Climate Dynamics manuscript No.
(will be inserted by the editor)

Regional Arctic sea-ice prediction: Potential versus operational seasonal forecast skill

Mitchell Bushuk · Rym Msadek · Michael Winton · Gabriel Vecchi · Xiaosong Yang · Anthony Rosati · Rich Gudgel

Received: date / Accepted: date

Abstract Seasonal predictions of Arctic sea ice on regional spatial scales are a pressing need for a broad group of stakeholders, however, most assessments of predictability and forecast skill to date have focused on pan-Arctic sea-ice extent (SIE). In this work, we present the first direct comparison of perfect model (PM) and operational (OP) seasonal prediction skill for regional Arctic SIE within a common dynamical prediction system. This assessment is based on two complementary suites of seasonal prediction ensemble experiments performed with a global coupled climate model. First, we present a suite of PM predictability experiments with start dates spanning the calendar year, which are used to quantify the potential regional SIE prediction skill of this system. Second, we assess the system's OP prediction skill for detrended regional SIE using a suite of retrospective initialized seasonal forecasts spanning 1981-2016. In nearly all Arctic regions and for all target months, we find a substantial skill gap between PM and OP predictions of regional SIE. The PM experiments reveal that regional winter SIE is potentially predictable at lead times beyond 12 months, substantially longer than the skill of their OP counterparts. Both the OP and PM predictions display a spring prediction skill barrier for regional summer SIE forecasts, indicating a fundamental predictability limit for summer regional predictions. We find that a similar barrier exists for pan-Arctic sea-ice volume predictions, but is not present for predictions of pan-Arctic SIE. The skill gap identified in this work indicates a promising potential for future improvements in regional SIE predictions.

Mitchell Bushuk
Geophysical Fluid Dynamics Laboratory, NOAA, Princeton, New Jersey, USA;
University Corporation for Atmospheric Research, Boulder, Colorado, USA
E-mail: mitchell.bushuk@noaa.gov

Rym Msadek
CNRS/CERFACS, CECI UMR 5318, Toulouse, France

Michael Winton, Xiaosong Yang, Anthony Rosati, Rich Gudgel
Geophysical Fluid Dynamics Laboratory, NOAA, Princeton, New Jersey, USA

Gabriel A. Vecchi
Department of Geosciences, Princeton University, Princeton, NJ, USA; Princeton Environmental Institute, Princeton University, Princeton, NJ, USA

Keywords Sea ice · Seasonal predictability · Arctic

1 Introduction

Rapid changes in Arctic sea-ice extent (SIE), thickness (SIT), and age over the satellite era, and their implications for a broad group of stakeholders, have led to a burgeoning research interest in seasonal-to-interannual predictability and prediction skill of Arctic sea ice. Over the past decade, substantial progress in sea-ice prediction science has been made, including the first seasonal predictions of sea ice made using coupled global climate models (GCMs) [75,18,65,52,54,59,7,20,33,27,12,5,15], the first prognostic estimates of potential sea-ice prediction skill using "perfect model" approaches [45,35,6,70,30,21,22], diagnostic studies quantifying timescales and identifying key sources of sea-ice predictability [44,4,36,17,21,11,9,16,13,10], the development of novel statistical techniques for sea-ice forecasting [29,28,50,71,68,63,43,74,81,77,60], and the creation of the sea-ice prediction network (SIPN, [68,7]), which collects and communicates predictions of September Arctic SIE (see http://www.arcus.org/sipn/sea-ice-outlook).

A crucial finding that has emerged from this body of work is that current seasonal forecasts of pan-Arctic SIE made with operational (OP) prediction systems could be substantially improved. State-of-the-art dynamical prediction systems, based on fully-coupled GCMs and initial conditions (ICs) constrained by observations, can skillfully predict detrended pan-Arctic summer SIE at 1-6 month lead times and winter SIE at 1-11 month lead times depending on the prediction system used [75,18,65,52,54,59,7,20,33,27]. These OP skill estimates are based on retrospective predictions (hindcasts), in which the fixed prediction system is run using only data available prior to the forecast initialization date. Perfect model (PM) studies, based on ensembles of model runs initialized from nearly identical ICs, complement these findings by providing estimates of the upper limits of prediction skill within a given GCM. These idealized experiments provide skill estimates in the case of perfectly known model physics and perfect ICs, and therefore are considered to be an upper bound to the prediction skill achievable in an OP system. PM studies show that pan-Arctic SIE and sea-ice volume (SIV) are predictable at 12-36 and 24-48 month lead times, respectively, highlighting a significant skill gap between PM and OP predictions [45,35,6,70,30,21].

The principal focus of Arctic sea-ice predictability research has been pan-Arctic SIE, a quantity of minimal utility at stakeholder-relevant spatial scales. As prospects for skillful seasonal sea-ice prediction systems become more realistic, it is paramount for sea-ice predictability science to address the regional scales required by future forecast users, which include northern communities, shipping industries, fisheries, wildlife management organizations, ecotourism, and natural resource industries [42]. Initial steps towards understanding Arctic regional predictability have been made, but many knowledge gaps remain. The PM study of [21] demonstrated a potential for skillful regional SIE predictions in the HadGEM1.2 GCM, finding greatest predictability for winter SIE in the Labrador, Greenland-Iceland-Norwegian (GIN), and Barents Seas (at lead times of 1.5-2.5 years) and lower predictability for summer SIE (skill at lead times of 2-4 months). [66] showed skillful OP predictions of detrended sea-ice retreat and advance dates, with notably high skill for ice-advance date predictions in the Labrador Sea/Baffin Bay, Beaufort Sea, Laptev/East Siberian Seas, Chukchi Sea, and Hudson Bay (3-5 month leads for detrended anomalies). The work of [46] reported skillful OP predictions of detrended sea-ice area up to 6 month lead times in the Barents/Kara Seas and the Northeast passage region. [12] provided the first comprehensive assessment of OP regional SIE predictions, reporting detrended SIE skill at lead times of 5-11 months in the Labrador, GIN, and Barents Seas, and 1-4 months in the Laptev, East Siberian, Chukchi, Beaufort, Okhotsk, and Bering Seas. This work attributed the high winter SIE skill of the North Atlantic to initialization of subsurface ocean temperature anomalies, and the summer SIE skill to initialization of SIT anomalies. Using two different OP seasonal prediction systems, [20] and [27] both found that improved SIT ICs led to improvements in regional predictions of summer sea ice. On longer timescales, [80] demonstrated that decadal sea-ice trends in the North Atlantic are predictable, due to dynamical predictability of thermohaline circulation variations.

While the gap between PM and OP prediction skill suggests a potential for improved OP predictions, it is important to note that the PM and OP studies cited above were performed with different GCMs. Since each GCM has unique model physics and a resulting unique set of model biases, this precludes a direct quantitative assessment of the PM/OP skill gap. In this study, we present the first formal comparison of PM and OP Arctic sea-ice prediction skill within the same GCM-based prediction system. In order to provide an "apples-to-apples" skill comparison, we first address the general problem of how to make a robust comparison between PM and OP skill. PM and OP studies often utilize different metrics to quantify prediction skill, or use different definitions for metrics with the same name [34]. In this study, we begin by introducing a consistent set of PM and OP skill metrics, which can be computed analogously for both PM and OP prediction applications. These metrics are specifically designed to allow for a robust comparison between PM and OP skill.

In this work, we perform a suite of PM experiments initialized from six start months spanning the calendar year and from six start years spanning different initial SIV states. This experimental design provides better seasonal coverage than earlier PM studies, allowing for an evaluation of PM skill for all target months and lead times of 0-35 months. We also consider a suite of retrospective OP predictions made with the same model, initialized on the first of each month from January 1981–December 2016. Using these complementary experiments, we directly compare PM and OP prediction skill for regional Arctic SIE, providing a quantitative assessment of the gap between current and potential Arctic seasonal-to-interannual prediction skill.

The plan of this paper is as follows. In section 2, we describe the experimental design and introduce prediction skill metrics that allow for a direct comparison between PM and OP skill. Section 3 presents predictability results for pan-Arctic SIV and SIE. In section 4, comparisons between PM and OP skill are made for fourteen Arctic regions. We conclude in section 5.

2 Experimental Design and Prediction Skill Metrics

2.1 The Dynamical Model

This study is based on experiments performed with the Geophysical Fluid Dynamics Laboratory Forecast-oriented Low Ocean Resolution (GFDL-FLOR) GCM. FLOR is a fully-coupled global atmosphere-ocean-sea ice-land model, which employs a relatively high resolution of 0.5° in the atmosphere and land components and a lower resolution of 1° in the ocean and sea-ice components [72]. The choice of a coarser resolution for the ocean and sea-ice components was made for computational efficiency, as this model was developed for seasonal prediction applications requiring ensemble integrations and many start dates, and for consistency with the ocean and sea ice components of GFDL-CM2.1 [23], which is the basis of the assimilation system with which the initial conditions for the OP predictions are generated. The sea-ice component of FLOR is the sea-ice simulator version 1 (SIS1, [23]), which utilizes an elastic-viscous-plastic rheology to compute the internal ice stresses [37], a modified Semtner 3-layer thermodynamic scheme with two ice layers and one snow layer [78], and a subgrid-scale ice-thickness distribution with 5 thickness categories [2]. FLOR's ocean component is the Modular Ocean Model version 5 (MOM5, [31]), which uses a rescaled geopotential height coordinate (z*, [32]) with 50 vertical levels. The atmospheric component of FLOR is Atmospheric Model version 2.5 (AM2.5, [24]), which uses a cubed-sphere finite-volume dynamical core [49, 62] with 32 vertical levels, and the land component of FLOR is Land Model, version 3 (LM3, [53]).

2.2 The Control Integration

The perfect model (PM) experiments described in the following subsection are branched from a 300-year control integration of FLOR, which uses radiative forcing and land use conditions that are representative of 1990. This 300-year control integration ("the new control run") was initialized from year 800 of another 1400-year 1990 control run (henceforth "the original control run"), which had been previously run on a now-decommissioned high-performance computing cluster. The new control run and PM experiments were run on a new computing cluster, which does not bitwise reproduce numerical solutions obtained on the previous cluster but does reproduce the climate mean state and variability. The original control run shows clear signs of model spin up, with a notable adjustment occurring in the first 500 years of the run (see the evolution of SIV anomalies in Fig. 1a). After roughly year 600, the model reaches a statistically steady equilibrium for the variables of interest in this study. The new control run was initialized from the well-equilibrated year 800 of the original control run, and does not show signs of model drift over the 300-year integration period (see Fig. 1a). Centennial-timescale drift of Arctic SIE and SIV associated with model spin up is a ubiquitous feature across GCMs (e.g., see Fig. 1 of [22]) and has the potential to significantly bias PM skill results. These potential skill biases are particularly relevant for regional sea ice, as a drifting climatology can cause a formerly high-variability region to shift to a low-variability region as it becomes ice covered or ice free, and vice versa. Therefore, the well-equilibrated control run shown in Fig. 1a is a crucial feature of this regional sea-ice study. Henceforth, we will refer to the new 300-year control run simply as "the control run."

We evaluate the FLOR sea-ice model biases using monthly-averaged passive microwave satellite SIC observations from the National Snow and Ice Data Center (NSIDC) processed using the NASA Team Algorithm (dataset ID: NSIDC-0051, [14]). We also consider SIT data from the Pan-Arctic Ice Ocean Modeling and Assimilation System (PIOMAS, [82]), an ice-ocean reanalysis that agrees quite well with available in situ and satellite thickness observations [64]. For comparison with FLOR, both the NSIDC and PIOMAS data were regridded onto the FLOR sea-ice grid. The pan-Arctic SIE climatology of FLOR has fairly good agreement with satellite observations, with a slight low bias in August–October and good agreement in other months (see Fig. S1a). The model biases are more pronounced when considering SIC spatial patterns. FLOR's winter SIC has negative biases (too little sea ice) in the Labrador, Okhotsk, and Bering Seas, and positive biases (too much sea ice) in the Greenland-Iceland-Norwegian (GIN) and Barents Seas (Fig. S2a-c). The summer SIC pattern is dominated by a negative bias wrapping the Alaskan and Eurasian coastlines, and a positive bias in the northern GIN and Barents Seas (Fig. S2d-f). Compared to PIOMAS, FLOR has a substantial thin bias of 0.5–1m at most central Arctic gridpoints (Fig. S3) and a lower pan-Arctic SIV in all months of the year (Fig. S1b). The spatial biases in SIC variability are largely dictated by biases in the mean ice-edge position, which result in dipole bias patterns in the SIC standard deviation fields (Fig. S4). One notable exception to this is the Labrador Sea during winter, in which FLOR has less SIC variability throughout the region.

2.3 Perfect Model Predictability Experiments

The 300-year control simulation serves as the baseline for our PM predictability experiments. Using this run, we choose a number of start dates, initialize a twelve-member initial condition ensemble for each start date, and run these ensembles forward in time for three years. A novel aspect of our experimental design is the choice of start dates with uniform seasonal coverage. Prior PM studies have focused primarily on January, May, and July start dates [22]. In this study, for each start year, we initialize ensembles on January 1, March 1, May 1, July 1, September 1, and November 1 (see Table 1 for a summary of the PM experiments). This uniform seasonal coverage allows us to investigate the lead-dependence of seasonal forecast skill and to make a clean quantitative comparison with the OP prediction skill reported in [12]. These start dates also allow us to identify optimal initialization months for given regions or target months of interest. In order to assess how predictability varies with the initial SIV state, we choose start years based on SIV anomalies, selecting two high volume years, two typical volume years, and two low volume years. The high/low volume years are years in which the SIV anomaly exceeds $\pm 1.2\sigma$ in all months of the year, and the typical volume years have SIV anomalies with absolute value less than 0.25σ in all months of the year (see Fig. 1b). The SIV standard deviation of the FLOR control run ($\sigma = 1.1\text{e}12\ \text{m}^3$) is comparable to the detrended SIV standard deviation of PIOMAS ($\sigma = 1.3\text{e}12$ m^3), indicating that the chosen high/low SIV anomalies have similar magnitude to those in the PIOMAS record. The start years are chosen at least 20 years apart,

Table 1 Summary of GFDL-FLOR PM experiments

Start year	Volume State	Start Months	Ensemble members	Integration time
839	High	Jan, Mar, May, Jul, Sep, Nov	12	3 years
874	Low	Jan, Mar, May, Jul, Sep, Nov	12	3 years
898	Typical	Jan, Mar, May, Jul, Sep, Nov	12	3 years
933	High	Jan, Mar, May, Jul, Sep, Nov	12	3 years
981	Low	Jan, Mar, May, Jul, Sep, Nov	12	3 years
1008	Typical	Jan, Mar, May, Jul, Sep, Nov	12	3 years

so that each start year of ensembles can be considered independent of other start years.

A key aspect of PM experiments is the availability of model restart files which can be used to construct an ensemble of initial conditions. In the control run, restart files were saved at monthly frequency, which allows us to initialize an ensemble from any month of the year. The ensembles were constructed by adding a random spatially uncorrelated Gaussian perturbation with standard deviation 10^{-4}K to the SST field at each ocean gridpoint. This ensemble generation technique mirrors the protocol used in the APPOSITE experiments [21, 70, 22]. Our PM experiments were run with 12 ensemble members, which is the ensemble size used for GFDL's initialized seasonal predictions (see following subsection). This suite of experiments, consisting of six start years, six start months per start year, 12 ensemble members per start month, and 3 years of integration time, totals 1296 years of model integration.

In each ensemble experiment, the ensemble members are initialized infinitesimally close to one another and diverge over time due to the chaotic dynamics of the system (see Fig. 1c). The rate at which this ensemble divergence occurs provides information on the inherent predictability of the system, quantifying the timescale at which a skillful prediction could be made in the case of perfect ICs and perfectly known model physics. In subsection 2.6, we present a set of metrics used to quantity the prediction skill of PM predictability experiments.

2.4 Retrospective Seasonal Prediction Experiments

As a complement to the PM experiments, we analyze the seasonal prediction skill of a suite of retrospective OP prediction experiments made using the FLOR model. These twelve-member ensemble predictions are initialized on the first of each month from January 1981–December 2016, and integrated for one year. The initial conditions come from GFDL's Ensemble Coupled Data Assimilation (ECDA; [83, 84]) System, which is based on the ensemble adjustment Kalman filter [1]. The ECDA system assimilates satellite sea-surface temperatures (SST), subsurface temperature and salinity data, and atmospheric reanalysis data from National Centers for Environmental Prediction [12]. Note that while this system does not explicitly assimilate sea-ice data, the sea-ice state in the coupled assimilation is constrained via surface heat fluxes associated with assimilation of SST and surface-air temperature data. This assimilation system captures the climatology, long-term trend, and interannual variability of pan-Arctic SIE with reasonable fidelity [54]. These FLOR retrospective seasonal predictions have been used to ex-

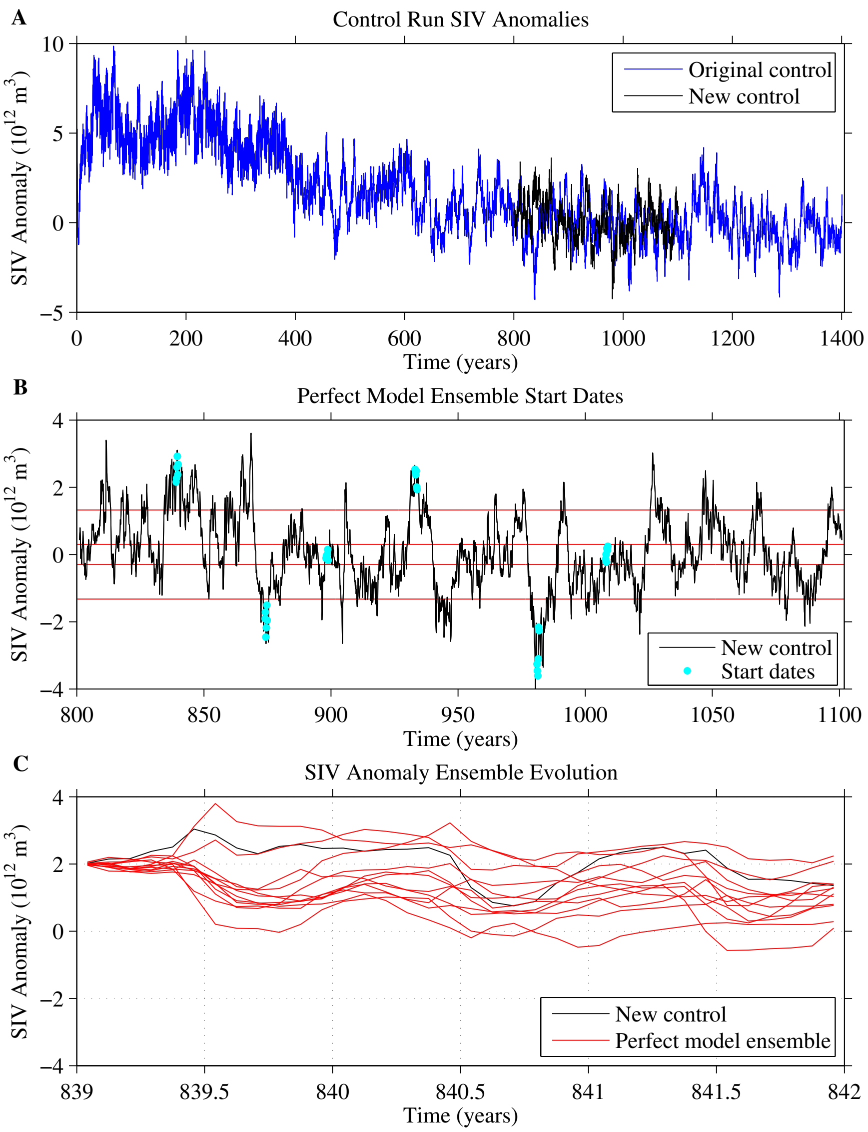

Fig. 1 Experimental setup for PM experiments. (A): Arctic SIV anomaly timeseries from the original and new 1990 control runs. The new control is initialized from year 800 of the original control. (B): Start dates for PM ensemble experiments (cyan dots). The new control is used to define thresholds to select high/low/typical SIV years. The $\pm 1.2\sigma$ levels and $\pm 0.25\sigma$ levels are indicated by horizontal red lines. (C): Evolution of volume anomalies from an ensemble initialized on January 1 of year 839. The black line shows the control run realization.

amine pan-Arctic [54] and regional [12] SIE prediction skill in addition to a diverse set of other climate prediction applications, including regional SST [67], tropical cyclones [72, 55], temperature and precipitation over land [39, 38], and extratropical storm tracks [79]. Using FLOR for both the PM and OP predictions allows us to make a clean "apples-to-apples" comparison between operational and potential prediction skill within the same prediction system.

2.5 Operational Prediction Skill Metrics

We assess the skill of the OP predictions using the anomaly correlation coefficient (ACC) and the mean-squared skill score (MSSS). We let o and p be observed and predicted values, respectively, of a time series of interest, for example pan-Arctic SIE. We let τ be the forecast lead time, o_j be the observed value at time j, K be the number of years in the observed timeseries, and N be the number of prediction ensemble members. We let $p_{ij}(\tau)$ be the predicted value given by the ith ensemble member initialized τ months prior to time j. Our lead τ prediction of o_j is given by the ensemble-mean prediction $\langle p_j(\tau)\rangle$, where:

$$\langle p_j(\tau)\rangle = \frac{1}{N}\sum_{i=1}^{N} p_{ij}(\tau). \tag{1}$$

We let $\bar{\cdot}$ denote the time-mean over the K samples. The ACC is given by the Pearson correlation coefficient between the predicted and observed timeseries:

$$ACC(\tau) = \frac{\sum_{j=1}^{K}\left(\langle p_j(\tau)\rangle - \overline{p(\tau)}\right)\left(o_j - \bar{o}\right)}{\sqrt{\sum_{j=1}^{K}\left(\langle p_j(\tau)\rangle - \overline{p(\tau)}\right)^2}\sqrt{\sum_{j=1}^{K}\left(o_j - \bar{o}\right)^2}}. \tag{2}$$

The mean-squared error (MSE) is given by

$$MSE(\tau) = \frac{\sum_{j=1}^{K}\left(\langle p_j(\tau)\rangle - o_j\right)^2}{K}, \tag{3}$$

and the MSE of a climatological forecast $\bar{o}$ is given by

$$MSE_{clim} = \frac{\sum_{j=1}^{K}(\bar{o} - o_j)^2}{K}. \tag{4}$$

The MSSS [56] is a skill score based on a comparison between MSE and MSE_{clim}, and is given by

$$MSSS(\tau) = 1 - \frac{MSE(\tau)}{MSE_{clim}}. \tag{5}$$

The MSSS is directly related to the ACC via the decomposition of [56], which shows that

$$MSSS(\tau) = ACC^2(\tau) - \left(ACC(\tau) - \frac{\sigma_p}{\sigma_o}\right)^2 - \frac{(\overline{p(\tau)} - \bar{o})^2}{\sigma_o^2}, \tag{6}$$

where the last two terms are negative definite and correspond to the conditional and unconditional forecast biases, respectively, and σ is the standard deviation of the given time series. The unconditional bias term is related to the mean offset between the observed and predicted time series, whereas the conditional bias term represents the degree to which the slope of the regression line between these time series deviates from 1 (i.e. the degree to which predictions are underconfident or overconfident).

Since the focus of this study is the initial-value predictability of Arctic sea ice, we assess prediction skill relative to a linear trend reference forecast. Specifically, we detrend the regional SIE time series' using a linear trend forecast which is updated each year using all available past data [60,12] and compute OP ACC and MSSS values using these detrended data. This differs from the approach used in other hindcast studies, which compute detrended anomalies using linear or quadratic trends based on the full hindcast period, providing an *a posteriori* assessment of detrended prediction skill [75,18,65,52,54,59,33,27]. A drawback to this full-hindcast period approach is that the detrended anomaly of a given year relies upon future information, and therefore the linear trend reference forecast does not represent a viable forecasting strategy. The approach employed here ameliorates this issue, by computing a linear trend forecast each year using all available past data (we assume a linear trend of zero for the first three hindcast years). After this detrending, the OP ACC and MSSS can be cleanly compared to the PM ACC and MSSS, respectively. Note that we also computed detrended regional SIE prediction skill using linear and quadratic trends computed over the full hindcast period, and found that regional prediction skill is relatively insensitive to the choice of detrending method.

2.6 Perfect Model Skill Metrics

We next introduce a set of predictability metrics, which are used to judge the prediction skill of the PM experiments. These metrics utilize a technique commonly used in the PM literature [19,34] in which each ensemble member in turn is taken to be the "truth" and the remainder of the ensemble is used to predict this "truth" member. In order to facilitate a clean comparison between OP and PM skill, we define our PM skill metrics in analogy to the OP skill metrics presented in the previous section. Note that these metrics differ somewhat from other metrics commonly used in the PM predictability literature [19,61,34], and offer conceptual advantages when comparing to OP prediction skill (see Appendix 6.2 for a discussion of how these metrics relate to other commonly used definitions). In particular, these PM metrics can be compared directly with their OP analogues, while other commonly used PM metrics cannot.

We let x be a timeseries of interest, for example pan-Arctic SIE or SIV. We let $x_{ij}(\tau)$ be the prediction of x from start date j and ensemble member i at lead time τ. Suppose that we have M ensemble start dates, with each ensemble consisting of N members (in this study $M = 6$ and $N = 12$). We now motivate a definition for the PM MSE. Suppose that ensemble member i is the synthetic observation (the "truth" member). We use the remaining $N-1$ ensemble members to predict this synthetic observation. Specifically, we take the ensemble mean of these $N-1$ members as our prediction of x_{ij}. As a notation, we let $\mathbf{x}_{\hat{i}j}$ be a vector of ensemble members from the jth ensemble with the ith member removed:

$$\mathbf{x}_{\hat{i}j} = (x_{1j}, \ldots, x_{i-1j}, x_{i+1j}, \ldots, x_{Nj}), \tag{7}$$

and let $\langle\cdot\rangle$ denote the ensemble mean operator. Thus, $\langle \mathbf{x}_{\hat{i}j}(\tau)\rangle$ is our prediction of x_{ij}, and has a squared error of $(\langle \mathbf{x}_{\hat{i}j}(\tau)\rangle - x_{ij}(\tau))^2$. Letting each ensemble member take a turn as the truth and averaging over all ensemble members (N) and ensemble start dates (M), we obtain the mean-squared error (MSE):

$$MSE(\tau) = \frac{\sum_{j=1}^{M}\sum_{i=1}^{N}\Big(\langle \mathbf{x}_{\hat{i}j}(\tau)\rangle - x_{ij}(\tau)\Big)^2}{MN}. \tag{8}$$

This metric is the PM analogue to the OP MSE defined in Eqn. 3. This MSE formula satisfies a necessary condition for forecast reliability [41,58,40,48,76], which states that the MSE of ensemble-mean forecasts is equal to the mean intra-ensemble variance, σ_e^2, up to a scaling factor related to the finite ensemble size. Specifically, we show in Appendix 6.1 that

$$MSE(\tau) = \frac{N}{N-1}\sigma_e^2(\tau), \tag{9}$$

where

$$\sigma_e^2(\tau) = \frac{1}{M}\sum_{j=1}^{M}\frac{1}{N-1}\sum_{i=1}^{N}\Big(\langle \mathbf{x}_j(\tau)\rangle - x_{ij}(\tau)\Big)^2, \tag{10}$$

and $\langle \mathbf{x}_j(\tau)\rangle$ is the ensemble mean of the jth ensemble.

We can now define a PM MSSS, given by

$$MSSS(\tau) = 1 - \frac{MSE(\tau)}{\sigma_c^2}, \tag{11}$$

where σ_c^2 is the climatological variance of x computed from the control run. σ_c^2 is the MSE of a climatological reference forecast, which can be seen by replacing the ensemble-mean forecast in Eqn. 8 with μ, the monthly climatological mean of the control run. In practice, computing the climatological variance from the control run is more robust than using Eqn. 8, due to the relatively small number of start dates used in most PM studies. MSSS values close to one indicate high PM skill and a value of zero indicates no prediction skill relative to a climatological forecast. The MSSS is closely related to the potential prognostic predictability (PPP, [61]), and can be interpreted analogously (see Appendix 6.2).

We also consider root-mean squared error (RMSE)

$$RMSE(\tau) = \sqrt{MSE(\tau)}, \tag{12}$$

which quantifies the error in physical units, and the normalized RMSE (NRMSE),

$$NRMSE(\tau) = \frac{RMSE(\tau)}{\sigma_c}, \tag{13}$$

which normalizes the RMSE by the RMSE of a climatological forecast. NRMSE values close to zero indicate skillful PM predictions and a value of one indicates no prediction skill relative to a climatological forecast. The MSSS is directly related to the NRMSE via

$$MSSS(\tau) = 1 - (NRMSE(\tau))^2. \tag{14}$$

This RMSE definition provides a more natural comparison with OP RMSE than the definition of [19] (which includes an additional factor of $\sqrt{2}$), reducing potential for confusion when interpreting PM RMSE values (see Appendix 6.2).

We define the ACC as the correlation between predicted and "observed" anomalies, where each ensemble member x_{ij} takes a turn as the "truth" and the ensemble means $\langle \mathbf{x}_{\hat{i}j}(\tau)\rangle$ are used to predict these synthetic observations:

$$ACC(\tau) = \frac{\sum_{j=1}^{M}\sum_{i=1}^{N}\Big(\langle \mathbf{x}_{\hat{i}j}(\tau)\rangle - \mu(\tau)\Big)\Big(x_{ij}(\tau) - \mu(\tau)\Big)}{\sqrt{\sum_{j=1}^{M}\sum_{i=1}^{N}\Big(\langle \mathbf{x}_{\hat{i}j}(\tau)\rangle - \mu(\tau)\Big)^2}\sqrt{\sum_{j=1}^{M}\sum_{i=1}^{N}\Big(x_{ij}(\tau) - \mu(\tau)\Big)^2}}. \tag{15}$$

Note that the anomalies are computed relative to $\mu(\tau)$, which is the climatological value of x at lead time τ computed using the control run. In a non-stationary climate, μ is a function of start date j. Given that the control run considered in this study has a statistically steady climate, we drop the j dependence in this formula. ACC values near 1 indicate high PM skill, and values of zero indicate no skill relative to a climatological forecast.

2.7 Significance Testing

Throughout the manuscript, we assess statistical significance using a 95% confidence level. The statistical significance of the PM RMSE, NRMSE, and MSSS values is assessed using an F-test based on the F_{MN-1,s^*-1} distribution, where M and N are the number of start dates and ensemble members from the PM experiments, respectively, and s^* is the effective number of degrees of freedom in the control run, given by $s^* = s\frac{1-r(\Delta t)^2}{1+r(\Delta t)^2}$ where s is the number of samples in the control run and $r(\Delta t)$ is the lag-1 year autocorrelation computed from the control run [8]. For the initialized forecast RMSE, NRMSE, and MSSS values, we use an F-test based on the F_{K^*-1,K^*-1} distribution. Here K^* is given by $K^* = K\frac{1-r_1(\Delta t)r_2(\Delta t)}{1+r_1(\Delta t)r_2(\Delta t)}$, where $K = 35$ is the number of years in the retrospective forecast experiments and $r_1(\Delta t)$ and $r_2(\Delta t)$ are the lag-1 year autocorrelation values for each time series.

We assess whether the PM ACC values are significantly greater than zero based on a t-test with $MN - 2$ degrees of freedom. Similarly, we assess the OP ACC values using a t-test with $K^* - 2$ degrees of freedom. Scatterplots of predicted vs observed regional SIE show that the assumptions of linearity and homoscedasticity are satisfied in all regions except for the Central Arctic, which is fully ice-covered for many of the verification years. When directly comparing PM and OP forecast ACC, we use the OP forecast significance threshold, which is the higher (more conservative) threshold of the two.

3 Pan-Arctic Predictability

3.1 Pan-Arctic SIV

We begin by investigating the ensemble evolution and PM prediction skill for pan-Arctic SIV. As an example, Fig. 2 shows the ensemble evolution of SIV anomalies for ensembles initialized in year 839, a high volume year. As the ensembles evolve in time, they progressively diverge under the chaotic dynamics of the system. This divergence occurs on a timescale of years for pan-Arctic SIV: After three years of integration, most ensemble members have retained a portion of their initial positive SIV anomaly, indicating that SIV is predictable beyond three-year lead times in this model. The rate of ensemble divergence also has a clear seasonal dependence. In particular, the ensemble members diverge rapidly over the months of May–July, and experience a much slower rate of divergence over the late summer, fall, and winter months (for example, compare the May initialized ensemble to the July initialized ensemble). This qualitative behavior is consistent with the physical expectation that the positive ice-albedo feedback should drive rapid ensemble divergence during the months of maximum solar insolation. Conversely, negative feedbacks active in fall and winter should act to reduce ensemble divergence, possibly even leading to ensemble convergence. These feedbacks include the negative feedback between ice growth and ocean entrainment ([51], ice growth increases the amount of heat entrained into the mixed layer, reducing ice growth rates), ice growth and ice thickness ([3], thin ice has larger growth rates than thick ice), and ice strength and ice thickness ([57], thin, weak ice has a greater propensity for thickening via ice convergence and for open-water formation via ice divergence, which leads to increased thermodynamic growth).

The PM skill metrics help to quantify the qualitative impressions obtained from Fig. 2. In Fig. 3, we plot the PM RMSE, NRMSE, ACC, and MSSS for pan-Arctic SIV. Note that each of these curves is computed over all six start years. Each of these metrics shows statistically significant prediction skill for SIV to lead times beyond 36 months, consistent with earlier PM studies [6, 70, 21, 30, 22]. We find that error growth rates and normalized error growth rates, as indicated by the slopes of the RMSE and NRMSE curves, respectively, vary strongly with target month. For both RMSE and NRMSE, the largest error growth occurs in May–July, which is followed by a sharp decrease in error growth in August and September. These low error growth rates continue into the fall and winter seasons, reaching their lowest values in the months of January–April (the error growth rates are negative in the winters of the second and third years). This is followed by rapid error growth in May as the melt season begins, and the error growth cycle roughly repeats again. Similar behavior is also observed in the ACC and MSSS metrics, with precipitous decreases in skill from May–July and much slower skill declines for the remainder of the year. The MSSS, and to a lesser extent the ACC, display a winter reemergence of prediction skill in years two and three, in which the winter skill values are higher than the skill of the previous summer.

The clear seasonality of SIV error growth rates highlights the crucial importance of initialization month in Arctic SIV predictions. In particular, there is a significant skill gap between predictions initialized prior to June and those initialized post June, suggesting a melt season “predictability barrier” for SIV. These results demonstrate that this barrier lies somewhere between May 1 and July 1,

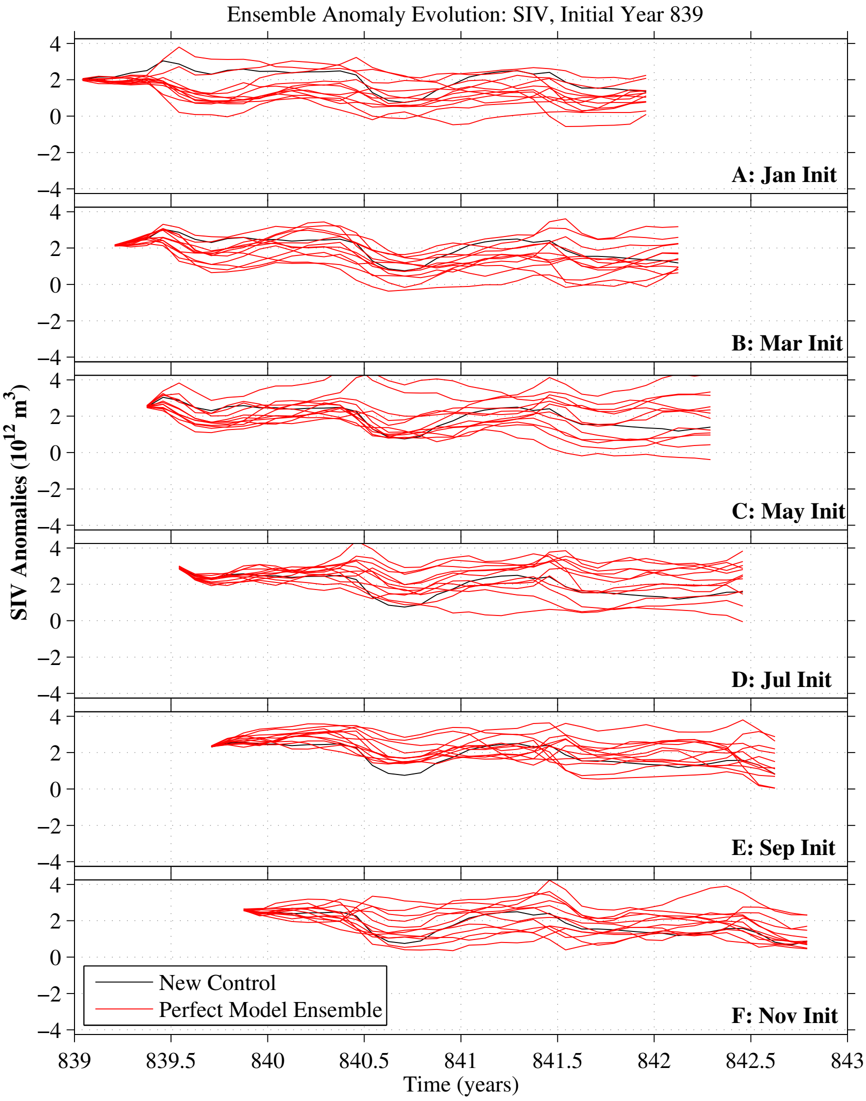

Fig. 2 Temporal evolution of sea ice volume anomalies for ensembles initialized in year 839 and months (A) January; (B) March; (C) May; (D) July; (E) September; (F) November. The control run realization is shown in black.

but further experiments are required to pinpoint its precise date. In other words, how far into the melt season must a prediction be initialized in order to avoid the unpredictable effects of atmospheric chaos, melt onset variability, and ice-albedo feedbacks? It is important to note that while this melt season predictability barrier is quite stark for SIV, it is less clearly defined for predictions of pan-Arctic SIE (see subsection 3.4, ahead).

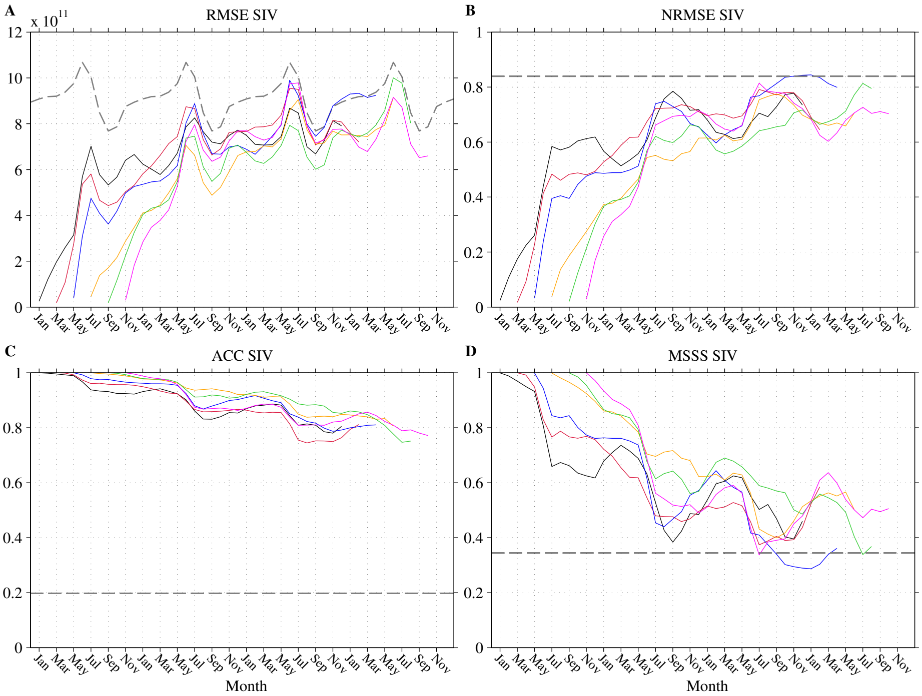

Fig. 3 Pan-Arctic SIV PM prediction skill for different initialization months. Shown here are the temporal evolutions of (A) RMSE; (B) NRMSE; (C) ACC; and (D) MSSS. The curves are colored based on their initialization month. The gray dashed lines indicate the 95% threshold for statistical significance. Note that the RMSE significance level is not constant due to the seasonal cycle in pan-Arctic SIV standard deviation.

3.2 State-dependence of predictability

Next, we consider the state-dependence of SIV predictability, asking: Does the initial SIV state have an influence on SIV predictability characteristics? In Fig. 4, we plot SIV predictability metrics for each initial month binned into high, low, and typical volume states. For the skill metrics based on ensemble spread (RMSE, NRMSE, and MSSS), we find no clear dependence on the volume state; however, the ACC metric shows a striking difference between the high/low volume states and the typical volume states. This result is consistent with the findings of [22] and can be explained via the ACC formula given in Eqn. 15. For the high/low volume ensembles, the ensemble means retain positive/negative anomalies over some timescale as the model relaxes towards its climatology (e.g. Fig. 2a), and the ensemble members fluctuate randomly around this ensemble mean. Therefore, the high/low ensembles each contribute positive values to the numerator of Eqn. 15, since both the synthetic observations and synthetic predictions have like-signed anomalies. On the other hand, the typical-anomaly ensembles fluctuate randomly around a near-zero anomaly state, making both positive and negative contributions to the numerator of Eqn. 15, and producing an ACC that is close to zero. A similar

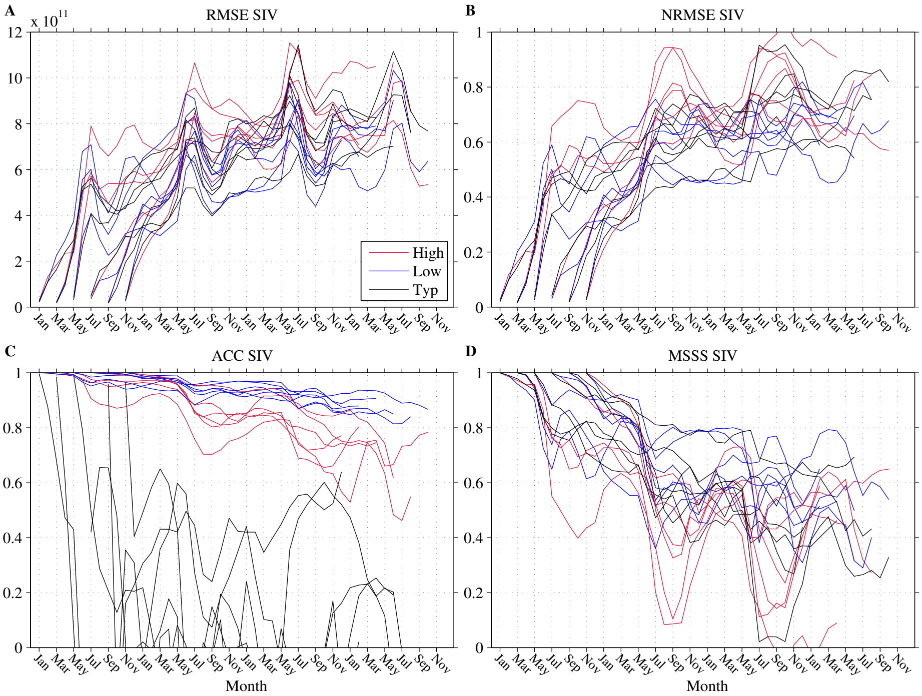

Fig. 4 PM prediction skill (A: RMSE; B: NRMSE; C: *ACC*; D: MSSS) for pan-Arctic SIV in high (red curves), low (blue curves), and typical (black curves) volume states for different initialization months.

ACC state-dependency holds for pan-Arctic SIE and other variables (not shown).

3.3 An unbiased estimate of perfect model *ACC*

Because the PM *ACC* is strongly state dependent, the *ACC* computed using Eqn. 15 will be highly sensitive to the set of start dates chosen for a given PM study. This is an important caveat to consider when evaluating PM *ACC*: If start dates are not drawn randomly from the climatological distribution of states, the *ACC* estimates will have systematic biases. For example, in this study, start dates were selected specifically to have high, low, and typical volume states (see Fig. 1b). These states do not obey the climatological distribution of volume states, as four of the six have notably large anomalies. Since large-anomaly states have higher *ACC* values, our *ACC* estimates are likely biased high due to the non-random sampling of start dates used in this study.

To remedy this issue, we appeal to the decomposition of [56], which relates the MSSS to the *ACC* (see Eqn. 6). In a PM framework, predictions are free of conditional and unconditional biases, therefore [56] suggests that the identity $MSSS = ACC^2$ should hold for PM predictions [70,34]. However, we find that PM $MSSS$ is not equal to ACC^2 (e.g. see Fig. 6, ahead). Why is this? The decomposition of [56] is a mathematical identity, which holds identically when the climatological mean and variance are computed "in sample" (i.e. using the available samples from the PM experiments, and not the control run values). In Eqns. 11 and 15, the climatological mean and variance are computed using the control run. If the start dates are non-randomly sampled, the control run mean and variance will be biased relative to the "in sample" mean and variance. This results in a breakdown of the decomposition of [56]. Since the $MSSS$ shows much less sensitivity to start date than the ACC, it is less prone to sampling bias, and provides a more robust assessment of PM skill. We use this fact to define an unbiased estimate of the ACC, ACC_U, which can be cleanly compared to OP ACC values:

$$ACC_U = \sqrt{MSSS}. \tag{16}$$

The ACC_U is the value the ACC would have if the decomposition of [56] held, which is the case when the PM states are sampled from the climatological distribution. Therefore, up to the independence of $MSSS$ with respect to start date, this formula provides an ACC estimate which is insensitive to start-date sampling error. In the following section, we directly compare OP ACC and PM ACC_U. Note that we could also directly compare OP and PM predictions based on MSSS values. If this comparison is made, many of the skill structures present in OP ACC are degraded and the PM/OP skill gap is larger than the gap based on ACC, due to conditional biases in the OP predictions (not shown). For these reasons, we make our skill comparisons using OP ACC, which provides a lower bound on the PM/OP skill gap.

3.4 Pan-Arctic SIE predictability

In this subsection, we compare the PM and OP prediction skill of pan-Arctic SIE. Figure 5 shows the evolution of RMSE, NRMSE, ACC, and MSSS for different initialization months for both PM and OP predictions of pan-Arctic SIE. Figure 6 takes a different vantage point, plotting the skill as a function of target month (the month we are trying to predict) and forecast lead time. These "target month" style PM skill plots are a unique contribution of this study, made possible by our choice of equally-spaced initialization months spanning the calendar year. Previous PM studies have typically focussed on January and/or July initializations, not providing enough initial-month "resolution" to construct a target-month style plot. These plots allow for a systematic study of the skill dependence on target month, initial month, and lead time. Note that we have PM predictions initialized at two-month intervals. For example, for target month January, we have predictions for all even lead times, from lead-0 through lead-34 (note that a lead-0 prediction is defined as the January-mean value from a prediction initialized on January 1). To obtain skill estimates for the odd lead times, we perform a linear interpolation between the even-lead values. This method provides reasonable results, as most skill variations occur over lead times of many months (see Fig. 6).

We find a striking gap between the PM and OP prediction skill for pan-Arctic SIE. While the OP predictions have statistically significant ACC at lead times of 0–5 months depending on the target month (Fig. 6c), the PM predictions have statistically significant ACC and ACC_U up to lead times of 35 months, for all

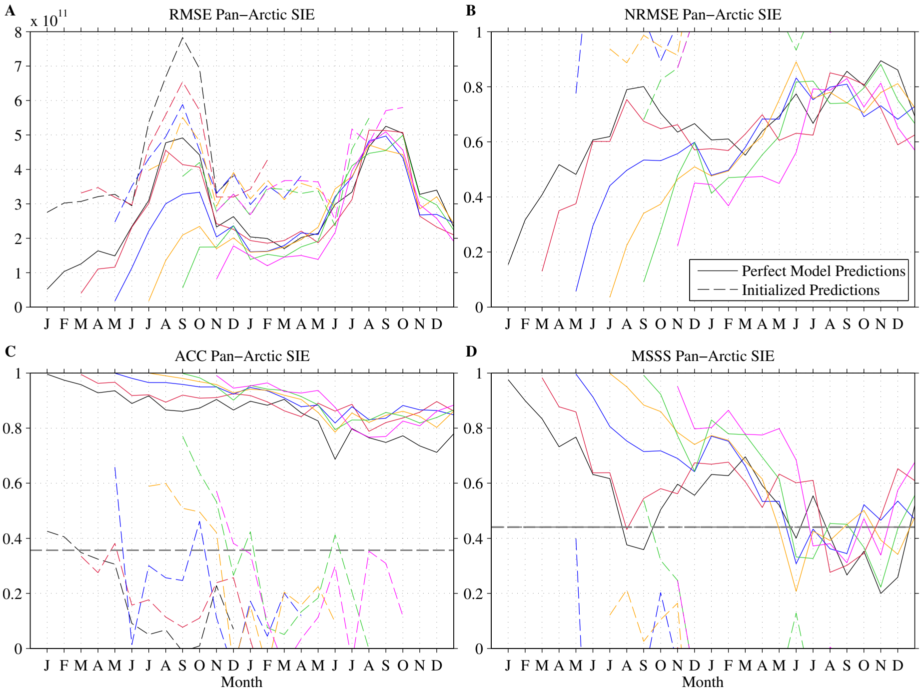

Fig. 5 Comparison of PM (solid lines) and OP (dashed lines) prediction skill (A: RMSE; B: NRMSE; C: *ACC*; D: MSSS) for pan-Arctic SIE for different initialization months. The 95% significance levels for *ACC* and MSSS are plotted as dashed gray lines.

target months (Fig. 6a,b). It is important to note that PM skill should be considered an upper limit of prediction skill, and may overestimate the skill achievable in reality (see discussion in Section 4.3, ahead). Nevertheless, the skill gap shown in Fig. 5 and 6 suggests that substantial skill improvements are possible in current OP prediction systems. In particular, Fig. 5 shows large differences in lead-0 skill, indicating that the OP predictions likely suffer from initialization errors and/or initialization shocks. These lead-0 predictions could presumably be improved by assimilating more observational data, improving data assimilation techniques, and expanding existing observational networks. In addition, we find that the loss of skill in the OP predictions occurs much more rapidly than in the PM experiments. This rapid loss of skill likely results from a combination of i) model physics errors; ii) model drift associated with initialization shock; and iii) differences between the model and nature in their underlying predictability, possibly resulting in an overestimated upper limit of predictability in the PM experiments.

Comparing Fig. 6a and 6b, we find that pan-Arctic SIE ACC is higher than ACC_U, consistent with our *a priori* expectation from subsection 3.3. ACC and ACC_U offer similar qualitative conclusions, but have quantitative differences when assessing limits of predictability. For the skill comparisons throughout the remainder of the paper, we will use the ACC_U values when comparing to OP prediction ACC. The PM skill shows a clear seasonality, with higher skill for winter SIE predictions than summer SIE. As a reference-level for a "highly skillful" prediction,

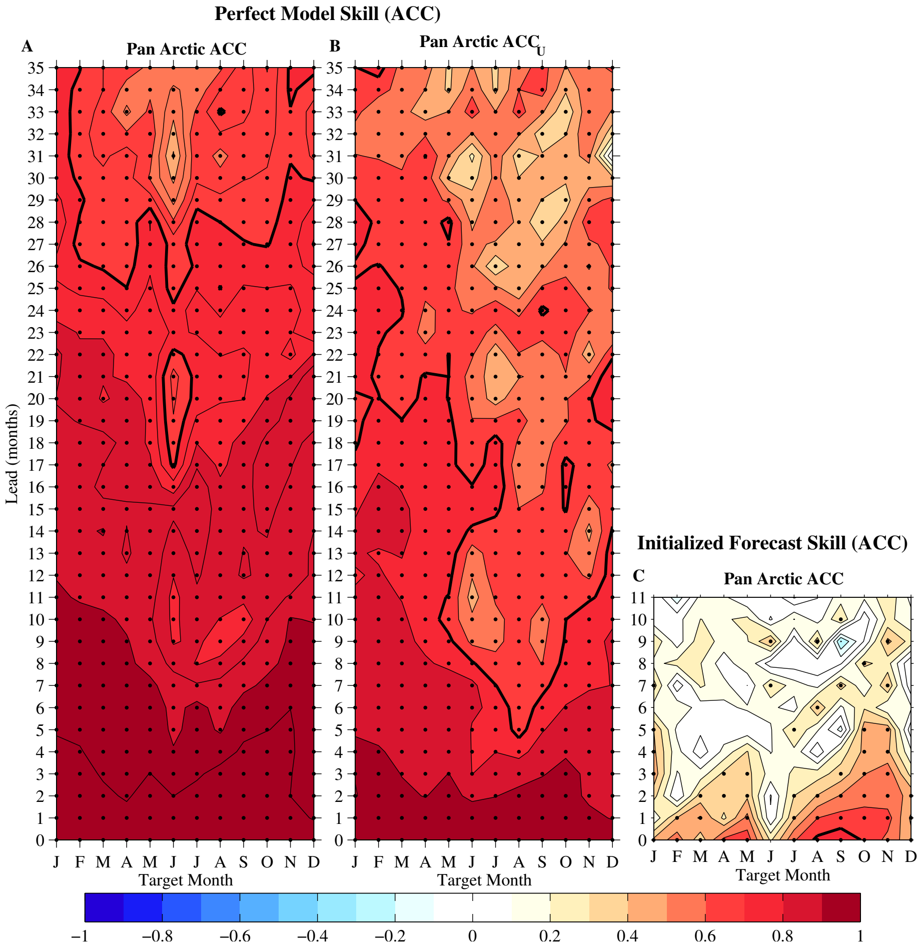

Fig. 6 Comparison of PM and OP prediction skill for pan-Arctic SIE, plotted as a function of target month and forecast lead time. Panel A shows PM ACC computed using Eqn. 15, Panel B shows PM unbiased ACC, defined as $ACC_U = \sqrt{MSSS}$, and Panel C shows ACC from the OP prediction experiments. The thick black lines indicate the ACC=0.7 contours. Dots indicate months in which the ACC values are statistically significant at the 95% confidence level.

we have marked the $ACC = 0.7$ contour in Fig. 6, as this is the level at which half the variance of the observed signal can be predicted. This shows that half the winter SIE variance is predictable at 18-26 month lead times, whereas the analogous limits for summer SIE are 5-11 months.

The study of [21] found evidence of a May "predictability barrier" for pan-Arctic SIE, in which predictions initialized in May lost skill more rapidly in the first four months than those initialized in January or July. In this model, there is no clear evidence of such a barrier, as the error growth rates over the first four months are similar for all initialization months (see Fig. 5b,d). Also, a May predictability barrier would result in a diagonal ACC_U feature corresponding to initial month May in Fig. 6b, which is not seen. This lies in contrast to SIV, which shows clear evidence of a melt-season predictability barrier (see Fig. 3). Interestingly, the OP predictions of summer SIE show evidence of a spring prediction skill barrier, with lower skill for forecasts initialized prior to May. A similar feature is also seen in SIE persistence forecasts (see Fig. S8), suggesting that SIE persistence is a key source of skill for the OP predictions, whereas the PM predictions presumably benefit from other sources of predictability, such as perfect SIT ICs, which extend skill beyond this barrier. We find that both PM and OP predictions show spring skill barriers in certain regions, which we explore in Section 4 ahead.

4 Regional Sea-Ice Predictability

4.1 SIC Predictability

In this section, we move to smaller spatial scales, exploring the ability of this model to make skillful predictions at the regional and gridpoint scale. In Fig. 7, we plot PM MSSS values for SIC for different target months and lead times of 0–14 months. We find that for all target months, the lead-0 SIC predictions are highly skillful, indicating a year-round potential for regional-scale sub-seasonal sea-ice predictions in this model. The loss of SIC predictability with lead time is highly dependent on the region and target month. We observe a clear difference between summer and winter SIC predictions, with summer predictions losing most of their skill beyond six-month lead times and winter predictions retaining skill beyond 14-month lead times. This long-lead winter prediction skill is notably high in the Barents and GIN Seas, with lower values in the Labrador, Bering, and Okhotsk Seas. The SIC prediction skill for even target months and odd lead times has analogous skill characteristics (not shown).

To synthesize the information of Fig. 7, we introduce a "predictable area" metric, defined as

$$\text{Predictable area}(\tau) = \frac{\int MSSS(x, y, \tau)dA}{\int MSSS(x, y, \tau = 0)dA}, \tag{17}$$

which is the area integral of the SIC MSSS for a given target month, normalized by its lead-0 value. Fig. 8 shows the evolution of SIC predictable area with lead time. We find that predictions of summer and winter SIC lose predictability at a similar rate over the first 3 months, after which the rates of predictability loss begin to diverge. At lead times beyond 6 months, the winter SIC predictions (target months December–May) have higher predictable area values than summer SIC predictions (target months June–November). Consistent with the pan-Arctic SIE results, this shows that there is a greater potential for skillful long-lead predictions of winter SIC compared with summer SIC.

4.2 Regional SIE predictability

Next, we consider the predictability of regional SIE, providing a direct comparison between PM and OP regional SIE prediction skill. Regional SIE is likely a more "forgiving" metric than SIC, as it is less sensitive to local-scale ice dynamics associated with unpredictable atmospheric forcing. The region definitions follow those

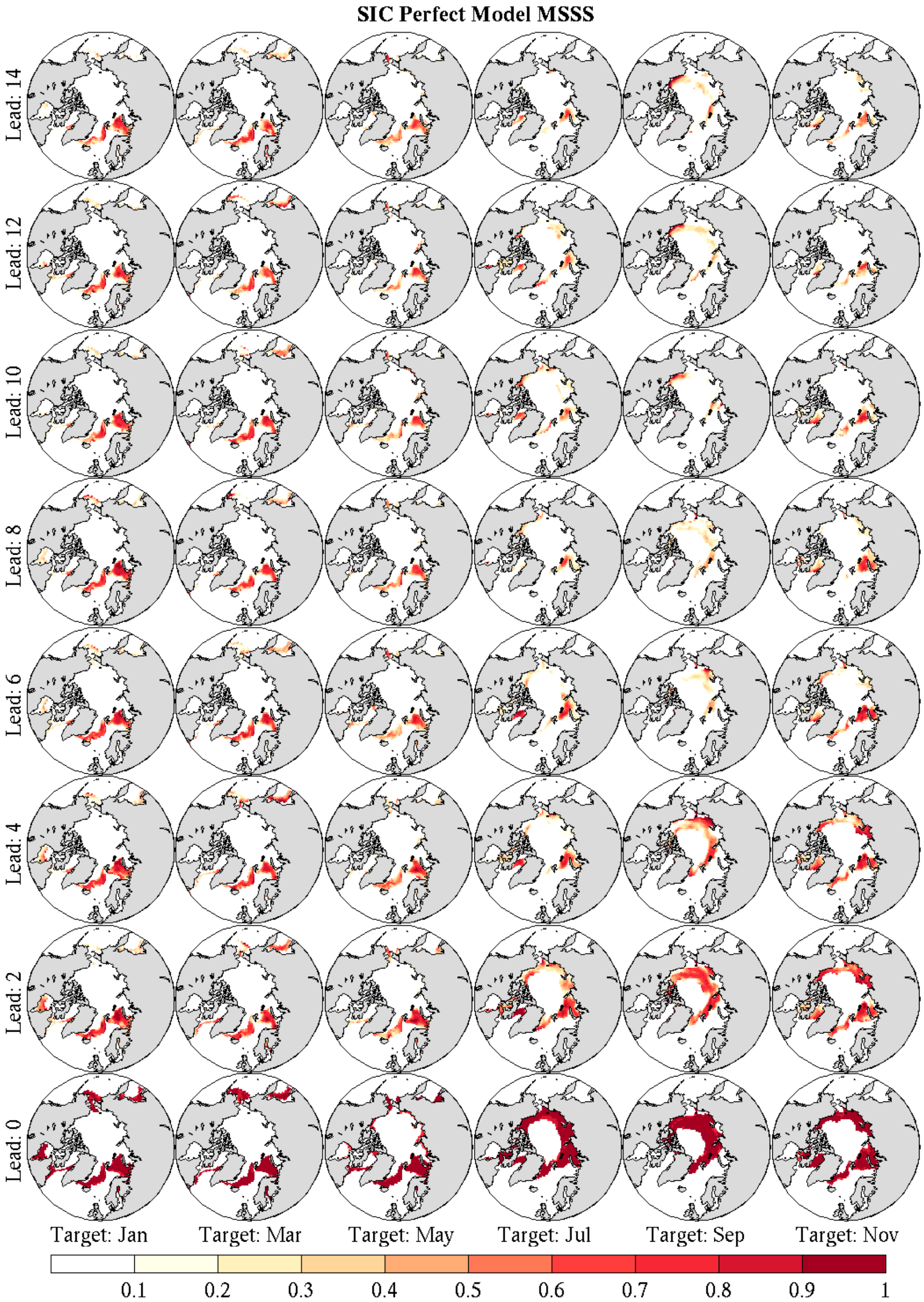

Fig. 7 SIC PM MSSS for different target months and lead times of 0–14 months. A mask has been applied such that only gridpoints with SIC standard deviation greater than 10% are plotted.

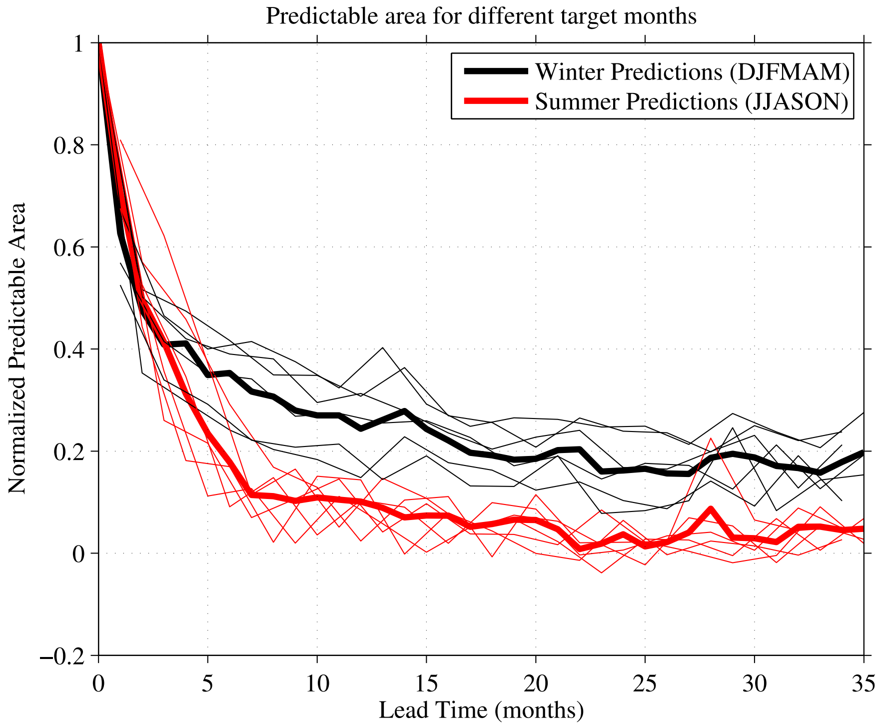

Fig. 8 SIC predictable area (see Eqn. 17) vs lead time. Each winter target month (December–May) is plotted as a thin black curve and each summer target month (June–November) is plotted as a thin red curve. The thick black and red curves are the mean over all winter and summer target months, respectively.

used in [21,12] (See Fig. S5). We find that for nearly all regions and all target months, there is a substantial gap between PM and OP prediction skill, indicating a potential for large improvements in regional SIE predictions (see Figs. 9–11). We also find that the ACC skill structures are broadly similar between the PM and OP predictions. This correspondence indicates that OP prediction skill is partially governed by the fundamental predictability limits found in the PM experiments, and that common physical mechanisms underlie the prediction skill of both PM and OP predictions. Finally, we find that the regional differences in PM prediction skill generally mirror the skill differences found in the OP SIE predictions.

In both the PM and OP predictions, the highest regional prediction skill is found for winter SIE in the North Atlantic sector (see Fig. 9). PM predictions in the Barents and GIN Seas are highly skillful (defined here as $ACC \geq 0.7$; a prediction capable of capturing more than half the variance) at lead times beyond 24 months. This lies in contrast to the OP predictions, which have statistically significant skill in these regions at lead times of 5–11 months, but are not highly skillful. In both PM and OP predictions, regional SIE skill in the North Pacific sector is lower than that of the North Atlantic. This suggests that the Bering Sea and Sea of Okhotsk are fundamentally less predictable, lacking the potential for highly skillful predictions beyond 12-month lead times. Compared with the

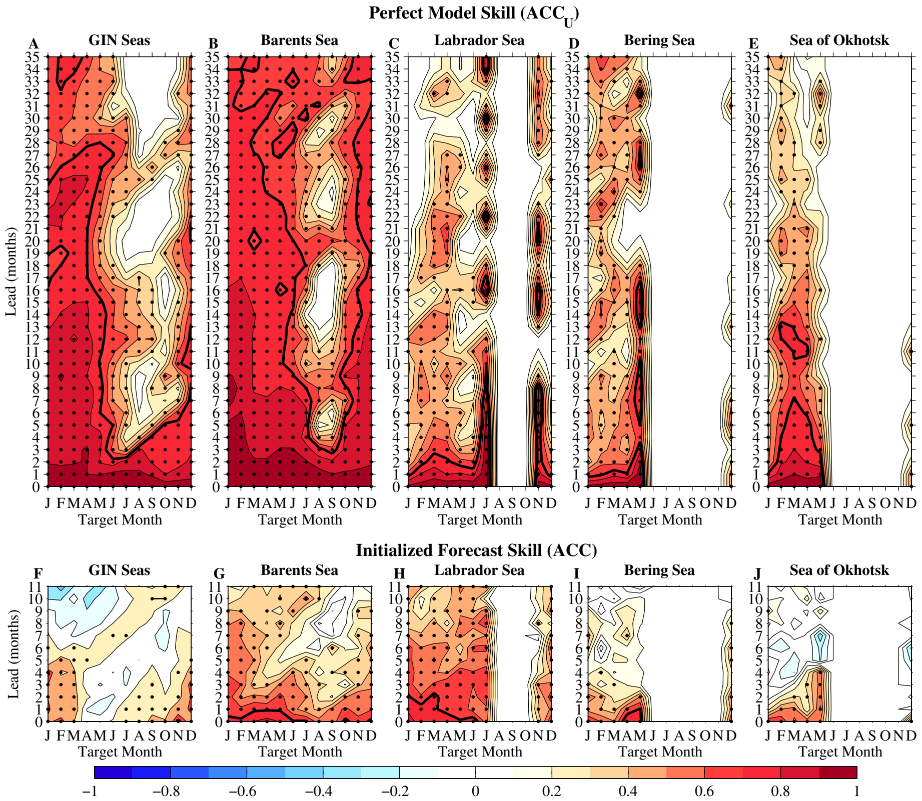

Fig. 9 Comparison of PM prediction skill (ACC_U) and OP prediction skill (ACC) for Arctic regional SIE for the GIN, Barents, Labrador, and Bering Seas and the Sea of Okhotsk. ACC values are plotted as a function of target month and forecast lead time, and are only plotted for target months with SIE standard deviation greater than 0.03×10^6 km^2. The thick black lines indicate the ACC=0.7 contours. Dots indicate months in which the ACC values are statistically significant at the 95% confidence level.

large PM/OP skill gap found in other regions, the Labrador Sea is an exception, showing similar PM and OP skill. The PM skill of this model may underestimate the fundamental limit of Labrador SIE predictability, as this model has too little SIC variability in this region (See Fig. S4). This SIC variability bias likely results from excessive deep open ocean convection in the Labrador sea, which restricts sea-ice variability in this region. Indeed, the study of [21] found that the Labrador Sea had the longest duration of predictability in HadGEM1.2, suggesting that model formulation and biases may strongly affect Labrador Sea predictability estimates.

The study of [12] identified a spring prediction skill barrier in the Laptev, East Siberian and Beaufort Seas, in which summer SIE prediction skill dropped off sharply for OP forecasts initialized prior to May, May, and June, respectively (see Fig. 10g,10h,10j). Interestingly, the PM forecasts show a similar skill barrier in these regions, with highly skillful summer SIE predictions for forecasts initialized May 1 and later, and a clear drop-off in skill for predictions initialized before this (see Fig. 10b,10c,10e). The diagonal ACC contours in these regions indicate that

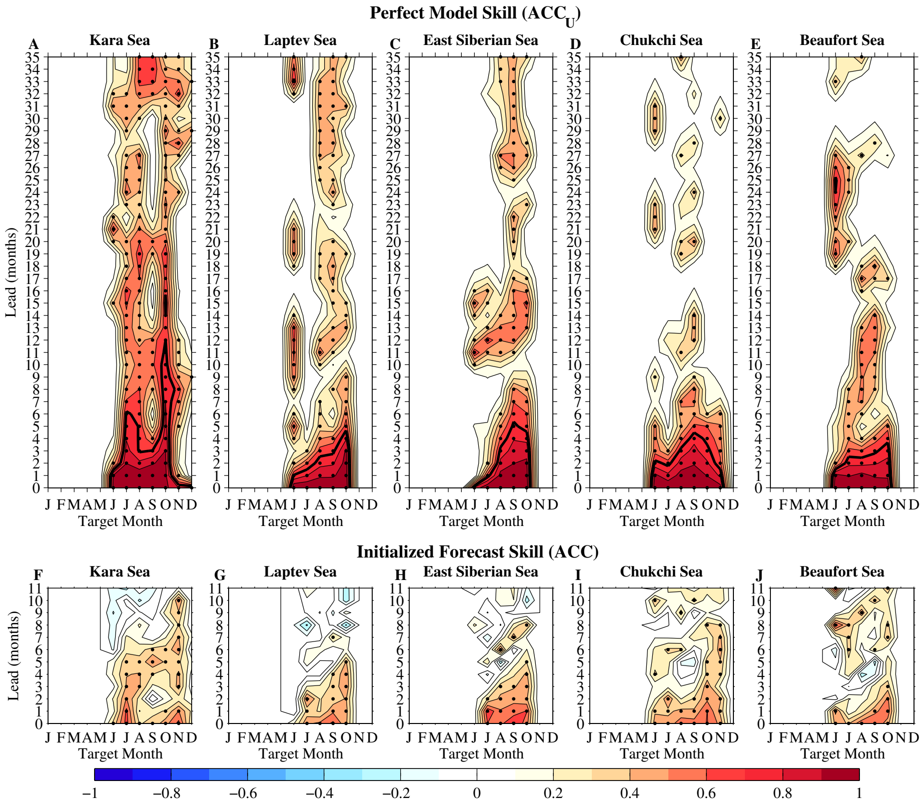

Fig. 10 Comparison of PM prediction skill (ACC_U) and OP prediction skill (ACC) for Arctic regional SIE for the Kara, Laptev, East Siberian, Chukchi, and Beaufort Seas.

summer SIE skill tends to be roughly constant for a given initialization month. The fact that the spring prediction skill barrier is present in both OP and PM predictions suggests that it is a fundamental predictability feature of this model, rather than resulting from IC errors in the OP predictions. In particular, the perfect SIT ICs in the PM experiments are not sufficient to overcome this spring barrier. Additional PM experiments using other GCMs are required to determine if the spring barrier is truly a feature present in nature. Summer SIE predictions in the Chukchi Sea are highly skillful at 2-4 month lead times in the PM experiments. While there is some diagonal structure in the Chukchi ACC plots, both the PM and OP predictions do not have a clearly defined spring barrier in this region. The Kara Sea has highly skillful PM predictions for summer and fall SIE at lead times of 2-11 months and also does not show a spring prediction skill barrier.

The Central Arctic has relatively low PM and OP prediction skill (see Fig. 11), whereas the Canadian Archipelago has slightly higher skill, with highly skillful PM forecasts of August and September SIE at 2-3 month lead times. The Canadian Archipelago results should be viewed with some caution, given the model's coarse resolution of this bathymetrically complex region. The PM forecasts have skill in predicting both melt season and growth season SIE anomalies in Hudson and

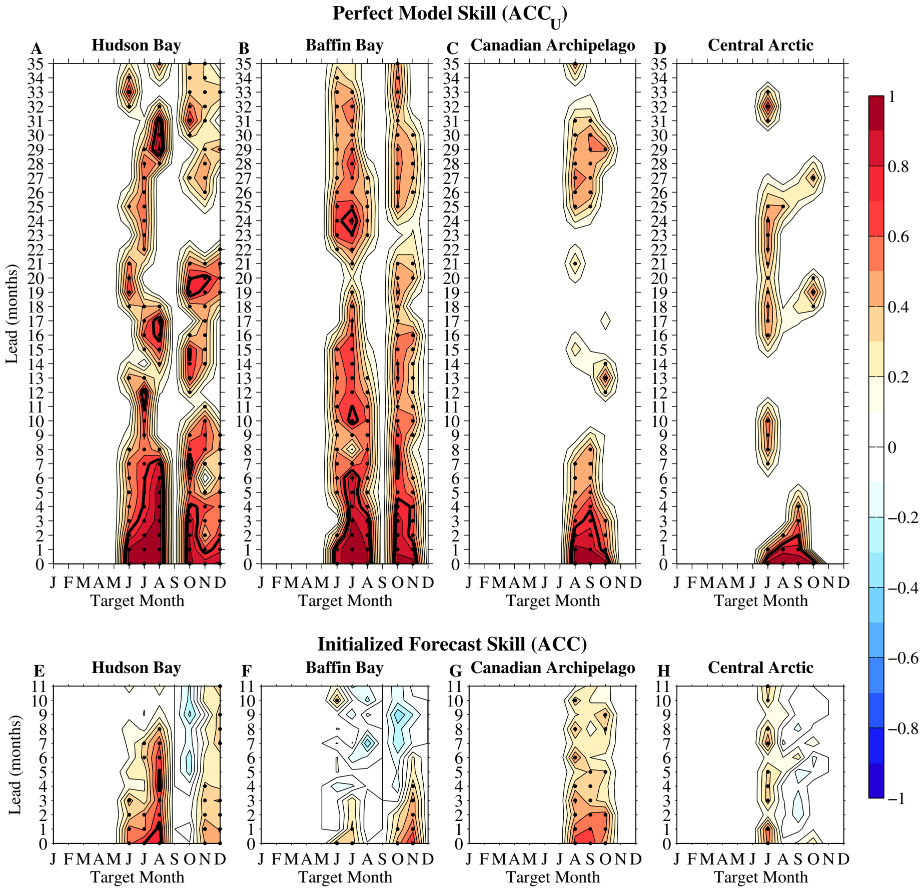

Fig. 11 Comparison of PM prediction skill (ACC_U) and OP prediction skill (ACC) for Arctic regional SIE for Hudson Bay, Baffin Bay, the Canadian Arctic Archipelago, and the Central Arctic.

Baffin Bay. In each of these regions, the melt season skill is higher than the growth season, suggesting that persistence of winter ice thickness anomalies is the greatest source of predictability in these regions. The Hudson and Baffin Bay OP skill is substantially lower than the PM skill, particularly for the growth season in Hudson Bay and the melt season in Baffin Bay. This skill discrepancy could possibly be reduced by directly assimilating SIT data in the OP system.

We also note that there are a small number of instances in which an isolated month shows OP skill but not PM skill (for example, lead-6 September predictions in the Barents Sea, lead-8 October/November predictions in the Chukchi Sea, and lead-4 November predictions in the Kara Sea). These instances tend to have fairly low skill ($ACC < 0.5$), suggesting that sampling errors in the OP predictions could be playing a role. Also, in some of these instances the PM skill does not decay monotonically with lead time, violating a property that we expect PM predictions to satisfy. This suggests that sampling errors in the PM predictions could also explain these discrepancies.

4.3 Interpretation of the PM/OP skill gap

The PM/OP skill gap demonstrated in Figs. 9-11 raises a natural question: To what extent can these PM skill estimates be realized in future OP prediction systems? In other words, is it valid to interpret the PM/OP skill gap as possible "room for improvement" in prediction skill? The work of [47] directly addresses these questions, providing a framework to assess the fidelity of PM skill estimates. [47] argue that the interpretation of the PM/OP skill gap as "room for improvement" relies on an implicit assumption that the observed and model-predicted time series' share the same statistical characteristics. In particular, they show that differences in PM skill between different models can largely be attributed to differences in temporal autocorrelation (persistence) and, by extension, argue that a model's temporal autocorrelation should be compared to observations before making inferences based on PM skill.

Following this, we compare the temporal autocorrelation of observed detrended regional SIE to the autocorrelation of the FLOR control run. Computing autocorrelation values for all target months and lead times of 0-35 months, we find that the model's regional SIE persistence characteristics are generally quite consistent with observed persistence (see Figs. S6-S8). In particular, we find strong agreement in the Laptev, East Siberian, Beaufort, Bering, Canadian Arctic Archipelago, Hudson Bay, Baffin Bay, and Central Arctic regions. This suggests that in these regions the PM skill provides a reliable estimate of the true upper limits of skill achievable in nature. In the Chukchi Sea, Kara Sea, and Sea of Okhotsk, the model autocorrelation values agree well with observations for lead times less than or equal to 6 months. For lead times beyond 6 months, the model has higher correlation values than observed, although the values are quite modest (less than 0.4). Since the majority of highly skillful PM predictions in these regions occur for lead times of 6 months or less, we conclude that the PM skill estimates are also quite reliable in these regions. We find a larger discrepancy in the GIN and Barents Seas, with the model displaying higher autocorrelation values than the observations, particularly for winters 1 and 2 years in advance of a given winter target month. This discrepancy could potentially arise due to the removal of low-frequency (period > 20 years) variability when the observed SIE is linearly detrended. However, we find that this cannot fully explain the discrepancy, as notable differences in autocorrelation remain present even if the model data is 20-year high-pass filtered. This suggests that the PM skill may overestimate the true upper limits of prediction skill in the Barents and GIN Seas. Conversely, we find that the model has lower autocorrelation values than detrended observations in the Labrador Sea, suggesting that the PM skill underestimates the true skill achievable in this region. This is consistent with the lack of a PM/OP skill gap in the Labrador Sea, and likely results from the model biases discussed in subsection 4.2. Finally, we find that the model's pan-Arctic SIE is substantially more persistent than detrended observations, suggesting that the PM skill overestimates the true upper limit of predictability for the pan-Arctic domain. Overall, these findings provide general confidence in the interpretation of the PM/OP skill gap as possible "room for improvement" in prediction skill, while highlighting some caveats that apply to the North Atlantic regions and the pan-Arctic domain.

5 Conclusions and Discussion

In this work, we have established the first direct comparison of perfect model (PM) and operational (OP) Arctic sea-ice prediction skill within a common prediction system. Using the GFDL-FLOR coupled GCM, we have performed two complementary suites of ensemble prediction experiments. The first is a suite of PM experiments, consisting of ensembles initialized in January, March, May, July, September, and November, and in high, low, and typical sea-ice volume (SIV) regimes. Secondly, we have utilized a suite of retrospective initialized OP predictions spanning 1981-2016 made with GFDL-FLOR. The skill comparison between these OP predictions and the PM experiments forms the basis of this study.

In order to make a robust skill comparison, we have introduced a set of PM skill metrics, defined in analogy with metrics used in OP prediction applications. These metrics were designed to allow for an "apples-to-apples" PM/OP skill comparison, and offer conceptual advantages over other commonly used PM skill metrics. We have found that PM skill metrics based on ensemble spread (RMSE, NRMSE, MSSS) do not have a clear dependence on the SIV state, whereas the ACC is clearly higher in high/low volume states compared with typical volume states. This state-dependency can lead to biased ACC estimates if start dates are not sampled from the climatological distribution. We have defined an unbiased ACC, ACC_U, which does not suffer from this sampling bias. All comparisons with OP prediction skill in this study were made using ACC_U. The unbiased ACC metric may be broadly useful for PM studies, since many of these studies do not sample start dates from the climatological distribution of states. Using these PM and OP skill metrics, we have investigated the predictability of pan-Arctic SIV, pan-Arctic SIE, and regional Arctic SIE.

This study has shown that PM predictions of pan-Arctic SIV and SIE have statistically significant skill for all target months and lead times up to 35 months (the length of our PM simulations). The PM predictions of pan-Arctic SIE are highly skillful ($ACC_U \geq 0.7$) at leads of 18–26 months for winter SIE predictions and leads of 5–11 months for summer SIE predictions. In contrast, OP predictions of pan-Arctic SIE have statistically significant skill at lead times of 0–5 months, and are not highly skillful beyond lead-0. This notable skill gap indicates that pan-Arctic SIE predictions could be improved in all months of the year, with particularly large opportunities for improvements in winter SIE predictions. Given that winter sea ice covaries strongly with the NAO (e.g. [26]) and that SIC anomalies can force an NAO response [25,69], improving winter SIE predictions has the potential to improve winter NAO predictions. For example, recent work by [73] shows that fall SIC is an important predictor of the winter NAO index, attributing their NAO skill to persistence of fall SIC conditions.

The uniform seasonal coverage of PM start dates employed by this study has allowed us to shed additional light on the spring predictability barrier for pan-Arctic SIE proposed by [21]. We have found that PM predictions of pan-Arctic SIV display a spring predictability barrier related to rapid error growth during the early melt season, in which predictions initialized prior to June lose skill much faster than those initialized post June. Unlike SIV, we have found that pan-Arctic SIE does not display a clear spring predictability barrier. This finding, which may be model-dependent, suggests that there is not an optimal month in which to initialize pan-Arctic SIE predictions. While the spring barrier is not present for pan-Arctic

SIE, we have found clear evidence of spring predictability barriers in certain Arctic regions. In particular, the Laptev, East Siberian, and Beaufort Seas each display spring prediction skill barriers in both the PM and OP predictions, suggesting that these barriers are a fundamental predictability feature of these regions. These barriers suggest that summer SIE predictions in these regions should be initialized May 1 or later, since skill is substantially lower for predictions initialized prior to May 1.

In nearly all Arctic regions, we have identified substantial skill gaps between PM and OP predictions of Arctic regional SIE. While their absolute skill values are different, the PM and OP regional predictions generally display similar correlation skill structures, indicating that similar physical mechanisms are contributing to both PM and OP skill. We have found that PM predictions in the Barents and GIN Seas are highly skillful at lead times beyond 24 months, whereas OP predictions have statistically significant skill at 5-11 months but are not highly skillful beyond 1 month lead times. In both the PM and OP predictions, the North Pacific sector has lower winter SIE skill than these North Atlantic regions, suggesting that the North Pacific is fundamentally less predictable. This finding is consistent with the PM study of [21] and the statistical prediction study of [81], and is relevant for fisheries industries active in these regions that could benefit from skillful winter SIE predictions.

We have found that regional winter SIE is generally more predictable than summer SIE. PM predictions of regional summer SIE in the Laptev, East Siberian, Chukchi, and Beaufort Seas are highly skillful at leads of 1–5 months, displaying similar correlation structures to their OP counterparts. The PM/OP skill gap suggests that substantial improvements are possible at these 1–5 month lead times, but that long-lead skillful predictions are not possible in these regions. This finding is relevant for the predictability of summer shipping lanes along the Northern Sea Route, implying that these lanes could be skillfully predicted from May 1, but not earlier.

This study has identified a striking skill gap between OP and PM predictions made with the GFDL-FLOR model, suggesting that skillful long-lead predictions of SIE are possible in many regions of the Arctic. The large gap in lead-0 prediction skill indicates a clear potential for improved predictions via improved initialization. Additionally, the rapid decay of OP prediction skill relative to the PM experiments indicates that improved model physics and/or more balanced ICs are required in future prediction systems. It is important to note that these findings are based upon a single GCM and similar studies with other seasonal prediction systems are required to solidify these results. This work has provided a robust comparison of regional PM and OP prediction skill, but has not investigated the physical mechanisms underlying this skill. Future work exploring these mechanisms, and identifying the key modeling and observational deficiencies in current dynamical prediction systems, is required in order to close the gap between PM and OP skill identified in this study.

Acknowledgements This paper is dedicated to Walter Bushuk. We thank two anonymous reviewers for constructive comments which improved the manuscript. We also acknowledge Olga Sergienko and Hiroyuki Murakami for comments on a preliminary version of the manuscript. We thank Seth Underwood, Bill Hurlin, and Chris Blanton for assistance in setting up the model experiments. M. Bushuk was supported by NOAA's Climate Program Office, Climate Variability and Predictability Program (Award GC15-504).

6 Appendix

6.1 Reliability condition for ensemble forecasts

Claim: The PM MSE given by Eqn. 8 satisfies the necessary condition for forecast reliability:

$$MSE(\tau) = \frac{N}{N-1}\sigma_e^2(\tau). \tag{18}$$

Proof: The mean intra-ensemble variance, σ_e^2, is given by

$$\sigma_e^2(\tau) = \frac{1}{M}\sum_{j=1}^{M}\frac{1}{N-1}\sum_{i=1}^{N}\Big(\langle \mathbf{x}_j(\tau)\rangle - x_{ij}(\tau)\Big)^2, \tag{19}$$

where $\langle \mathbf{x}_j(\tau)\rangle$ is the ensemble mean of the jth ensemble. The MSE is given by

$$MSE(\tau) = \frac{\sum_{j=1}^{M}\sum_{i=1}^{N}\Big(\langle \mathbf{x}_{\hat{i}j}(\tau)\rangle - x_{ij}(\tau)\Big)^2}{MN}. \tag{20}$$

First, we note a relation between the ensemble mean $\langle \mathbf{x}_j(\tau)\rangle$ and the ensemble mean with the ith member removed $\langle \mathbf{x}_{\hat{i}j}(\tau)\rangle$. These ensemble means are defined respectively as

$$\langle \mathbf{x}_j(\tau)\rangle = \frac{1}{N}\sum_{k=1}^{N} x_{kj}(\tau), \tag{21}$$

and

$$\langle \mathbf{x}_{\hat{i}j}(\tau)\rangle = \frac{1}{N-1}\sum_{k\neq i}^{N} x_{kj}(\tau), \tag{22}$$

and are related by:

$$\langle \mathbf{x}_j(\tau)\rangle = \frac{1}{N}\sum_{k=1}^{N} x_{kj}(\tau) = \frac{x_{ij}(\tau)}{N} + \frac{1}{N}\sum_{k\neq i}^{N} x_{kj}(\tau) = \frac{x_{ij}(\tau)}{N} + \frac{N-1}{N}\langle \mathbf{x}_{\hat{i}j}(\tau)\rangle. \tag{23}$$

Therefore,

$$\sigma_e^2(\tau) = \frac{\sum_{j=1}^{M}\sum_{i=1}^{N}\Big(\langle \mathbf{x}_j(\tau)\rangle - x_{ij}(\tau)\Big)^2}{M(N-1)} \tag{24}$$

$$= \frac{\sum_{j=1}^{M}\sum_{i=1}^{N}\Big(\frac{1}{N}x_{ij}(\tau) + \frac{N-1}{N}\langle \mathbf{x}_{\hat{i}j}(\tau)\rangle - x_{ij}(\tau)\Big)^2}{M(N-1)} \tag{25}$$

$$= \frac{\sum_{j=1}^{M}\sum_{i=1}^{N}\Big(\frac{N-1}{N}\langle \mathbf{x}_{\hat{i}j}(\tau)\rangle - \frac{N-1}{N}x_{ij}(\tau)\Big)^2}{M(N-1)} \tag{26}$$

$$= \left(\frac{N-1}{N}\right)^2 \frac{\sum_{j=1}^{M}\sum_{i=1}^{N}\Big(\langle \mathbf{x}_{\hat{i}j}(\tau)\rangle - x_{ij}(\tau)\Big)^2}{M(N-1)} \tag{27}$$

$$= \frac{N-1}{N} \frac{\sum_{j=1}^{M} \sum_{i=1}^{N} \left(\langle \mathbf{x}_{\hat{i}j}(\tau) \rangle - x_{ij}(\tau) \right)^2}{MN} \tag{28}$$

$$= \frac{N-1}{N} MSE(\tau). \tag{29}$$

6.2 Relation of perfect model skill metrics to other metrics

6.2.1 PPP

A commonly used PM skill metric is the potential prognostic predictability (PPP, [61]), which compares the ensemble variance, $\sigma_e^2(\tau)$, to the climatological variance, σ_c^2. The PPP is defined as

$$PPP(\tau) = 1 - \frac{\sigma_e^2(\tau)}{\sigma_c^2}, \tag{30}$$

which has a similar form to the MSSS defined in Eqn. 11. Since $MSE = \frac{N}{N-1}\sigma_e^2$, for any finite N, $MSSS < PPP$ and $MSSS \to PPP$ as $N \to \infty$. For most typical values of N, the PPP and MSSS will be quite similar and share the same qualitative interpretations. However, we believe that the $MSSS$ metric provides a more natural comparison with the $MSSS$ metric used in OP predictions. In the PPP formulation, the ensemble mean $\langle x_j \rangle$ is used to predict a given truth member x_{ij}. This implies that the prediction has knowledge of the observed value, since the x_{ij} truth member is included in the ensemble mean computation. This is an undesirable property for a skill metric, and will tend to bias skill scores high. The MSSS does not suffer from this issue, as only non-truth members are used to predict a given truth member.

6.2.2 RMSE

In the PM MSE formula given in Eqn. 8, we have used the $(N-1)$-member ensemble mean to predict a given truth member. In general, we could use an E-member ensemble mean to make this prediction, where $1 \leq E \leq N-1$. It can be shown that an MSE based on an E-member ensemble mean satisfies $MSE = \frac{E+1}{E}\sigma_e^2$, where the proof uses the Central Limit Theorem and follows the same approach as that of [41]. The formula in 6.1 is the special case when $E = N-1$. The PM RMSE definition of [19], uses 1-member ensembles to predict a given truth member, and therefore satisfies $MSE = 2\sigma_e^2$. At long lead times, the PM RMSE of [19] converges to $\sqrt{2}\sigma_c$ (note that this is strictly true only if the normalization of $MN(N-1)-1$ used in [19] is replaced with $MN(N-1)$).

This factor of $\sqrt{2}$ is a potential source of confusion, since in the PM literature a “no skill” forecast has $RMSE = \sqrt{2}\sigma_c$, whereas in the OP literature a “no skill” (climatological) forecast has an RMSE of σ_c. This can lead to confusion when quoting PM RMSE in physical units, or when comparing PM and OP RMSE values (e.g. as done in [7,70]). In particular, the RMSE values obtained via the formula of [19] are too large, since they do not benefit from ensemble averaging. If ensemble means are used for the PM prediction, this issue is greatly ameliorated, since the PM RMSE values converge to $\sqrt{\frac{N}{N-1}}\sigma_c$, allowing for cleaner comparison with OP predictions.

References

1. Anderson, J.L.: An ensemble adjustment Kalman filter for data assimilation. Monthly weather review **129**(12), 2884–2903 (2001)
2. Bitz, C., Holland, M., Weaver, A., Eby, M.: Simulating the ice-thickness distribution in a coupled climate model. J. Geophys. Res.: Oceans **106**(C2), 2441–2463 (2001)
3. Bitz, C., Roe, G.: A mechanism for the high rate of sea ice thinning in the Arctic Ocean. J. Climate **17**(18), 3623–3632 (2004)
4. Blanchard-Wrigglesworth, E., Armour, K.C., Bitz, C.M., DeWeaver, E.: Persistence and inherent predictability of Arctic sea ice in a GCM ensemble and observations. J. Climate **24**, 231–250 (2011)
5. Blanchard-Wrigglesworth, E., Barthélemy, A., Chevallier, M., Cullather, R., Fučkar, N., Massonnet, F., Posey, P., Wang, W., Zhang, J., Ardilouze, C., et al.: Multi-model seasonal forecast of Arctic sea-ice: forecast uncertainty at pan-Arctic and regional scales. Climate Dynamics **49**(4), 1399–1410 (2017)
6. Blanchard-Wrigglesworth, E., Bitz, C., Holland, M.: Influence of initial conditions and climate forcing on predicting Arctic sea ice. Geophys. Res. Lett. **38**(18) (2011)
7. Blanchard-Wrigglesworth, E., Cullather, R., Wang, W., Zhang, J., Bitz, C.: Model forecast skill and sensitivity to initial conditions in the seasonal Sea Ice Outlook. Geophys. Res. Lett **42**(19), 8042–8048 (2015)
8. Bretherton, C.S., Widmann, M., Dymnikov, V.P., Wallace, J.M., Bladé, I.: The effective number of spatial degrees of freedom of a time-varying field. J. Climate **12**(7), 1990–2009 (1999)
9. Bushuk, M., Giannakis, D.: Sea-ice reemergence in a model hierarchy. Geophys. Res. Lett. **42**, 5337–5345 (2015)
10. Bushuk, M., Giannakis, D.: The seasonality and interannual variability of Arctic sea-ice reemergence. J. Climate **30**, 4657–4676 (2017)
11. Bushuk, M., Giannakis, D., Majda, A.J.: Arctic sea-ice reemergence: The role of large-scale oceanic and atmospheric variability. J. Climate **28**, 5477–5509 (2015)
12. Bushuk, M., Msadek, R., Winton, M., Vecchi, G., Gudgel, R., Rosati, A., Yang, X.: Skillful regional prediction of Arctic sea ice on seasonal timescales. Geophys. Res. Lett. **44** (2017)
13. Bushuk, M., Msadek, R., Winton, M., Vecchi, G., Gudgel, R., Rosati, A., Yang, X.: Summer enhancement of Arctic sea-ice volume anomalies in the September-ice zone. J. Climate **30**, 2341–2362 (2017)
14. Cavalieri, D.J., Parkinson, C.L., Gloersen, P., Zwally, H.J.: Sea ice concentrations from Nimbus-7 SMMR and DMSP SSM/I-SSMIS Passive Microwave Data, Version 1. NASA DAAC at the Natl. Snow and Ice Data Cent. (1996). DOI 10.5067/8GQ8LZQVL0VL
15. Chen, Z., Liu, J., Song, M., Yang, Q., Xu, S.: Impacts of assimilating satellite sea ice concentration and thickness on Arctic sea ice prediction in the NCEP Climate Forecast System. Journal of Climate **30**(21), 8429–8446 (2017)
16. Cheng, W., Blanchard-Wrigglesworth, E., Bitz, C.M., Ladd, C., Stabeno, P.J.: Diagnostic sea ice predictability in the pan-Arctic and US Arctic regional seas. Geophys. Res. Lett **43**(22) (2016)
17. Chevallier, M., Salas y Mélia, D.: The role of sea ice thickness distribution in the Arctic sea ice potential predictability: A diagnostic approach with a coupled GCM. J. Climate **25**(8), 3025–3038 (2012)
18. Chevallier, M., Salas y Mélia, D., Voldoire, A., Déqué, M., Garric, G.: Seasonal forecasts of the pan-Arctic sea ice extent using a GCM-based seasonal prediction system. J. Climate **26**(16), 6092–6104 (2013)
19. Collins, M.: Climate predictability on interannual to decadal time scales: The initial value problem. Climate Dyn. **19**, 671–692 (2002)
20. Collow, T.W., Wang, W., Kumar, A., Zhang, J.: Improving Arctic sea ice prediction using PIOMAS initial sea ice thickness in a coupled ocean–atmosphere model. Mon. Wea. Rev. **143**(11), 4618–4630 (2015)
21. Day, J., Tietsche, S., Hawkins, E.: Pan-Arctic and regional sea ice predictability: Initialization month dependence. J. Climate **27**(12), 4371–4390 (2014)
22. Day, J.J., Goessling, H.F., Hurlin, W.J., Keeley, S.P.: The Arctic predictability and prediction on seasonal-to-interannual timescales (APPOSITE) data set version 1. Geoscientific Model Development **9**(6), 2255 (2016)

23. Delworth, T.L., Broccoli, A.J., Rosati, A., Stouffer, R.J., Balaji, V., Beesley, J.A., Cooke, W.F., Dixon, K.W., Dunne, J., Dunne, K., et al.: GFDL's CM2 global coupled climate models. Part I: Formulation and simulation characteristics. J. Climate **19**(5), 643–674 (2006)
24. Delworth, T.L., Rosati, A., Anderson, W., Adcroft, A.J., Balaji, V., Benson, R., Dixon, K., Griffies, S.M., Lee, H.C., Pacanowski, R.C., et al.: Simulated climate and climate change in the GFDL CM2. 5 high-resolution coupled climate model. J. Climate **25**(8), 2755–2781 (2012)
25. Deser, C., Magnusdottir, G., Saravanan, R., Phillips, A.: The effects of North Atlantic SST and sea ice anomalies on the winter circulation in CCM3. Part II: Direct and indirect components of the response. Journal of Climate **17**(5), 877–889 (2004)
26. Deser, C., Walsh, J.E., Timlin, M.S.: Arctic sea ice variability in the context of recent atmospheric circulation trends. J. Climate **13**, 617–633 (2000)
27. Dirkson, A., Merryfield, W.J., Monahan, A.: Impacts of sea ice thickness initialization on seasonal Arctic sea ice predictions. Journal of Climate **30**(3), 1001–1017 (2017)
28. Drobot, S.D.: Using remote sensing data to develop seasonal outlooks for Arctic regional sea-ice minimum extent. Remote Sensing of Environment **111**(2-3), 136–147 (2007)
29. Drobot, S.D., Maslanik, J.A., Fowler, C.: A long-range forecast of Arctic summer sea-ice minimum extent. Geophysical Research Letters **33**(10) (2006)
30. Germe, A., Chevallier, M., y Mélia, D.S., Sanchez-Gomez, E., Cassou, C.: Interannual predictability of Arctic sea ice in a global climate model: Regional contrasts and temporal evolution. Climate Dynamics **43**(9-10), 2519–2538 (2014)
31. Griffies, S.: Elements of the Modular Ocean Model (MOM), GFDL Ocean Group Technical Report. Tech. Rep. No. 7, NOAA/Geophysical Fluid Dynamics Laboratory (2012)
32. Griffies, S.M., Winton, M., Donner, L.J., Horowitz, L.W., Downes, S.M., Farneti, R., Gnanadesikan, A., Hurlin, W.J., Lee, H.C., Liang, Z., et al.: The GFDL CM3 coupled climate model: characteristics of the ocean and sea ice simulations. Journal of Climate **24**(13), 3520–3544 (2011)
33. Guemas, V., Chevallier, M., Dqu, M., Bellprat, O., Doblas-Reyes, F.: Impact of sea ice initialisation on sea ice and atmosphere prediction skill on seasonal timescales. Geophys. Res. Lett **43**(8), 3889–3896 (2016)
34. Hawkins, E., Tietsche, S., Day, J.J., Melia, N., Haines, K., Keeley, S.: Aspects of designing and evaluating seasonal-to-interannual Arctic sea-ice prediction systems. Quarterly Journal of the Royal Meteorological Society **142**(695), 672–683 (2016)
35. Holland, M.M., Bailey, D.A., Vavrus, S.: Inherent sea ice predictability in the rapidly changing Arctic environment of the Community Climate System Model, version 3. Climate Dynamics **36**(7-8), 1239–1253 (2011)
36. Holland, M.M., Stroeve, J.: Changing seasonal sea ice predictor relationships in a changing arctic climate. Geophys. Res. Lett **38**(18) (2011)
37. Hunke, E., Dukowicz, J.: An elastic-viscous-plastic model for sea ice dynamics. J. Phys. Oceanogr. **27**(9), 1849–1867 (1997)
38. Jia, L., Yang, X., Vecchi, G., Gudgel, R., Delworth, T., Fueglistaler, S., Lin, P., Scaife, A.A., Underwood, S., Lin, S.J.: Seasonal prediction skill of northern extratropical surface temperature driven by the stratosphere. Journal of Climate **30**(1), 4463–4475 (2017)
39. Jia, L., Yang, X., Vecchi, G.A., Gudgel, R.G., Delworth, T.L., Rosati, A., Stern, W.F., Wittenberg, A.T., Krishnamurthy, L., Zhang, S., et al.: Improved seasonal prediction of temperature and precipitation over land in a high-resolution GFDL climate model. J. Climate **28**(5), 2044–2062 (2015)
40. Johnson, C., Bowler, N.: On the reliability and calibration of ensemble forecasts. Monthly Weather Review **137**(5), 1717–1720 (2009)
41. Jolliffe, I.T., Stephenson, D.B.: Forecast verification: a practitioner's guide in atmospheric science–2nd Ed. John Wiley & Sons (2012)
42. Jung, T., Gordon, N.D., Bauer, P., Bromwich, D.H., Chevallier, M., Day, J.J., Dawson, J., Doblas-Reyes, F., Fairall, C., Goessling, H.F., et al.: Advancing polar prediction capabilities on daily to seasonal time scales. Bull. Amer. Meteor. Soc. (2016). DOI 10.1175/BAMS-D-14-00246.1
43. Kapsch, M.L., Graversen, R.G., Economou, T., Tjernström, M.: The importance of spring atmospheric conditions for predictions of the Arctic summer sea ice extent. Geophysical Research Letters **41**(14), 5288–5296 (2014)
44. Kauker, F., Kaminski, T., Karcher, M., Giering, R., Gerdes, R., Voßbeck, M.: Adjoint analysis of the 2007 all time Arctic sea-ice minimum. Geophysical Research Letters **36**(3) (2009)

45. Koenigk, T., Mikolajewicz, U.: Seasonal to interannual climate predictability in mid and high northern latitudes in a global coupled model. Climate dynamics **32**(6), 783–798 (2009)
46. Krikken, F., Schmeits, M., Vlot, W., Guemas, V., Hazeleger, W.: Skill improvement of dynamical seasonal Arctic sea ice forecasts. Geophys. Res. Lett (2016)
47. Kumar, A., Peng, P., Chen, M.: Is there a relationship between potential and actual skill? Monthly Weather Review **142**(6), 2220–2227 (2014)
48. Leutbecher, M., Palmer, T.N.: Ensemble forecasting. Journal of Computational Physics **227**(7), 3515–3539 (2008)
49. Lin, S.J.: A vertically Lagrangian finite-volume dynamical core for global models. Monthly Weather Review **132**(10), 2293–2307 (2004)
50. Lindsay, R., Zhang, J., Schweiger, A., Steele, M.: Seasonal predictions of ice extent in the Arctic Ocean. Journal of Geophysical Research: Oceans **113**(C2) (2008)
51. Martinson, D.G.: Evolution of the Southern Ocean winter mixed layer and sea ice: Open ocean deepwater formation and ventilation. Journal of Geophysical Research: Oceans **95**(C7), 11,641–11,654 (1990)
52. Merryfield, W., Lee, W.S., Wang, W., Chen, M., Kumar, A.: Multi-system seasonal predictions of Arctic sea ice. Geophys. Res. Lett **40**(8), 1551–1556 (2013)
53. Milly, P.C., Malyshev, S.L., Shevliakova, E., Dunne, K.A., Findell, K.L., Gleeson, T., Liang, Z., Phillipps, P., Stouffer, R.J., Swenson, S.: An enhanced model of land water and energy for global hydrologic and earth-system studies. Journal of Hydrometeorology **15**(5), 1739–1761 (2014)
54. Msadek, R., Vecchi, G., Winton, M., Gudgel, R.: Importance of initial conditions in seasonal predictions of Arctic sea ice extent. Geophys. Res. Lett. **41**(14), 5208–5215 (2014)
55. Murakami, H., Vecchi, G.A., Delworth, T.L., Wittenberg, A.T., Underwood, S., Gudgel, R., Yang, X., Jia, L., Zeng, F., Paffendorf, K., et al.: Dominant role of subtropical pacific warming in extreme Eastern Pacific hurricane seasons: 2015 and the future. Journal of Climate **30**(1), 243–264 (2017)
56. Murphy, A.H.: Skill scores based on the mean square error and their relationships to the correlation coefficient. Monthly weather review **116**(12), 2417–2424 (1988)
57. Owens, W.B., Lemke, P.: Sensitivity studies with a sea ice-mixed layer-pycnocline model in the Weddell sea. J. Geophys. Res: Oceans (1978–2012) **95**(C6), 9527–9538 (1990)
58. Palmer, T., Buizza, R., Hagedorn, R., Lawrence, A., Leutbecher, M., Smith, L.: Ensemble prediction: a pedagogical perspective. ECMWF newsletter **106**, 10–17 (2006)
59. Peterson, K.A., Arribas, A., Hewitt, H., Keen, A., Lea, D., McLaren, A.: Assessing the forecast skill of Arctic sea ice extent in the GloSea4 seasonal prediction system. Climate Dynamics **44**(1-2), 147–162 (2015)
60. Petty, A.A., Schröder, D., Stroeve, J., Markus, T., Miller, J., Kurtz, N., Feltham, D., Flocco, D.: Skillful spring forecasts of September Arctic sea ice extent using passive microwave sea ice observations. Earth's Future **5**(2), 254–263 (2017)
61. Pohlmann, H., Botzet, M., Latif, M., Roesch, A., Wild, M., Tschuck, P.: Estimating the decadal predictability of a coupled AOGCM. Journal of Climate **17**(22), 4463–4472 (2004)
62. Putman, W.M., Lin, S.J.: Finite-volume transport on various cubed-sphere grids. Journal of Computational Physics **227**(1), 55–78 (2007)
63. Schröder, D., Feltham, D.L., Flocco, D., Tsamados, M.: September Arctic sea-ice minimum predicted by spring melt-pond fraction. Nature Climate Change (2014)
64. Schweiger, A., Lindsay, R., Zhang, J., Steele, M., Stern, H., Kwok, R.: Uncertainty in modeled Arctic sea ice volume. J. Geophys. Res.: Oceans **116**(C8) (2011)
65. Sigmond, M., Fyfe, J., Flato, G., Kharin, V., Merryfield, W.: Seasonal forecast skill of Arctic sea ice area in a dynamical forecast system. Geophys. Res. Lett. **40**(3), 529–534 (2013)
66. Sigmond, M., Reader, M., Flato, G., Merryfield, W., Tivy, A.: Skillful seasonal forecasts of Arctic sea ice retreat and advance dates in a dynamical forecast system. Geophys. Res. Lett. **43** (2016)
67. Stock, C.A., Pegion, K., Vecchi, G.A., Alexander, M.A., Tommasi, D., Bond, N.A., Fratantoni, P.S., Gudgel, R.G., Kristiansen, T., OBrien, T.D., et al.: Seasonal sea surface temperature anomaly prediction for coastal ecosystems. Progress in Oceanography **137**, 219–236 (2015)
68. Stroeve, J., Hamilton, L.C., Bitz, C.M., Blanchard-Wrigglesworth, E.: Predicting September sea ice: Ensemble skill of the SEARCH sea ice outlook 2008–2013. Geophys. Res. Lett. **41**(7), 2411–2418 (2014)

69. Sun, L., Deser, C., Tomas, R.A.: Mechanisms of stratospheric and tropospheric circulation response to projected Arctic sea ice loss. Journal of Climate **28**(19), 7824–7845 (2015)
70. Tietsche, S., Day, J., Guemas, V., Hurlin, W., Keeley, S., Matei, D., Msadek, R., Collins, M., Hawkins, E.: Seasonal to interannual Arctic sea ice predictability in current global climate models. Geophys. Res. Lett. **41**(3), 1035–1043 (2014)
71. Tivy, A., Howell, S.E., Alt, B., Yackel, J.J., Carrieres, T.: Origins and levels of seasonal forecast skill for sea ice in Hudson Bay using Canonical Correlation Analysis. Journal of Climate **24**(5), 1378–1395 (2011)
72. Vecchi, G.A., Delworth, T., Gudgel, R., Kapnick, S., Rosati, A., Wittenberg, A.T., Zeng, F., Anderson, W., Balaji, V., Dixon, K., et al.: On the seasonal forecasting of regional tropical cyclone activity. J. Climate **27**(21), 7994–8016 (2014)
73. Wang, L., Ting, M., Kushner, P.: A robust empirical seasonal prediction of winter NAO and surface climate. Scientific Reports **7**(1), 279 (2017)
74. Wang, L., Yuan, X., Ting, M., Li, C.: Predicting summer Arctic sea ice concentration intraseasonal variability using a vector autoregressive model*. J. Climate **29**(4), 1529–1543 (2016)
75. Wang, W., Chen, M., Kumar, A.: Seasonal prediction of Arctic sea ice extent from a coupled dynamical forecast system. Mon. Wea. Rev. **141**(4), 1375–1394 (2013)
76. Weigel, A.P., Liniger, M.A., Appenzeller, C.: Seasonal ensemble forecasts: Are recalibrated single models better than multimodels? Monthly Weather Review **137**(4), 1460–1479 (2009)
77. Williams, J., Tremblay, B., Newton, R., Allard, R.: Dynamic preconditioning of the minimum September sea-ice extent. J. Climate **29**(16), 5879–5891 (2016)
78. Winton, M.: A reformulated three-layer sea ice model. J. Atmos. Oceanic Technol. **17**(4), 525–531 (2000)
79. Yang, X., Vecchi, G.A., Gudgel, R.G., Delworth, T.L., Zhang, S., Rosati, A., Jia, L., Stern, W.F., Wittenberg, A.T., Kapnick, S., et al.: Seasonal predictability of extratropical storm tracks in GFDLs high-resolution climate prediction model. Journal of Climate **28**(9), 3592–3611 (2015)
80. Yeager, S.G., Karspeck, A.R., Danabasoglu, G.: Predicted slowdown in the rate of Atlantic sea ice loss. Geophysical Research Letters **42**(24) (2015)
81. Yuan, X., Chen, D., Li, C., Wang, L., Wang, W.: Arctic sea ice seasonal prediction by a linear markov model. J. Climate **29**(22), 8151–8173 (2016)
82. Zhang, J., Rothrock, D.: Modeling global sea ice with a thickness and enthalpy distribution model in generalized curvilinear coordinates. Mon. Wea. Rev. **131**(5), 845–861 (2003)
83. Zhang, S., Harrison, M., Rosati, A., Wittenberg, A.: System design and evaluation of coupled ensemble data assimilation for global oceanic climate studies. Mon. Wea. Rev. **135**(10), 3541–3564 (2007)
84. Zhang, S., Rosati, A.: An inflated ensemble filter for ocean data assimilation with a biased coupled GCM. Mon. Wea. Rev. **138**(10), 3905–3931 (2010)