This is a post-peer-review, pre-copyedit version of an article published in Climate Dynamics. The final authenticated version is available online at: https://doi.org/10.1007/s00382-020-05273-8

Summer predictions of Arctic sea ice edge in multi-model seasonal re-forecasts

Authors: Lauriane Batté[1], Ilona Välisuo[1], Matthieu Chevallier[1,2], Juan C. Acosta Navarro[3], Pablo Ortega[3] and Doug Smith[4]

Abstract

In this study, the forecast quality of 1993-2014 summer seasonal predictions of five global coupled models, of which three are operational seasonal forecasting systems contributing to the Copernicus Climate Change Service (C3S), is assessed for Arctic sea ice. Beyond the Pan-Arctic sea ice concentration and extent deterministic re-forecast assessments, we use sea ice edge error metrics such as the Integrated Ice Edge Error (IIEE) and Spatial Probability Score (SPS) to evaluate the advantages of a multi-model approach.

Skill in forecasting the September sea ice minimum from late April to early May start dates is very limited, and only one model shows significant correlation skill over the period when removing the linear trend in total sea ice extent. After bias and trend-adjusting the sea ice concentration data, we find quite similar results between the different systems in terms of ice edge forecast errors. The highest values of September ice edge error in the 1993-2014 period are found for the sea ice minima years (2007 and 2012), mainly due to a clear overestimation of the total extent. Further analyses of deterministic and probabilistic skill over the Barents-Kara, Laptev-East Siberian and Beaufort-Chukchi regions provide insight on differences in model performance.

For all skill metrics considered, the multi-model ensemble, whether grouping all five systems or only the three operational C3S systems, performs among the best models for each forecast time, therefore confirming the interest of multi-system initiatives building on model diversity for providing the best forecasts.

Affiliations: (1) CNRM UMR 3589, Université de Toulouse, Météo-France, CNRS, Toulouse, France; (2) Direction des Opérations pour la Prévision, Département Marine et Océanographie, Météo-France, Toulouse, France; (3) Barcelona Supercomputing Centre (BSC), Barcelona, Spain; (4) Met Office Hadley Centre, Exeter, UK

Correspondence: lauriane.batte@meteo.fr

ORCID (L Batté): 0000-0002-7903-9762

ORCID (I Välisuo): 0000-0002-8665-4990

ORCID (M. Chevallier): 0000-0003-2033-166X

ORCID (J. C. Acosta Navarro):0000-0001-5375-0639

ORCID (P. Ortega): 0000-0002-4135-9621

ORCID (D. Smith): 0000-0001-5708-694X

1. Introduction

In recent decades, Arctic sea ice extent has significantly decreased, while exhibiting important year-to-year variability, sparking interest in the polar prediction community in the provision of forecasts for sea ice conditions at the sub-seasonal to decadal time scales.

Arctic sea ice extent anomalies have both local and remote impacts. At a local or regional scale, sea ice formation or melt can have consequences on the livelihood of local communities, shipping activities, fisheries and infrastructure safety (Eicken, 2013). Moreover, several works (see e.g. Vihma, 2014 and Jung et al. 2015 and references therein) have suggested evidence for remote effects of sea ice cover on atmospheric variability in the midlatitudes, and the accelerated warming of the Arctic is thought by some studies to bring more persistent Northern Hemisphere weather regimes favouring extremes (Coumou et al. 2018, Francis et al. 2018), although this is still widely debated (Blackport et al. 2019). One example of such remote effects is the link between autumn sea ice concentration over the Barents-Kara seas and the subsequent winter North Atlantic Oscillation index which could explain a significant part of climate variability over the North Atlantic sector at weekly to annual time scales (Garcia-Serrano et al. 2015).

Mechanisms for Arctic sea ice predictability have been highlighted by past studies (see Guemas et al. 2016 for a review). These include the advection of anomalous sea ice conditions, the atmospheric circulation over the Arctic, ocean heat transport and thermohaline circulation, but also persistence of anomalies in the initial sea ice state. In the last decade, potential predictability studies using global coupled models (GCMs) in a perfect model framework, such as that of the APPOSITE project (Tietsche et al. 2014) have provided estimates of theoretical seasonal-to-decadal sea ice prediction skill limits in current-generation climate models. Using different metrics, Day et al. (2014) and Chevallier and Salas y Mélia (2012) estimated e-folding times of 2-5 months for total sea ice area in GCMs, depending on the initial date, and consistent with that of observations such as NSIDC (Blanchard-Wrigglesworth et al. 2011). For sea ice volume, significant levels of potential predictability were found up to three years ahead (Day et al. 2014, Cruz-Garcia et al. 2019). Yet the skill of initialized seasonal hindcasts often fall short of these potential predictability estimates (Guemas et al. 2016).

Reliable prediction of total Arctic sea ice extent is in itself a challenge, but a correct sea ice extent value may mask some large compensating errors in the presence or absence of ice. Indeed, sea ice predictability and prediction skill may vary depending on the region of interest (Germe et al. 2014, Bushuk et al. 2017), with different processes at play. Cruz-Garcia et al. (2019) highlighted using EC-

Earth perfect-model simulations that predictability in the Atlantic sector peripheral seas was linked to local sea surface temperature and ocean heat content anomalies. In the case of summer sea ice predictions, initial sea ice thickness is found to be a precursor for sea ice extent over the East Siberia, Laptev, Beaufort and Chukchi Seas (Bushuk et al. 2017; Bushuk et al. 2019). To circumvent the possible overestimation of skill using total sea ice extent, more challenging metrics that assess ice edge errors were suggested by Goessling et al. (2016) and Goessling and Jung (2018) to evaluate model skill in forecasting the position of the sea ice.

With thinning ice and a warmer atmosphere over the region, the melt season is particularly challenging to forecast at such extended time ranges, where drivers of variability are dominated by chaotic processes (Serreze and Stroeve, 2015; Olonscheck et al. 2019). Past studies have found that SIE potential predictability estimated from GCM simulations drops faster in predictions initialized from May than from July (Day et al. 2014), although these conclusions seem to be model-dependent for the pan-Arctic SIE (see e.g. Bushuk et al. 2019). Bonan et al. (2019) found evidence for this loss in predictive capacity in GCMs in the Arctic marginal seas between June and May starts by analyzing correlation between sea ice area and sea ice volume from previous months of CMIP5 preindustrial control runs. This springtime "predictability barrier" is also consistent with evaluations of empirical forecasts based on observational data (Walsh et al. 2019). Moreover, initialized predictions often perform at substantially lower skill levels than those estimated in potential predictability studies (Guemas et al. 2016; Bushuk et al. 2019).

One approach to try and bridge the gap between potential and actual forecast skill is to combine single-model forecasts into a multi-model ensemble. Since 2008, the Sea Ice Outlook initiative (Blanchard-Wrigglesworth et al. 2017) has collected several sources of forecasts (statistical, dynamical and heuristic) for the September Arctic sea ice minimum extent at three to one month lead times. A recent study by Wayand et al. (2019) has demonstrated the current capabilities - and limitations - of such state of the art forecasting systems in predicting sea ice concentration and thickness during the 2018 melt season and the better performance of the multi-model on sub-seasonal time scales. At the seasonal scale, past studies have demonstrated the interest of combining individual forecasting systems into a multi-model ensemble for atmospheric fields (e.g. Hagedorn et al. 2005) as a way of improving the signal-to-noise ratio of ensemble forecasts. Merryfield et al. (2013) showed that the combination of CanSIPS and CFSv2 seasonal forecast systems led in most cases to improved sea ice concentration forecast skill over the Arctic. Dirkson et al. (2019a) recently provided new evidence of the additional skill of multi-model combinations over single models for September sea ice concentration using six different state-of-the-art seasonal forecasting systems. In the framework of the H2020-APPLICATE project, which aims to broaden the understanding of linkages between the Arctic region and the Northern Hemisphere mid-latitudes and improve models over these regions, several seasonal re-forecasts were run using state-of-the-art coupled climate models initialized in May and November, over a period covering at least 22 years. These were evaluated alongside re-forecasts from operational climate prediction centers involved in the project, contributing to the Copernicus Climate Change Services (C3S) initiative. We focus here on predictions initialized at the end of April or May to assess the skill of these models in forecasting summer Arctic sea ice concentration and extent anomalies. We also further investigate the added value of a multi-model approach for sea ice forecasts grouping the models from the APPLICATE project.

A complete description of the models and the skill metrics used for the evaluation is presented in section 2. Section 3 describes the main results at a pan-Arctic scale, whereas section 4 focuses on specific Arctic seas or regions. Limitations to the study and future developments for the different forecast systems are discussed in section 5.

2. Models and methods

2.1 Seasonal re-forecasts and reference data

The present study focuses on boreal summer re-forecasts initialized on May 1 or late April. Seasonal re-forecasts from five GCMs are evaluated: the model developed jointly by Centre National de Recherches Météorologiques (CNRM) and Cerfacs for the sixth phase of the Coupled Model Intercomparison Project (CMIP6) called CNRM-CM6-1; European Consortium Earth system model version 3.2 (hereafter EC-Earth3.2), as well as re-forecasts provided in the framework of the Copernicus Climate Change Service (C3S) from three operational systems: Met Office fifth global seasonal forecasting system GloSea5, Météo-France sixth generation seasonal forecast system System 6 and the European Centre for Medium-Range Weather Forecasts fifth generation seasonal forecast system SEAS5. Table 1 presents information on these different sets of seasonal re-forecasts regarding the coupled model components, their resolution, and the initial conditions for the ocean and ice components. All re-forecasts use the Nucleus for European Modelling of the Ocean (NEMO, Madec et al. 2017) ocean model, albeit with different model versions and settings; however, the sea ice components differ amongst the GCMs.

The CNRM-CM6-1 GCM is described in detail in Voldoire et al. (2019). This version of the GCM uses Arpege-Climate v6.3 for the atmosphere and NEMO 3.6 – Gelato v6 for the ocean and sea ice. The land surface component is Surfexv8. Coupling between atmosphere/land and ocean is called in the Surfex interface using the OASIS-MCT code. This GCM was used to run seasonal re-forecast experiments initialized on May 1st 1993-2014. An ensemble of 30 members was constructed by combining three ocean and sea ice initial conditions with 10 initial perturbations of the ERA-Interim atmospheric initial conditions for Arpege. The ocean and sea ice components are initialized from a run constrained towards the Mercator Ocean International Glorys 2V4 reanalysis (Ferry et al. 2010), using the same NEMO-Gelato model versions as in the GCM. The sea ice component adjusts to the atmospheric forcing and ocean constraints, except for sea ice concentration which is relaxed towards Glorys.

Re-forecasts from a second GCM, EC-Earth3.2, are also evaluated in this study. EC-Earth3.2 is based on ECMWF's atmospheric circulation model IFS (cycle 36r4) and land surface model H-Tessel coupled with the OASIS-3 coupler to the ocean model NEMO 3.6 including the Louvain-la-Neuve Sea Ice Model LIM3 (Vancoppenolle et al., 2009). As for CNRM-CM6-1, re-forecasts are initialized on May 1st 1993-2014. The ensemble size is of 10 members, generated with random perturbations of ERA-Interim initial conditions. The LIM3 model is initialized in these re-forecasts using a standalone NEMO-LIM3 run forced with atmospheric fluxes calculated from the Drakkar Forcing Set (DFS, Brodeau et al., 2010), and assimilating sea ice concentrations using an Ensemble Kalman Filter approach (Massonnet et al., 2014). The ocean initial conditions are interpolated from the ocean reanalysis system ORAS4 reanalysis (Balmaseda et al., 2013).

Alongside these re-forecasts, three operational seasonal forecast systems from the Copernicus Climate Change Service (C3S) are analyzed for the same forecast times. ECMWF SEAS5 (Johnson et al., 2019) is based on the IFS cy43r1 atmospheric model directly coupled to the NEMO 3.4 ocean and LIM2 sea ice (Fichefet et al., 1997) components. The 25-member ensemble is generated using both initial condition and stochastic perturbations of the atmosphere. Ocean and sea ice are initialized from the ocean reanalysis system ORA-S5 part of the operational OCEAN5 analysis (Zuo et al., 2019).

Met Office GloSea5 (MacLachlan et al. 2015) re-forecasts were also analyzed, building a 28-member ensemble re-forecast from the 7-member ensembles initialized on the 9th, 17th, 25th of April and 1st of May. GloSea5 also uses NEMO 3.4 but the Los Alamos sea ice model CICE 4.1. Note that these two forecast systems use a higher resolution ocean and sea ice (1/4°) than the other models in this study.

A third system from the C3S program was also included in the analysis, Météo-France seasonal forecast system 6 (MF-Sys6). This system is based on a very similar model version of the CNRM-CM coupled model as CNRM-CM6-1 described previously, but runs at a higher resolution in the atmosphere. 25 ensemble members are generated using atmospheric stochastic perturbations (Batté and Déqué, 2016) and a lagged initialization, with 12 members initialized on the 20th of April, 12 on the 25th of April, and one control member on the 1st of May. Ocean and sea ice initial conditions are derived similarly to CNRM-CM6-1 from a run constrained towards Glorys 2V4, except for sea ice concentration which evolves freely in the NEMO-Gelato run to initialize MF-Sys6.

Beyond the coupled GCMs used, the re-forecasts compared in this study use different initialization strategies and ensemble generation techniques. This can impact the ensemble spread and forecast quality.

The study evaluates monthly mean sea ice extent (SIE) derived from sea ice concentration (SIC), using the 15% SIC threshold to define presence or absence of sea ice. These fields are compared to reference data provided by the National Snow and Ice Data Center (NSIDC) version 4, based on brightness temperature (Cavalieri et al. 1996).

Throughout the re-forecast period, NSIDC SIC data is missing in some areas north of 85°N. This could have some influence on skill evaluations especially when computing area-averaged scores, we therefore chose to consider gridpoint SIC data from 45°N to 85°N, masking out regions from 85°N to 90°N in our computations over the Pan-Arctic region.

2.2 Re-forecast bias adjustment

Due to model imperfection and initial error growth, re-forecasts based on GCMs are prone to systematic errors and drift when forecast time increases. This makes bias adjustment of re-forecasts a necessary step before the evaluation of forecast quality using the metrics described in the following section.

For Pan-Arctic SIE, we chose to bias-correct the ensemble mean SIE of each individual model against NSIDC using a leave-one-out cross-validation bias correction. The metrics shown therefore evaluate the skill of the model SIE anomalies versus NSIDC SIE anomalies, irrespective of the mean SIE bias.

In the case of metrics based on Arctic sea ice edge position, we compare in this study two straightforward methods for bias-adjusting the grid-point SIC values. The first method consists in bias-correcting (BC) the SIC values using (as for total SIE) leave-one-out cross-validation against NSIDC SIC. This simple method has some caveats, since for bounded fields such as SIC values it can yield values outside the theoretical range. We simply correct "out of bounds" values by setting negative SIC values to 0 and SIC values higher than 100% to 100%.

The second method uses also a leave-one-out cross-validation, but to trend-adjust (TA) the data: we adjust the SIC of each model (either ensemble mean or member) at a given grid-point as well as

NSIDC data for a linear trend. In this study we chose to remove the linear trend and then compute anomalies with respect to the 1993-2014 mean. The obtained SIC values are then adjusted to the [0,1] range before computation of the indices described in the following section. Note that more elaborate trend-adjustment techniques for SIC have been introduced in past works such as Dirkson et al. (2019b).

2.3 Re-forecast evaluation metrics

After evaluating (and removing) the mean model biases for SIC and SIE with respect to NSIDC, two types of verification metrics are used in this study. In section 3.2, total Pan-Arctic SIE re-forecast skill is evaluated according to forecast time using standard deterministic scores such as root mean square error (RMSE) and correlation. Benchmark skill for SIE is assessed using the persistence of April SIE anomalies.

We then focus on the skill of the models in representing the position of the sea ice edge, using the Integrated Ice Edge Error (IIEE, Goessling et al. 2016) and its probabilistic counterpart, the Spatial Probability Score (SPS, Goessling and Jung, 2018). These metrics take into account possible error compensations between overestimation and underestimation of the presence of ice over different basins of the Arctic, and therefore present a more complete analysis of the ability of GCMs to predict sea ice concentration at a seasonal time scale.

The IIEE is computed to evaluate the total spatial extent of errors in the position of the sea ice edge. The IIEE is the sum of areas where the presence of sea ice, defined with a 15% SIC threshold, is overestimated (O) and underestimated (U) with respect to reference data. Following Goessling et al. (2016), the IIEE is decomposed into two terms, namely misplacement error (ME) and absolute extent error (AEE), as follows:

$$IIEE = O + U = |O - U| + 2 \cdot min(O, U) = AEE + ME$$

The absolute error corresponds to the total Pan-Arctic SIE error when this metric is computed over the region, whereas the misplacement error shows the compensation between areas with overestimation and areas with underestimation.

In the case of the IIEE, two benchmark re-forecasts are considered depending on the bias-adjustment technique used. The benchmark re-forecast IIEE is computed for comparison with bias corrected re-forecasts (trend-adjusted re-forecasts, respectively) using a leave-one-out climatology (linear trend-adjusted climatology, respectively) of SIC NSIDC data.

A natural extension to the IIEE is used to examine the skill of probabilistic forecasts for presence of sea ice at a grid point level. The SPS consists of a spatial integral of the Brier Score for the probabilistic event of SIC exceeding the 15% threshold. With NSIDC data as a reference, and under the assumption that reference data is "perfect" and therefore not accounting for observational uncertainty, the SPS is formulated as follows:

$$SPS = \iint \left(P_{SIC_f > 0.15}(x, y) - 1_{SIC_o > 0.15}(x, y) \right)^2 dxdy$$

In this study, probabilities are computed by counting the fraction of ensemble members exceeding the 15% concentration threshold (with or without trend-adjustment), and then bias-corrected using leave-one-out cross-validation. For the benchmark probability re-forecasts, we consider probabilities based on the 21 other years of the re-forecast period, either with or without trend-adjustment.

The Brier Score (Brier, 1950) and its decomposition into reliability, resolution and uncertainty components (Murphy, 1972) are computed over regional seas for the probabilistic event of SIC exceeding the 15% concentration threshold. The positively-oriented Brier Skill Score (BSS) is used to determine model skill over using a simple climatology to forecast this probability. In this framework, reliability diagrams plotting binned forecast probabilities against mean observed frequencies for the event help estimate the conditional bias in probability space of the ensembles, and quantify how trustworthy these systems are on average over the re-forecast period (Weisheimer and Palmer, 2014).

Note that these metrics can be sensitive to the ensemble size of each model re-forecast, and differences in skill should therefore be interpreted with caution - in particular, for the EC-Earth 3.2 model, 10 members were available, significantly less than the 25-30 member ensemble sizes of the other models in this study.

2.4 Multi-model combination

The advantages of using a multi-model approach in seasonal forecasting have been demonstrated in many studies focusing on the predictability of atmospheric fields (e.g. Hagedorn et al., 2005). We compute here a simple multi-model combination of the different model re-forecasts by first bias-adjusting each model individually (either with the BC or TA methods), and then combining the members of each of the five models into an unweighted multi-model ensemble. This ensemble is called MME in what follows. Most models studied have a similar ensemble size, except for the EC-Earth 3.2 model. With the unweighted ensemble approach used in this study, this model is thus under-represented with respect to the others in the MME.

Since two of the operational systems considered in this study show higher correlation values and lower root mean square errors than in the other model re-forecasts, we also examine the skill of a multi-model restricted to the operational C3S systems, called C3S MME.

3. Pan-Arctic scale results

This section describes the ability and deficiencies of current state-of-the-art seasonal forecasting systems in reproducing summer Arctic sea ice concentration variability from May initializations.

3.1 Systematic errors in sea ice concentration and extent

Before focusing on integrated indices of hindcast quality, often computed after bias-correcting the individual ensemble forecasts, we first assess the model quality in terms of systematic errors in the raw model outputs for sea ice concentration.

Figure 1 shows the mean bias over the re-forecast period of month 1 (May) and month 5 (September) SIC with respect to NSIDC. Red areas show where SIC is too low in the models, whereas blue areas highlight where model have excessive SIC. All models show a common low bias in Labrador Sea SIC, already present in the reanalyses used to initialize the re-forecasts (Chevallier et al. 2017). Elsewhere, from the first month of simulation, the systems exhibit different behaviors. CNRM-CM6-1 has too low SIC along the ice edge in the Greenland sea, a feature shared with the operational MF-Sys6 which relies on a similar version of the CNRM-CM coupled model and initial conditions of the ocean and sea ice. The three other systems show too high SIC in the Iceland and Nordic seas at month 1. At longer lead times, both sets of re-forecasts based on CNRM-CM exhibit a substantially different bias than the other models, with too little SIC over most of the Arctic. This is likely due to the initialization strategy for the model, for which even at the initial stage, sea ice thickness is often too low. During the melt season, this results in an excessive reduction of SIC over most of the Arctic, and a subsequent loss in predictability. EC-Earth 3.2, SEAS5 and GloSea5 show similar patterns of systematic errors for September, particularly over the Beaufort-Chukchi and East Siberian sectors where SIC is too high at the end of the melt season. The largest differences between these three models are found north of the Greenland, Iceland and Norwegian (GIN) seas and Barents-Kara seas in September, where GloSea5 slightly under-estimates and SEAS5 slightly over-estimates SIC, while EC-Earth 3.2 has biases of opposite sign between the Barents and Kara sectors.

So as to evaluate the impact of these biases on total Pan-Arctic SIE re-forecasts, as well as the model spread 5 months after initialization, we show in Fig. 2 box-and-whisker plots of Pan-Arctic September SIE computed with raw model outputs (before bias-correction) for SIC. For each year of the common re-forecast period, the boxes show the interquartile range and spread of ensemble members, compared to SIE computed from NSIDC SIC data. Alongside this analysis, we also compute the linear trend in mean September SIE for each model as well as for NSIDC. Values are shown in Table 2.

The models exhibit different characteristics: consistent with results from Fig. 1, CNRM-CM6-1 (Fig. 2(a)) and MF-Sys6 (Fig. 2(e)) show a clear underestimation of September SIE for most years of the re-forecast, whereas the three other models show values in the observed range. However, SEAS5 SIE values are comparable to NSIDC in the beginning of the re-forecast period but are then overestimated with respect to NSIDC after 2006, due to a too weak negative trend in the re-forecast (about one third of the linear trend estimated in NSIDC). The four other models also underestimate the amplitude of the negative trend, but much less so, with values ranging from -83,000 to -96,000 squared kilometers per year of loss in Pan-Arctic SIE. This underestimation of the negative trend in SIE was also found, although over a different re-forecast period and using a different model, by Wang et al. (2013). In the case of GloSea5 and EC-Earth 3.2, SIE computed from NSIDC data are inside the range of the ensemble for almost all years of the re-forecast period. The spread of SEAS5 appears to be slightly lower than the other two operational seasonal re-forecast systems, GloSea5 and MF-Sys6. This could be due to the burst initialization strategy for SEAS5, whereas the other systems used a lagged ensemble approach with different ocean (and therefore sea ice) initial conditions.

Two consequences arise from these analyses. First of all, for most systems, it appears necessary to bias-correct the SIC values since large systematic errors are found (sometimes related to errors present from month 1 onwards). Second of all, the strong negative trend in SIC and hence SIE values means that in skill scores such as correlation, the trend may have an impact on results. In Table 2, results both with and without detrending SIE values are shown for each model and persistence of April SIE anomalies. Although some models have very large biases which translate into high RMSE before bias removal, they all exhibit correlation values before linear detrending above 0.65, with most systems reaching approximately 0.8. However, when removing the linear trend, it appears that most of this apparent skill is in fact related to correctly capturing the sign -and part of the amplitude -of the trend over the region. Levels of skill unrelated to trend are much more modest.

Unless mentioned otherwise, the skill evaluations presented in what follows are therefore computed for sea ice concentration ensemble re-forecasts that are linearly-detrended and bias corrected in cross-validation mode, as described in section 2.2. At this stage, we note that the choice of a linear trend may have some influence on results, but the short re-forecast period made more elaborate trend computations hazardous.

3.2 Pan-Arctic sea ice extent skill

As a first glimpse of the skill of different systems in re-forecasting sea ice conditions, we focus on Pan-Arctic sea ice extent RMSE and correlation over the 1993-2014 re-forecast period are shown in Fig. 3, and results for September summarized in Table 2.

The skill of individual systems is compared to a multi-model ensemble (MME) grouping all ensemble members of each system together (without weighting individual systems but with equal weight for each member). The skill of the MME is shown in black. Scores can be compared to a simple persistence approach (persisting SIE anomalies from April to the following months) for which results are shown in magenta. Most systems exhibit fairly similar levels of skill, both for RMSE and correlation. RMSE is maximum in September when SIE is at the minimum of the seasonal cycle. Correlation drops (as expected) with lead time, from above 0.8 in May to near-zero correlation for two of the models in October, namely CNRM-CM6-1 and EC-Earth 3.2, although in the case of the latter, this may be due to the smaller ensemble size. The three operational systems generally exhibit significant levels of correlation with NSIDC data at a 6-month lead time, although MF-Sys6 drops below the 95% significance threshold for August and September SIE (see Table 2 for September RMSE and correlation values). As expected from the results of the individual models, the C3S-MME (in orange) outperforms the MME for both metrics in the long forecast times. All models show higher skill than persistence, although the score for persistence is included inside the range of uncertainty of the scores (based on a χ^2 for RMS and a Fisher test for correlation) after 2 months forecast time in most cases (not shown). This is likely related to the limited number of re-forecast years in the evaluation.

Although not strictly comparable due to different re-forecast years, the results found for SIE correlation and RMSE are consistent with previous works: Wang et al. (2013) and Msadek et al. (2014) found similar performances with other re-forecast systems in terms of SIE correlation. Msadek et al. (2014) also showed evidence that skill tends to be lower in recent decades than over a longer re-forecast period spanning also the 1980s. More recently, Bushuk et al. (2019) showed with the GFDL-FLOR model a sharp drop in summer pan-Arctic SIE anomaly correlation for May initializations as early as June (see their Fig. 5).

3.3 Sea ice edge forecast quality

While seasonal forecasts of Pan-Arctic sea ice can provide some indication of below-average or above-average presence of sea ice, these may not be the most relevant indicators for potential end-users of seasonal forecast information. Among these users, some are most interested in the exact position of the sea ice edge, or its probability of presence along shipping routes or near the climatological sea ice edge (Melia et al. 2017).

We therefore evaluate the skill of the different models in representing the position of the sea ice edge (based on monthly averages) by computing the IIEE metric introduced by Goessling et al. (2016). This is first done after correcting SIC for systematic errors with a simple cross-validation bias removal.

Figure 4 shows the IIEE for each individual model for September 1993-2014, as well as for the MME and C3S-MME (after individual model bias correction). Results for the different models are quite similar, with IIEE increasing during the re-forecast period, mainly due to an increase in AEE. This positive trend in AEE is consistent with the models under-estimating the negative trend in SIE discussed previously.

Peaks in IIEE are found in 2007 and 2012 for each system, indicating that all models missed to some extent the record low SIE for both of these years. Conversely, in the first half of the re-forecast period, most models exhibit their highest IIEE for 1996 for which SIE was the highest of the 1993-2014 period. IIEE is (by construction) very sensitive to errors in forecast extrema. These results can be expected given the low predictability of such extrema, partly due to atmospheric conditions at synoptic scales which are inherently unpredictable at such large forecast times. However, in the case of the 2012 minimum, past studies using observational data and GCM experiments suggest that the role of an extreme summer storm over the Arctic was minor compared to sea ice preconditioning and warmer near-surface atmospheric temperature conditions during the summer season (Zhang et al. 2013, Guemas et al. 2013).

The operational systems show generally less variability in the misplacement error than CNRM-CM6-1 and EC-Earth 3.2, apart from MF-Sys6 in the first half of the re-forecast period. This suggests that for the former, skill evaluations based on RMSE of Pan-Arctic SIE are giving a rather accurate picture of the model capacity to predict the sea ice edge position, whereas for the latter two systems, the Pan-Arctic SIE may "hide" some compensation between areas where SIC is overestimated and where it is underestimated.

For some systems, IIEE tends to grow during the re-forecast period, which may be related to the strong decrease in total SIE during 1993-2014. We therefore re-compute the IIEE score after trend-adjusting the SIC as described in section 2.2. Results are shown in Fig. 5. With this SIC trend adjustment, the minimum over the period is 2007 (and 2012 no longer appears as a year with low SIE). All models miss the 2007 anomaly with a large AEE. As found previously using SIC data corrected for the mean bias, the AEE is the largest contribution (on average) for September IIEE in all systems.

In order to evaluate the impact of trend-adjustment on IIEE, and also extend the analysis to the other months of the re-forecasts, we show in Fig. 6 the mean evolution as a function of forecast time of the IIEE in the different models and both MMEs considered, using bias-corrected SIC data (left) and trend-adjusted SIC data (right). Trend adjustment does improve the mean IIEE values for most systems, although some seem to benefit far more from this technique than others. For instance, focusing again on the month of September, the CNRM-CM6-1 model forecasts are clearly improved, whereas EC-Earth 3.2 and SEAS5 IIEE are only slightly reduced. It is also worth noticing in Fig. 6 that both the MME (in black) and C3S MME (in orange) exhibit very similar results in terms of IIEE, and improve all the individual forecasts for almost every forecast month, irrespective of the bias adjustment technique used. When compared to SIE RMSE values in Fig. 3, IIEE values (which are dominated by the AEE term) exhibit substantially higher values. Although the computation method for total sea ice extent differs between Fig. 3 and Figs. 5-6, this suggests that further improvements would be found with more sophisticated bias correction and trend-adjustment techniques.

The IIEE peaks in September when the total SIE is lowest, and the opposite sign in the evolution of average IIEE and SIE during summer is quite striking. Some models did seem to exhibit (to some degree) a return of skill in terms of correlation and RMSE between September and October (see Fig. 3) which is also suggested by the decrease in IIEE.

Ensemble forecasts bear the advantage that information can be provided in probabilistic form to potential users. This is most useful when the forecast is associated with a potential risk and corresponding losses for the user, as different courses of action may be undertaken depending on a given probability threshold, and at the time scales considered in this study, forecasts are rarely yes/no answers as they bear intrinsic uncertainties. We therefore focus in the following paragraph on the probabilistic extension of the IIEE, the SPS.

3.4 Probabilistic re-forecasts of sea ice edge

We compute the SPS using monthly SIC re-forecasts and a 15% SIC threshold for presence or absence of sea ice, with two approaches to bias-adjust the model data over the re-forecast period. The first method used is a grid-point bias correction of the probabilities for each model (or multi-model) to exceed the 0.15 threshold, using leave-one-out cross-validation. Probabilities exceeding 1 or below 0 are readjusted to 1 or 0, respectively. The second method uses the same adjustment of probabilities, but computes these after applying the trend adjustment to the SIC values. Figure 7 shows results according to the forecast month for each individual model as well as the 5-model and C3S MMEs. Consistent with deterministic results for the IIEE, the MMEs rank among the best models (low SPS) for each forecast time with the first adjustment, and tend to outperform all the individual systems after the second adjustment, but skill scores are not significantly better. Year-to-year values for SPS are not shown, since very limited inter-annual variability in SPS is found - setting aside the 1996, 2007 or 2012 cases during which SIE over the Pan-Arctic region reached local extrema.

As found previously, the SIC trend adjustment technique helps further improve skill levels. This is particularly striking in the case of the CNRM-CM6-1 model, which suffered from large systematic errors in SIC at longer forecast times, and shows that despite these issues some predictive skill remains.

In most cases, by comparing Fig. 7 with Fig. 6, it appears that the SPS values are clearly lower than the corresponding IIEE. This suggests that in areas where uncertainty is high, the spread in the models tends to reduce the probabilities for presence of sea ice, so models are generally not too overconfident. These considerations prompted the analysis of model reliability shown in the following section, which focuses on the skill at a regional level.

4. Regional skill

In this section we focus on skill of the different models over different key regions of the Arctic. Based on previous results, we target our analysis on the IIEE for sub-basins, as well as the Brier Score and reliability and resolution components to characterize probabilistic skill. Note that the SPS is a spatially weighted Brier Score for the event of SIC exceeding the 0.15 threshold set in this study to define the presence of sea ice. Our analysis focuses on the extended Beaufort-Chukchi Seas and Laptev-East Siberian Seas sectors, as well as the Barents-Kara region, where all models exhibit strong biases at forecast month 5.

Figure 8 shows the models and multi-models IIEE after trend-adjustment as a function of forecast time over the Beaufort-Chukchi Seas (a), Laptev-East Siberian Seas (b), and the Barents and Kara seas (c). For the Beaufort-Chukchi and Laptev-East Siberian seas, all models exhibit similar evolutions with forecast time, quite similar to the total SIE over the region. As for the total Pan-Arctic region, IIEE is maximum when the SIE is minimum, and then drops in October. May IIEE is close to zero for each system, suggesting that models are correctly initialized as fully ice-covered over these regions and error grows quite slowly initially. In the Laptev-East Siberian seas sector, the October IIEE is very similar for all systems due to the annual cycle of sea ice extent over the region: only a few years in NSIDC data show some ice-free areas in these seas, and they are generally not captured by the different forecasting systems at such long forecast ranges (not shown). The MME and C3S MME IIEE values nearly overlap (black and orange lines), indicating that the CNRM-CM6-1 and EC-Earth 3 models do not necessarily provide additional value to the multi-model approach in these areas. Skill is very limited compared to re-forecasts based on a simple linear trend climatology (Clim in Fig. 8), although more systems outperform this empirical forecast over the Beaufort-Chukchi sector than over the Laptev-East Siberian seas.

Over the Barents and Kara sector, the IIEE evolution exhibits a quite different behavior. From the first month of the re-forecasts, some errors in the ice edge are found in the different systems, leading to a first peak of IIEE in July for which the IIEE amounts to almost half of the total SIE over the area. However, the error of the linear trend climatology forecast is higher than each system from May to July, and higher than most systems up to October. This suggests that although predictability of summer sea ice over the Barents and Kara seas is very limited, dynamical systems do provide some information on SIE beyond simple empirical forecasts.

We evaluate the probabilistic skill in forecasting the presence of ice by plotting reliability diagrams for each region, alongside the Brier Skill Score (BSS) and reliability and resolution components of the Brier Score for the event of SIC in the grid cell exceeding 0.15. The probabilistic forecasts are evaluated after bias correcting or trend-adjusting the SIC data as previously described for the Pan-Arctic SPS. Since over the 1993-2014 period, most of the Barents-Kara Seas region is ice-free in September, we show results for the Beaufort-Chukchi and Laptev-East Siberian Seas regions only. Results for the C3S operational re-forecasts and the C3S MME for September over the Beaufort-Chukchi region are shown in Fig. 9. Comparing the top and bottom rows, we find that trend adjustment improves the reliability and resolution of the forecasts. This translates into higher BSS for each system. Unlike SEAS5 and GloSea5 (Fig. 9 e-f), MF-Sys6 (g) almost systematically underestimates the probabilities of presence of ice, whereas the other systems tend to have too high forecast probability values with respect to the observed occurrence of the event (SEAS5 more dramatically so than GloSea5).

Over the Laptev and Siberian Seas (Fig. 10), trend adjustment noticeably improves the reliability of all systems considered, with reliability diagrams closely fitting the perfect reliability diagonal. Out of the three operational systems, MF-Sys6 is the one which is most improved after trend adjustment, since using a simple bias correction led in this case to practically no skill over climatology in predicting the presence of ice over the region. All systems (including the C3S MME) have very similar levels of resolution after trend adjustment, demonstrating the interest of correcting for the trend in sea ice concentration before formulating probabilistic forecasts for the presence of ice.

5. Summary and discussion

In this study, a comprehensive multi-system ensemble was evaluated for boreal summer predictions of sea ice by grouping re-forecasts from three operational systems with two ensembles with current generation GCMs (CNRM-CM6-1 and EC-Earth 3.2). The common re-forecast period, 1993-2014, coincides with the highest trends in sea ice concentration and extent over the Pan-Arctic region. The focus of this study was on sea ice concentration and extent, and using metrics designed to assess the ability of models to represent the position of the sea ice edge. A companion study by Acosta Navarro et al. (2020) examines the link in these models between fall Arctic sea ice and Northern Hemisphere boreal winter atmospheric seasonal forecast skill.

Models exhibit diverse levels of forecast quality and ability to reproduce tendencies in sea ice extent estimated with NSIDC data. Beyond this comparison, a multi-model approach either grouping all five models or the three operational C3S systems does not lead to major improvements, especially with respect to the best systems, besides reliability and resolution components when investigating probabilistic skill over the Beaufort-Chukchi and Laptev-East Siberian seas. However, either the MME or C3S MME rank systematically among the two best models at all lead times and cases examined, which pleads in favor of model diversity, and is consistent with pan-Arctic and regional evaluations of probabilistic skill of single-models and multi-models discussed in Dirkson et al. (2019a). Yet some model deficiencies leading to strong biases or errors in trends do seem to alter the MME skill. This highlights the need for a careful bias correction and trend adjustment of current state-of-the-art forecasting systems, as a necessary first step before using such predictions. These results confirm the limited predictability of summer Arctic sea ice with current state-of-the-art GCMs, especially at longer forecast times such as five months ahead of the September sea ice minimum. They are consistent with recent results suggesting a “spring predictability barrier” in prediction skill (e.g. Bonan et al. 2019).

One major limitation to statistical post-processing and adjustments of forecasts is the very restricted number of years available for the evaluation of seasonal forecast biases and skill. The use of linear trends over longer time periods may quickly show some limitations, especially with bounded variables such as sea ice concentration. Director et al. (2017) emphasized the limitations related to such bias correction techniques which can lead to unrealistic sea ice edges, and designed a contour shifting method which corrects using linear regression the position of the sea ice edge. Dirkson et al. (2019b) found additional improvements in terms of probabilistic skill scores when fitting the sea ice concentration distribution to a parametric distribution and applying a trend-adjusted quantile mapping correction. These methods would likely further enhance skill scores of the systems evaluated in this study. However, given the coarse common spatial resolution used, the restricted number of re-forecast years, and the same statistical treatment applied to our benchmark forecasts, we are confident the bias and trend adjustment applied yield results that are representative of the actual capacity of these models to forecast summer Arctic sea ice over simple empirical approaches. Another source of possible error in the estimation of skill levels is the reference data used. The NSIDC data is known to have some uncertainties, in particular during the summer season where melt ponds can be interpreted as ice free areas, but was chosen so as to provide a fair comparison between systems (since none were initialized directly from this dataset).

The diverse levels of skill likely arise from differences in the sea ice initialization and modeling strategies, as suggested by recent works on the S2S scale by Zampieri et al. (2018) and Wayand et al. (2019). In particular, the results found for re-forecasts based on CNRM-CM (either CNRM-CM6-1 or MF-Sys6) show substantially lower skill than in previous studies (e.g. Chevallier et al. 2013). Ongoing evaluation of these systems show that they exhibit from the first month of the re-forecasts lower sea ice thickness than reference datasets. The initialization of sea ice thickness has been identified by recent works as a source of predictability on seasonal time scales, either by direct assimilation (Blockley et al., 2018) or constraining SIT with SIC (Kimmritz et al., 2019). Other important processes for the melt season, such as melt ponds, are still only partially represented in models used in this study. Some pathways for improvement of current systems are currently explored in the framework of the APPLICATE project and will hopefully contribute to better and more robust forecasts of Arctic sea ice at the seasonal time scale in years to come.

Acknowledgements and data

This study was partly funded by the H2020-APPLICATE project, EU grant number 727862. Copernicus C3S seasonal re-forecast data for ECMWF SEAS5, Met Office GloSea5 and Météo-France System 6 were retrieved using the Mars API. The reforecasts with the EC-Earth 3.2 and CNRM-CM6-1 models were run as part of the H2020-APPLICATE project and data is available on the APPLICATE data portal or upon request.

JCAN acknowledges the Spanish Ministry of Science, Innovation and Universities for the personal grant Juan de la Cierva FJCI-2017-34027, PRACE for awarding access to MareNostrum at Barcelona Supercomputing Center (BSC), and ESA/CMUG-CCI3 for financial support. PO work was funded by the Ramon y Cajal grant RYC-2017-22772.

The authors wish to acknowledge F. Massonnet who developed the sea ice assimilation technique used to initialize EC-Earth 3.2, and S. Tietsche for pointing to the relevant SEAS5 data.

All plots and graphs were realized using R, and some skill evaluations and data detrending were computed using the s2dverification package available on CRAN (Manubens et al. 2018).

The authors would also like to thank two anonymous reviewers, whose feedback helped substantially improve this article.

This is a post-peer-review, pre-copyedit version of an article published in Climate Dynamics. The final authenticated version is available online at: http://dx.doi.org/10.1007/s00382-020-05273-8

References

Acosta-Navarro, J. C., P. Ortega, L. Batté, D. Smith, P.-A. Bretonnière, V. Guemas et al. (2020) Link between autumnal Arctic Sea ice and Northern Hemisphere winter forecast skill, *Geophys. Res. Lett.*, accepted.

Balmaseda, M. A., K. Mogensen, and A. T. Weaver (2013) Evaluation of the ECMWF ocean reanalysis system ORAS4, *Q. J. Roy. Meteor. Soc*., 139, 1132–1161, doi: 10.1002/qj.2063

Batté, L. and M. Déqué (2016) Randomly correcting model errors in the ARPEGE-Climate v6. 1 component of CNRM-CM: applications for seasonal forecasts, *Geosc. Mod. Dev.* 9 (6), doi: 10.5194/gmd-9-2055-2016.

Blackport, R., J. A. Screen, K. van der Wiel and R. Bintanja (2019) Minimal influence of reduced Arctic sea ice on coincident cold winters in mid-latitudes. *Nature Climate Change*, 9(9), 697-704.

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. Clim.* 24: 231–250.

Blanchard-Wrigglesworth, E., A. Barthélemy, M. Chevallier, R. Cullather, N. S. Fuckar et al. (2017) Multi-model seasonal forecast of Arctic sea-ice: forecast uncertainty at pan-Arctic and regional scales. *Clim. Dyn.*, doi:10.1007/s00382-016-3388-9

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, 3419–3438. doi:10.5194/tc-12-3419-2018

Bonan, D. B., M. Bushuk, and M. Winton (2019). A spring barrier for regional predictions of summer Arctic sea ice. *Geophys. Res. Lett.* 46, 5937–5947. doi:10.1029/2019GL082947

Brodeau, L., Barnier, B., Treguier, A. M., Penduff, T., & Gulev, S. (2010). An ERA40-based atmospheric forcing for global ocean circulation models. *Ocean Modelling*, 31(3-4), 88-104.

Bushuk, M. et al. (2017) Skillful regional prediction of Arctic sea ice on seasonal timescales. *Geophys. Res. Lett.* 44, 4953–4964, doi:10.1002/2017GL073155

Bushuk, M., R. Msadek, M. Winton, G. Vecchi et al. (2019) Regional Arctic sea–ice prediction:

potential versus operational seasonal forecast skill. *Clim. Dyn.*, 52 (5-6), 2721-2743, doi: 10.1007/s00382-018-4288-y

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. Boulder, Colorado USA. *NASA National Snow and Ice Data Center Distributed Active Archive Center.* doi: 10.5067/8GQ8LZQVL0VL. (Accessed 20 February 2017)

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. Clim.* 25: 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. Clim.* 26(16): 6092-6104, doi: 10.1175/JCLI-D-12-00612.1.

Chevallier, M., G.C. Smith, F. Dupont et al. (2017). Intercomparison of the Arctic sea ice cover in global ocean–sea ice reanalyses from the ORA-IP project. *Clim Dyn* 49: 1107, doi: 10.1007/s00382-016-2985-y

Coumou, D., G. Di Capua, S. Vavrus, L. Wang and S. Wang (2018) The influence of Arctic amplification on mid-latitude summer circulation. *Nature Comm.* 9, 2959, doi: 10.1038/s41467-018-05256-8

Cruz-Garcia, R., V. Guemas, M. Chevallier and F. Massonnet (2019) An assessment of regional sea ice predictability in the Arctic ocean. *Clim. Dyn.*, doi: 10.1007/s00382-018-4592-6, *in press.*

Day, J.J., S. Tietsche and E. Hawkins (2014) Pan-Arctic and Regional Sea Ice Predictability: Initialization Month Dependence. *J. Clim.* 27: 4371–4390, doi: 10.1175/JCLI-D-13-00614.1.

Director, H. M., A. E. Raftery and C. M. Bitz (2017) Improved Sea Ice Forecasting through Spatiotemporal Bias Correction. *J. Clim.*, 30, 9493-9510, doi:10.1175/JCLI-D-17-0185.1

Dirkson, A., B. Denis, and W. J. Merryfield (2019a) A multimodel approach for improving seasonal probabilistic forecasts of regional Arctic sea ice. *Geophys. Res. Lett.*, 46, 10,844–10,853, doi:10.1029/2019GL083831

Dirkson, A., W. J. Merryfield and A. H. Monahan (2019b) Calibrated probabilistic forecasts of Arctic sea ice concentration. *J. Clim.*, 32, 1251-1271, doi:10.1175/JCLI-D-18-0224.1

Eicken, H. (2013) Arctic sea ice needs better forecasts. *Nature* **497**, 431–433, doi:10.1038/497431a

Ferry, N., Parent, L., Garric, G., Barnier, B., et al. (2010), 'Mercator global Eddy permitting ocean reanalysis GLORYS1V1 : Description and results', Mercator-Ocean Quarterly Newsletter 36, 15–27. https://www.mercator-ocean.fr/sciences-publications/mercator-ocean-journal/newsletter-36-data-assimilation-and-its-application-to-ocean-reanalyses/

Fichefet, T. and M. A. Maqueda (1997) Sensitivity of a global sea ice model to the treatment of ice thermodynamics and dynamics, *J. Geophys. Res.-Oceans*, 102, 12609–12646, doi: 10.1029/97JC00480

Francis, J. A., N. Skific, S. J. Vavrus (2018) North American Weather Regimes Are Becoming More Persistent: Is Arctic Amplification a Factor? *Geophys. Res. Lett.* 45, 11,414-11,422, doi:10.1029/2018GL080252

García-Serrano, J., et al. (2015) On the predictability of the winter Euro-Atlantic climate: lagged influence of autumn Arctic sea ice. *Journal of Climate* 28 (13): 5195-5216.

Germe, A., M. Chevallier, D. Salas y Mélia, E. Sanchez-Gomez and C. Cassou (2014) Interannual predictability of Arctic sea ice in a global climate model: regional contrasts and temporal evolution. *Climate Dynamics*, **43** (9-10), 2519-2538, doi: 10.1007/s00382-014-2071-2

Goessling, H.F. et al. (2016) Predictability of the Arctic sea ice edge. *Geophys. Res. Lett.* 43, 1642-1650, doi:10.1002/2015GL067232

Goessling, H.F. and T. Jung (2018) A probabilistic verification score for contours: Methodology and application to Arctic ice-edge forecasts. *Q. J. R. Meteorol. Soc.*, doi:10.1002/qj.3242, *in press.*

Guemas V., F. Doblas-Reyes, A. Germe, M. Chevallier and D. Salas y Mélia (2013) September 2012 Arctic sea ice minimum : Discriminating between sea ice memory, the August 2012 extreme storm and prevailing warm conditions [in "Explaining Extreme Events of 2012 from a Climate Perspective"], *Bull. Amer. Meteor. Soc.*, 94 (9), S20-S22, doi: 10.1175/BAMS-D-13-00085.1

Guemas, V., E. Blanchard-Wrigglesworth, M. Chevallier, J. J. Day, M. Déqué et al. (2016) A review on Arctic sea-ice predictability and prediction on seasonal to decadal time-scales. *Q. J. R. Meteorol.*

Soc. 142: 546-561, doi: 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 A: Dynamic Meteorology and Oceanography,* 57A, 219-233, doi: 10.3402/tellusa.v57i3.14657

Johnson, S., T. N. Stockdale, L. Ferranti, M. A. Balmaseda, F. Molteni et al. (2019) SEAS5: the new ECMWF seasonal forecast system. *Geosci. Model Dev.*, 12, 1087–1117, doi: 10.5194/gmd-12-1087-2019

Jung, T., et al. (2015) Polar lower-latitude linkages and their role in weather and climate prediction. *Bull. Amer. Meteor. Soc.*, **96**, ES197–ES200, doi: 10.1175/BAMS-D-15-00121.1.

Kimmritz, M., F. Counillon, L. H., 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 Sy.*, 11, 4147–4166. doi:10.1029/2019MS001825

MacLachlan, C., A. Arribas, K. A. Peterson, A. Maidens, D. Fereday, A. A. Scaife, M. Gordon, M. Vellinga, A. Williams, R. E. Comer, J. Camp, P. Xavier and G. Madec (2015) Global Seasonal forecast system version 5 (GloSea5): a high-resolution seasonal forecast system. *Q. J. R. Meteorol. Soc.*, 141, 1072-1084, doi:10.1002/qj.2396

Madec G., R. Bourdallé-Badie, P.-A. Bouttier et al (2017) NEMO ocean engine. doi: 10.5281/ZENODO.1472492

Manubens, N., L.-P. Caron, A. Hunter, O. Bellprat, E. Exarchou et al. (2018) An R package for climate forecast verification. *Env. Mod. Soft.,* 103, 29-42, doi: 10.1016/jenvsoft.2018.01.018

Massonnet, F., H. Goosse, T. Fichefet and F. Counillon (2014), Calibration of sea ice dynamic parameters in an ocean-sea ice model using an ensemble Kalman filter, *J. Geophys. Res. Oceans*, 119, 4168–4184, doi:10.1002/2013JC009705

Melia N., K. Haines, E. Hawkins and J. J. Day (2017) Towards seasonal Arctic shipping route predictions. *Environ. Res. Lett.*, 12, 084005, doi: 10.1088/1748_9326/aa7a60

Merryfield, W. J., 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, doi: 10.1002/grl.50317.

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

Olonscheck, D., T. Mauritsen and D. Notz (2019) Arctic sea-ice variability is primarily driven by atmospheric temperature fluctuations. *Nature Geoscience,* 12: 430–434, doi: 10.1038/s41561-019-0363-1.

Serreze, M. C. and J. Stroeve (2015) Arctic sea ice trends, variability and implications for seasonal ice forecasting. *Phil. Trans. R. Soc. A*, 373: 20140159. doi:10.1098/rsta.2014.0159.

Tietsche, S., J.J. Day, V. Guemas, W. J. Hurlin, S.P.E. Keeley et al. (2014) Seasonal to interannual Arctic sea ice predictability in current global climate models. *Geophys. Res. Lett.*, 41, 1035-1043, doi: 10.1002/2013GL058755.

Vancoppenolle, M., T. Fichefet, H. Goosse, S. Bouillon, G. Madec, and M.A. Morales Maqueda (2009) Simulating the mass balance and salinity of Arctic and Antarctic sea ice. 1. Model description and validation. *Ocean Modelling*, 27, 33-53, doi : 10.1016/j.oceamod.2008.10.005.

Vihma, T. (2014) Effects of Arctic sea ice decline on weather and climate: a review. *Surv. Geophys.* **35**, 1175–1214.

Voldoire, A., D. Saint-Martin, S. Sénési, B. Decharme, A. Alias et al. (2019) Evaluation of CMIP6 DECK experiments with CNRM-CM6-1 *J. Adv. Model. Earth Sy.*, doi:10.1029/2019MS001683, *in press.*

Walsh J. E., J. S. Stuart and F. Fetterer (2019) Benchmark seasonal prediction skill estimates based on regional indices. *The Cryosphere*, 13, 1073–1088, doi:10.5194/tc-13-1073-2019

Wayand N. E., C. M. Bitz and E. Blanchard-Wrigglesworth (2019) A year-round subseasonal-to-seasonal sea ice prediction portal. *Geophys. Res. Lett.*, 46, doi: 10.1029/2018GL081565, *in press.*

Weisheimer A., and T. N. Palmer (2014) On the reliability of seasonal climate forecasts. *J. R. Soc. Interface*, 11: 20131162. doi:10.1098/rsif.2013.1162

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, doi:10.1002/grl.50190.

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 Sci.*, 15, 779–808, doi:10.5194/os-15-779-2019

Tables

Table 1: Characteristics of the seasonal re-forecasts evaluated. All systems are initialized with ERA-Interim for the atmosphere component.

Model/System	CNRM-CM6-1	EC-Earth 3.2.2	SEAS5	GloSea5	MF-Sys6
Atmosphere	Arpege 6.3	IFS Cy36r4	IFS Cy43r1	UM v6	Arpege 6.2
Ocean	NEMO 3.6	NEMO 3.6	NEMO 3.4	NEMO 3.4	NEMO 3.6
Sea ice	Gelato v6	LIM3	LIM2	CICE 4.1	Gelato v6
Atmospheric resolution	~1.4° 91 levels	~0.7° 91 levels	36 km 91 levels	~0.7° 85 levels	~0.5° 91 levels
Ocean/ice resolution	1° 75 levels	1° 75 levels	0.25° 75 levels	0.25° 75 levels	1° 75 levels
Sea ice initial conditions	Gelato-NEMO run constrained towards Glorys 2V4 (Mercator)	Forced LIM3-NEMO run with ENKF SIC assimilation	ORA-S5	NEMOVAR	Gelato-NEMO run constrained towards Glorys 2V4 (Mercator)
Ensemble size	30	10	25	28*	25*

** All re-forecasts are initialized on the 1st of the month, except for GloSea5 for which 7 members from the 9th, 17th and 25th of April as well as 7 from the 1st of May are grouped into a 28-member ensemble, and MF-Sys6 for which 12 members from the 20th and 25th of April are grouped with one member from the 1st of May into a 25-member ensemble.*

Table 2: Linear trend of SIE and RMSE of SIE after linear detrending (in thousands of km^2/year) and correlation over 1993-2014 for September over the Pan Arctic region in the different models studied, the MME and C3S MME. Scores are computed against NSIDC SIC data.

Model	CNRM-CM6-1	EC-Earth 3.2.2	SEAS 5	GloSea5	MF-Sys6	MME	C3S MME	Reference*
Sept. SIE linear trend	-84.1	-84.7	-44.0	-83.2	-95.9	-78.0	-74.7	-130.1
Sept. SIE RMSE	2443.6	666.0	957.1	625.4	2472.5	1208.6	902.8	947.7
Sept. SIE RMSE (detrended)	702.4	595.8	543.8	535.9	584.3	570.2	543.1	693.7
Sept. SIE Correlation	0.66	0.78	0.79	0.84	0.80	0.81	0.83	0.34
Sept. SIE Correlation (detrended)	-0.09	0.15	0.35	0.38	0.17	0.21	0.35	-0.25

** Reference scores (RMSE, correlation) are computed for the persistence of April NSIDC SIE anomalies (magenta lines in Fig. 3). Reference trend is computed with NSIDC data.*

Figures

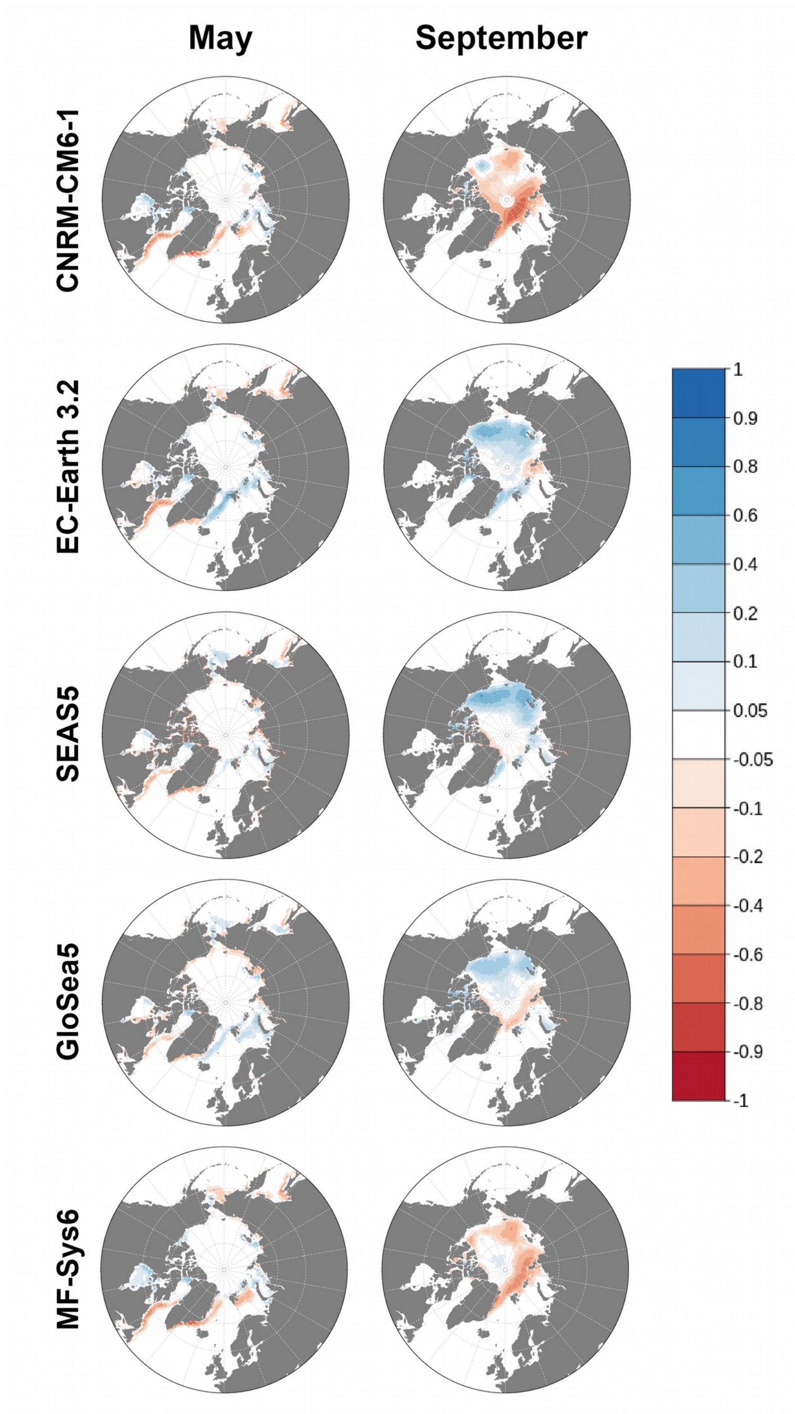

Fig. 1 Mean bias in monthly mean sea ice concentration with NSIDC in May (forecast month 1, left column) and September (forecast month 5, right column) for each of the coupled systems

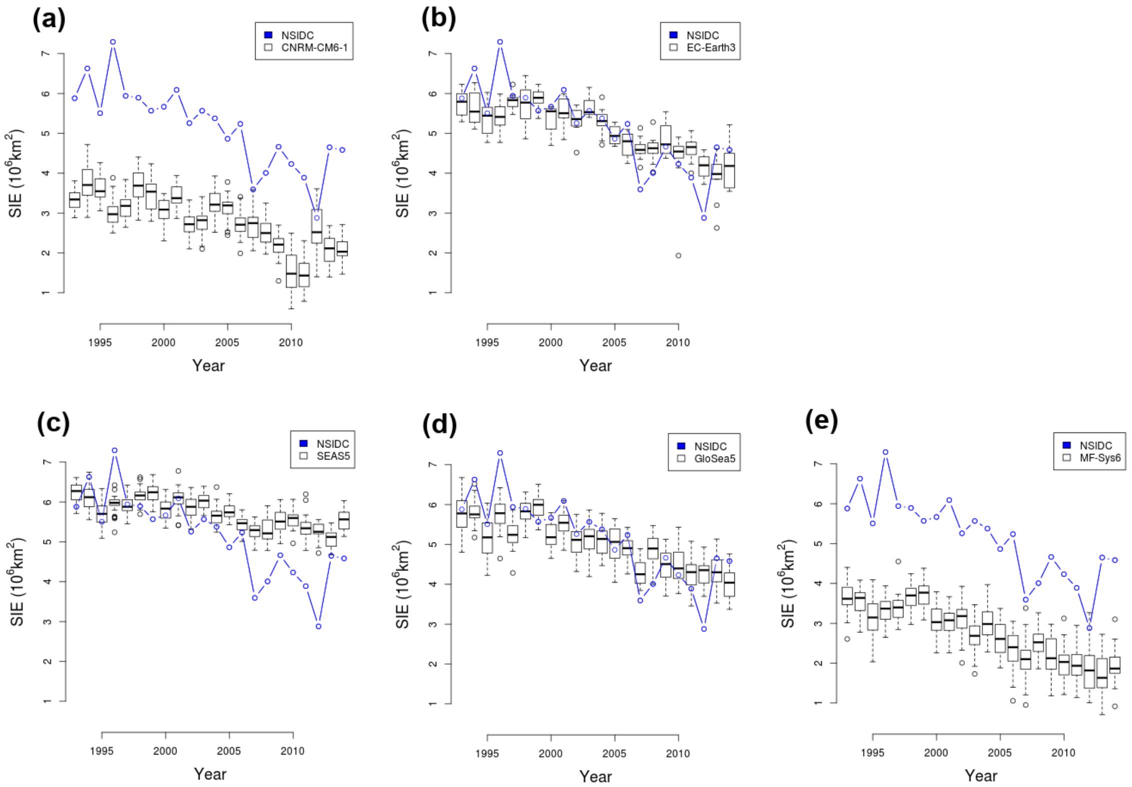

Fig. 2 Box-and-whisker plots representing September SIE values for the re-forecast ensembles in each of the models (a) CNRM-CM6-1, (b) EC-Earth 3.2, (c) SEAS5, (d) GloSea5 and (e) MF-Sys6 compared to NSIDC data (in blue). The boxes show the interquartile range of the ensembles, the thick black line is the ensemble median, and whiskers show the range of the ensemble up to 1.5 σ, and dots represent outliers beyond this range

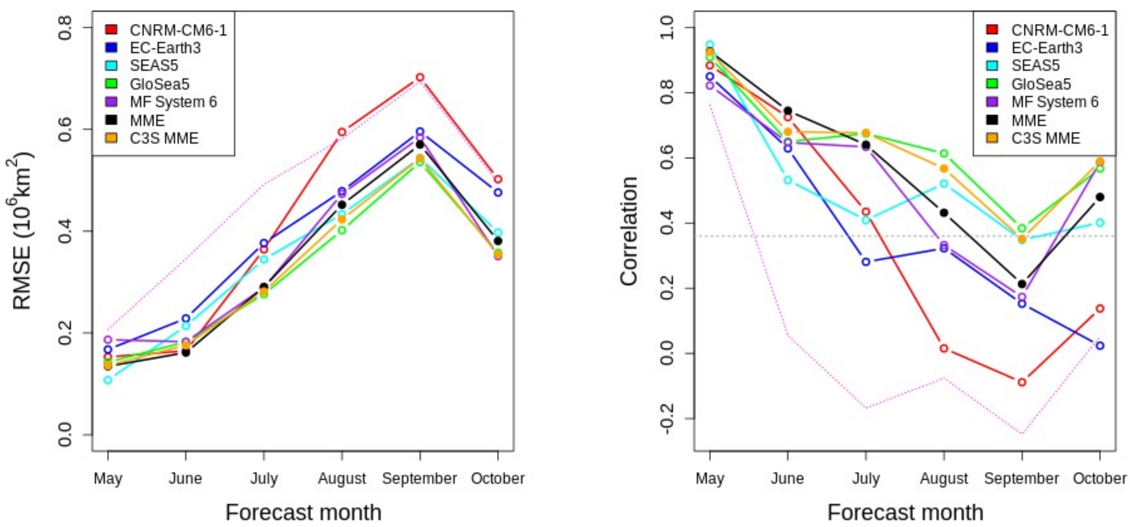

Fig. 3 Evolution according to forecast month of detrended pan-Arctic SIE RMSE (left) and correlation (right) with NSIDC reference data for the individual models (colored lines, open circles) and multi-model ensembles (filled circles). The multi-model ensemble (MME) is shown in black and the C3S MME in orange. Skill levels of the persistence of April anomalies are shown with a thin dotted magenta line. For correlation (right), a thin grey dotted line shows the 95% confidence threshold (0.36) computed using a one-sided t-test accounting for observational dependence of samples

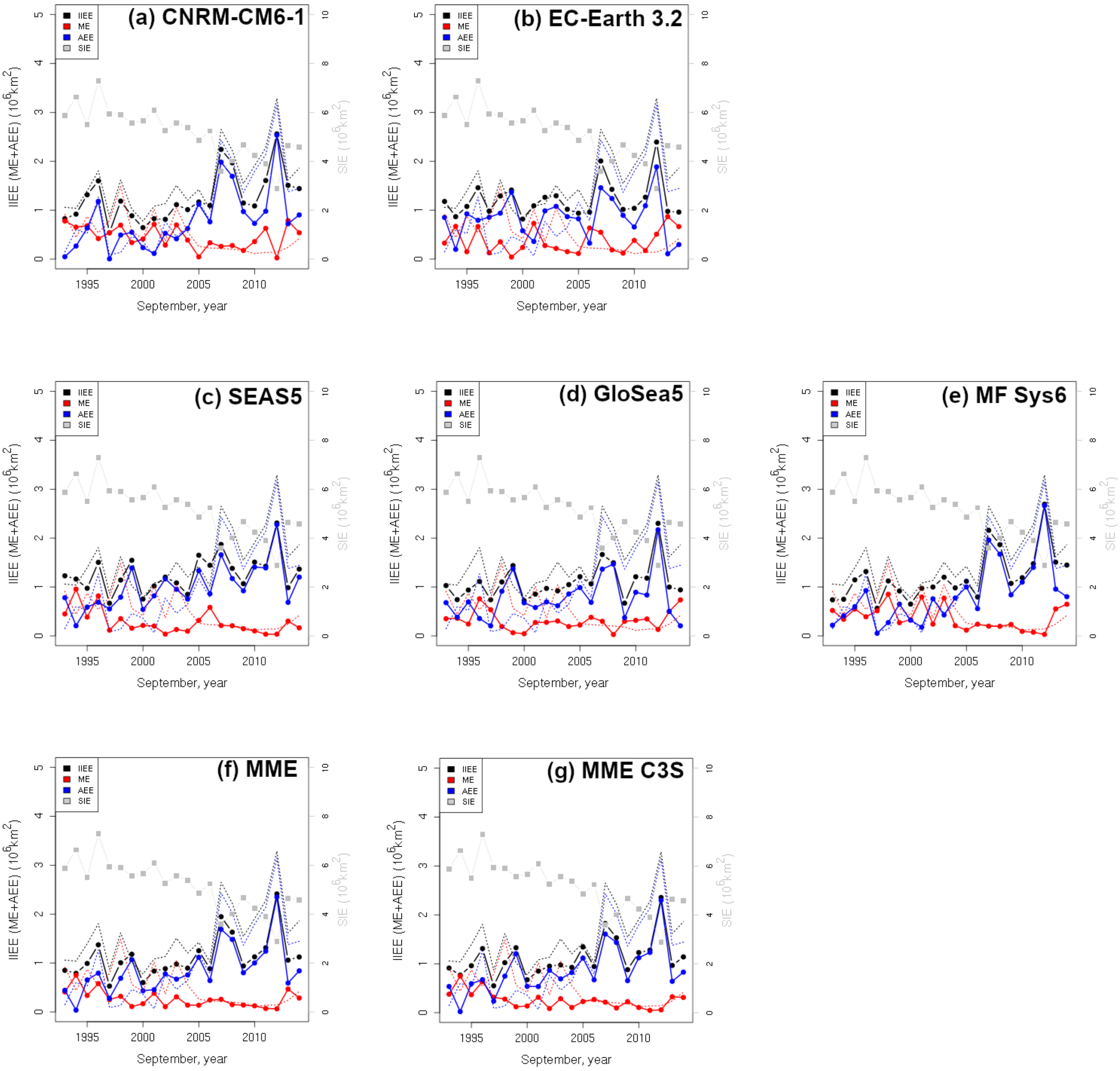

Fig. 4 IIEE (black, in millions of km2) and decomposition in ME (red) and AEE (blue) with respect to NSIDC data for September 1993 to 2014 in re-forecasts initialized in May with (a) CNRM-CM6-1, (b) EC-Earth 3.2, (c) SEAS5, (d) GloSea5 and (e) MF-Sys6. (f) Same as (a-e) but for a multi-model ensemble grouping all ensemble members of each individual system (after individual bias correction of SIC). (g) Same as (f) but for the C3S operational systems (c-e). In each graph, thin dashed lines show the corresponding IIEE, ME and AEE values for a leave-one-out climatology based on NSIDC data, and the grey line shows the reference SIE (y-axis on the right hand side)

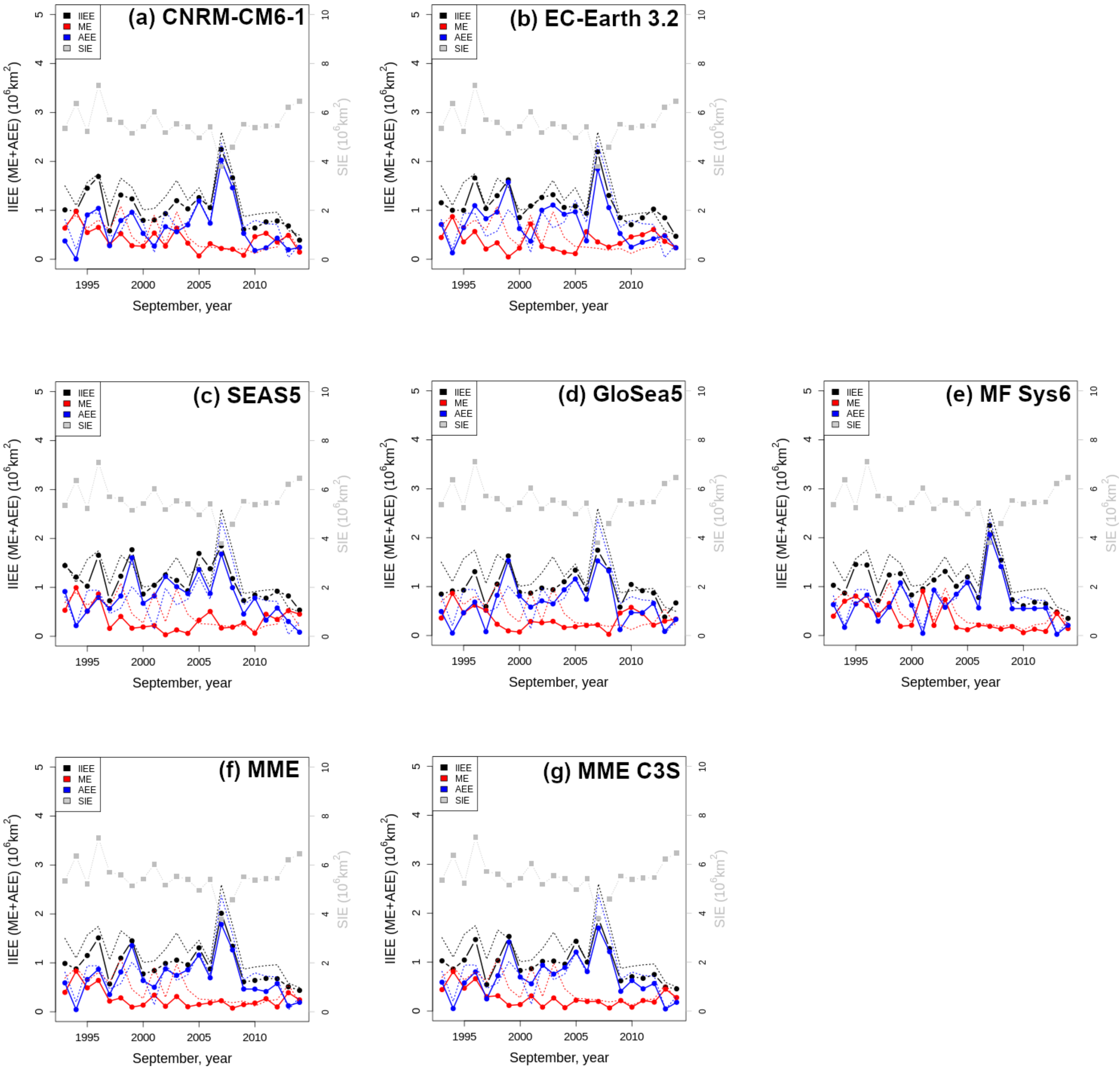

Fig. 5 Same as Fig 4 but for IIEE computed after trend-adjusting SIC data at the gridpoint level. The dashed lines (IIEE, ME and AEE values of a forecast based on climatology) and SIE shown in grey are also computed from trend-adjusted SIC NSIDC data

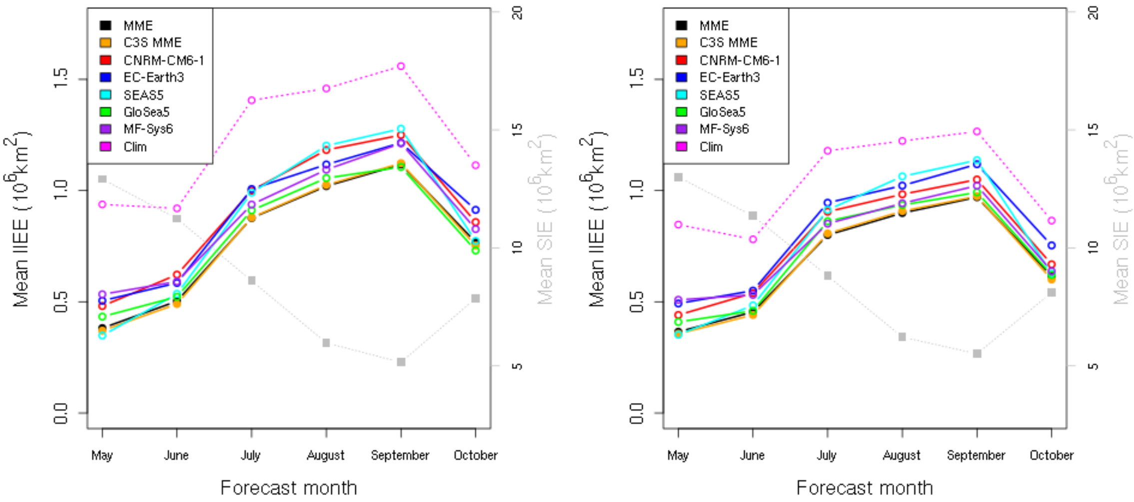

Fig. 6 Evolution according to forecast time of mean IIEE computed using bias-corrected SIC data (left) and trend-adjusted SIC data (right) over the 1993-2014 re-forecast period for each model and the MME (in black) and C3S MME (in orange). Mean IIEE of the climatology forecasts (respectively, leave-one-out and trend-adjusted climatologies using NSIDC SIC data) are also shown (Clim, dotted magenta line). The monthly mean average SIE over 1993-2014 is shown in grey dotted lines (right y-axis)

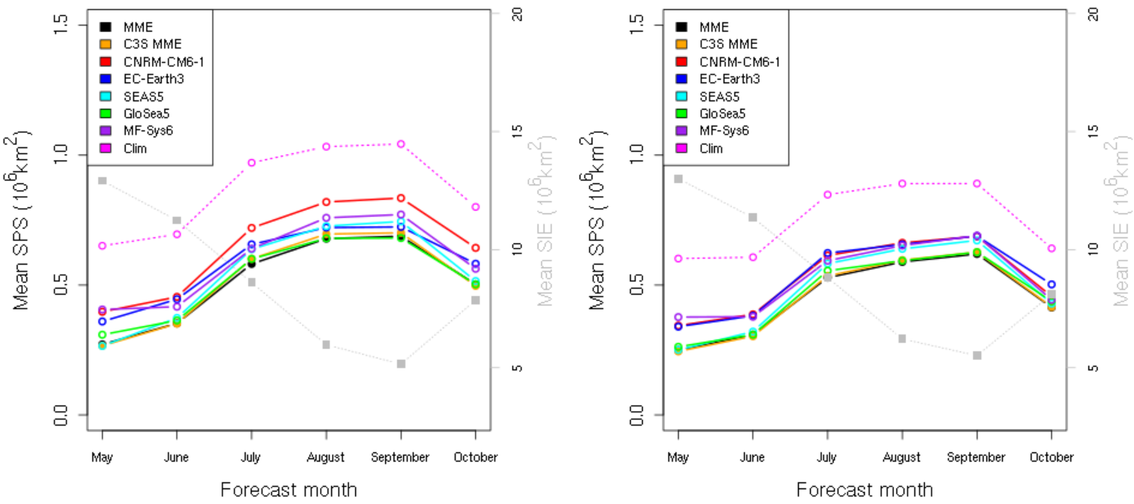

Fig. 7 Left: mean SPS over 1993-2014 according to forecast month for each system and the MME (in black) and C3S MME (in orange), and 1993-2014 monthly mean SIE computed with NSIDC SIC data (in grey dotted line, right y-axis). SPS is computed after bias-correcting the probabilities of SIC exceeding 0.15. The benchmark forecast (Clim, in magenta) is based on a leave-one-out probability forecast using the other years of the 1993-2014 period. Right: same as left-hand-side figure but after additionally trend-adjusting SIC values before computing probabilities; the Clim forecast is in this case a linear trend adjusted leave-one-out probability forecast.

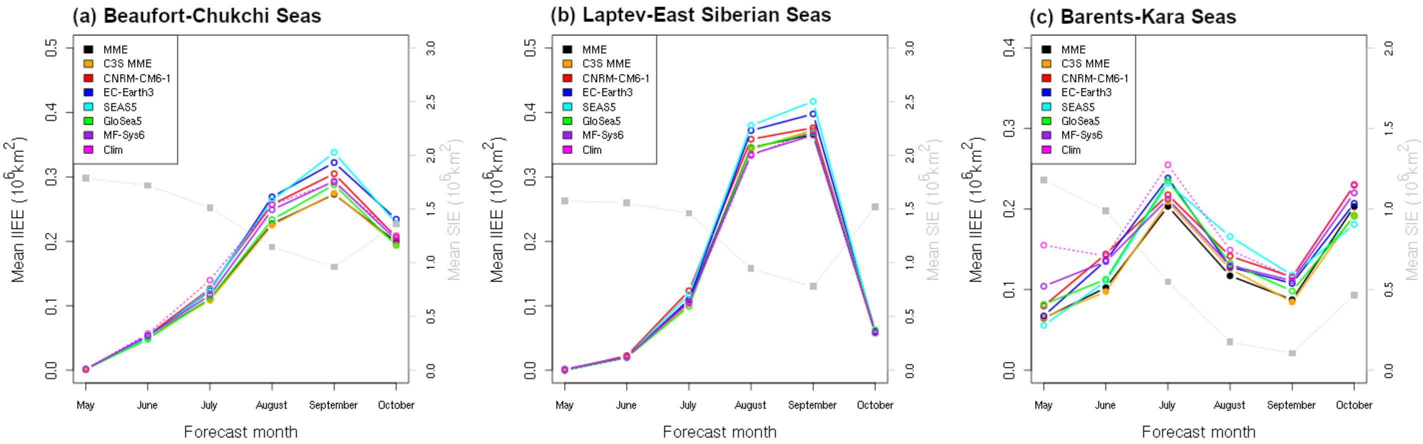

Fig. 8 IIEE computed for trend-adjusted SIC re-forecasts for each model and the 5-model and C3S MME, over the extended Beaufort-Chukchi Seas region (a), the extended Laptev-East Siberian Seas region (b) and the Barents-Kara Seas region (c). IIEE for a benchmark climatology forecast based on linear trend-adjusted SIC is also plotted (in magenta dashed lines). Total SIE over the regions are shown in grey (right y-axis values). The y-axis values differ between graphs (a-b) and (c).

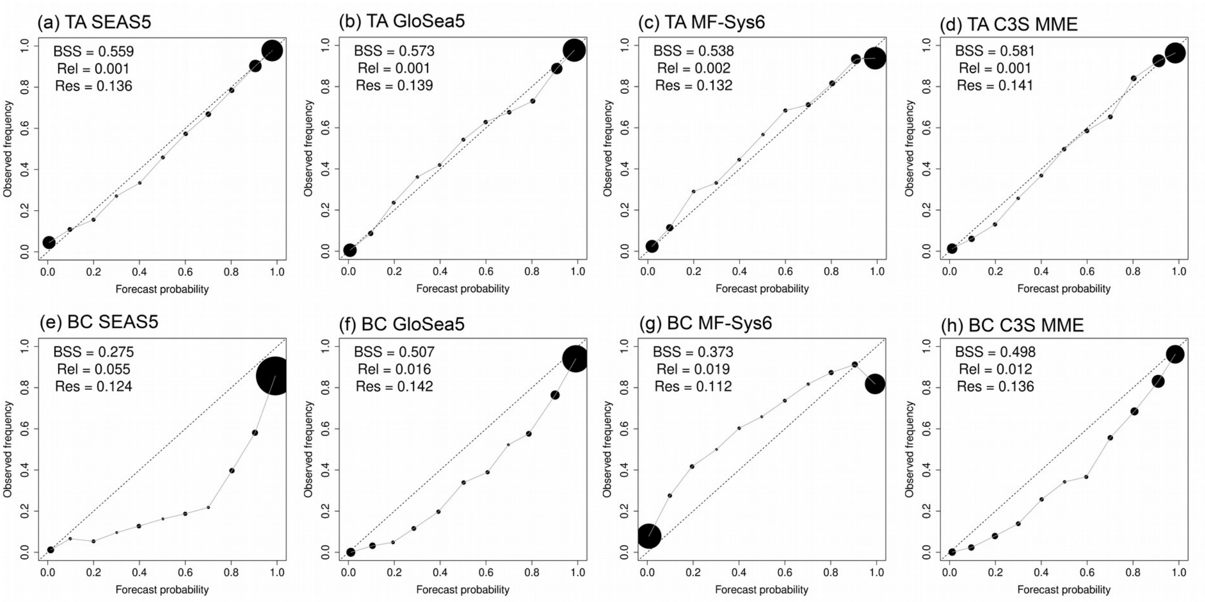

Fig. 9 Reliability diagrams (observed frequency for binned forecast probabilities) for September mean SIC exceeding the 0.15 threshold computed for grid cells of the Beaufort-Chukchi seas region, using trend-adjusted (a-d) and bias-corrected (e-h) SIC ensemble re-forecasts initialized in May 1993-2014 for the three operational systems and the C3S MME. The size of the dots is proportional to the population of each bin. Reference data is NSIDC. The Brier Skill Score as well as reliability and resolution components of the Brier Score are shown in the top left corner of each diagram

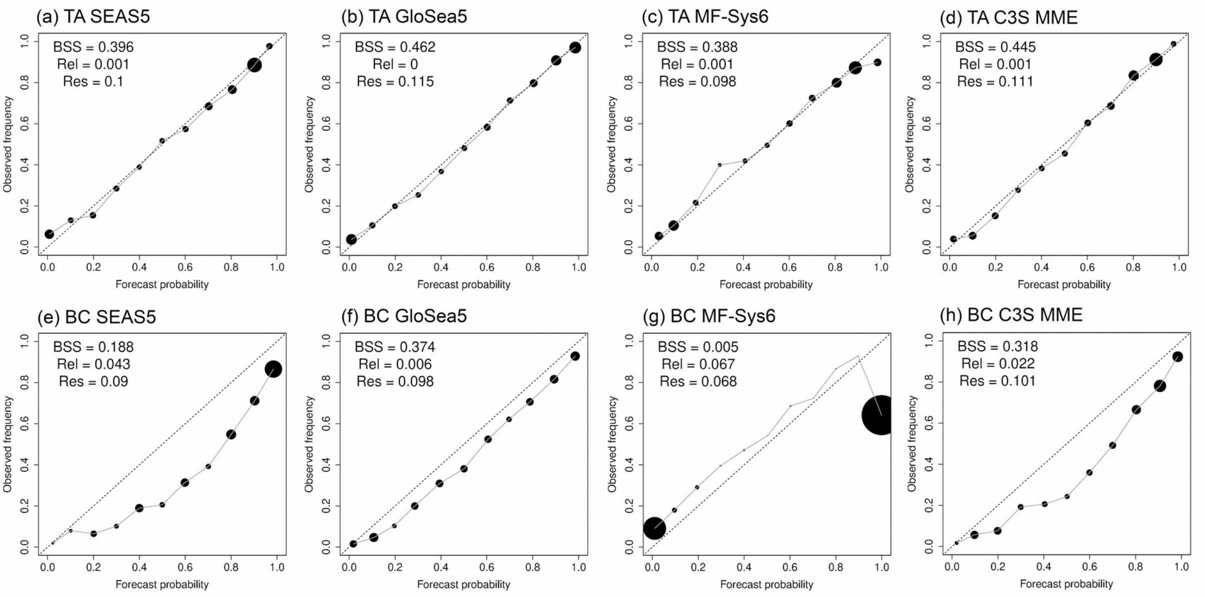

Fig. 10 Same as Fig. 9 but for the Laptev-East Siberian Seas region