Skill of Seasonal Arctic Sea Ice Extent Predictions Using the North American Multimodel Ensemble

K. J. HARNOS

NOAA/Climate Prediction Center, College Park, and Innovim, LLC, Greenbelt, Maryland

M. L'HEUREUX

NOAA/Climate Prediction Center, College Park, Maryland

Q. DING

Department of Geography, and Earth Research Institute, University of California, Santa Barbara, Santa Barbara, California

Q. ZHANG

NOAA/Climate Prediction Center, College Park, Maryland

(Manuscript received 10 November 2017, in final form 9 November 2018)

ABSTRACT

Previous studies have outlined benefits of using multiple model platforms to make seasonal climate predictions. Here, reforecasts from five models included in the North American Multimodel Ensemble (NMME) project are utilized to determine skill in predicting Arctic sea ice extent (SIE) during 1982–2010. Overall, relative to the individual models, the multimodel average results in generally smaller biases and better correlations for predictions of total SIE and year-to-year (Y2Y), linearly, and quadratically detrended variability. Also notable is the increase in error for NMME predictions of total September SIE during the mid-1990s through 2000s. After 2000, observed September SIE is characterized by more significant negative trends and increased Y2Y variance, which suggests that recent sea ice loss is resulting in larger prediction errors. While this tendency is concerning, due to the possibility of models not accurately representing the changing trends in sea ice, the multimodel approach still shows promise in providing more skillful predictions of Arctic SIE over any individual model.

1. Introduction

September Arctic sea ice has decreased by more than 10% per decade since satellite observations began in 1979 (Stroeve et al. 2012; Comiso et al. 2008). The historical record of sea ice satellite observations is relatively short (Serreze and Stroeve 2015); however, there is a clear decline in September sea ice extent (SIE) since 1979 with a steepening in the trend since about 1997. There are a number of interdependent factors that contribute to declining SIE and include reduced sea ice thickness, longer melt seasons, increased solar radiation absorption, and changes in the high-latitude atmospheric circulation (Serreze and Stroeve 2015; Ding et al. 2017). With the multitude of environmental, geopolitical, and economic implications of sea ice decline, it is imperative to improve sea ice predictions through improved global climate model simulations.

Previous work shows evidence for skillful predictions of seasonal SIE using both statistical and dynamical methods. Initial attempts at seasonal SIE prediction by Walsh (1980) showed that a statistical method, using empirical orthogonal functions, resulted in significant skill out to only 1- to 2-month lead times for regional sea ice north of Alaska in both winter and summer months. Lindsay et al. (2008) also used a statistical approach to assess how well pan-Arctic and regional September SIE could be predicted for a lead time of up to one year. Their results show that the trend accounts for a large amount of the skill. They also document that, for forecast lead times less than 2 months, sea ice concentration

Corresponding author: Kirstin Harnos, kirstin.harnos@noaa.gov

DOI: 10.1175/JCLI-D-17-0766.1

is the most important predictor while ocean temperature is more important for longer lead times.

Dynamical models also show the ability to skillfully predict seasonal SIE, which depends on initial conditions, data assimilation, and model physics (e.g., Msadek et al. 2014; Chevallier et al. 2013; Wang et al. 2013; Blanchard-Wrigglesworth et al. 2015, 2017; Bushuk et al. 2017; Sigmond et al. 2013; Merryfield et al. 2013b). Understanding predictability and skill in the dynamical models has been previously investigated using the "perfect model" approach in which the initial conditions and model physics are assumed to be perfect. These studies (e.g., Blanchard-Wrigglesworth et al. 2015, 2017; Holland et al. 2011; Day et al. 2014) conclude that predictability in SIE is largely influenced by the sea ice thickness initialization and improvements in skill can be achieved through improved initial conditions. Hindcast studies (e.g., Msadek et al. 2014; Bushuk et al. 2017; Chevallier et al. 2013; Wang et al. 2013) use real-world initial conditions to determine actual skill between the forecast and the observations, while sources of prediction error are more difficult to isolate as they may be due to either initial condition or model physics errors (Blanchard-Wrigglesworth et al. 2015). Many studies attribute the majority of skillful predictions of individual modeling systems to the long-term trend of sea ice; however, there is still significant skill once the trends are removed mainly due to the influence of the initial conditions (Sigmond et al. 2013; Wang et al. 2013; Chevallier et al. 2013; Holland et al. 2011).

While covariability exists between sea ice and the atmospheric circulation, in general the chaotic nature of high-latitude internal atmospheric variability acts to limits skill (Wettstein and Deser 2014; Serreze and Stroeve 2015; Ding et al. 2017). As the community continues to address questions of inherent predictability and how to enhance prediction skill in the models, one approach to improve seasonal-to-interannual predictions is by combining multiple models together (e.g., Guemas et al. 2016; DelSole et al. 2014; Kirtman et al. 2014). Previous work (Merryfield et al. 2013b; Blanchard-Wrigglesworth et al. 2017) has shown higher skill in predicting SIE for multimodel ensembles than an individual model. For seasonal forecasting, Hagedorn et al. (2005) investigated improved skill using the multimodel approach and concluded that the increase is achieved by reducing model errors and improving forecast consistency. Using hindcasts from nine coupled climate models of El Niño–Southern Oscillation (ENSO), DelSole et al. (2014) show that enhanced skill is due to the reduction of noise and the combination of signals from a diverse set of models. Model diversity results in the cancellation of less meaningful signals and enhances the signals that are more skillful. Interestingly, they conclude that the combination of diverse model signals contributes more toward enhancing skill than the reduction of noise by ensemble averaging.

In recent years, the Sea Ice Prediction Network (SIPN; Stroeve et al. 2014) has collected predictions from a variety of dynamical, statistical, and heuristic systems to predict September sea ice. The results for 2008–13, informed by hundreds of submissions, show that higher skill forecasts of the SIE ensemble median occur when the observed SIE is near the long-term trend (Stroeve et al. 2014). The SIPN project has also found that dynamical SIE forecasts could be improved by adopting postprocessing methods to correct for model bias, because individual model uncertainty is the lead contributor to forecast uncertainty (Blanchard-Wrigglesworth et al. 2017). Merryfield et al. (2013b) demonstrated the potential usefulness of the multiplatform approach to sea ice prediction. The average of two coupled dynamical systems, the Climate Forecast System version 2 (CFSv2) and the Canadian Seasonal to Interannual Prediction System (CanSIPS), outperformed both individual systems overall in Arctic sea ice area forecasts. Guemas et al. (2016) provide a comprehensive review on the current state of sea ice predictability and recommend an increase in ensemble sizes, an examination of more case studies, and comparisons of multimodel systems to individual models. As recommended, this paper examines a comprehensive multimodel prediction system, which is currently used for operational outlooks at the NOAA Climate Prediction Center.

The North American Multimodel Ensemble (NMME) prediction experiment (Kirtman et al. 2014) originates from a multiagency team that collects and organizes global model data on a mostly uniform spatial and temporal scale. Previously the NMME has been shown to improve forecast skill for temperature, precipitation, sea surface temperature (Becker et al. 2014), and ENSO (Barnston et al. 2017; Tippett et al. 2017). Given the potential impacts of a changing Arctic and the real-time success of the NMME in improving prediction, the goal here is to explore the use of the NMME in seasonal forecasts of Arctic sea ice. The skill of Arctic SIE prediction is assessed from five NMME models and the multimodel average. In particular, the total SIE (which includes the long-term trend), year-to-year SIE, and detrended variability of SIE reforecast simulations will be examined.

2. Models, methods, and verification

a. Model description and methodology

There are five models that currently provide monthly mean sea ice reforecasts to the NMME for an overlapping period of 1982–2010: CanSIPS, which is separated into its two components, Canadian Climate Model versions 3 and 4 (CanCM3 and CanCM4; Merryfield et al. 2013a), the Forecast-Oriented Low Ocean Resolution model (FLORB-01; Msadek et al. 2014), the

CFSv2 (Saha et al. 2014), and the Community Climate System Model version 4 (CCSM4; Jahn et al. 2012). CanCM3, CanCM4, and FLORB-01 archives were accessed through the online NMME Phase-II database hosted by the National Center for Atmospheric Research (NCAR). The model providers supplied the CFSv2 and CCSM4 data. To be included in the NMME, model data are processed by NCAR from its native resolution to a common $1° \times 1°$ grid. However, sea ice is not considered a required NMME variable so there are some caveats. For one, the CFSv2 sea ice data are not archived in the NMME NCAR database and are not processed onto the $1° \times 1°$ grid. Finally, while FLORB-01 data are $1° \times 1°$, they are currently archived at NCAR on their native model tripolar grid. In this study, SIE was calculated from the files as delivered by data providers. Table 1 summarizes the number of ensemble members, resolution, and simulation years from each model. To maintain consistency with previous CFSv2 sea ice studies (Wang et al. 2013), only 16 ensemble members are included in the analysis; 10 to 12 members are available from the other four models. Each hindcast simulation is initialized and run out to 9 (CFSv2) or 12 months (CanCM3, CanCM4, FLORB-01, and CCSM4), although skill will only be evaluated for the shared nine forecast leads. Forecast lead refers to the number of months since the simulation began. For example, a 2-month lead forecast in September refers to a simulation initialized in July. Henceforth, "NMME" will refer to the average of the ensemble means (a total of 58 members) from five component models for bias examination. Skill metrics are calculated using the individual model anomalies where the model climatology at each lead time is removed. This has been shown previously to provide a first-order removal of the model systematic biases (Becker et al. 2014). Following the same ensemble procedures as the operational NMME forecasts for ENSO, temperature, and precipitation, the ensemble average is equally weighted. Previous work by DelSole et al. (2013) shows statistically for temperature and precipitation that equally weighted multimodel forecasts were similar in skill to those forecasts made using unequal weighted schemes. From a global perspective, they found that, including high latitudes, unequal weighting schemes added value to only a small fraction of the globe. The effects of weighting have not been explicitly explored for sea ice forecasts and are beyond the scope of the study here.

SIE is defined as the total area of grid boxes with sea ice concentration of at least 15%. Hindcast predictions are verified against observational data derived from the National Snow and Ice Data Center NASA bootstrap algorithm (Cavalieri et al. 1996; Comiso 2000). Since the NASA bootstrap is not used as an initialization source for any of the NMME models, it provides a slightly more independent verification dataset. Further, Notz (2014) conducted a comparison of the NASA bootstrap and NASA team algorithms, the two most widely used data sources for verification, and found that differences in SIE were present mainly in areas of low sea ice concentration. Because the total area of low concentration is small, both of the NASA algorithms produce similar SIE values and therefore selecting between the two products has little effect on the verifications.

TABLE 1. Summary of the individual NMME models.

Model	Simulation years	Resolution	Ensemble members
CanCM3	1981–2013	$1° \times 1°$	10
CanCM4	1981–2013	$1° \times 1°$	10
FLORB-01	1981–2013	$1° \times 1°$ (tripolar grid)	12
CFSv2	1982–2011	$2.5° \times 2.5°$	16
CCSM4	1982–2015	$1° \times 1°$	10

The prediction skill is evaluated using the anomaly correlation coefficient (ACC), bias, and root-mean-square error (RMSE). ACC measure the similarity, or the strength of the fit, between the observed and predicted anomalies, whereas the error is evaluated using the bias (forecast minus observations) and RMSE (square root of the average of the squared difference between the forecast and observations). Bias contains the sign of the error (overestimate or underestimate), but positive and negative forecast errors can cancel out. In contrast, RMSE does not indicate the sign of the error, but can provide an average magnitude of error. Each metric is tested for statistical significance at the 95% level using a t test.

Total SIE values and any associated skill are heavily dominated by the long-term trend (Wang et al. 2013; Chevallier et al. 2013; Sigmond et al. 2013). For certain users interested in seasonal predictions outside of the trend, the change of SIE from one year to the next, or interannual variability, is more valuable. However, the low-frequency trend in SIE loss is challenging to predict because SIE has a significant nonlinear component, reflecting the steepening of recent declines (Serreze and Stroeve 2015; Swart et al. 2015; Lindsay and Schweiger 2015). Isolating this lower-frequency signal is not a trivial undertaking, especially in real-time prediction. For the purposes of isolating variability outside of the trend, or detrending the data, three approaches are presented. Given the lack of clear future trajectory, one method is to simply subtract out the standard least squares linear fit over the reforecast period. However, this method does not capture nonlinear change in the trend, which is still included in the residual. With the increasingly steeper SIE

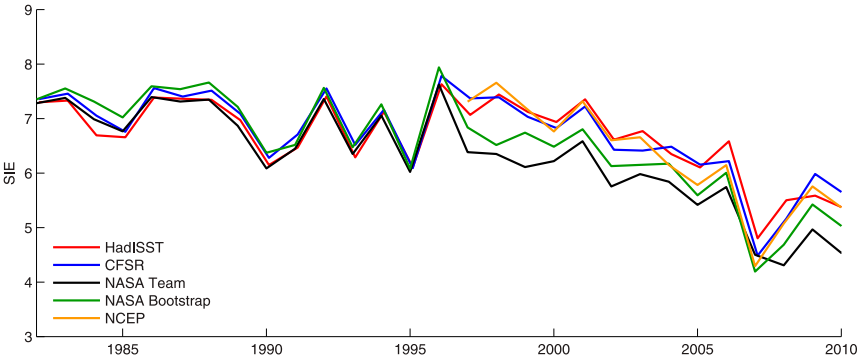

FIG. 1. September time series from 1982 to 2010 of initialization data sources for NMME models. SIE units in 10^6 km^2.

trends (Serreze and Stroeve 2015; Swart et al. 2015; Lindsay and Schweiger 2015), the use of a linear fit to capture the variability signal can lead to an overestimation in skill depending on the evaluation time period (Dirkson et al. 2017). Detrending SIE using a quadratic fit was recently presented in Dirkson et al. (2017) and this study will extend their results to four additional models. Alternatively, the "year-to-year" (Y2Y) SIE calculation established in Wang et al. (2013) does not assume an a priori fit and evaluates the skill of the models in predicting SIE from one year to the next, which can involve changes irrespective of the longer-term trend. Y2Y SIE is computed as the difference in SIE of year $n + 1$ minus year n (Wang et al. 2013). Note that the final year in the record is not included in the Y2Y analyses due to a lack of "final year + 1" information. Y2Y errors may also detect systematic model biases if values are consistently of the same sign.

b. Sea ice initialization

The model's initial conditions are a leading source of SIE prediction skill, especially at shorter lead times (e.g., Chevallier and Salas-Mélia 2012; Blanchard-Wrigglesworth et al. 2015; Msadek et al. 2014). Because this study is utilizing models from several institutions, documenting the differing approaches to initialization may provide insights into why models perform differently. The National Centers for Environmental Prediction (NCEP) CFSv2 hindcast initial conditions are provided by Climate Forecast System Reanalysis (CFSR; Saha et al. 2010; Wang et al. 2013). While CFSR assimilates both in situ and satellite measurements for atmosphere, land, and ocean values, sea ice concentrations are solely from satellite observations. More specifically, the source of sea ice concentration from 1979 through December 1996 is from the NASA team algorithm (Cavalieri et al. 1996). The NCEP operational ice analysis (Grumbine 1996) is used from January 1997 to the present. Sea ice thickness is determined by the thermodynamic balance between the sea ice, ocean, and atmosphere (Wang et al. 2013). The NCAR CCSM4 model, run from the University of Miami, also uses CFSR as the initial conditions for their model (Infanti and Kirtman 2016).

The Hadley Centre Sea Ice and Sea Surface Temperature dataset (HadISST1.1; Rayner et al. 2003) is utilized as the sea ice initialization source in both CanCM3 and CanCM4, which are two different climate models from Environment Canada's CanSIPS forecast system. The initial conditions are provided from an assimilation system that uses atmospheric inputs from the ERA-Interim reanalysis, oceanic inputs from OISST, and sea ice from HadISST1.1 (Merryfield et al. 2013a). Sea ice concentrations from the assimilation system are then relaxed toward the observations to reduce model input biases (Merryfield et al. 2013a). Thickness is initialized from a seasonally varying climatology produced from a previous CanSIPS model version (Merryfield et al. 2013a). Finally, the Geophysical Fluid Dynamics Laboratory (GFDL) FLORB-01 system obtains its sea ice concentration and thickness initial conditions from the assimilation system that ingests oceanic and atmospheric data from the NCEP reanalysis (Msadek et al. 2014; Jia et al. 2015; Bushuk et al. 2017). Sea ice observations are not used in the creation of FLORB-01 sea ice initial conditions. Msadek et al. (2014) notes that the initial conditions supplied to FLORB-01 from the data assimilation system are biased slightly low when compared to observations, which may heavily influence FLORB-01 biases presented in a later section.

Figure 1 presents a time series of September SIE from 1982 to 2010 from CFSR, HadISST1.1, the NASA team, NASA bootstrap, and the NCEP operational analysis. As seen in Notz (2014), the bootstrap and team algorithms are similar throughout the period with the team algorithm slightly lower than the bootstrap. In comparison, before

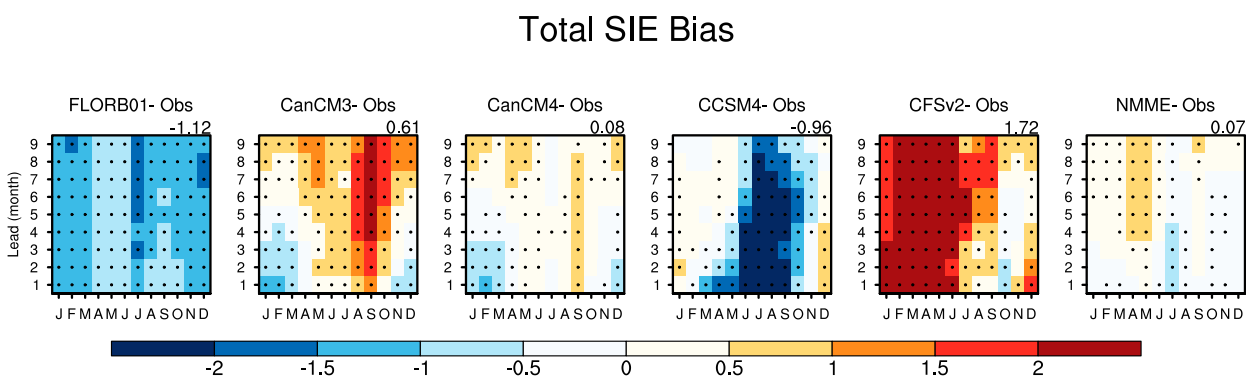

FIG. 2. Model bias (model minus observations) including the NMME as a function of target month vs lead time (months) for total SIE (10^6 km^2). Dotted overlay indicates 95% *t*-test significance.

September 1997 HadISST1.1 and CFSR are similar to the NASA algorithms. This is due to the fact that they are derived directly from the NASA team algorithm (Cavalieri et al. 1996). After 1997, HadISST1.1 and CFSR follow the NCEP operational analysis more closely while the NASA algorithms fall slightly lower than NCEP. This result is unsurprising since CFSR and HadISST1.1 switch to the real-time NCEP operational analysis as their algorithm source after 1997 (Grumbine 1996). These three datasets show larger SIE values on average 0.48 million km^2 or about 7% greater than both of the NASA algorithms. The underestimation of the SIE trend in HadISST1.1 has been shown to decrease ACC scores for total SIE anomalies (Sigmond et al. 2013). Wang et al. (2013) also document that the change in CFSR results in a weaker SIE trend in the CFSv2 hindcasts after 1997.

3. Results

a. SIE climatology and prediction skill

Figure 2 presents the total SIE bias among the five different individual models and the NMME average and significance at the 95% level is indicated by the dotted overlay. Interestingly, the results show a striking lack of common biases between the individual models. The overall bias in the total SIE is smallest for CanCM4 with an average bias of 0.08 million km^2 over all months and leads. More specifically, the biases run from a significant underestimation of 1.2 million km^2 or 7.8% less SIE in 1-month lead February prediction to an overestimation of 0.9 million km^2 or 6.6% more SIE in January at the 8- to 9-month leads. Biases during September are slightly larger than other summer and fall months and are not statistically significant until 7- and 8-month lead times. CanCM3, with a total bias of 0.6 million km^2, generally underpredicts SIE for shorter leads during the winter and overpredicts for almost all leads in the summer and early fall and is significant over almost all months and leads. As with CanCM4, September has the largest differences, although biases in CanCM3 are much larger with values over 2 million km^2 or 35% more SIE at lead times greater than 4 months. Similar to what is shown in Merryfield et al. (2013b), both of the Canadian models share the same general pattern of biases, including negative biases in short lead winter and largest positive biases in the fall. Given that these models have the same sea ice initialization, but differing atmospheric models (Merryfield et al. 2013b), the similar patterns in the SIE bias imply an important role for the initial conditions over model physics, noting that these models also share the same ocean components, which may also influence the bias patterns.

FLORB-01 consistently underpredicts SIE for all months and leads on average by 1.1 million km^2, which equates to approximately 10% less SIE than the observations, especially for forecast targets outside of the spring (April–June). As noted in Msadek et al. (2014), FLORB-01 initial conditions are negatively biased possibly due to imperfect model physics or uncertainty in the satellite data. The continuation of the negative bias in FLORB-01 into the longer lead times also points to issues with the model physics. The largest negative bias in any individual model stands out in CCSM4, which significantly underpredicts the observations by values between 2.8 and 4.0 million km^2 for shorter leads during February through May and all leads for forecast targets in June through October. This underprediction equates to as much as 45% less ice than the observed values. In contrast, CFSv2 has the largest positive biases of up to 29% more SIE than observations. This model largely overpredicts winter and spring SIE with values on average between 2.5 and 3.5 million km^2 or 20% to 25% over estimations, while showing smaller biases during the second half of the year. Given that past studies have shown multimodel averages improve prediction by mitigating large biases (Merryfield et al. 2013b; Blanchard-Wrigglesworth et al. 2015), the

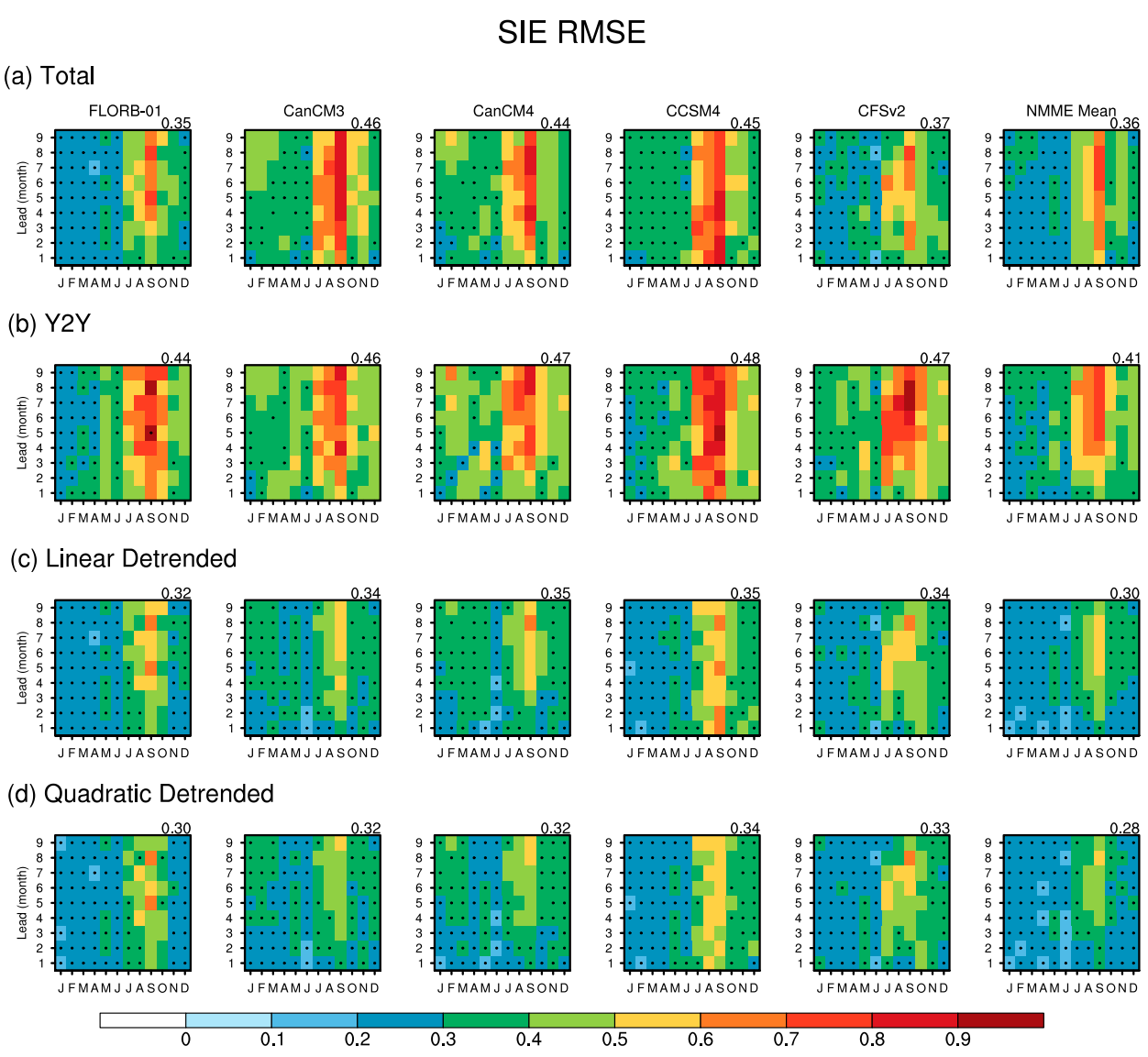

FIG. 3. RMSE values for (a) total SIE, (b) Y2Y SIE, (c) linear detrended, and (d) quadratic detrended SIE as a function of target month vs lead time (month). Dotted overlay indicates 95% *t*-test significance. The numeric value in the upper right of each panel represents the average RMSE over all months and lead times.

NMME average performed well with smaller biases for most lead times and target months relative to the individual models with an average bias of 0.07 million km^2 or 2.8% more SIE than observed. Results from the total SIE bias are consistent with previous studies using multimodel averages (e.g., Merryfield et al. 2013b; Blanchard-Wrigglesworth et al. 2015).

Another skill metric, the RMSE, is shown in Fig. 3 for the total SIE (Fig. 3a), Y2Y (Fig. 3b), detrended SIE using linear fit (Fig. 3c), and detrended SIE using quadratic fit (Fig. 3d) with significance at the 95% level indicated by the dotted overlay. As stated previously, the ability to isolate and remove the long-term trend in SIE is not a trivial undertaking. The three detrending methods presented here represent differing approaches for capturing the interannual variability. RMSE values are generally larger for Y2Y variability when compared to both detrended RMSE methods. This is possibly due to Y2Y representing the year-to-year variability only, while the detrended methods incorporate the year-to-year variability as well as some component of the long-term trend that is not removed (Wang et al. 2013). However, as mentioned earlier, subtracting out a linear fit does not remove all aspects of the longer-term trend, which may be nonlinear. Also, while the quadratic fit was shown previously in Dirkson et al. (2017) and Fučkar et al. (2016) to provide a

more accurate removal of the long-term trend, there is still some potential for trends to influence the results if future trends depart from the selected fit. In real-time prediction, one does not know the influence of the trend for a given forecast, so for real-time purposes tracking the Y2Y SIE skill is a useful way to examine changes without an a posteriori knowledge of the trend.

For all three metrics and the total, the error values are greatest in the summer and fall, especially in September (Fig. 3). This is consistent with past results indicating that individual models (CFSv2: Wang et al. 2013; CanSIPS: Merryfield et al. 2013b; FLORB-01: Msadek et al. 2014) have largest errors in predicting variability for target months in late summer and fall when SIE is at its minimum. While the overall features are similar for RMSE, using the NMME average reduces the RMSE for all three detrending metrics; however, it does not show improvement over all models for the total SIE. Also similar to all three detrending methods and the total, FLORB-01 has the lowest error in the individual models with CCSM4 having the largest (CanCM3 has the largest RMSE in the total and CCSM4 has the second largest). Averaged over all months and leads, the individual model Y2Y values range from 0.44 in FLORB-01 to 0.48 in CCSM4. NMME Y2Y RMSE is 0.41 or around a 10% reduction. Linear detrending gives RMSE values ranging from 0.32 in FLORB-01 to 0.35 in both CanCM4 and CCSM4. NMME values are slightly lower at 0.30, again around a 10% reduction. Quadratic detrending is similar to the other metrics in that FLORB-01 has the lowest individual model RMSE at 0.30 and CCSM4 has the highest value at 0.34. NMME has the lowest RMSE of the three methods at an average of 0.28. The specific months and leads of improvement are dependent on the model, with the NMME showing the most improvement over FLORB-01 and CCSM4 at shorter leads for summer targets and all leads for the fall. Further, the NMME exhibits improvement over CanCM3 and CanCM4 for the late winter and spring at all leads and also for most months and leads in CFSv2.

The influence of trend in prediction skill is quite evident when analyzing SIE ACC (Fig. 4). There is a stark drop in correlation and significance when comparing the total SIE (Fig. 4a) to the Y2Y and detrended SIEs (Figs. 4b,c,d). Averaged over all leads and months, FLORB-01, CFSv2, and NMME have the highest total ACC with values averaging 0.77, 0.76, and 0.78, respectively. CanCM3, CanCM4, and CCSM4 ACC are around 30% lower with values of 0.52, 0.58, and 0.56. Speculating as to the cause of the lower ACC scores, for CanCM3 and CanCM4, the values may be due to issues with the underestimation of the trends in HadISST1.1 (Merryfield et al. 2013b). Although Merryfield et al. (2013a) also note the weak trend in CanCM3 when run without initialization, issues in the initial conditions do not provide the whole explanation. Since CCSM4 uses the same initialization source as CFSv2, the lower values may point to the model physics, rather than an initialization issue. As with the RMSE, ACC values of detrended SIE are generally 30%–40% more skillful than the Y2Y values, which range from 0.12 in CanCM3 to 0.27 in FLORB-01, with NMME at 0.28. Linear detrended scores are the highest of the three variability methods with highest values in FLORB-01 and CFSv2 at 0.36 and lowest ACC in CanCM4 at 0.23. As seen in all the previous metrics, the NMME score improves over the individual models with a value of 0.40, or approximately a 10% improvement over FLORB-01 and CFSv2 and a 40% improvement over CanCM4. The quadratic detrended values range similarly to the linear scores, with FLORB-01 at 0.36, and a slightly lower bottom score of 0.22 in both CanCM3 and CCSM4. As such, the NMME is slightly lower at 0.38. While the decrease in ACC values from total SIE to interannual variability SIE is consistent with previous studies (e.g., Wang et al. 2013; Sigmond et al. 2013; Merryfield et al. 2013b), it is clear that it is more difficult for models to predict the variability than the total SIE anomalies, which potentially benefit from some combination of their ability to capture the longer-term trends. For total, Y2Y, and detrended SIEs, the NMME predictions result in generally higher ACC compared to the individual models. The higher ACC with the NMME is most evident for predictions of Y2Y variability, especially for lead times greater than ~3 months. The NMME also noticeably improves upon CanCM3, CanCM4, and CCSM4 for predictions of total SIE.

While the NMME generally improves the prediction skill of total, Y2Y, and detrended SIE over any individual model, the amplitude of the benefit will vary depending on which skill metric is selected. The reduction in total SIE bias is likely due to the cancellation of the large positive and negative biases seen in the individual models. When evaluating skill based on the ACC, there is a more substantial benefit from using the NMME over individual models, particularly with respect to Y2Y or detrended variability. Overall, these results support the idea that averaging the models in NMME results in greater skill due to the addition of predictable signals (DelSole et al. 2014).

Finally, while not an official contribution to NMME, statistical methods have been shown to provide skillful predictions of SIE (Msadek et al. 2014; Blanchard-Wrigglesworth et al. 2011, 2015). For comparison, the same metrics are applied to a damped persistence statistical model (van den Dool 2006). This model develops coefficients using linear regression with SIE values from the NASA team algorithm. Predictions for each month

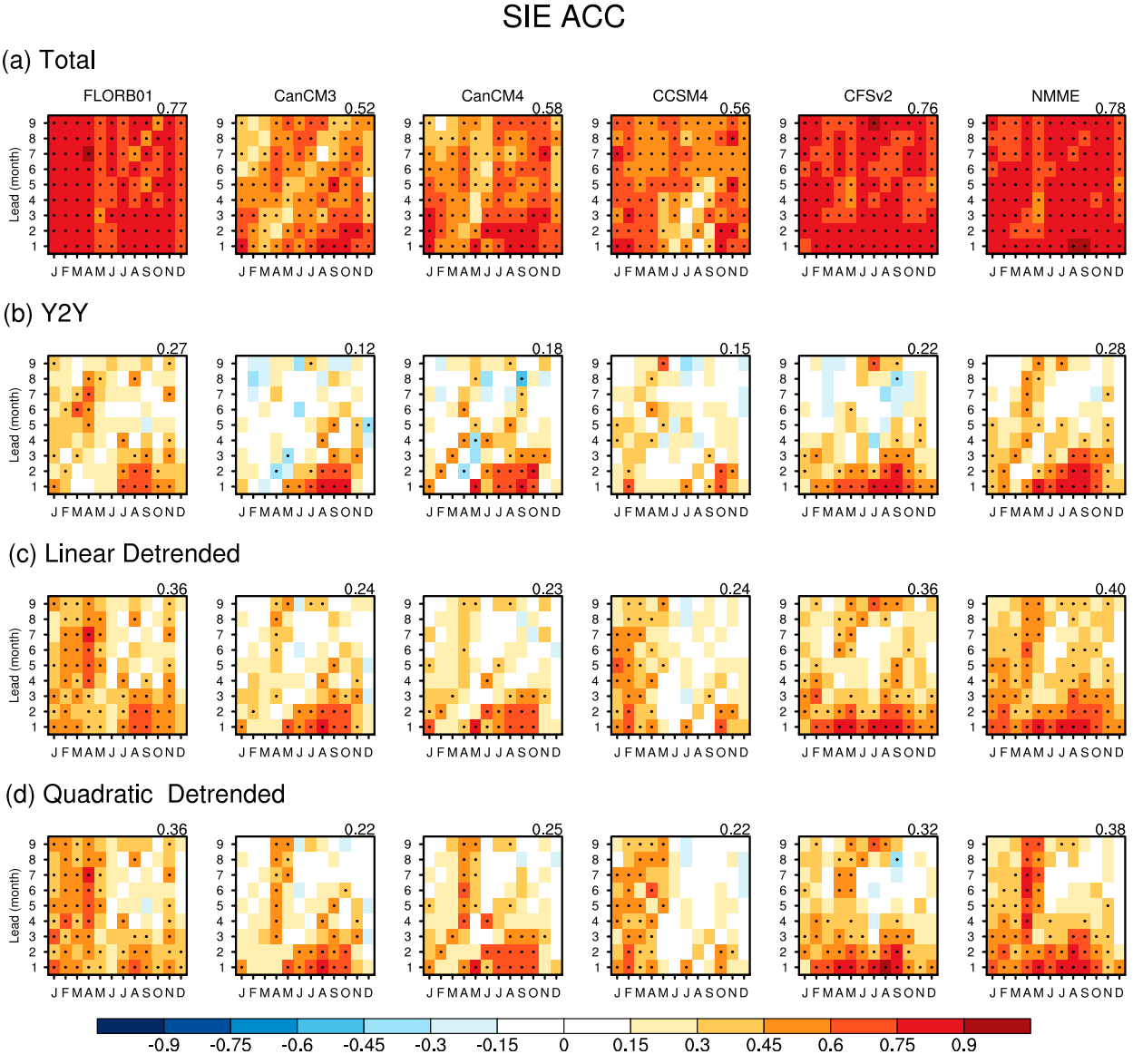

FIG. 4. ACC values for (a) total SIE, (b) Y2Y SIE, (c) linear detrended, and (d) quadratic detrended SIE as a function of target month vs lead time (months). Dotted overlay indicates 95% *t*-test significance. Numeric value on upper right of each panel represents the average ACC over all months and lead times.

and lead time are created by applying the anomaly from the linear trend of the initial month to the future predicated month, scaled by the regression coefficients (van den Dool 2006). Previous studies have shown that damped persistence generally has lower RMSE than dynamical model predictions in the summer months (Blanchard-Wrigglesworth et al. 2015) and higher detrended ACC in winter especially at longer leads (Msadek et al. 2014). However, Sigmond et al. (2013) and Merryfield et al. (2013b) conclude that forecast skill in CanCM3 and CanCM4 is enhanced compared to persistence. When compared to the NMME results in Figs. 3 and 4, damped persistence shows a number of interesting features, some contrary to the previous findings. RMSE scores for Y2Y (Fig. 5a) and both detrended SIE methods (Figs. 5b,c), with values of 0.44, 0.33, and 0.31, respectively, are higher than the NMME (Fig. 3; 0.41, 0.30, and 0.28) for all leads and months. This is contrary to the previous result of Blanchard-Wrigglesworth et al. (2015), who found that the damped persistence only slightly outperforms the multimodel mean for summer initializations; however, the study uses SIPN models from 2009 to 2014 for the multimodel mean and state.

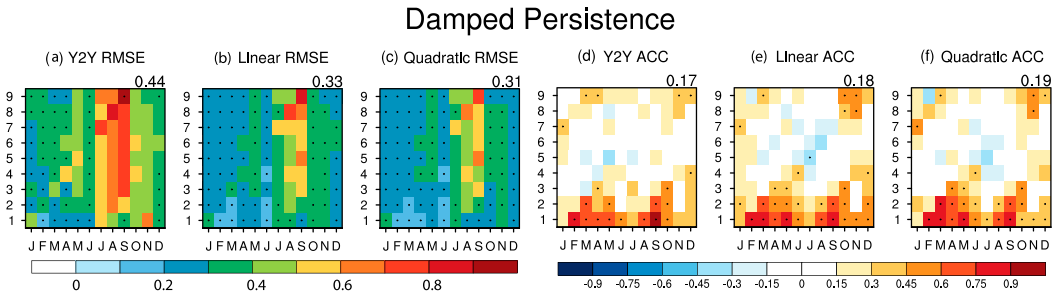

FIG. 5. Damped persistence SIE analysis for (a) Y2Y RMSE, (b) linear detrended RMSE, (c) quadratic detrended RMSE, (d) Y2Y ACC, (e) linear detrended ACC, and (f) quadratic detrended ACC. Dotted overlay indicates 95% *t*-test significance. The numeric value in the upper right of each panel represents the average value over all months and lead times.

The ACC of Y2Y (Fig. 5d) and detrended (Figs. 5e,f) SIE show the same pattern of value decrease with lead time when compared to the ACC as the dynamical models (Fig. 4). The largest values between 0.4 and 0.8 at lead 1 quickly fall to values less than 0.3 beyond lead month 3. The higher skill values using damped persistence for the longer lead winter season, as shown in Msadek et al. (2014), are not reflected in our results, which show that the ACC values of detrended SIE in NMME are generally higher for all leads in the winter. The NMME values are also higher in the midrange leads of 3–7 months where NMME values are positive and the damped persistence shows near zero or slightly negative ACC values. Over all leads and months, the damped persistence values are approximately half the NMME ACC averages with values of 0.17, 0.18, and 0.19 for Y2Y, linear, and quadratic detrended scores. The results are consistent with Sigmond et al. (2013) showing that dynamical models and the multimodel mean are more skillful than persistence. In fact, NMME has better skill over not only the damped persistence model, but also the individual models.

b. March and September SIE anomaly time series

While the previous section focuses on the SIE skill metrics averaged over 1982–2010, here analyses examine how the predictions have evolved during that time period. Especially given the recent acceleration in the SIE trends (Serreze and Stroeve 2015), it is important to question whether the models have been able to capture these changes. Because September and March are when SIE reaches its minimum and maximum, they are the focus of the following sections. These two months are displayed in Fig. 6, which shows the anomaly time series of the individual models and NMME mean with observations (red line) overlaid for total and Y2Y SIE. Given the results of the previous section show highest skill for forecasts up to 5-month lead, these five lead times are displayed leading up to the March or September forecast target month. A single individual member (gray lines) from a model is analogous to the observational time series. Arctic SIE observations contain some element of unpredictable noise plus a predictable signal from the initial conditions and drivers such as anthropogenic climate change and sea surface temperatures (Comiso et al. 2008; Parkinson 2014). Given enough members, the ensemble means (ranges shown by blue shading) from each model result in noise cancellation and therefore isolate the signal. That the observations sometimes deviate from the spread of the ensemble mean is expected, as the real-world SIE variability also contains noisy fluctuations that are neither forced nor predictable. However, within a well-calibrated model, it is expected that the observational data should lie within the spread of individual members. A corollary of this is that, ideally, the variance of well-sampled observations should match the average of the variances from each ensemble member. If there are clear divergences from these aspects, then providing reliable, probabilistic predictions of Arctic sea ice is a challenge that may necessitate improving the model itself or applying suitable statistical corrections.

1) MARCH SIE

For predictions of March total SIE (Fig. 6a), the observations mostly fall within the spread of the individual members for each model and the NMME. In general, the observations are located within the upper edge of the spread during the early part of the record, and then during the later period they are on the lower edge of the spread. Because of averaging, it is not expected that the ensemble mean would share the same amplitude as the observations. However, the fact that the observations lie in the upper range (or outside) the spread of individual members during the beginning of the period and at the bottom range at the end indicates potential deficiencies in reproducing

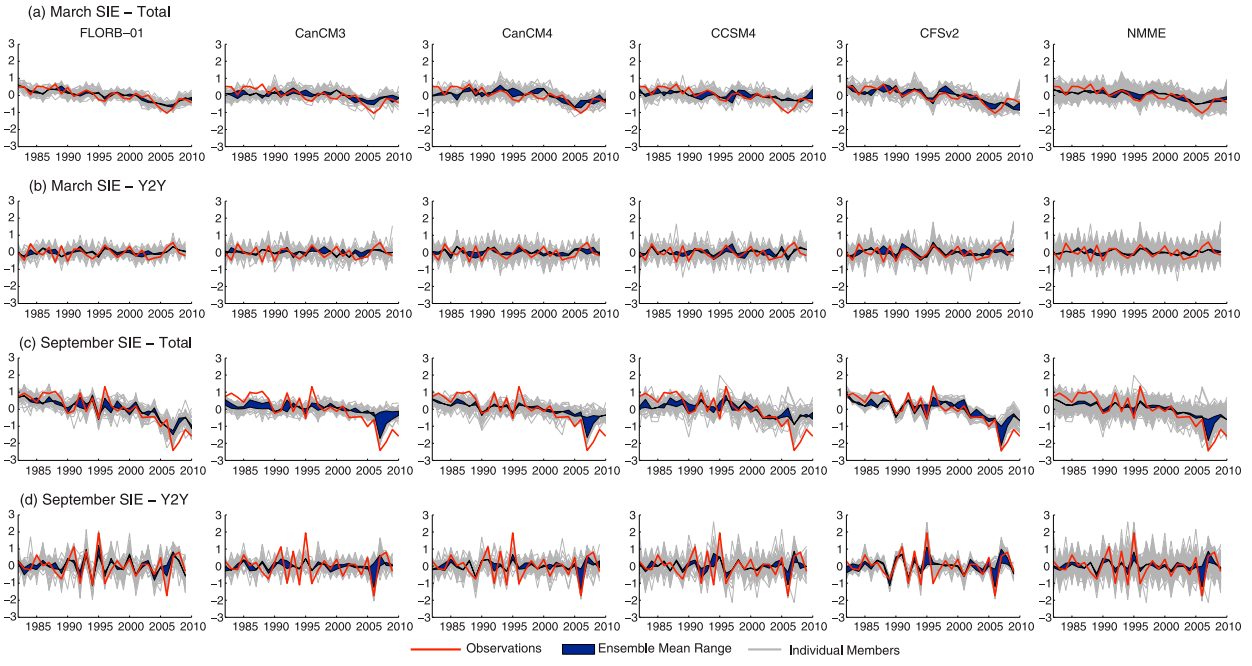

FIG. 6. Time series of SIE observation anomalies (red line) with the range of ensemble mean anomalies (1- to 5-month lead time; blue shading) and individual ensemble member anomalies (gray lines) for each individual model and the NMME mean: (a) total March SIE, (b) Y2Y March SIE, (c) total September SIE, and (d) Y2Y September SIE. Units for SIE are 10^6 km^2.

observed SIE trends. Figure 7 shows the linear trend for each lead time and observations over 1982–2010. For March targets after lead 1, the CFSv2 ensemble mean (blue line) and spread of individual members (gray lines) consistently capture the observed trend especially after lead 1. FLORB-01 also appears to reproduce the observations with smaller errors in lead 1 relative to CFSv2. Finally, with the exception of one or two individual members at lead 5, the trends from CanCM3, CanCM4, CCSM4, and the NMME are all less negative than the observations. The smaller trends in CanCM3 and CanCM4 have been previously documented as a consequence of the initialization source of HadISST1.1 (Merryfield et al. 2013b). Not only do these trend values indicate a slower decline in SIE, they are around a quarter to half as steep as the observations. The ensemble mean from CCSM4 even has a slight positive trend at lead 1. To examine how the spread, or variance, of the individual members varies by lead time, Fig. 8 compares the standard deviation of the individual members for each model and the NMME mean as a bar graph to the observations (straight dashed line). Overall, each system is relatively similar, but slightly underdispersive to the observed total March SIE value.

For March Y2Y SIE, the observed trends lie consistently within the individual member spread (Fig. 6b). However, in contrast to the March total SIE, the variance (Fig. 8) of the individual members is almost double the observed variance across the models and the NMME at all lead times. While it is possible that the observational dataset itself has too little variance due to measurement or algorithm errors, it also indicates that the models are overdispersive for March Y2Y forecasts, most likely resulting in underconfident probabilistic forecasts. Thus, properly calibrating the forecast probabilities is an important avenue for future work.

2) SEPTEMBER SIE

Predictions for September total SIE anomalies (Fig. 6c) show that, across all models, there are periods of time when the observations clearly lie outside of the spread of the model predictions. This is particularly true in 1996 and in the years following the large sea ice melt in 2007, when the NMME and models underestimated the degree of sea ice loss. With decreased sea ice thickness, longer melt seasons, and warmer winters (Serreze and Stroeve 2015) contributing to a steeper trend in the later record, it is possible the underestimated trend in the NMME is increasingly a function of poor initial conditions (e.g., lack of sea ice thickness information playing a larger role). Similar to March total SIE anomalies (Fig. 6a), the observations start near the upper edge of the ensemble spread during the early part of the record, and then reside near the lower edge of the spread near the end of the period. Again this

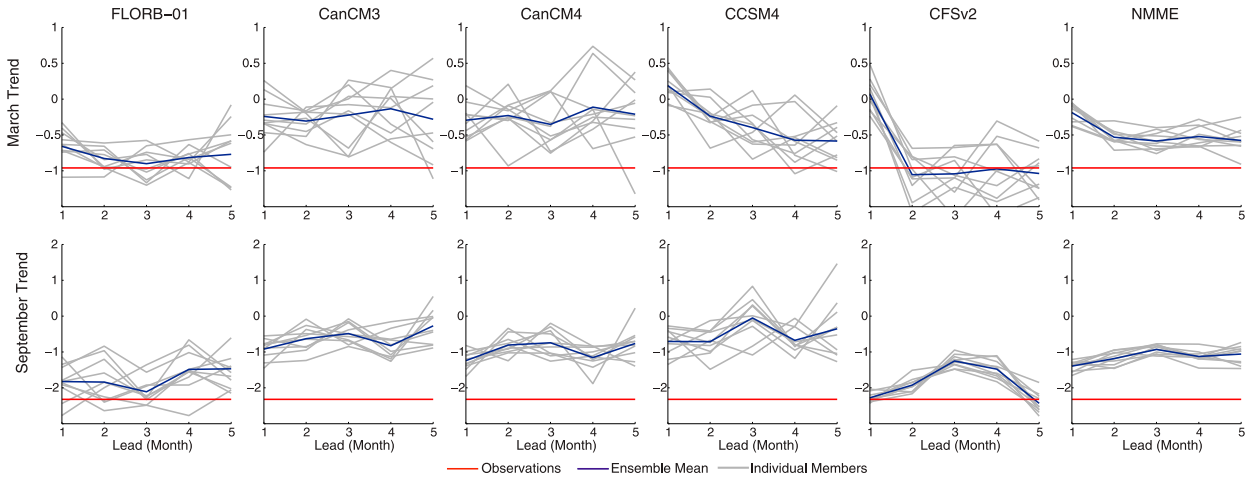

FIG. 7. Trend vs lead time from the time series in Fig. 6. The ensemble mean is shown in blue with the individual ensemble members in gray. The red horizontal line is the observed trend. Units for SIE trend are 10^6 km^2.

indicates that the individual models and NMME mean are not capturing the magnitude of the SIE decrease in the most recent years. As for March, FLORB-01 and CFSv2 are closest to the observed September trend. Also, CanCM3, CanCM4, CCSM4, and NMME individual members and ensemble mean have trends of roughly half of the observed values.

Contrary to March, the variance of individual models and the NMME ensemble for total September SIE (Fig. 8) is generally half the observed variance at all lead times. This is in part due to the inability of the models to capture the observed trends. While for March Y2Y SIE, the models show an overdispersion, for total September SIE the models suggest underdispersion. As a result of the underdispersion and errors in CFSv2, Collow et al. (2015) reran CFSv2 based on a modified version of the model that assimilates sea ice data from PIOMAS, which includes detailed information on sea ice thickness and a more realistic marine stratus cloud scheme. While the 1982–2010 CFSv2 reforecasts were not regenerated and are not shown here, their modified CFSv2 has produced significant improvement for real-time September SIE predictions indicating the importance of including sea ice thickness into model initializations.

Finally, for Y2Y September SIE (Fig. 6d), the observations largely lie within the spread of the individual model member forecasts including the more extreme Y2Y years (e.g., 1996, 2007). Similar to the Y2Y March values, the observations generally follow the center of the spread, with the ensemble mean following the same pattern as the observations. Clearly, the two largest Y2Y changes in 1996 and 2007 were not captured by the ensemble mean in every model, but they largely lie within the manifold of individual ensemble members, indicating that they were partially a consequence of natural internal variability. The models mostly capture the observed Y2Y variance (Fig. 8), with CanCM3 and FLORB-01 more significantly departing from the observations.

c. *Changes in forecast skill*

The final analysis seeks to examine the temporal changes in the forecast skill over the hindcast period. The time series in Fig. 9 show sliding 10-yr windows of the ACC and RMSE skill metrics for the months of September and March and for both total or Y2Y SIE. For example, the label 1995 indicates the range from 1986 to 1995. Each line represents a different forecast lead time, with darker blue lines indicating shorter leads and lighter yellows indicating the longer leads. The *t*-test significance at the 95% level is represented as filled circles within the significant windows. In the left column of Fig. 9, ACC values for March (both total and Y2Y) show considerable variability in correlations over time, with little dependence on lead times. Significance is found when total ACC values are larger than around 0.5 for lead 1, 4, and 7 before 2005 and most leads after 2006. Likewise, March Y2Y ACC is significant at lead 7 before 2000 and lead 4 in 2002 and 2003. In contrast, September ACC demonstrates some dependence on forecast lead time, with relatively constant, high, and significant ACC through lead 4. While ACC are lower, generally not significant, and more variable at longer lead times. When comparing ACC averages for 1983–2002, 1993–2012, and 2003–12 in the FLORB-01 predecessor, Msadek et al. (2014) found a decrease in September anomaly correlation for linearly detrended SIE after 2002. They attributed this skill degradation to decreasing ice thickness, which is less predictable (Holland et al. 2011). Within the sliding windows, a similar decrease is not found in September Y2Y correlation with the

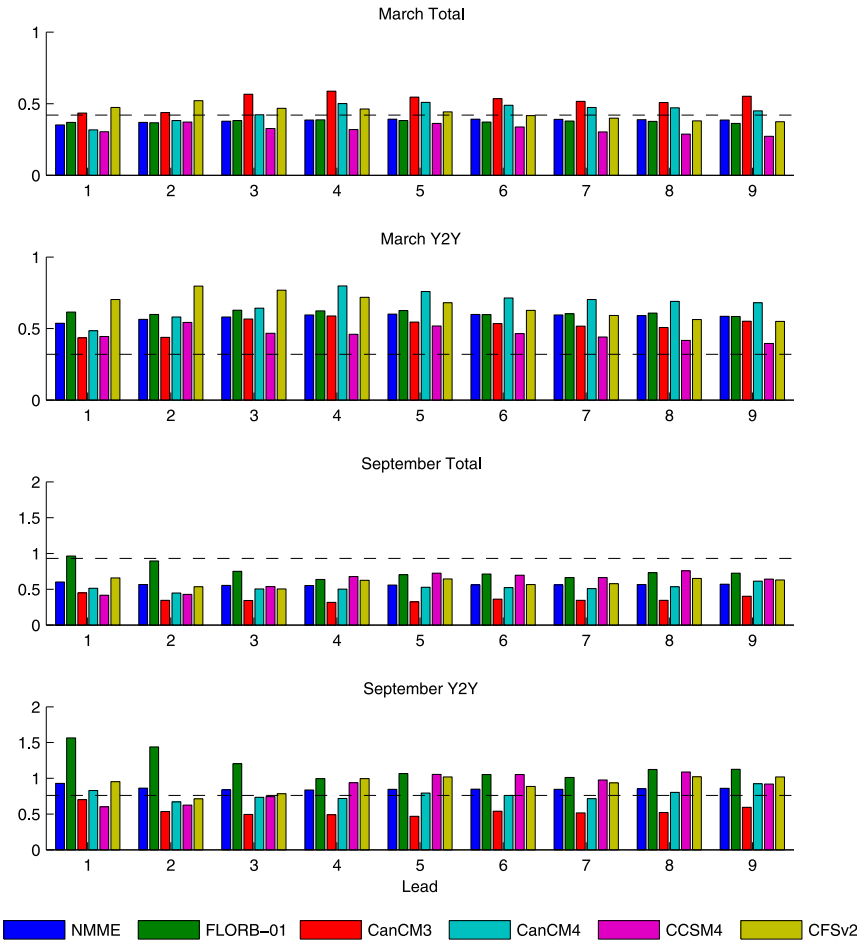

FIG. 8. Standard deviation of individual ensemble members averaged for each lead time for each of the models (colors). The black dashed line is the observational value. Units for SIE standard deviation are 10^6 km^2.

NMME, but instead notice a significant increase in total RMSE over the most recent decade.

For March, there is little to no lead-time dependence in RMSE (Fig. 9, right column) with values between 0.2 and 0.5 for all windows and leads. Significance is only seen for the first three leads before 2000 and after 2006 for Y2Y RMSE. This is likely due to smaller trends and Y2Y variability during the month when SIE is maximized. RMSE values for September, in contrast to March, are more dependent lead time, with smaller errors for shorter forecast leads. There are also changes in the RMSE values across the period of record with significance throughout most of the timespan. The RMSE for total September SIE are relatively low and constant between 0.4 and 0.6 during most of the record early period. Starting in the 2007 window, there is an increase to over 0.6 with longer leads near 0.8 until the end of the data record. In contrast, the Y2Y errors decrease toward the end of the record. Because Y2Y changes are mostly independent of longer-term trends, this suggests that errors in the prediction of total SIE are increasing over time due to the NMME not adequately capturing the trend in recent years. However, this conclusion is made with a relatively short data record. It is also interesting to note that the total September RMSE is occasionally larger than the Y2Y RMSE. This is also corroborated by the results of Fig. 6, which clearly demonstrate the deviation of the observations from the predictions for September in the later portion of the time period. Overall, the implication is the trend contributes to these errors in NMME, especially given the significant acceleration in SIE loss documented in several past studies (Stroeve et al. 2012; Comiso et al. 2008).

4. Summary

This study is the first to make use of a currently operational multimodel forecasting system, the NMME, to

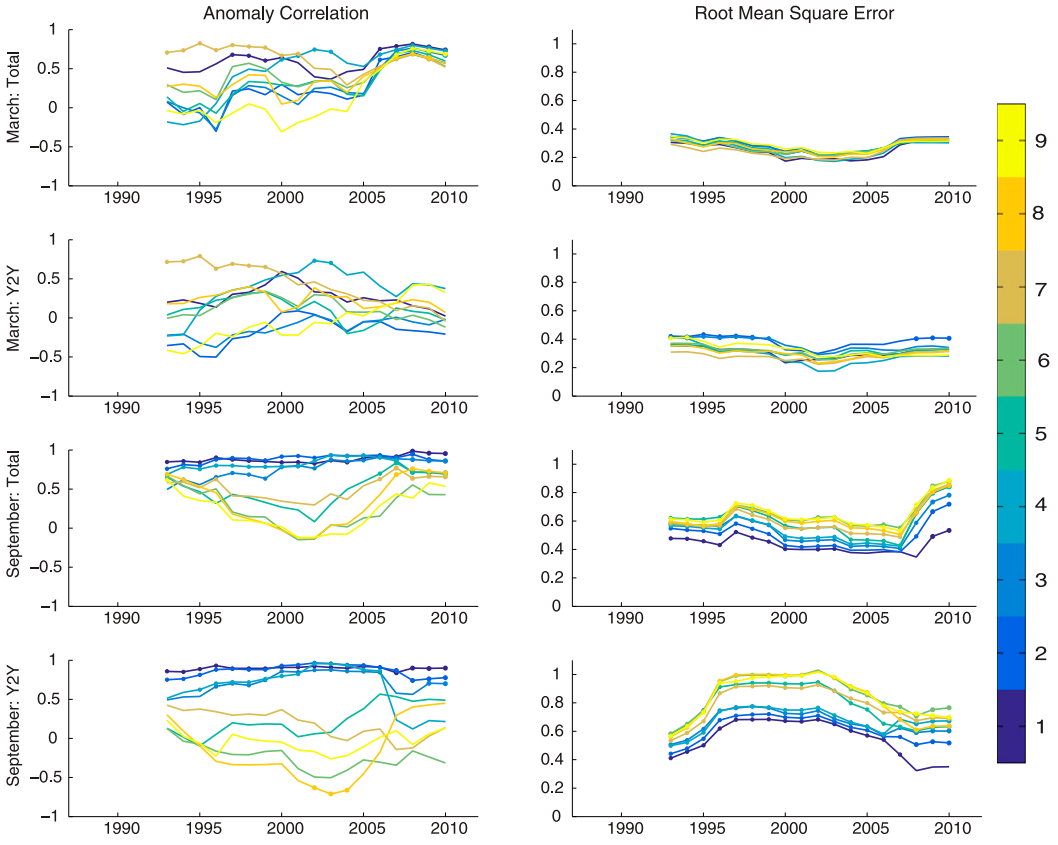

FIG. 9. Time series of (left) ACC and (right) RMSE for NMME ensemble mean at each lead time out to 9 months (line colors). Dots indicate 95% statistically significant values.

assess the skill of seasonal SIE. To our knowledge, this study also uniquely documents, using a long reforecast record (1982–2010) how the prediction skill has changed over time, finding that September total SIE errors have increased in the latter part of the record. Results indicate that, as a real-time system, complemented by long reforecasts, the NMME would be able to improve upon the overall skill of individual model systems. It also would be equipped to flag recent changes in skill, which is instrumental given the rapidly changing conditions in the Arctic and heightens the odds that past skill may not be an indicator of the future.

A major advantage of the NMME system is that the multimodel average provides more skillful predictions over any individual model (e.g., Merryfield et al. 2013b; Blanchard-Wrigglesworth et al. 2015). The NMME average demonstrates lower bias for predictions of total SIE, reduced RMSE for both Y2Y and detrended variability, and increased correlations of Y2Y SIE variability. In fact, NMME gives around a 10% reduction in RMSE for all the metrics examined. Given these improvements, future work should focus on isolating the common sources of skill, which may be determined by their ability to initialize sea ice and oceanic conditions and also capture atmosphere–ocean sea ice coupling over preceding seasons (e.g., Ding et al. 2017; Blanchard-Wrigglesworth et al. 2011; Sigmond et al. 2013). While detrended (both linear and quadratic) SIE is used in most studies to quantify interannual variability (e.g., Msadek et al. 2014; Wang et al. 2013; Merryfield et al. 2013b; Sigmond et al. 2013; Dirkson et al. 2017), the accelerating trend in SIE makes it difficult to separate interannual variability from longer-term, fitted trends (Serreze and Stroeve 2015). Although Y2Y SIE is slightly less skillful than detrended SIE, this measure isolates real-time interannual variability independent of an a posteriori fitted linear trend.

Initialization has been shown to heavily influence prediction skill in individual models (Blanchard-Wrigglesworth et al. 2015; Msadek et al. 2014). Each model contributing to the NMME has a different approach to creating the initial sea ice conditions. While this study did not rerun any hindcast simulations with improved initializations, findings presented here corroborate past studies that have shown that forecast biases may relate to the initial conditions (Msadek et al. 2014). However, biases due to imperfect model physics cannot be ignored especially

since biases remain for most of the models long after initialization. Here, results show CanCM3 and CM4 have similar biases across target month and lead times. The fact that these models have differing atmospheric physics but use the same initialization points to the influence of sea ice initialization in contributing to model bias, with a note that the ocean model physics may influence as well. Msadek et al. (2014) also shows the importance of initialization when FLORB-01 is compared to a lower-resolution ocean model. SIE still has similar skill despite the different models. This suggests that the quality of operational sea ice observations and assimilation systems needs to be improved in order to increase SIE skill. Chevallier et al. (2017) also note that biases within reanalyses have an impact on the forecast simulations. One way to improve initialization may be to focus on inputs of sea ice thickness. Day et al. (2014) outline the influence of sea ice thickness and determine that accurate sea ice thickness initialization impacts forecasts up to 8-month lead. Blanchard-Wrigglesworth et al. (2011) and Chevallier and Salas-Mélia (2012) show that, due to persistence in sea ice thickness, the summer sea ice minimum may have predictability at lead times of 6 months or longer. This improvement in prediction is also documented in Collow et al. (2015) and Dirkson et al. (2017) when initializing CFSv2 and CanCM3with PIOMAS sea ice thickness resulted in significant improvement in predictions.

September anomaly time series show that models tend to overestimate SIE during the latter part of the 1982–2010 hindcasts, especially after the 2007 September sea ice minimum (L'Heureux et al. 2008). Blanchard-Wrigglesworth et al. (2015) point out that more recent summer sea ice extent may be less predictable in the more recent period possibly due to reduced sea ice thickness (Holland et al. 2011). Also, the switch in initial condition sources from the NASA team algorithm to the NCEP operational analysis in 1997 may have influenced the biases seen in the recent periods, as noted in the CFSv2 (Wang et al. 2013). The struggle of the models to predict the following year SIE change, along with the increasingly larger errors for September SIE predictions in recent decades, suggests that prediction of the trend remains a fundamental challenge for most coupled modeling systems. However, the short reforecast record after the 2007 decrease in skill should be noted when extrapolating a continued trend. Skill degradation in models over the recent decades is a cause for concern and should be monitored in real-time forecast runs and extensions of the hindcast.

While this study gives an overview of NMME sea ice for the entire Arctic domain, future work could focus on evaluating the skill of regional sea ice extent and concentration. Already, using SIPN models, Blanchard-Wrigglesworth et al. (2017) note that forecast uncertainty is highest along the Arctic coastlines. Goessling et al. (2016) also suggest focusing on the predictability of the sea ice edge, but their results show that the ice edge is less predictable than SIE, especially in September. Finally, Bushuk et al. (2017) and Sigmond et al. (2016) show that skillful regional prediction of SIE is highly sensitive to specific regions, but generally exceeds the skill of a persistence forecast.

Acknowledgments. We are thankful for support from NOAA Climate Program Office/CVP Grant NA15OAR4310162. The NMME project is supported by NOAA, NSF, NASA, and DOE, with help from NCEP, IRI, and NCAR personnel who maintain the NMME archive.

REFERENCES

Barnston, A.G., M. K. Tippett, M. Ranganathan, and M. L'Heureux, 2017: Deterministic skill of ENSO predictions from the North American Multimodel Ensemble *Climate Dyn.*, https://doi.org/10.1007/s00382-017-3603-3.

Becker, E., H. M. van den Dool, and Q. Zhang, 2014: Predictability and forecast skill in NMME. *J. Climate*, **27**, 5891–5906, https://doi.org/10.1175/JCLI-D-13-00597.1.

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

——, R. I. Cullather, W. Wang, J. Zhang, and C. M. Bitz, 2015: Model forecast skill and sensitivity to initial conditions in the seasonal Sea Ice Outlook. *Geophys. Res. Lett.*, **42**, 8042–8048, https://doi.org/10.1002/2015GL065860.

——, and Coauthors, 2017: Multi-model seasonal forecast of Arctic sea-ice: Forecast uncertainty at pan-Arctic and regional scales. *Climate Dyn.*, **49**, 1399–1410, https://doi.org/10.1007/s00382-016-3388-9.

Bushuk, M., R. Msadek, M. Winton, G. A. Vecchi, R. Gudgel, A. Rosati, and X. Yang, 2017: Skillful regional prediction of Arctic sea ice on seasonal timescales. *Geophys. Res. Lett.*, **44**, 4953–4964, https://doi.org/10.1002/2017GL073155.

Cavalieri, D. J., C. L. Parkinson, P. Gloersen, and H. J. Zwally, 1996 (updated yearly): Sea ice concentrations from *Nimbus-7* SMMR and DMSP SSM/I-SSMIS passive microwave data, version 1. [1981–2015]. NASA National Snow and Ice Data Center Distributed Active Archive Center, accessed 19 May 2016, https://doi.org/10.5067/8GQ8LZQVL0VL.

Chevallier, M., and D. Salas-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**, 3025–3038, https://doi.org/10.1175/JCLI-D-11-00209.1.

——, ——, A. Voldoire, and M. Déqué, 2013: Seasonal forecasts of the pan-Arctic sea ice extent using a GCM-based seasonal prediction system. *J. Climate*, **26**, 6092–6104, https://doi.org/10.1175/JCLI-D-12-00612.1.

——, and Coauthors, 2017: Intercomparison of the Arctic sea ice cover in global ocean–sea ice reanalyses from the ORA-IP project. *Climate Dyn.*, **49**, 1107–1136, https://doi.org/10.1007/s00382-016-2985-y.

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**, 4618–4630, https://doi.org/10.1175/MWR-D-15-0097.1.

Comiso, J. C., 2000 (updated 2015): Bootstrap sea ice concentrations from *Nimbus-7* SMMR and DMSP SSM/I-SSMIS, version 2 [1981–2015]. NASA National Snow and Ice Data Center Distributed Active Archive Center, accessed 19 May 2016, https://doi.org/10.5067/J6JQLS9EJ5HU.

——, C. L. Parkinson, R. Gersten, and L. Stock, 2008: Accelerated decline in the Arctic sea ice cover. *Geophys. Res. Lett.*, **35**, L01703, https://doi.org/10.1029/2007GL031972.

Day, J. J., E. Hawkins, and S. Tietsche, 2014: Will Arctic sea ice thickness initialization improve seasonal forecast skill? *Geophys. Res. Lett.*, **41**, 7566–7575, https://doi.org/10.1002/2014GL061694.

DelSole, T., X. Yang, and M. K. Tippett, 2013: Is unequal weighting significantly better than equal weighting for multi-model forecasting? *Quart. J. Roy. Meteor. Soc.*, **139**, 176–183, https://doi.org/10.1002/qj.1961.

——, J. Nattala, and M. K. Tippett, 2014: Skill improvement from increased ensemble size and model diversity. *Geophys. Res. Lett.*, **41**, 7331–7342, https://doi.org/10.1002/2014GL060133.

Ding, Q., and Coauthors, 2017: Influence of the recent high-latitude atmospheric circulation change on summertime Arctic sea ice. *Nat. Climate Change*, **7**, 289–295, https://doi.org/10.1038/nclimate3241.

Dirkson, A., W. J. Merryfield, and A. Monahan, 2017: Impacts of sea ice thickness initialization on seasonal Arctic sea ice predictions. *J. Climate*, **30**, 1001–1017, https://doi.org/10.1175/JCLI-D-16-0437.1.

Fučkar, N. S., V. Guemas, N. C. Johnson, F. Massonnet, and F. J. Doblas-Reyes, 2016: Clusters of interannual sea ice variability in the Northern Hemisphere. *Climate Dyn.*, **47**, 1527–1543, https://doi.org/10.1007/s00382-015-2917-2.

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**, 1642–1650, https://doi.org/10.1002/2015GL067232.

Grumbine, R. W., 1996: Automated passive microwave sea ice concentration analysis at NCEP. NCEP Ocean Modeling Branch Tech. Note 120, 13 pp., http://polar.ncep.noaa.gov/seaice/icegroup.shtml.

Guemas, V., and Coauthors, 2016: A review on Arctic sea ice predictability and prediction on seasonal-to-decadal timescales. *Quart. J. Roy. Meteor. Soc.*, **142**, 546–561, https://doi.org/10.1002/qj.2401.

Hagedorn, R., F. J. Doblas-Reyes, and T. N. Palmer, 2005: The rationale behind the success of multi-model ensembles in seasonal forecasting—I. Basic concept. *Tellus*, **57A**, 219–233, https://doi.org/10.1111/j.1600-0870.2005.00103.x.

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. *Climate Dyn.*, **36**, 1239–1253, https://doi.org/10.1007/s00382-010-0792-4.

Infanti, J. M., and B. P. Kirtman, 2016: Prediction and predictability of land and atmosphere initialized CCSM4 climate forecasts over North America. *J. Geophys. Res.*, **121**, 12 690–12 701, https://doi.org/10.1002/2016JD024932.

Jahn, A., and Coauthors, 2012: Late twentieth-century simulation of Arctic sea ice and ocean properties in the CCSM4. *J. Climate*, **25**, 1431–1452, https://doi.org/10.1175/JCLI-D-11-00201.1.

Jia, L., and Coauthors, 2015: Improved seasonal prediction of temperature and precipitation over land in a high-resolution GFDL climate model. *J. Climate*, **28**, 2044–2062, https://doi.org/10.1175/JCLI-D-14-00112.1.

Kirtman, B. P., and Coauthors, 2014: The North American Multi-Model Ensemble (NMME): Phase-1 seasonal to interannual prediction; phase-2 toward developing intra-seasonal prediction. *Bull. Amer. Meteor. Soc.*, **95**, 585–601, https://doi.org/10.1175/BAMS-D-12-00050.1.

L'Heureux, M. L., A. Kumar, G. D. Bell, M. S. Halpert, and R. W. Higgins, 2008: Role of the Pacific–North American (PNA) pattern in the 2007 Arctic sea ice decline. *Geophys. Res. Lett.*, **35**, L20701, https://doi.org/10.1029/2008GL035205.

Lindsay, R., and A. Schweiger, 2015: Arctic sea ice thickness loss determined using subsurface, aircraft, and satellite observations. *Cryosphere*, **9**, 269–283, https://doi.org/10.5194/tc-9-269-2015.

——, J. Zhang, A. J. Schweiger, and M. A. Steele, 2008: Seasonal predictions of ice extent in the Arctic Ocean. *J. Geophys. Res.*, **113**, C02023, https://doi.org/10.1029/2007JC004259.

Merryfield, M. J., W.-S. Lee, G. J. Boer, V. V. Kharin, J. F. Scinocca, G. M. Flato, R. S. Ajayamohan, and J. C. Fyfe, 2013a: The Canadian Seasonal to Interannual Prediction System. Part I: Models and initialization. *Mon. Wea. Rev.*, **141**, 2910–2945, https://doi.org/10.1175/MWR-D-12-00216.1.

——, ——, W. Wang, and A. Kumar, 2013b: Multi-system seasonal predictions of Arctic sea ice. *Geophys. Res. Lett.*, **40**, 1551–1556, https://doi.org/10.1002/grl.50317.

Msadek, R., G. A. Vecchi, M. Winston, and R. G. Gudgel, 2014: Importance of initial conditions in seasonal predictions of Arctic sea ice extent. *Geophys. Res. Lett.*, **41**, 5208–5215, https://doi.org/10.1002/2014GL060799.

Notz, D., 2014: Sea-ice extent and its trend provide limited metrics of model performance. *Cryosphere*, **8**, 229–243, https://doi.org/10.5194/tc-8-229-2014.

Parkinson, C. L., 2014: Global sea ice coverage from satellite data: Annual cycle and 35-yr trends. *J. Climate*, **27**, 9377–9382, https://doi.org/10.1175/JCLI-D-14-00605.1.

Rayner, N. A., D. E. Parker, E. B. Horton, C. K. Folland, L. V. Alexander, D. P. Rowell, E. C. Kent, and A. Kaplan, 2003: Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *J. Geophys. Res.*, **108**, 4407, https://doi.org/10.1029/2002JD002670.

Saha, S., and Coauthors, 2010: The NCEP Climate Forecast System Reanalysis. *Bull. Amer. Meteor. Soc.*, **91**, 1015–1057, https://doi.org/10.1175/2010BAMS3001.1.

——, and Coauthors, 2014: The NCEP Climate Forecast System version 2. *J. Climate*, **27**, 2185–2208, https://doi.org/10.1175/JCLI-D-12-00823.1.

Serreze, M. C., and J. Stroeve, 2015: Arctic sea ice trends, variability, and implications for seasonal ice forecasting. *Philos. Trans. Roy. Soc.*, **373A**, 20140159, https://doi.org/10.1098/rsta.2014.0159.

Sigmond, M., J. C. Fyfe, G. M. Flato, V. V. Kharin, and W. J. Merryfield, 2013: Seasonal forecast skill of Arctic sea ice area in dynamical forecast system. *Geophys. Res. Lett.*, **40**, 529–534, https://doi.org/10.1002/grl.50129.

——, M. C. Reader, G. M. Flato, W. J. 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**, 12 457–12 465, https://doi.org/10.1002/2016GL071396.

Stroeve, J. C., V. Kattsov, A. Barrett, M. Serreze, T. Pavlova, M. Holland, and W. N. Meier, 2012: Trends in Arctic sea ice

extent for CMIP5, CMIP3, and observations. *Geophys. Res. Lett.*, **39**, L16502, https://doi.org/10.1029/2012GL052676.

——, 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**, 2411–2418, https://doi.org/10.1002/2014GL059388.

Swart, N. C., Fyfe, J. C., Hawkins, E., Kay, J. E. and Jahn, 2015: Influence of internal variability on Arctic sea-ice trends. *Nat. Climate Change*, **5**, 86–89, https://doi.org/10.1038/nclimate2483.

Tippett, M. K., M. Ranganathan, M. L'Heureux, A. G. Barnston, and T. DelSole, 2017: Assessing probabilistic predictions of ENSO phase and intensity from the North American Multimodel Ensemble. *Climate Dyn.*, https://doi.org/10.1007/s00382-017-3721-y.

van den Dool, H., 2006: *Empirical Methods in Short-Term Climate Prediction*. Oxford University Press, 240 pp.

Walsh, J. E., 1980: Empirical orthogonal functions and the statistical predictability of sea ice extent. *Sea Ice Processes and Models,* R. S. Pritchard, Ed., University of Washington Press, 373–384.

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**, 1375–1394, https://doi.org/10.1175/MWR-D-12-00057.1.

Wettstein, J. J., and C. Deser, 2014: Internal variability in projections of twenty-first-century Arctic sea ice loss: Role of the large-scale atmospheric circulation. *J. Climate*, **27**, 527–550, https://doi.org/10.1175/JCLI-D-12-00839.1.