Predicting September Arctic Sea Ice: A Multi-Model Seasonal Skill Comparison

Mitchell Bushuk,[a] Sahara Ali,[b] David A. Bailey,[c] Qing Bao,[d] Lauriane Batté,[e] Uma S. Bhatt,[ub] Edward Blanchard-Wrigglesworth,[uw] Ed Blockley,[mo] Gavin Cawley,[uea] Junhwa Chi,[pknu] François Counillon[no] , Philippe Goulet Coulombe,[uq] Richard I. Cullather,[gmao,umd] Francis X. Diebold,[up] Arlan Dirkson,[eccc1] Eleftheria Exarchou,[bsc] Maximilian Göbel,[bu] William Gregory,[pu] Virginie Guemas,[cnrm] Lawrence Hamilton,[unh] Bian He,[d] Sean Horvath,[nsidc] Monica Ionita,[nos,mos] Jennifer E. Kay,[cu] Eliot Kim,[gmao] Noriaki Kimura,[aori] Dmitri Kondrashov,[ucla] Zachary M. Labe,[pu] WooSung Lee,[cccma] Younjoo J. Lee,[nps] Cuihua Li,[la] Xuewei Li,[dlut,sysu] Yongcheng Lin,[sysu] Yanyun Liu,[cpc,ert] Wieslaw Maslowski,[nps] François Massonnet,[uclv] Walter N. Meier,[nsidc] William J. Merryfield,[cccma] Hannah Myint,[uclahm] Juan C. Acosta Navarro,[ec] Alek Petty,[umd] Fangli Qiao,[fio] David Schröder,[cpom] Axel Schweiger,[apluw] Qi Shu,[fio] Michael Sigmond,[cccma] Michael Steele,[apluw] Julienne Stroeve,[um] Nico Sun,[cryo] Steffen Tietsche,[ecmwf] Michel Tsamados,[ucl] Keguang Wang,[nmi] Jianwu Wang,[b] Wanqiu Wang,[cpc] Yiguo Wang,[no] Yun Wang,[sysu] James Williams,[giss] Qinghua Yang,[sysu] Xiaojun Yuan,[la] Jinlun Zhang,[apluw] and Yongfei Zhang[pu]

[a] *National Oceanic and Atmospheric Administration/Geophysical Fluid Dynamics Laboratory, Princeton, New Jersey, USA*

[b] *Department of Information Systems, University of Maryland Baltimore County, Maryland, USA*

[c] *National Center for Atmospheric Research, Boulder, Colorado, USA*

[d] *State Key Laboratory of Numerical Modeling for Atmospheric Sciences and Geophysical Fluid Dynamics (LASG), Institute of Atmospheric Physics (IAP), Chinese Academy of Sciences, Beijing, 100029, China*

[e] *Direction de la Climatologie et des Services Climatiques, Météo-France, Toulouse, France*

[ub] *Geophysical Institute, Dept. of Atmospheric Sciences, University of Alaska Fairbanks*

[uw] *Department of Atmospheric Sciences, University of Washington, Seattle, USA*

Early Online Release: This preliminary version has been accepted for publication in *Bulletin of the American Meteorological Society*, may be fully cited, and has been assigned DOI 10.1175/BAMS-D-23-0163.1. The final typeset copyedited article will replace the EOR at the above DOI when it is published.

[mo] *Met Office, FitzRoy Road, Exeter, United Kingdom*

[uea] *School of Computing Sciences, University of East Anglia, Norwich, United Kingdom*

[pknu] *Division of Data Information Sciences, Pukyong National University, Busan, Korea*

[no] *Nansen Environmental and Remote Sensing Center, Bjerknes Centre for Climate Research, Bergen, Norway*

[uq] *Departement des Sciences Économiques, Université du Québec à Montréal, Montréal, Canada*

[gmao] *Global Modeling and Assimilation Office, NASA Goddard Space Flight Center, Greenbelt, MD, USA*

[aori] *Atmosphere and Ocean Research Institute, The University of Tokyo, Kashiwa, Japan*

[up] *Departments of Economics, Finance and Statistics, University of Pennsylvania, Philadelphia, USA*

[eccc1] *Meteorological Research Division, Environment and Climate Change Canada, Dorval, QC, Canada*

[bsc] *Barcelona Supercomputing Centre, Barcelona, Spain*

[bu] *Bocconi University, Milan, Italy*

[pu] *Atmospheric and Oceanic Sciences Program, Princeton University, Princeton, New Jersey, USA*

[cnrm] *CNRM, Météo-France, CNRS, Université de Toulouse, Toulouse, France*

[unh] *Department of Sociology, University of New Hampshire, Durham, New Hampshire, USA*

[nos] *Alfred Wegener Institute Helmholtz Center for Polar and Marine Research, Bremerhaven, Germany*

[mos] *Faculty of Forestry, "Stefan cel Mare" University of Suceava, Suceava, Romania*

[cu] *Department of Atmospheric and Oceanic Sciences and Cooperative Institute for Research in Environmental Sciences, University of Colorado Boulder, Boulder, Colorado, USA*

[ucla] *Department of Atmospheric and Oceanic Sciences,University of California, Los Angeles, USA*

[cccma] *Canadian Centre for Climate Modelling and Analysis, Environment and Climate Change Canada, Victoria, BC, Canada*

[nps] *Department of Oceanography, Naval Postgraduate School, Monterey, California, USA*

[la] *Lamont-Doherty Earth Observatory, Columbia University, Palisades, New York, USA*

[dlut] *State Key Laboratory of Coastal and Offshore Engineering, Dalian University of Technology, Dalian 116023, China*

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0163.1.

[sysu] *School of Atmospheric Sciences, Sun Yat-sen University, and Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai), Zhuhai 519082, China*

[cpc] *NOAA/NWS/NCEP/Climate Prediction Center, College Park, Maryland, USA*

[ert] *Earth Resources Technology Inc, Laurel, Maryland, USA*

[uclv] *Earth and Climate Centre, Earth and Life Institute, Université catholique de Louvain, Louvain-la-Neuve, Belgium*

[nsidc] *National Snow and Ice Data Center, Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, Colorado, USA*

[uclahm] *Institute of the Environment and Sustainability, University of California, Los Angeles, Los Angeles, California, USA*

[ec] *European Commission, Joint Research Centre, Ispra, Italy*

[umd] *Earth System Science Interdisciplinary Center, University of Maryland, Maryland, USA*

[fio] *First Institute of Oceanography (FIO), Ministry of Natural Resources, Qingdao, China*

[cpom] *CPOM, Dep. of Meteorology, University of Reading, U.K.*

[apluw] *Applied Physics Laboratory/Polar Science Center, University of Washington, Seattle, USA*

[um] *University of Manitoba, CEOS, Winnipeg, Manitoba, Canada*

[aori] *Atmosphere and Ocean Research Institute, The University of Tokyo, Kashiwa, Japan*

[cryo] *Citizen Scientist, CryosphereComputing.com*

[ecmwf] *European Centre for Medium-Range Weather Forecasts, Bonn, Germany*

[ucl] *Centre for Polar Observation and Modelling, Earth Sciences, University College London, London, UK*

[nmi] *Division of Ocean and Ice, Norwegian Meteorological Institute, Oslo, Norway*

[giss] *NASA Goddard Institute of Space Studies, New York, New York, USA*

Corresponding author: Mitch Bushuk, mitchell.bushuk@noaa.gov

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0163.1.

ABSTRACT: This study quantifies the state-of-the-art in the rapidly growing field of seasonal Arctic sea ice prediction. A novel multi-model dataset of retrospective seasonal predictions of September Arctic sea ice is created and analyzed, consisting of community contributions from 17 statistical models and 17 dynamical models. Prediction skill is compared over the period 2001–2020 for predictions of Pan-Arctic sea ice extent (SIE), regional SIE, and local sea ice concentration (SIC) initialized on June 1, July 1, August 1, and September 1. This diverse set of statistical and dynamical models can individually predict linearly detrended Pan-Arctic SIE anomalies with skill, and a multi-model median prediction has correlation coefficients of 0.79, 0.86, 0.92, and 0.99 at these respective initialization times. Regional SIE predictions have similar skill to Pan-Arctic predictions in the Alaskan and Siberian regions, whereas regional skill is lower in the Canadian, Atlantic, and Central Arctic sectors. The skill of dynamical and statistical models is generally comparable for Pan-Arctic SIE, whereas dynamical models outperform their statistical counterparts for regional and local predictions. The prediction systems are found to provide the most value added relative to basic reference forecasts in the extreme SIE years of 1996, 2007, and 2012. SIE prediction errors do not show clear trends over time, suggesting that there has been minimal change in inherent sea ice predictability over the satellite era. Overall, this study demonstrates that there are bright prospects for skillful operational predictions of September sea ice at least three months in advance.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0163.1.

SIGNIFICANCE STATEMENT: The observed decline of Arctic sea ice extent has created an emerging need for predictions of sea ice on seasonal timescales. This study provides a comparison of September Arctic sea ice seasonal prediction skill across a diverse set of dynamical and statistical prediction models, quantifying the state-of-the-art in the rapidly growing sea ice prediction research community. We find that both dynamical and statistical models can skillfully predict September Arctic sea ice 0–3 months in advance on Pan-Arctic, regional, and local spatial scales. Our results demonstrate that there are bright prospects for skillful operational seasonal predictions of Arctic sea ice and highlight a number of crucial prediction system design aspects to guide future improvements.

CAPSULE: September Arctic sea ice seasonal prediction skill is compared across a diverse set of dynamical and statistical prediction models, quantifying the state-of-the-art in the rapidly growing sea ice prediction community.

1. Introduction

The rapid decline of summer Arctic sea ice over the satellite era (Fig. 1) has led to increased socioeconomic activity in the region and an emerging need for skillful predictions of sea ice conditions (Jung et al. 2016; Wagner et al. 2020). Following the then-record-setting 2007 September Arctic sea ice extent (SIE) minimum, a new research subfield emerged focused on scientific understanding of sea ice predictability and prediction. At the core of this research community has been the Sea Ice Outlook (SIO), which collects, analyzes, and synthesizes real-time seasonal predictions of September Pan-Arctic SIE (Stroeve et al. (2014); see arcus.org/sipn/sea-ice-outlook). From 2008–present, the SIO has collected predictions of September SIE initialized on June 1, July 1, and August 1, months that span the summer Arctic melt season. The SIO began additionally collecting September 1 initialized predictions in 2021. The number of annual SIO submissions has grown steadily over time, with approximately 40 groups submitting predictions in recent years. These submissions are provided by an international community of polar scientists and employ a diverse mix of dynamical modeling, statistical, and heuristic approaches.

In parallel to the growth of the SIO, a body of work on sea ice predictability has been developed, which underpins the expectation that sea ice could be predictable on seasonal timescales. Coupled global climate models (GCMs) have been used to estimate the upper limits of sea ice predictability

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0163.1.

based on "perfect model" ensemble experiments, which quantify potential prediction skill in the case of perfectly known initial conditions, forcing, and model physics. These studies have shown that, with typical sample sizes, Arctic SIE potential predictability is statistically significant up to 12-36 months in advance (Koenigk and Mikolajewicz 2009; Blanchard-Wrigglesworth et al. 2011b; Holland et al. 2011; Tietsche et al. 2014; Day et al. 2014; Bushuk et al. 2019; Holland et al. 2019), however they may overestimate nature's true predictability limits due to the overly persistent SIE anomalies present in most modern GCMs (Blanchard-Wrigglesworth and Bushuk 2019; Giesse et al. 2021). The inherent predictability of Arctic sea ice is determined by a competition between the slowly evolving predictable components of the ice-ocean-land system and the comparatively unpredictable variability of the atmosphere (Tietsche et al. 2016). A number of physical mechanisms for summer Arctic SIE predictability have been demonstrated. These include the persistence and reemergence of SIE and sea ice concentration (SIC) anomalies (Blanchard-Wrigglesworth et al. 2011a; Bushuk and Giannakis 2015; Ordoñez et al. 2018; Giesse et al. 2021; Zhang et al. 2021), the persistence and advection of sea ice thickness (SIT) anomalies (Holland et al. 2011; Blanchard-Wrigglesworth et al. 2011b; Chevallier and Salas y Mélia 2012; Krumpen et al. 2013; Blanchard-Wrigglesworth and Bitz 2014; Day et al. 2014; Collow et al. 2015; Massonnet et al. 2015; Guemas et al. 2016; Williams et al. 2016; Blanchard-Wrigglesworth et al. 2017; Bushuk et al. 2017b; Dirkson et al. 2017; Blockley and Peterson 2018; Holland et al. 2019; Bonan et al. 2019; Brunette et al. 2019; Babb et al. 2019; Ponsoni et al. 2020; Babb et al. 2020; Balan-Sarojini et al. 2021), ocean heat transport and persistence of upper ocean heat content anomalies (Serreze et al. 2016; Lenetsky et al. 2021; Bushuk et al. 2022), melt onset and summer ice-albedo feedback processes (Schröder et al. 2014; Kapsch et al. 2014; Liu et al. 2015; Landy et al. 2015; Cox et al. 2016; Zhan and Davies 2017; Kwok et al. 2018; Bushuk et al. 2020), and summertime atmospheric circulation patterns (Ding et al. 2017, 2019; Baxter et al. 2019; Baxter and Ding 2022). Taken together, these studies have laid critical groundwork, showing that sea ice should be potentially predictable on seasonal timescales.

Have modern prediction systems capitalized upon this potential predictability and produced skillful predictions of observed Arctic sea ice? There is a tension in the sea ice prediction literature regarding this question. On one hand, a number of studies have evaluated the performance of September SIE predictions submitted in real-time to the SIO and found that these predictions

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0163.1.

have only a modest skill advantage relative to a baseline linear trend prediction (Stroeve et al. 2014; Blanchard-Wrigglesworth et al. 2015; Hamilton and Stroeve 2016; Lukovich et al. 2021; Blanchard-Wrigglesworth et al. 2023). The initial assessment performed by Stroeve et al. (2014) on SIO predictions submitted over the period of 2008–2013 found that, regardless of method, predictions struggled to capture years with large SIE anomalies relative to the linear trend. These initial findings have been largely corroborated over the longer assessment periods of 2008–2015 and 2008–2022 considered by Hamilton and Stroeve (2016) and Blanchard-Wrigglesworth et al. (2023), respectively. Blanchard-Wrigglesworth et al. (2023) found that the SIO multi-model median prediction has similar skill to a damped anomaly persistence forecast from July 1 and August 1 initialization dates, and is slightly more skillful than damped persistence from June 1. They found that the skill of individual models was lower than the multi-model median skill and had worse skill than damped persistence.

On the other hand, there has been a recent proliferation of studies that document the development of seasonal prediction systems capable of skillfully predicting detrended September Arctic SIE anomalies. These skill assessments are based on retrospective seasonal predictions (also known as hindcasts or reforecasts), which use a fixed initialization and modeling formulation to make seasonal predictions of past observations using only data that would have been available at the time of initialization. Many dynamical prediction systems, which are based on initialized coupled dynamical models, have recently shown skillful seasonal predictions of detrended September Arctic SIE anomalies (Wang et al. 2013; Sigmond et al. 2013; Chevallier et al. 2013; Merryfield et al. 2013; Msadek et al. 2014; Peterson et al. 2015; Collow et al. 2015; Sigmond et al. 2016; Guemas et al. 2016; Bushuk et al. 2017a; Dirkson et al. 2017, 2019; Kimmritz et al. 2019; Harnos et al. 2019; Batté et al. 2020; Shu et al. 2021; Bushuk et al. 2022; Zhang et al. 2022; Martin et al. 2023). Simultaneously, many statistical prediction systems, which leverage empirical relationships in past observational data, have also demonstrated skillful detrended SIE predictions (Drobot et al. 2006; Lindsay et al. 2008; Schröder et al. 2014; Kapsch et al. 2014; Yuan et al. 2016; Williams et al. 2016; Serreze et al. 2016; Petty et al. 2017; Kondrashov et al. 2018; Brunette et al. 2019; Ionita et al. 2019; Walsh et al. 2019; Gregory et al. 2020; Andersson et al. 2021; Chi et al. 2021; Horvath et al. 2021). Both dynamical and statistical predictions (see subsection 2b ahead) have been shown to

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0163.1.

outperform the damped persistence forecast in most cases. This discrepancy between retrospective and real-time prediction skill represents a key tension in the sea ice prediction literature.

While many dynamical and statistical prediction systems have documented "skillful" SIE predictions, it is arguably more important to consider the quantitative level of skill and whether such predictions could provide value to end users (Murphy 1993). The sea ice prediction community gathered for a Sea Ice Outlook Contributors Forum in 2021 where this and many other issues were discussed (Steele et al. 2021). Many workshop attendees expressed a need to rigorously quantify the current state-of-the-art across modern sea ice prediction systems. Unfortunately, this quantitative skill comparison is challenging due to differences in the evaluation time period and skill metrics considered across different studies and the relatively short period of real-time SIO predictions. This knowledge gap led to a key outcome of the SIO Forum–the expressed need for an "apples-to-apples" skill comparison of modern dynamical and statistical sea ice prediction systems. This community intercomparison of sea ice prediction skill forms the basis of the present study.

The outline for this paper is as follows. In section 2, we describe a retrospective prediction data request that was sent to the SIO contributor community, summarize the prediction methodologies used by the 35 groups who contributed predictions, and outline our methods for assessing prediction skill against multiple observational products. In section 3, we assess Pan-Arctic September SIE prediction skill across dynamical and statistical models and consider whether SIE prediction skill has changed over time. In section 4, we consider smaller spatial scales, evaluating regional SIE prediction skill in five Arctic regions and comparing Pan-Arctic and regional performance. Finally, we assess prediction skill for local SIC and ice-edge predictions in section 5. We discuss our findings in section 6, focusing on the key elements of successful sea ice prediction systems and the skill differences between retrospective and real-time predictions. Conclusions and a future outlook are presented in section 7.

2. Methods

a. Retrospective Prediction Data Request

In order to facilitate a direct "apples-to-apples" skill comparison of SIO models, a data request for retrospective predictions of September Arctic sea ice was sent to the SIO contributor community in

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0163.1.

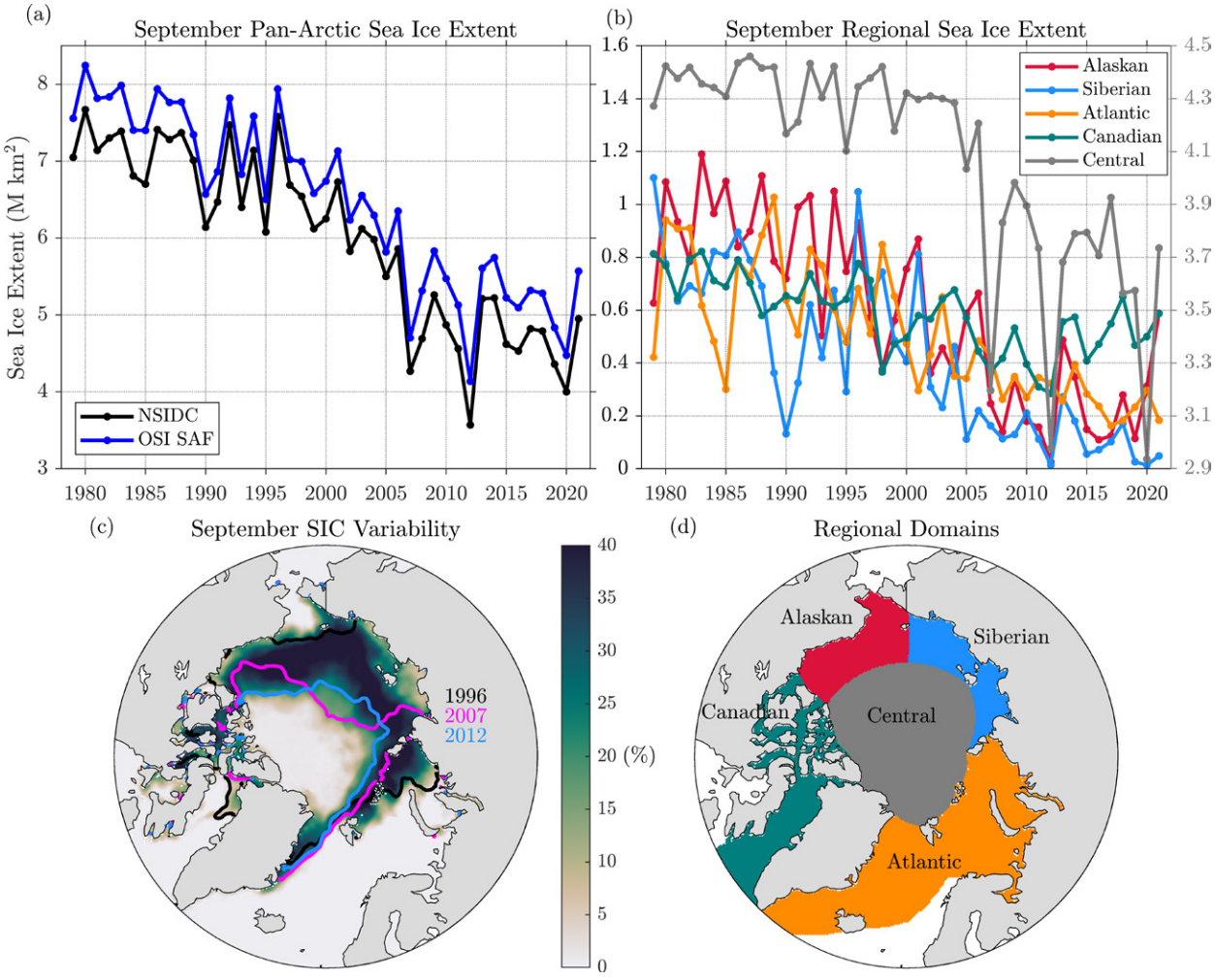

FIG. 1. Observations of (a) September Pan-Arctic SIE from NSIDC v3 and OSI SAF v2.1 Sea Ice Indices; (b) September regional SIE from NSIDC CDR SIC (G02202); and (c) September-mean SIC interannual standard deviation and sea ice edge positions in the extreme years of 1996 (black), 2007 (magenta), and 2012 (blue). Panel (d) shows the regional domain definitions for the Alaskan, Siberian, Atlantic, Canadian, and Central Arctic regions. Note that the Central Arctic time series in panel (b) is plotted using a shifted y-axis on the right (values in gray).

early 2022. The data request was for retrospective predictions initialized on the SIO initialization dates of June 1, July 1, August 1, and September 1, and spanning a minimum period of 2001–2020. The requested target variables were September monthly-mean Pan-Arctic SIE, regional SIE, and gridded SIC fields. Pan-Arctic SIE is defined as the area of all Northern Hemisphere grid cells covered by at least 15% SIC. We define monthly-mean SIE following the NSIDC Sea Ice

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0163.1.

Index convention, which defines the monthly-mean extent as the monthly mean of the daily SIE values. Regional SIE was requested for four regional domains: the Alaskan Seas (Chukchi and Beaufort), Siberian Seas (East Siberian and Laptev), Atlantic Seas (Kara, Barents, and Greenland), and Canadian Seas (Canadian Archipelago and Baffin Bay). These regions were defined based on a recently updated NSIDC region mask, which has better agreement with the regional definitions used by the International Hydrographic Organization (Meier and Stewart (2023); see map in Fig. 1d). We also later derived Central Arctic regional SIE using the submitted Pan-Arctic and regional SIE values by taking their difference. SIO contributors were invited to submit any combination of the requested target variables and initialization dates along with metadata describing the design of their prediction system. We also requested submission of individual ensemble members, if applicable, and the initial SIC and SIT conditions used for dynamical predictions. Contributors were informed that the NSIDC sea ice index and SIC climate data record would be the official verification products, but we also utilize OSI SAF observations for verification in this study (see subsection 2c ahead). For groups that only provided SIC predictions, Pan-Arctic and regional SIE were computed on the native model grid and a post-processing was applied to remove biases (see ahead), including those related to land-sea mask differences.

Retrospective prediction contributions were received from 17 statistical models, 17 dynamical models, and 1 heuristic prediction (see summary of submitted data in Table 1). These contributions span 11 countries across Europe, Asia, and North America, and provide a total of 2807 individual predictions of September Pan-Arctic SIE (1267 statistical; 1526 dynamical; 14 heuristic). All data have been subsequently formatted into a common format and made publicly available via an online repository (https://zenodo.org/doi/10.5281/zenodo.10124346). The online repository also contains scripts for processing the raw data, computing skill metrics, and producing all figures for this study. This is the most comprehensive dataset of multi-model Arctic sea ice predictions that has been assembled to date, and is intended to provide an open community resource for future sea ice prediction research. In this study, we will focus on ensemble-mean sea ice predictions in order to compare ensemble and deterministic contributions, since this is the primary focus of the SIO and allows for the largest set of models to be compared.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0163.1.

TABLE 1. Summary of submitted retrospective prediction data. Target variables are Pan-Arctic SIE (P), Regional SIE (R), SIC (S), and the number of ensemble members (e) is indicated in parentheses. The variables that are bias corrected are shown in parentheses in the Bias Correction column.

Name	Forecast Method	Time period	Initialization Dates	Target Variables	Bias Correction
AWI	Statistical	2000–2021	JJA	P	No
BDAL	Statistical	2001–2021	JJAS	P	No
Cawley	Statistical	2001–2021	June	PR	No
CPOM	Statistical	1984–2021	JJ	P	No
CPOM-UCL	Statistical	1993–2020	JJAS	PR	No
CSU	Statistical	2011–2021	A	P	No
GSFC_Petty	Statistical	1990–2021	JJAS	PR	No
Damped Persistence	Statistical	1990–2021	JJAS	PRS	No
Horvath	Statistical	2001–2020	JJAS	PRS	Yes (S)
KOPRI	Statistical	2001–2021	JJAS	PRS	No
Lamont	Statistical	2013–2021	JJAS	PRS	Yes (PR)
MetNo-sparse-st	Statistical	2000–2020	JJAS	P	No
Nico Sun	Statistical	2000–2021	JJAS	PRSe(3)	No
SYSU/SML-KNN	Statistical	2000–2020	JJAS	PRS	No
SYSU/SML-MLM	Statistical	1980–2020	JJAS	PRS	No
Trend Climatology	Statistical	1990–2021	JJAS	PRS	No
UCLA	Statistical	2012–2021	JJAS	PR	No
UMBC-REU	Statistical	2000–2020	S	PS	No
UPenn	Statistical	2000–2021	JJAS	P	No
BCCR	Dynamical	2003–2021	A	PRSe(10)	Yes (PR)
CNRM	Dynamical	1993–2016	JJAS	PRSe(25)	Yes (PR)
CPC CFSm5	Dynamical	2006–2021	JJAS	PRSe(4)	Yes (S)
CPC CFSv2	Dynamical	1991–2021	JJAS	PRSe(4)	Yes (S)
ECCC-CanSIPSv2	Dynamical	1990–2021	JJAS	PRSe(20)	Yes (PRS)
EC-Earth	Dynamical	1981–2014	June	PRSe(10)	No
ECMWF SEAS5	Dynamical	1993–2021	JJAS	PRSe(25)	Yes (PR)
FGOALS-f2	Dynamical	2000–2021	JJAS	PRS	No
FIO-ESM	Dynamical	2000–2021	JJAS	PRS	Yes (PR)
GFDL-FLOR	Dynamical	1981–2020	JJAS	PRSe(12)	Yes (PR)
GFDL-SPEAR	Dynamical	1993–2021	JJAS	PRSe(15)	Yes (PR)
GFDL-SPEAR-IceDA	Dynamical	1992–2021	JJAS	PRSe(15)	Yes (PR)
Met Office	Dynamical	1993–2016	JJA	Pe(21)	Yes (P)
NASA GMAO	Dynamical	1981–2021	JJA	PRSe(10)	No
PIOMAS-CFS	Dynamical	2000–2020	JJAS	PRSe(4)	No
RASM	Dynamical	2001–2021	JJAS	PRSe(10)	No
UCLouvain	Dynamical	2006–2019	June	PRSe(10)	No
NCAR-CU Sea Ice Pool	Heuristic	2008–2021	June	P	No

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0163.1.

b. Statistical and Dynamical Prediction Systems

The submitted predictions can be grouped into two main categories - dynamical and statistical predictions. Dynamical predictions are based on numerical dynamical models that are initialized from observationally constrained initial conditions and integrated forward in time. Statistical predictions are based on empirical predictor-predictand relationships and are trained using past observational or reanalysis data. It should also be noted that the distinction between dynamical and statistical methods is not perfect - for example, many dynamical models use statistical post-processing techniques to bias correct their predictions and many statistical models are trained on reanalysis-based predictor data. There is also one submitted "heuristic" prediction from the NCAR / University of Colorado sea ice pool. This office pool collects September SIE predictions each summer on June 1 from NCAR / CU scientists, and serves as a useful 'human expert assessment' baseline to compare against the skill of dynamical and statistical models (Hamilton et al. 2014).

Table 2 summarizes the dynamical prediction systems, which come in three main varieties: fully-coupled global models, fully-coupled regional models driven by specified lateral boundary conditions, and ice-ocean models driven by specified atmospheric forcing. Fully-coupled global models are the most common model formulation, likely because many centers have carefully developed these models for climate modelling applications. Regional models offer the advantage of substantial computational savings, allowing for Arctic simulations at higher resolution, but come with the additional challenges of requiring high-quality boundary conditions and significant research investment in model development. The ice-ocean models that use specified atmospheric forcing are driven either using atmospheric fields from another prediction system or using reanalysis atmospheric fields from previous years. The spatial ice-ocean resolution of the global dynamical models range from 0.25° to 2.8° nominal horizontal resolution, whereas the two submitted regional models have 0.08° and 0.3° nominal resolutions, respectively. The horizontal atmospheric resolutions employed range from 0.4° to 2.8°. Most of the dynamical prediction systems incorporate observations of SIC (11 of 17 systems), sea-surface temperature (SST; 14 systems), ocean temperature and salinity (T/S) profiles (13 systems), and reanalysis atmospheric data (15 systems) into their initialization procedure. A number of systems also initialize their models using observed sea level anomaly (SLA) data (4 systems), and SIT data (2 systems). A variety of different data assimilation techniques are employed including 3DVAR, 4DVAR, strongly and weakly coupled ensemble

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0163.1.

Kalman filters, nudging, optimal interpolation, and reanalysis-forced ice-ocean runs. Most of the dynamical models are ensemble prediction systems, and their deterministic SIO prediction is taken as the ensemble mean. Note that for SIE predictions, SIE is first computed for each ensemble member and then averaged to form the ensemble mean.

The methodologies of each statistical prediction system are summarized in Table 3. A variety of different methods are employed, including standard statistical techniques such as linear regression, multiple regression, autoregressive models, and more complex methods including convolutional neural networks, Gaussian process regression, multivariate linear Markov models, long short-term memory networks, and harmonic decomposition. Most models include a sea ice predictor variable–typically SIE or SIC–and some models also include thermodynamic ocean variables and dynamic and thermodynamic atmospheric variables as predictors (see Table 3, column 2). The reader is reminded that the predictand variables are provided in Table 1, column 5. All submitted statistical models are trained using past data only. Some prediction systems choose to specify a designated training period (e.g. 1979–2000) and use a fixed statistical model to predict all future years (e.g. 2001–2021). Other systems re-train their model each successive year using all available past data (e.g. predict 2001 based on 1979–2000 data, predict 2002 based on 1979–2001 data, etc.). As such, we are unable to disentangle the relative skill from the sophistication of the statistical approach versus other aspects of the statistical forecast (e.g. the use of training data).

Many of the systems perform a post-processing of their predictions in order to correct systematic biases present in their retrospective predictions (see Table 1). The bias correction methods employed are relatively simple, such as correction of the mean bias, correction of the trend, or a linear regression adjustment. Some systems bias correct their SIE time series directly, whereas others correct the SIC spatial fields. We note that some bias correction methods require computing anomalies relative to a climatology, which may implicitly incorporate future data. This is a standard approach for retrospective prediction assessment, but may artificially increase prediction skill (Risbey et al. 2021).

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0163.1.

TABLE 2. Summary of Dynamical Prediction Models. Acronyms used are: Sea ice concentration (SIC), sea ice thickness (SIT), sea surface temperature (SST), subsurface temperature and salinity (T/S) profiles, sea level anomaly (SLA), Ensemble Kalman Filter (EnKF), ECMWF Reanalysis (ERA), Integrated Forecasting System (IFS), Ocean Reanalysis System (ORAS), National Centers for Environmental Prediction (NCEP), Climate Forecast System Reanalysis (CFSR), Modern-Era Retrospective analysis for Research and Applications (MERRA), Japanese Reanalysis (JRA), Climate Prediction Center (CPC), Numerical Weather Prediction (NWP), Forecast Ocean Assimilation Model (FOAM), HadISST2 combined with Canadian Ice Service Charts (Had2CIS).

Name	Initialization Data	Brief Method Description and Reference
BCCR	SIC, SST, T/S profiles	Fully-coupled global with 1° ice-ocean, 2° atmosphere; initialized from strongly coupled ice-ocean EnKF (Kimmritz et al. 2019)
CNRM	SST, T/S profiles, ERA-Int/ERA5	Fully-coupled global with 0.25° ice-ocean, 0.5° atmosphere; initialized from forced ice-ocean with T/S nudging, atmo. from ERA5 (http://www.umr-cnrm.fr/IMG/pdf/system7-technical.pdf; Voldoire et al. (2019))
CPC CFSv2	SIC, SST, T/S profiles, CFSR	Fully-coupled global with 0.5° ice-ocean, 1° atmosphere; initialized from CFSR (Wang et al. 2013; Saha et al. 2014).
CPC CFSm5	SIC, SST, T/S profiles, CFSR	Fully-coupled global with 0.5° ice-ocean, 1° atmosphere; initialized from CFSR and CPC sea ice intialization system (Liu et al. 2019; Collow et al. 2019)
ECCC-CanSIPSv2	SIC, SST, T/S profiles, Era-Int	Two fully-coupled global models: CanCM4i with 2.8° ice-atmosphere, 1° ocean; initialized from Had2CIS SIC, nudged run, and offline ocean T assimilation from ORAP5; GEM-NEMO with 1° ice-ocean, 1.4° atmosphere; initialized from Had2CIS SIC and ORAP5 (Lin et al. 2020).
EC-Earth	SST, T/S profiles, Era-Int	Fully-coupled global with 1° ice-ocean, 1° atmosphere; initialized from ORAS4 in the ocean, atmo. from Era-Int (Hazeleger et al. 2012)
ECMWF SEAS5	SIC, SST, T/S profiles, SLA, ERA5/IFS	Fully-coupled global with 0.25° ice-ocean, 0.4° atmosphere; initialized from ERA5/IFS 4DVAR and ORAS5/OCEAN5 (Johnson et al. 2019; Zuo et al. 2019)
FGOALS-f2	T profiles, JRA55	Fully-coupled global with 1° ice-ocean, 1° atmosphere; initialized from nudged run (Li et al. 2021)
FIO-ESM	SIC, SIT, SST, SLA	Fully-coupled global with 1° ice-ocean, 1° atmosphere; initialized from weakly coupled EnKF (Qiao et al. 2013; Chen et al. 2016; Shu et al. 2021)
GFDL-FLOR	SST, T/S profiles, NCEP-2	Fully-coupled global with 1° ice-ocean, 0.5° atmosphere; initialized from weakly coupled EnKF (Msadek et al. 2014; Bushuk et al. 2017a)
GFDL-SPEAR	SIC, SST, T/S profiles, CFSR	Fully-coupled global with 1° ice-ocean, 0.5° atmosphere; initialized from weakly coupled EnKF and nudged run (Bushuk et al. 2022)
GFDL-SPEAR-IDA	SIC, SST, T/S profiles, CFSR	Fully-coupled global with 1° ice-ocean, 1° atmosphere; initialized from weakly coupled EnKF, sea ice EnKF, and nudged run (Zhang et al. 2022)
Met Office	SIC, SST, T/S profiles, SLA, Met Office NWP	Fully-coupled global with 0.25° ice-ocean, 0.6° atmosphere; initialized from Met Office 4DVAR and FOAM/NEMOVAR (Blockley et al. 2014; MacLachlan et al. 2015)
NASA-GMAO	SIC, SST, T/S profiles, SLA, MERRA-2	Fully-coupled global with 0.5° ice-ocean, 0.5° atmosphere; initialized from weakly coupled EnKF (Molod et al. 2020)
PIOMAS-CFS	SIC, SIT, SST, CFSR/CFS	Regional ice-ocean with 0.3° ice-ocean forced with atmospheric fields from CFS forecasts; initialized via nudging and optimal interpolation (Zhang and Rothrock 2003; Zhang et al. 2008)
RASM	CFSR	Fully-coupled regional with 0.08° ice-ocean and 0.5° atmosphere forced with CFS operational forecasts; initialized from RASM hindcast run nudged to CFSR (air temperature and winds) above 540 hPa (Cassano et al. 2017)
UCLouvain	JRA55	Global sea ice-ocean NEMO3.6/LIM3 with 1° resolution forced with JRA55 atmospheric forcing from the ten previous years; initialized from forced ice-ocean run (Rousset et al. 2015; Barthélemy et al. 2018)

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0163.1.

TABLE 3. Summary of Statistical Prediction Models. Acronyms used for training/intialization data are: Sea ice concentration (SIC), sea ice extent (SIE), sea ice thickness (SIT), sea ice velocity (SIU), melt pond area (MPA), sea surface temperature (SST), ocean heat content (OHC), ocean temperature (OT), 2m air temperature (SAT), downwelling longwave radiation (LWDN), downwelling shortwave radiation (SWDN), net surface heat flux (NSHF), sea level pressure (SLP), surface pressure (PS), geopotential height (Z), surface wind (USURF/VSURF), winds at geopotential height level (UZ/VZ), specific humidity (q), rain rate (RR), snowfall rate (SR), precipitable water content (PWC), Icelandic Low (IL), Arctic Oscillation (AO).

Name	Training/Initialization Data	Brief Method Description and Reference
AWI	SIE, SAT, LWDN, USURF, VSURF, PWC, SLP, SST, 700m OHC, 100m OT	Stability maps and multiple regression (Ionita et al. 2019).
BDAL	SIE, SST, PS, USURF, VSURF, qSURF, SAT, SWDN, LWDN, RR, SR	Long Short Term Memory (LSTM) model (Ali et al. 2021)
Cawley	SIE	Gaussian process regression (Williams and Rasmussen 2006)
CPOM	MPA	Spatially-weighted linear regression model (Schröder et al. 2014)
CPOM-UCL	SIC, SST	Complex networks and Gaussian process regression (Gregory et al. 2020)
CSU	SAT, SIC, SIT, SST, IL, AO	Multiple regression
Damped Persistence	SIE/SIC	Damped anomaly persistence
GSFC_Petty	SIC/SIE	Spatially-weighted linear regression model (Petty et al. 2017).
Horvath	SIC, SIT, SAT, LWDN, SWDN, SIU	Linear mixed effects regression (Horvath et al. 2021).
KOPRI	SIC	Convolutional LSTM model with perceptual loss function (Chi and Kim 2017; Chi et al. 2021).
Lamont	SIC, SST, SAT, Z300, UZ300, VZ300	Multivariate linear Markov model (Yuan et al. 2016).
MetNo-sparse-st	SIE	Autoregressive model with adaptive order
Nico Sun	SIC	SIC persistence and past-year analogues
SYSU/SML-KNN	SIC, NSHF	K-nearest neighbor algorithm (Lin et al. 2023)
SYSU/SML-MLM	SIC, SST, SAT, NSHF	Multivariate linear Markov model (Zeng et al. 2023)
Trend Climatology	SIE/SIC	Linear trend
UCLA	SIE	Data-adaptive harmonic decomposition (Chekroun and Kondrashov 2017; Kondrashov et al. 2018).
UMBC-REU	SIC, SST, SP, USURF, VSURF, qSURF, SAT, SWDN, LWDN, RR, SR	Convolutional Neural Network model (Kim et al. 2021)
UPenn	SIE	Feature-engineered linear regression (Diebold and Göbel 2022; Diebold et al. 2023)

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0163.1.

c. Observational Verification

Consistent with the SIO evaluation, we verify Pan-Arctic SIE predictions against the NSIDC Sea Ice Index, Version 3 (Fetterer et al. 2017), which is based on the NASA Team retrieval algorithm. We also verify Pan-Arctic SIE predictions against the OSI SAF Sea Ice Index, Version 2.1 (OSI-420), which uses the Bristol/Bootstrap retrieval algorithm (Lavergne et al. 2019). SIC predictions are verified against the NOAA/NSIDC Climate Data Record (CDR) of SIC, version 4 (G02202; Meier et al. (2021)) and the OSI SAF SIC CDR, release 3 (OSI-450a; EUMETSAT Ocean and Sea Ice Satellite Application Facility (2022)). Both of these products use a spatial interpolation to gap fill the polar observational hole. We also use the NSIDC and OSI SAF CDR SIC data to compute regional SIE using the recently updated NSIDC Arctic region mask (Meier and Stewart 2023). We perform all SIC analysis on the 25km NSIDC polar stereographic north grid and regrid each model's SIC data to the NSIDC grid using bilinear interpolation and NSIDC's CDR land-sea mask. In cases where the model land-sea boundary lies within the NSIDC ocean domain, nearest neighbor extrapolation to the NSIDC grid is used.

d. Skill Metrics

We quantify prediction skill using the anomaly correlation coefficient (ACC) and root mean squared error (RMSE) between predicted and observed time series, which are commonly used metrics in the sea ice prediction literature. The ACC is the temporal correlation between predicted and observed time series and is defined as

$$\mathrm{ACC}(\tau) = \frac{\sum_{i=1}^{N} \left(p_i(\tau) - \overline{p(\tau)}\right)\left(o_i - \bar{o}\right)}{\sqrt{\sum_{i=1}^{N} \left(p_i(\tau) - \overline{p(\tau)}\right)^2} \sqrt{\sum_{i=1}^{N} \left(o_i - \bar{o}\right)^2}}, \tag{1}$$

where p_i is the model prediction for year i, o_i is the observed value, τ is the forecast lead time, N is the number of years, and the overbar indicates a temporal mean. The RMSE is given by

$$\mathrm{RMSE}(\tau) = \sqrt{\frac{\sum_{i=1}^{N} \left(p_i(\tau) - o_i\right)^2}{N}}. \tag{2}$$

In order to isolate trend-independent prediction skill, we compute detrended skill metrics, which remove a linear trend from both predicted and observed time series prior to computing the skill

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0163.1.

metrics. The detrended ACC is defined as

$$\mathrm{ACC}_{\mathrm{detrend}}(\tau) = \frac{\sum_{i=1}^{N} \left(p_i(\tau) - p_i^L(\tau)\right)\left(o_i - o_i^L\right)}{\sqrt{\sum_{i=1}^{N} \left(p_i(\tau) - p_i^L(\tau)\right)^2}\sqrt{\sum_{i=1}^{N} \left(o_i - o_i^L\right)^2}}, \tag{3}$$

where $p_i^L(\tau)$ and o_i^L are linear trend fits to the predicted and observed time series. Note that $p_i^L(\tau)$ is a function of lead time τ, since each lead time will have its own linear trend prediction. Similarly, the detrended RMSE is defined as

$$\mathrm{RMSE}_{\mathrm{detrend}}(\tau) = \sqrt{\frac{\sum_{i=1}^{N} \left(\left(p_i(\tau) - p_i^L(\tau)\right) - \left(o_i - o_i^L\right)\right)^2}{N}}. \tag{4}$$

Note that, unlike RMSE, the detrended RMSE has no contribution from mean bias, since this bias is subtracted off during the detrending procedure, but it does have contributions from conditional biases (predicting the incorrect amplitude of anomalies). Another commonly used metric in the sea ice prediction literature is the mean-squared error skill score (MSSS), which is connected to the ACC and RMSE via the decomposition of Murphy (1988). In particular, the squared ACC skill provides an upper bound on the MSSS and can be interpreted as the variance explained by a regression-adjusted forecast that is free of conditional and mean biases.

In order to facilitate an "apples-to-apples" skill comparison, we focus most of our analysis on the 2001–2020 time period, which is the period with the most submitted predictions (see Table 1). Note that some models were only able to submit predictions for a portion of this time period, which may bias their skill results. Specifically, 24 models submitted predictions for the full 2001-2020 period and 31 models submitted at least 14 years of predictions. We also include figures in the Supplementary Material showing prediction skill metrics computed over the full time period submitted by each model (Figs. S5, S6). We emphasize that the overall conclusions of the study are unchanged if the full time period is used for computing skill.

e. Reference and Multi-Model Predictions

We compare model prediction skill to two reference predictions; a linear trend climatology and a damped anomaly persistence forecast. The linear trend climatology prediction, o_i^L, is computed for a given year i by computing a linear fit to September SIE using all available past data (i.e.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0163.1.

1979 to year $i-1$) and evaluating the linear function for year i. The damped anomaly persistence forecast uses the linear trend climatology prediction and adds a scaled observed anomaly at the initialization time. The damped anomaly persistence forecast is given by

$$DP_i(\tau) = o_i^L + r(o'_{\text{Sep}}, o'_\tau)\frac{\sigma_{\text{Sep}}}{\sigma_\tau} o'_i(\tau), \tag{5}$$

where $o'_i(\tau)$ is the observed anomaly in year i at lead time τ, $r(o'_{\text{Sep}}, o'_\tau)$ is the correlation between September anomalies and anomalies at lead time τ, and σ_{Sep} and σ_τ are the standard deviations of these respective anomalies (Van den Dool 2007). "Anomalies" here are detrended anomalies (i.e. they are computed relative to the linear trend climatology). The observed anomaly is computed using the daily observation immediately prior to the initialization date (for example, the June 1 observed anomaly is taken as the May 31 anomaly). The linear trend climatology, lagged correlation and standard deviation values are updated each year using all available past data. Henceforth, we refer to the damped anomaly persistence forecast as damped persistence.

We also compute a multi-model median prediction, which is the median predicted value across all models for each year and each lead time. The multi-model median prediction is only computed for years with at least 10 models available (years 1993–2021), in order to reduce the impact of sampling bias.

3. Pan-Arctic Predictions

a. September Pan-Arctic SIE Prediction Skill

We begin by assessing the ability of models to predict September Pan-Arctic SIE, which is the flagship prediction target of the SIO. Figure 2 shows time series of NSIDC observed September SIE (black) and multi-model median predictions (red) from initialization dates of June 1–September 1. The red shading indicates the interquartile range (middle 50%) of individual model ensemble-mean predictions. We find that the multi-model median prediction has high skill across SIO lead times, capturing both the observed SIE trend and interannual variations over the period 1993–2021. The ACC values, which include a substantial trend contribution, are greater than 0.9 for all lead times, whereas the detrended ACC values range from 0.66–0.97. The RMSE values of the multi-model median prediction are substantially smaller than the observed detrended standard deviation (0.54

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0163.1.

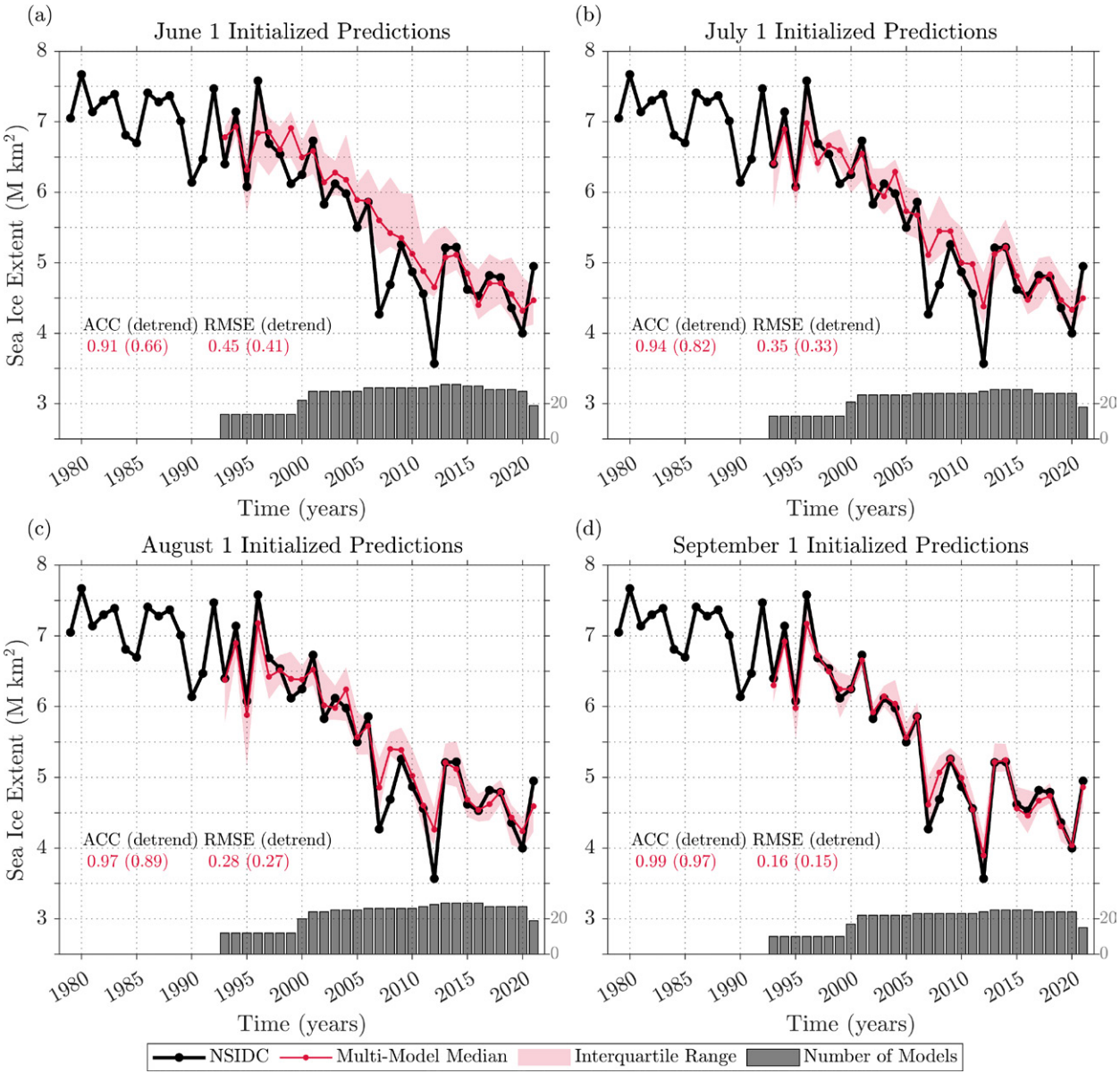

FIG. 2. Multi-model predictions of Pan-Arctic SIE initialized on (a) June 1, (b) July 1, (c) August 1, and (d) September 1. The multi-model predictions are based on a multi-model median (red). Red shading indicates the interquartile range (middle 50%) of individual model predictions. Skill metrics computed over 1993–2021 are shown in red text, with detrended skill in parentheses. The number of models available for each year is indicated by the grey bars at the bottom of each plot and grey text on the right y-axis. Multi-model predictions are only plotted for years with at least 10 models available.

million km^2), indicating prediction skill relative to the trend climatology prediction. We find that the multi-model predictions become more confident (decreased inter-model spread) as the lead

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0163.1.

time decreases, and also capture SIE anomalies with greater skill. For example, better predictions of the extreme 1996, 2007, and 2012 SIE anomalies are made from July 1 than June 1, and similar improvements are seen in the forecasts from August 1 and September 1, respectively. We note that the retrospective skill of the multi-model median prediction is considerably higher than the skill of multi-model median real-time predictions submitted to the SIO (see supplementary Fig. S1). We return to this point in the discussion section (section 6b). Versions of Fig. 2 for each individual model submitted can be viewed on github (https://github.com/MitchBushuk/SIO_review_paper).

Next, we take a more granular view and explore the prediction skill of individual models. Figures 3 and 4 show the prediction skill of the dynamical and statistical models, respectively. We find that the majority of dynamical and statistical models are skillful at SIO lead times, outperforming the trend climatology prediction (dashed grey line). The models also generally outperform damped persistence (solid grey line) from June 1 and July 1, whereas damped persistence provides a more challenging benchmark from August 1 and September 1, with about half the models beating damped persistence from August 1 and most models losing to damped persistence from September 1. While there is a large spread in skill across models, we find that the majority of models have detrended ACC values that exceed 0.4 from June 1, and 0.5 from July 1 onward, the latter of which is a commonly used practical threshold for useful forecast skill. The fact that this broad set of models, which employ diverse prediction methodologies and input datasets, are generally skillful at SIO lead times shows that useful real-time multi-month predictions of September sea ice should be achievable.

The very high skill of damped persistence from September 1 (detrended ACC of 0.98) indicates that interannual fluctuations of September-mean SIE are essentially "locked in" by September 1. This high skill demonstrates that the key source of predictability from September 1 is the multi-week persistence of SIE anomalies, which have particularly high persistence values at the time of the summer minimum (Blanchard-Wrigglesworth et al. 2011a). Since these SIE anomalies are observable in near-real time, dynamical prediction systems should, in principle, be able to initialize predictions using these data and capture this source of predictability. However, we find that the majority of dynamical models are less skillful than damped persistence from September 1, which indicates that they are making errors in their sea ice initial conditions and/or have substantial short term forecast drift that is not adequately post-processed in the forecasts. The most

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0163.1.

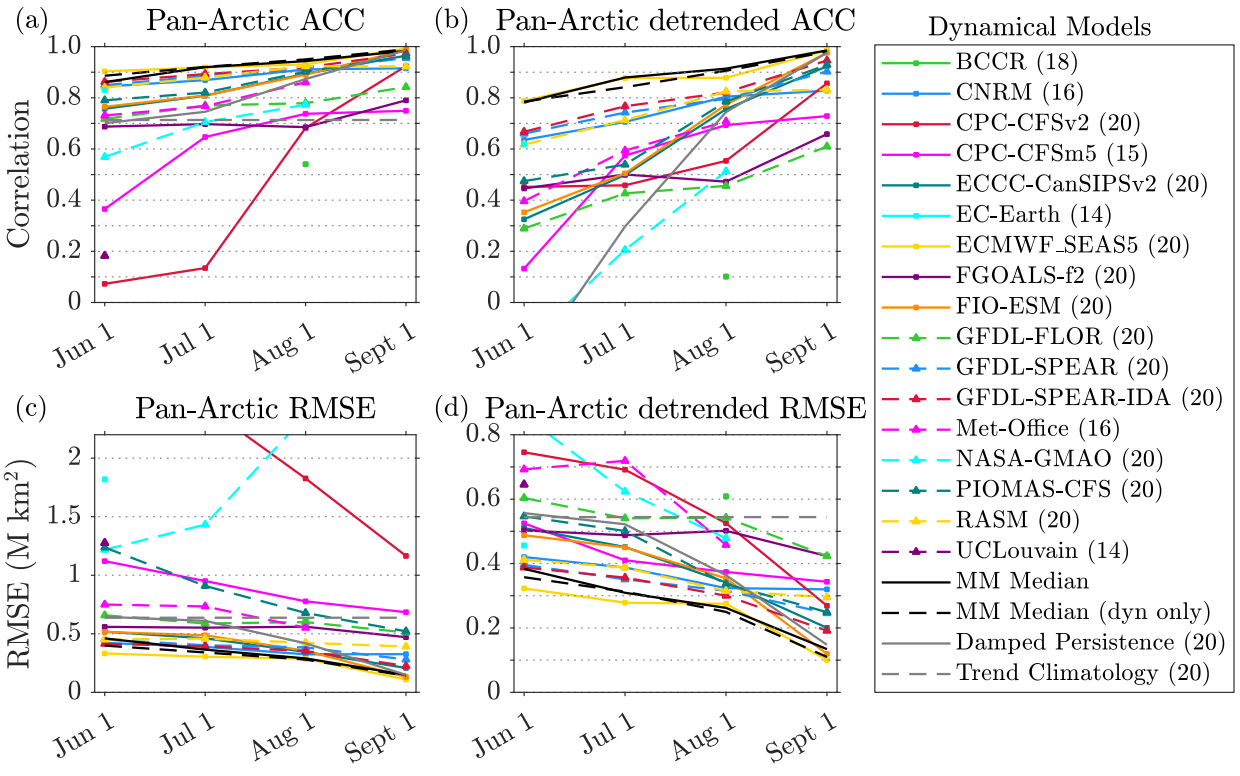

FIG. 3. Dynamical model prediction skill for September Pan-Arctic SIE computed over the period 2001–2020. Individual models are shown in colors, multi-model predictions are shown in black, and reference predictions are shown in grey. Skill metrics are plotted for each available initialization time (June 1–September 1) and are computed for both full (a,c) and detrended (b,d) time series. The numbers in parentheses in the legend indicate the number of years available from each model over the 2001-2020 time period. Note that the isolated markers in the plots correspond to models that submitted a single initialization month.

skillful dynamical models from September 1 are comparable to the damped persistence benchmark, suggesting that these systems are successfully assimilating sea ice concentration or other related observations. Similarly, the most skillful statistical models are similar to damped persistence from September 1 and most statistical models have lower skill than this benchmark despite, in principle, having access to the same SIE observations as used by the damped persistence forecast. This lower skill likely results from a combination of factors, such as some models using monthly rather than daily data and some models including other predictor variables besides SIE which may negatively impact September 1 skill in favor of higher skill at longer lead times. We also note that training and verifying the damped persistence forecast on different datasets can provide a useful

Accepted for publication in *Bulletin of the American Meteorological Society*. DOI 10.1175/BAMS-D-23-0163.1.

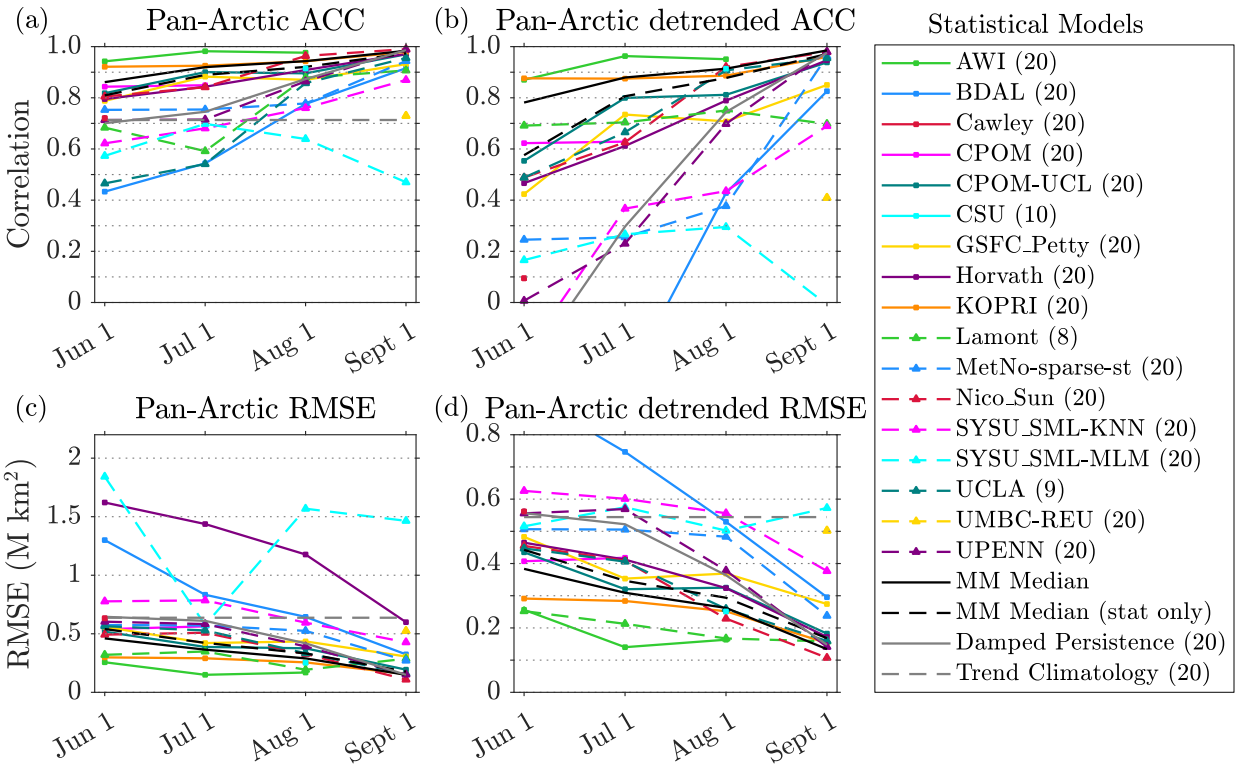

FIG. 4. Statistical model prediction skill for September Pan-Arctic SIE computed over the period 2001–2020. Individual models are shown in colors, multi-model predictions are shown in black, and reference predictions are shown in grey. Skill metrics are plotted for each available initialization time (June 1–September 1) and are computed for both full (a,c) and detrended (b,d) time series. The numbers in parentheses in the legend indicate the number of years available from each model over the 2001-2020 time period.

measure of observational uncertainty. We find detrended ACC values of 0.96, 0.95, and 0.95 based on training/verification pairs of NSIDC/OSI SAF, OSI SAF/NSIDC, and OSI SAF/OSI SAF, respectively, which are slightly lower than the value of 0.98 for NSIDC/NSIDC reported above.

Moving to longer lead times, we find that slightly more than half the models outperform damped persistence from August 1, and nearly all the models outperform damped persistence from June 1 and July 1. This indicates that the models are successfully capturing other sources of predictability at these lead times, potentially including SIT anomaly persistence, surface albedo anomalies and ice-albedo feedback, surface air temperature anomalies, and atmospheric circulation patterns. Taken as a whole, the Pan-Arctic skill of the dynamical and statistical models are broadly similar,

Accepted for publication in *Bulletin of the American Meteorological Society*. DOI 10.1175/BAMS-D-23-0163.1.

however, the model spread precludes definitive statements on which class of method is preferable for Pan-Arctic predictions.

The multi-model median prediction has high skill, with detrended ACC values exceeding 0.75 for all SIO lead times. The multi-model median skill is higher than nearly all individual models, suggesting that this prediction benefits from cancellation of random errors across prediction systems, which is a common finding across a variety of prediction applications including the SIO (e.g., Hagedorn et al. (2005); DelSole et al. (2014); Harnos et al. (2019); Blanchard-Wrigglesworth et al. (2023)) as well as the Southern Ocean counterpart of the SIO, the SIPN South ensemble (Massonnet et al. 2023). We also note that the skill of a multi-model median prediction based only on dynamical models is similar to the skill of the multi-model median based on all models, whereas the median prediction based only on statistical models has lower skill.

Of the dynamical models, ECMWF SEAS5 stands out as having particularly high Pan-Arctic prediction skill, achieving comparable skill to the multi-model median. There are also two statistical models that are high-skill outliers: the AWI model, which employs a multiple regression based on stability maps, and the KOPRI model, which uses a convolutional long-short term memory model. We note that the skill levels of the AWI and KOPRI models are roughly equal to the upper limit of Pan-Arctic SIE predictability as estimated by perfect model GCM experiments (compare with the July 1 initialized forecast skill in Fig. 1 of Tietsche et al. (2014)). We return to the possible sources of prediction skill across the individual systems in Section 6a.

We also verify the predictions using the OSI SAF sea ice index (see supplementary Figs. S3,S4). The OSI SAF sea ice index has a higher mean value than the NSIDC sea ice index (see Fig. 1a), but the indices otherwise have a close agreement, with ACC of 1.00, detrended ACC of 0.98, and detrended RMSE of 0.10 million km^2. Consistent with this close agreement, we find that the skill values are not sensitive to the choice of verification product, and that the choice of verification product does not affect the qualitative conclusions regarding Pan-Arctic skill. The main difference between the NSIDC and OSI SAF-verified skill metrics occurs for the RMSE skill, since this metric is affected by the mean offset between the products, whereas the other skill metrics are not.

The heuristic prediction submitted from the NCAR / CU sea ice pool provides a useful 'human expert assessment' baseline for Pan-Arctic SIE prediction skill. We find that this June 1 heuristic prediction has no skill (ACC=-0.18; detrended ACC=-0.39) over their submission period of 2008–

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0163.1.

2021 (see Hamilton et al. (2014), and more recent figure here: https://bit.ly/3MscjmL), emphasizing the inherent challenges in human-based assessments.

b. Is Prediction Skill Changing Over Time?

Earlier theoretical work has shown that Arctic sea ice predictability is dependent on the mean climate state (Holland et al. 2011; Holland and Stroeve 2011; Cheng et al. 2016; Holland et al. 2019). While some have argued that the recent observed trends towards a thinner and more mobile ice pack may reduce inherent summer sea ice predictability, Holland et al. (2019) show that changes in sea ice predictability characteristics are highly non-monotonic under climate change and sea ice predictability actually reaches a local maximum in the CESM1 model in the 2010s decade. We can use the retrospective prediction dataset to investigate this question by analyzing the evolution of prediction errors over time across the multi-model dataset.

Figure 5a shows the multi-model mean of single-model detrended Pan-Arctic SIE absolute errors plotted as a function of time (horizontal axis) and initialization date (colors). We find that the error time series do not display clearly identifiable trends, but are punctuated by large errors in the extreme sea ice years of 1996, 2007, and 2012, which respectively had high, low, and low sea ice extents (Serreze and Stroeve 2015; Kay et al. 2008; Zhang et al. 2013). The trends in prediction errors are not significantly different from zero (at the 95% confidence level) for any initialization month. This finding suggests that there has not been a detectable change in sea ice predictability since 1990.

As expected, we find that the SIE errors increase with lead time, but the error reduction between lead times changes from year to year. For example, 2005 has similar errors across lead times, 2007 shows a similar reduction for each successive initialization month, and 2012 shows large reductions from June to July and from August to September, but little change from July to August. These differences are likely related to the particular synoptic conditions of each summer - for example the August 1 error is particularly large in 2012, likely because the great Arctic cyclone, which peaked on August 6 (Simmonds and Rudeva 2012) and led to rapid sea ice loss in August, was not predicted (or its impact on sea ice) by seasonal prediction systems (Yamagami et al. 2018).

Earlier work has shown that sea ice predictions typically struggle in "hard to predict years" with large SIE anomalies (Stroeve et al. 2014), sometimes related to atmospheric conditions such as

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0163.1.

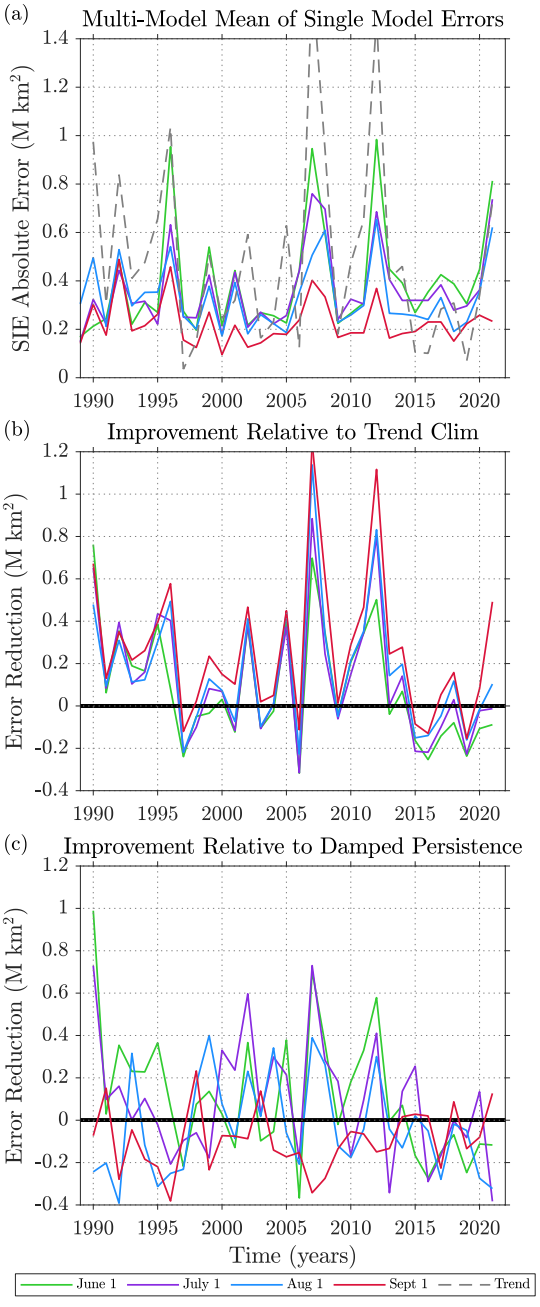

FIG. 5. Pan-Arctic SIE prediction errors versus time. Panel (a) shows the multi-model mean of single-model absolute detrended SIE errors for different years (horizontal axis) and initialization dates (colors). The absolute SIE error of the trend climatology reference prediction is shown in grey. Panels (b) and (c) show the improvement in absolute error relative to the trend climatology and damped persistence predictions, respectively (positive values indicate improvement, negative values indicate degradation).

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0163.1.

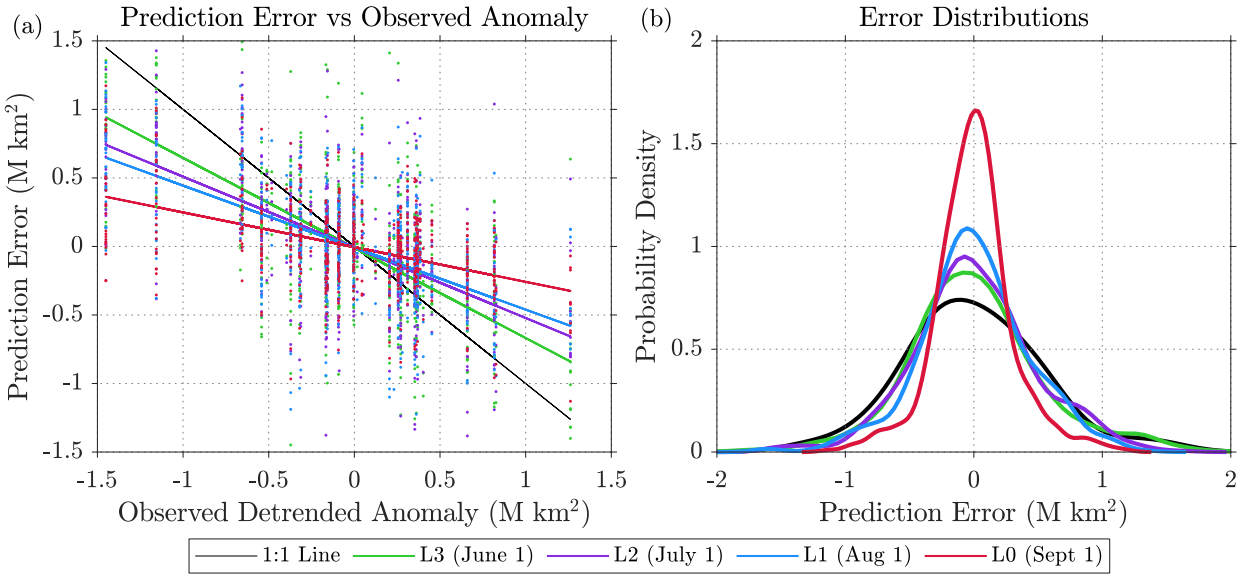

FIG. 6. Relation between detrended prediction errors and observed detrended SIE anomalies. Each dot in panel (a) shows the error for a particular model and target year (errors are plotted over all available years 1980–2021). The dots are colored according to lead time and the colored lines are linear fits. The grey line indicates the error of the linear trend fit to the observations. Panel (b) shows the distribution of errors for different lead times, with the black curve showing the distribution of observed detrended anomalies.

late-summer cyclones (Lukovich et al. 2021; Finocchio et al. 2022). Theoretically, however, some extreme years exhibit seasonal predictability (Tietsche et al. 2013). These large errors in extreme SIE years have been characterized as a major shortcoming of sea ice prediction systems. However, Fig. 5a shows that the linear trend prediction makes much larger errors in these years compared with the prediction systems (compare dashed grey line to colored lines). Figures 5b and 5c show the skill improvement of the model-based predictions relative to the trend climatology and damped persistence predictions, respectively, with positive values indicating error reductions. We find that the time-mean error reductions are generally positive, indicating that the prediction systems typically provide better skill than the reference forecasts, with the exception of the September 1 damped persistence forecast. Moreover, the extreme SIE years of 1996, 2007, and 2012 stand out as years in which the prediction systems provide the largest skill improvements over the linear trend prediction. This challenges the typical interpretation that prediction systems "failed" in these extreme SIE years. Rather, it is precisely these extreme years that the prediction systems provide the most value added relative to basic reference forecasts.

Accepted for publication in *Bulletin of the American Meteorological Society*. DOI 10.1175/BAMS-D-23-0163.1.

We next investigate the error characteristics of individual model predictions in Fig. 6. Figure 6a shows prediction errors from individual models and target years plotted against the observed detrended SIE anomalies in those years. In low SIE years, the models generally overpredict the observed SIE (positive errors), and the models generally underpredict in high SIE years (negative errors). The distribution of errors (Fig. 6b) is relatively symmetric about zero for all initialization times, suggesting that high and low SIE anomalies are similarly difficult to predict. Q-Q plots reveal that the error distributions for all initialization times have symmetric heavy tails compared with a Gaussian distribution, suggestive of outlier models with large errors (not shown). The linear fits to the prediction errors in Fig. 6a (colored lines) have decreasing slopes as the initialization date approaches September, and are bracketed by the 1:1 line (a no skill prediction) and the y=0 line (a perfect prediction). If September SIE were entirely unpredictable, we would expect the errors to lie on the 1:1 line, whereas if it were perfectly predicable, we would expect the errors to lie on the y=0 line. Thus, the decreasing slopes as the initialization date approaches September shows that inherent SIE predictability increases as the lead time decreases. We also find that the prediction error distributions become progressively more peaked around zero as the lead time decreases (Fig. 6b).

4. Regional Predictions

a. September Regional SIE Prediction Skill

The prediction systems skillfully predict Pan-Arctic SIE, but how do they perform on the regional and local scales that users ultimately require? In Figs. 7 and 8, we plot detrended regional SIE skill for the dynamical and statistical models, respectively, in the five regional domains shown in Fig. 1d. The skill metrics for full regional SIE time series are shown in Figs. S9 and S10.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0163.1.

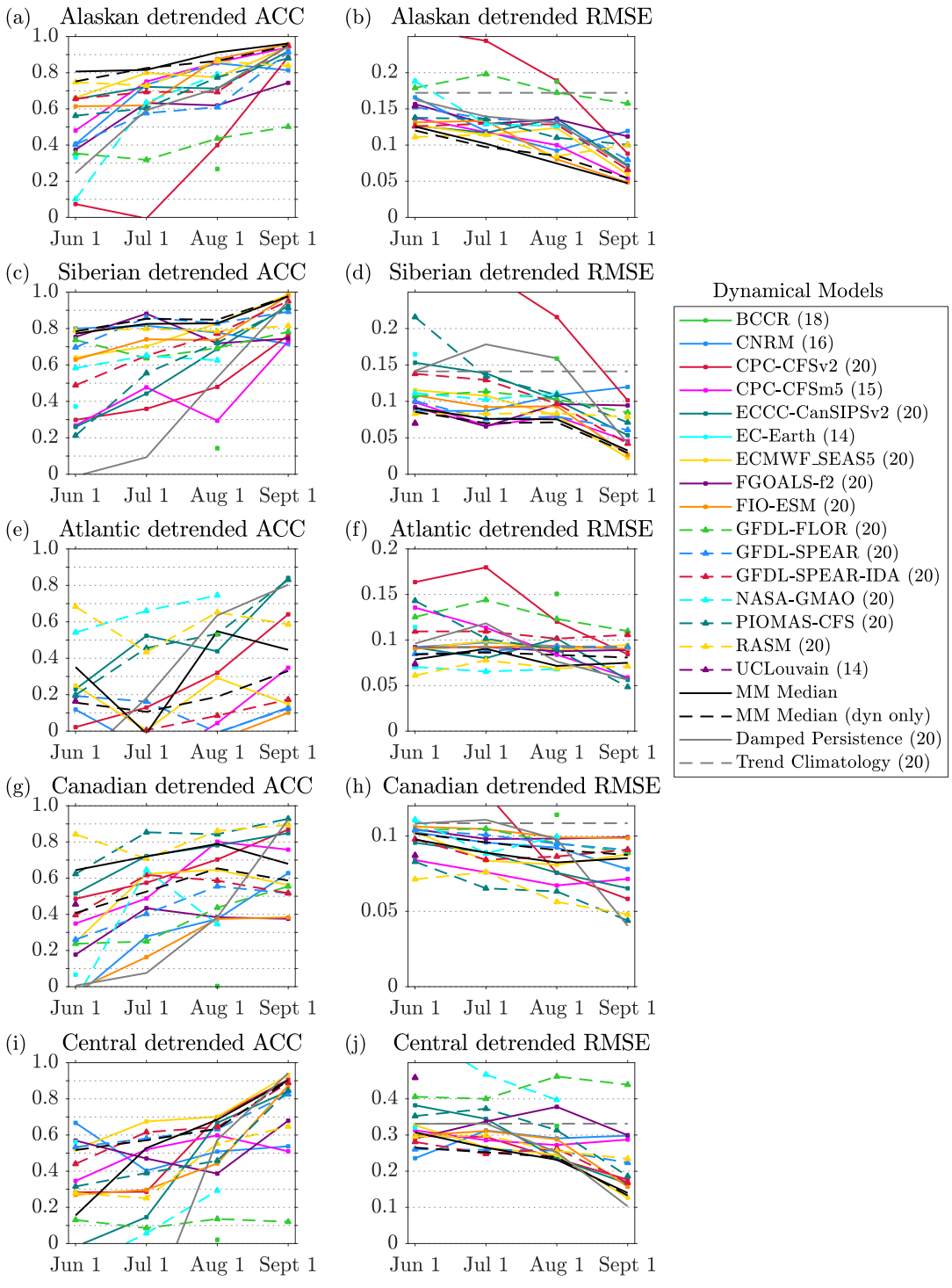

FIG. 7. Dynamical model prediction skill for September regional SIE computed over the period 2001–2020 in the Alaskan (a,b), Siberian (c,d), Atlantic (e,f), Canadian (g,h), and Central (i,j) regions (shown in Fig. 1d). Individual models are shown in colors, multi-model predictions are shown in black, and reference predictions are shown in grey. Skill metrics are plotted for each available initialization time (June 1–September 1) and are computed for detrended time series. The numbers in parentheses in the legend indicate the number of years available from each model over the 2001–2020 time period.

Accepted for publication in *Bulletin of the American Meteorological Society*. DOI 10.1175/BAMS-D-23-0163.1.

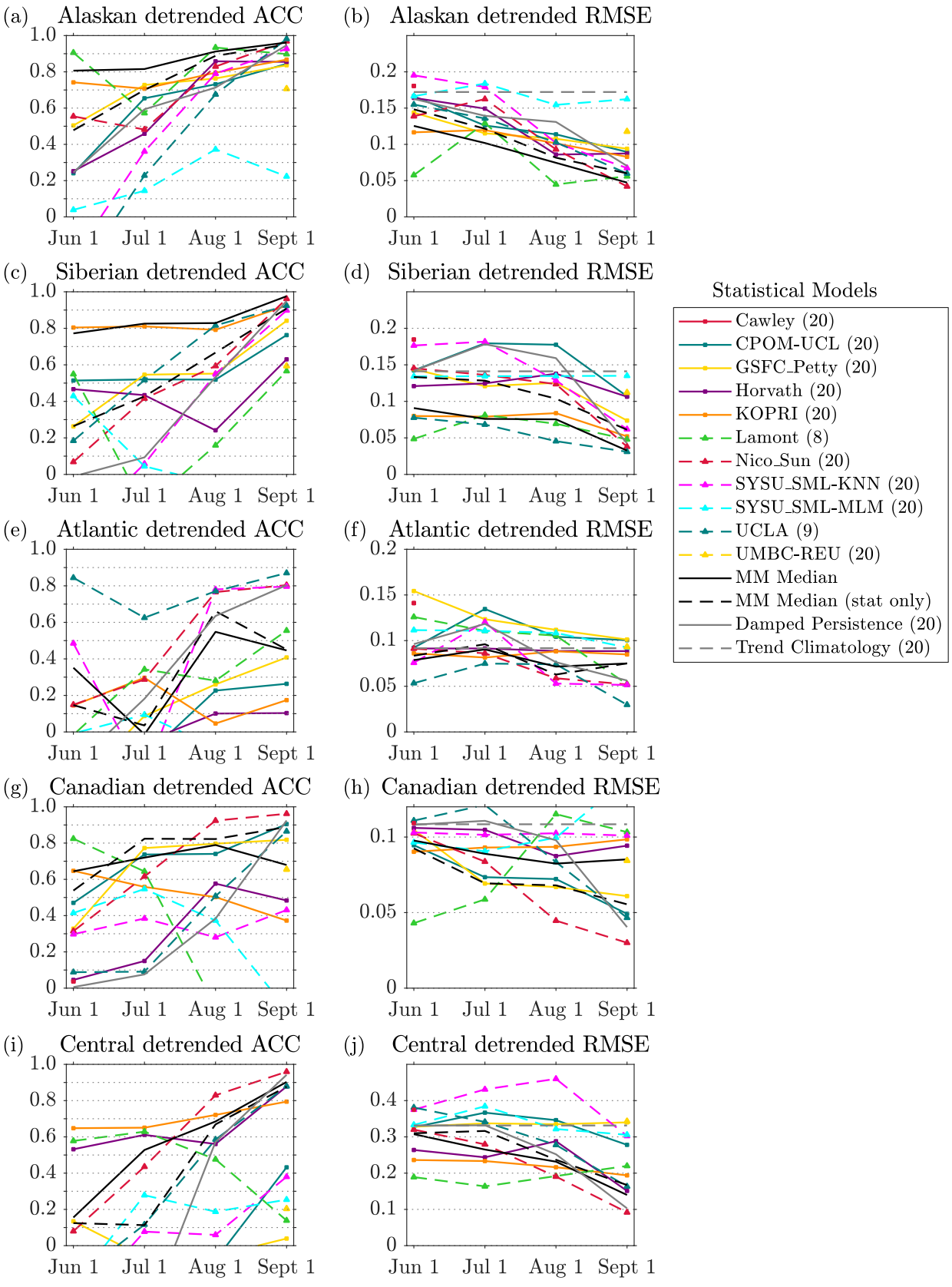

FIG. 8. As in Fig. 7 but for statistical models.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0163.1.

We find that both dynamical and statistical models have detrended regional skill, but the level of skill is regionally variable. The highest skill is found in the Alaskan and Siberian sectors, in which the multi-model median detrended ACC exceeds 0.75 at SIO lead times. Unlike the Pan-Arctic skill results, there is a notable difference between dynamical and statistical model performance in these regions (compare panels a–d of Figs. 7 and 8). Taken as a whole, the dynamical models outperform the statistical models in the Alaskan and Siberian regions, however the KOPRI statistical model has high skill in both regions at a level comparable to the most skillful dynamical models. The dynamical models also outperform the statistical models in the Central Arctic domain (compare panels i–j), whereas the skill differences are more modest in the Canadian and Atlantic regions (panels e–h). Interestingly, the superior regional SIE skill of dynamical models does not clearly translate into better Pan-Arctic skill relative to statistical models.

The model skill is lowest in the Atlantic region for both dynamical and statistical models. This is likely because Atlantic September SIE variations result from SIE variability occurring in the northern portions of the Greenland, Barents, and Kara Seas, which are driven by anomalies in sea ice export that are challenging to predict (Kwok 2008). The Canadian Archipelago is also well known as a difficult to predict region due to its complex network of channels and straits. Encouragingly, the majority of statistical and dynamical models show detrended prediction skill in this region, albeit at a generally lower skill level than dynamical models in the Alaskan and Siberian sector. Of the dynamical models, RASM has high skill in the Canadian region, potentially related to its relatively high horizontal resolution compared to other systems. This higher resolution provides both a more accurate representation of complex land geometry and a more realistic representation of sea ice dynamical and thermodynamical processes. The skill in the Central Arctic domain is the second lowest next to the Atlantic. The Central Arctic SIE time series is dominated by large anomalies in 2007, 2012, and 2020 (Fig. 1b), which suggests that the models generally struggled to capture the Central Arctic anomalies in these years.

Relative to the damped persistence benchmark, the models perform quite skillfully for regional SIE. Analogous to Pan-Arctic SIE, regional SIE damped persistence is highly skillful for September 1 forecasts and provides a stringent benchmark that most dynamical and statistical models fail to beat. The models perform more favorably at longer lead times. In the Alaskan, Siberian and Canadian regions, the majority of models outperform damped persistence from June 1, July 1, and

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0163.1.

August 1 initialization dates. In the Central Arctic, most models beat damped persistence from June 1 and July 1. In the Atlantic sector, the models are notably less skillful than damped persistence from August 1, suggesting a deficiency in the models in representing summertime Atlantic SIE. These regional skill results are insensitive to the verification product–the same conclusions hold if OSI SAF observations are used for verification (see Figs. S11 and S12).

b. Relation Between Pan-Arctic and Regional Skill

Are models more skillful at predicting Pan-Arctic or regional SIE? Do models with high Pan-Arctic skill also have high regional skill? We investigate these questions in Fig. 9, which plots regional vs Pan-Arctic detrended ACC for each model, colored by lead time. In most regions, the majority of predictions lie below the 1:1 line, indicating that regional SIE skill is generally lower than Pan-Arctic skill. The Alaskan region is the most skillfully predicted region, with 46% of predictions lying above the 1:1 line. The damped persistence prediction also lies above the 1:1 line (square markers) indicating that the Alaskan region may have high inherent predictability. The Siberian and Canadian regions are also predicted fairly well, with 37% and 32% of predictions exceeding Pan-Arctic skill, respectively. The performance is notably worse in the Atlantic and Central sectors, as each of these regions only has 12% of predictions that exceed Pan-Arctic skill. We also find that the regional skill differences across models are related to their Pan-Arctic skill differences. For example, the R^2 values between regional and Pan-Arctic detrended ACC are 0.59, 0.48, and 0.49 in the Alaskan, Siberian, and Central regions. Regional skill is more decoupled from Pan-Arctic skill in the Canadian and Atlantic regions, with R^2 values of 0.25 and 0.05, respectively.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0163.1.

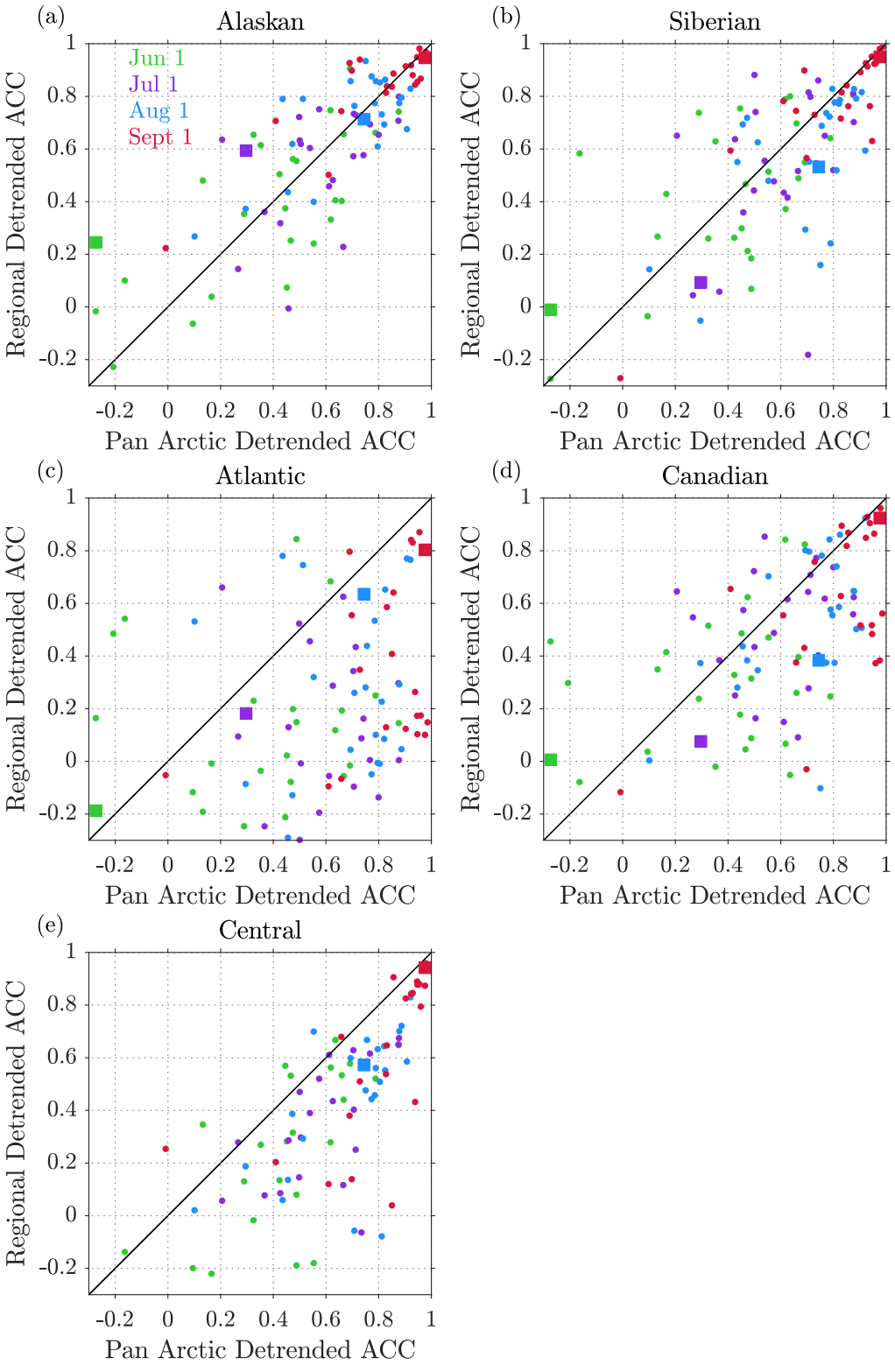

FIG. 9. Relationship between regional and pan-Arctic prediction skill in the Alaskan (a), Siberian (b), Atlantic (c), Canadian (d), and Central (e) regions over the 2001–2020 period. Each dot shows detrended ACC values for an individual model, colored by lead time. Square markers indicate the damped persistence forecast skill, and the 1:1 line is shown in black.

Accepted for publication in *Bulletin of the American Meteorological Society*. DOI 10.1175/BAMS-D-23-0163.1.

5. Sea Ice Concentration Predictions

a. September SIC Prediction Skill

Finally, we consider the ability of models to predict September sea ice variations on the local scale. Figures 10 and 11 show SIC skill metrics for the dynamical and statistical models that submitted SIC predictions, respectively. These metrics are first computed locally and then area-averaged over the zone of September SIC variability, defined as all grid points where the September SIC standard deviation exceeds 10% (see Fig. 1c). The gap between full and detrended SIC skill is quite small, consistent with the fact that observed SIC variability is dominated by interannual rather than trend-based variance (84% and 16% of the total variance, respectively). Compared to the skill levels for Pan-Arctic and regional SIE, the SIC skill scores are lower, consistent with a larger role for unpredictable local-scale dynamics and the fact that, unlike SIE, SIC predictions do not benefit from error compensation (i.e., the cancellation of over and under-estimations). This lower predictability is also reflected by the damped persistence forecast, which is skillful from September 1 but drops off quite rapidly for earlier initialization dates. Interestingly, a handful of models (ECMWF SEAS5, CPC CFSv2, ECCC-CanSIPSv2, FIO-ESM, GFDL-SPEAR-IDA, PIOMAS-CFS, Nico Sun) outperform damped persistence from September 1, which was not the case for Pan-Arctic or regional SIE. This suggests that some models are extracting additional skill from their ability to skillfully predict the atmospheric state over early September and the corresponding local SIC response.

The multi-model median SIC prediction has detrended ACC values above 0.5 for initialization dates of July 1 and later, and falls off for predictions made on June 1. Most individual models exceed the 0.5 detrended ACC threshold from September 1, a handful exceed it from August 1, and all models lie below 0.5 from July 1 and June 1. While the relatively small number of statistical models (7) precludes definitive statements, the dynamical models generally have higher SIC skill than the statistical models. The SIC skill scores are very similar when predictions are verified against OSI SAF SIC observations (see Figs. S13 and S14). Again, we find that the ECMWF SEAS5 model stands out amongst the dynamical models (see solid gold line in Fig. 10) and the KOPRI model stands out amongst the statistical models (see solid orange line in Fig. 11). The fact that these models also perform well at the local scale increases our confidence in their strong Pan-

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0163.1.

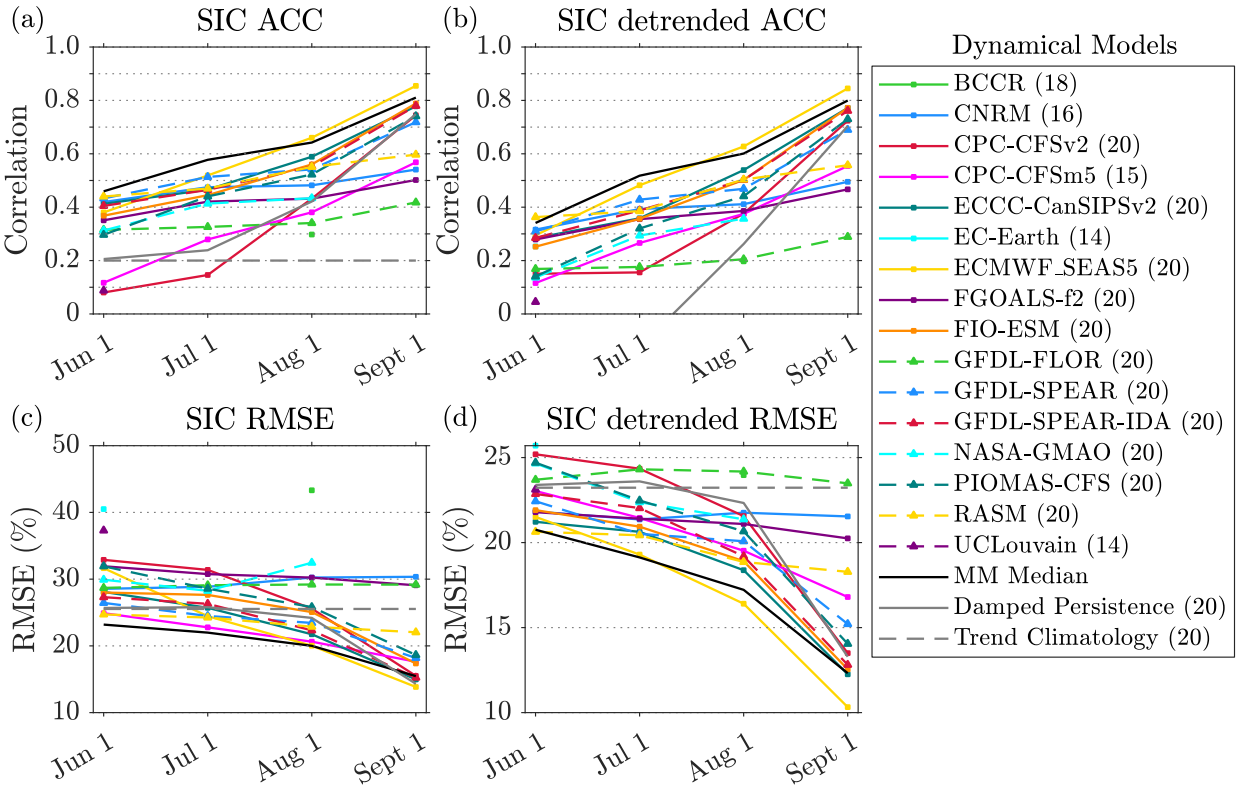

FIG. 10. Dynamical model prediction skill for September SIC computed over the period 2001–2020. Skill values are averaged over the region in which observed SIC standard deviation is greater than 10%. Individual models are shown in colors, multi-model predictions are shown in black, and reference predictions are shown in grey. Skill metrics are plotted for each available initialization time (June 1–September 1). The numbers in parentheses in the legend indicate the number of years available from each model over the 2001–2020 time period.

Arctic and regional performance. It is notable that the ECMWF system also stood out as the best performing system for subseasonal (0–45 day) sea ice predictions in the multi-model comparison study of Zampieri et al. (2018). Interestingly, the ECMWF model has a notable bias from June 1 resulting in SIC RMSE values that are larger than the trend climatology and most other models (see Fig. 10c). Nevertheless, ECMWF maintains detrended prediction skill at this lead time for SIC, regional SIE, and Pan-Arctic SIE, suggesting that this is mostly a linear bias that can be removed.

We also consider the integrated ice-edge error (IIEE), which is the areal integral of local sea ice extent errors (Goessling et al. 2016). The IIEE has contributions from both absolute extent errors (Pan-Arctic SIE errors) and ice edge misplacement errors. Note that we do not consider

Accepted for publication in *Bulletin of the American Meteorological Society*. DOI 10.1175/BAMS-D-23-0163.1.

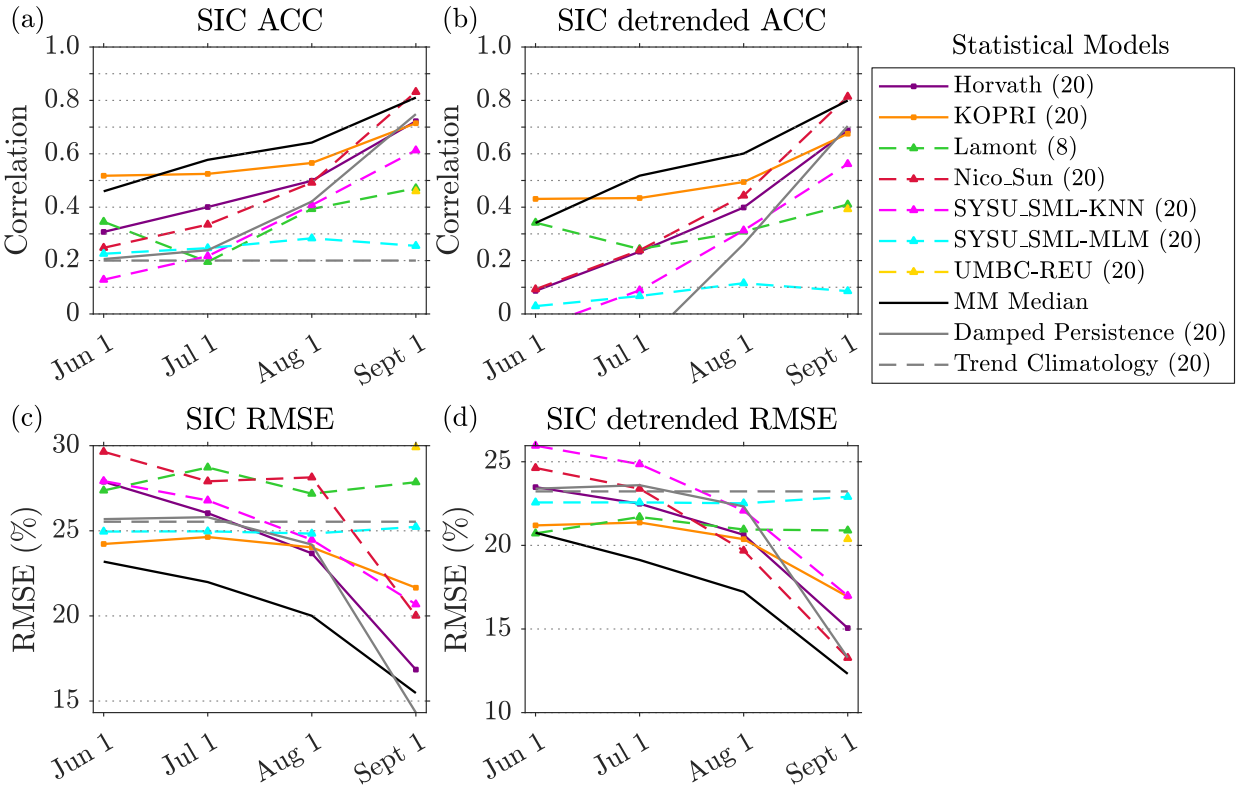

FIG. 11. As in Fig. 10 but for statistical models.

the probabilistic version of the IIEE, i.e. the spatial probability score of Goessling and Jung (2018)), since this study focuses on deterministic (non-ensemble) sea ice predictions. The IIEE for dynamical and statistical models is shown in Fig. 12. Note that no detrending or additional bias correction has been applied in computing the IIEE metrics. Relative to the trend climatology prediction, we find that most models are skillful when initialized on August 1 and September 1. Approximately half of the models outperform the trend climatology from July 1, and most models lose to this benchmark from June 1. The multi-model median ice edge prediction is skillful at SIO lead times relative to both damped persistence and the trend climatology prediction. The median prediction is more skillful than the individual model predictions, with the exception of the KOPRI model which maintains low IIEE at June 1 and July 1 initialization dates.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0163.1.

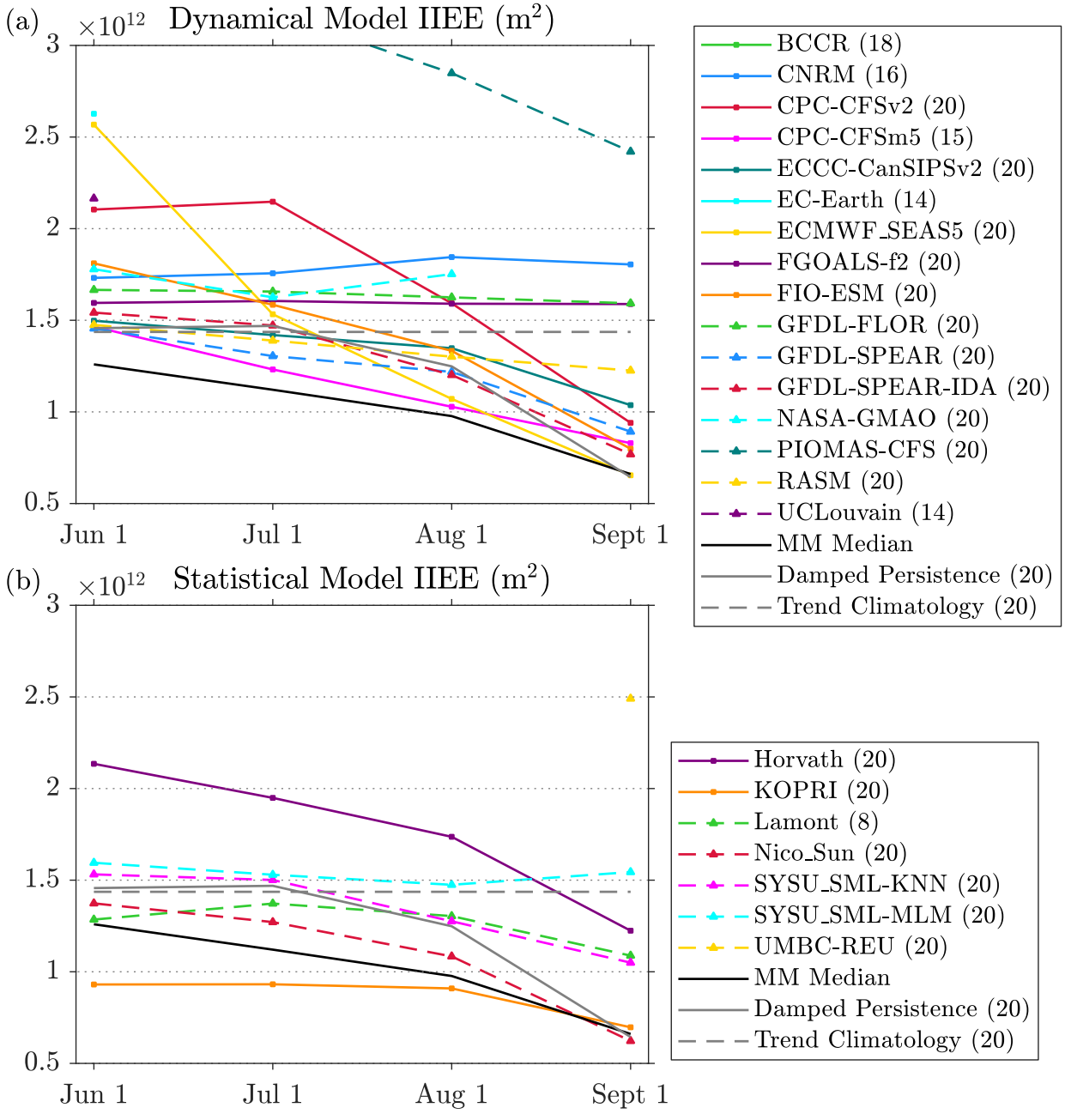

FIG. 12. Integrated ice edge error for September SIC predictions for dynamical and statistical models over the period 2001–2020. Individual models are shown in colors, multi-model predictions are shown in black, and reference predictions are shown in grey. Skill metrics are plotted for each available initialization time (June 1–September 1). The numbers in parentheses in the legend indicate the number of years available from each model over the 2001–2020 time period.

Accepted for publication in *Bulletin of the American Meteorological Society*. DOI 10.1175/BAMS-D-23-0163.1.

6. Discussion

a. Elements of Successful Sea Ice Prediction Systems

In addition to quantifying and comparing prediction skill across models, the retrospective prediction dataset can also be used to understand the key system design choices that underpin skillful sea ice predictions. This "meta-analysis" can allow members of the sea ice prediction community to learn from one another, and incorporate these lessons into development of their own prediction systems. We consider the average pan-Arctic SIE prediction skill of models grouped according to various system design choices. These results should be viewed with some caution, given the relatively small number of models available for each group and the possibility of other confounding factors contributing to skill differences. Nevertheless, this unique dataset can offer insights into the key factors that determine skill differences between models. We first discuss aspects of the dynamical prediction systems and follow with a discussion of the statistical systems.

Figure 13 shows the averaged pan-Arctic SIE prediction skill of dynamical models grouped according to their initialization data and their ice-ocean and atmospheric horizontal resolution. Consistent with earlier work assessing the impact of SIC data assimilation on seasonal prediction skill (e.g., Zhang et al. (2022)), we find that the models that assimilate SIC have superior skill from September 1, and that SIC assimilation has less of an impact from June 1, July 1, and August 1 (Fig. 13a). Similarly, the models that assimilate SST have superior September 1 skill, likely because SST assimilation provides a strong constraint on the sea ice edge position and also provides predictability for the ice growth that occurs during the latter half of September (Fig. 13b). The SST-assimilating models do not show a clear difference for July and August predictions, and have insignificant differences from June 1. Only two models assimilate SIT information, which precludes us from analyzing the impact of SIT data assimilation on seasonal prediction skill.

It is commonly suggested that high-resolution dynamical models should be more skillful than their low-resolution counterparts (e.g. Vecchi et al. (2014); Prodhomme et al. (2016); Kirtman et al. (2017)), but this has not been carefully demonstrated for sea ice prediction applications before. We find that the models with higher ice-ocean resolution (defined here as ice-ocean grid spacing less than 0.4°) have higher skill than the low-resolution models at all SIO lead times (Fig. 13c), suggesting that there is indeed value to using higher-resolution prediction systems. It is important

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0163.1.

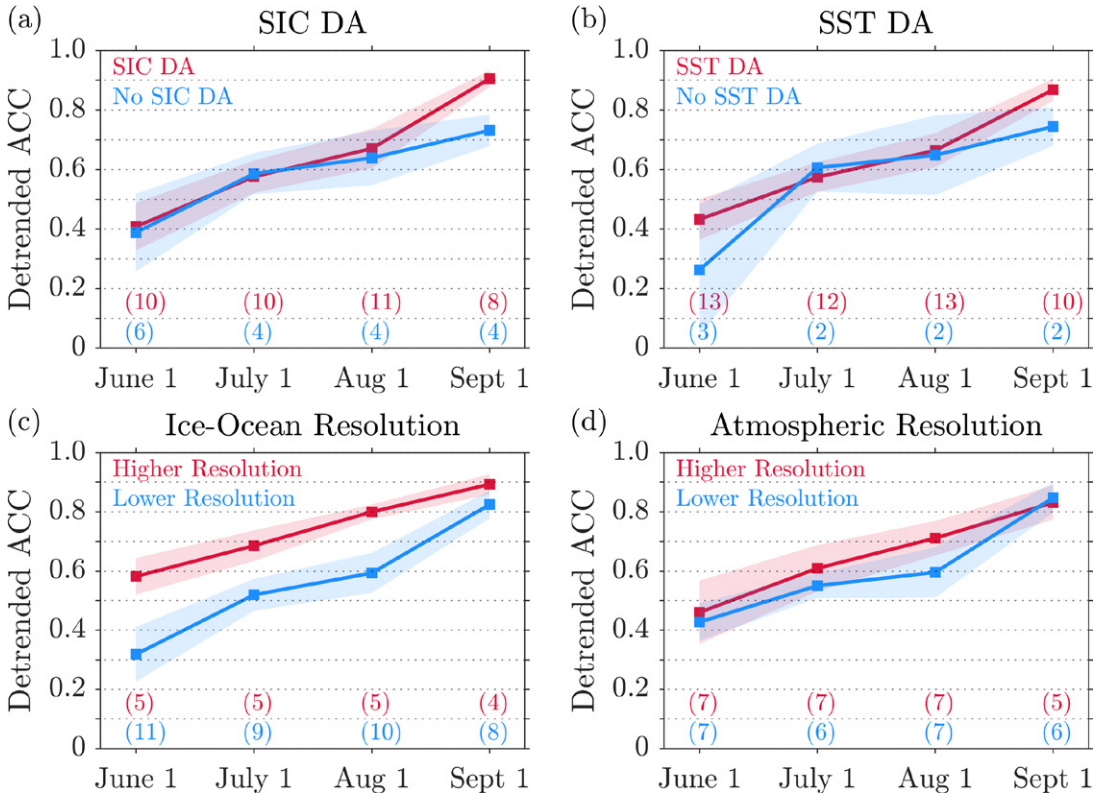

FIG. 13. Average detrended Pan-Arctic ACC for dynamical models grouped according to various system design choices: models that include SIC data assimilation (a), SST data assimilation (b), high versus low ice-ocean horizontal resolution (c), and high versus low atmospheric horizontal resolution (d). The numbers in parentheses indicate the number of models in each group. Shading indicates 68% confidence intervals based on bootstrapped distributions of 1000 realizations.

to note that this finding is based on a small set of models and there could be other confounding factors at play; for example, modeling groups capable of running high-resolution predictions tend to be better resourced and may have also placed additional focus on other aspects of their prediction systems. The impact of higher horizontal atmospheric resolution (defined here as atmospheric grid spacing less than 0.6°) is smaller than that of ice-ocean resolution (Fig. 13d). The higher atmospheric resolution models have higher skill from July 1 and August 1, but the differences are not significant, and they show similar skill to the lower resolution models at other initialization times.

In terms of sea ice physics, the majority of models use an (elastic) viscous plastic rheology (Hibler 1979; Hunke and Dukowicz 1997) and include a prognostic ice-thickness distribution (Bitz et al.

Accepted for publication in *Bulletin of the American Meteorological Society*. DOI 10.1175/BAMS-D-23-0163.1.

2001). Interestingly, the most skillful dynamical model–ECMWF SEAS5–uses a relatively simple sea ice physics formulation based on the Louvain-la-Neuve sea-ice model version 2 (LIM2; Fichefet and Maqueda (1997)), which uses a single thickness category and does not include prognostic melt ponds. The strong performance of ECMWF SEAS5 suggests that sea ice physics complexity may not be of leading-order importance for seasonal sea ice predictions.

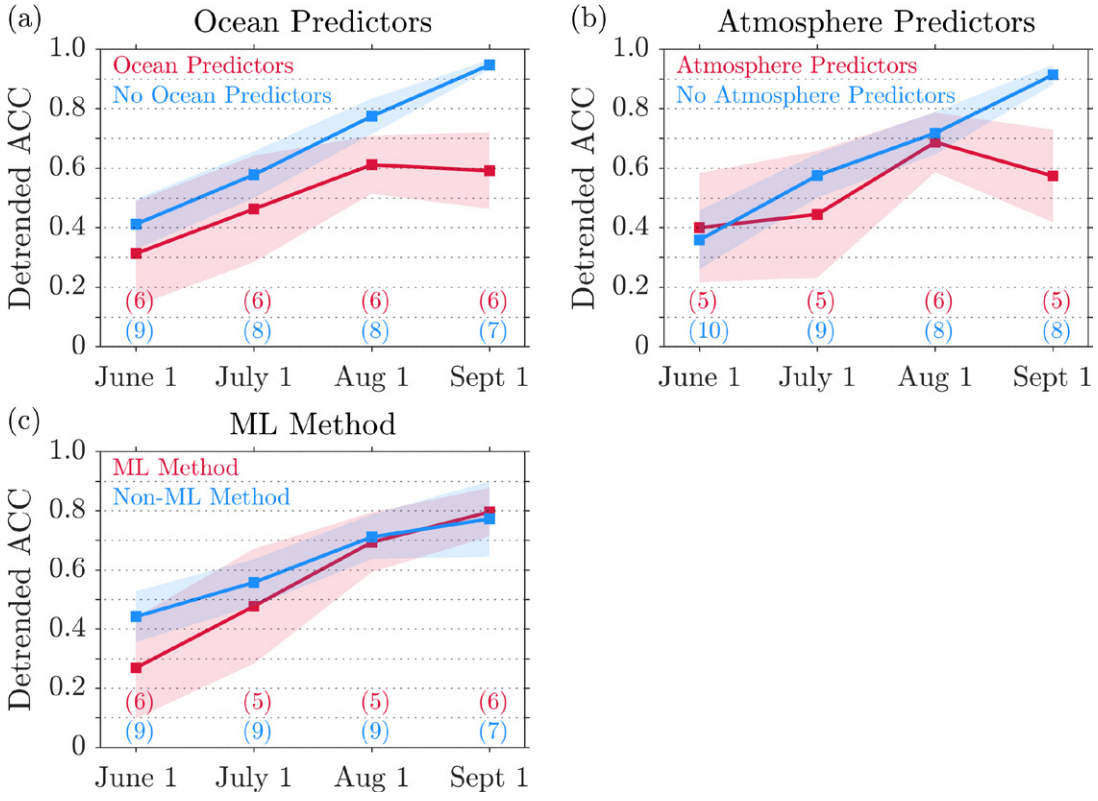

FIG. 14. Average detrended Pan-Arctic ACC for statistical models grouped according to various system design choices: models that include ocean predictors (a), atmosphere predictors (b), and machine learning versus non-machine learning methods (c). The numbers in parentheses indicate the number of models in each group. Shading indicates 68% confidence intervals based on bootstrapped distributions of 1000 realizations.

Figure 14 shows the prediction skill of statistical models grouped according to their predictor/input data and their prediction methodology. All statistical models include a sea ice predictor variable (see Table 3) and a number of models also include predictors from the ocean and atmosphere. Figure 14a–b examines the impact of these oceanic and atmospheric predictors. Unexpectedly, we find that the models that withhold ocean and atmosphere predictors tend to outperform the models that include these predictors. The statistical models without ocean predictors show

Accepted for publication in *Bulletin of the American Meteorological Society*. DOI 10.1175/BAMS-D-23-0163.1.

higher skill at all SIO lead times, and the models without atmospheric predictors have higher skill from July 1 and September 1. We note that a handful of the statistical models that include ocean and atmosphere predictors have quite low skill generally. This affects the average skill shown in Fig. 14, thus this result should be viewed with some caution.

While these findings are somewhat contrary to our understanding of seasonal sea ice predictability, the inclusion of additional physically-relevant predictor variables within a given statistical model may actually degrade predictions due to the 'curse of dimensionality' problem (e.g., Bellman and Kalaba (1959)), which can lead to overfitting. Indeed, although it is possible to mitigate these effects with regularization techniques and/or sufficient training data samples, the former option may not be available to many of the statistical models submitted to this inter-comparison, and overcoming the limitations of typically fewer than 43 years (samples) of training data is potentially only achievable with more sophisticated data augmentation approaches, such as transfer learning (e.g., Andersson et al. (2021)).

There has been a recent proliferation of machine learning (ML) methods applied to seasonal prediction problems, including for Arctic sea ice. How do the skill of these ML methods compare to other statistical techniques? Figure 14c shows this comparison, and we find that there is no clear skill advantage for ML-based prediction techniques. The skill of ML and Non-ML methods is similar for August and September initialization dates, whereas the Non-ML methods have slightly higher skill for June and July initializations, but the differences are not significant. We note that the classification of "ML methods" is somewhat equivocal. Here, our classification of ML methods is based on self-identification by seven models, and includes methods based on convolutional neural networks, Gaussian process regression, long short-term memory networks, the k-nearest neighbors algorithm, and data adaptive harmonic decomposition. Of the top five performing statistical models at each SIO lead time, there is a roughly equal split between ML and non-ML models, suggesting that one class of methods is not clearly superior to the other.

b. Differences between Real-Time and Retrospective Prediction Skill

Consistent with the tension in the sea ice prediction literature discussed in the Introduction, the retrospective predictions analyzed in this paper have higher skill than the real-time predictions submitted to the SIO. Why is this? This skill difference could potentially result from a number

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0163.1.

of interacting factors, which we describe in more detail below: i) Different evaluation periods; ii) Non-stationarity of prediction systems; iii) Model selection bias; iv) Bias correction; v) Systematic differences between real-time and retrospective prediction methodologies; vi) Implicit or erroneous use of future data; and vii) Changes in inherent predictability.

Regarding the different evaluation periods (factor i), we confirm that the multi-model median skill difference is still present when skill is evaluated over the period of 2008–2021, which is common to both the retrospective and real-time SIO predictions (see Figs. S1, S7, S8). A natural question is: Can one directly compare retrospective and real-time SIO forecasts submitted from the same system? Such a comparison is difficult to make due to the non-stationarity of prediction systems (factor ii). In particular, many groups have updated and improved their systems during the SIO period. These groups have submitted their best-performing system to the SIO retrospective skill comparison, which in some cases differ from the SIO predictions that were submitted in real-time. Also, while many of the groups have submitted predictions to the real-time SIO at some point, few groups have submitted predictions using a "frozen" system over a sustained period of more than 5 years. These factors complicate the comparison between real time and retrospective predictions.

Selection bias (factor iii) may also play a role in the skill difference. Given that submission to the retrospective comparison is voluntary, the skill scores in this paper may be biased towards better performing models. Also, models with retrospective prediction suites have likely used this retrospective skill information to assess and improve their prediction systems. Knowing the skill of one's method is not a requirement to submit to the real-time SIO, which may contribute to poorer performance of real-time SIO predictions. Access to a retrospective prediction suite also allows for a quantification of a model's lead-dependent prediction bias, which allows for more effective bias correction. This may contribute to the higher skill found in models that submitted retrospective forecasts (factor iv). Figure S1 shows that the main differences in multi-model median prediction errors occur over the early portion of the SIO period (from 2008–2014). The SIO received fewer annual submissions over this period, including a notable fraction from heuristic methods (see Fig. S2), which may have degraded the skill of the real-time SIO median prediction relative to the more recent SIO period (2015–2021).

It is well known that retrospective skill of seasonal-to-interannual climate predictions tends to be higher than real-time skill (e.g., Goddard et al. (2013); Risbey et al. (2021)), and this is likely

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0163.1.

a contributor to the real-time/retrospective SIO skill differences (factor v). These skill differences can arise due to certain observations being available for retrospective forecasts but not in real-time (e.g. subsurface T/S ocean profiles collected from ships) or due to real-time predictions relying on real-time reanalyses and satellite data that has not been rigorously quality controlled. Skill differences can also arise due to the implicit use of future data in retrospective predictions, which may unrealistically inflate their skill scores (factor vi). For example, dynamical models are often tuned to match the climatology and trends of the observational record. Statistical models need to take care to not include future data in their training procedure (e.g., computing EOFs using only past data), which is often called the "data leakage" problem in machine learning methods (Kapoor and Narayanan 2023). Also, both dynamical and statistical prediction systems are constructed based on their performance predicting past anomalies, which may result in an "overfitting" to the observational record. Additionally, standard bias correction approaches (e.g., Manzanas et al. (2019)) require computing anomalies relative to a climatology which is typically computed using the full record. This approach implicitly uses future data and may artificially increase skill (Risbey et al. 2021).

Finally, skill differences between real-time and retrospective predictions could arise due to changes in inherent sea ice predictability if the real-time SIO period had inherently lower predictability (factor vii). The earlier analysis shown in Fig. 5a suggests that prediction errors have not changed substantially over the SIO period, thus factor vii is unlikely a dominant contributor to the skill differences. In summation, our analysis suggests that the most likely contributors to the real-time versus retrospective skill differences are a combination of model selection bias, bias correction differences, and systematic differences between real-time and retrospective predictions. The skill gaps associated with model selection and bias correction could be addressed fairly straightforwardly, by using only the subset of prediction systems with proven skill and by ensuring that these systems utilize retrospective predictions to bias correct their forecasts. The skill gaps associated with systematic differences between real-time and retrospective predictions are more challenging to address, as these require modifications to the observing network and may also be influenced by inherent biases present in the prediction system development process.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0163.1.

7. Conclusions and Future Outlook

This work has produced and analyzed a novel multi-model retrospective seasonal sea ice prediction dataset, consisting of community contributions from 17 statistical models, 17 dynamical models, and 1 heuristic prediction. The majority of contributing models provided retrospective predictions of September Arctic sea ice initialized on the Sea Ice Outlook (SIO) initialization dates of June 1, July 1, August 1, and September 1, spanning a minimum period of 2001–2020 (see Table 1). The statistical and dynamical model submissions employ a wide range of prediction methodologies ranging from linear regression, to Markov models, to deep learning techniques, to coupled regional and global models with data assimilation (see Tables 2 and 3).

Our overarching key finding is that this diverse set of seasonal prediction models can skillfully predict September Arctic sea ice at SIO lead times on Pan-Arctic, regional, and local scales. These results demonstrate that useful real-time multi-month predictions of September sea ice are likely within reach in the coming years. We have shown that the majority of models have detrended ACC values for Pan-Arctic sea ice extent (SIE) which exceed 0.5 at SIO lead times, and that the multi-model median prediction exceeds 0.75 at SIO lead times. Regional SIE skill is similar to Pan-Arctic SIE skill in the Alaskan and Siberian regions, whereas skill is lower in the Atlantic, Canadian, and Central Arctic regions. The multi-model median detrended ACC exceeds 0.75 at all SIO lead times in the Alaskan and Siberian regions, exceeds 0.6 in the Canadian sector, and falls below 0.5 for certain lead times in the Atlantic and Central Arctic. We have found that the regional skill differences across models are related to their Pan-Arctic skill differences, especially in the Alaskan, Siberian, and Central Arctic regions. The models also have skill in predicting local sea ice concentration (SIC), however this local quantity is much more challenging to predict than Pan-Arctic or regional SIE. The multi-model median SIC prediction has detrended ACC values above 0.5 from July 1–September 1 and below 0.4 from June 1. The high skill of the Pan-Arctic and regional multi-model median predictions is slightly lower than lead 0–3 month predictions of the winter El Nino Southern Oscillation index, which have ACC skill above 0.9 for a number of individual models (see Fig. 7 of Barnston et al. (2012)).

We have investigated whether sea ice prediction errors have changed over time, and determined that there are no statistically significant trends in prediction errors over the period since 1990. This suggests, but does not prove, that there has been minimal change in inherent sea ice predictability

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0163.1.

over this period. We have found that models generally exhibit their largest errors in extreme sea ice years (i.e., 1996, 2007, 2012), however they also provide the most "added value" over baseline trend climatology and damped anomaly persistence forecasts in these years. This finding challenges the interpretation of earlier studies which stated that prediction systems perform particularly poorly in extreme SIE years (e.g., Stroeve et al. (2014)).

We have found that the skill of dynamical and statistical models is generally comparable for Pan-Arctic SIE, whereas dynamical models tend to outperform their statistical counterparts when evaluated on the regional and local scale. It is important to note that there are individual statistical models that have high levels of skill for both regional SIE and local SIC, which are competitive with the most skillful dynamical models. Analysis of the design aspects of the dynamical prediction systems revealed higher skill in models that i) assimilate SIC; ii) assimilate sea surface temperature; iii) use higher ice-ocean horizontal resolution (finer that 0.4°), and iv) use higher atmospheric horizontal resolution. A similar analysis of the statistical prediction systems revealed skill degradation in models that i) include ocean predictors and ii) include atmospheric predictors, potentially associated with overfitting. We also found that statistical models based on machine learning methods had no clear skill advantage over other statistical techniques. The retrospective predictions evaluated in this study have higher prediction skill than real-time predictions submitted to the SIO. We speculate that these skill differences result from a number of interacting factors, with the most likely contributors being model selection bias, bias correction techniques, and systematic differences between real-time and retrospective prediction methodologies.

This study demonstrates that there are bright prospects for skillful seasonal predictions of Arctic sea ice made using both dynamical and statistical prediction models. We anticipate that the multi-model retrospective prediction dataset produced by this study will motivate additional future research by the sea ice prediction community. Natural future directions include assessment of probabilistic forecast skill using the submitted ensemble predictions, comparison of initial conditions across prediction systems and their relation to prediction skill, analysis of the mechanisms of predictability being captured by different systems, exploration of the role of subseasonal-to-seasonal atmospheric prediction skill in determining sea ice skill, analysis of forecast errors, understanding the importance of prediction system design choices, and the construction of a "consensus" real-time SIO prediction based on a skill-weighted multi-model mean. We also hope that the findings of

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0163.1.

this study motivate new targeted experiments and development efforts within individual sea ice prediction systems. This study has focused on September sea ice predictions, but similar skill intercomparisons are required for other months of the year, particularly winter freeze-up months which are characterized by very different predictability mechanisms. Another route for future investigation is a comparison of Arctic and Antarctic sea ice prediction skill, making use of the multi-model SIPN-South seasonal prediction dataset (Massonnet et al. 2023). The past 15 years have featured many breakthroughs in the field of sea ice prediction and predictability. Community intercomparison of sea ice prediction systems, combined with new observations, improved coupled models, new statistical techniques, deepened stakeholder input, improved dissemination of forecast products, and theoretical predictability research, provide key pathways for continuing to advance this field over the coming decade.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0163.1.

Acknowledgments. We acknowledge the community building efforts of the Sea Ice Prediction Network and the Sea Ice Outlook, which were supported by the National Science Foundation (PLR-1303938; OPP-1748308; OPP-1749081; OPP-1751363; OPP-1748953; OPP-1748325; OPP-1331083) and the Office of Naval Research (N00014-13-1-0793). JS was supported by NSFGEO-NERC Advancing Predictability of Sea Ice: Phase 2 of the Sea Ice Prediction Network (SIPN2) NE/R017123/1. EB-W acknowledges support from NSF grant OPP-1751363. SA and JW acknowledge the support from National Science Foundation (OAC-1942714). Yiguo Wang acknowledges the Norges Forskningsrad (Grant No. 328886) and the Trond Mohn stiftelse (Grant No. BFS2018TMT01). QY, XL, YL and YW acknowledge the National Key R&D Program of China (No. 2022YFE0106300), the National Natural Science Foundation of China (No. 42106233). EB was supported by the Met Office Hadley Centre Climate Programme funded by DSIT. ZL acknowledges support under CIMES award NA18OAR4320123. FM and this project received funding from the BELSPO project RESIST. We thank Mike Winton and Andrew Ross for helpful comments on a preliminary draft of this manuscript.

Data availability statement. Retrospective prediction data for all models and code to process and analyze data and make figures are available via an online repository (https://zenodo.org/doi/10.5281/zenodo.10124346). The NSIDC sea ice index version 3 is available from https://nsidc.org/data/seaice_index/. The OSI SAF sea ice index v2.1 is available from https://osi-saf.eumetsat.int/products/osi-420. The NSIDC CDR SIC data is available from https://nsidc.org/data/g02202. The OSI SAF SIC CDR data is available from https://osi-saf.eumetsat.int/products/osi-450-a.

References

Ali, S., Y. Huang, X. Huang, and J. Wang, 2021: Sea ice forecasting using attention-based ensemble lstm. *International Conference on Machine Learning (ICML)*, https://doi.org/https://www.climatechange.ai/papers/icml2021/50.

Andersson, T. R., and Coauthors, 2021: Seasonal Arctic sea ice forecasting with probabilistic deep learning. *Nature communications*, **12 (1)**, 1–12.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0163.1.

Babb, D., J. Landy, D. Barber, and R. Galley, 2019: Winter sea ice export from the Beaufort Sea as a preconditioning mechanism for enhanced summer melt: A case study of 2016. *J. Geophys. Res.: Oceans*, **124 (9)**, 6575–6600.

Babb, D., J. Landy, J. Lukovich, C. Haas, S. Hendricks, D. Barber, and R. Galley, 2020: The 2017 reversal of the Beaufort Gyre: Can dynamic thickening of a seasonal ice cover during a reversal limit summer ice melt in the Beaufort Sea? *J. Geophys. Res.: Oceans*, **125 (12)**, e2020JC016 796.

Balan-Sarojini, B., S. Tietsche, M. Mayer, M. Balmaseda, H. Zuo, P. De Rosnay, T. Stockdale, and F. Vitart, 2021: Year-round impact of winter sea ice thickness observations on seasonal forecasts. *The Cryosphere*, **15 (1)**, 325–344.

Barnston, A. G., M. K. Tippett, M. L. L'Heureux, S. Li, and D. G. DeWitt, 2012: Skill of real-time seasonal ENSO model predictions during 2002–11: Is our capability increasing? *Bulletin of the American Meteorological Society*, **93 (5)**, 631–651.

Barthélemy, A., H. Goosse, T. Fichefet, and O. Lecomte, 2018: On the sensitivity of Antarctic sea ice model biases to atmospheric forcing uncertainties. *Climate Dynamics*, **51**, 1585–1603.

Batté, L., I. Välisuo, M. Chevallier, J. C. A. Navarro, P. Ortega, and D. Smith, 2020: Summer predictions of Arctic sea ice edge in multi-model seasonal re-forecasts. *Clim. Dyn.*, **54 (11)**, 5013–5029.

Baxter, I., and Q. Ding, 2022: An optimal atmospheric circulation mode in the Arctic favoring strong summertime sea ice melting and ice-albedo feedback. *Journal of Climate*, **35 (20)**, 6627–6645.

Baxter, I., and Coauthors, 2019: How tropical Pacific surface cooling contributed to accelerated sea ice melt from 2007 to 2012 as ice is thinned by anthropogenic forcing. *Journal of Climate*, **32 (24)**, 8583–8602.

Bellman, R., and R. Kalaba, 1959: On adaptive control processes. *IRE Transactions on Automatic Control*, **4 (2)**, 1–9.

Bitz, C., M. Holland, A. Weaver, and M. Eby, 2001: Simulating the ice-thickness distribution in a coupled climate model. *J. Geophys. Res.: Oceans*, **106 (C2)**, 2441–2463.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0163.1.

Blanchard-Wrigglesworth, E., K. C. Armour, C. M. Bitz, and E. DeWeaver, 2011a: Persistence and inherent predictability of Arctic sea ice in a GCM ensemble and observations. *J. Climate*, **24**, 231–250.

Blanchard-Wrigglesworth, E., C. Bitz, and M. Holland, 2011b: Influence of initial conditions and climate forcing on predicting Arctic sea ice. *Geophys. Res. Lett.*, **38 (18)**.

Blanchard-Wrigglesworth, E., and C. M. Bitz, 2014: Characteristics of Arctic sea-ice thickness variability in GCMs. *J. Climate*, **27 (21)**, 8244–8258.

Blanchard-Wrigglesworth, E., and M. Bushuk, 2019: Robustness of Arctic sea-ice predictability in GCMs. *Clim. Dyn.*, **52 (9)**, 5555–5566.

Blanchard-Wrigglesworth, E., M. Bushuk, F. Massonnet, L. C. Hamilton, C. M. Bitz, W. N. Meier, and U. S. Bhatt, 2023: Forecast skill of the Arctic Sea Ice Outlook 2008–2022. *Geophys. Res. Lett*, **50 (6)**, e2022GL102 531.

Blanchard-Wrigglesworth, E., R. Cullather, W. Wang, J. Zhang, and C. Bitz, 2015: Model forecast skill and sensitivity to initial conditions in the seasonal Sea Ice Outlook. *Geophys. Res. Lett*, **42 (19)**, 8042–8048.

Blanchard-Wrigglesworth, E., and Coauthors, 2017: Multi-model seasonal forecast of Arctic sea-ice: forecast uncertainty at pan-Arctic and regional scales. *Clim. Dyn.*, **49 (4)**, 1399–1410.

Blockley, E., and Coauthors, 2014: Recent development of the Met Office operational ocean forecasting system: an overview and assessment of the new global FOAM forecasts. *Geoscientific Model Development*, **7 (6)**, 2613–2638.

Blockley, E. W., and K. A. Peterson, 2018: Improving Met Office seasonal predictions of Arctic sea ice using assimilation of CryoSat-2 thickness. *The Cryosphere*, **12 (11)**, 3419–3438.

Bonan, D., M. Bushuk, and M. Winton, 2019: A spring barrier for regional predictions of summer Arctic sea ice. *Geophys. Res. Lett.*, **46**, 5937–5947.

Brunette, C., B. Tremblay, and R. Newton, 2019: Winter coastal divergence as a predictor for the minimum sea ice extent in the Laptev Sea. *J. Climate*, **32 (4)**, 1063–1080.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0163.1.

Bushuk, M., and D. Giannakis, 2015: Sea-ice reemergence in a model hierarchy. *Geophys. Res. Lett.*, **42**, 5337–5345.

Bushuk, M., R. Msadek, M. Winton, G. Vecchi, R. Gudgel, A. Rosati, and X. Yang, 2017a: Skillful regional prediction of Arctic sea ice on seasonal timescales. *Geophys. Res. Lett.*, **44**, 4953–4964.

Bushuk, M., R. Msadek, M. Winton, G. Vecchi, R. Gudgel, A. Rosati, and X. Yang, 2017b: Summer enhancement of Arctic sea-ice volume anomalies in the September-ice zone. *J. Climate*, **30**, 2341–2362.

Bushuk, M., R. Msadek, M. Winton, G. Vecchi, X. Yang, A. Rosati, and R. Gudgel, 2019: Regional Arctic sea-ice prediction: Potential versus operational seasonal forecast skill. *Clim. Dyn.*, **52 (5)**, 2721–2743.

Bushuk, M., M. Winton, D. B. Bonan, E. Blanchard-Wrigglesworth, and T. Delworth, 2020: A mechanism for the Arctic sea ice spring predictability barrier. *Geophys. Res. Lett.*, 1–13, https://doi.org/e2020GL088335.

Bushuk, M., and Coauthors, 2022: Mechanisms of regional Arctic sea ice predictability in two dynamical seasonal forecast systems. *J. Climate*, **35 (13)**, 4207–4231.

Cassano, J. J., and Coauthors, 2017: Development of the Regional Arctic System Model (RASM): Near-surface atmospheric climate sensitivity. *Journal of Climate*, **30 (15)**, 5729–5753.

Chekroun, M. D., and D. Kondrashov, 2017: Data-adaptive harmonic spectra and multilayer Stuart-Landau models. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, **27 (9)**, 093 110, https://doi.org/10.1063/1.4989400.

Chen, H., X. Yin, Y. Bao, and F. Qiao, 2016: Ocean satellite data assimilation experiments in FIO-ESM using ensemble adjustment Kalman filter. *Science China Earth Sciences*, **59 (3)**, 484–494.

Cheng, W., E. Blanchard-Wrigglesworth, C. M. Bitz, C. Ladd, and P. J. Stabeno, 2016: Diagnostic sea ice predictability in the pan-Arctic and US Arctic regional seas. *Geophys. Res. Lett*, **43 (22)**.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0163.1.

Chevallier, M., and D. Salas y Mélia, 2012: 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.

Chevallier, M., D. Salas y Mélia, A. Voldoire, M. Déqué, and G. Garric, 2013: Seasonal forecasts of the pan-Arctic sea ice extent using a GCM-based seasonal prediction system. *J. Climate*, **26 (16)**, 6092–6104.

Chi, J., J. Bae, and Y.-J. Kwon, 2021: Two-stream convolutional long-and short-term memory model using perceptual loss for sequence-to-sequence Arctic sea ice prediction. *Remote Sensing*, **13 (17)**, 3413.

Chi, J., and H.-c. Kim, 2017: Prediction of Arctic sea ice concentration using a fully data driven deep neural network. *Remote Sensing*, **9 (12)**, 1305.

Collow, T. W., Y. Liu, W. Wang, A. Kumar, and D. DeWitt, 2019: Develop improved seasonal and week 3/4 sea ice outlook. *NOAA's Climate Prediction Center Stakeholder Meeting*, 39.

Collow, T. W., W. Wang, A. Kumar, and J. Zhang, 2015: Improving Arctic sea ice prediction using PIOMAS initial sea ice thickness in a coupled ocean–atmosphere model. *Mon. Wea. Rev.*, **143 (11)**, 4618–4630.

Cox, C. J., T. Uttal, C. N. Long, M. D. Shupe, R. S. Stone, and S. Starkweather, 2016: The role of springtime Arctic clouds in determining autumn sea ice extent. *J. Climate*, **29 (18)**, 6581–6596.

Day, J., E. Hawkins, and S. Tietsche, 2014: Will Arctic sea ice thickness initialization improve seasonal forecast skill? *Geophys. Res. Lett.*, **41 (21)**, 7566–7575.

DelSole, T., J. Nattala, and M. K. Tippett, 2014: Skill improvement from increased ensemble size and model diversity. *Geophys. Res. Lett.*, **41 (20)**, 7331–7342.

Diebold, F., and M. Göbel, 2022: A benchmark model for fixed-target Arctic sea ice forecasting. *Economics Letters*, **215**, 110 478.

Diebold, F., M. Göbel, and P. Goulet Coulombe, 2023: Assessing and comparing fixed-target forecasts of Arctic sea ice: glide charts for feature-engineered linear regression and machine learning models. *Energy Economics*, **124**, 106 833.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0163.1.

Ding, Q., and Coauthors, 2017: Influence of high-latitude atmospheric circulation changes on summertime Arctic sea ice. *Nature Climate Change*, **7 (4)**, 289–295.

Ding, Q., and Coauthors, 2019: Fingerprints of internal drivers of Arctic sea ice loss in observations and model simulations. *Nat. Geosci.*, **12 (1)**, 28.

Dirkson, A., B. Denis, and W. Merryfield, 2019: A multimodel approach for improving seasonal probabilistic forecasts of regional Arctic sea ice. *Geophys. Res. Lett.*, **46 (19)**, 10 844–10 853.

Dirkson, A., W. J. Merryfield, and A. Monahan, 2017: Impacts of sea ice thickness initialization on seasonal Arctic sea ice predictions. *J. Climate*, **30 (3)**, 1001–1017.

Drobot, S. D., J. A. Maslanik, and C. Fowler, 2006: A long-range forecast of Arctic summer sea-ice minimum extent. *Geophys. Res. Lett.*, **33 (10)**.

EUMETSAT Ocean and Sea Ice Satellite Application Facility, 2022: OSI SAF Global sea ice concentration climate data record 1978-2020 (v3.0, 2022). Data extracted from OSI SAF FTP server: (1979-2020,) (Arctic,) accessed March 27, 2023, https://doi.org/10.15770/EUM_SAF_OSI_0013.

Fetterer, F., K. Knowles, W. N. Meier, M. Savoie, and A. K. Windnagel., 2017: Sea ice index, version 3. National Snow and Ice Data Center, URL https://nsidc.org/data/G02135/versions/3, https://doi.org/10.7265/N5K072F8.

Fichefet, T., and M. M. Maqueda, 1997: Sensitivity of a global sea ice model to the treatment of ice thermodynamics and dynamics. *Journal of Geophysical Research: Oceans*, **102 (C6)**, 12 609–12 646.

Finocchio, P. M., J. D. Doyle, and D. P. Stern, 2022: Accelerated sea ice loss from late summer cyclones in the new Arctic. *Journal of Climate*, **35 (23)**, 7751–7769.

Giesse, C., D. Notz, and J. Baehr, 2021: On the origin of discrepancies between observed and simulated memory of Arctic sea ice. *Geophys. Res. Lett.*, **48 (11)**, e2020GL091 784.

Goddard, L., and Coauthors, 2013: A verification framework for interannual-to-decadal predictions experiments. *Climate Dynamics*, **40**, 245–272.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0163.1.

Goessling, H., and T. Jung, 2018: A probabilistic verification score for contours: Methodology and application to Arctic ice-edge forecasts. *Quarterly Journal of the Royal Meteorological Society*, **144 (712)**, 735–743.

Goessling, H. F., S. Tietsche, J. J. Day, E. Hawkins, and T. Jung, 2016: Predictability of the Arctic sea ice edge. *Geophys. Res. Lett.*, **43 (4)**, 1642–1650.

Gregory, W., M. Tsamados, J. Stroeve, and P. Sollich, 2020: Regional september sea ice forecasting with complex networks and Gaussian processes. *Weather and Forecasting*, **35 (3)**, 793–806.

Guemas, V., M. Chevallier, M. Déqué, O. Bellprat, and F. Doblas-Reyes, 2016: Impact of sea ice initialisation on sea ice and atmosphere prediction skill on seasonal timescales. *Geophys. Res. Lett*, **43 (8)**, 3889–3896.

Hagedorn, R., F. J. Doblas-Reyes, and T. N. Palmer, 2005: The rationale behind the success of multi-model ensembles in seasonal forecastingÑI. Basic concept. *Tellus A: Dynamic Meteorology and Oceanography*, **57 (3)**, 219–233.

Hamilton, L. C., C. M. Bitz, E. Blanchard-Wrigglesworth, M. Cutler, J. Kay, W. N. Meier, J. Stroeve, and H. Wiggins, 2014: Sea ice prediction has easy and difficult years. *Witness the Arctic*, **18 (GSFC-E-DAA-TN16146)**.

Hamilton, L. C., and J. Stroeve, 2016: 400 predictions: The SEARCH sea ice outlook 2008–2015. *Polar Geography*, **39 (4)**, 274–287.

Harnos, K., M. L'Heureux, Q. Ding, and Q. Zhang, 2019: Skill of seasonal Arctic sea ice extent predictions using the North American Multimodel Ensemble. *J. Climate*, **32 (2)**, 623–638.

Hazeleger, W., and Coauthors, 2012: EC-Earth V2. 2: description and validation of a new seamless Earth system prediction model. *Climate dynamics*, **39**, 2611–2629.

Hibler, W. D., 1979: A dynamic thermodynamic sea ice model. *Journal of physical oceanography*, **9 (4)**, 815–846.

Holland, M. M., D. A. Bailey, and S. Vavrus, 2011: Inherent sea ice predictability in the rapidly changing Arctic environment of the Community Climate System Model, version 3. *Clim. Dyn.*, **36 (7-8)**, 1239–1253.

Accepted for publication in *Bulletin of the American Meteorological Society*. DOI 10.1175/BAMS-D-23-0163.1.

Holland, M. M., L. Landrum, D. Bailey, and S. Vavrus, 2019: Changing seasonal predictability of Arctic summer sea ice area in a warming climate. *J. Climate*, **32 (16)**, 4963–4979.

Holland, M. M., and J. Stroeve, 2011: Changing seasonal sea ice predictor relationships in a changing Arctic climate. *Geophys. Res. Lett*, **38 (18)**.

Horvath, S., J. Stroeve, and B. Rajagopalan, 2021: A linear mixed effects model for seasonal forecasts of Arctic sea ice retreat. *Polar Geography*, **44 (4)**, 297–314.

Hunke, E., and J. Dukowicz, 1997: An elastic-viscous-plastic model for sea ice dynamics. *J. Phys. Oceanogr.*, **27 (9)**, 1849–1867.

Ionita, M., K. Grosfeld, P. Scholz, R. Treffeisen, and G. Lohmann, 2019: September Arctic sea ice minimum prediction–a skillful new statistical approach. *Earth Syst. Dynam.*, **10**, 189–203, https://doi.org/10.5194/esd-10-189-2019.

Johnson, S. J., and Coauthors, 2019: SEAS5: the new ECMWF seasonal forecast system. *Geoscientific Model Development*, **12 (3)**, 1087–1117.

Jung, T., and Coauthors, 2016: Advancing polar prediction capabilities on daily to seasonal time scales. *Bull. Amer. Meteor. Soc.*, **97 (9)**, 1631–1647, https://doi.org/10.1175/BAMS-D-14-00246.1.

Kapoor, S., and A. Narayanan, 2023: Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, **4 (9)**.

Kapsch, M.-L., R. G. Graversen, T. Economou, and M. Tjernström, 2014: The importance of spring atmospheric conditions for predictions of the Arctic summer sea ice extent. *Geophys. Res. Lett.*, **41 (14)**, 5288–5296.

Kay, J. E., T. L'Ecuyer, A. Gettelman, G. Stephens, and C. O'Dell, 2008: The contribution of cloud and radiation anomalies to the 2007 Arctic sea ice extent minimum. *Geophys. Res. Lett.*, **35 (8)**.

Kim, E., P. Kruse, S. Lama, J. Bourne, M. Hu, S. Ali, Y. Huang, and J. Wang, 2021: Multi-task deep learning based spatiotemporal Arctic sea ice forecasting. *IEEE International Conference on Big Data*, https://doi.org/10.1109/BigData52589.2021.9671491.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0163.1.

Kimmritz, M., F. Counillon, L. H. Smedsrud, I. Bethke, N. Keenlyside, F. Ogawa, and Y. Wang, 2019: Impact of ocean and sea ice initialisation on seasonal prediction skill in the Arctic. *J. Adv. Model. Earth Syst.*, **11 (12)**, 4147–4166.

Kirtman, B. P., N. Perlin, and L. Siqueira, 2017: Ocean eddies and climate predictability. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, **27 (12)**.

Koenigk, T., and U. Mikolajewicz, 2009: Seasonal to interannual climate predictability in mid and high northern latitudes in a global coupled model. *Clim. Dyn.*, **32 (6)**, 783–798.

Kondrashov, D., M. D. Chekroun, and M. Ghil, 2018: Data-adaptive harmonic decomposition and prediction of Arctic sea ice extent. *Dynamics and Statistics of the Climate System*, **3 (1)**, dzy001.

Krumpen, T., M. Janout, K. Hodges, R. Gerdes, F. Girard-Ardhuin, J. Hölemann, and S. Willmes, 2013: Variability and trends in Laptev Sea ice outflow between 1992–2011. *The Cryosphere*, **7**, 1–15.

Kwok, R., 2008: Summer sea ice motion from the 18 GHz channel of AMSR-E and the exchange of sea ice between the Pacific and Atlantic sectors. *Geophys. Res. Lett.*, **35 (3)**.

Kwok, R., G. Cunningham, and T. Armitage, 2018: Relationship between specular returns in CryoSat-2 data, surface albedo, and Arctic summer minimum ice extent. *Elem. Sci. Anth.*, **6 (1)**, 1–10.

Landy, J. C., J. K. Ehn, and D. G. Barber, 2015: Albedo feedback enhanced by smoother Arctic sea ice. *Geophys. Res. Lett.*, **42 (24)**, 10–714.

Lavergne, T., and Coauthors, 2019: Version 2 of the EUMETSAT OSI SAF and ESA CCI sea-ice concentration climate data records. *The Cryosphere*, **13 (1)**, 49–78.

Lenetsky, J. E., B. Tremblay, C. Brunette, and G. Meneghello, 2021: Subseasonal predictability of Arctic ocean sea ice conditions: Bering strait and Ekman-driven ocean heat transport. *J. Climate*, **34 (11)**, 4449–4462.

Li, J., and Coauthors, 2021: Dynamical seasonal prediction of tropical cyclone activity using the FGOALS-f2 ensemble prediction system. *Weather and Forecasting*, **36 (5)**, 1759–1778.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0163.1.

Lin, H., and Coauthors, 2020: The Canadian seasonal to interannual prediction system version 2 (CanSIPSv2). *Weather and Forecasting*, **35 (4)**, 1317–1343.

Lin, Y., and Coauthors, 2023: Optimization of the k-nearest-neighbors model for summer Arctic sea ice prediction. *Frontiers in Marine Science*, **10**, 1260 047, https://doi.org/10.3389/fmars.2023.1260047.

Lindsay, R., J. Zhang, A. Schweiger, and M. Steele, 2008: Seasonal predictions of ice extent in the Arctic Ocean. *J. Geophys. Res.: Oceans*, **113 (C2)**.

Liu, J., M. Song, R. M. Horton, and Y. Hu, 2015: Revisiting the potential of melt pond fraction as a predictor for the seasonal Arctic sea ice extent minimum. *Environ. Res. Lett.*, **10 (5)**, 054 017.

Liu, Y., W. Wang, A. Kumar, and T. Collow, 2019: Assessment of CPC sea ice initialization system (CSIS) and CPC weekly experimental sea ice forecasts. *NOAA's Climate Prediction Center Stakeholder Meeting*, **22**, 24.

Lukovich, J. V., J. C. Stroeve, A. Crawford, L. Hamilton, M. Tsamados, H. Heorton, and F. Massonnet, 2021: Summer extreme cyclone impacts on Arctic sea ice. *Journal of Climate*, **34 (12)**, 4817–4834.

MacLachlan, C., and Coauthors, 2015: Description of GloSea5: the Met Office high resolution seasonal forecast system. *Q.J.R. Met. Soc.*, **141**, 1072–01 084.

Manzanas, R., J. M. Gutiérrez, J. Bhend, S. Hemri, F. J. Doblas-Reyes, V. Torralba, E. Penabad, and A. Brookshaw, 2019: Bias adjustment and ensemble recalibration methods for seasonal forecasting: A comprehensive intercomparison using the C3S dataset. *Clim. Dyn.*, **53 (3)**, 1287–1305.

Martin, J., A. Monahan, and M. Sigmond, 2023: Improved Seasonal Forecast Skill of Pan-Arctic and Regional Sea Ice Extent in CanSIPS Version 2. *Weather and Forecasting*, **38 (10)**, 2029–2056, https://doi.org/10.1175/WAF-D-22-0193.1.

Massonnet, F., T. Fichefet, and H. Goosse, 2015: Prospects for improved seasonal Arctic sea ice predictions from multivariate data assimilation. *Ocean Modelling*, **88**, 16–25.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0163.1.

Massonnet, F., and Coauthors, 2023: SIPN South: Six years of coordinated seasonal Antarctic sea ice predictions. *Frontiers in Marine Science*, **10**, 1148 899.

Meier, W., and Coauthors, 2021: 2020 Sea Ice Outlook Post-Season Report. *Sea Ice Prediction Network*, URL https://www.arcus.org/sipn/sea-ice-outlook/2020/post-season.

Meier, W. N., and J. S. Stewart, 2023: Arctic and Antarctic regional masks for sea ice and related data products, Version 1. National Snow and Ice Data Center, URL https://nsidc.org/data/NSIDC-0780/versions/1, https://doi.org/10.5067/CYW3O8ZUNIWC.

Merryfield, W., W.-S. Lee, W. Wang, M. Chen, and A. Kumar, 2013: Multi-system seasonal predictions of Arctic sea ice. *Geophys. Res. Lett*, **40 (8)**, 1551–1556.

Molod, A., and Coauthors, 2020: GEOS-S2S version 2: The GMAO high-resolution coupled model and assimilation system for seasonal prediction. *Journal of Geophysical Research: Atmospheres*, **125 (5)**, e2019JD031 767.

Msadek, R., G. Vecchi, M. Winton, and R. Gudgel, 2014: Importance of initial conditions in seasonal predictions of Arctic sea ice extent. *Geophys. Res. Lett.*, **41 (14)**, 5208–5215.

Murphy, A. H., 1988: Skill scores based on the mean square error and their relationships to the correlation coefficient. *Mon. Wea. Rev.*, **116 (12)**, 2417–2424.

Murphy, A. H., 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. *Weather and forecasting*, **8 (2)**, 281–293.

Ordoñez, A. C., C. M. Bitz, and E. Blanchard-Wrigglesworth, 2018: Processes controlling Arctic and Antarctic sea ice predictability in the Community Earth System Model. *J. Climate*, **31 (23)**, 9771–9786.

Peterson, K. A., A. Arribas, H. Hewitt, A. Keen, D. Lea, and A. McLaren, 2015: Assessing the forecast skill of Arctic sea ice extent in the GloSea4 seasonal prediction system. *Clim. Dyn.*, **44 (1-2)**, 147–162.

Petty, A. A., D. Schröder, J. Stroeve, T. Markus, J. Miller, N. Kurtz, D. Feltham, and D. Flocco, 2017: Skillful spring forecasts of September Arctic sea ice extent using passive microwave sea ice observations. *Earth's Future*, **5 (2)**, 254–263.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0163.1.

Ponsoni, L., F. Massonnet, D. Docquier, G. Van Achter, and T. Fichefet, 2020: Statistical predictability of the Arctic sea ice volume anomaly: identifying predictors and optimal sampling locations. *The Cryosphere*, **14 (7)**, 2409–2428.

Prodhomme, C., L. Batté, F. Massonnet, P. Davini, O. Bellprat, V. Guemas, and F. J. Doblas-Reyes, 2016: Benefits of increasing the model resolution for the seasonal forecast quality in EC-Earth. *Journal of Climate*, **29 (24)**, 9141–9162.

Qiao, F., Z. Song, Y. Bao, Y. Song, Q. Shu, C. Huang, and W. Zhao, 2013: Development and evaluation of an Earth System Model with surface gravity waves. *Journal of Geophysical Research: Oceans*, **118 (9)**, 4514–4524.

Risbey, J. S., and Coauthors, 2021: Standard assessments of climate forecast skill can be misleading. *Nature Communications*, **12 (1)**, 4346.

Rousset, C., and Coauthors, 2015: The Louvain-La-Neuve sea ice model LIM3. 6: global and regional capabilities. *Geoscientific Model Development*, **8 (10)**, 2991–3005.

Saha, S., and Coauthors, 2014: The NCEP climate forecast system version 2. *J. Climate*, **27 (6)**, 2185–2208.

Schröder, D., D. L. Feltham, D. Flocco, and M. Tsamados, 2014: September Arctic sea-ice minimum predicted by spring melt-pond fraction. *Nat. Clim. Change*, **4 (5)**, 353–357.

Serreze, M. C., A. D. Crawford, J. Stroeve, A. P. Barrett, and R. A. Woodgate, 2016: Variability, trends, and predictability of seasonal sea ice retreat and advance in the chukchi sea. *J. Geophys. Res.: Oceans*, **121 (10)**, 7308–7325.

Serreze, M. C., and J. Stroeve, 2015: Arctic sea ice trends, variability and implications for seasonal ice forecasting. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **373 (2045)**, 20140 159.

Shu, Q., F. Qiao, J. Liu, Z. Song, Z. Chen, J. Zhao, X. Yin, and Y. Song, 2021: Arctic sea ice concentration and thickness data assimilation in the FIO-ESM climate forecast system. *Acta Oceanologica Sinica*, **40 (10)**, 65–75.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0163.1.

Sigmond, M., J. Fyfe, G. Flato, V. Kharin, and W. Merryfield, 2013: Seasonal forecast skill of Arctic sea ice area in a dynamical forecast system. *Geophys. Res. Lett.*, **40 (3)**, 529–534.

Sigmond, M., M. Reader, G. Flato, W. Merryfield, and A. Tivy, 2016: Skillful seasonal forecasts of Arctic sea ice retreat and advance dates in a dynamical forecast system. *Geophys. Res. Lett.*, **43**.

Simmonds, I., and I. Rudeva, 2012: The great Arctic cyclone of August 2012. *Geophys. Res. Lett.*, **39 (23)**.

Steele, M., and Coauthors, 2021: Moving sea ice prediction forward via community intercomparison. *Bulletin of the American Meteorological Society*, **102 (12)**, E2226–E2228.

Stroeve, J., L. C. Hamilton, C. M. Bitz, and E. Blanchard-Wrigglesworth, 2014: Predicting September sea ice: Ensemble skill of the SEARCH sea ice outlook 2008–2013. *Geophys. Res. Lett.*, **41 (7)**, 2411–2418.

Tietsche, S., E. Hawkins, and J. J. Day, 2016: Atmospheric and oceanic contributions to irreducible forecast uncertainty of Arctic surface climate. *J. Climate*, **29 (1)**, 331–346.

Tietsche, S., D. Notz, J. H. Jungclaus, and J. Marotzke, 2013: Predictability of large interannual Arctic sea-ice anomalies. *Clim. Dyn.*, **41 (9-10)**, 2511–2526.

Tietsche, S., and Coauthors, 2014: Seasonal to interannual Arctic sea ice predictability in current global climate models. *Geophys. Res. Lett.*, **41 (3)**, 1035–1043.

Van den Dool, H., 2007: *Empirical Methods in Short-Term Climate Prediction*. Oxford Univ. Press, Oxford, U. K.

Vecchi, G. A., and Coauthors, 2014: On the seasonal forecasting of regional tropical cyclone activity. *J. Climate*, **27 (21)**, 7994–8016.

Voldoire, A., and Coauthors, 2019: Evaluation of CMIP6 deck experiments with CNRM-CM6-1. *Journal of Advances in Modeling Earth Systems*, **11 (7)**, 2177–2213.

Wagner, P. M., and Coauthors, 2020: Sea-ice information and forecast needs for industry maritime stakeholders. *Polar Geography*, **43 (2-3)**, 160–187.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0163.1.

Walsh, J. E., J. S. Stewart, and F. Fetterer, 2019: Benchmark seasonal prediction skill estimates based on regional indices. *The Cryosphere*, **13 (4)**, 1073–1088.

Wang, W., M. Chen, and A. Kumar, 2013: Seasonal prediction of Arctic sea ice extent from a coupled dynamical forecast system. *Mon. Wea. Rev.*, **141 (4)**, 1375–1394.

Williams, C. K., and C. E. Rasmussen, 2006: *Gaussian processes for machine learning*, Vol. 2. MIT press Cambridge, MA.

Williams, J., B. Tremblay, R. Newton, and R. Allard, 2016: Dynamic preconditioning of the minimum September sea-ice extent. *J. Climate*, **29 (16)**, 5879–5891.

Yamagami, A., M. Matsueda, and H. L. Tanaka, 2018: Predictability of the 2012 great Arctic cyclone on medium-range timescales. *Polar Science*, **15**, 13–23.

Yuan, X., D. Chen, C. Li, L. Wang, and W. Wang, 2016: Arctic sea ice seasonal prediction by a linear markov model. *J. Climate*, **29 (22)**, 8151–8173.

Zampieri, L., H. F. Goessling, and T. Jung, 2018: Bright prospects for Arctic sea ice prediction on subseasonal time scales. *Geophys. Res. Lett.*, **45 (18)**, 9731–9738.

Zeng, J., Q. Yang, X. Li, X. Yuan, M. Bushuk, and D. Chen, 2023: Reducing the spring barrier in predicting summer Arctic sea ice concentration. *Geophys. Res. Lett.*, **50 (8)**, e2022GL102 115.

Zhan, Y., and R. Davies, 2017: September Arctic sea ice extent indicated by June reflected solar radiation. *J. Geophys. Res.: Atmospheres*, **122 (4)**, 2194–2202.

Zhang, J., R. Lindsay, A. Schweiger, and M. Steele, 2013: The impact of an intense summer cyclone on 2012 Arctic sea ice retreat. *Geophys. Res. Lett.*, **40 (4)**, 720–726.

Zhang, J., and D. Rothrock, 2003: Modeling global sea ice with a thickness and enthalpy distribution model in generalized curvilinear coordinates. *Mon. Wea. Rev.*, **131 (5)**, 845–861.

Zhang, J., M. Steele, R. Lindsay, A. Schweiger, and J. Morison, 2008: Ensemble 1-year predictions of Arctic sea ice for the spring and summer of 2008. *Geophys. Res. Lett.*, **35 (8)**.

Zhang, Y.-F., M. Bushuk, M. Winton, B. Hurlin, X. Yang, T. Delworth, and L. Jia, 2021: Assimilation of satellite-retrieved sea ice concentration and prospects for september predictions of Arctic sea ice. *J. Climate*, **34 (6)**, 2107–2126.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0163.1.

Zhang, Y.-F., and Coauthors, 2022: Subseasonal-to-seasonal Arctic sea ice forecast skill improvement from sea ice concentration assimilation. *J. Climate*, **35 (13)**, 4233–4252.

Zuo, H., M. A. Balmaseda, S. Tietsche, K. Mogensen, and M. Mayer, 2019: The ECMWF operational ensemble reanalysis–analysis system for ocean and sea ice: a description of the system and assessment. *Ocean science*, **15 (3)**, 779–808.

Accepted for publication in *Bulletin of the American Meteorological Society.* DOI 10.1175/BAMS-D-23-0163.1.