JGR Oceans

RESEARCH ARTICLE
10.1029/2021JC017545

Special Section:
Advancing prediction of coastal marine ecosystems

Seasonal Predictability of Sea Ice and Bottom Temperature Across the Eastern Bering Sea Shelf

K. A. Kearney[1,2], M. Alexander[3], K. Aydin[2], W. Cheng[1,4], A. J. Hermann[1,4], G. Hervieux[3,5], and I. Ortiz[1,2]

[1]University of Washington, Cooperative Institute for Climate Ocean and Ecosystem Studies, Seattle, WA, USA, [2]NOAA Alaska Fisheries Science Center, Seattle, WA, USA, [3]NOAA Earth System Research Laboratory, Boulder, CO, USA, [4]NOAA Pacific Marine Environmental Laboratory, Seattle, WA, USA, [5]University of Colorado, Cooperative Institute for Research in Environmental Sciences, Boulder, CO, USA

Key Points:

- Dynamic forecasting with a regional model can predict summer bottom temperature on the eastern Bering Sea shelf 3–4 months in advance
- Most of the bottom temperature predictability comes from persistence; a multi-model dynamic forecast may extend predictability 1–2 months
- Sea ice presents a prediction barrier, with low sea ice and bottom temp. Skill if forecasts begin before or during the ice season (October–February)

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

K. A. Kearney,
kelly.kearney@noaa.gov

Citation:

Kearney, K. A., Alexander, M., Aydin, K., Cheng, W., Hermann, A. J., Hervieux, G., & Ortiz, I. (2021). Seasonal predictability of sea ice and bottom temperature across the eastern Bering Sea shelf. *Journal of Geophysical Research: Oceans*, *126*, e2021JC017545. https://doi.org/10.1029/2021JC017545

Received 5 MAY 2021
Accepted 18 OCT 2021

Abstract Seasonal sea ice plays a key role in shaping the ecosystem dynamics of the eastern Bering Sea shelf. In particular, it leads to the formation of a characteristic pool of cold water that covers the bottom of the shelf from winter through summer; the extent of this cold pool is often used as a management index for distribution, productivity, recruitment, and survival of commercially important fish and shellfish species. Here, we quantify our ability to seasonally forecast interannual variability in Bering Sea bottom temperature and sea ice extent. Retrospective forecast simulations from two global forecast models are downscaled using a regional ocean model; the retrospective forecast simulations include 9-month to 12-month forecasts spanning 1982–2010. We find that dynamic forecasting can predict summer bottom temperatures across the eastern Bering Sea shelf with lead times of up to 4 months. The majority of the prediction skill derives from the persistence signal, and a persistence forecast is comparably skillful to the dynamic forecast at these lead times. However, forecast skill of sea ice advance and retreat is low when a forecast model is initialized before or during the ice season (October–February); this limits the ability of either dynamic or persistence models to predict summer bottom temperatures when initialized across the late fall to early spring months.

Plain Language Summary The Bering Sea region is home to a large number of important fisheries. Fluctuations in fish distribution, abundance, and productivity can be influenced by environmental factors; in the Bering Sea, the temperature of the deep shelf water can affect where groundfish are located, how much of their preferred prey is available, and how many juvenile fish and shellfish survive each year. In this study, we test whether we can successfully predict whether bottom water will be colder or warmer than average in a given year. We do this by running forecast models for previous years and comparing the model output to observations collected during the same period. We find that we can skillfully predict summer bottom temperature if we start a forecast four or fewer months in advance of the summer period. However, beginning the forecast during the winter (when sea ice is present) or during the prior fall results in low skill. The 4-month forecasts can be useful when planning for summer surveys and to provide advance notice to the North Pacific Fishery Management Council of any unusual summer conditions that may affect the quotas they set for the upcoming fishing seasons.

1. Introduction

The eastern Bering Sea shelf is home to a highly productive and commercially important ecosystem. The wide, shallow shelf, long growing season, tidal mixing, and ice-influenced stratification lead to high primary productivity across the shelf and slope, which in turn supports productivity in both pelagic and benthic ecosystems. The high productivity supports a wide variety of both subsistence and commercial fisheries, including nearly half of U.S. landings (Fissel et al., 2017; National Marine Fisheries Service, 2017).

Sea ice plays a key role in shaping the ecosystems of the Bering Sea. Influenced by northeasterly winds, ice advances via both local formation and advective processes from the Bering Strait region into the Bering Sea, with much of the eastern shelf at least partially covered by ice beginning in late fall (October–November) through early spring (March–April). The timing of ice onset and retreat and extent of sea ice can vary significantly from year to year depending on the position and strength of the Aleutian Low and Siberian High

pressure systems, as well as ocean conditions (Stabeno et al., 2001). As ice advances, the freezing process and resulting brine rejection lead to the formation of cold, salty, dense bottom water underneath the ice (Stabeno et al., 2001). In the spring, warming of the surface waters and melting of sea ice sharply stratify the water column over much of the shelf region, isolating this bottom water from surface heating and mixing. As a result, this signature cold water mass, referred to as the cold pool, can persist well into the summer months (Stabeno et al., 2001; Zhang et al., 2012).

Bottom temperature across the Bering Sea shelf is often used as an ecosystem status indicator for living marine resource management (Mueter & Litzow, 2008). In particular, a metric known as the cold pool index, defined as the proportion of the southeastern shelf that has a bottom temperature below a particular threshold (0–2°C depending on application), has been selected by stock assessment authors and other experts as a key indicator related to the growth and recruitment of several managed species (O'Leary et al., 2020; Thorson, Ciannelli & Litzow, 2020; Thorson et al., 2021). As part of the ecosystem-based fisheries management approach in Alaska, the cold pool extent is specifically included in the annual ecosystem status report for the eastern Bering Sea, which summarizes environmental trends and their links to fishery management (Zador et al., 2017). The usefulness of the cold pool index stems from its relationship to both direct and indirect drivers of fish recruitment and survival. The size and location of the cold pool influence vertical mixing and stratification in the water column (Stabeno et al., 2012), and also correlates with the latitudinal and longitudinal distribution of the groundfish community, including several important commercial species (Kotwicki et al., 2005; Mueter & Litzow, 2008; Spencer, 2008; Stevenson & Lauth, 2019); this, in turn, changes the spatial distribution of fishing effort (Haynie & Pfeiffer, 2012).

The large spatial extent, remote location, and presence of seasonal sea ice have historically restricted the collection of environmental data on the eastern Bering Sea shelf to the late spring to early fall months. Recent studies have begun using novel technology, such as sail drones (Mordy et al., 2017), profiling moorings (prawlers) (Duffy-Anderson et al., 2019; Stabeno et al., 2017), and under-ice pop-up buoys (Stabeno et al., 2020) to increase observations year round. In addition, hydrodynamic modeling can supplement these new observing systems. For the past decade, a regional ocean model of the Bering Sea driven by historical reanalysis data has been used to simulate the physical and biogeochemical properties of the shelf region, providing more complete spatial and temporal coverage than is possible through direct observation. These reanalysis-driven hindcast simulations capture many of the observed horizontal and vertical patterns of water movement, mixing, and stratification, as well as the temperature and salinity signatures of various water masses throughout the Bering Sea (Hermann et al., 2016; Kearney, 2021; Kearney et al., 2020). In addition, hindcasts of the cold pool index have shown strong predictive power for the recruitment of economically important Bering Sea groundfish (Eisner et al., 2020; Holsman et al., 2016; Thorson, Cheng, et al., 2020).

Given the success of these hindcast simulations, the possibility of seasonal predictions of the cold pool has generated substantial interest within the Bering Sea fisheries management community. Currently, stock assessments of Bering Sea groundfish rely heavily on data collected during the annual summer groundfish survey, a bottom trawl survey that samples the number, size, condition, and diet of many commercially important groundfish and shellfish, alongside environmental data like surface and bottom temperature (Stevenson & Lauth, 2012). The surveys take place from June to August, leaving a short window to analyze and incorporate the new data into the stock assessment models; the data and models must be completed by October for the North Pacific Fishery Management Council to set management quotas for the upcoming calendar year. Predictions of cold pool extent provided in advance of the annual summer surveys could be used to guide upcoming surveys and sampling plans, and they could provide early warnings and additional context for potential environmental changes in the survey-derived environmental data that informs management decisions.

In this study, we use a series of 9-month to 12-month retrospective forecast simulations to assess the seasonal predictability of biologically important physical properties across the eastern Bering Sea shelf. We primarily focus on bottom temperature due to its widespread use within stock assessments and other fisheries-related models in this region, for example, those used to estimate essential fish habitat (Laman et al., 2018). We also examine related properties, including sea ice advance and retreat.

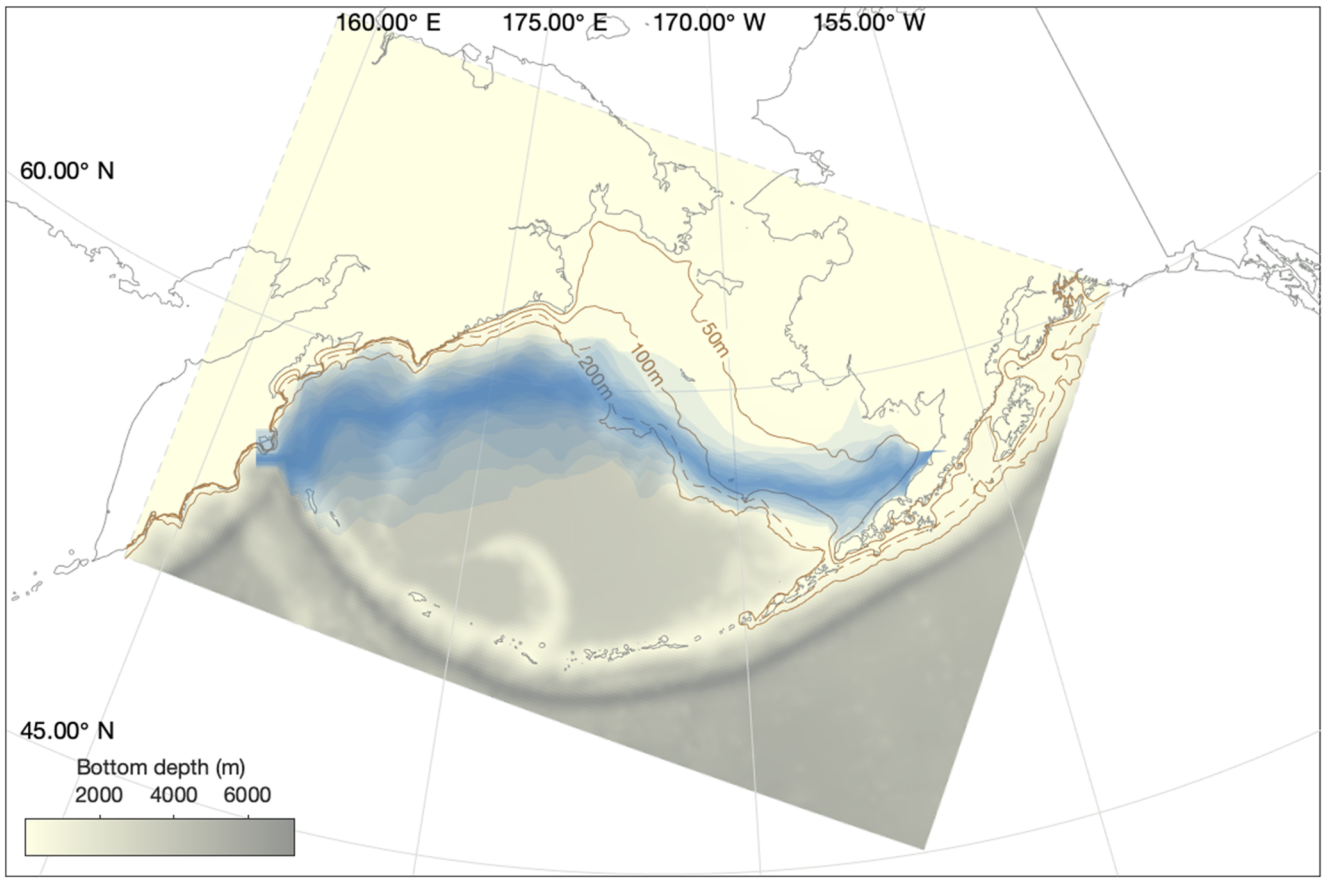

Figure 1. Map of the Bering Sea region, with bathymetry from the Bering10K model. Blue shading indicates the range of maximum ice extent between 1980 and 2018; the darkest blue region encapsulates the 45th–55th percentile, and the lightest the full range. Brown contours indicate the 50-, 100-, and 200-m isobaths; the dashed line is the Bering10K 200-m isobath.

2. Methods

2.1. Study Domain and Dynamical Downscaling

Our study focuses on the eastern Bering Sea shelf (Figure 1). This region consists of a wide, shallow continental shelf, approximately 500 km across and with an average depth of 70 m. The southeastern shelf can be subdivided into three regions with distinct patterns of mixing and stratification; the inner domain (<50 m) is well-mixed year-round, the middle domain (50–100 m) is well-mixed during the winter but thermally stratified in spring and summer, and the outer domain (100–200 m) experiences seasonal stratification similar to deeper oceanic waters (Coachman, 1986; Kachel et al., 2002).

With horizontal resolution on the order of a degree, most global-scale models do not fully resolve the processes leading to the formation of the cold pool on the Bering Sea shelf. They tend to underestimate the interannual variability in temperature across this shelf region. In particular, coarse-resolution global-scale models fail to produce shelf bottom water with temperatures below approximately 2°C (Kearney et al., 2020). To better capture the evolution of the physical processes that are important to the formation of the cold pool, we use a high-resolution regional model, known as Bering10K, to dynamically downscale a number of coarser-resolution global-scale models. Bering10K is an implementation of the Regional Ocean Modeling System (ROMS), a free-surface, primitive equation hydrographic model (Haidvogel et al., 2008; Shchepetkin & McWilliams, 2005). Its domain spans the Bering Sea and the northern Gulf of Alaska, with 10-km horizontal resolution and 30 terrain-following depth levels. For this study, we use a Bering10K variant that includes ocean and sea ice modules, with the fully coupled exchange of heat and freshwater fluxes

between these two components. Further details of the model implementation and performance can be found in Hermann et al. (2016) and Kearney et al. (2020).

The dynamical downscaling process involves using atmospheric and ocean output from a global-scale model as surface and lateral boundary conditions, respectively, for the regional model. In this study, we apply this dynamic downscaling to several sets of global-scale models, including reanalysis-driven historical hindcasts and retrospective seasonal forecasts. In all cases, bulk formulae were used to relate atmospheric forcing from the atmospheric component of the global models to surface stress, freshwater fluxes, and heat fluxes (Fairall et al., 1996). Current velocities, temperature, salinity, and sea surface height from the global models provided lateral boundary conditions for the open southern and eastern boundaries of the model domain, applied with a hybrid nudging–radiation scheme (Marchesiello et al., 2001). Freshwater runoff due to river input was reconstructed from observed river discharge from Alaskan and Russian rivers (Kearney, 2019); river freshwater input adds an additional surface freshwater flux (in addition to precipitation), and is distributed across model grid points near the coast based on river mouth location with an e-folding scale of 20 km.

The first simulation used within this study, referred to as the hindcast simulation, runs from 1982 to 2018 under forcing from the Climate Forecast System Reanalysis (CFSR) (Saha et al., 2010) data set. This simulation is used to demonstrate the Bering10K model's ability to simulate past interannual variability in bottom temperature, for initialization of the retrospective forecast simulations, and as a reference point for forecast skill assessment. Detailed information regarding this simulation's skill in capturing observed surface and bottom temperature can be found in Kearney (2021); we also present a short summary of this validation in Section 2.3.2.

The remaining simulations are seasonal retrospective forecast (henceforth abbreviated as “reforecast”) simulations. This term refers to a forecast simulation initialized from a historical date rather from the present, allowing comparison of the forecast results to observations for skill assessment purposes. Our downscaled simulations were driven by the global-scale reforecasts from a subset of the North American Multi-Model Ensemble (NMME) (Kirtman et al., 2014). While six modeling centers contributed at least one forecast model to the NMME suite of seasonal-to-interannual forecasts, our use was limited to those that archived the atmospheric and ocean variables necessary to perform dynamical downscaling. Two parent models met these criteria: CFSv2 (National Centers for Environmental Prediction, Climate Forecast System, version 2) (Saha et al., 2014) and Canadian Meteorological Center, Canadian Coupled Climate Model, version 4 (CanCM4) (Merryfield et al., 2013). For the NMME experiment, both models contributed retrospective forecast simulations spanning 1982 through 2010. The CFSv2 simulations spanned a forecast period of 9 months, while the CanCM4 simulations extended to 12 months. From each parent model, we downscaled simulations initialized at the beginning of each of the 12 months per year within the 29-year reforecast period, totaling 348 initialization dates. We downscaled three ensemble members associated with each initialization date. The CanCM4 model provided 10 ensemble members for each initialization date, with reforecasts initialized on the first of each month and with the ensemble members differentiated following the initialization method in Merryfield et al. (2013). We downscaled ensemble members 1–3 from each initialization date. The CFS model ran their reforecast simulations every 5 days, differentiating ensemble members by staggering the forecast initialization times by 6 hr; we downscaled the initialization dates corresponding to the closest date preceding the first of each month and using the 0600, 1200, and 1800 UTC start times. In total, this sums to 1044 Bering10K simulations per parent model (29 years × 12 monthly initialization dates within each year × 3 ensemble members per initialization date).

As in the hindcast simulations, atmospheric surface boundary conditions and lateral ocean boundary conditions were extracted from the relevant parent model output variables. Bias correction was used to remove systematic biases from the atmospheric and oceanic boundary data extracted from the two parent models before being used within the regional ROMS configuration. For both parent models, bias was calculated by comparing the mean climatological forecast state of each initialization month/ensemble member/lead time combination to a reanalysis-based monthly climatological data set. For CanCM4, bias was evaluated using ERA-interim (Berrisford, Dee, et al., 2011; Berrisford, Kållberg, et al., 2011) and SODA 3.4.1 (Carton et al., 2019; Carton, Chepurin, & Chen, 2018; Carton, Chepurin, Chen & Grodsky, 2018) as historical reference data sets for the atmospheric and oceanic variables, respectively, with forecast climatologies

constructed from the CanCM4 reforecast simulation input data sets from 1982-2010. For CFS, we used preexisting monthly climatological data products derived from the CFSR and CFS reforecast simulations over the 1999–2010 period (Saha et al., 2014); these were available for both the atmospheric and oceanic variables. The resulting spatially varying bias correction data sets were then applied to all 29 years worth of reforecast simulations. Neither parent model provided estimates of reforecasted freshwater runoff, so for river input forcing, we used a climatologically averaged version of the river data set from our hindcast, applying the same monthly mean values from the 50-year Kearney (2019) data set to all reforecast simulations.

Forecast initialization fields for all state variables, including ocean and sea ice fields, were extracted from the Bering10K hindcast simulation at the corresponding initialization date. We opted to use the Bering10K hindcast state, as opposed to ocean and sea ice state variables from the global-scale parent models, for a few reasons. First, previous comparisons between both the downscaled, CFSR-forced Bering10K hindcast and the coarser resolution CFSR ocean model itself indicated that the coarse resolution model was unable to capture some of the key characteristics of the cold pool on the shelf (Kearney et al., 2020). Most notably, the coarse resolution model produced a much more diffuse cold pool with bottom water temperatures that never dropped much below 2°C (note that while CFSR does assimilate temperature profile data, only a small number of these profiles are found in the Bering shelf region; Saha et al., 2010). Because seasonal predictions derive a large part of their skill from initial conditions, we opted to use the model that most closely matched the observations. Use of the hindcast for initial conditions also eliminates any spinup effects due to mismatch between initial conditions and the preferred state of the downscaled ocean model. Finally, from a practical standpoint, use of the hindcast model minimized the amount of ocean data that needed to be retrieved and reformatted from the parent models. We do note that this initialization choice may introduce some mismatch between the atmospheric forcing and ocean state, particularly for the CanCM4-forced simulations. An ideal setup would use a downscaled CanCM4 hindcast simulation for those reforecasts (using the assimilative portions of each CanCM4 ensemble member), but that data set was not easily accessible at the time of this study.

2.2. Validation Data Sets

Surface and bottom temperature measurements are sampled annually each summer across the eastern Bering Sea shelf (EBS) as part of the National Marine Fisheries Service (NMFS) Eastern Bering Sea Continental Shelf Bottom Trawl Survey. This data set, which spans 1982 through 2019, provides the most continuous spatially resolved data set of bottom temperature available for the Bering Sea shelf region where our study is focused. Net trawls are conducted at fixed survey stations located across the shelf at 20 nautical mile resolution (Figure 2a). From 1982 to 1989, temperature data were collected via expendable bathythermographs (XBTs). More recent surveys use digital bathythermograph recorders attached to the headrope of the bottom trawl net (BRANCKER RBR XL-200 Micro BTs recorded at 6-s intervals for the 1993–2001 surveys, and a Sea-Bird SBE-39 bathythermograph continuous data recorder at 3-s intervals for 2002–present). Temperature is then averaged over the on-bottom and surface portions of the trawl (which covers a mean distance of 2.75 n. mi) to produce a single value for bottom and surface temperature per station per year. See Buckley et al. (2009) and Lauth et al. (2019) for full details of temperature data collection and post-processing. This data set was used directly in our study to assess survey-replicated forecast skill (see Section 2.3.3), and also as a validation data set for hindcast simulations (see Section 2.3.1).

While the trawl-derived data set provides the most continuous spatially resolved record of EBS bottom temperatures, it measures only summer conditions, with sampling dates spread over a 3-month period each year. Survey station locations and approximate sampling order remain similar from year to year, but actual sampling date can still vary significantly across years (31 ± 17 days, depending on the station), particularly in the northern part of the shelf (Kearney, 2021). This makes it difficult to disentangle the temporal factors from spatial variations when quantifying summer-forecast model skill.

Because of the lack of a consistent, seasonally resolved bottom temperature data set across the spatial and temporal domain that would be required to conduct a thorough, year-round skill assessment, we instead rely on the Bering10K model hindcast simulation (1982–2010) as the primary standard against which forecast skill is assessed. We elaborate on that decision in the following sections.

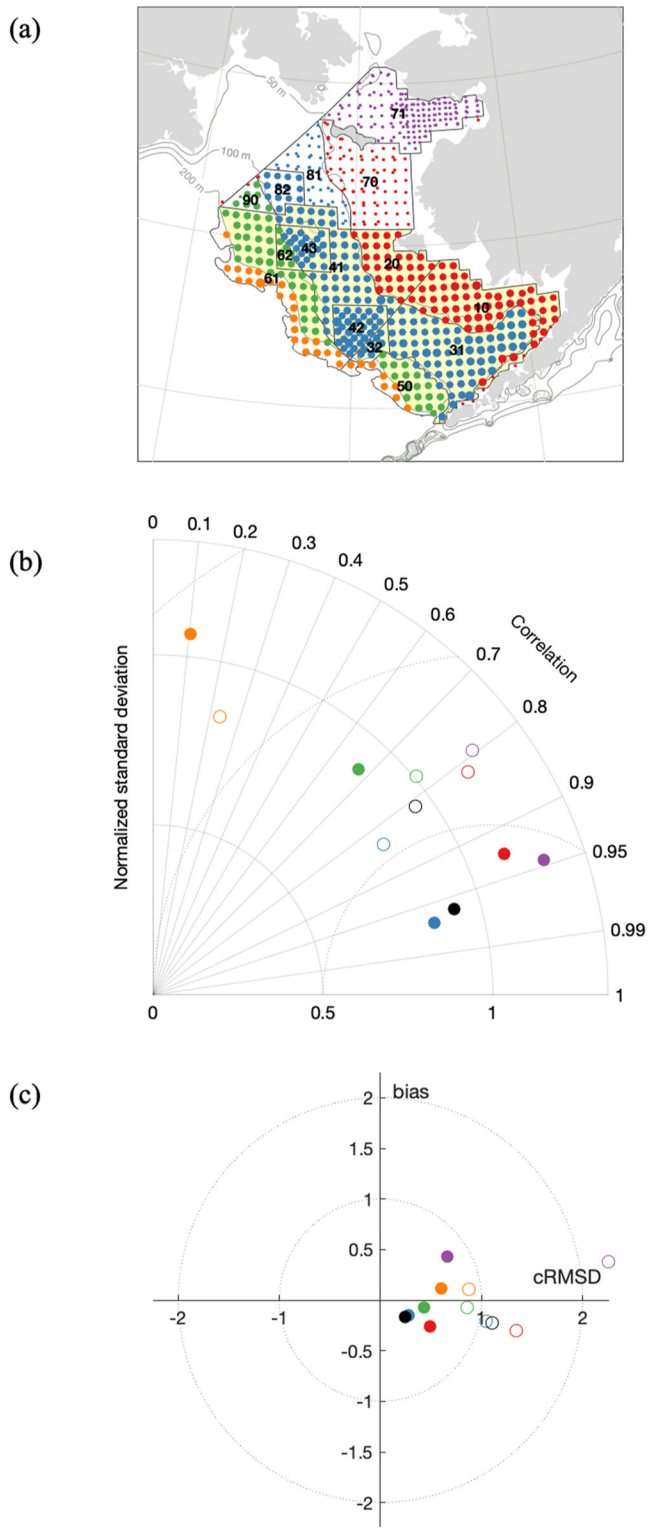
(a)
50 m
100 m
200 m
74
81
70
82
90
43
62
41
20
61
10
42
32
31
50
(b)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
0.95
0.99
1
Correlation
Normalized standard deviation
0
0.5
1
(c)
bias
cRMSD
2
1.5
1
0.5
-0.5
-1
-1.5
-2
-2
-1
1
2

Figure 2.

2.3. Seasonal Forecast Skill Assessment

Forecast model skill is typically quantified by comparing reforecast output directly to observed data or to a gridded reanalysis product (e.g., Jacox et al., 2017; Stock et al., 2015). However, as stated in the previous section, measurements of Bering Sea bottom temperature are limited, and the non-synoptic nature of those measurements introduce additional complexities when used to assess forecast model skill. Therefore, our skill assessment is divided into three steps.

First, we quantify the Bering10K hindcast simulation's skill in reproducing the mean and interannual variability in the size, location, and intensity of the cold pool relative to that seen in the survey-measured temperatures described in the previous section. This step summarizes a more thorough assessment of surface and bottom temperature in the Bering10K hindcast simulations as described in Kearney et al. (2020) and Kearney (2021). We repeat those results here because (a) the results provide the justification necessary to use the hindcast as a forecast validation data set, (b) regional variations in hindcast skill influence which regions we focus on in the forecast skill assessment, and (c) regional variation in survey sampling patterns, noted during the hindcast skill assessment, influence the interpretation of forecast skill as measured against the survey data.

The second part of our skill assessment quantifies the predictive skill of the two dynamic forecast models using the hindcast simulation as the validation data set. When using this model-derived data set as our measure of "truth," we are quantifying the ability of the parent forecast models to provide the large-scale information necessary for a skillful regional forecast, assuming that the dynamical downscaling provided by the Bering10K model serves only to resolve higher-resolution features and does not affect the forecast skill; in other words, it assumes that the downscaling step is perfect. The majority of our results focus on this set of skill metrics since they can be calculated for any target date and lead time and they cover the entire model domain.

The final part of our skill assessment quantifies the skill of the two dynamic forecast models using the survey data itself as the validation data set. Because survey data is only collected in the summer, this skill assessment is limited to forecast target dates that match the survey sample collection dates, and the skill can only be spatially resolved across the portion of the shelf where data has been regularly collected.

Further details of the metrics used for each of these three analyses are as follows.

2.3.1. Hindcast Validation Metrics

Surface and bottom temperature values within the Bering10K hindcast simulation have previously been validated against the groundfish trawl-derived temperature data set described in Section 2.2 (Kearney et al., 2020; Kearney, 2021). Because these results are key to evaluating the skill of this seasonal re-forecast experiment, we repeat the primary methods and results from these previous validation studies for context here.

To assess the skill of the hindcast simulation relative to the groundfish trawl temperature data, bottom temperature points were subsampled in time and space according to the trawl survey; that is, we extract data from the hindcast simulation using the closest model grid point (10-km resolution) and closest output time step (weekly resolution) to each data point in the groundfish data set. Model bottom temperature was defined as the mean temperature over the bottom 5 m of each model grid cell. This definition accounts for the time-varying thickness of the simulated vertical coordinate layers, which adjust in response to variations in bottom depth and the height of the free surface. The 5-m bottom layer thickness approximates the

Figure 2. The groundfish survey sites (panel (a), colored dots scaled by number of times sampled) were divided into five regions: inner shelf (red), middle shelf (blue), outer shelf (green), shelf break (orange), and northern (purple). Also shown are the survey strata polygons (dark gray, labeled by strata number) and model isobaths. Skill metrics for hindcasted bottom temperature versus observations are depicted through (b) Taylor and (c) target diagram; the former summarizes the relationship between normalized standard deviation (radial axis), correlation (angular axis), and centered RMSD (radial axis centered on nstd = 1, corr = 1), and the latter the relationship between centered RMSD (*x*-axis), bias (*y*-axis), and total RMSD (distance from origin). Skill was assessed across all points in each region (open markers) and for regional averages over time (filled markers). Colors correspond to the regions in (a); black circles indicate statistics across the entire southern portion of the shelf (light yellow region).

depth of the temperature sensor during the bottom trawls (net heights average between 2.5 and 7 m above the seafloor, depending on equipment used [Nichol et al., 2007]).

Sampling locations were separated into five biophysical domains (Figure 2), based on common biophysical properties of the survey strata polygons used in the survey design. The southern portion of the shelf was divided into three regions following the three mixing regimes: inner shelf (strata 10–20, 70), middle shelf (strata 31–43, 81–82), and outer shelf (strata 50–62, 90). The outer shelf was further subdivided based on the bathymetry of the Bering10K model. The Bering10K model uses bathymetric smoothing to avoid errors in the horizontal pressure gradient that are characteristic of sigma-coordinate models like ROMS in areas of steep topography (Sikirić et al., 2009). Because of this, the modeled shelf is slightly narrower than the real-world one. We reflect this difference in shelf width by subdividing the outer shelf region (defined by sampling strata polygons bounded by the real-world 100-m and 200-m isobaths) into two using the model's 200-m contour as a dividing line, and we refer to the outer set of points as the shelf break mismatch region. A fifth region encompasses data from Norton Sound and north of St. Lawrence (strata 71). Finally, we define an overlapping region, referred to herein as the southeastern Bering Sea (SEBS), that encompasses strata 10–62 minus the shelf break mismatch portion of the outer shelf. This is the region typically used for management-oriented, survey-derived calculations of the cold pool index. In many management contexts, strata 82 and 90 are also considered part of the SEBS region (with strata 70, 71, and 81 designated as the northern Bering Sea shelf, or NBS); however, we opted not to include strata 82 or 90 in our SEBS polygon since sampling in these strata did not begin until 1987.

Validation statistics, including correlation, root mean squared difference, relative standard deviation, and bias were calculated between the observed bottom temperature data and the hindcast survey-replicated bottom temperature values on both a point by point basis and as a regionally averaged time series.

2.3.2. Forecast Skill Assessment Versus Hindcast

Forecast skill was calculated for bottom temperature and sea ice coverage following Stock et al. (2015) and Jacox et al. (2017), using anomaly correlation coefficients (ACCs) between the forecast and hindcast anomalies:

$$ACC(m,t) = \frac{\sum_{\alpha=1}^{N}(F'_{\alpha}(t,m) \times H'_{\alpha}(t,m))}{\sqrt{\sum_{\alpha=1}^{N} F'_{\alpha}(t,m)^2 \sum_{\alpha=1}^{N} H'_{\alpha}(t,m)^2}}, \quad (1)$$

where $F'_{\alpha}(t,m)$ and $H'_{\alpha}(t,m)$ are forecast and hindcast anomalies, respectively, for initialization month m, lead time t, in sample year α. Both hindcast and forecast anomalies were calculated relative to the lead-dependent climatologies across the 29 sample years, correcting for any mean error between the simulations. Dynamical forecast ACC values were calculated for each individual ensemble member from the two parent models, for the ensemble-mean forecast from each parent model, and for the multimodel mean across ensemble members from both parent models. All ACC values were calculated for individual grid points and for the SEBS regional average (Figure 2a, yellow shaded region). ACC values were also calculated for a persistence forecast, which assumed that the hindcast anomaly from the month before the initialization month would persist across all lead times.

Significance of ACC values was assessed as in Stock et al. (2015) following Bretherton et al. (1999), and tested whether the dynamic and persistence forecast ACC values were significantly greater than 0 and whether the dynamic forecast ACC values were significantly greater than the persistence forecast ACC values. Positive values of ACC indicate some forecast success relative to mean climate. In addition, ACC values of 0.5 (Roads, 1986; Stock et al., 2015) to 0.6 (Hollingsworth et al., 1980) have been determined empirically to represent the level at which a forecast achieves some synoptic skill. We use 0.5 as an approximate threshold for skill within this study when interpreting ACC values.

2.3.3. Forecast Skill Assessment Versus Groundfish Trawl Observations

Forecast skill was also assessed versus the summer values collected during the groundfish survey. ACC values were calculated for each survey sample location within the southeastern Bering Sea shelf region (Figure 2, strata 10–62) following Equation 1, with anomalies of each year defined as the observed value at

a station minus the long-term mean of all samples collected at that station. For each forecast simulation, the model simulation was subsampled from the closest grid point (10-km resolution) and model output time (daily resolution) to the survey sample points collected in the survey the following summer (using survey trawl midpoint location and trawl start time to match location and time, respectively). For the purposes of "following summer," we defined the cutoff of each survey to be the end of August, so forecasts initialized from September of year $x-1$ through August of year x are evaluated against the year x summer survey; for the May–August initialization dates, stations were eliminated from the calculations if their survey sample date preceded the dynamic forecast initialization date.

3. Results

3.1. Hindcast Validation

Performance statistics for the hindcast simulation show high agreement between the observed and simulated bottom temperatures (Figure 2). On the southeastern shelf, regionally averaged annual bottom temperature tracks the observed temperatures with a correlation of 0.96, a bias of −0.17°C, a centered root mean square difference (cRMSD) of 0.24°C, and normalized standard deviation of 0.923. Bias in the inner domain is slightly higher than in the middle domain (−0.26°C and −0.14°C, respectively) and correlation is slightly lower (0.92 and 0.97, respectively). The outer shelf, which experiences relatively low variability over time, has the lowest bias of the regions at −0.07°C, but also the lowest correlation of 0.67. While point-by-point statistics show lower skill than the regionally averaged time series, skill values remain high across most of the southeastern shelf.

Bias in the north is higher than in the south, at 0.43°C, though correlation, at 0.95, is comparable to the inner and middle shelf regions. The high correlation in this region, however, is likely in part due to the low number of samples and wide spread in sampling dates in this region. Even the most well-sampled northern stations have only been measured in a maximum of five different years, with a range in sampling dates up to a month apart from one year to the next. The correlation values therefore reflect a combination of skill in capturing the probable seasonal change over the range of sampling dates as well as skill in simulating interannual variability between years. The scarcity of data and lack of any day-of-year overlap between years limits us from identifying and removing the portion of the observed variability that may come from seasonality as opposed to interannual variability. Because our confidence in these skill metrics is low in the lightly sampled northern regions, we limit the majority of our forecast skill assessment to the more densely sampled southeastern portion of the shelf.

The hindcast validation metrics also indicate low correlation between the model and observations in the shelf break mismatch region. This is as expected, given that these points compare real-world on-shelf samples (influenced by interannual variations in the cold pool) to simulated slope and basin samples (which experience very little interannual variability in bottom temperature). For this reason, the shelf break mismatch region is also removed from the SEBS polygon for all forecast skill assessments.

Seasonal artifacts, such as those highlighted in the sparsely sampled northern region, may be present throughout the groundfish survey data, and may influence both the hindcast skill metrics and the survey-based forecast metrics. Due to the stratification that isolates the cold pool from surface warming across the middle and outer shelf regions, seasonal influence is likely minimal for the majority of stations there. However, the stations in the inner region, which is well-mixed in summer, are more likely to be influenced by such seasonal warming or cooling.

3.2. Forecast Skill Assessment

Before assessing the forecast model skill via the ACC metric, we verified that the mean monthly state of the forecast models did not show any major deviations from the hindcast simulation (Figure S1 in Supporting Information S1). Both forecast models showed only small monthly mean biases in bottom temperature across the SEBS region, with a mean absolute bias of 0.35°C and 0.2°C for CanCM4 and CFS, respectively. Interannual variability within the forecast models also matched closely to that seen in the hindcast simulation, with monthly standard deviations across the 29-year period ranging from 0.5 to 0.8°C. Both forecast

models tended to slightly underestimate interannual variability during the winter months (December–February), which are the months with the lowest year-to-year variability, but showed normalized standard deviations near 1.0 during the summer months when the interannual variability is highest.

Forecast skill ACC values for regionally averaged SEBS bottom temperature for each initialization date and lead time combination are presented in Figures 3 and 4. All models, including both dynamic models and the persistence forecast, show a similar pattern in skill. When initialized during typically ice-free months (April 1 to October 1), skill values are close to 1.0 at a lead time of 1 month, and remain at these high levels until the next ice season begins in mid-October. At the onset of the ice season, ACC values drop sharply. When initialized during the ice season, the skill between models is more varied. During the ice season, the persistence forecast sees ACC values over a 0.5 skill threshold for up to 3 months lead time if initialized in December and January, but only for a 1–2-month lead in the earlier or later winter month initializations. The CFS winter-initialized forecasts perform poorly with the exception of the January-initialized simulation; the ACC values here drop more slowly over the first 3 lead months than either December-initialized or February-initialized forecasts, remain low for another 3 months, and then rebound above 0.5 in lead months 7–9. The CanCM4 model, on the other hand, performs poorly for November-initialized and January-initialized forecasts, but shows consistently near 0.5 ACC values for the full 12-month simulation when initialized in December. October-initialized CFS simulation ACC values fall below the 0.5 threshold, but like December, remain consistently higher than November-initialized or January-initialized forecasts. Overall, the ensemble-mean forecast for each model tends to perform better than most, but not all, individual ensemble members. On average, the CanCM4 model tends to outperform the CFSv2 model, with the exception of January initializations. The multimodel mean ACC tracks the more skillful of the two parent model ensemble average values, with slightly lower ACC values than whichever contributing model performed better at a given time (Figure 4).

For spring and summer initializations (April through August), despite the high ACC values, the dynamic forecasts rarely perform better than the simple persistence forecast (Figure 5). In this case, we're defining "better" as either having an ACC value significantly greater than that of the persistence forecast, or achieving an ACC significantly greater than 0 when the persistence forecast ACC is not significantly greater than 0. The exception to this summer-initialization pattern is for May-initialized forecasts, where the dynamic forecasts are better for 1–3-month lead times, with significantly higher ACC values at 1–2-month lead times. For forecasts initialized in late winter (Jan through March), the dynamic forecast ACC values are often better than persistence by the above definition, particularly in the multimodel mean, and including during the target forecast months of June through August. However, the ACC values during that target period are low, below the 0.5 threshold for forecast skill, and are not significantly higher than those of the persistence forecast. At longer lead times of 8–12 months, the dynamic forecasts also typically outperform the persistence forecast, but again the ACC values are low, crossing the 0.5 threshold in only one instance (the CanCM4 December initialization with a lead of 5–11 months).

Forecast skill for metrics related to ice coverage is very low for most initialization and lead month pairings (Figure 6). Both models show ACC values significantly greater than 0 for short lead times from January to June. CFS significantly outperforms a simple persistence forecast for up to 3-month lead times with January initialization, 2-month lead times with May initialization, and 1-month lead times for February and April initialization; CanCM4 ACC values are only significantly higher than persistence for 1-month lead times in February, May, and June. For most of the initialization/lead pairings, neither the dynamic nor the persistence forecast is a better predictor than a climatological estimate, using the same definition for "better" as for bottom temperature, though the multimodel ensemble performs best for short lead times when initialized in late winter to spring (February–June).

Spatial variations in summer bottom temperature forecast skill relative to the hindcast simulation can be seen across the southeastern shelf region (Figure 7). ACC values along the outermost shelf region rise to a skillful level with the longest lead times, followed by the southern portion of the shelf and then the middle shelf. By April, the dynamic forecast ACC values are high across most of the shelf. Skill near the inner front (i.e., the tidal front separating the well-mixed inner domain from the thermally stratified middle domain) lags the rest of the shelf in reaching high ACC values. Spatial patterns for all lead/lag pairings can be seen in Figures S1 and S2 in Supporting Information S1.

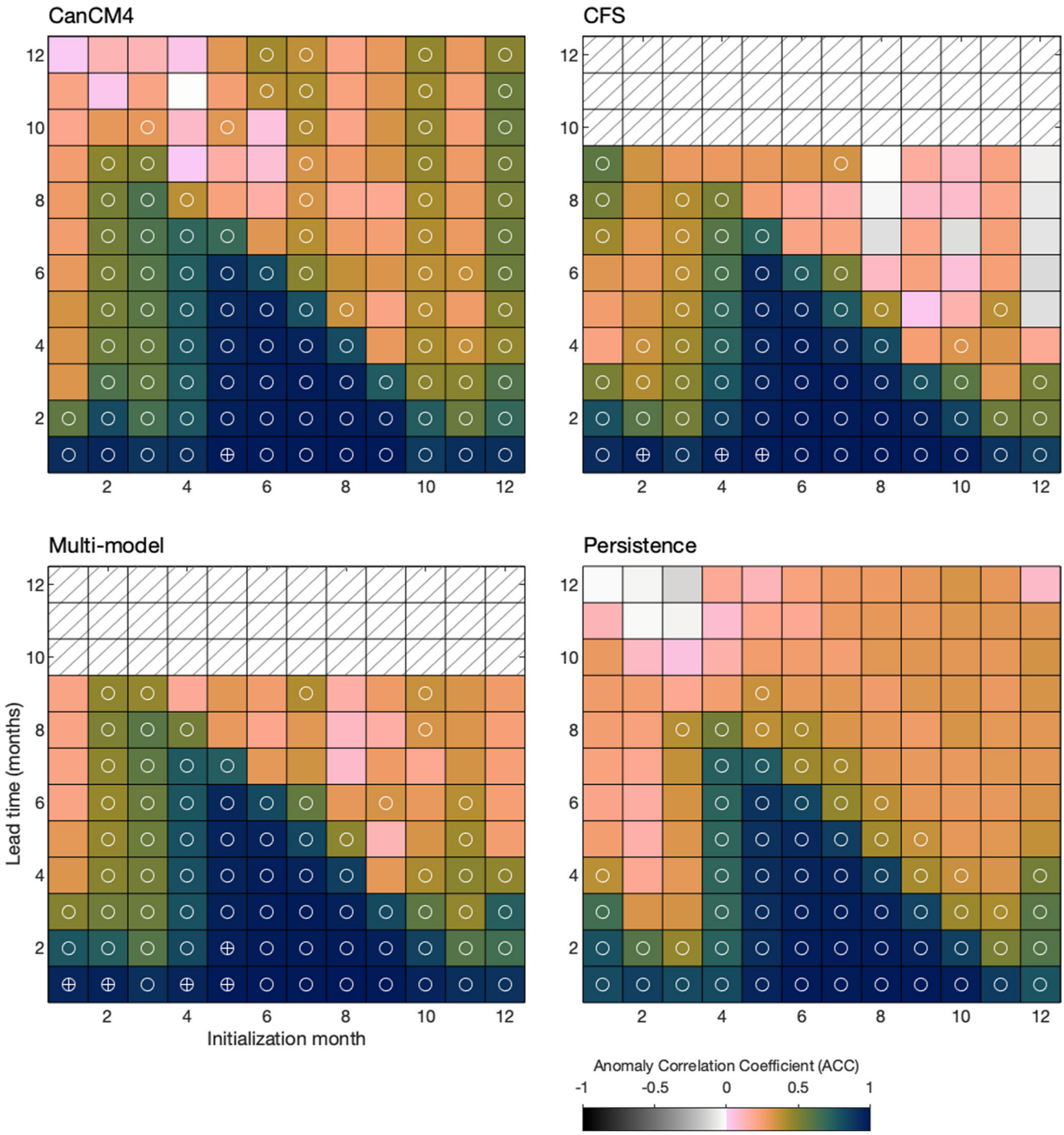

Figure 3. Anomaly correlation coefficient (ACC) of forecasted bottom temperature compared to hindcasted bottom temperature, averaged over the southeastern Bering shelf (SEBS) region. Colors indicate ACC versus initialization month (*x*-axis) and lead time (*y*-axis). White circles indicate points where the forecast ACC is significantly greater than 0 ($p = 0.05$), and white plus symbols indicate points where the dynamic forecast ACC is significantly greater than the persistence forecast ACC ($p = 0.05$).

ACC metrics calculated using the in-situ survey observations themselves as opposed to the hindcast simulation values are a bit more complicated to interpret (Figure 8), due to the non-synoptic nature of the observations and the year-to-year variations in sampling dates. The spatial variations in skill at any given initialization date can originate due to forecast skill differences in different locations, at different times of the year,

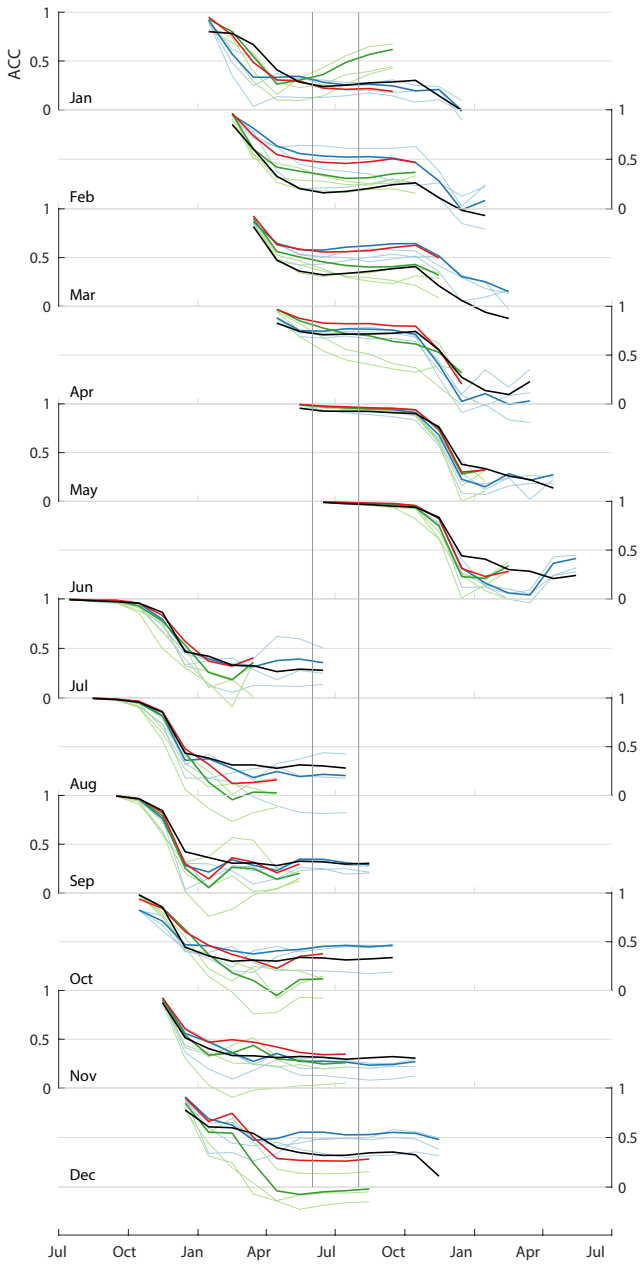

Figure 4. Anomaly correlation coefficient versus forecast date for each initialization date ranging from January (top) to December (bottom). Light-colored lines indicate ACC from each individual ensemble member from the CanCM4 (blue) and CFS (green) simulations, while dark lines indicate ACC for the ensemble-averaged forecast from each parent model. Also shown are the ACC of the multi-model-averaged forecast (red) and the persistence forecast (black). Vertical gray lines indicate the target forecast period of June 1 to August 1.

and at different lead times. Given the limited number of observations at any given point, we are unable to fully disentangle these potential sources for mismatch between the forecast and observations. Within this analysis, skill is highest in the inner and middle shelf regions. This pattern is similar to that seen in the hindcast-based metrics; higher ACC values at earlier initialization dates in the Bristol Bay region may reflect the early sampling dates, and hence shorter forecast lead times, of these stations. Stations in the shallowest parts of the inner domain also have higher ACC values, especially at longer lead times. Because this inner domain region is well-mixed year-round, skill in this region may derive in part from the forecast model's skill in capturing climatological patterns rather than interannual variability, since a station's anomaly from the mean combines anomaly from climatology (i.e., whether the year is a cooler or warmer one) as well as anomaly due to sampling time (i.e., samples collected earlier in the year are typically cooler than those collected later in summer). Stations falling in the shelf break mismatch region show very low skill, even with short lead times. This is to be expected, given that in the Bering10K model, these stations are located beyond the shelf break, and therefore do not experience the ice-influenced variations in bottom temperature seen on the real shelf.

Within a fisheries management context, bottom temperature forecasts would be most useful if they could successfully predict whether conditions in the following year might change relative to the current year. Therefore, we also evaluated whether each dynamic forecast model performed well as a binary classifier for changing summer conditions. We used a cutoff value of 3°C mean July bottom temperature in the SEBS region to separate "warm" from "cold" years, and then classified each year between 1982 and 2010 as positive if conditions changed from warm to cold or vice versa from July 1 to the next and negative if conditions remained stable. We then checked whether each of the six dynamic forecast ensemble members correctly classified these conditions in each of the years. We repeated the calculations separately across both parent models and for each parent model independently, as well as across all years and across only cold and only warm years. Results of this binary classification are summarized in Figure 9 and Tables S1–S3 in Supporting Information S1. See Figure S4 in Supporting Information S1 for a more detailed breakdown of hindcast and forecast classification by year. For lead times of 12 to 5 months, the accuracy of the classifier remains near 0.5, indicating an equal chance of a correct or incorrect prediction. Accuracy then increases sharply at the 4-month lead time, plateauing just below 1.0 with a 1–3-month lead. Precision for true negatives (no change) is higher than for true positives (change) for all lead times. There were no clear differences in the accuracy of the prediction between cold years and warm years.

4. Discussion

Bottom temperature across the eastern Bering Sea shelf is largely influenced by the formation of sea ice. During the summer months, mean bottom temperatures range from approximately 2°C to 6°C, peaking in late September. In the winter, with the contribution of near-freezing seawater associated with ice formation, the mean drops to between −1°C and 2°C (Figure S5 in Supporting Information S1). The coldest temperatures are typically seen in late March to early April, just before sea ice begins to retreat from the shelf.

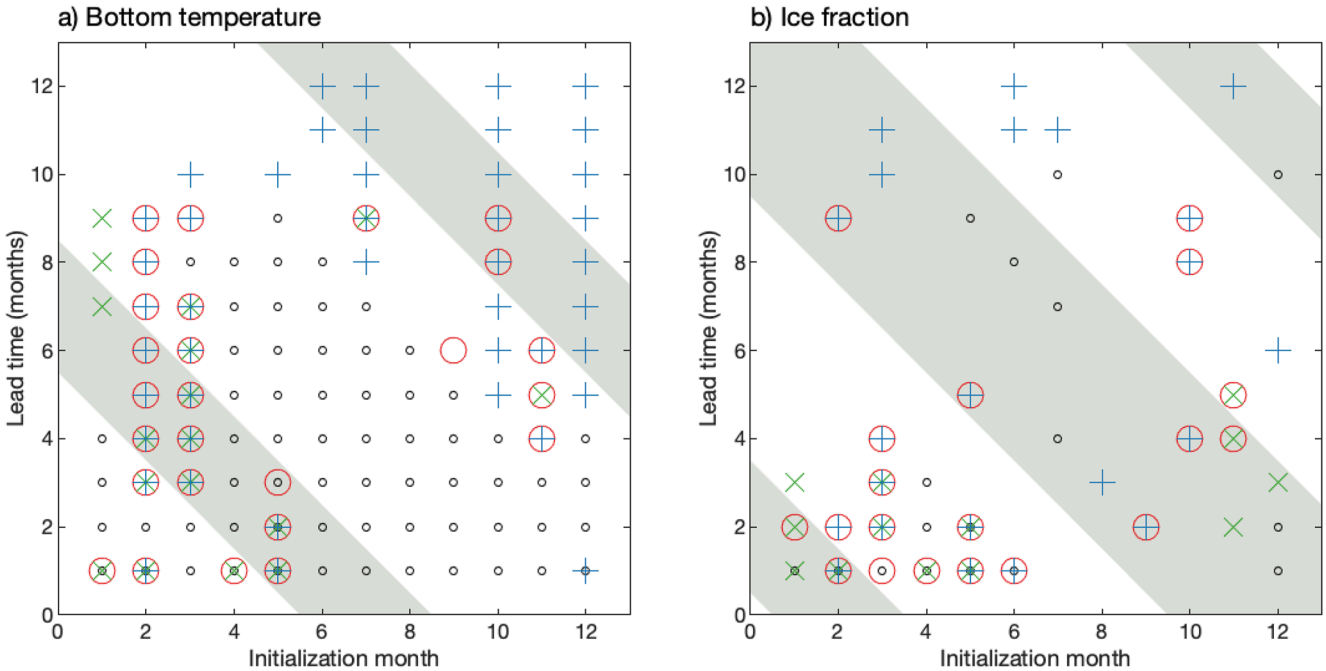

Figure 5. Predictor skill for SEBS (a) bottom temperature and (b) ice coverage. Large symbols indicate where the multimodel ensemble mean (red circles), CanCM4 ensemble mean (blue plus signs), and CFS ensemble mean (green x's) are better than the persistence forecast, with better being defined as ACC > 0 ($p = 0.05$) and either persistence ACC is not greater than 0 ($p = 0.05$) or the dynamic ACC > persistence ACC ($p = 0.05$). Small black circles indicate where the persistence forecast ACC > 0 (p = 0.05). An unmarked location implies that none of these forecast models significantly outperform a climatological estimate. Gray-shaded regions indicate the target forecast periods for each: June–August for bottom temperature, and October–March for ice coverage.

The reforecast experiments revealed that the majority of the predictability in the bottom temperature signature on the eastern Bering Sea shelf comes from the persistence signal. While bottom temperature anomalies fluctuate widely during the winter months, the anomaly at the time of the final ice retreat is generally maintained from early spring until the onset of ice the following winter (Figure S5 in Supporting Information S1). As a result, a seasonal persistence forecast performs very well when initialized after the ice season and maintains skill through the onset of the following ice season.

Both dynamic forecast models we tested were also able to skillfully predict bottom temperatures when initialized after the ice season. Using our most generous definition, the two-model ensemble average forecast performed better than a simple persistence forecast model, particularly in the late winter and early spring (April–May initializations). However, the increased benefit of this multimodel dynamic forecast over a persistence model may be marginal; anomaly correlation coefficient values were often low and not significantly higher than those of the persistence forecast. Considered independently, neither the CFS nor CanCM4 simulations alone outperformed the persistence forecast.

The persistence and dynamic forecasts both demonstrated very low skill when attempting to predict summer bottom temperature anomalies ahead of the ice season (i.e., initialized in the fall of the previous year). The persistence signal that is so dominant during the spring-to-summer restratification months is erased with the onset of winter and advance of new sea ice. The dynamic seasonal forecast models do not appear to capture the short time scale events, such as specific storm events that may impact ice formation or retreat, that lead to the high variation of bottom temperature anomalies during the ice season. This is apparent in the low skill for forecasting the advance and retreat of ice itself. For most of the ice season, the forecast models show skill in predicting fractional ice coverage only at one-month lead times. Only after the ice has begun its final retreat in early spring, that is, in April-initiated forecasts or beyond, do the forecast models show skill beyond one month, with skill extending to up to 3 months (until ice disappears for the year). Neither the direction nor the magnitude of bottom temperature anomalies during the ice season appears to correlate with the eventual spring and summer anomalies (Figure S5 in Supporting Information S1). A damped persistence forecast, which assumes the persistence anomaly decays to zero over time using a

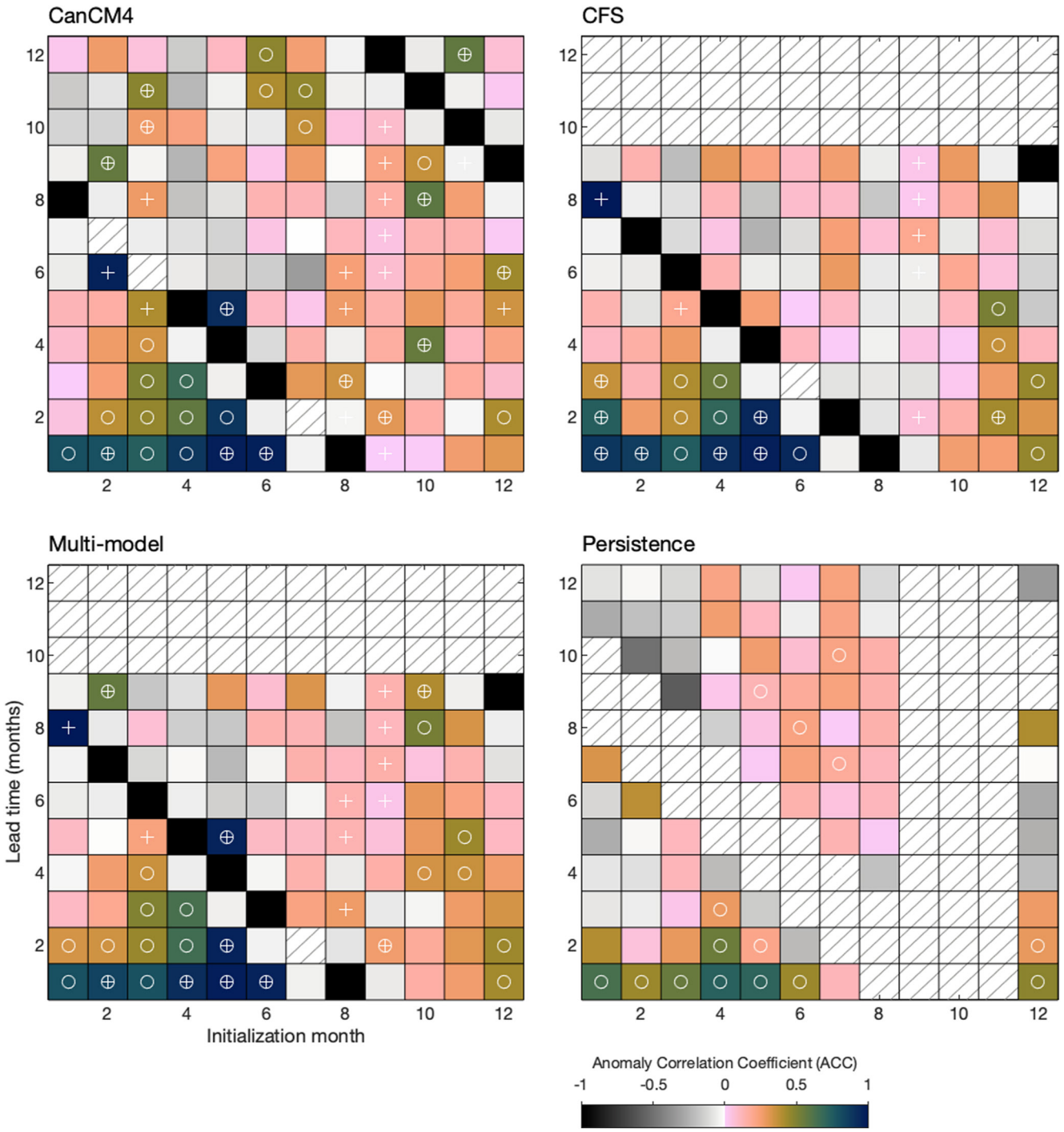

Figure 6. Anomaly correlation coefficient (ACC) of forecasted fractional ice coverage over the southeastern Bering shelf (SEBS) region, compared to hindcasted ice coverage. Colors indicate ACC versus initialization month (*x*-axis) and lead time (*y*-axis). White circles indicate points where the forecast ACC is significantly greater than 0 ($p = 0.05$), and white plus symbols indicate points where the dynamic forecast ACC is significantly greater than the persistence forecast ACC ($p = 0.05$). Hatched regions indicate where ACC is undefined; this results when the denominator of Equation 1 is 0, for example, months where no ice has been present across either the forecast or hindcast record.

timescale defined by the autocorrelation within the hindcast simulation, was also evaluated. This damped persistence forecast showed nearly identical skill values as the persistence forecast, suggesting that a climatological forecast was equally useful (or rather, not) compared to the assumption of persistence when forecasting at lead times that crossed the ice season barrier.

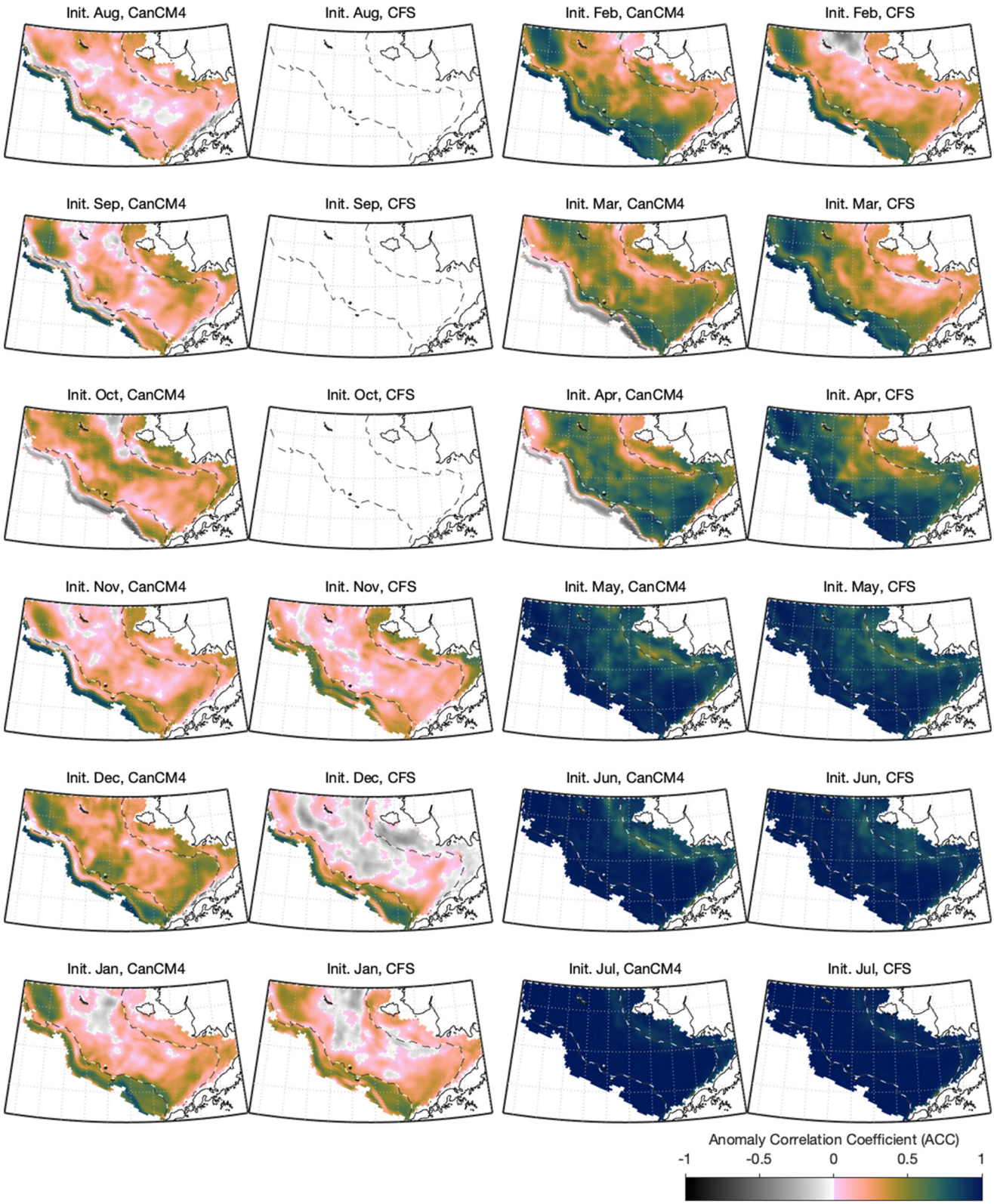

Figure 7. ACC of forecasted July bottom temperature across the southeastern Bering Sea shelf, with initialization dates ranging from August (12-month lead) to July (1-month lead). Dashed gray-and-white line indicates the maximum extent of the 2°C cold pool over the 29-year simulation period.

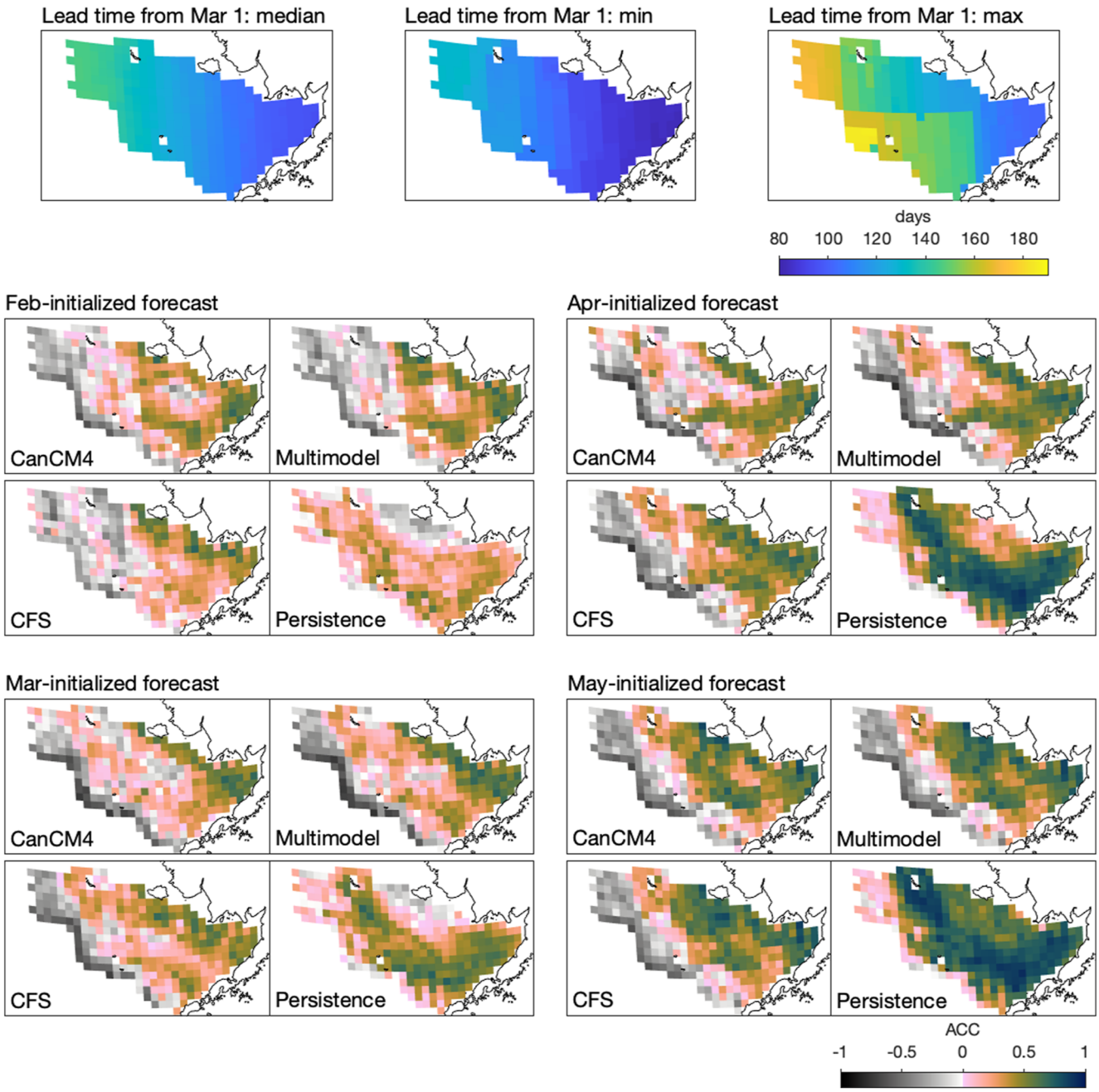

Figure 8. Top row: Median, minimum, and maximum lead times, in days, between a March 1-initialized forecast and the sampling dates at each station location over the 1982–2010 surveys. Bottom: ACC of forecasted survey-replicated bottom temperature across the southeastern Bering Sea shelf, with initialization dates ranging from February (approximately 3–5 months lead) to May (0–2 months lead).

The low skill we see with respect to sea ice extent predictability echos the patterns seen in more comprehensive analyses of pan-Arctic sea ice predictability (e.g., Blanchard-Wrigglesworth et al., 2011) and regional analyses (e.g., Cheng et al., 2016; Day et al., 2014). Seasonal ice in marginal regions like the Bering Sea tends to be thinner and more mobile than multi-year ice (Francis & Hunter, 2007); it is therefore more influenced by local ocean and atmospheric circulation patterns rather than the larger-scale advective processes that lend predictability beyond persistence to other Arctic regions (Guemas et al., 2016). Sea ice extent in the Bering Sea in particular appears to respond primarily to anomalies in easterly winds associated with the location of the Aleutian Low pressure system (Francis & Hunter, 2007); this is in contrast to other regions

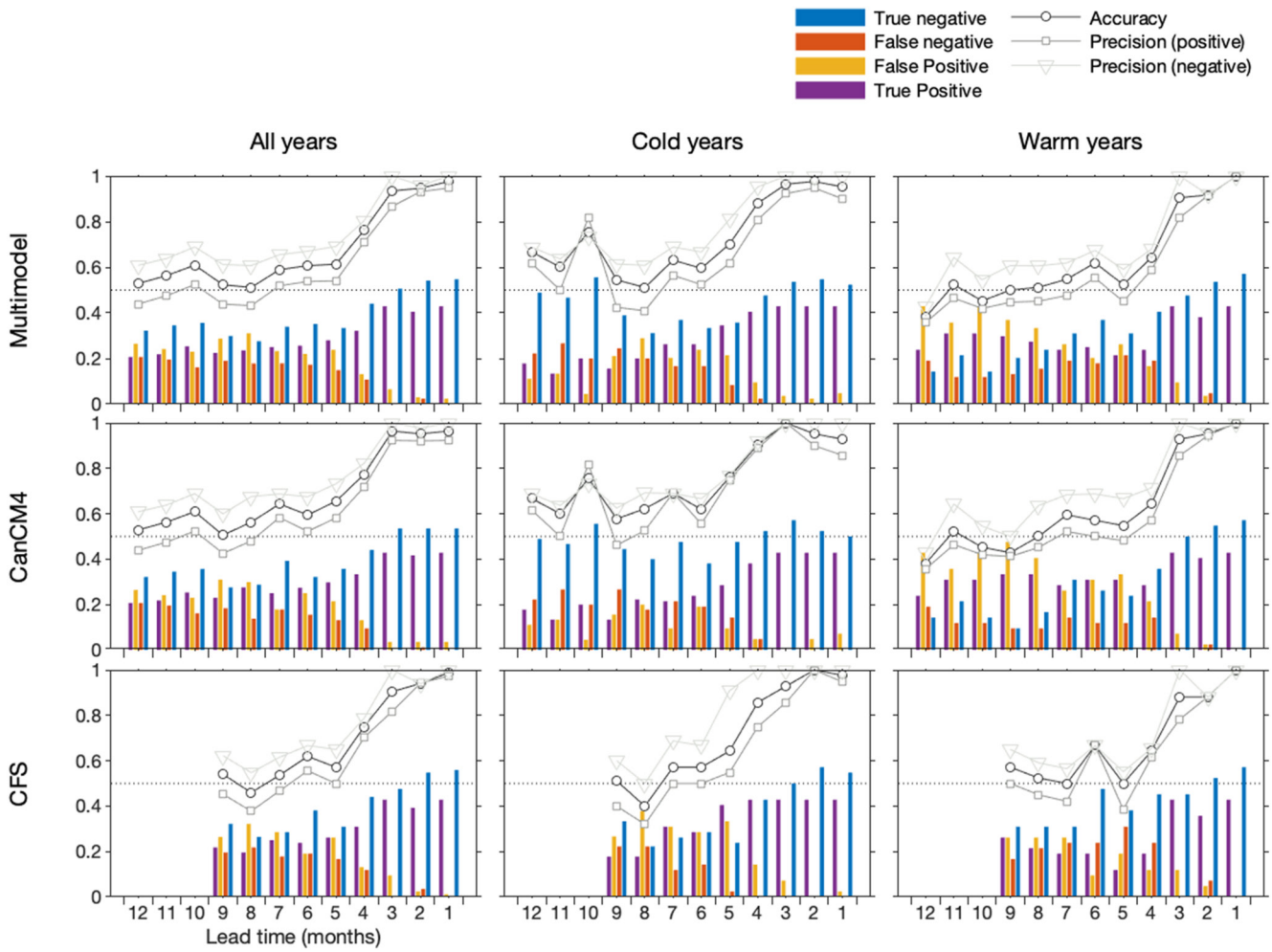

Figure 9. Dynamic forecast binary classification statistics, where a positive classification indicates changing temperature conditions. Accuracy is defined as the total rate of correct forecast classification (either positive or negative), precision (positive) indicates the fraction of correct positive classifications out of total positive classifications, and precision (negative) the fraction of correct negative classifications out of total negative classifications. Axis row indicates which ensemble members were included in each calculation, and axis column indicates whether the classifier was attempting to categorize conditions following any year, only cold years, or only warm years.

more influenced by ocean sea surface temperature. Therefore, low predictability of atmospheric circulation beyond 1–2 weeks is an inherent barrier to sea ice predictability of the region. A direct comparison of the CFS forecast winds used here with their hindcast (reanalysis) counterpart in fact demonstrates very low skill; for example, monthly average January winds predicted from the previous end of September/beginning of October have ACC less than 0.2 for both easterly and northerly components over the Bering Sea, with even lower values by February and March (Figures S6–S8 in Supporting Information S1). A number of studies have found that predictability in the Pacific sector of the Arctic, including the Bering Sea and Sea of Okhotsk, is lower than the Atlantic sector across a range of prediction methods, including both dynamic and statistical simulation (Bushuk et al., 2017, 2019; Yuan et al., 2016). By comparing perfect model simulations to operational model simulations, Bushuk et al. (2019) demonstrated that there remains a significant skill gap between the abilities of the current generation of seasonal to decadal forecast models and the theoretical upper limit of sea ice predictability. This suggests that seasonal forecasts of sea ice could improve as our understanding of and ability to simulate the underlying processes controlling ice extent improves. However, they also noted that the Bering Sea and Sea of Okhotsk demonstrated the lowest skill across both the operational model and perfect model simulations, suggesting that the Pacific sector regions are fundamentally less predictable than their Atlantic counterparts.

From the standpoint of eastern Bering Sea living marine resource management, being able to predict switches from good to poor conditions (for many managed fish species, cold to warm shifts leading to changes in the abundance and availability of prey) is more crucial than being able to predict any other situations (warm-to-warm, cold-to-cold, or warm-to-cold). Within our binary classification analysis, a true positive indicated a correct forecast of an upcoming change in conditions, while a false negative represented incorrectly predicting stability when a change was upcoming. In this context, the worst-case scenario would be a false negative during a cold year (center column, Figure 9). While our simulations showed a low prevalence of false negatives up to 5 months in advance, this came at the cost of frequent false positives (i.e., predictions of shifts when none occurred) at that 5-month lead. In general, for lead times beyond four months, a research and management framework that considers the large envelope of uncertainty in these predictions and is able to adapt to either changing or stable conditions will likely be more robust than attempting to rely on specific forecast values. In stock assessments, in particular, forecast information can be considered alongside other ecosystem information and sources of uncertainty using risk tables (Dorn & Zador, 2020).

The need for and benefits of climate-informed ecosystem approaches to fisheries management are increasingly being recognized (Heenan et al., 2015; Holsman et al., 2019; Ogier et al., 2016). However, the cost of running a dynamically downscaled multimodel forecast within an operational, management-oriented simulation framework is high, not only in terms of computing resources, but in the amount of time and effort required to acquire atmospheric and oceanic boundary condition data from multiple external modeling centers, bias correct and reformat the data for use in our regional model, and run the seasonal forecast simulations. On the other hand, adding a persistence forecast to our existing hindcast simulation framework, which is updated three times per year to within a few weeks of real time, would be relatively simple. Based on the marginal increased skill gained from a dynamic forecast relative to a persistence forecast, and the significant difference in effort to run it, the persistence forecast is the more efficient choice to be used within this framework at this time. Such a persistence forecast is capable of providing a skillful, spatially resolved prediction of the cold pool as early as April. Within the management cycle of the North Pacific Fishery Management Council, this is early enough to inform the management teams and the council of any unusual conditions that may affect the upcoming summer surveys or inform the quota-setting decisions at the October council meetings.

Many studies have also used statistical techniques, such as linear inverse modeling (LIM), to successfully forecast environmental anomalies at both global and regional scales. These techniques have been particularly useful in regions where anomalies are influenced by basin-scale variability such as the El Niño Southern Oscillation (ENSO) and Pacific Decadal Oscillation (PDO) (Jacox et al., 2020, and citations within). Similar techniques may be able to be used in the Bering Sea region to improve upon the existing persistence forecast.

The ensemble of simulations used in this study spanned only two parent models. While our original experiment design called for downscaling the full set of seven models that participated in the NMME exercise (Kirtman et al., 2014), our dynamical downscaling methods required far more atmospheric and oceanic input variables than many of the global-scale simulations had archived. This mismatch in data archiving needs between the original NMME simulations and regional applications has been acknowledged by many of the participating seasonal forecast modeling centers, and several centers plan to archive a greater number of variables in future simulations. As more seasonal forecast models become available, and as the skill at capturing processes relevant to sea ice dynamics within those models improve, the resulting downscaled multimodel ensemble may provide sea ice and cold pool forecasts across longer lead times.

Data Availability Statement

Data sets for this research, including the summarized simulation output underlying all figures in this study, can be accessed via Kearney et al. (2021) (https://doi.org/10.5281/zenodo.4735496).

Acknowledgments
This project was funded by the NOAA OAR Climate Program Office's Modeling, Analysis, Predictions, and Projections (MAPP) program under Award number NA17OAR4310104. This publication is partially funded by the Joint Institute for the Study of the Atmosphere and Ocean (JISAO) under NOAA Cooperative Agreement NA15OAR4320063, Contribution no. 2020-1142. This is PMEL contribution number 5257 and EcoFOCI-1009. Partial support for G. Hervieux and M. Alexander was provided by NOAA's Integrated Ecosystem Assessment (IEA) Program.

References

Berrisford, P., Dee, D. P., Poli, P., Brugge, R., Fielding, M., Fuentes, M., & Simmons, A. (2011). *The ERA-Interim archive Version 2.0 (*. 1) (p. 23). Retrieved from https://www.ecmwf.int/node/8174

Berrisford, P., Kållberg, P., Kobayashi, S., Dee, D., Uppala, S., Simmons, A. J., et al. (2011). Atmospheric conservation properties in ERA-Interim. *Quarterly Journal of the Royal Meteorological Society*, *137*(659), 1381–1399. https://doi.org/10.1002/qj.864

Blanchard-Wrigglesworth, E., Armour, K. C., Bitz, C. M., & Deweaver, E. (2011). Persistence and inherent predictability of arctic sea ice in a GCM ensemble and observations. *Journal of Climate*, *24*(1), 231–250. https://doi.org/10.1175/2010JCLI3775.1

Bretherton, C. S., Widmann, M., Dymnikov, V. P., Wallace, J. M., & Bladé, I. (1999). The effective number of spatial degrees of freedom of a time-varying field. *Journal of Climate*, *12*(7), 1990–2009. https://doi.org/10.1175/1520-0442(1999)012<1990:tenosd>2.0.co;2

Buckley, T. W., Greig, A., & Boldt, J. L. (2009). *Describing summer pelagic habitat over the continental shelf in the eastern Bering Sea, 1982-2006* (U.S. Dep. Commer., NOAA Tech. Memo. Nos. NMFS-AFSC-196) (p. 49).

Bushuk, M., Msadek, R., Winton, M., Vecchi, G., Yang, X., Rosati, A., & Gudgel, R. (2019). Regional Arctic sea–ice prediction: Potential versus operational seasonal forecast skill. *Climate Dynamics*, *52*(5–6), 2721–2743. https://doi.org/10.1007/s00382-018-4288-y

Bushuk, M., Msadek, R., Winton, M., Vecchi, G. A., Gudgel, R., Rosati, A., & Yang, X. (2017). Skillful regional prediction of Arctic Sea ice on seasonal timescales. *Geophysical Research Letters*, *44*(10), 4953–4964. https://doi.org/10.1002/2017GL073155

Carton, J. A., Chepurin, G. A., & Chen, L. (2018). SODA3: A new ocean climate reanalysis. *Journal of Climate*, *31*(17), 6967–6983. https://doi.org/10.1175/jcli-d-18-0149.1

Carton, J. A., Chepurin, G. A., Chen, L., & Grodsky, S. A. (2018). Improved global net surface heat flux. *Journal of Geophysical Research: Oceans*, *123*(5), 3144–3163. https://doi.org/10.1002/2017JC013137

Carton, J. A., Penny, S. G., & Kalnay, E. (2019). Temperature and salinity variability in the SODA3, ECCO4r3, and ORAS5 ocean reanalyses, 1993-2015. *Journal of Climate*, *32*(8), 2277–2293. https://doi.org/10.1175/JCLI-D-18-0605.1

Cheng, W., Blanchard-Wrigglesworth, E., Bitz, C. M., Ladd, C., & Stabeno, P. J. (2016). Diagnostic sea ice predictability in the pan-Arctic and U.S. Arctic regional seas. *Geophysical Research Letters*, *43*(22), 11688–11696. https://doi.org/10.1002/2016GL070735

Coachman, L. (1986). Circulation, water masses, and fluxes on the southeastern Bering Sea shelf. *Continental Shelf Research*, *5*(1–2), 23–108. https://doi.org/10.1016/0278-4343(86)90011-7

Day, J. J., Tietsche, S., & Hawkins, E. (2014). Pan-arctic and regional sea ice predictability: Initialization month dependence. *Journal of Climate*, *27*(12), 4371–4390. https://doi.org/10.1175/JCLI-D-13-00614.1

Dorn, M. W., & Zador, S. G. (2020). A risk table to address concerns external to stock assessments when developing fisheries harvest recommendations. *Ecosystem Health and Sustainability*, *6*(1), 1813634. https://doi.org/10.1080/20964129.2020.1813634

Duffy-Anderson, J. T., Stabeno, P., Andrews, A. G., Cieciel, K., Deary, A., Farley, E., et al. (2019). Responses of the northern Bering Sea and southeastern Bering Sea pelagic ecosystems following record-breaking low winter sea ice. *Geophysical Research Letters*, *46*(16), 9833–9842. https://doi.org/10.1029/2019GL083396

Eisner, L. B., Yasumiishi, E. M., Andrews, A. G., & O'Leary, C. A. (2020). Large copepods as leading indicators of walleye pollock recruitment in the southeastern Bering Sea: Sample-Based and spatio-temporal model (VAST) results. *Fisheries Research*, *232*, 105720. https://doi.org/10.1016/j.fishres.2020.105720

Fairall, C. W., Bradley, E. F., Rogers, D. P., Edson, J. B., & Young, G. S. (1996). Bulk parameterization of air-sea fluxes for Tropical Ocean-Global Atmosphere Coupled-Ocean Atmosphere Response Experiment. *Journal of Geophysical Research*, *101*(C2), 3747–3764. https://doi.org/10.1029/95JC03205

Fissel, B., Dalton, M., Garber-Yonts, B., Haynie, A., Kasperski, S., Lee, J., & Wise, S. (2017). *Stock Assessment and Fishery Evaluation Report for the Groundfish Fisheries of the Gulf of Alaska and Bering Sea/Aleutian Islands Area: Economic status of the groundfish fisheries off Alaska, 2016* (No. NPFMC Economic SAFE). North Pacific Fishery Management Council.

Francis, J. A., & Hunter, E. (2007). Drivers of declining sea ice in the Arctic winter: A tale of two seas. *Geophysical Research Letters*, *34*(17), 1–5. https://doi.org/10.1029/2007GL030995

Guemas, V., Blanchard-Wrigglesworth, E., Chevallier, M., Day, J. J., Déqué, M., Doblas-Reyes, F. J., et al. (2016). A review on Arctic sea-ice predictability and prediction on seasonal to decadal time-scales. *Quarterly Journal of the Royal Meteorological Society*, *142*(695), 546–561. https://doi.org/10.1002/qj.2401

Haidvogel, D. B., Arango, H., Budgell, W. P., Cornuelle, B. D., Curchitser, E., Di Lorenzo, E., et al. (2008). Ocean forecasting in terrain-following coordinates: Formulation and skill assessment of the Regional Ocean Modeling System. *Journal of Computational Physics*, *227*(7), 3595–3624. https://doi.org/10.1016/j.jcp.2007.06.016

Haynie, A. C., & Pfeiffer, L. (2012). Why economics matters for understanding the effects of climate change on fisheries. *ICES Journal of Marine Science*, *69*(7), 1160–1167. https://doi.org/10.1093/icesjms/fss021

Heenan, A., Pomeroy, R., Bell, J., Munday, P. L., Cheung, W., Logan, C., et al. (2015). A climate-informed, ecosystem approach to fisheries management. *Marine Policy*, *57*, 182–192. https://doi.org/10.1016/j.marpol.2015.03.018

Hermann, A. J., Curchitser, E. N., Hedstrom, K., Cheng, W., Bond, N. A., Wang, M., et al. (2016). Projected future biophysical states of the Bering Sea. *Deep Sea Research Part II: Topical Studies in Oceanography*, *134*, 30–47. https://doi.org/10.1016/j.dsr2.2015.11.001

Hollingsworth, A., Arpe, K., Tiedtke, M., Capaldo, M., & Savijärvi, H. (1980). The Performance of a medium-range forecast model in winter-impact of physical parameterizations. *Monthly Weather Review*, *108*(11), 1736–1773. https://doi.org/10.1175/1520-0493(1980)108<1736:tpoamr>2.0.co;2

Holsman, K. K., Hazen, E. L., Haynie, A., Gourguet, S., Hollowed, A., Bograd, S. J., et al. (2019). Towards climate resiliency in fisheries management. *ICES Journal of Marine Science*, *76*(5), 1368–1378. https://doi.org/10.1093/icesjms/fsz031

Holsman, K. K., Ianelli, J., Aydin, K., Punt, A. E., & Moffitt, E. A. (2016). A comparison of fisheries biological reference points estimated from temperature-specific multi-species and single-species climate-enhanced stock assessment models. *Deep-Sea Research Part II: Topical Studies in Oceanography*, *134*, 360–378. https://doi.org/10.1016/j.dsr2.2015.08.001

Jacox, M. G., Alexander, M. A., Siedlecki, S., Chen, K., Kwon, Y. O., Brodie, S., et al. (2020). Seasonal-to-interannual prediction of North American coastal marine ecosystems: Forecast methods, mechanisms of predictability, and priority developments. *Progress in Oceanography*, *183*, 102307. https://doi.org/10.1016/j.pocean.2020.102307

Jacox, M. G., Alexander, M. A., Stock, C. A., & Hervieux, G. (2017). On the skill of seasonal sea surface temperature forecasts in the California Current System and its connection to ENSO variability. *Climate Dynamics*, *53*(12), 7519–7533. https://doi.org/10.1007/s00382-017-3608-y

Kachel, N. B., Hunt, G. L., Salo, S. A., Schumacher, J. D., Stabeno, P. J., & Whitledge, T. E. (2002). Characteristics and variability of the inner front of the southeastern Bering Sea. *Deep-Sea Research Part II: Topical Studies in Oceanography*, *49*(26), 5889–5909. https://doi.org/10.1016/S0967-0645(02)00324-7

Kearney, K. A. (2019). *Freshwater input to the Bering Sea, 1950–2017* (NMFS-AFSC-388). NOAA Tech. Memo.
Kearney, K. A. (2021). *Temperature data from the eastern Bering Sea continental shelf bottom trawl survey as used for hydrodynamic model validation and comparison* (U.S. Dep. Commer., NOAA Tech. Memo. No. NMFS-AFSC-415).
Kearney, K. A., Alexander, M., Aydin, K., Cheng, W., Hermann, A., Hervieux, G., & Ortiz, I. (2021). *Supporting data: Seasonal predictability of sea ice and bottom temperature across the eastern Bering Sea shelf*. Zenodo. https://doi.org/10.5281/zenodo.4735496
Kearney, K. A., Hermann, A., Cheng, W., Ortiz, I., & Aydin, K. (2020). A coupled pelagic-benthic-sympagic biogeochemical model for the Bering Sea: Documentation and validation of the BESTNPZ model (v2019.08.23) within a high-resolution regional ocean model. *Geoscientific Model Development*, *13*(2), 597–650. https://doi.org/10.5194/gmd-13-597-2020
Kirtman, B. P., Min, D., Infanti, J. M., Kinter, J. L., Paolino, D. A., Zhang, Q., et al. (2014). The North American multimodel ensemble: Phase-1 seasonal-to-interannual prediction; phase-2 toward developing intraseasonal prediction. *Bulletin of the American Meteorological Society*, *95*(4), 585–601. https://doi.org/10.1175/BAMS-D-12-00050.1
Kotwicki, S., Buckley, T. W., Honkalehto, T., & Walters, G. (2005). Variation in the distribution of walleye pollock (*Theragra chalcogramma*) with temperature and implications for seasonal migration. *Fishery Bulletin*, *103*(4), 574–587.
Laman, E. A., Rooper, C. N., Turner, K., Rooney, S., Cooper, D. W., & Zimmermann, M. (2018). Using species distribution models to describe essential fish habitat in alaska. *Canadian Journal of Fisheries and Aquatic Sciences*, *75*(8), 1230–1255. https://doi.org/10.1139/cjfas-2017-0181
Lauth, R. R., Dawson, E. J., & Conner, J. (2019). *Results of the 2017 eastern and northern Bering Sea continental shelf bottom trawl survey of groundfish and invertebrate fauna* (U.S. Dep. Commer., NOAA Tech. Memo. No. NMFS-AFSC-396).
Marchesiello, P., McWilliams, J. C., & Shchepetkin, A. (2001). Open boundary conditions for long-term integration of regional oceanic models. *Ocean Modelling*, *3*(1–2), 1–20. https://doi.org/10.1016/S1463-5003(00)00013-5
Merryfield, W. J., Lee, W. S., Boer, G. J., Kharin, V. V., Scinocca, J. F., Flato, G. M., et al. (2013). The Canadian seasonal to interannual prediction system. Part I: Models and initialization. *Monthly Weather Review*, *141*(8), 2910–2945. https://doi.org/10.1175/MWR-D-12-00216.1
Mordy, C. W., Cokelet, E. D., De Robertis, A., Jenkins, R., Kuhn, C. E., Lawrence-Slavas, N., & Wangen, I. (2017). Saildrone surveys of oceanography, fish, and marine mammals in the Bering Sea. *Oceanography*, *30*(2), 28–31. https://doi.org/10.5670/oceanog.2017.230
Mueter, F. J., & Litzow, M. A. (2008). Sea ice retreat alters the biogeography of the Bering Sea continental shelf. *Ecological Applications*, *18*(2), 309–320. https://doi.org/10.1890/07-0564.1
National Marine Fisheries Service. (2017). *Fisheries economics of the United States, 2015* (U.S. Dep. Commer., NOAA Tech. Memo. No. NMFS-F/SPO-170).
Nichol, D. G., Honkalehto, T., & Thompson, G. G. (2007). Proximity of Pacific cod to the sea floor: Using archival tags to estimate fish availability to research bottom trawls. *Fisheries Research*, *86*(2–3), 129–135. https://doi.org/10.1016/j.fishres.2007.05.009
Ogier, E. M., Davidson, J., Fidelman, P., Haward, M., Hobday, A. J., Holbrook, N. J., et al. (2016). Fisheries management approaches as platforms for climate change adaptation: Comparing theory and practice in Australian fisheries. *Marine Policy*, *71*, 82–93. https://doi.org/10.1016/j.marpol.2016.05.014
O'Leary, C. A., Thorson, J. T., Ianelli, J. N., & Kotwicki, S. (2020). Adapting to climate-driven distribution shifts using model-based indices and age composition from multiple surveys in the walleye pollock (*Gadus chalcogrammus*) stock assessment. *Fisheries Oceanography*, *29*(6), 541–557. https://doi.org/10.1111/fog.12494
Roads, J. O. (1986). Forecasts of time averages with a numerical weather prediction model. *Journal of the Atmospheric Sciences*, *43*(9), 871–893. https://doi.org/10.1175/1520-0469(1986)043<0871:fotawa>2.0.co;2
Saha, S., Moorthi, S., Pan, H. L., Wu, X., Wang, J., Nadiga, S., et al. (2010). The NCEP climate forecast system reanalysis. *Bulletin of the American Meteorological Society*, *91*(8), 1015–1058. https://doi.org/10.1175/2010BAMS3001.1
Saha, S., Moorthi, S., Wu, X., Wang, J., Nadiga, S., Tripp, P., et al. (2014). The NCEP climate forecast system version 2. *Journal of Climate*, *27*(6), 2185–2208. https://doi.org/10.1175/JCLI-D-12-00823.1
Shchepetkin, A. F., & McWilliams, J. C. (2005). The regional oceanic modeling system (ROMS): A split-explicit, free-surface, topography-following-coordinate oceanic model. *Ocean Modelling*, *9*(4), 347–404. https://doi.org/10.1016/j.ocemod.2004.08.002
Sikirić, M. D., Janeković, I., & Kuzmić, M. (2009). A new approach to bathymetry smoothing in sigma-coordinate ocean models. *Ocean Modelling*, *29*(2), 128–136. https://doi.org/10.1016/j.ocemod.2009.03.009
Spencer, P. D. (2008). Density-independent and density-dependent factors affecting temporal changes in spatial distributions of eastern Bering Sea flatfish. *Fisheries Oceanography*, *17*(5), 396–410. https://doi.org/10.1111/j.1365-2419.2008.00486.x
Stabeno, P. J., Bond, N. A., Kachel, N. B., Salo, S. A., & Schumacher, J. D. (2001). On the temporal variability of the physical environment over the south-eastern Bering Sea. *Fisheries Oceanography*, *10*(1), 81–98. https://doi.org/10.1046/j.1365-2419.2001.00157.x
Stabeno, P. J., Duffy-Anderson, J. T., Eisner, L. B., Farley, E. V., Heintz, R. A., & Mordy, C. W. (2017). Return of warm conditions in the southeastern Bering Sea: Physics to fluorescence. *PLOS One*, *12*(9), 1–16. https://doi.org/10.1371/journal.pone.0185464
Stabeno, P. J., Farley, E. V., Kachel, N. B., Moore, S., Mordy, C. W., Napp, J. M., et al. (2012). A comparison of the physics of the northern and southern shelves of the eastern Bering Sea and some implications for the ecosystem. *Deep-Sea Research Part II: Topical Studies in Oceanography*, *65–70*, 14–30. https://doi.org/10.1016/j.dsr2.2012.02.019
Stabeno, P. J., Mordy, C. W., & Sigler, M. F. (2020). Seasonal patterns of near-bottom chlorophyll fluorescence in the eastern Chukchi Sea: 2010–2019. *Deep-Sea Research Part II: Topical Studies in Oceanography*, *177*, 104842. https://doi.org/10.1016/j.dsr2.2020.104842
Stevenson, D. E., & Lauth, R. R. (2012). Latitudinal trends and temporal shifts in the catch composition of bottom trawls conducted on the eastern Bering Sea shelf. *Deep-Sea Research Part II: Topical Studies in Oceanography*, *65–70*, 251–259. https://doi.org/10.1016/j.dsr2.2012.02.021
Stevenson, D. E., & Lauth, R. R. (2019). Bottom trawl surveys in the northern Bering Sea indicate recent shifts in the distribution of marine species. *Polar Biology*, *42*(2), 407–421. https://doi.org/10.1007/s00300-018-2431-1
Stock, C. A., Pegion, K., Vecchi, G. A., Alexander, M. A., Tommasi, D., Bond, N. A., et al. (2015). Seasonal sea surface temperature anomaly prediction for coastal ecosystems. *Progress in Oceanography*, *137*, 219–236. https://doi.org/10.1016/j.pocean.2015.06.007
Thorson, J. T., Arimitsu, M. L., Barnett, L. A., Cheng, W., Eisner, L. B., Haynie, A. C., et al. (2021). Forecasting community reassembly using climate-linked spatio-temporal ecosystem models. *Ecography*, *44*, 1–625. https://doi.org/10.1111/ecog.05471
Thorson, J. T., Cheng, W., Hermann, A. J., Ianelli, J. N., Litzow, M. A., O'Leary, C. A., & Thompson, G. G. (2020). Empirical orthogonal function regression: Linking population biology to spatial varying environmental conditions using climate projections. *Global Change Biology*, *26*(8), 4638–4649. https://doi.org/10.1111/gcb.15149
Thorson, J. T., Ciannelli, L., & Litzow, M. A. (2020). Defining indices of ecosystem variability using biological samples of fish communities: A generalization of empirical orthogonal functions. *Progress in Oceanography*, *181*, 102244. https://doi.org/10.1016/j.pocean.2019.102244

Yuan, X., Chen, D., Li, C., Wang, L., & Wang, W. (2016). Arctic sea ice seasonal prediction by a linear Markov model. *Journal of Climate*, *29*(22), 8151–8173. https://doi.org/10.1175/JCLI-D-15-0858.1

Zador, S. G., Holsman, K. K., Aydin, K. Y., & Gaichas, S. K. (2017). Ecosystem considerations in Alaska: The value of qualitative assessments. *ICES Journal of Marine Science*, *74*(1), 421–430. https://doi.org/10.1093/icesjms/fsw144

Zhang, J., Woodgate, R., & Mangiameli, S. (2012). Towards seasonal prediction of the distribution and extent of cold bottom waters on the Bering Sea shelf. *Deep-Sea Research Part II: Topical Studies in Oceanography*, *65–70*, 58–71. https://doi.org/10.1016/j.dsr2.2012.02.023