations.

Each member is generated by applying small perturbations to the atmospheric initial state. Because CESM-LE is uninitialized, each member exhibits year-to-year ocean internal variability such that the temporal evolution is not necessarily consistent with observations. Instead, the ensemble mean captures the externally forced component of variability. As such, CESM-LE serves two important roles in this work: (1) it provides a baseline for assessing how much of the forecast skill in CESM-DPLE and ROMS-DOWN arises from external forcing versus initialization, and (2) it illustrates the level of multi-year predictability expected solely from forced long-term trends or externally driven decadal variations.

CESM-DPLE has been widely used to assess decadal predictability of the Atlantic and global climate system, and has demonstrated robust skill in the subpolar North Atlantic (Yeager et al., 2018; Athanasiadis et al., 2020). In CESM-DPLE, high-frequency atmospheric output (daily or sub-daily) is archived for only a subset of ensemble members; the implications of this constraint for the regional downscaling design are described in Section 2.4.

2.3 ROMS Hindcast (ROMS-HIND)

To evaluate the performance of our regional ocean model configuration, we use a hindcast simulation (ROMS-HIND). ROMS-HIND is based on the Regional Ocean Modeling System (ROMS), a free-surface hydrostatic, primitive equations numerical model (Shchepetkin and McWilliams, 2005). The model domain encompasses the NES, adjacent slope waters and the Gulf Stream. It extends from 30°N to 55°N and from 80°W to 35°W (Figure 1). The northern boundary of ROMS-HIND and ROMS-DOWN is set within the subpolar gyre so that the boundary condition from CESM-DPLE may contain highly skillful prediction (Yeager et al., 2018). The model has a ~~variable~~ 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. ~~The minimum depth was set to~~ A minimum depth of 10 m ~~to~~ was imposed by modifying the bathymetry prior to the simulation to avoid wetting and drying of grid cells. At the open boundaries, Chapman boundary conditions were used for the free surface, Shchepetkin lateral boundary conditions for the depth-averaged momentum, and all other fields use radiation boundary conditions (Chapman, 1985; Mason et al., 2010).

~~A realistic topography is implemented in the domain~~ The model bathymetry is derived from the ETOPO1 Global Relief Model (Figure 1a-b). ~~We used data from the ETOPO1 Global Relief Model.~~ To reduce the pressure gradient errors due to the steep bathymetry, the bathymetry was smoothed with the Batteen and Miller (2009) method using a maximum roughness value of R_{xo}=0.25 and R_{x1}=7. Several sensitivity experiments were conducted in which R_{xo} was varied, producing different degrees of bathymetry smoothing and shifts in the shelf break position relative to the coast (Figure A1). The configuration presented here was selected because it yielded the most realistic Gulf Stream path—including separation near Cape Hatteras—while minimizing pressure gradient errors and maintaining a realistic shelf circulation.

Surface atmospheric forcing is obtained from the European Centre for Medium-Range Weather Forecasts (ECMWF) Reanalysis v5 (ERA5) (Hersbach et al., 2020). The ERA5 analysis provides hourly estimates for a large number of atmospheric, ocean-wave, and land-surface quantities. ERA5 combines vast amounts of historical observations into global estimates using advanced modeling and data assimilation systems. ROMS-HIND is forced with hourly data including zonal and meridional wind, shortwave and longwave radiation, cloud, atmospheric pressure, rain, and humidity via COARE 3.5 bulk flux formula

(Edson et al., 2013). The initial and open boundary ocean conditions are produced from daily 1/12°GLORYS12v1 output. The model is forced with monthly ~~climatology river transport and temperature produced from 1992 to 2018~~ climatological river discharge and temperature at 21 river mouths along the NES~~(Dai, 2021)~~, derived from the global river reanalysis dataset of Dai (2021). Tidal forcing is included in the model ~~, with only the dominant tidal harmonics~~ using the dominant semidiurnal tidal constituents (M2 ~~,~~ and S2) ~~added~~ from the TPXO09 global tidal model (Egbert and Erofeeva, 2002), providing the primary tidal mixing necessary to realistically represent the circulation over the Shelf (Chen and He, 2015). The simulation was performed from January 2004 to December 2010 which overlaps with the satellite and reanalysis datasets used for validation and with the ROMS-DOWN ~~simulation~~ predictions described below.

2.4 ROMS Downscaling (ROMS-DOWN)

To perform dynamical downscaling for multi-year prediction on the NES, we employ a regional forecast system (ROMS-DOWN) based on the same ROMS configuration used in ROMS-HIND. Unlike ROMS-HIND, which is initialized and forced solely by reanalysis products, ROMS-DOWN is initialized and forced by anomalies from CESM-DPLE. The atmospheric forcing is constructed as the 1993–2024 ERA5 monthly climatology ~~for 1993-2024~~ plus daily CESM-DPLE anomaly fields~~computed by removing the mean values from~~ , where the daily anomalies are computed relative to a daily climatology obtained by interpolating the 1993~~to~~ –2010 ~~, while ocean initial condition~~ monthly climatology. Similarly, the ocean initial and lateral boundary conditions are ~~produced from~~ constructed from the 1993–2021 GLORYS monthly climatology ~~for 1993-2021~~ plus monthly CESM-DPLE ocean anomaly fields computed relative to the ~~1993-2010 mean.~~ 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. 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. River input is prescribed identically to ROMS-HIND using monthly climatological river transport and temperature at 21 river mouths along the NES. Because river discharge is prescribed as a repeating climatology, interannual freshwater variability is not included in ROMS-DOWN. This "anomaly downscaling" strategy reduces the impact of mean-state biases in CESM-DPLE while allowing large-scale signals to propagate into the high-resolution regional domain. ~~Note that the northern boundary of ROMS-HIND and ROMS-DOWN is set within the subpolar gyre so that the boundary condition from CESM-DPLE may contain highly skillful prediction (Yeager et al., 2018).~~

ROMS-DOWN consists of 20 start years (1992–2011), each initialized on January 1st with 10 ensemble members and integrated for 8 years, yielding a total of 200 eight-year forecasts. Our downscaling experimental design is constrained by the availability of high-frequency CESM-DPLE atmospheric output. The required daily (or 3-hourly) atmospheric forcing fields are archived only for a limited subset of CESM-DPLE ensemble members and initialization years; as a result, only 10 members provide the necessary forcing for our study period. This output constraint dictated the choice of a 10-member ROMS-DOWN. To assess whether the ensemble size of 10 members and the number of start dates of 20 are sufficient for NES multi-year prediction, sensitivity tests were conducted using the larger CESM-DPLE ensemble set (62 start years and 40

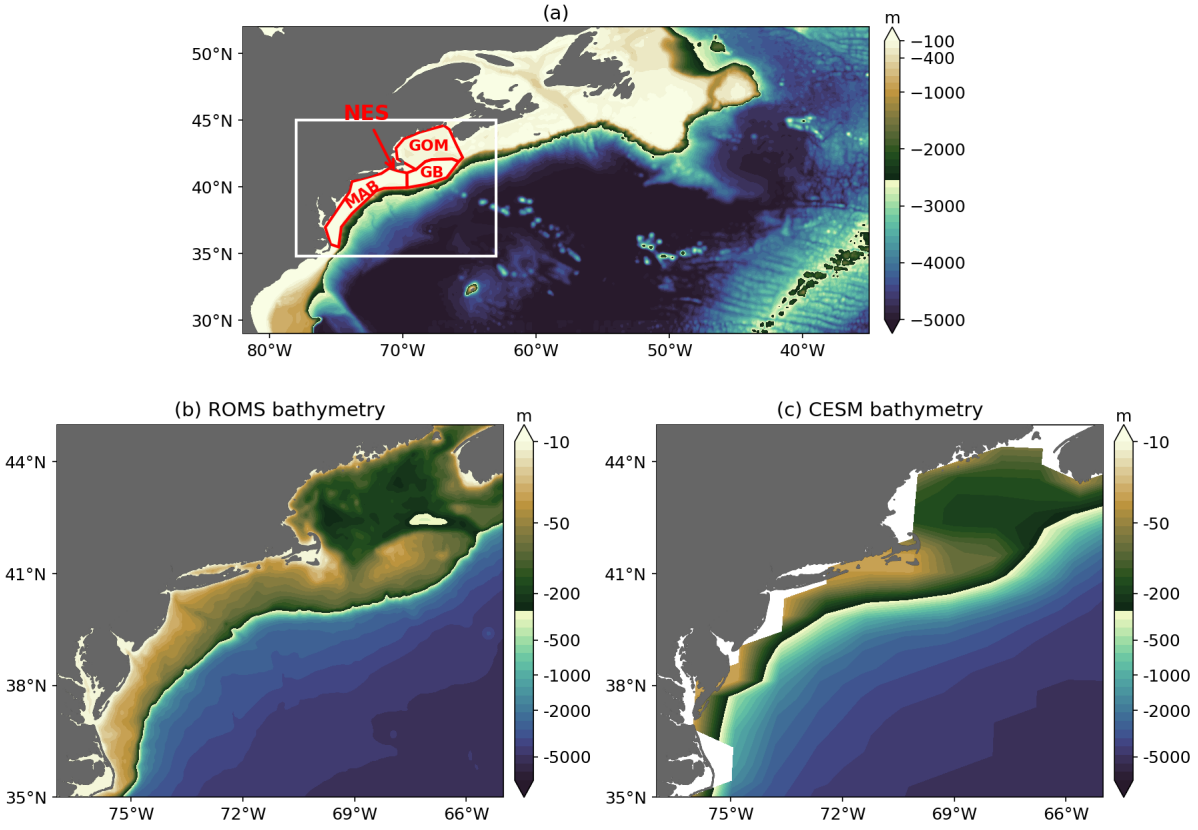


Figure 1. Study area. (a) ROMS model domain showing bathymetry~~and~~. The white box denotes the ~~three~~ analysis region shown in panel (b). (b) Bathymetry within the analysis region. The U.S. Northeast Shelf (NES) is defined as the shelf ~~provinces~~ region shallower than 200 m and consists of the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

ensemble members). The results suggest that 10 members and 20 start dates are adequate to capture the ensemble-mean forecast signal over the NES region (Figure A2), indicating that the ensemble design provides a sufficient sample to characterize forecast spread and uncertainty. By construction, ROMS-DOWN preserves the global predictability inherited from CESM-DPLE while resolving smaller-scale shelf processes that are absent in the coarse global simulations (Figure 1).

To ensure consistent comparisons, all datasets are bilinearly interpolated to the ROMS grid and evaluated using a common land-sea mask. For regional skill assessment (Section 2.5), metrics are aggregated over the entire NES~~and its subregions—the~~, 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.

2.5 Skill metrics

We evaluate prediction performance with two deterministic metrics—the Pearson anomaly correlation coefficient (ACC) and the mean square skill score (MSSS)—applied to ensemble-mean annual anomalies, and one probabilistic metric—the Brier Score (BrS)—applied to ensemble probabilities of tercile events following Hervieux et al. (2019). Skill is assessed using the GLORYS ocean reanalysis. 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.

For each variable, we focus on their anomalies based on annual mean and the ensemble mean for each prediction year, which is calculated as:

$$f_{i\tau} = \frac{1}{n_e}\sum_{j=1}^{n_e} f_{ij\tau}. \tag{1}$$

where f is the annual mean anomaly, $i = 1, \ldots, N$ indicates the model initialization years (N=20), $j = 1, \ldots, n_e$ indicates each ensemble member (n_e=10), and τ is the prediction lead year (LY1-LY8 in this work). 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.

To ~~remove the effect of model drift and isolate predictable variability, we subtract a lead-year-dependent climatology computed across all start years(i. e., independent of the start years) (Yeager et al., 2018, 2023),~~ 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). Specifically, the anomaly is defined as:

$$\hat{f}_{i\tau} = f_{i\tau} - \overline{f}_{\tau}, \qquad \overline{f}_{\tau} = \frac{1}{N}\sum_{i=1}^{N} f_{i\tau}, \tag{2}$$

Analogously, we sample the observations with 20 initial years and 8 lead years to be parallel to the prediction, with the same observation values repeated for different pairs of initial and lead years, e.g., 1995 observation is used for LY1 of 1995 initialization and LY2 of 1994 initialization. For the observations, $o_{i\tau}$, $\hat{o}_{i\tau} = o_{i\tau} - \overline{o}_{\tau}$. 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.

All deterministic skill metrics are computed from $(\hat{f}_{i\tau}, \hat{o}_{i\tau})$ using the N start-year pairs available at each lead years. For probabilistic metrics, we use the full ensemble to form event probabilities (Section 2.5.3) following Hervieux et al. (2019).

2.5.1 Pearson anomaly correlation coefficient (ACC)

ACC measures phase agreement between prediction and observed anomalies:

$$\mathrm{ACC}_\tau = \frac{\sum_{i=1}^{N} \hat{f}_{i\tau}\, \hat{o}_{i\tau}}{\sqrt{\sum_{i=1}^{N} \hat{f}_{i\tau}^2} \sqrt{\sum_{i=1}^{N} \hat{o}_{i\tau}^2}}. \tag{3}$$

Values range from -1 to 1, with positive values indicating that predictions track observed interannual variations.

2.5.2 Mean-square skill score (MSSS)

MSSS quantifies amplitude of error reduction relative to a climatological reference, defined here as the variance of the observed anomalies:

$$\mathrm{MSSS}_\tau = 1 - \frac{\sum_{i=1}^{N} \left(\hat{f}_{i\tau} - \hat{o}_{i\tau}\right)^2}{\sum_{i=1}^{N} \hat{o}_{i\tau}^2}. \tag{4}$$

Because the denominator equals the observed variance, $\mathrm{MSSS} > 0$ denotes improvement over climatology and $\mathrm{MSSS} < 0$ degradation. We report MSSS alongside ACC to jointly characterize phase and amplitude bias behavior.

2.5.3 Brier Score for tercile events

Following Hervieux et al. (2019), we assess prediction probability errors using the Brier Score (BrS). For each grid cell, lead year τ, and event definition E (upper or lower tercile of the observed anomaly distribution over the verification period), the ensemble prediction probability for start year i is

$$f_i = \frac{\#\{\text{members predicting } E\}}{n_e} \in [0,1], \tag{5}$$

The corresponding observation indicator is defined as $o_i \in \{0,1\}$, where $o_i = 1$ if the observation for year i falls in the specified event and $o_i = 0$ otherwise. The Brier Score averages the squared probabilistic error:

$$\mathrm{BrS}_\tau = \frac{1}{N} \sum_{i=1}^{N} \left(f_{i\tau} - o_{i\tau}\right)^2. \tag{6}$$

Lower BrS values indicate higher forecast reliability and resolution. Unless noted otherwise, we report BrS separately for upper- and lower-tercile events.

3 Results

3.1 Mean state and variability assessment

To assess how the regional and global prediction systems reproduce the mean spatial structure of key ocean variables on the NES, we show the 2004–2010 time-mean (and ensemble-mean) distributions of SST, SSH, sea surface salinity (SSS), bottom temperature (T_{bottom}), and 0–200 m heat content (HC_{200m}). Observational references are provided by OISST for SST and AVISO for SSH, with GLORYS used for all variables. Note that the GLORYS are compared against OISST and AVISO, while ROMS-HIND, ROMS-DOWN, and CESM-DPLE are compared against GLORYS for the quantification of biases using RMSE (Figure 2). Spatial maps of the mean bias (model minus GLORYS) for each variable are provided in Figure A3 to illustrate the spatial structure and sign of these errors. Across all variables, both ROMS-HIND and ROMS-DOWN show markedly improved mean structures and reduced mean biases compared to CESM-DPLE. For SST, CESM-DPLE exhibits a pronounced warm bias on the shelf lacking any sign of shelf break front, with a root-mean-square error (RMSE) of 3.35°C, primarily due to the overshooting Gulf Stream separation near Cape Hatteras (e.g., Chassignet and Marshall, 2008), whereas ROMS simulations better capture the sharp cross-shelf temperature gradient and Gulf Stream position, with RMSE values below 1.3°C. SSH patterns reveal that CESM-DPLE overestimates sea level gradient across the shelfbreak (RMSE = 0.22 m), while both ROMS configurations align more closely with AVISO and GLORYS (RMSE ∼0.08–0.09 m). Similarly, CESM-DPLE produces overly salty shelf waters (SSS RMSE = 1.89), whereas ROMS-HIND and ROMS-DOWN reproduce the fresher shelf and the offshore gradient (SSS RMSE ∼1.3–1.4). Improvements in the regional models are also evident in subsurface variables. For bottom temperature, CESM-DPLE lacks a realistic shelf structure due to its coarse bathymetry and simplified coastlines (Figure 1), resulting in overly warm and spatially smooth bottom fields (RMSE = 2.22°C), while the ROMS simulations capture the shelf patterns and the transition to slope waters by more realistically reproducing dynamical constraint imposed by the bottom topography. The 0-200 m heat content (HC_{200m}) bias is similarly reduced in the ROMS simulations (RMSE = 1.3–1.9 GJ m^{-2}) relative to CESM-DPLE (RMSE = 2.31 GJ m^{-2}).

To further quantify errors in temporal variability, the temporal ~~variability errors, Figure 3 shows the NES-wide and subregional average of temporal RMSE~~ RMSE is first calculated at each grid point ~~for the ROMS simulations and CESM-DPLE (~~using annual anomalies relative to GLORYS ~~), while spatial maps of the temporal RMSE are provided~~ 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. Consistent with the spatial mean biases (Figure 2), CESM-DPLE exhibits substantially larger temporal variation errors across all variables and subregions of the NES. For SST, errors exceed 6°C on average across the shelf, whereas ROMS-HIND and ROMS-DOWN reduce errors to less than 2°C in all subregions. For SSH, CESM-DPLE shows RMSE values greater than 27 cm, while both ROMS configurations remain below 8 cm, after removing the spatial means—defined as the average over the analysis region shown in Figure 2—from both simulated and observed fields to eliminate reference-level differences when deriving SSH. Salinity errors are also reduced in the ROMS simulations compared to CESM-DPLE. Performance improvements in ROMS are likewise evident in subsurface variables, including bottom temperature and upper 200 m heat content, where RMSE values are reduced by roughly half compared to

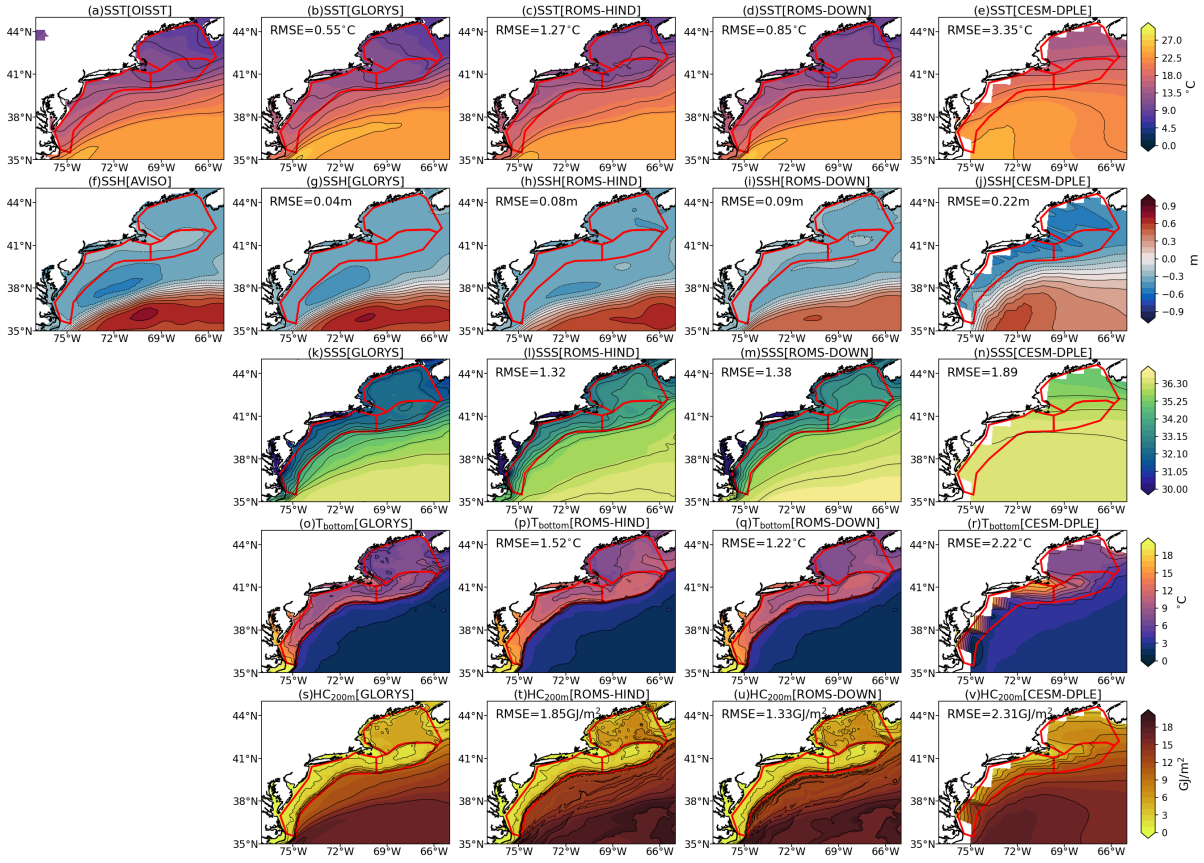


Figure 2. Time-mean (2004–2010) spatial distributions of (a–e) SST, (f–j) SSH, (k–n) SSS, (o–r) T_{bottom}, and (s–v) HC_{200m} from observational products and model simulations. The observational reference is shown in the first column, followed by GLORYS, ROMS-HIND, ROMS-DOWN, and CESM-DPLE. Spatial RMSE values relative to observations for GLORYS or relative to GLORYS for ROMS-HIND, ROMS-DOWN and CESM-DPLE are indicated in the upper-left corner of each panel. For ROMS-DOWN and CESM-DPLE, only lead year 1 (LY1) forecasts overlapping 2004–2010 are used, and results are shown as ensemble means based on 10 and 40 members, respectively. NES and its subregions are outlined in red.

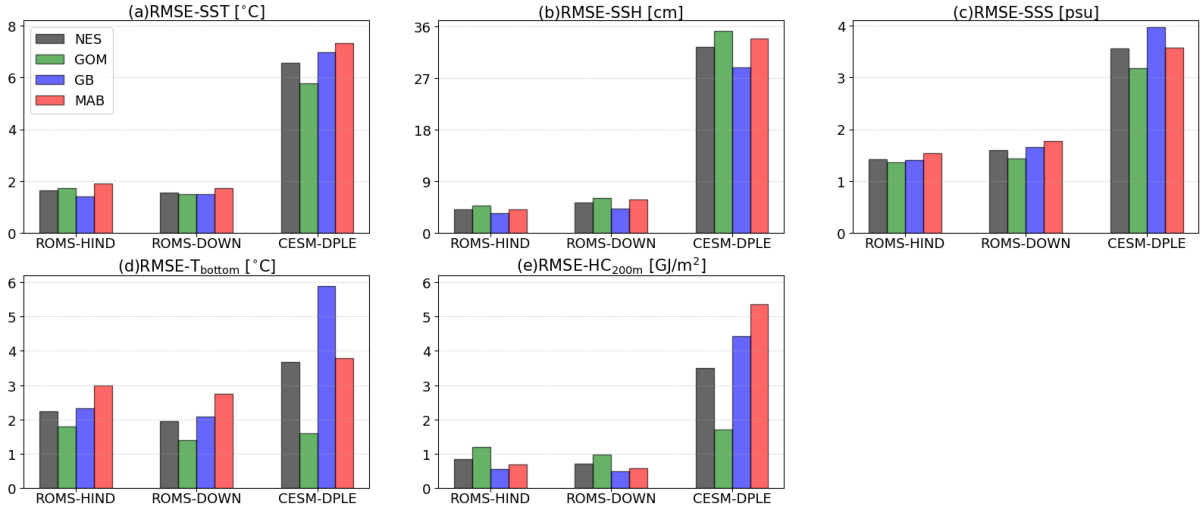


Figure 3. ~~Subregion-mean~~ Spatial average of the temporal RMSE (2004–2010; relative to GLORYS) ~~on the~~ calculated at each grid point within each shelf subregion for (a) SST, (b) SSH, (c) SSS, (d) T_{bottom}, and (e) HC_{200m} in ROMS-HIND, ROMS-DOWN, and CESM-DPLE. NES (gray bars) represents the entire U.S. Northeast Shelf. The 3 subregions are delineated in Figure 1a.

CESM-DPLE. One notable exception is the Gulf of Maine, where CESM-DPLE performs comparably to ROMS (Figure 3d,e). This may indicate that variability in these subsurface fields during this period is partly influenced by large-scale signals, such as NAO-related changes in basin-scale or regional water properties and circulations (Fratantoni and Pickart, 2007; Mountain, 2012), which the global model can capture to some degree despite the unrealistic topography due to its coarse horizontal resolution, although why CESM-DPLE performs as well as ROMS in this region is uncertain.

This model comparison shows that the high-resolution regional model reduces mean-state and variability biases in both surface and subsurface variables across the NES. These improvements are associated with a more skillful representation of shelf–slope bathymetry, slope currents, and Gulf Stream position, which are better represented in the high-resolution model framework (Gawarkiewicz et al., 2012; Saba et al., 2016; Chassignet and Xu, 2017; Guo et al., 2022b). In addition, the anomaly-downscaling framework helps maintain a realistic mean state while reducing large-scale climatological biases present in the global system, improving the representation of regional ocean conditions. 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.

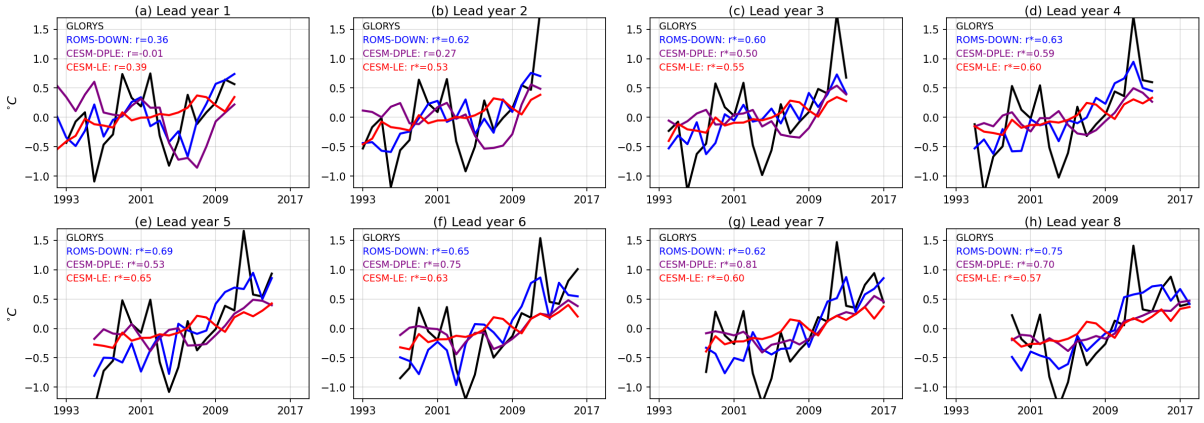


Figure 4. Time series of SST anomalies averaged over the entire NES shelf from GLORYS (black), ROMS-DOWN (blue), CESM-DPLE (purple), and CESM-LE (red) for lead years 1–8. Anomaly correlation coefficient skills (r) against GLORYS are indicated in the top left of each panel, with r* indicating significance at the 95% level based on a t-test with effective degrees of freedom accounting for autocorrelation.

3.2 Assessment of SST anomaly prediction skill

To examine the predictive skill of water properties in the NES, we first assessed the ACC for SST anomalies for lead years 1–8 (LY1–LY8) in ROMS-DOWN, CESM-DPLE, and CESM-LE. The time series for each lead year is constructed by stringing together the corresponding lead year prediction values from each start year simulation. The ensemble-mean annual SST anomaly time series averaged over the entire NES shelf (Figure 4) shows pronounced interannual and lower-frequency variability, including strong warming anomalies between the early 2000s and mid-2010s. ROMS-DOWN generally tracks the GLORYS anomalies more closely across all lead years, whereas CESM-DPLE ~~exhibits~~ tends to exhibit weaker amplitude and less agreement in the timing of extremes, particularly at ~~shorter leads~~the shorter lead times. Different from ROMS-DOWN and CESM-DPLE, the uninitialized CESM-LE captures much of the long-term warming signal but severely underestimates interannual and decadal variability, as expected by design.

Figure 5 shows the spatial distribution of ACC over the shelf, while Figure 6 summarizes the NES-wide and subregional (GOM, GB, and MAB) mean ACC as a function of lead year, computed from spatially averaged anomaly time series. ROMS-DOWN exhibits enhanced skill generally at early lead years, with widespread positive and statistically significant correlations across the shelf in LY2–LY5. Improvements are particularly notable in the GOM and MAB, where ROMS-DOWN consistently outperforms CESM-DPLE (Figure 6b,d). However, a consistent feature across both the regional and global prediction systems is the relatively weak ACC of SST at lead year 1 compared to longer forecast years. This reduced skill likely reflects a stronger interannual SST variability and the relatively smaller contribution of low-frequency signals during the forecast year-1 period (Figure 4a), and may also be related to initialization error in CESM-DPLE (Yeager et al., 2018). At longer lead times (LY6–LY8), CESM-DPLE retains greater skill, which is largely attributed to the increasing influence of externally forced low-frequency variability (Figure 4f-h). The uninitialized CESM-LE also shows surprisingly high correlations across much

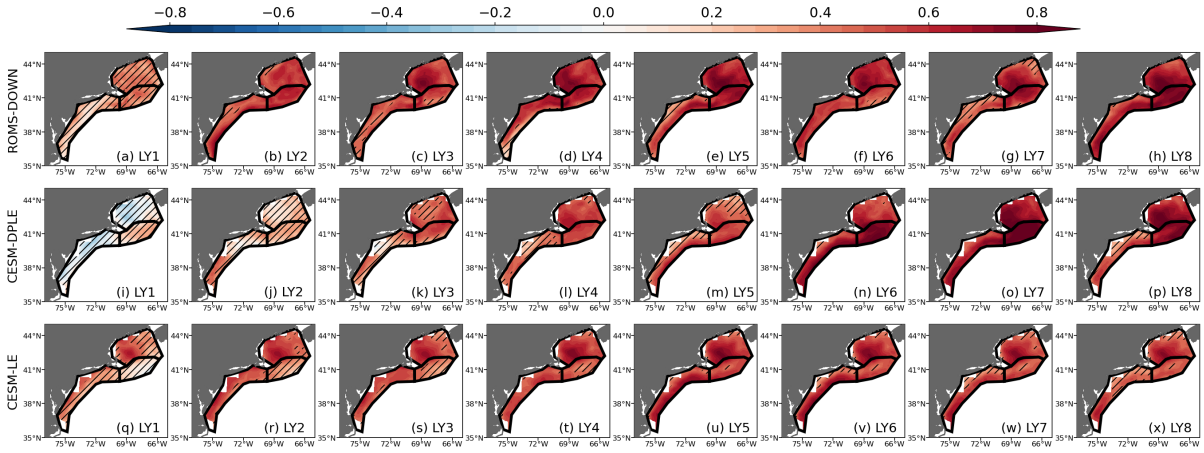


Figure 5. Anomaly correlation coefficients (ACC) for SST for lead years 1–8 (LY1–LY8) from (a–h) ROMS-DOWN, (i–p) CESM-DPLE, and (q–x) CESM-LE. ACC is computed relative to GLORYS reanalysis. Subregions outlined in black represent GOM, GB, and MAB, respectively. Hatched areas indicate regions where the ACC is not statistically significant at the 95% confidence level.

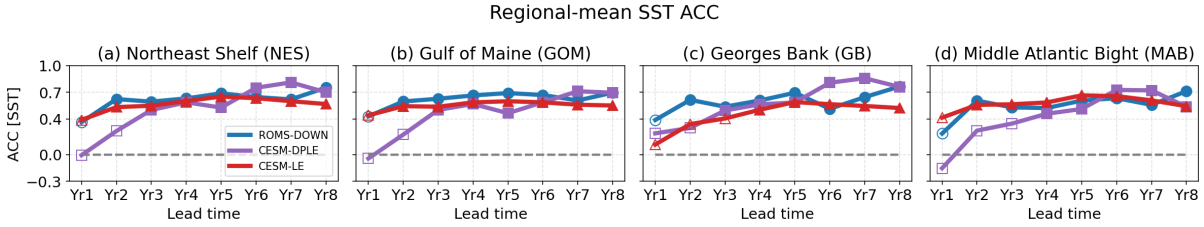


Figure 6. ACC of the subregion-mean SST as a function of lead year (LY1–LY8) for (a) NES, (b) GOM, (c) GB, and (d) MAB. ACC values are shown for ROMS-DOWN (blue), CESM-DPLE (purple), and CESM-LE (red). Filled markers denote statistically significant correlations at the 95% confidence level. Grey dashed line represents zero ACC skill.

of the shelf. This elevated skill arises because SST variability on the NES is strongly dominated by the externally forced warming trend (Chen et al., 2020); thus, CESM-LE—of which the ensemble mean contains primarily the forced response and limited internal variability—can still reproduce much of the observed low-frequency signal. As a result, CESM-LE provides an estimate of the predictable externally forced component that both CESM-DPLE and ROMS-DOWN inherit at longer lead times.

In addition to ACC, we also evaluate forecast performance using the ensemble-mean mean square skill score (MSSS), which measures error reduction relative to a climatological reference. The spatial distribution (Figure A5) and regional averages (Figure A6) of MSSS for SST show patterns broadly consistent with the ACC results, with generally positive skill across much of the NES except at the earliest lead years. ROMS-DOWN exhibits modest improvements over CESM-DPLE, particularly at shorter lead times, while skill at longer leads reflects the influence of low-frequency variability. Overall, the ACC and MSSS

scores suggest that much of the SST forecast skill on the NES is linked to predictable large-scale signals and long-term trends, while ROMS-DOWN improves the prediction skill by better representing shelf-scale variability and captures higher-frequency fluctuations, particularly at shorter leads.

Because the long-term warming trend dominates SST variability in the NES, both ACC and MSSS show relatively high scores in the coarse-resolution CESM-DPLE, with ROMS-DOWN providing modest improvements. 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. To evaluate this event probabilistic skill, we calculate the BrS for upper- and lower-tercile SST anomalies (Figure 7) for CESM-DPLE and ROMS-DOWN. Consistent with the deterministic metrics, ROMS-DOWN generally produces lower (i.e., better) BrS than CESM-DPLE across most of the NES, indicating higher reliability in predicting probability of extreme events.

At longer lead times (LY6–LY8), BrS improves as externally forced low-frequency variability becomes the dominant source of skill—especially for upper-tercile warm SST events in both ROMS-DOWN and CESM-DPLE—but ROMS-DOWN still outperforms CESM-DPLE. Recall that, in terms of ACC and MSSS, CESM-DPLE shows comparable or even higher values on these leads (Figures 6 and A6), largely because it captures the low-frequency warming trend. The SST anomaly time series also highlights pronounced interannual variations during these long lead years (e.g., strong cooling around 2004 and warming around 2012) superimposed on the trend. Although CESM-DPLE retains high ACC and MSSS on these leads, its failure to reproduce these interannual fluctuations is revealed by poorer Brier Score performance for tercile events (Figure 7). Furthermore, compared to warm events, CESM-DPLE produces systematically higher (i.e., poor) BrS for lower-tercile (cold) events than for upper-tercile (warm) events (Figure 7). This asymmetry suggests that much of the apparent event-level skill in CESM-DPLE arises from its ability to capture the externally forced warming trend, which favors the prediction of warm extremes. And cold events may be more strongly influenced by regional processes such as shelf–slope exchange, episodic intrusions of subpolar waters, and wind-driven vertical mixing and entrainment associated with mixed layer deepening. Because such processes depend on realistic bathymetry, vertical resolution, and mixing parameterizations, they are better represented in the high-resolution regional model than in the coarse-resolution global system. As a result, CESM-DPLE is less able to reproduce cold extremes, leading to poorer probabilistic skill (higher BrS) for lower-tercile events.

3.3 ~~Improved~~ Predictability of NES Water ~~Property Predictability through Downscaling~~Properties

Building on the SST prediction skill assessment, we next examine whether dynamical downscaling improves prediction skill for other key NES water properties, including SSH, SSS, T_{bottom}, and HC_{200m}. ~~Figures 8 show~~ 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

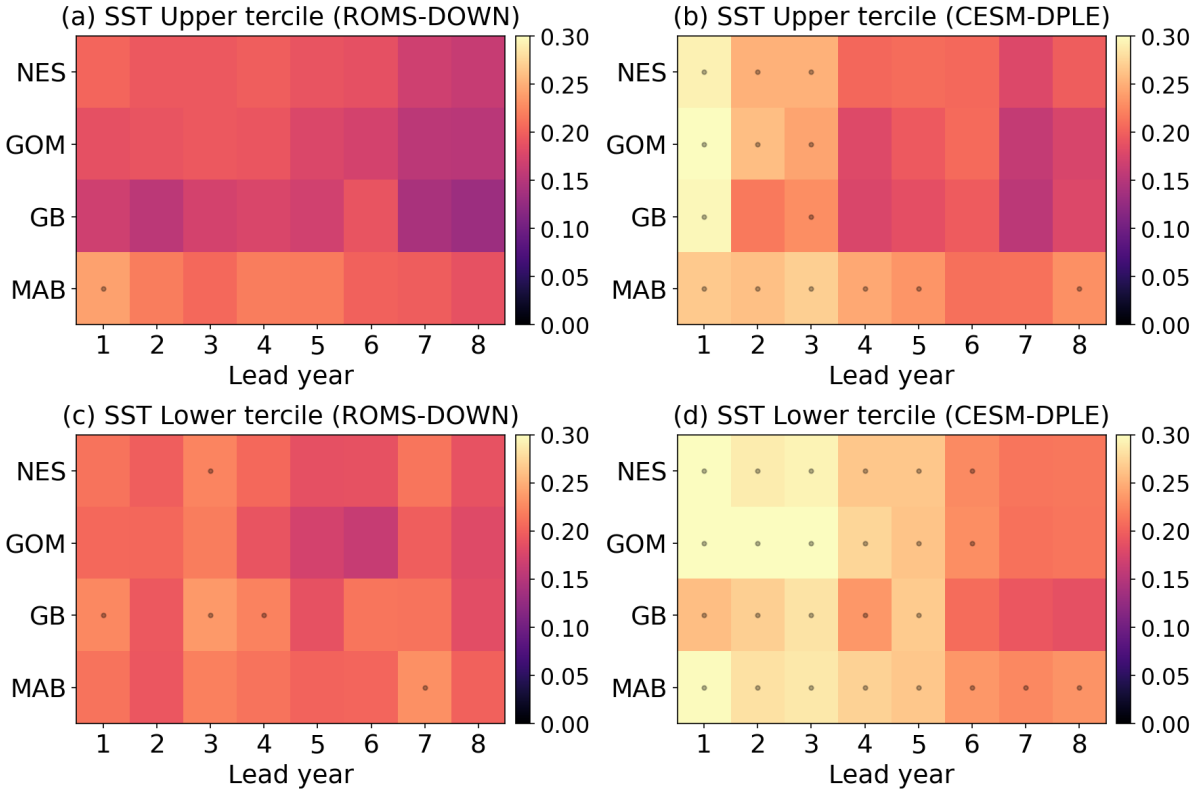


Figure 7. Brier Score for probabilistic predictions of SST tercile events on the NES. Panels (a) and (c) show the Brier Score for upper- and lower-tercile SST events, respectively, in ROMS-DOWN; panels (b) and (d) show the corresponding results for CESM-DPLE. Scores are calculated for the NES, GOM, GB, and MAB as a function of lead year (LY1–LY8). Lower values indicate higher prediction skill. Black dots denote values with BrS $\geq$ 2/9, the no-skill reference value for tercile events (i.e., a climatological forecast that assigns equal probability (1/3) to each category yields an expected Brier Score of 2/9, so values at or above this level indicate no improvement over climatology)

.

summarizes the mean prediction skill over the NES during the early (LY2–5) and late (LY6–8) forecast periods. 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).

For SST, the skill differences (Figure 8a,e) further confirm that ROMS-DOWN's improvements are most pronounced during the first five lead years, reflecting our regional model's ability to better resolve shelf dynamics, Gulf Stream position, and mesoscale variability. Similar enhancements are evident for SSS in lead years 1-6 to a generally greater extent than SST

(Figure 8b,f), consistent with the pronounced interannual variability in SSS (e.g., Figure 10), which is better represented in ROMS-DOWN. The BrS difference indicates that ROMS-DOWN provides higher reliability in predicting cold and freshening events in SST and SSS (Figure 8m,n), whereas CESM-DPLE consistently shows weaker skill for lower-tercile anomalies that are influenced by unresolved regional processes mentioned above in section 3.2. For subsurface properties, ROMS-DOWN again improves skill relative to CESM-DPLE. For T_{bottom}, improvements are most evident in the lead years 1-5, when bottom-layer variability is strongly influenced by shelf bathymetry and stratification, processes that are not well represented in CESM-DPLE. For HC_{200m}, the advantage is most pronounced in MSSS, with widespread positive differences persisting across all lead years. The ~~improved performance~~ better agreement of the high-resolution regional model with GLORYS may be related to its finer vertical resolution, more realistic shelf bathymetry, and improved representation of vertical mixing and entrainment processes which together likely contribute to a more accurate simulation of the vertical structure and timing of variability.

The magnitude and persistence of these skill differences also vary across subregions (Figure 8), particularly for variables and lead times where regional processes are expected to play a larger role. In the Gulf of Maine, skill improvements are generally more modest—especially for subsurface variables—likely reflecting the strong influence of upstream slope-water variability that is partly captured by CESM-DPLE. In contrast, the Middle Atlantic Bight tends to exhibit larger improvements in ROMS-DOWN for salinity and subsurface variables, especially at early lead times. This likely reflects the poorer performance of CESM-DPLE in this region, where Gulf Stream overshooting (Figure 2) and the narrow shelf bathymetry (Figure 1) are not well represented. Skill over Georges Bank is more variable and generally weaker, possibly due to strong tidal mixing and locally driven circulation variability that limit deterministic predictability and reduce the contribution of low-frequency, predictable signals.

Across metrics, the complementarity of ACC, MSSS, and BrS provides important insights into the sources of predictability. ACC reveals where models capture (or fail to capture) the phasing of anomalies. MSSS shows reductions in mean bias and error amplitude, while BrS directly evaluates the reliability of probability for discrete events. Together, these metrics suggest that while CESM-DPLE has skill at longer leads (LY6-LY8) through its representation of externally forced low-frequency variability, ROMS-DOWN tends to improve prediction skill at shorter leads and in shelf regions where regional dynamics matter most. ~~In summary, dynamical downscaling can improve the predictability of NES water properties across both deterministic and probabilistic skills~~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. The improved representation of shelf–slope exchange, freshwater pathways, and vertical stratification in ROMS-DOWN may contribute to reduced mean-state biases, enhanced skill in bottom and upper-ocean heat content, and improved reliability for extreme events. While CESM-DPLE retains an advantage during periods when externally forced signals dominate—such as the shelf-wide warming observed over the analysis period (Figure 4)—ROMS-DOWN appears to provide greater regional skill by better representing interannual-to-decadal variability.

Table 1. Comparison of mean prediction skill between ROMS-DOWN and CESM-DPLE over the NES. Skill metrics are averaged over lead years (LY) 2–5 and 6–8 to summarize the early- and late-lead forecast performance, respectively. For ACC and MSSS, larger values indicate higher deterministic skill, whereas for BrS, smaller values indicate higher probabilistic skill. Bold values indicate the model with the higher skill (higher ACC and MSSS, lower BrS) for each lead-year range.

Variable	Metric	LY2–5		LY6–8	
		ROMS-DOWN	CESM-DPLE	ROMS-DOWN	CESM-DPLE
SST	ACC	**0.63**	0.47	0.67	**0.75**
	MSSS	**0.29**	0.14	0.29	**0.33**
	Upper BrS	**0.19**	0.23	**0.17**	0.19
	Lower BrS	**0.20**	0.28	**0.20**	0.22
SSS	ACC	**0.56**	0.00	**0.37**	0.31
	MSSS	**0.13**	-0.21	-0.23	**-0.06**
	Upper BrS	**0.24**	0.32	**0.26**	0.31
	Lower BrS	**0.22**	0.34	**0.27**	0.31
T_{bottom}	ACC	**0.50**	0.20	0.61	**0.67**
	MSSS	**0.13**	-0.04	0.20	**0.21**
	Upper BrS	**0.22**	0.28	**0.20**	0.26
	Lower BrS	**0.24**	0.29	**0.22**	0.24
HC_{200m}	ACC	**0.56**	0.20	0.62	**0.66**
	MSSS	**0.21**	-0.76	**0.26**	0.09
	Upper BrS	**0.20**	0.29	**0.19**	0.24
	Lower BrS	**0.23**	0.29	**0.21**	0.22

3.4 Contributions from different timescales

To further investigate the sources of improved prediction skill in ROMS-DOWN, we decompose the anomaly correlation coefficient (ACC) into contributions from the linear trend, low-frequency variability (periods $>$ 10 years), and higher-frequency variability. This decomposition is performed by sequentially removing components from both the predicted and observed anomaly time series (Figure A7) and recomputing ACC at each step. Specifically, ACC is first calculated using the full anomaly time series. To quantify the relative contributions to total forecast skill, we use the squared ACC as a measure of explained anomaly variance. The contribution from the linear trend is estimated from the reduction in squared ACC after removing a linear trend from both the predicted and observed series. The detrended series is then high-pass filtered using a Butterworth filter with a 10-year cutoff period, applied independently to the predicted and observed lead-year time series using reflection padding to reduce edge effects. The squared ACC of the filtered series represents the contribution from variability on timescales shorter

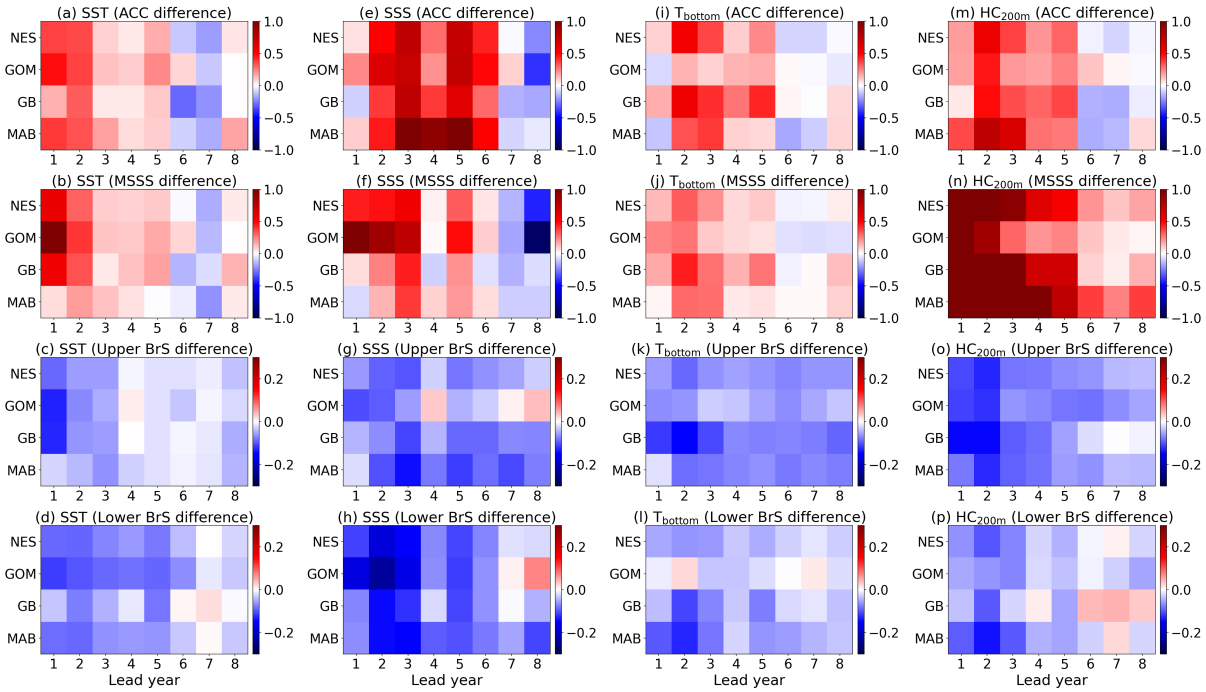


Figure 8. Differences in prediction skill between ROMS-DOWN and CESM-DPLE across the NES for multiple water properties. Panels (a–d) show ACC differences for SST, SSS, T_{bottom}, and HC_{200m}; panels (e–h) show MSSS differences. Positive values (red) indicate higher skill in ROMS-DOWN, while negative values (blue) indicate higher skill in CESM-DPLE. Panels (i–l) show Brier Score (BrS) differences for upper-tercile events, and panels (m–p) show BrS differences for lower-tercile events. Negative values (blue) indicate lower BrS and thus greater probabilistic skill in ROMS-DOWN relative to CESM-DPLE, while positive values (red) denote better performance in CESM-DPLE. Values are averaged across the full shelf (NES) and subregions (GOM, GB, MAB) as a function of lead year (LY1–LY8).

than 10 years. The remaining contribution, attributed to decadal and longer-timescale variability, is defined as the residual difference between the squared ACC of the detrended series and that of the high-pass filtered series. This decomposition provides a quantitative assessment of the relative roles of externally forced trends versus internal climate variability in shaping the overall skill.

Taking SST as an example, the decomposition of the time series (Figure A7) and the corresponding ACC contributions (Figure 9) show that a large fraction of the prediction skill in ROMS-DOWN is attributable to the linear trend, particularly at longer lead times (LY6–LY7). This suggests that much of the correlation skill derives from the externally forced warming signal, consistent with the finding that CESM-DPLE and CESM-LE also perform well when evaluated at longer leads. However, the linear trend over this short period may also reflect low-frequency internal variability rather than purely external forcing, which may partly explain the reduction of trend-related ACC at LY8 in the GB and MAB regions (Figure 9h). At shorter lead times (LY1–LY5), however, interannual-to-decadal variability contribute more to SST forecast skill, suggesting that the regional downscaling adds value by better resolving internal variability driven by regional processes that are not well represented in the

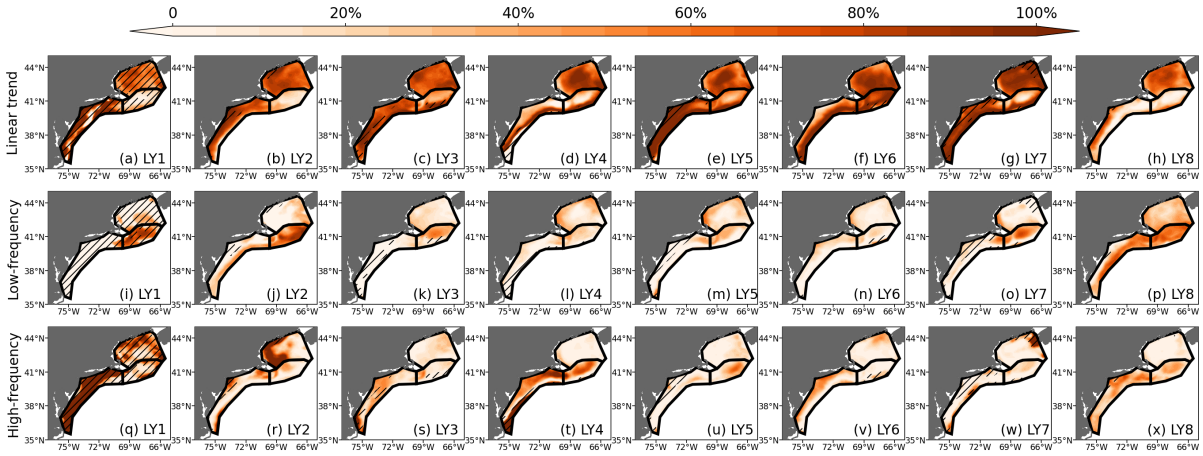


Figure 9. Decomposition of ACC(SST) in ROMS-DOWN. The total ACC is partitioned into the relative contributions (%) from (a–h) the linear trend, (i–p) low-frequency variability (periods $>$ 10 years), and (q–x) high-frequency variability (periods $<$ 10 years). Contributions are calculated as percentage of the squared ACC associated with each component. Hatched areas mask regions where the total ACC is not statistically significant at the 95% confidence level.

coarse-resolution global model. Even though the contribution from high-frequency variability is generally small compared to the trend component, it plays a non-negligible role in certain subregions. In the MAB and GB, for example, where shelf–slope exchange, Gulf Stream and mesoscale activity strongly influence shelf water conditions, high-frequency contributions exceed 20% of the total ACC in the early lead years (Figure 9q-t). This highlights the capacity of the high-resolution model to represent fine-scale dynamics that are not resolved in the coarse global system. Overall, this decomposition emphasizes that while much of the skill in SST predictions arises from predictable large-scale and trend-driven signals, the improvement of ROMS-DOWN lies in its ability to capture interannual-to-decadal variability and some higher-frequency processes that influence the shelf. It also suggests that high ACC or MSSS values do not necessarily imply that the models are skillfully capturing regional variability at all timescales, but may instead reflect dominance of low-frequency trend signals. The added value of regional downscaling may become more pronounced on shorter timescales, such as seasonal predictions, where regional dynamics and improved initialization could play a larger role.

Unlike SST, whose forecast skill primarily arises from externally forced low-frequency variability, the sources of SSS forecast skill reflect complex contributions among trend, low-frequency, and high-frequency variability. Because SSS in the NES is strongly influenced by smaller-scale shelf dynamics and regional freshwater input (Gawarkiewicz et al., 2022; Ryan et al., 2024), salinity anomalies exhibit pronounced interannual-to-decadal variations on the shelf, with relatively smaller contributions from the long-term trend or low-frequency variability, particularly during lead years 1-4 (Figures 10 and 11). Both CESM-DPLE and the uninitialized CESM-LE exhibit weak SSS variability and do not reproduce the observed shelf-wide freshening and salinification events (Figure 10). In particular, CESM-LE exhibits very weak SSS variability. For example, the freshening event captured in GLORYS in 1998 (Wallace et al., 2018)—associated with advection of freshwater along

the ~~slope sea~~ Slope Sea (not shown)—was skillfully represented only in ROMS-DOWN in all lead times including 1998, i.e., LY1-LY6. These events are absent in CESM-DPLE and CESM-LE, suggesting that although the large-scale interannual variability is present in the global model, its expression on the shelf is not well represented. The improved performance in ROMS-DOWN likely reflects its more realistic representation of shelf bathymetry and coastal processes, which allow large-scale salinity anomalies to be more effectively translated onto the shelf. ~~In ROMS-DOWN, variability in the~~ Variability in Gulf Stream position and meandering modulates salinity in the ~~slope sea, which in turn~~ Slope Sea, which subsequently influences shelf salinity through water mass exchange. ~~In addition, coastal processes such as~~ Although freshwater inflow from ~~estuaries and rivers are represented in ROMS-DOWN through prescribed monthly climatological forcing, which improves~~ rivers and estuaries contributes to the mean salinity structure ~~on the shelf. However, because river discharge does not vary interannually~~ through the prescribed climatological river forcing, the absence of interannual river variability in our experimental design ~~, the improved~~ indicates that the enhanced SSS skill in ROMS-DOWN primarily ~~reflects~~ arises from its ability to dynamically translate ~~large-scale boundary~~ basin-scale salinity anomalies onto the shelf~~through realistic shelf–slope exchange and along-shelf advection processes (Figure 8b,f)~~. This mechanism is further discussed using the 1998 freshening event in Section 3.2. The enhanced ACC skill of SSS in ROMS-DOWN is particularly evident in the MAB (Figure 8b), where the interplay between coastal freshwater plumes and slope water intrusions drives relatively strong interannual variations in salinity (Ryan et al., 2024), consistent with the notable contribution from high-frequency variability to the overall ACC skill in the MAB (Figure 11q–x). Although a larger contribution from higher-frequency variability does not necessarily imply higher overall ACC, it indicates that the model effectively represents a broader spectrum of salinity variability, including processes that the coarse-resolution global model fails to capture. The larger skill improvement in this region likely also reflects limitations in CESM-DPLE's representation of shelf bathymetry and offshore Gulf Stream structure, which reduce the fidelity of shelf salinity variability, whereas the higher-resolution regional configuration provides a more realistic representation of these processes.

4 ~~Discussion~~

3.1 Prediction of two key shelf processes

Skillful multi-year prediction of physical conditions on the NES has the potential to support ecosystem-based management by providing advance information on environmental variability that influences marine habitats, productivity, and fisheries resources. Here we discuss the prediction of two key physical phenomena on the NES, MAB cold pool and the GOM slope-water mixture, because they translate basin-scale signals into shelf conditions that have direct ecological and management importance. The cold pool regulates summer bottom temperatures and stratification across major demersal fishing grounds, while the relative fractions of different slope-water sources entering the GOM through the Northeast Channel control deep hydrography and nutrient supply (Link et al., 2002; Townsend et al., 2010; Miles et al., 2021). Both features are tracked as ecosystem indicators in regional assessments and are central to understanding habitat for commercially important species (Pershing et al., 2015). Evaluating prediction skill for these two process-based indicators therefore tests whether regional

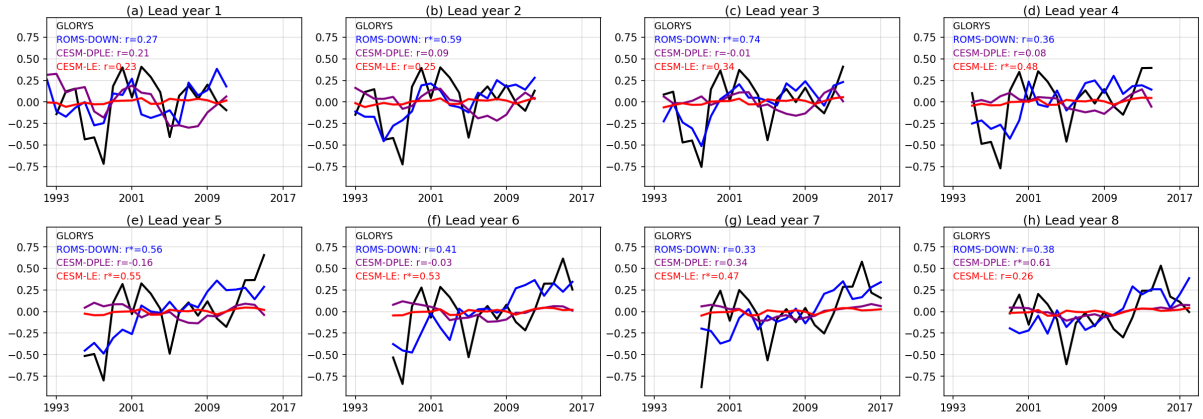


Figure 10. Time series of SSS anomalies averaged over the entire NES shelf from GLORYS (black), ROMS-DOWN (blue), CESM-DPLE (purple), and CESM-LE (red) for lead years 1–8. Correlation coefficients (r) and significance levels (p-values) between each model and GLORYS are indicated in the top left of each panel, with r* indicating significance at the 95% level.

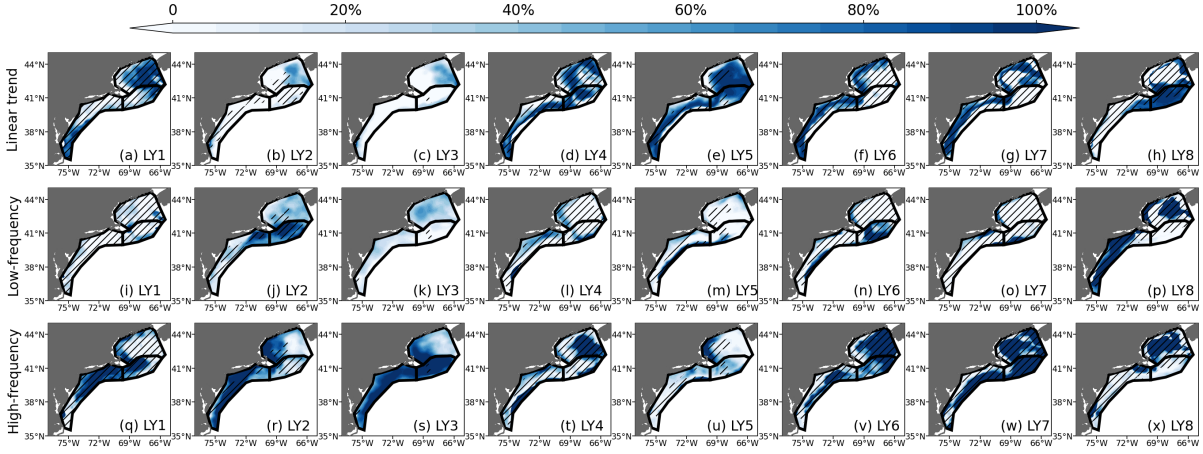


Figure 11. ~~Same as Figure 9, but for~~ Decomposition of ACC(SSS) in ROMS-DOWN. The total ACC is partitioned into the relative contributions (%) from (a–h) the linear trend, (i–p) low-frequency variability (periods > 10 years), and (q–x) high-frequency variability (periods < 10 years). Contributions are calculated as percentage of the squared ACC associated with each component. Hatched areas mask regions where the total ACC is not statistically significant at the 95% confidence level.

downscaling not only improves large-scale skill metrics but also provides information that is decision-relevant for fisheries and ecosystem management.

3.1.1 Prediction of the MAB cold pool

Although CESM-DPLE provides relatively skillful large-scale predictions of SST and upper-ocean heat content in particular in longer lead years, it does not resolve important processes on the NES. One notable example is the MAB cold pool, a distinct subsurface thermal feature that integrates interactions among stratification, air-sea heat flux, advective heat fluxes, and cross-shelf exchange (Bigelow, 1933; Fratantoni and Pickart, 2007; Lentz, 2017; Chen and Curchitser, 2020) which is not well represented in CESM-DPLE. The coarse horizontal and vertical resolution of the global model with Gulf Stream overshooting prevents it from capturing the sharp vertical stratification and cross-shelf temperature gradients required for cold pool formation, resulting in a weak and diffuse temperature field that does not clearly represent the observed cold pool structure (Figure 2r). Given that ROMS-DOWN demonstrates better performance and improved forecast skill for key water properties on the MAB shelf, it is important to evaluate whether these improvements translate into a more realistic simulation of the MAB cold pool, a feature fundamental to the physical and ecological dynamics on NES (Fratantoni and Pickart, 2007; Miles et al., 2021).

The cold pool forms each spring and summer when surface heating and stratification isolate the dense, cold winter bottom water from the warmer surface layer. It typically occupies the northern shelf region in MAB and persists through early fall until mixing erodes the vertical temperature gradient. Following Chen and Curchitser (2020), the cold pool can be identified in GLORYS and ROMS-DOWN using three criteria: (1) temperature below 10 °C and salinity below 34 psu, (2) a well-developed stratification, and (3) location within the MAB shelf region between the 20 m and 200 m isobaths. These criteria allow for an objective comparison between GLORYS and model simulations in terms of cold pool structure and variability. For comparison, Figure 12 shows the annual mean, depth-averaged temperature over the identified cold pool region from GLORYS and the ensemble-mean ROMS-DOWN simulations for 2004–2008 (the cold pool is not reproduced at all in CESM-DPLE based on the above criteria). In general, ROMS-DOWN reproduces the spatial extent and structure of the cold pool with good agreement to GLORYS, capturing the tongue-shaped feature that extends southward along the shelf and the persistent cold core centered around 38–40°N. The regional model also captures more skillfully the interannual variability in cold pool intensity and volume, which is absent in the coarse-resolution global system (Figure 13a-b). Specifically, ROMS-DOWN captures the observed warming and shrinking in cold pool in recent years (Friedland et al., 2022), consistent with GLORYS. Several key factors contribute to the improved representation of the cold pool in ROMS-DOWN, including its enhanced ability to simulate vertical thermal structure, shelf–slope exchange, and local heat advections that control cold pool formation and decay. These processes are sensitive to bathymetry and vertical resolution, which ROMS-DOWN is capable of realistically capturing.

In addition, the ACCs of cold pool temperature and volume in ROMS-DOWN show modest but positive correlations across most lead times, which indicates some degree of multi-year predictability of cold pool characteristics in the dynamical downscaling system (Figure 13c). Although the ACC magnitude of cold pool volume is relatively low, the persistence of positive correlations through the eight lead years suggests that ROMS-DOWN captures part of the variance in cold pool evolution.

However, while dynamical downscaling clearly improves the representation of the cold pool's structure and variability, it does not necessarily translate into statistically high predictability. Several factors likely limit the prediction skill even in the high-resolution regional framework. For example, the boundary and surface forcing from CESM-DPLE inherit large-scale biases and initialization errors from the global system (Yeager et al., 2018), particularly in the Gulf Stream path and subsurface heat content. In addition, air–sea heat flux variability and wind-driven mixing—both of which play a central role in winter cold pool formation and its summer erosion—contain substantial unpredictable interannual variability that limits deterministic forecast skill in the model system (Chen and Curchitser, 2020). Given the importance of vertical diffusion in driving temperature anomalies in the cold pool region (Chen and Curchitser, 2020), uncertainties in vertical mixing parameterizations in the model may further degrade multi-year predictability. Moreover, mesoscale processes such as Gulf Stream warm core rings can interact with the continental shelf (Chen et al., 2014b; Gawarkiewicz et al., 2018) and influence cold-pool evolution, which is challenging to model without data assimilation. The initialization of the cold pool is also crucial for driving its interannual variability (Chen and Curchitser, 2020). Such mesoscale and internal processes exhibit intrinsic variability that limits deterministic prediction. Therefore, while dynamical downscaling enhances physical realism, the inherent complexity and internally generated nature of the system continue to constrain its multi-year predictability.

3.1.2 Prediction of slope water mixing in GOM

A key component of NES shelf-water variability is driven by boundary inflows and the mixing of distinct slope-water masses entering through several channels, with the Northeast Channel being one of the most important pathways. The GOM is strongly influenced by the NES large-scale cyclonic circulation system. This includes the Labrador Current which brings cold, nutrient-rich water along the shelf and into the GOM through the Northeast Channel. Therefore, waters of distinct origins converge and mix within the GOM (Pringle, 2006; Townsend et al., 2010). The deep inflow through the Northeast Channel consists primarily of slope waters that are a mixture of warm, saline Warm Slope Water (WSW) originated from the Gulf Stream and cold, fresh Labrador Slope Water (LSW) advected southwestward from the subpolar North Atlantic (Mountain, 2012). These water masses, together with the overlying Scotian Shelf Water (SSW), combine to form the subsurface properties of the GOM, exerting strong control on regional temperature, salinity, stratification, and nutrient variability. The relative proportions of WSW and LSW entering the GOM vary on interannual to decadal timescales and are linked to large-scale climate forcing and Gulf Stream variability (Mountain, 2012; Seidov et al., 2021). Therefore, understanding and predicting the variability of slope-water mixing is fundamental to assessing shelf hydrographic predictability and its downstream ecological impacts.

Following Mountain (2012), the three end-members—WSW, LSW, and SSW—form a well-defined mixing triangle. Annual-mean subsurface (150–200 m) temperature–salinity (T–S) distributions within the Northeast Channel are shown in Figure 14 for GLORYS, ROMS-DOWN, and CESM-DPLE, covering the period 1992–2011 across all lead years of the forecast systems. GLORYS occupies a broad range primarily between the WSW and LSW end-members, consistent with substantial variability in the slope-water composition entering the GOM shown from the observational dataset in Mountain (2012). CESM-DPLE, however, exhibits an overly saline bias, likely associated with proximity of the Gulf Stream due to the overshooting. ROMS-DOWN reproduces a wider spread of T–S values, spanning much of the observed range between WSW and LSW, while its

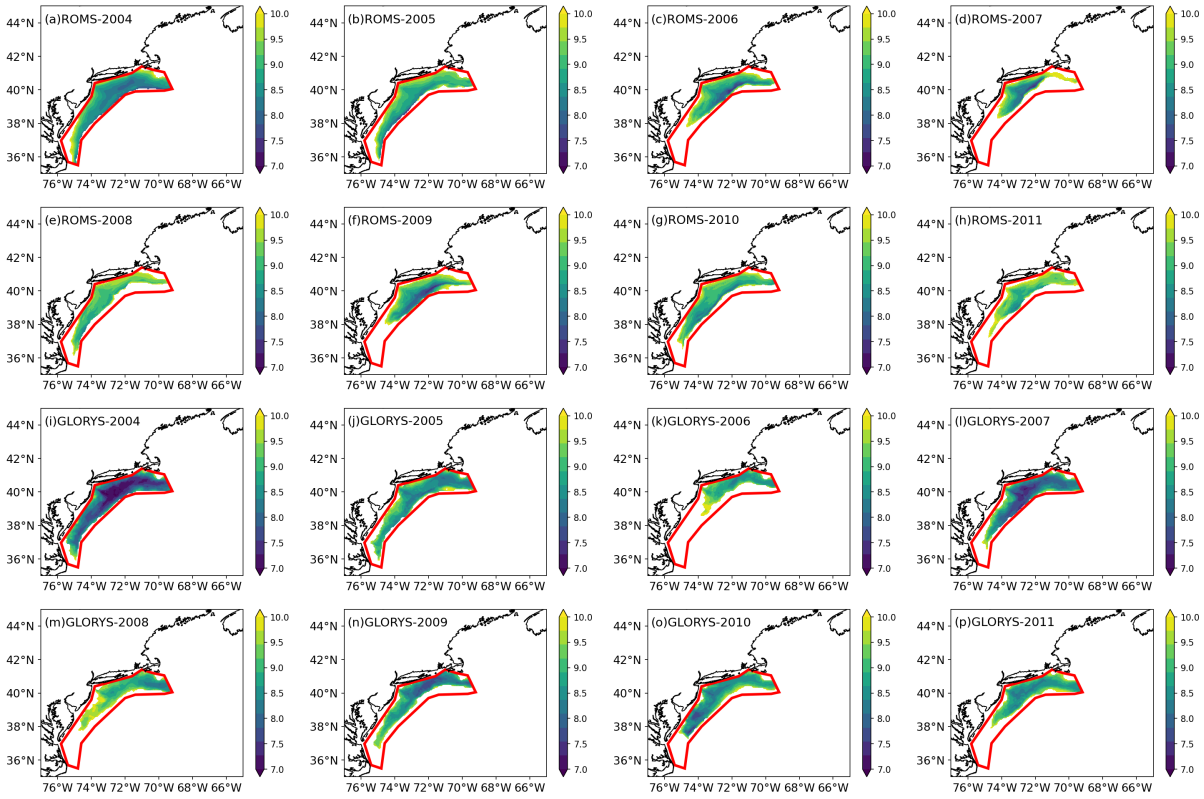


Figure 12. Annual mean vertically averaged temperature within the identified MAB cold pool layer from (a–h) the ROMS-DOWN averaged across the ensemble members and lead years and (i–~~o~~p) GLORYS for 2004–2011.

ensemble mean lies closer to the GLORYS distribution. This suggests that the regional model more skillfully represents the processes that control intermediate and deep water-mass variability in the GOM. The interannual variability of the slope-water mixture is influenced by large-scale circulation changes associated with the NAO (Mountain, 2012), as well as by oceanic intrinsic variability, including Gulf Stream meandering and eddy intrusions along the slope (Seidov et al., 2021). The improved simulation in ROMS-DOWN likely reflects its finer resolution and its ability to resolve bathymetry steering, shelf–slope exchange, and Gulf Stream variations that modulate the inflow through the Northeast Channel.

To further evaluate the model's ability to predict changes in this slope-water mixture, the forecast skill of the Labrador Slope Water fraction (%LSW) near the Northeast Channel is examined in Figure 15. The overall positive ACC values of %LSW indicate that ROMS-DOWN retains some multi-year predictive capability for the proportion of LSW entering the GOM. The persistence of skill throughout the prediction period suggests that the model captures large-scale processes governing LSW variability, such as the southwestward transport of subpolar waters and their modulation by the NAO (Mountain, 2012). However, given the relatively short verification period, part of the increasing skill at longer leads may also reflect the influence of low-frequency variability or a trend-like signal in slope-water properties, potentially associated with broader climate-driven

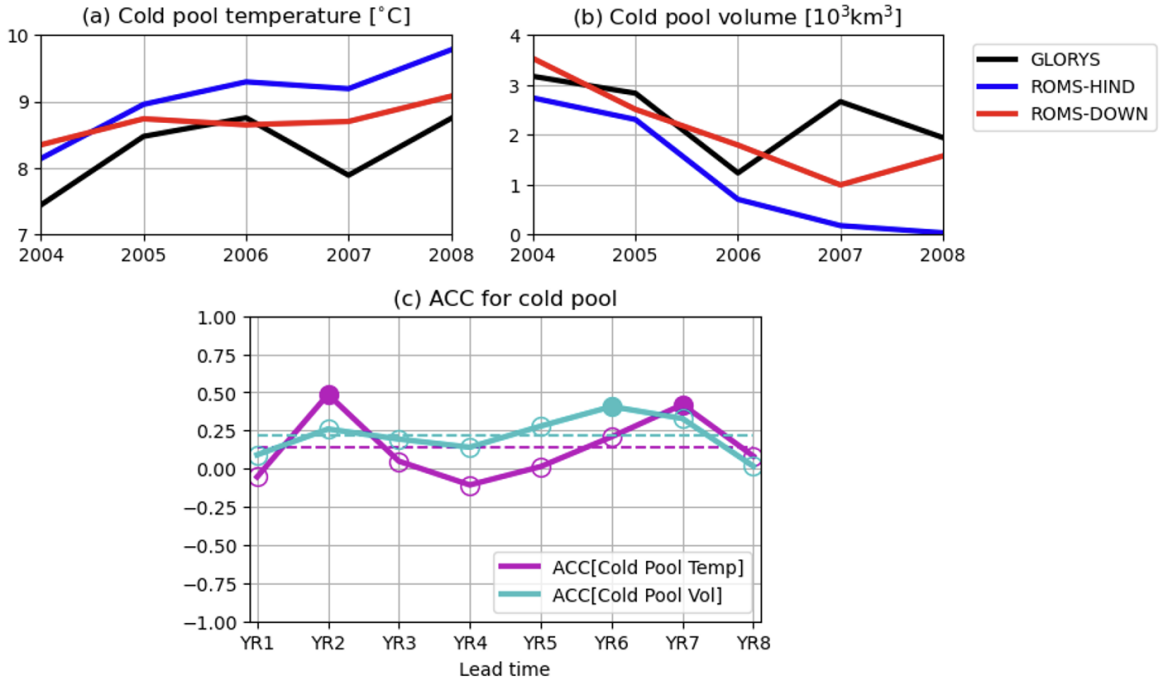


Figure 13. (a) Annual mean cold pool temperature and (b) cold pool volume from GLORYS (black), ROMS-HIND (blue), and ROMS-DOWN (red) averaged across the ensemble members and lead years during 2004–2011. (c) ACC of cold pool temperature and cold pool volume in ROMS-DOWN as a function of lead year (LY1–LY8). The dashed line indicates the mean ACC over all the lead years. Only filled markers are statistically significant at the 95% confidence level.

555 changes. The relatively higher skill of LSW compared to more locally influenced shelf-water properties implies that this signal is more strongly controlled by predictable basin-scale dynamics, which are preserved through initialization from CESM-DPLE and transmitted into the regional system.

Accurately predicting the relative contribution of %LSW also depends on the model's ability to simulate the background hydrographic structure of the shelf. Even though CESM-DPLE captures the large-scale forcing and its influence on Labrador 560 Water transport to some extent, the representation of %LSW entering the GOM can still be biased by unrealistic local shelf conditions, such as the intrusion of warm, saline water associated with the Gulf Stream's overshooting into the region or the lack of a Northeast Channel due to coarse-resolution bathymetry. Although the MSSS values of %LSW are modest, which indicates limited amplitude accuracy, ROMS-DOWN maintains variability in slope-water composition over several forecast years. The slight offset in magnitude between ROMS-DOWN and GLORYS (e.g., saltier water masses in Figure 14) likely 565 reflects residual biases in the boundary conditions inherited from CESM-DPLE.

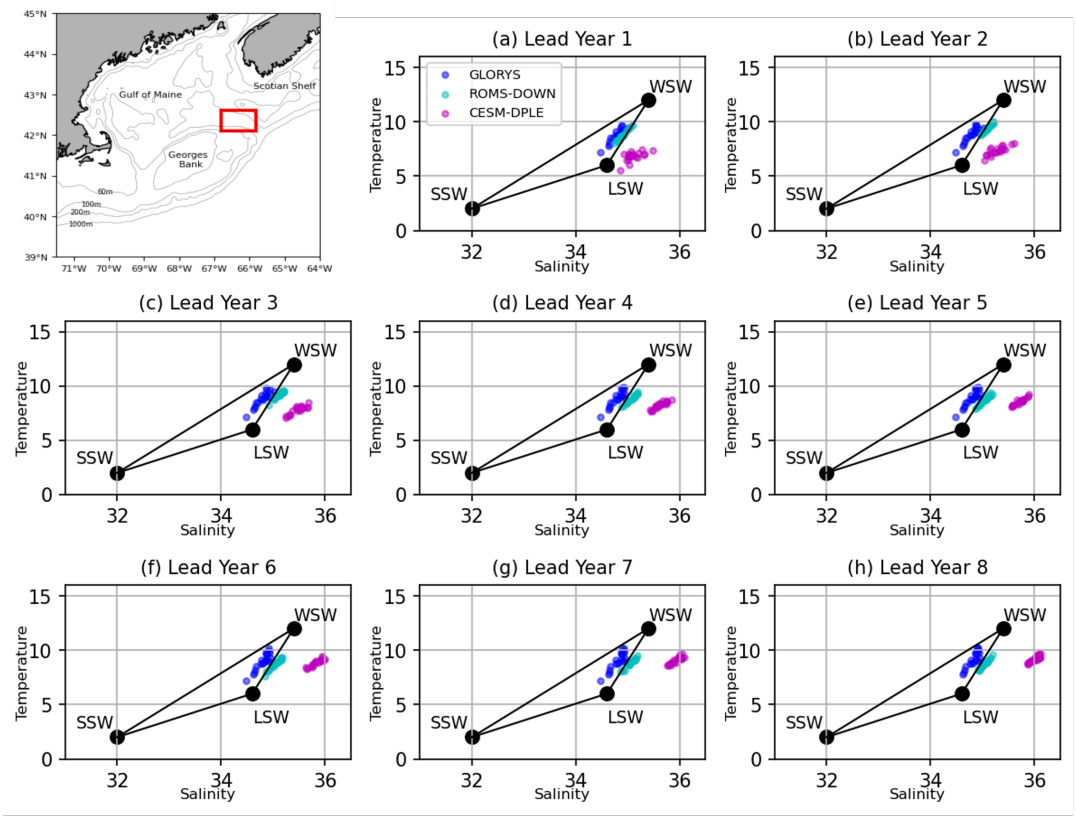


Figure 14. Temperature–salinity (T–S) diagrams showing slope water masses and their mixing in the GOM. The inset map (upper left) shows the analysis region near the Northeast Channel in the GOM (red box). Water-mass end-members include Warm Slope Water (WSW), Labrador Slope Water (LSW), and Scotian Shelf Water (SSW), following Mountain (2012). Colored dots represent depth-averaged (150-200 m) annual-mean T–S values within the analysis region from GLORYS (blue), ROMS-DOWN (cyan), and CESM-DPLE (magenta). Each dot corresponds to a different year. For ROMS-DOWN and CESM-DPLE, values are ensemble means for each year.

3.2 Source of multiyear predictability: a case study of the 1998 freshening event

The improved multiyear prediction skill of SSS in ROMS-DOWN suggests that regional processes play an important role in translating large-scale salinity variability onto the NES shelf. To illustrate this mechanism, we consider the pronounced freshening event observed in 1998 (Figure 10). The spatial distribution of SSS anomalies in GLORYS (Figure 16a) shows a strong negative anomaly extending along the shelf and slope, particularly in the MAB and GOM. ROMS-DOWN reproduces both the magnitude and spatial structure of this anomaly (Figure 16b), whereas CESM-DPLE fails to capture this freshening event (Figure 16c). This is consistent with the improved interannual prediction skill of SSS in ROMS-DOWN and suggests that the regional model better represents key processes associated with this event.

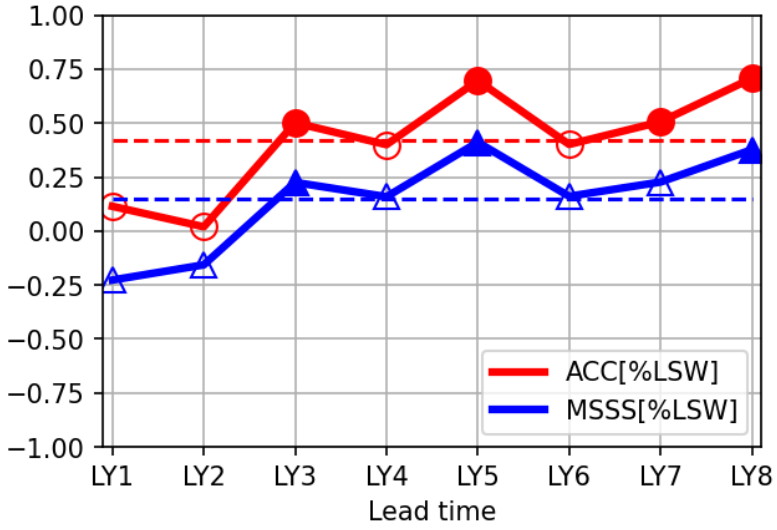


Figure 15. ACC (red) and MSSS (blue) of Labrador Slope Water fraction (%LSW) in ROMS-DOWN as a function of lead year (LY1–LY8). The dashed lines indicate the mean skill over the prediction period. Filled markers are statistically significant correlations at the 95% confidence level.

Analysis of the preconditioning leading up to the event (not shown) reveals that a fresh anomaly develops in the slope waters one to two years prior to 1998 in both GLORYS and ROMS-DOWN. This indicates a large-scale origin of the signal perhaps from outside of our ROMS-DOWN domain. Consistent with this, the fraction of LSW at the entrance of the GOM (Figure 16d) as discussed in section 3.1.2 shows a pronounced increase in both GLORYS and ROMS-DOWN in 1998, which suggests enhanced southward transport of subpolar waters (currents shown in Figure 16a-b) as a key driver of the event. However, CESM-DPLE does not capture the %LSW variation (Figure 16d) nor the equatorward shelf and slope currents (Figure 16c), indicating that although large-scale signals may be present, their expression on the shelf is not well represented in the coarse-resolution global system. The strong relationship between %LSW and salinity anomaly in 1998 further suggests that basin-scale forcing, potentially linked to NAO variability (Mountain, 2012), ~~provide~~ provides an important source of predictability for NES properties.

To assess whether the predictability of this event is attributable to large-scale anomalies advected from the subpolar gyre—which are prescribed as the northern boundary condition for ROMS-DOWN—we examine SSS anomalies over a narrow area near the northern boundary (Figure 16e). ROMS-DOWN closely follows CESM-DPLE in both phase and magnitude of boundary SSS anomalies, indicating that the large-scale salinity signal is inherited from CESM-DPLE forcing by design. Despite this, CESM-DPLE fails to reproduce the corresponding shelf freshening due to lack of equatorward advection along the slope and shelf break (Figure 16c). The large-scale salinity signal weakens as it propagates into the coastal region in CESM-DPLE, likely due to its limited representation of shelf–slope bathymetry and along-shelf transport. In contrast, ROMS-DOWN more effectively preserves and translates the inherited large-scale signal into a realistic regional response.

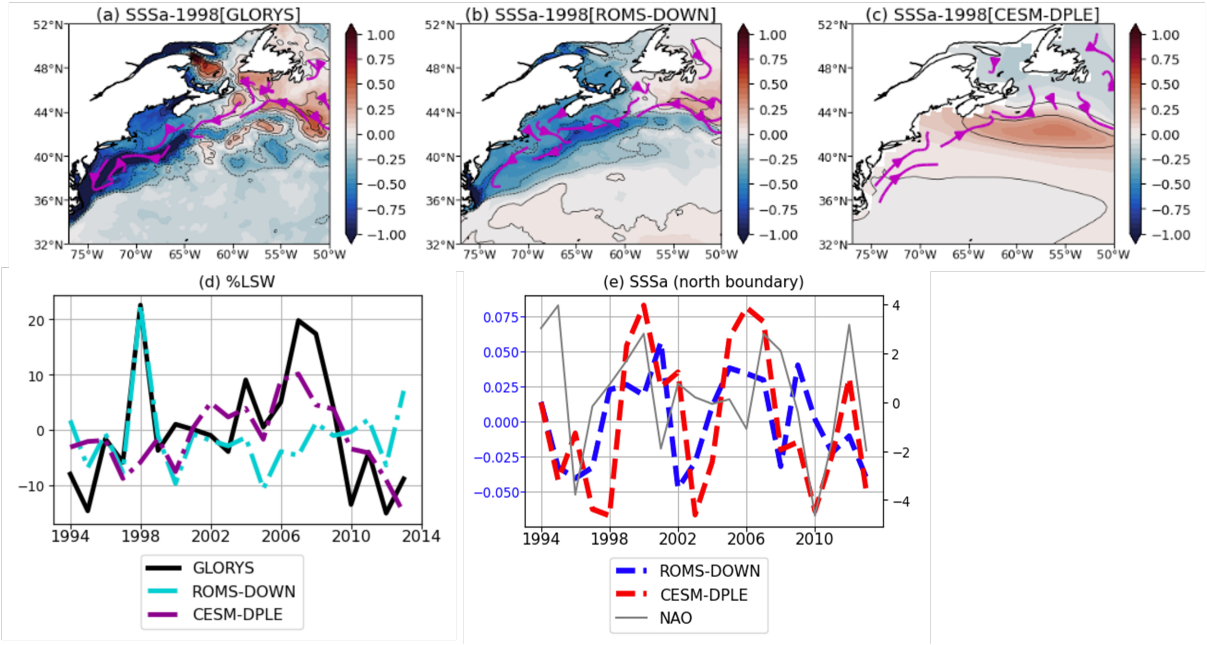


Figure 16. (a–c) Spatial maps of SSS anomalies in 1998 (a) GLORYS, (b) ROMS-DOWN, and (c) CESM-DPLE. Arrows indicate along-shelf and slope currents. (d) Time series of LSW fraction in the Northeast Channel of GOM from GLORYS, ROMS-DOWN, and CESM-DPLE. (e) Time series of SSS anomalies averaged over a narrow area near the northern boundary of ROMS-DOWN from CESM-DPLE and ROMS-DOWN, along with their relationship to the NAO. CESM-DPLE and ROMS-DOWN results are ensemble means from the 1994 start year.

This case study indicates that the improved interannual prediction skill in ROMS-DOWN arises from a combination of inherited large-scale predictability and enhanced representation of regional processes, which together enable a more accurate prediction of basin-scale anomalies onto the NES~~shelf~~.

3.3 ~~Limitations and uncertainties in the regional downscaling prediction system~~

4 Discussion

While dynamical downscaling improves the physical realism and regional predictability of NES water properties, several limitations remain that constrain overall forecast skill.

First, ~~the prediction skill in SSH is notably weak (~~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 the regional system.

A second limitation arises from the experimental design. Each ROMS-DOWN ensemble is integrated for only 8 prediction lead years and 20 start years due to limited availability of the daily atmospheric variables from CESM-DPLE to force the ROMS-DOWN as well as limited computational resources. With pronounced interannual to decadal variability, the sliding 20-year windows for different lead time prediction time series (Figures 4 and 10) are dictated by different low-frequency temporal evolutions, e.g., 1992-2021 for LY1 vs. 1999-2018 for LY8, and thus the prediction skill differences may not be solely due to the difference in lead years. For example, ~~prediction skill at longer leads may appear higher than at shorter leads because the 20-year time windows used for verification of the~~ the verification periods for longer lead years ~~are more dominated by~~ contain a larger contribution from the long-term ~~trend~~ 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. h). This sampling difference introduces uncertainty when comparing lead-year-dependent skill across variables and regions. Extending the number of start-year simulations and lengthening each simulation would provide a more comprehensive assessment of uncertainty and improve statistical confidence in the skill estimates.

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 ~~Georges Bank~~ GB and parts of the ~~Gulf of Maine~~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 ~~tides in GLORYS may contribute to modest regional differences in skill magnitude, it~~ 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.

Overall, while our dynamical downscaling approach enhances the representation of shelf processes and improves multi-year forecast skill, it does not yet overcome the inherent limits of predictability set by internal variability, model biases in both the global and regional models, and finite ensemble sampling. Future efforts should focus on extending hindcast periods, increasing ensemble size, ~~and~~ 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—~~which~~ may yield additional gains in multi-year prediction skill on the NES.

5 Summary

This study evaluates the improvement in multi-year prediction skill achieved through dynamic downscaling on the U.S. Northeast Shelf using a high-resolution regional model forced by anomaly fields from the global prediction using CESM-DPLE system. The assessment focuses on mean-state fidelity, forecast skill metrics, and the representation of key regional phenomena, including the MAB cold pool and slope-water mixing near the Northeast Chanel. Model results are compared against the GLORYS ocean reanalysis to assess both deterministic (anomaly correlation coefficient, mean squared skill score) and probabilistic (Brier Score) performance.

ROMS-DOWN substantially reduces mean-state biases in SST, SSH, SSS, bottom temperature, and upper-ocean heat content relative to CESM-DPLE. Improvements are pronounced over the shelf and slope regions, where the higher horizontal resolution allows for a more realistic representation of cross-shelf gradients, bathymetry, and offshore Gulf Stream position. Both deterministic and probabilistic metrics demonstrate that ROMS-DOWN improves forecast skill, particularly during the first five forecast years (~~LY1–LY5~~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. The BrS further shows that ROMS-DOWN provides greater reliability in predicting the probability of upper- and lower-tercile events than the global system, highlighting its strength in capturing event-scale variability. Decomposition of the forecast skill indicates that the sources of predictability differ among water properties. For example, much of the SST skill arises from the externally forced warming trend, which both ROMS-DOWN and CESM-DPLE capture well. Part of this trend-related skill may also reflect low-frequency internal variability projected onto the relatively short analysis period. In contrast, the ACC skill for SSS is primarily linked to interannual-to-decadal variability, where ROMS-DOWN exhibits improved multi-year prediction capability, likely because its higher-resolution bathymetry and more realistic shelf representation allow large-scale salinity variability inherited from CESM-DPLE to be more accurately ~~projected~~ transferred onto the shelf.

The MAB cold pool, a thermally stratified bottom layer that strongly influences regional temperature structure and marine ecosystems, is skillfully reproduced in ROMS-DOWN but not resolved at all in CESM-DPLE, manifested in the weaker T_{bottom} prediction skill in the global system (Figure 8c,g,k). The regional model captures the cold pool's spatial structure and its interannual variability in strength and volume, which demonstrates an improved representation of stratification and shelf–slope exchange processes. However, despite this enhanced physical realism, cold-pool predictability remains moderate, likely constrained by boundary forcing biases, parameterization uncertainties, and the intrinsic variability of mesoscale and synoptic processes. In the GOM, ROMS-DOWN also more accurately represents the mixing between shelf and slope waters entering through the Northeast Channel, reproducing the observed range of slope-water temperature and salinity variability more skillfully than CESM-DPLE. The global system exhibits a warm and saline bias, likely linked to bias in Gulf Stream separation near Cape Hatteras. The forecast skill of the LSW fraction in the Northeast Channel remains consistently positive, reflecting the influence of predictable large-scale circulation patterns—such as the westward transport of subpolar waters and their modulation by climate forcings such as the NAO. While amplitude errors remain, ROMS-DOWN captures physically

meaningful multi-year variability in slope-water composition, underscoring the importance of dynamical downscaling for linking large-scale climate signals with regional hydrographic variability.

Our results complement recent work on seasonal and decadal forecasts for the NES region. Initialized decadal prediction studies show that starting from observed climate states improves regional circulation and heat-transport signals, enabling skill for North Atlantic SST and related high-latitude metrics (e.g., (Yeager et al., 2015, 2018)). On a seasonal scale, dynamically downscaled regional forecasts have achieved better skill on the shelf by resolving fine-scale processes along the NES (Ross et al., 2024). Extending to decadal scales, recent downscaled predictions suggest a near-term pause in shelf warming (Koul et al., 2024). Here we assess prediction skill on interannual timescales in a regional dynamical downscaling framework for the NES. While initialized climate predictions from subseasonal to decadal timescales have been widely studied (Meehl et al., 2021; O'Kane et al., 2023), the extent to which large-scale predictability can be translated into improved regional forecasts for the NES through dynamical downscaling remains less explored. Our work on multi-year prediction ~~timescale~~ fills a critical gap between seasonal and decadal forecasts~~and is particularly~~ . 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 ~~in the area, where decisions often require~~ because it provides outlooks beyond seasonal forecasts ~~and~~ while remaining more actionable than decadal projections. We demonstrate that dynamical downscaling enhances the representation and multiyear prediction of NES hydrography by bridging global-scale forcing with regional processes that shape shelf variability. ROMS-DOWN improves mean-state fidelity, early-lead forecast skill, and event-level reliability. Although the predictability of features such as the shelf sea level anomaly and cold pool remains constrained by intrinsic variability and forcing uncertainties, the results highlight the potential of regional dynamical downscaling as a critical tool for advancing multi-year climate predictions and their applications to coastal and ecosystem management on the NES. Future efforts should include coordinated comparisons of regional downscaling systems in the NES (e.g., ROMS and Modular Ocean Model version 6 (MOM6) in (Ross et al., 2024)). Because no single model is free of bias, systematic intercomparison across models, domain configurations, and boundary forcing strategies is essential for identifying robust sources of predictability and quantifying uncertainties. Such efforts will improve confidence in downscaled climate predictions for coastal applications.

Code and data availability. The sea surface temperature data used in this study are publicly available from Reynolds et al. (2002). Sea surface height data are available from CMEMS (Assessed on 30 Aug 2026b). The GLORYS reanalysis data can be accessed via CMEMS (Assessed on 30 Aug 2026a). The model data and analysis code have been deposited in a Zenodo repository Guo et al. (2026).

Author contributions. Y.G. conducted model simulations and data analysis and wrote the original draft. A.C-T. led the model simulations and contributed to analysis. K.C., Y.K., and S.P. contributed to data analysis and result interpretation. Y.K., K.C., H.S., P.F., M.A., and V.S. provided resources, funding acquisition, administration, and supervision. All authors contributed to result interpretation, reviewing and editing the manuscript.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. The authors acknowledge support from the NOAA Climate Variability and Predictability (CVP) Program (NA20OAR4310482-T1-01) and the Cooperative Institute for the North Atlantic Region (NA19OAR4320074). KC was also supported by National Science Foundation (NSF) Ocean Science Division under Grant OCE–2241407. HS thanks for the support from the Uehiro Center for the Advancement of Oceanography (UC·AO) at the University of Hawai'i at Mānoa. YG thanks for the support from Illinois State University.

~~The authors acknowledge support from the NOAA Climate Variability and Predictability (CVP) Program (NA20OAR4310482-T1-01) and the Cooperative Institute for the North Atlantic Region (NA19OAR4320074). KC was also supported by National Science Foundation (NSF) Ocean Science Division under Grant OCE–2241407. YG was also supported by funding from Illinois State University.~~

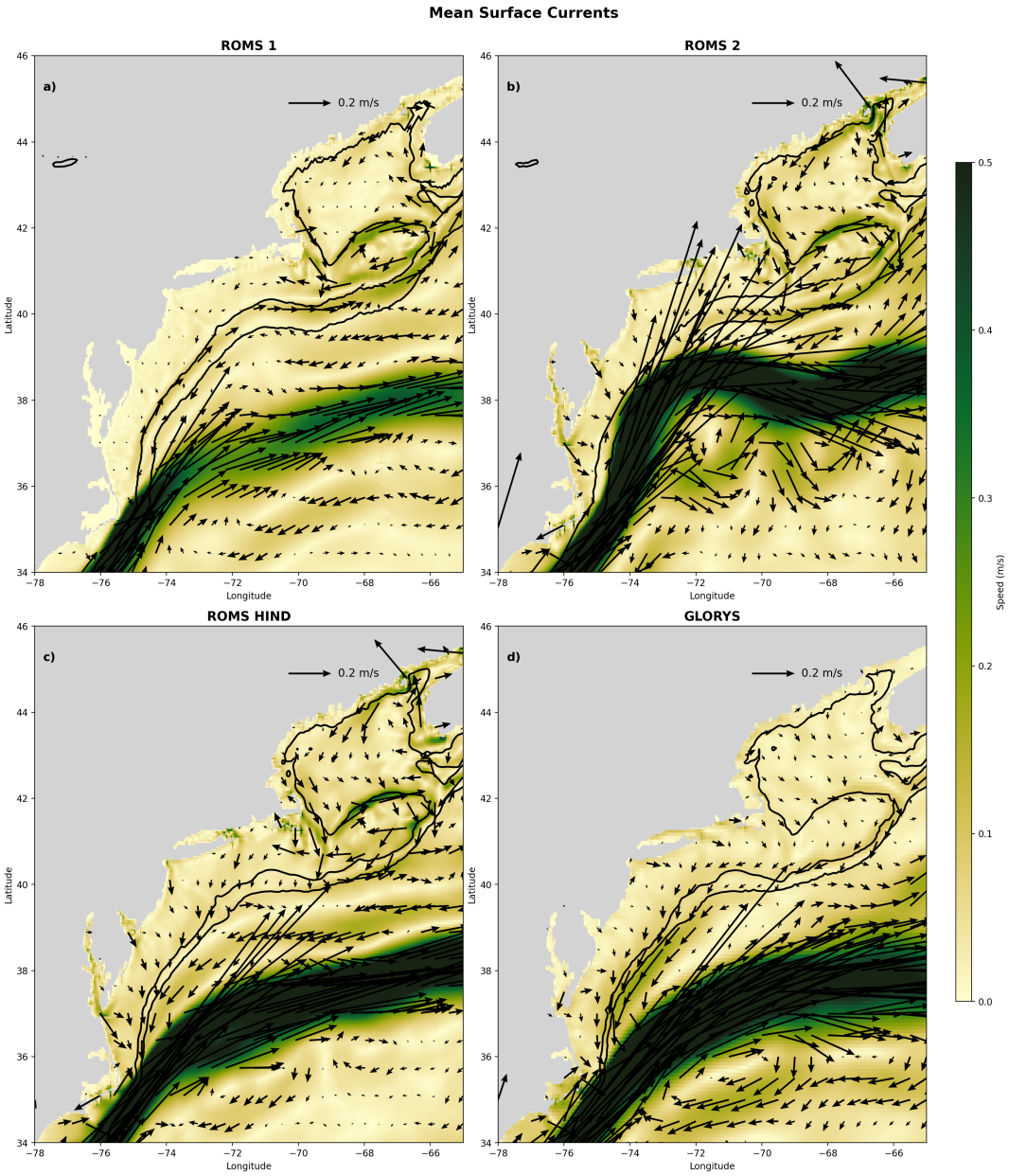


Figure A1. Five-year mean (2004–2008) surface velocity vectors and current speed (m s^{-1}) over the NES from three ROMS configurations and the GLORYS reanalysis: (a) ROMS 1 (Batteen and Miller, 2009) (smoothing, $r_{x0} = 0.1$), showing a shoreward-shifted 100-m isobath around GB, a weakened equatorward shelf flow, and a weakened Gulf Stream; (b) ROMS 2 (Shapiro filter, $r_{x0} = 0.1$), with an offshore-displaced 1000-m isobath near Cape Hatteras and Gulf Stream overshooting; (c) ROMS-HIND, using updated shelf bathymetry ($r_{x0} = 0.25$), which is more consistent with ETOPO1, yielding improved shelf circulation and Gulf Stream pathway; and (d) GLORYS reanalysis for reference (also based on ETOPO1 bathymetry). Black contours denote the 100-m and 1000-m isobaths only. Note that the shelfbreak/slope current is poleward in ROMS 1, and to a lesser extent in ROMS 2, which is likely due to the bathymetry smoothing near the Cape Hatteras, in contrast to the more realistic equatorward flow in ROMS-HIND and GLORYS.

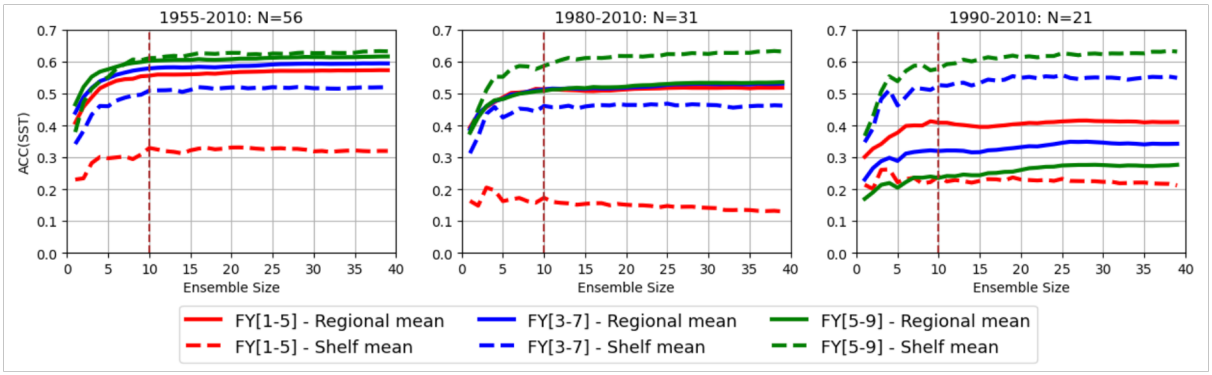


Figure A2. Sensitivity of Anomaly correlation coefficients (ACC) forecast skill in CESM-DPLE to the number of ensemble members and the number of initialization years. Each panel shows the ACCs of SST averaged over the entire domain (solid lines) and just the shelf (dashed lines) for different ensemble sizes, based on three sets of number (N) of initialization years.

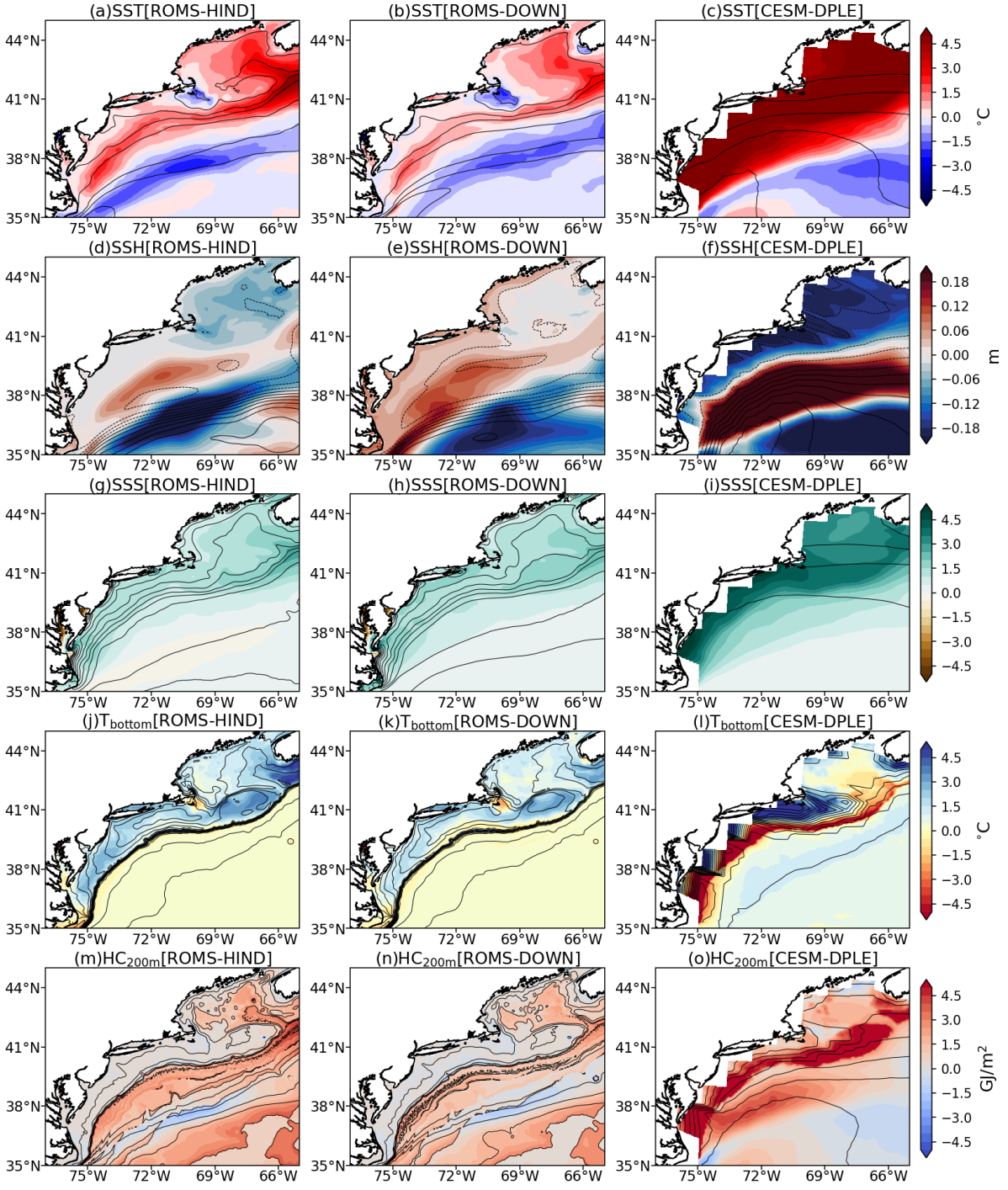


Figure A3. Spatial distribution of mean biases (model minus GLORYS) for (a–c) SST, (d–f) SSH, (g–i) SSS, (j–l) T_{bottom}, and (m–o) HC_{200m} for ROMS-HIND (left column), ROMS-DOWN (middle column), and CESM-DPLE (right column). Biases are calculated using annual-mean fields over the period 2004–2010. Contours indicate the corresponding climatological fields.

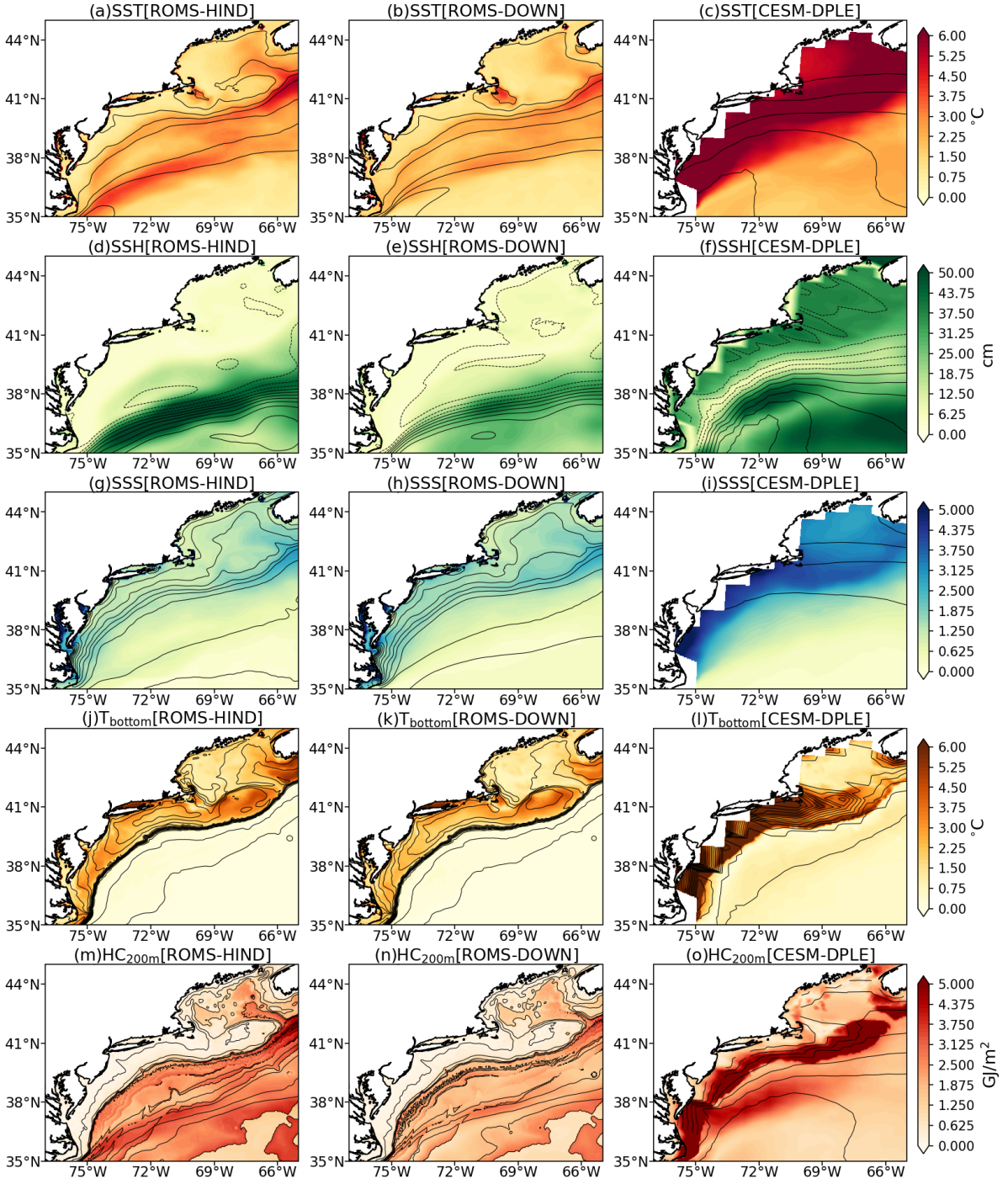


Figure A4. Spatial distribution of temporal root-mean-square error (RMSE) relative to GLORYS for (a–c) SST, (d–f) SSH, (g–i) SSS, (j–l) T_{bottom}, and (m–o) HC_{200m} from ROMS-HIND (left column), ROMS-DOWN (middle column), and CESM-DPLE (right column). RMSE is calculated using annual-mean anomalies over the analysis period. Contours indicate the corresponding climatological fields.

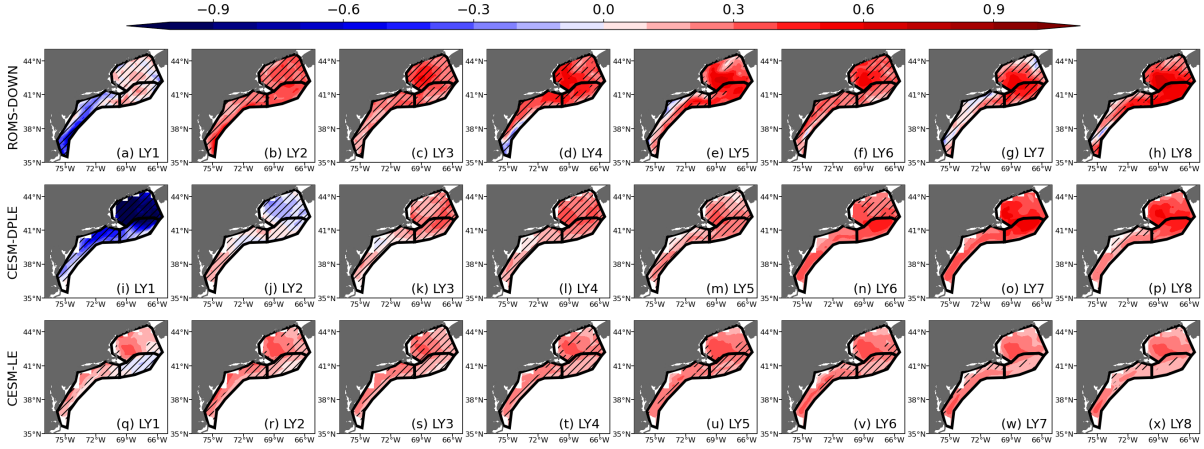


Figure A5. Mean square skill score (MSSS) of SST for lead years 1–8 (LY1–LY8) from (a–h) ROMS-DOWN, (i–p) CESM-DPLE, and (q–x) CESM-LE. MSSS is computed relative to GLORYS reanalysis. Subregions outlined in black represent GOM, GB, and MAB, respectively. Hatched areas indicate regions where the MSSS is not statistically significant at the 95% confidence level.

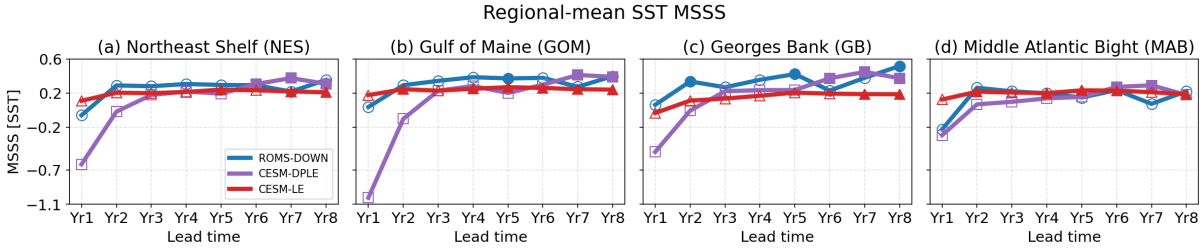


Figure A6. MSSS of the subregion-mean SST as a function of lead year (LY1–LY8) for (a) NES, (b) GOM, (c) GB, and (d) MAB. MSSS values are shown for ROMS-DOWN (blue), CESM-DPLE (purple), and CESM-LE (red). Filled markers denote statistically significant values at the 95% confidence level. Grey dashed line denotes zero MSSS.

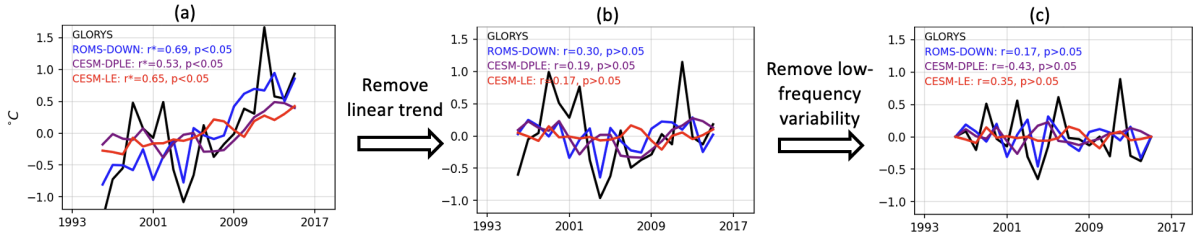


Figure A7. Similar to Figure 4e, but showing the decomposition of the SST anomaly time series. (a) Original SST anomaly time series from GLORYS (black), ROMS-DOWN (blue), CESM-DPLE (purple), and CESM-LE (red) for lead year 5, which are identical to Figure 4e. (b) SST anomaly time series after removing the linear trend, from which the linear-trend contribution to ACC (Figure 11e) is derived. (c) SST anomaly time series containing only the high-frequency component (periods $<$ 10 years) obtained using a high-pass filter, from which the high-frequency contribution to ACC (Figure 11u) is derived. Correlation coefficients (r) and significance levels (p values) between each model and GLORYS are indicated in each panel.

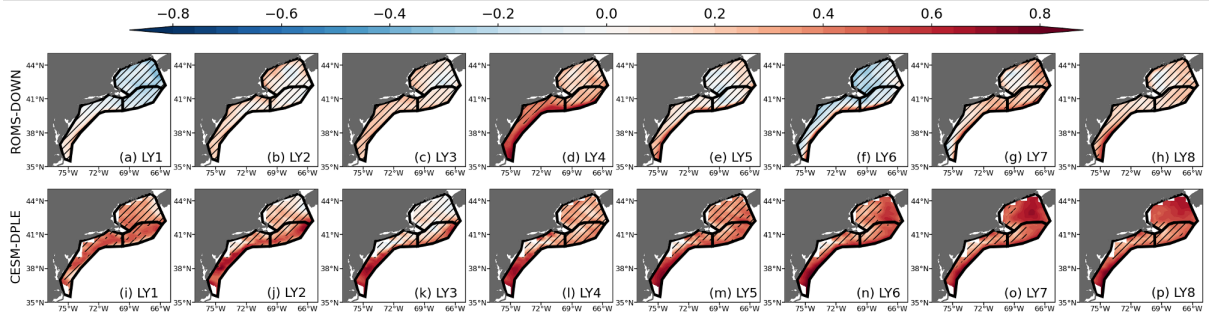


Figure A8. ACC for SSH for lead years 1–8 (LY1–LY8) from (a–h) ROMS-DOWN and (i–p) CESM-DPLE. ACC is computed relative to GLORYS reanalysis. Subregions outlined in black represent GOM, GB, and MAB, respectively. Hatched areas indicate regions where the ACC is not statistically significant at the 95% confidence level.

References

Athanasiadis, P. J., Yeager, S., Kwon, Y.-O., Bellucci, A., Smith, D. W., and Tibaldi, S.: Decadal predictability of North Atlantic blocking and the NAO, NPJ Climate and Atmospheric Science, 3, 20, 2020.

Batteen, M. L. and Miller, H. A.: Process-oriented modeling studies of the 5500-km-long boundary flow off western and southern Australia, Continental Shelf Research, 29, 702–718, 2009.

Bigelow, H. B.: Studies of the waters on the continental shelf, Cape Cod to Chesapeake Bay. I: The cycle of temperature, Pap. Phys. Oceanogr. Meteorol., 2, 135, 1933.

Böhm, E., Hopkins, T., Pietrafesa, L., and Churchill, J.: Continental slope sea level and flow variability induced by lateral movements of the Gulf Stream in the Middle Atlantic Bight, Progress in Oceanography, 70, 196–212, 2006.

Castillo-Trujillo, A. C., Kwon, Y.-O., Fratantoni, P., Chen, K., Seo, H., Alexander, M. A., and Saba, V. S.: An evaluation of eight global ocean reanalyses for the Northeast US Continental shelf, Progress in Oceanography, 219, 103 126, 2023.

Chapman, D. C.: Numerical treatment of cross-shelf open boundaries in a barotropic coastal ocean model, Journal of Physical oceanography, 15, 1060–1075, 1985.

Chapman, D. C. and Beardsley, R. C.: On the origin of shelf water in the Middle Atlantic Bight, Journal of Physical Oceanography, 19, 384–391, 1989.

Chassignet, E. and Marshall, D.: Gulf Stream separation in numerical ocean models, Geophysical Monograph Series, 177, 2008.

Chassignet, E. P. and Xu, X.: Impact of horizontal resolution (1/12 to 1/50) on Gulf Stream separation, penetration, and variability, Journal of Physical Oceanography, 47, 1999–2021, 2017.

Chen, K. and He, R.: Mean circulation in the coastal ocean off northeastern North America from a regional-scale ocean model, Ocean Science, 11, 503–517, 2015.

Chen, K., Gawarkiewicz, G. G., Lentz, S. J., and Bane, J. M.: Diagnosing the warming of the Northeastern US Coastal Ocean in 2012: A linkage between the atmospheric jet stream variability and ocean response, Journal of Geophysical Research: Oceans, 119, 218–227, 2014a.

Chen, K., He, R., Powell, B. S., Gawarkiewicz, G. G., Moore, A. M., and Arango, H. G.: Data assimilative modeling investigation of Gulf Stream Warm Core Ring interaction with continental shelf and slope circulation, Journal of Geophysical Research: Oceans, 119, 5968–5991, 2014b.

Chen, K., Gawarkiewicz, G., and Yang, J.: Mesoscale and submesoscale shelf-ocean exchanges initialize an advective marine heatwave, Journal of Geophysical Research: Oceans, 127, e2021JC017 927, 2022.

Chen, Z. and Curchitser, E. N.: Interannual variability of the Mid-Atlantic Bight cold pool, Journal of Geophysical Research: Oceans, 125, e2020JC016 445, 2020.

Chen, Z., Kwon, Y.-O., Chen, K., Fratantoni, P., Gawarkiewicz, G., and Joyce, T.: Long-term SST variability on the Northwest Atlantic continental shelf and slope, Geophysical Research Letters, 47, e2019GL085 455, 2020.

Christensen, H. M., Berner, J., and Yeager, S.: The value of initialization on decadal timescales: state-dependent predictability in the CESM decadal prediction large ensemble, Journal of Climate, 33, 7353–7370, 2020.

Claret, M., Galbraith, E. D., Palter, J. B., Bianchi, D., Fennel, K., Gilbert, D., and Dunne, J. P.: Rapid coastal deoxygenation due to ocean circulation shift in the northwest Atlantic, Nature climate change, 8, 868–872, 2018.

CMEMS: Global Ocean Physics Reanalysis [Dataset]. E.U. Copernicus Marine Service Information (CMEMS). Marine Data Store (MDS). DOI: https://doi.org/10.48670/moi-00021. [Dataset]., Assessed on 30 Aug 2026a.

CMEMS: Global Ocean Gridded L 4 Sea Surface Heights And Derived Variables Reprocessed 1993 Ongoing. E.U. Copernicus Marine Service Information (CMEMS). Marine Data Store (MDS). DOI: https://doi.org/10.48670/moi-00148. [Dataset]., Assessed on 30 Aug 2026b.

Dai, A.: Hydroclimatic trends during 1950–2018 over global land, Climate Dynamics, 56, 4027–4049, 2021.

Davis, X. J., Joyce, T. M., and Kwon, Y.-O.: Prediction of silver hake distribution on the Northeast US shelf based on the Gulf Stream path index, Continental Shelf Research, 138, 51–64, 2017.

Edson, J. B., Jampana, V., Weller, R. A., Bigorre, S. P., Plueddemann, A. J., Fairall, C. W., Miller, S. D., Mahrt, L., Vickers, D., and Hersbach, H.: On the exchange of momentum over the open ocean, Journal of Physical Oceanography, 43, 1589–1610, 2013.

Egbert, G. D. and Erofeeva, S. Y.: Efficient inverse modeling of barotropic ocean tides, Journal of Atmospheric and Oceanic technology, 19, 183–204, 2002.

Ezer, T.: Can the Gulf Stream induce coherent short-term fluctuations in sea level along the US East Coast? A modeling study, Ocean Dynamics, 66, 207–220, 2016.

Fratantoni, P. S. and Pickart, R. S.: The western North Atlantic shelfbreak current system in summer, Journal of Physical Oceanography, 37, 2509–2533, 2007.

Friedland, K. D., Miles, T., Goode, A. G., Powell, E. N., and Brady, D. C.: The Middle Atlantic Bight Cold Pool is warming and shrinking: indices from in situ autumn seafloor temperatures, Fisheries Oceanography, 31, 217–223, 2022.

Friedland, K. D., Du Pontavice, H., Palter, J., Townsend, D. W., Fratantoni, P., Silver, A., and Gangopadhyay, A.: Regime change in northwest Atlantic sea surface temperatures revealed using a quantile approach, Regional Studies in Marine Science, 71, 103 398, 2024.

Friedland, K. D., Scopel, L. C., Yang, X., Gaichas, S. K., and Rokosz, K. J.: Species richness in the Northeast US Continental Shelf ecosystem: Climate-driven trends and perturbations, PLOS Climate, 4, e0000 557, 2025.

Gartland, J., Gaichas, S. K., and Latour, R. J.: Spatiotemporal patterns in the ecological community of the nearshore Mid-Atlantic Bight, Marine Ecology Progress Series, 704, 15–33, 2023.

Gawarkiewicz, G., Todd, R. E., Zhang, W., Partida, J., Gangopadhyay, A., Monim, M.-U.-H., Fratantoni, P., Mercer, A. M., and Dent, M.: The changing nature of shelf-break exchange revealed by the OOI Pioneer Array, Oceanography, 31, 60–70, 2018.

Gawarkiewicz, G., Fratantoni, P., Bahr, F., and Ellertson, A.: Increasing frequency of mid-depth salinity maximum intrusions in the middle Atlantic bight, Journal of Geophysical Research: Oceans, 127, e2021JC018 233, 2022.

Gawarkiewicz, G. G., Todd, R. E., Plueddemann, A. J., Andres, M., and Manning, J. P.: Direct interaction between the Gulf Stream and the shelfbreak south of New England, Scientific reports, 2, 553, 2012.

Gonçalves Neto, A., Langan, J. A., and Palter, J. B.: Changes in the Gulf Stream preceded rapid warming of the Northwest Atlantic Shelf, Communications Earth & Environment, 2, 74, 2021.

Guo, Y., Bachman, S., Bryan, F., and Bishop, S.: Increasing trends in oceanic surface poleward eddy heat flux observed over the past three decades, Geophysical Research Letters, 49, e2022GL099 362, 2022a.

Guo, Y., Bishop, S., Bryan, F., and Bachman, S.: A global diagnosis of eddy potential energy budget in an eddy-permitting ocean model, Journal of Physical Oceanography, 52, 1731–1748, 2022b.

Guo, Y., Bishop, S., Bryan, F., and Bachman, S.: Mesoscale variability linked to interannual displacement of Gulf Stream, Geophysical Research Letters, 50, e2022GL102 549, 2023.

Guo, Y., Castillo-Trujillo, A. C., Chen, K., Kwon, Y.-O., Perkins, S., Seo, H., Fratantoni, P., Alexander, M., and Saba, V.: Data for "Global ocean pCO2 variation regimes: spatial patterns and the emergence of a hybrid regime". Zenodo. https://doi.org/10.5281/zenodo.20115945. [Dataset]., 2026.

Helmholtz Ctr Ocean Res, G.: North Atlantic simulations in Coordinated Ocean-ice Reference Experiments phase II (CORE-II). Part II: Inter-annual to decadal variability, Ocean Modelling, 97, 65–90, 2016.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., et al.: The ERA5 global reanalysis, Quarterly journal of the royal meteorological society, 146, 1999–2049, 2020.

Hervieux, G., Alexander, M., Stock, C., Jacox, M., Pegion, K., Becker, E., Castruccio, F., and Tommasi, D.: More reliable coastal SST forecasts from the North American multimodel ensemble, Climate Dynamics, 53, 7153–7168, 2019.

Huang, B., Liu, C., Banzon, V., Freeman, E., Graham, G., Hankins, B., Smith, T., and Zhang, H.-M.: Improvements of the daily optimum interpolation sea surface temperature (DOISST) version 2.1, Journal of Climate, 34, 2923–2939, 2021.

Hunke, E. C., Lipscomb, W. H., Turner, A. K., Jeffery, N., and Elliott, S.: Cice: the los alamos sea ice model documentation and software user's manual version 4.1 la-cc-06-012, T-3 Fluid Dynamics Group, Los Alamos National Laboratory, 675, 500, 2010.

Hurrell, J. W., Holland, M. M., Gent, P. R., Ghan, S., Kay, J. E., Kushner, P. J., Lamarque, J.-F., Large, W. G., Lawrence, D., Lindsay, K., et al.: The community earth system model: a framework for collaborative research, Bulletin of the American Meteorological Society, 94, 1339–1360, 2013.

Jacox, M. G., Alexander, M. A., Siedlecki, S., Chen, K., Kwon, Y.-O., Brodie, S., Ortiz, I., Tommasi, D., Widlansky, M. J., Barrie, D., et al.: Seasonal-to-interannual prediction of North American coastal marine ecosystems: Forecast methods, mechanisms of predictability, and priority developments, Progress in Oceanography, 183, 102 307, 2020.

Jean-Michel, L., Eric, G., Romain, B.-B., Gilles, G., Angélique, M., Marie, D., Clément, B., Mathieu, H., Olivier, L. G., Charly, R., et al.: The Copernicus global 1/12 oceanic and sea ice GLORYS12 reanalysis, Frontiers in Earth Science, 9, 698 876, 2021.

Karmalkar, A. V. and Horton, R. M.: Drivers of exceptional coastal warming in the northeastern United States, Nature Climate Change, 11, 854–860, 2021.

Kay, J. E., Deser, C., Phillips, A., Mai, A., Hannay, C., Strand, G., Arblaster, J. M., Bates, S., Danabasoglu, G., Edwards, J., et al.: The Community Earth System Model (CESM) large ensemble project: A community resource for studying climate change in the presence of internal climate variability, Bulletin of the American Meteorological Society, 96, 1333–1349, 2015.

Kelly, K. A., Small, R. J., Samelson, R., Qiu, B., Joyce, T. M., Kwon, Y.-O., and Cronin, M. F.: Western boundary currents and frontal air–sea interaction: Gulf Stream and Kuroshio Extension, Journal of Climate, 23, 5644–5667, 2010.

Koul, V., Ross, A. C., Stock, C., Zhang, L., Delworth, T., and Wittenberg, A.: A predicted pause in the rapid warming of the Northwest Atlantic Shelf in the coming decade, Geophysical Research Letters, 51, e2024GL110 946, 2024.

Kwon, Y.-O., Alexander, M. A., Bond, N. A., Frankignoul, C., Nakamura, H., Qiu, B., and Thompson, L. A.: Role of the Gulf Stream and Kuroshio–Oyashio systems in large-scale atmosphere–ocean interaction: A review, Journal of Climate, 23, 3249–3281, 2010.

Lawrence, D. M., Oleson, K. W., Flanner, M. G., Thornton, P. E., Swenson, S. C., Lawrence, P. J., Zeng, X., Yang, Z.-L., Levis, S., Sakaguchi, K., et al.: Parameterization improvements and functional and structural advances in version 4 of the Community Land Model, Journal of Advances in Modeling Earth Systems, 3, 2011.

Lellouche, J.-M., Bourdalle-Badie, R., Greiner, E., Garric, G., Melet, A., Bricaud, C., Legalloudec, O., Hamon, M., Candela, T., Regnier, C., et al.: The Copernicus global 1/12° oceanic and sea ice reanalysis, in: EGU General Assembly Conference Abstracts, pp. EGU21–14 961, 2021.

Lentz, S. J.: Seasonal warming of the Middle Atlantic Bight Cold Pool, Journal of Geophysical Research: Oceans, 122, 941–954, 2017.

Link, J. S., Brodziak, J. K., Edwards, S. F., Overholtz, W. J., Mountain, D., Jossi, J. W., Smith, T. D., and Fogarty, M. J.: Marine ecosystem assessment in a fisheries management context, Canadian Journal of Fisheries and Aquatic Sciences, 59, 1429–1440, 2002.

Loder, J. W., Petrie, B., and Gawarkiewicz, G.: The coastal ocean of Northeastern North America: A large-scale view (1, W), The sea: ideas and observations on progress in the study of the seas, 1998.

Lucey, S. M. and Nye, J. A.: Shifting species assemblages in the northeast US continental shelf large marine ecosystem, Marine Ecology Progress Series, 415, 23–33, 2010.

Mason, E., Molemaker, J., Shchepetkin, A. F., Colas, F., McWilliams, J. C., and Sangrà, P.: Procedures for offline grid nesting in regional ocean models, Ocean Modelling, 35, 1–15, https://doi.org/10.1016/j.ocemod.2010.05.007, 2010.

McHenry, J., Welch, H., Lester, S. E., and Saba, V.: Projecting marine species range shifts from only temperature can mask climate vulnerability, Global Change Biology, 25, 4208–4221, 2019.

Meehl, G. A., Goddard, L., Murphy, J., Stouffer, R. J., Boer, G., Danabasoglu, G., Dixon, K., Giorgetta, M. A., Greene, A. M., Hawkins, E., et al.: Decadal prediction: can it be skillful?, Bulletin of the American Meteorological Society, 90, 1467–1486, 2009.

Meehl, G. A., Richter, J. H., Teng, H., Capotondi, A., Cobb, K., Doblas-Reyes, F., Donat, M. G., England, M. H., Fyfe, J. C., Han, W., et al.: Initialized Earth System prediction from subseasonal to decadal timescales, Nature Reviews Earth & Environment, 2, 340–357, 2021.

Miles, T., Murphy, S., Kohut, J., Borsetti, S., and Munroe, D.: Offshore wind energy and the Mid-Atlantic Cold Pool: a review of potential interactions, Marine Technology Society Journal, 55, 72–87, 2021.

Mountain, D. G.: Labrador slope water entering the Gulf of Maine—response to the North Atlantic Oscillation, Continental Shelf Research, 47, 150–155, 2012.

Nye, J. A., Joyce, T. M., Kwon, Y.-O., and Link, J. S.: Silver hake tracks changes in Northwest Atlantic circulation, Nature communications, 2, 412, 2011.

O'Kane, T. J., Scaife, A. A., Kushnir, Y., Brookshaw, A., Buontempo, C., Carlin, D., Connell, R. K., Doblas-Reyes, F., Dunstone, N., Förster, K., et al.: Recent applications and potential of near-term (interannual to decadal) climate predictions, Frontiers in Climate, 5, 1121 626, 2023.

Park, S., Bretherton, C. S., and Rasch, P. J.: Integrating cloud processes in the Community Atmosphere Model, version 5, Journal of Climate, 27, 6821–6856, 2014.

Pershing, A. J., Alexander, M. A., Hernandez, C. M., Kerr, L. A., Le Bris, A., Mills, K. E., Nye, J. A., Record, N. R., Scannell, H. A., Scott, J. D., et al.: Slow adaptation in the face of rapid warming leads to collapse of the Gulf of Maine cod fishery, Science, 350, 809–812, 2015.

Pringle, J. M.: Sources of variability in Gulf of Maine circulation, and the observations needed to model it, Deep Sea Research Part II: Topical Studies in Oceanography, 53, 2457–2476, 2006.

Reynolds, R. W., Rayner, N., Smith, T., Stokes, D., and Wang, W.: NOAA Optimum Interpolation (OI) Sea Surface Temperature (OISST) Analysis, Version 2 [Dataset]. https://psl.noaa.gov/data/gridded/data.noaa.oisst.v2.html, 2002.

Reynolds, R. W., Smith, T. M., Liu, C., Chelton, D. B., Casey, K. S., and Schlax, M. G.: Daily high-resolution-blended analyses for sea surface temperature, Journal of climate, 20, 5473–5496, 2007.

Ross, A. C., Stock, C. A., Adcroft, A., Curchitser, E., Hallberg, R., Harrison, M. J., Hedstrom, K., Zadeh, N., Alexander, M., Chen, W., et al.: A high-resolution physical-biogeochemical model for marine resource applications in the Northwest Atlantic (MOM6-COBALT-NWA12 v1. 0), Geoscientific Model Development Discussions, 2023, 1–65, 2023.

Ross, A. C., Stock, C. A., Koul, V., Delworth, T. L., Lu, F., Wittenberg, A., and Alexander, M. A.: Dynamically downscaled seasonal ocean forecasts for North American East Coast ecosystems, Ocean Science, 20, 1631–1656, 2024.

Ryan, S., Ummenhofer, C. C., and Gawarkiewicz, G. G.: Seasonal and interannual salinity variability on the Northeast US continental shelf: Insights from satellite sea surface salinity and implications for stratification, Journal of Geophysical Research: Oceans, 129, e2024JC021 534, 2024.

Saba, V. S., Griffies, S. M., Anderson, W. G., Winton, M., Alexander, M. A., Delworth, T. L., Hare, J. A., Harrison, M. J., Rosati, A., Vecchi, G. A., et al.: Enhanced warming of the Northwest Atlantic Ocean under climate change, Journal of Geophysical Research: Oceans, 121, 118–132, 2016.

Seidov, D., Mishonov, A., and Parsons, R.: Recent warming and decadal variability of Gulf of Maine and Slope Water, Limnology and Oceanography, 66, 3472–3488, 2021.

Shchepetkin, A. F. and McWilliams, J. C.: The regional oceanic modeling system (ROMS): a split-explicit, free-surface, topography-following-coordinate oceanic model, Ocean modelling, 9, 347–404, 2005.

Smith, R., Jones, P., Briegleb, B., Bryan, F., Danabasoglu, G., Dennis, J., Dukowicz, J., Eden, C., Fox-Kemper, B., Gent, P., et al.: The parallel ocean program (POP) reference manual: Ocean component of the community climate system model (CCSM), Rep. LAUR-01853, 141, 1–141, 2010.

Townsend, D. W., Rebuck, N. D., Thomas, M. A., Karp-Boss, L., and Gettings, R. M.: A changing nutrient regime in the Gulf of Maine, Continental Shelf Research, 30, 820–832, 2010.

Wallace, E. J., Looney, L. B., and Gong, D.: Multi-decadal trends and variability in temperature and salinity in the Mid-Atlantic Bight, Georges Bank, and Gulf of Maine, Journal of Marine Research, 76, 163, 2018.

Xu, H., Kim, H.-M., Nye, J. A., and Hameed, S.: Impacts of the North Atlantic Oscillation on sea surface temperature on the Northeast US Continental Shelf, Continental Shelf Research, 105, 60–66, 2015.

Yeager, S., Danabasoglu, G., Rosenbloom, N., Strand, W., Bates, S., Meehl, G., Karspeck, A., Lindsay, K., Long, M., Teng, H., et al.: Predicting near-term changes in the earth system: a large ensemble of initialized decadal prediction simulations using the community earth system model, Bulletin of the American Meteorological Society, 99, 1867–1886, 2018.

Yeager, S. G., Karspeck, A. R., and Danabasoglu, G.: Predicted slowdown in the rate of Atlantic sea ice loss, Geophysical Research Letters, 42, 10–704, 2015.

Yeager, S. G., Rosenbloom, N., Glanville, A. A., Wu, X., Simpson, I., Li, H., Molina, M. J., Krumhardt, K., Mogen, S., Lindsay, K., et al.: The seasonal-to-multiyear large ensemble (SMYLE) prediction system using the Community Earth System Model version 2, Geoscientific Model Development, 15, 6451–6493, 2022.

Yeager, S. G., Chang, P., Danabasoglu, G., Rosenbloom, N., Zhang, Q., Castruccio, F. S., Gopal, A., Cameron Rencurrel, M., and Simpson, I. R.: Reduced Southern Ocean warming enhances global skill and signal-to-noise in an eddy-resolving decadal prediction system, npj Climate and Atmospheric Science, 6, 107, 2023.