Aspects of designing and evaluating seasonal-to-interannual Arctic sea-ice prediction systems

Aspects of designing and evaluating seasonal-to-interannual Arctic sea-ice prediction systems

Ed Hawkins[a]*, Steffen Tietsche[a], Jonathan J. Day[a], Nathanael Melia[a], Keith Haines[b], Sarah Keeley[c]

[a]*NCAS-Climate, Department of Meteorology, University of Reading, UK.*
[b]*Department of Meteorology, University of Reading, UK.*
[c]*European Centre for Medium-range Weather Forecasts, Reading, UK.*

*Correspondence to: Department of Meteorology, University of Reading, Reading. RG6 6BB. UK.
E-mail: e.hawkins@reading.ac.uk

Using lessons from idealised predictability experiments, we discuss some issues and perspectives on the design of operational seasonal to inter-annual Arctic sea-ice prediction systems. We first review the opportunities to use a hierarchy of different types of experiment to learn about the predictability of Arctic climate. We also examine key issues for ensemble system design, such as: measuring skill, the role of ensemble size and generation of ensemble members. When assessing the potential skill of a set of prediction experiments, using more than one metric is essential as different choices can significantly alter conclusions about the presence or lack of skill. We find that increasing both the number of hindcasts and ensemble size is important for reliably assessing the correlation and expected error in forecasts. For other metrics, such as dispersion, increasing ensemble size is most important. Probabilistic measures of skill can also provide useful information about the reliability of forecasts. In addition, various methods for generating the different ensemble members are tested. The range of techniques can produce surprisingly different ensemble spread characteristics. The lessons learnt should help inform the design of future operational prediction systems.

Key Words: Arctic; sea-ice; predictability; ensemble design

Received . . .

1. Introduction

Arctic sea-ice has shown a recent decline in extent, especially in summer, raising the possibility of increased usage of the region for shipping, resource extraction and tourism. Operational Arctic sea-ice prediction systems may help manage the risks and inform decisions about such usage of the Arctic region (Eicken 2013). However, information on what particular aspects of the sea-ice are required by users to inform their decisions is still sparse. In addition, skillful predictions of the sea-ice may also improve predictions of atmospheric variables (e.g. Scaife *et al.* 2014).

One developing method for making sea-ice predictions is using dynamical global climate models (GCMs), and the results of initial attempts at making such predictions are encouraging. For example, significant skill in retrospectively predicting (or 'hindcasting') September sea-ice extent has been demonstrated, but substantial issues remain (Sigmond *et al.* 2013; Wang *et al.* 2013; Chevallier *et al.* 2013; Merryfield *et al.* 2013; Guemas *et al.* 2014a; Msadek *et al.* 2014; Peterson *et al.* 2015).

Much of the prediction skill seen in these operational systems derives from predicting the long-term downward trend in sea-ice extent. However, users of the Arctic region would require predictions on shorter timescales, for example the coming season. For such forecasts, predicting the seasonal and year-to-year fluctuations in the ice is more important than the long-term trend, as recent observed sea-ice variations demonstrate. However, current operational GCM-based prediction systems show limited ability to make skillful predictions of these fluctuations in the September extent, especially more than 3-4 months ahead (Sigmond *et al.* 2013; Wang *et al.* 2013; Chevallier *et al.* 2013; Merryfield *et al.* 2013; Guemas *et al.* 2014a; Msadek *et al.* 2014; Peterson *et al.* 2015). Is this because the fundamental limits of predictability have already been reached, or because of model inadequacies and lack of observational data with which to initialise the predictions? Or a combination of both? It is hard to distinguish between these possibilities just using such operational hindcasts, and therefore difficult to decide what aspects to focus on to improve such predictions.

In addition, there are key questions about how to design operational Arctic sea-ice prediction systems. For example, how large an ensemble is required, and how should the different members be generated? How many hindcasts need to be performed to get a reliable estimate of the skill, and what metrics are appropriate to assess skill?

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/qj.2643

We examine some of these practical issues with examples from idealised predictions performed for the Arctic Predictability and Prediction on Seasonal-to-Interannual TimEscales (APPOSITE) project (see Tietsche *et al.* 2014, for a project description), with a focus on pan-Arctic sea-ice extent and volume. The advantage of this idealised approach is that some of the key questions outlined above can be considered without the complicating effects of uncertain observations and model inadequacies giving rise to biases. For example, Day *et al.* (2014b) examined an early-summer predictability 'barrier', whereby predictions started before June have limited skill for the subsequent September. This 'barrier' is seen in operational forecast systems (e.g. Sigmond *et al.* 2013) and is also present in similar idealised predictions, suggesting that there are indeed fundamental limits to predicting the sea-ice a few months ahead from certain times of year.

Here we offer further perspectives on the design of Arctic sea-ice prediction systems. In Section 2 we briefly consider the role of different types of experiment in informing the design of improved operational prediction systems. We consider issues of ensemble design and the generation of ensemble members in Section 3, and ways of evaluating predictive skill are discussed in Section 4. We summarise and discuss the recommendations and implications in Section 5.

2. Role of different experiment types

There are many different experiments with GCMs which can help inform about the predictability of the climate system in general, and which are applicable to the Arctic more specifically. These types of experiment can be viewed as complementary and as a hierarchy, moving towards operational predictions.

2.1. Terminology

Our ability to make accurate seasonal predictions is limited by the chaotic nature of the climate system. Even with a perfect model and near-perfect initial conditions, forecasts would diverge from the real world. The term *predictability* is used to describe the potential to make skillful predictions.

Different components of the climate system have different levels of predictability. For example, sea surface temperatures have larger predictability than land temperatures because they evolve on slower timescales.

The term predictability can also be applied to a GCM. In this case, the ability of the GCM to predict itself is measured - termed the 'perfect model' assumption. For example, a tiny perturbation to a single grid point is enough to produce a slightly different trajectory. The rate of divergence between such trajectories is a measure of predictability. But note that a GCM might show more or less predictability than the real climate system (Eade *et al.* 2014).

Predictability should not be confused with *skill* which usually refers to the ability of an imperfect model to make forecasts of the real world, though the terms are often used interchangably. The achieved skill will likely be smaller than the predictability of the real world.

2.2. Control simulations

Long 'control' simulations which have no changes in external forcing are useful to understand the properties of the GCM, such as the mean state, seasonal cycle and variability characteristics. Most control simulations are performed with 'pre-industrial' levels of greenhouse gases and other radiative forcings, which may be less informative for present-day Arctic conditions. However, there are an increasing number of present-day control simulations available, several of which were performed specifically for the APPOSITE project (Day *et al.* 2014b). In these simulations, the radiative forcings are held fixed at recent levels (e.g. 1990 or 2005), but the precise details vary between simulations (Tietsche *et al.* 2014).

Control simulations can be used to give an estimate of the predictability by using the diagnostic potential predictability metric (DPP, Boer 2000). Here we define DPP as the ratio of standard deviations of 5-year means and 1-year means (σ_5/σ_1), which is a crude estimate of potential skill. This is because it estimates the fraction of variance explained by lower frequency variability which is potentially more predictable than the high frequency variability. Diagnostic predictability measures, such as DPP, can be calculated in the same way for models and observations and provide a measure of predictability which can be used to directly compare the two.

Fig. 1 shows DPP for both sea-ice concentration and sea surface temperatures (SSTs) for the August-September-October season, in the present-day control simulations of various GCMs and for the linearly detrended HadISST observational dataset (during 1953-2010) (Rayner *et al.* 2003). In most cases, there is higher DPP near the ocean boundaries of the Arctic, i.e. Bering Strait and the Atlantic inflows, suggesting these regions are more predictable. However, the magnitude of potential predictability for both sea-ice and SSTs varies significantly across the GCMs and observations, suggesting the variability characteristics are rather different. There is also coincident potential predictability in both SSTs under the ice and in the sea-ice concentration itself, suggesting a possible role for the ocean in producing sea-ice variability. Alternatively, the sea-ice may be influencing SST variability.

In addition, lagged correlations have been used to determine Arctic predictability from a range of present-day control simulations (Blanchard-Wrigglesworth *et al.* 2011a; Chevallier and Salas-Mélia 2012; Day *et al.* 2014b). These analyses suggest that the GCMs tend to have higher predictability levels than similar estimates derived from the sea-ice observations. However, the short observational record and presence of long-term trends inhibits the comparison (Blanchard-Wrigglesworth *et al.* 2011a).

Control simulations provide a way to investigate predictor-predictand relationships where one of the variables is not well observed. For example, Chevallier and Salas-Mélia (2012) considered a range of predictors for September sea-ice extent in a GCM control simulation and found some potential skill for a few months. This type of analysis is useful to inform the design of empirical forecast methodologies which utilise the observations (e.g. Lindsay *et al.* 2008; Schroeder *et al.* 2014) and also what aspects of the climate system are potentially useful for initialisation.

2.3. Perfect model experiments

Another estimate of potential predictability is obtainable by running so-called 'perfect model' experiments. In these simulations, the GCM is used to predict itself using an ensemble of virtually identical initial conditions (with tiny perturbations only), allowing the chaotic nature of the climate system to amplify the perturbations.

These experiments are useful to determine the predictability inherent to the model as there are no complicating factors from model biases and assimilation of observations. This experimental design has been used extensively in ocean predictability studies (e.g. Griffies and Bryan 1997; Collins *et al.* 2006), and more recently applied in both control and externally forced Arctic predictability studies (Koenigk and Mikolajewicz 2009; Holland *et al.* 2011; Blanchard-Wrigglesworth *et al.* 2011b; Tietsche *et al.* 2014; Day *et al.* 2014b).

These so-called 'prognostic' estimates are normally assumed to provide an upper bound on the potential skill when using a

particular GCM to make predictions. However, Eade *et al.* (2014) suggested that certain aspects of climate variability may be more predictable in the real world than in the GCMs. If this is the case, then the 'perfect' predictability estimates could actually be lower than those found when using the same model to predict the real world.

2.4. Observation impact experiments

One set of approaches to testing the value of various observations in providing skill consists of adding or denying certain data types in forecasts or simulated forecasts. Such experiments are variously termed observing system experiments (OSEs) or observing system simulation experiments (OSSEs), depending on the exact approach adopted.

For example, operational assimilation techniques can be tested in a perfect model framework to highlight the sources of predictability. This has been demonstrated for the ocean (e.g. Dunstone and Smith 2010), and also in the Arctic (e.g. Tietsche *et al.* 2013a).

An example for OSEs is an experiment where the influence of certain data is removed from the initial conditions and predictions repeated to examine the importance of that data. This general 'data denial' technique has been applied in both idealised predictions and operational forecast systems.

For example, by removing information about the initial sea-ice thickness state in a set of idealised predictions, the importance of thickness for providing skill in perfect GCM prediction experiments has been demonstrated on seasonal timescales (Day *et al.* 2014a). This approach has highlighted that new observations and assimilation of sea-ice thickness would likely improve forecasts of the subsequent sea-ice conditions.

Alternatively, Jung *et al.* (2014) demonstrated the impact of reanalysis data from the Arctic in an operational forecast system, highlighting the need to maintain and improve observational coverage in the region.

2.5. Hindcasts and forecasts

Finally, GCMs can be used to try and predict the real climate on seasonal to decadal timescales in order to assess the predictability of the system. In this case, the skill is measured by performing retrospective forecasts of past cases (or 'hindcasts'). It is generally assumed that future predictive skill will be similar to hindcast skill. However, the interpretation of hindcast skill is complicated by the long-term trends, the changing observing network, and by the predictability itself changing over time (e.g. Holland *et al.* 2011).

Several operational Arctic hindcast and forecast systems have recently been developed, based on individual models (Sigmond *et al.* 2013; Wang *et al.* 2013; Chevallier *et al.* 2013; Germe *et al.* 2014; Msadek *et al.* 2014; Peterson *et al.* 2015) and a multi-model system (Merryfield *et al.* 2013), and they demonstrate encouraging levels of skill.

Such ensembles of hindcasts can also be used as potential predictability experiments by using each ensemble member in turn as the 'truth'. Wang *et al.* (2013) demonstrated larger potential predictability than actual predictive skill in their forecast system, suggesting that the limits of predictability have not yet been reached. However, their interpretation is slightly complicated by the forced trends, and potential 'drifts' or 'shocks' in the hindcasts when the GCM is initialised from observations.

2.6. Learning from the experimental hierarchy

The primary reason for considering such a hierarchy of experiments is to learn about the ability of each GCM to make predictions in simpler cases than in the real world. For example, perfect model predictability experiments are useful to understand the fundamental limits of predictability for the climate variable that you want to predict. This may give an upper bound on the ability *of that same GCM* to predict the real world (but see Otto *et al.* (2013) and Eade *et al.* (2014) for a discussion of some of these issues). Such techniques have been successfully applied to learn about predictability for other climate variables, such as the ocean circulation (e.g. Collins *et al.* 2006), and surface air temperature (e.g. Liu *et al.* 2012).

More specifically for the Arctic, if an operational system found that forecasts of September sea-ice extent started before May have low skill (e.g. Sigmond *et al.* 2013), it cannot be known without further experiments whether this is an inherent limit, or due to model biases or observational inadequacies. Using a set of similar perfect model experiments would enable a demonstration of whether there is an inherent predictability barrier for May forecasts. This has since been demonstrated for one GCM (Day *et al.* 2014b).

We also note that simple empirical (or statistical) methods are regularly used as useful 'benchmarks' for the GCM-based seasonal and decadal predictions of sea surface temperatures (e.g. Barnston *et al.* 1994; Ho *et al.* 2013) and have also been used for Arctic sea-ice (Lindsay *et al.* 2008; Schroeder *et al.* 2014). In addition, both Wang *et al.* (2013) and Merryfield *et al.* (2013) demonstrated that damped persistence hindcasts had some skill for predicting Arctic sea-ice extent a month or two ahead, but that the GCMs performed better. This finding has also been replicated in a perfect model experiment (Day *et al.* 2014b). Further development of these empirical techniques could prove valuable as a benchmark level of skill.

3. Issues of ensemble design

We now explore how some perfect model predictability experiments can help inform the design of operational prediction systems.

Assessments of the skill of seasonal forecast systems are performed using sets of ensemble hindcasts (e.g. Graham *et al.* 2005). Here we consider key questions of the number of hindcasts that might be necessary to reliably assess skill, as well as the number of ensemble members required and how the different members are generated.

3.1. Number of ensemble members and start years

Due to the large computational requirements of seasonal forecast systems, resource availability limits the number of ensemble members and number of start dates. Thus, it is important for forecast centres to know whether to prioritise running more start dates or more ensemble members (e.g. Buizza and Palmer 1998; Chen *et al.* 2013).

For seasonal predictions of the Arctic, little has been done to look at this trade off between the number of start dates and number of ensemble members. We consider these issues for predictions of pan-Arctic sea-ice extent (SIE) and volume (SIV) using a set of perfect model simulations with the HadGEM1.2 GCM (Johns *et al.* 2006; Shaffrey *et al.* 2009).

For ten different years, ensembles of sixteen members were initialised on each of Jan 1st, May 1st and July 1st with initial conditions from a control run, with the members differing only by spatial white noise applied to the SSTs (with magnitude $\sigma = 10^{-4}$ K). More details are given in Tietsche *et al.* (2014) and Day *et al.* (2014b). In this idealised situation we can consider questions of ensemble design without the issue of non-stationarity of observations or forecast error, albeit with a relatively small sample of start dates and ensemble members. We calculate the RMSE

between the ensemble mean and the reference control simulation, but considering different subsets of ensemble members and start years to examine the sensitivity of the RMSE to ensemble size and number of start dates respectively.

Figure 2 shows box and whisker plots of RMSE for sea-ice extent (top) and volume (bottom) from forecasts started on 1st July for the September mean (forecast month 3). When considering ensemble size (left), the quantiles are calculated from every combination of N ensemble members for all 10 start dates. For the start date panels (right), the quantiles are calculated from the RMSE values when using all 16 ensemble members and every combination of M start years. The crosses mark the RMSE using the full set of simulations ($M = 10$ start dates and $N = 16$ ensemble members).

It is clear that increasing both the number of start dates and ensemble members leads to a reduction in the uncertainty in the estimate of RMSE, with a slight indication that increasing start dates is more important than ensemble size. Forecasts for different individual years have very different error sizes, ranging from 0.17–0.51$\times 10^6$ km^2 for SIE. For different individual ensemble members the range is 0.19-0.48$\times 10^6$ km^2. However, the relative importance of the two aspects of ensemble size will vary with lead time, verification month, and climate variable.

For anomaly correlation (see Eqn. 4 later), increases in ensemble size can produce a significantly higher skill (Fig. 3), although this small effect has largely saturated with around 8 ensemble members for both SIE and SIV when considering 10 start dates. However, a larger ensemble size is likely to be more important for assessing the spread or dispersion of the prediction system (see Sect. 4.2). Note the increase in correlation for shorter lead times (compare rows in Fig. 3) highlights the levels of additional skill closer to the verification time (also see Day *et al.* (2014b)).

It should also be noted that we have only considered integrated properties of Arctic sea-ice, e.g. pan-Arctic extent. For regional forecasts where there is larger variability the ensemble sizes required will be larger. This type of analysis would be useful for operational centres to perform with their existing systems to inform future developments (e.g. Scaife *et al.* 2014).

3.2. Ensemble generation

Another key choice in ensemble design is how to generate the different ensemble members. In seasonal forecasting, it is often found that the ensemble spread is not large enough to encompass the subsequent observations (e.g. Weisheimer *et al.* (2009)), and various methods are used to generate additional ensemble spread to ensure the forecasts are 'reliable' (see Section 4.2 later), for example, singular vectors (e.g. Buizza 1997) or stochastic perturbations (e.g. Weisheimer *et al.* 2011; Juricke *et al.* 2013).

Here, we compare the impact of the simplest methods for generating ensemble initial conditions: (i) state-lagged, (ii) atmosphere-lagged, and (iii) SST-noise initialisation. Methods (ii) & (iii) are widely used in predictability studies. Previously, the differences which may arise from choosing one methodology over another have not been discussed for sea-ice predictions. Here, we show that for seasonal predictions of Arctic sea-ice, the methods do show some striking differences.

The state-lagged perturbation method (SL) takes the state of the GCM from days adjacent to the actual start date as initial conditions. This constitutes a sizable perturbation to the state of the atmosphere, but only a small perturbation to the state of the ocean and the sea-ice cover (see figure 7 in Tietsche *et al.* (2013b)). The atmosphere-lagged perturbation method (AL) applies the same lagged perturbation but only to the atmosphere, and the other components of the climate system remain unperturbed. Finally, the SST-noise perturbation (NSST) simply adds a tiny amount of spatially uncorrelated noise to the global sea-surface temperatures. This last method is essentially equivalent to assuming perfect knowledge of the initial conditions, as the magnitude of the noise ($\sigma = 10^{-4}$ K in our example) is smaller than any globally achievable measurement uncertainty. This methodology should produce the smallest ensemble spread, at least for short lead times.

Fig. 4 shows pan-Arctic SIV for a case study using three different ensembles started on 1st July of the same year with the MPI-ESM-LR GCM, with each of the perturbation methods described above applied. Note that the initial conditions for each ensemble are the same except for the added perturbation. It is evident that the state-lagged perturbation (Fig. 4a) creates a larger ensemble spread than the other two methods, even in the first lead month, because of the strong seasonal rate of change in Arctic sea-ice around the start date (1st July). The difference in SIV expected from the climatological seasonal cycle is about 200 km^3/day (or 0.2×10^{12}m^3/day) (Tietsche *et al.* 2014), which fully explains the spread of the ensemble in the first forecast month.

To add confidence and detail to the findings from the test ensemble shown above, we perform three more ensemble predictions in the same fashion, but starting in different years. This samples the variability in ensemble spread across different climate states, so that we can estimate robustness of the spread differences. Fig. 5 shows how ensemble spread develops over lead time for pan-Arctic SIE and SIV on average and how it varies between different years.

As concluded above, sea-ice volume ensemble spread is much higher for SL than for either AL or NSST. The same holds for sea-ice extent, albeit only for the first five months. We note that the differences between SL and AL may be smaller if the experiments were repeated at a time of year when the rate of change of sea-ice was smaller.

Additionally, we now see that there are in fact also differences between AL and NSST. As expected, AL spread is at least as large or larger than NSST spread at all times. For the first lead month, this difference is small but consistent, which illustrates the immediate impact of weather patterns on sea-ice extent and volume. For SIE, this difference between AL and NSST grows over the whole melt season (months 2 and 3), but then becomes marginal.

Interestingly, SIV spread after the first lead month shows the opposite behaviour: AL spread is not significantly higher for months 2 and 3, but during the freeze-up (months 4 to 6), AL spread grows much faster than NSST spread. This is a potentially important result. Although we cannot exclude the possibility that it is due to only sampling four different start years and a small ensemble size, we speculate that it arises because the differences in the atmospheric state present in AL but not in NSST have an impact on the seasonal ocean dynamics south of the summer sea-ice edge. When the ice edge advances to those regions during the freeze-up, the differences in the ocean state then translate into differences in sea-ice volume.

As ensemble spread is crucial to ensuring reliable forecasts (also see Section 4.2 later), operational prediction systems already consider the generation of ensemble members as a key component of their ensemble design. These results suggest that initial differences in the atmosphere, ocean and sea-ice state all contribute to increasing ensemble spread. In many existing forecast systems the sea-ice initial conditions are currently unperturbed, but prediction systems with multiple sea-ice initial conditions are now being developed (e.g. Guemas *et al.* 2014b).

4. Evaluating prediction ensembles

Once a set of hindcasts has been produced, there are further details to be decided about how to assess the skill. These issues include whether to use the ensemble mean and a deterministic (i.e. single prediction) measure of skill, or whether to examine the skill probabilistically (i.e. considering the whole predicted distribution). Using more 'perfect-model' examples, we illustrate some additional subtleties to be considered in any skill assessment.

4.1. Deterministic skill metrics

First, we consider deterministic skill metrics, which are based on analysing the skill of the ensemble mean. Even in a 'perfect-model' framework, such deterministic skill metrics can give ambiguous results because their value depends on how reference parameters like the climatological mean and standard deviation are defined. Here we illustrate this by directly comparing several choices of commonly used metrics and reference parameters.

4.1.1. Choice of reference period

We consider two choices of how to define the climatological mean μ_j and the climatological standard deviation σ_j:

1. The 'maximum knowledge'-approach (MK): The best linear fit to the whole control run serves as the time-dependent mean state; the variability is calculated from the time series after subtracting the linear fit.
2. The 'operational' approach (OP): In practice, at the time when an ensemble prediction is started, only observations from the last 30 years or so are available. Additionally, in the case of Arctic sea-ice, these observations have strong trends. To mimic this situation we base the climatology on the time series during the 30 years preceding the prediction start date: the climatological mean is the time series mean, and the climatological standard deviation is the standard deviation after linear detrending.

Only the operational approach is feasible in actual predictions of the observed climate, because it is impossible to utilize future observations to estimate μ_j and σ_j. However, comparing these two methods demonstrates how the predictability estimated using OP might mislead about the true predictability.

There are also two more caveats for predictability estimates obtained with MK: (i) it factors in the role of multi-decadal variability which may not be predictable, and so MK predictability estimates are likely to be biased high, and (ii) in the presence of strong secular trends, estimates of μ_j and σ_j obtained with MK might be contaminated by remote climate states that are not relevant for the time the ensemble prediction is started (Goosse *et al.* 2009).

4.1.2. Choice of metric

Several different predictability and skill metrics are widely used in seasonal to decadal prediction literature, including: potential prognostic predictability (Pohlmann *et al.* 2004), normalised root mean square error (Collins 2002; Collins *et al.* 2006), anomaly correlation coefficient (ACC, e.g. Goddard *et al.* 2013) and mean square error skill score (MSESS, e.g. Goddard *et al.* 2013).

While the first two have mainly been used in idealised studies of potential predictability, the latter two are standard measures for operational seasonal to decadal forecasts. Note that, in our idealised setup, ensemble forecasts have neither conditional nor unconditional biases, and hence there is a simple algebraic relationship between ACC and MSESS (Murphy 1988; Goddard *et al.* 2013). As a consequence, MSESS shows exactly the same information as ACC, and we do not discuss it separately.

For the perfect model approach employed here, there are in principle two different ways of verifying ensemble forecasts: (i) verification against the control run, and (ii) verification against any one ensemble member. The advantage of method (ii) is that the data can be used more efficiently by verifying against every ensemble member in turn, which increases the effective sample size and gives more robust estimates of the metrics (Collins 2002). All metrics defined below use method (ii), except for ACC_I (see below), which uses method (i).

We define the predictability metrics using the following notation: let x_{ij} be the value of sea-ice extent or volume for the i-th member of the j-th ensemble prediction, and μ_j and σ_j the climatological mean and standard deviation calculated from the control run at the time of the j-th ensemble prediction.

The *potential prognostic predictability* (PPP) compares average ensemble spread with the reference variability σ of the control run:

$$\mathrm{PPP} = 1 - \frac{\left\langle (x_{ij} - \bar{x}_j)^2 \right\rangle_{i,j}}{\left\langle \sigma_j^2 \right\rangle_j}, \tag{1}$$

where $\bar{x}_j = \left\langle x_{ij} \right\rangle_i$ is the ensemble mean of the j-th prediction ensemble, and $\left\langle \cdot \right\rangle_i$ denotes the expectation value, to be calculated by summing over the specified index with appropriate normalisation.

The *normalized RMSE* (NRMSE) compares forecast RMSE to climatological variability:

$$\mathrm{NRMSE} = 1 - \frac{\sqrt{\left\langle (x_{ij} - x_{kj})^2 \right\rangle_{i,j,k\neq i}}}{\sqrt{2\left\langle \sigma_j^2 \right\rangle_j}}, \tag{2}$$

where the denominator is the climatological RMSE between two independent realisations.

The *intra-ensemble anomaly correlation coefficient* measures the intra-ensemble correlation of predicted anomalies:

$$\mathrm{ACC}_I = \frac{\left\langle (x_{ij} - \mu_j)(x_{kj} - \mu_j) \right\rangle_{i,j,k\neq i}}{\left\langle (x_{ij} - \mu_j)^2 \right\rangle_{i,j}}, \tag{3}$$

where μ_j is the climatological mean at the time of the j-th ensemble prediction.

The *ensemble mean anomaly correlation coefficient* measures the correlation of the ensemble-mean predicted anomaly with the anomaly in the control run (pseudo-observations):

$$\mathrm{ACC}_M = \frac{\left\langle (\bar{x}_j - \mu_j)(x_r^* - \mu_j) \right\rangle_j}{\sqrt{\left\langle (\bar{x}_j - \mu_j)^2 \right\rangle_j \left\langle (x_r^* - \mu_j)^2 \right\rangle_j}}, \tag{4}$$

where x_r^* is the value of the control run at verification time (pseudo-observation).

NRMSE and ACC_M are regularly used for assessing operational predictions.

4.1.3. Sensitivity of estimated skill

Each GCM is likely to produce skill estimates which are sensitive to choice of metric and reference period. We illustrate these sensitivities using the HadGEM1.2 GCM (as in Section 3.1) because, for this model, the differences between the predictability estimates when choosing different reference periods and metrics are large.

For the MK approach (upper row in Fig. 6), there is a good agreement between the metrics, but for both SIE and SIV, the

NRMSE tends to lie below the other metrics. This is a trivial consequence of taking the square root in Eq. (2). Without it, PPP and NRMSE would be virtually identical (not shown). The other difference is that ACC_M tends to lie above the other estimates for SIE in the second lead year, but this might be due to insufficient sampling as for ACC_M only 10 data points are used to construct the correlation.

The picture changes quite dramatically when comparing the predictability estimates for the OP approach. While the ACC metrics are comparable with the MK approach, PPP and NRMSE are much lower. In fact, they quickly reach the limit of zero skill after a year or so, whereas for the MK approach they are significantly above zero for at least three years. This is because the estimated σ_j with the MK approach are generally much higher than estimates obtained with the OP approach (20 to 40% higher depending on season, not shown). As mentioned above, this difference corresponds to the choice whether to include decadal or even multi-decadal variability. However, OP might produce a pessimistic estimate of the true predictability.

These sensitivities highlight the need to consider multiple skill metrics in any assessment of predictability. For example, relying on timescales of predictability estimated from a single metric may not be robust. This comparison also quantifies how the apparent skill of an operational system might give a distorted impression of the true level of skill.

4.2. Probabilistic skill metrics

There are many different probabilistic skill metrics (e.g. Brier 1950; Candille and Talagrand 2005; Jolliffe and Stephenson 2012). Here, we discuss ensemble dispersion and the Brier Score as examples.

4.2.1. Dispersion

Reliability is an important and desirable property of any ensemble prediction system, i.e. that the forecast probabilities for a particular event are correct (e.g. Buizza 1997; Kumar *et al.* 2001; Weisheimer *et al.* 2011). A necessary requirement for reliability is that the ensemble has the correct dispersion, i.e. the ensemble spread matches the expected RMSE (see e.g. Jolliffe and Stephenson 2012).

Fig. 7a shows an example, using the GFDL CM3 GCM, of 'perfect-model' ensemble predictions of sea-ice extent and volume that have high skill, but nevertheless appear unreliable. On average, the ensemble predictions are under-dispersive: the RMSE of ensemble means is larger than the average ensemble spread for both SIE and SIV. The reason for this is insufficient sampling of start dates and ensemble members, as illustrated by dividing the complete set of 8 ensemble predictions into a set of 3 which are highly under-dispersive (Fig. 7b) and a set of 5 which have almost exactly the right amount of dispersion to be reliable (Fig. 7c).

This example illustrates that sampling error is a problematic issue for sample sizes widely used in seasonal-to-decadal prediction studies, and also highlights where a larger ensemble may help reduce the chance of a 'surprise' by more completely sampling the forecast distribution.

In addition, this example also suggests that the predictability of sea-ice may be state dependent, i.e. some situations may be more predictable than others (also see Fig. 2).

4.2.2. Brier Score

The Brier Score (BS, Brier (1950)) is often used to define probabilistic skill:

$$\mathrm{BS} = \frac{1}{N_t} \sum_t \left[f_t - o_t \right]^2 , \qquad (5)$$

where f_t is the forecast probability of a certain event, and o_t is a binary outcome (0 or 1) depending on whether the predicted event occurred or not. There are N_t forecasts over a time t.

The 'event' selected for our example is whether the sea ice state is above climatology. In this case, a trivial prediction of 50% chance for above climatology would produce BS=0.25. BS values smaller than this are therefore providing more information than a trivial climatological forecast.

Fig. 8 (top row, black lines) shows the BS for both pan-Arctic sea-ice extent and volume in the 'perfect' predictability experiments with MPI-ESM-LR. For sea-ice extent, the BS grows rapidly over the first forecast year, but is still providing additional probabilistic information in year 3 as the BS remains below 0.25. For sea-ice volume, the growth of BS is slower. For a regional averge of the eastern Arctic (bottom row), the BS is more noisy and the growth of BS is faster, as might be expected for a smaller region.

It is also possible to separate the BS into three components: reliability, resolution and uncertainty (Murphy 1973):

$$\mathrm{BS} = \mathrm{REL} - \mathrm{RES} + \mathrm{UNC} \qquad (6)$$

The reliability (REL) measures whether the forecast probabilities match the observed frequencies and should be close to zero for good forecasts. The resolution (RES) is a measure of how different the issued probabilities are from the relative frequencies, and is zero for a climatological forecast. The inherent uncertainty term (UNC) measures the frequency of the event being tested.

For this perfect model assessment, the uncertainty term would be 0.25 if the sample of forecasts was representative of the climatological frequency, and the reliability term should be zero as there are no model biases. The coloured lines in Fig. 8 show that the increase in BS is due to a reduction in resolution, as expected, i.e. the forecasts are becoming closer to a climatological forecast over time.

Considering the numerous ambiguities of real-world sea-ice and Arctic forecasts that arise from strong model biases and a rapidly changing mean climate (Sigmond *et al.* 2013; Msadek *et al.* 2014), we advocate the use of the BS score additionally to the widely used anomaly correlation. The additional benefit would be the ability to diagnose whether a low score is due to problems with resolution or a lack of reliability.

4.3. Recommendations

We have demonstrated here that even the evaluation of 'perfect' prediction ensembles can give ambiguous results that depend on the choice of the skill metric and the definition of the reference climatology. Therefore, we suggest to use a range of skill metrics to characterise the prediction ensembles, rather than relying on a single one.

Scrutinising ensembles from a probabilistic point of view might reveal problems that deterministic skill metrics do not pick up, such as a lack of reliability due to insufficient sampling. Additionally, we recommend employing probabilistic skill scores like the BS that allow a deeper understanding of a lack of skill by means of decomposition into uncertainty, reliability and resolution contributions.

5. Summary and conclusions

The retreat of Arctic sea-ice and growth of industry in the region has highlighted the need for improved predictive capability in the region (Eicken 2013). We have provided some perspectives on the design of Arctic sea-ice prediction systems, using idealised predictability experiments. To summarise:

1. The sources of sea-ice predictability can be studied with a hierarchy of experiments with forecast models, including: control simulations, perfect model predictability simulations, observing system experiments and hindcasts.
2. Care is needed when assessing predictability and hindcast skill. Different metrics and choices in the analysis can significantly alter any conclusions.
3. Ensemble design is a crucial aspect of reliably assessing any prediction system due to complications from limited ensemble size, short hindcast periods, the number of hindcasts available and ensemble generation.

We first note that much of this analysis has been on pan-Arctic integrated quantities. For operational predictions the regional distribution of ice is important, and integrated quantities can mask compensating errors in different regions (e.g. Tietsche *et al.* 2014). Fig. 8 highlighted that predictability was smaller for a regional average.

We also consider that it is imperative to issue real-time forecasts (e.g. Smith *et al.* 2013; Stroeve *et al.* 2014) to allow them to be tested completely out-of-sample. Appropriate empirical 'benchmarks' may also aid this assessment of predictive skill. For example, Schroeder *et al.* (2014) demonstrated that May melt pond fraction has significant skill for predicting September extent and Stroeve *et al.* (2014) highlighted that GCM forecasts were no better than empirical or heuristical predictions.

In addition, 'case studies' of predictability have been very useful in learning about ocean predictability (Robson *et al.* 2012) and could be explored further for sea-ice predictions; Guemas *et al.* (2013) is one such example.

Further work on considering how to correct sea-ice forecasts for model biases may also be required, especially when considering regional predictions due to the positive definite and non-Gaussian nature of sea-ice concentration and thickness.

Overall, it is clear that we have yet to reach the full potential for forecast skill of Arctic sea-ice. Additional observations, better assimilation techniques and improved models are all likely to increase predictive capabilities (also see Guemas *et al.* 2014a).

Acknowledgement

We thank Rym Msadek and Bill Hurlin for helping provide the GFDL-CM3 data. These analyses were performed as part of the UK NERC-funded APPOSITE project (grant NE/I029447/1).

References

Barnston AG, van den Dool HM, Rodenhuis DR, Ropelewski CR, Kousky VE, O'Lenic EA, Livezey RE, Zebiak SE, Cane MA, Barnett TP, Graham NE, Ji M, Leetmaa A. 1994. Long-lead seasonal forecasts? Where do we stand? *Bull. Amer. Meteor. Soc.* **75**: 2097–2114.

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

Blanchard-Wrigglesworth E, Bitz CM, Holland MM. 2011b. Influence of initial conditions and climate forcing on predicting Arctic sea ice. *Geophys. Res. Lett.* **38**: L18503, doi:10.1029/2011GL048807.

Boer GJ. 2000. A study of atmosphere-ocean predictability on long time scales. *Climate Dynamics* **16**: 469–477, doi:10.1007/s003820050340.

Brier GW. 1950. Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.* **78**: 1–3, doi:10.1175/1520-0493(1950)078¡0001:VOFEIT¿2.0.CO;2.

Buizza R. 1997. Potential forecast skill of ensemble prediction and spread and skill distributions of the ECMWF ensemble prediction system. *Mon. Wea. Rev.* **125**: 99–119.

Buizza R, Palmer TN, 1998. Impact of Ensemble Size on Ensemble Prediction. *Mon. Wea. Rev.* **126**: 2503-2518.

Candille G, Talagrand O. 2005. Evaluation of probabilistic prediction systems for a scalar variable. *Quarterly Journal of the Royal Meteorological Society* **131**(609): 2131–2150, doi:10.1256/qj.04.71.

Chen M, Wang W and Kumar A. 2013. Lagged Ensembles, Forecast Configuration, and Seasonal Predictions. *Mon. Wea. Rev.* **141**: 3477-3497, doi:10.1175/MWR-D-12-00184.1.

Chevallier M, Salas-Mélia D. 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, doi:10.1175/JCLI-D-11-00209.1.

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

Collins M. 2002. Climate predictability on interannual to decadal time scales: the initial value problem. *Clim. Dyn.* **19**(8): 671–692, doi:10.1007/s00382-002-0254-8.

Collins M, Botzet M, Carril AF, Drange H, Jouzeau A, Latif M, Masina S, Otteraa OH, Pohlmann H, Sorteberg A, Sutton R, Terray L. 2006. Interannual to decadal climate predictability in the North Atlantic: a multimodel-ensemble study. *J. Climate* **19**: 1195–1202, doi:10.1175/JCLI3654.1.

Day JJ, Hawkins E, Tietsche S. 2014a. Will arctic sea ice thickness initialization improve seasonal-to-interannual forecast skill? *Geophys. Res. Lett.* **41**: 7566, doi:10.1002/2014GL061694.

Day JJ, Tietsche S, Hawkins E. 2014b. Pan-Arctic and regional sea ice predictability: initialisation month dependence. *J. Climate* **27**: 4371, doi:10.1175/JCLI-D-13-00614.1.

Dunstone NJ, Smith DM. 2010. Impact of atmosphere and sub-surface ocean data on decadal climate prediction. *Geophys. Res. Lett.* **37**: L02 709, doi:10.1029/2009GL041609.

Eade R, Smith D, Scaife A, Wallace E, Dunstone N, Hermanson L, Robinson N. 2014. Do seasonal-to-decadal climate predictions underestimate the predictability of the real world? *Geophysical Research Letters* **41**: 5620–5628, doi:10.1002/2014GL061146.

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

Germe A, Chevallier M, Salas y Mlia D, Sanchez-Gomez E, Cassou C. 2014. Interannual predictability of arctic sea ice in a global climate model: regional contrasts and temporal evolution. *Climate Dynamics* : 1–20doi:10.1007/s00382-014-2071-2.

Goddard L, Kumar A, Solomon A, Smith D, Boer G, Gonzalez P, Kharin V, Merryfield W, Deser C, Mason S, Kirtman B, Msadek R, Sutton R, Hawkins E, Fricker T, Hegerl G, Ferro C, Stephenson D, Meehl G, Stockdale T, Burgman R, Greene A, Kushnir Y, Newman M, Carton J, Fukumori I, Delworth T. 2013. A verification framework for interannual-to-decadal predictions experiments. *Climate Dynamics* **40**: 245–272, doi:10.1007/s00382-012-1481-2.

Goosse H, Arzel O, Bitz CM, de Montety A, Vancoppenolle M. 2009. Increased variability of the Arctic summer ice extent in a warmer climate. *Geophysical Research Letters* **36**, doi:10.1029/2009GL040546.

Graham RJ, Gordon M, McLean PJ, Ineson S, Huddleston MR, Davey MK, Brookshaw A, Barnes RTH. 2005. A performance comparison of coupled and uncoupled versions of the Met Office seasonal prediction general circulation model. *Tellus A* **57**(A): 320–339.

Griffies SM, Bryan K. 1997. A predictability study of simulated North Atlantic multidecadal variability. *Climate Dyn.* **13**: 459–487.

Guemas V, Blanchard-Wrigglesworth E, Chevallier M, Day JJ, Deque M, Doblas-Reyes FJ, Fuckar NS, Germe A, Hawkins E, Keeley S, Koenigk T, Salas y Melia D, Tietsche S. 2014a. A review on arctic sea-ice predictability and prediction on seasonal to decadal time-scales. *Quarterly Journal of the Royal Meteorological Society* **in press**, doi:10.1002/qj.2401.

Guemas V, Doblas-Reyes F, Germe A, Chevallier M, Melia D. 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']. *BAMS* **94**: S20–S22.

Guemas V, Doblas-Reyes F, Mogensen K, Keeley S, Tang Y. 2014b. Ensemble of sea ice initial conditions for interannual climate predictions. *Climate Dynamics* : 1–17doi:10.1007/s00382-014-2095-7.

Ho CK, Hawkins E, Shaffrey L, Underwood FM. 2013. Statistical decadal predictions for sea surface temperatures: a benchmark for dynamical GCM predictions. *Climate Dynamics* **41**: 917–935, doi:10.1007/s00382-012-1531-9.

Holland MM, Bailey DA, Vavrus S. 2011. Inherent sea ice predictability in the rapidly changing Arctic environment of the Community Climate System Model, version 3. *Climate Dynamics* **36**: 1239–1253, doi:10.1007/s00382-010-0792-4.

Johns TC, Durman CF, Banks HT, Roberts MJ, McLaren AJ, Ridley JK, Senior CA, Williams KD, Jones A, Rickard GJ, Cusack S, Ingram WJ, Crucifix M, Sexton DMH, Joshi MM, Dong BW, Spencer H, Hill RSR, Gregory

JM, Keen AB, Pardaens AK, Lowe JA, Bodas-Salcedo A, Stark S, Searl Y. 2006. The new Hadley Centre Climate Model (HadGEM1): Evaluation of coupled simulations. *J. Climate* **19**: 1327–1353, doi:10.1175/JCLI3712.1.

Jolliffe IT, Stephenson DB (eds). 2012. *Forecast Verification: A Practicioner's Guide in Atmospheric Science*. Wiley-Blackwell.

Jung T, Kasper MA, Semmler T, Serrar S. 2014. Arctic influence on subseasonal midlatitude prediction. *Geophysical Research Letters* **41**(10): 3676–3680, doi:10.1002/2014GL059961.

Juricke S, Lemke P, Timmermann R, Rackow T. 2013. Effects of stochastic ice strength perturbation on arctic finite element sea ice modeling. *J. Climate* **26**: 3785–3802, doi:10.1175/JCLI-D-12-00388.1.

Koenigk T, Mikolajewicz U. 2009. Seasonal to interannual climate predictability in mid and high northern latitudes in a global coupled model. *Climate Dynamics* **32**: 783–798, doi:10.1007/s00382-008-0419-1.

Kumar A, Hoerling MP, Barnston AG. 2001. Seasonal Predictions, Probabilistic Verifications, and Ensemble Size. doi:10.1175/1520-0442(2001)014¡1671:SPPVAE¿2.0.CO;2.

Lindsay RW, Zhang J, Schweiger AJ, Steele MA. 2008. Seasonal predictions of ice extent in the arctic ocean. *JGR: Oceans* **113**, doi: 10.1029/2007JC004259.

Liu C, Haines K, Iwi A, Smith D. 2012. Comparing the UK Met Office Climate Prediction System (DePreSys) with idealized predictability in the HadCM3 model. *QJRMS* **138**: 81–90, doi:10.1002/qj.904.

Merryfield WJ, Lee WS, Wang W, Chen M, Kumar A. 2013. Multi-system seasonal predictions of Arctic sea ice. *Geophysical Research Letters* **40**: 1551–1556, doi:10.1002/grl.50317.

Msadek R, Vecchi GA, Winton M, Gudgel RG. 2014. Importance of initial conditions in seasonal predictions of arctic sea ice extent. *Geophysical Research Letters* **41**: 5208–5215, doi:10.1002/2014GL060799.

Murphy AH. 1973. A new vector partition of the probability score. *J. Appl. Meteor.* **12**: 595–600, doi:10.1175/1520-0450(1973)012¡0595:ANVPOT¿2.0.CO;2.

Murphy AH. 1988. Skill Scores Based on the Mean Square Error and Their Relationships to the Correlation Coefficient. *Mon. Weather Rev.* **116**(12): 2417–2424.

Otto FEL, Ferro CAT, Fricker TE, Suckling EB. 2013. On judging the credibility of climate predictions. *Climatic Change* **in press**, doi: 10.1007/s10584-013-0813-5.

Peterson K, Arribas A, Hewitt H, Keen A, Lea D, McLaren A. 2015. Assessing the forecast skill of arctic sea ice extent in the glosea4 seasonal prediction system. *Climate Dynamics* **44**: 147–162, doi:10.1007/s00382-014-2190-9.

Pohlmann H, Botzet M, Latif M, Roesch A, Wild M, Tschuck P. 2004. Estimating the Decadal Predictability of a Coupled AOGCM. *J. Clim.* **17**: 4463–4472.

Rayner NA, Parker DE, Horton EB, Folland CK, Alexander LV, Rowell DP, Kent EC, Kaplan A. 2003. Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *J. Geophys. Res.* **108**: 4407, doi:10.1029/2002JD002670.

Robson JI, Sutton R, Lohmann K, Smith D, Palmer M. 2012. Causes of the rapid warming of the North Atlantic ocean in the mid 1990s. *J. Climate* : in pressdoi:10.1175/JCLI-D-11-00443.1.

Scaife AA, Arribas A, Blockley E, Brookshaw A, Clark RT, Dunstone N, Eade R, Fereday D, Folland CK, Gordon M, Hermanson L, Knight JR, Lea DJ, MacLachlan C, Maidens A, Martin M, Peterson AK, Smith D, Vellinga M, Wallace E, Waters J, Williams A. 2014. Skillful long-range prediction of european and north american winters. *Geophysical Research Letters* **41**(7): 2514–2519, doi:10.1002/2014GL059637.

Schroeder D, Feltham DL, Flocco D, Tsamados M. 2014. September arctic sea-ice minimum predicted by spring melt-pond fraction. *Nature Climate Change* **4**: 353–357, doi:10.1038/nclimate2203.

Shaffrey LC, Stevens I, Norton WA, Roberts MJ, Vidale PL, Harle JD, Jrrar A, Stevens DP, Woodage MJ, Demory ME, Donners J, Clark DB, Clayton A, Cole JW, Wilson SS, Connolley WM, Davies TM, Iwi AM, Johns TC, King JC, New AL, Slingo JM, Slingo A, Steenman-Clark L, Martin GM. 2009. U.K. HiGEM: The New U.K. High-Resolution Global Environment ModelModel Description and Basic Evaluation. *J. Clim.* **22**: 1861–1896, doi:10.1175/2008JCLI2508.1.

Sigmond M, Fyfe JC, Flato GM, Kharin VV, Merryfield WJ. 2013. Seasonal forecast skill of Arctic sea ice area in a dynamical forecast system. *Geophysical Research Letters* **40**: 529–534, doi:10.1002/grl.50129.

Smith DM, Scaife AA, Boer GJ, Caian M, Doblas-Reyes FJ, Guemas V, Hawkins E, Hazeleger W, Hermanson L, Ho CK, Ishii M, Kharin V, Kimoto M, Kirtman B, Lean J, Matei D, Merryfield WJ, Muller WA, Pohlmann H, Rosati A, Wouters B, Wyser K. 2013. Real-time multi-model decadal climate predictions. *Climate Dynamics* **41**: 2875, doi:10.1007/s00382-012-1600-0.

Stroeve J, Hamilton LC, Bitz CM, Blanchard-Wrigglesworth E. 2014. Predicting september sea ice: Ensemble skill of the search sea ice outlook 2008-2013. *Geophysical Research Letters* **41**: 2411–2418, doi: 10.1002/2014GL059388.

Tietsche S, Day JJ, Guemas V, Hurlin WJ, Keeley S, Matei D, Msadek R, Collins M, Hawkins E. 2014. Seasonal to interannual Arctic sea-ice predictability in current GCMs. *Geophys. Res. Lett* **41**: 1035, doi: 10.1002/2013GL058755.

Tietsche S, Notz D, Jungclaus JH, Marotzke J. 2013a. Assimilation of sea-ice concentration in a global climate model - physical and statistical aspects. *Ocean Science* **9**: 19–36, doi:10.5194/os-9-19-2013.

Tietsche S, Notz D, Jungclaus JH, Marotzke J. 2013b. Predictability of large interannual Arctic sea-ice anomalies. *Clim. Dyn.* **41**(9): 2511–2526, doi: 10.1007/s00382-013-1698-8.

Wang W, Chen M, Kumar A. 2013. Seasonal prediction of Arctic sea ice extent from a coupled dynamical forecast system. *Mon. Wea. Rev.* **141**: 1375–1394, doi:10.1175/MWR-D-12-00057.1.

Weisheimer A, Doblas-Reyes FJ, Palmer TN, Alessandri A, Arribas A, Dqu M, Keenlyside N, MacVean M, Navarra A, Rogel P. 2009. ENSEMBLES: A new multi-model ensemble for seasonal-to-annual predictions: Skill and progress beyond DEMETER in forecasting tropical Pacific SSTs. *Geophysical Research Letters* **36**, doi:10.1029/2009GL040896.

Weisheimer A, Palmer TN, Doblas-Reyes FJ. 2011. Assessment of representations of model uncertainty in monthly and seasonal forecast ensembles. *Geophysical Research Letters* **38**(16), doi:10.1029/2011GL048123.

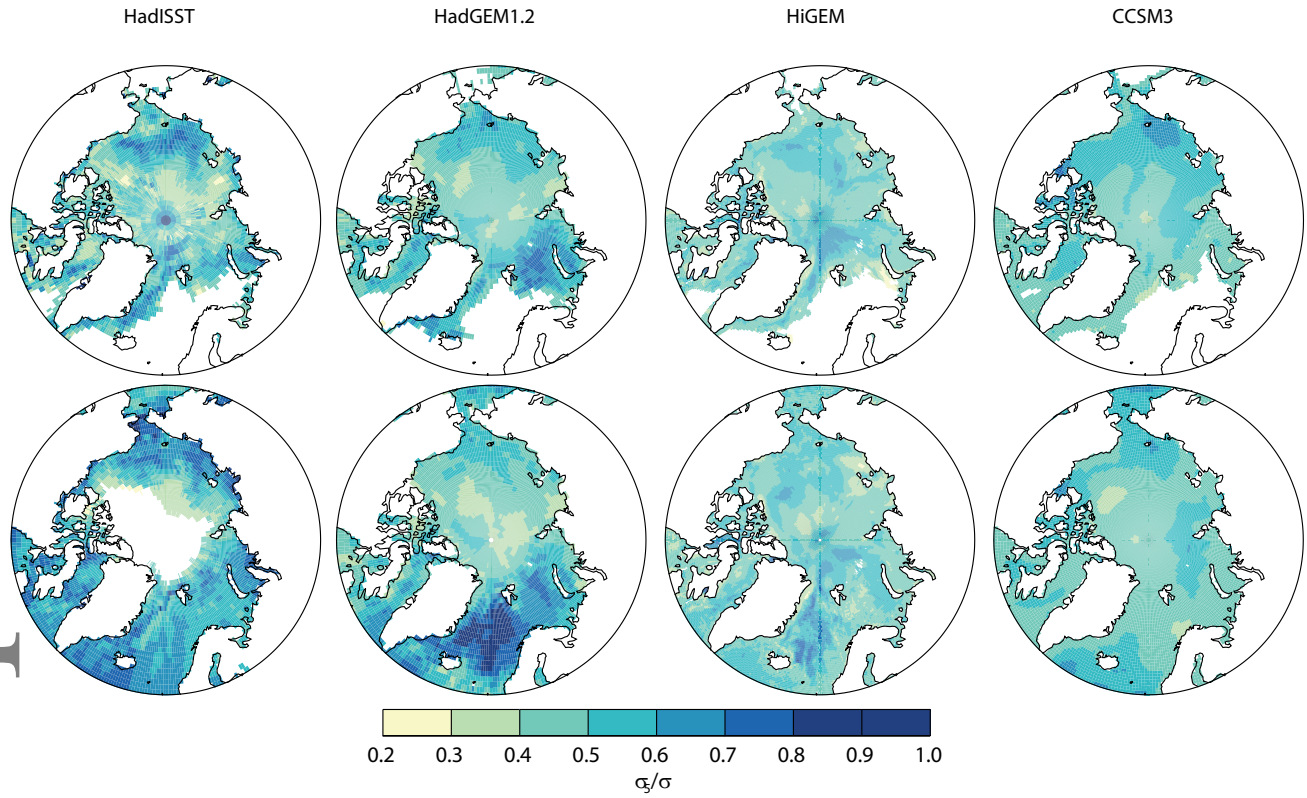

Figure 1. Diagnostic potential predictability (DPP, σ_5/σ_1) for August-September-October (ASO) sea-ice concentration (sic) and sea surface temperature (sst) using observations (left, HadISST, Rayner *et al.* (2003)) and present day control simulations for various GCMs as labelled.

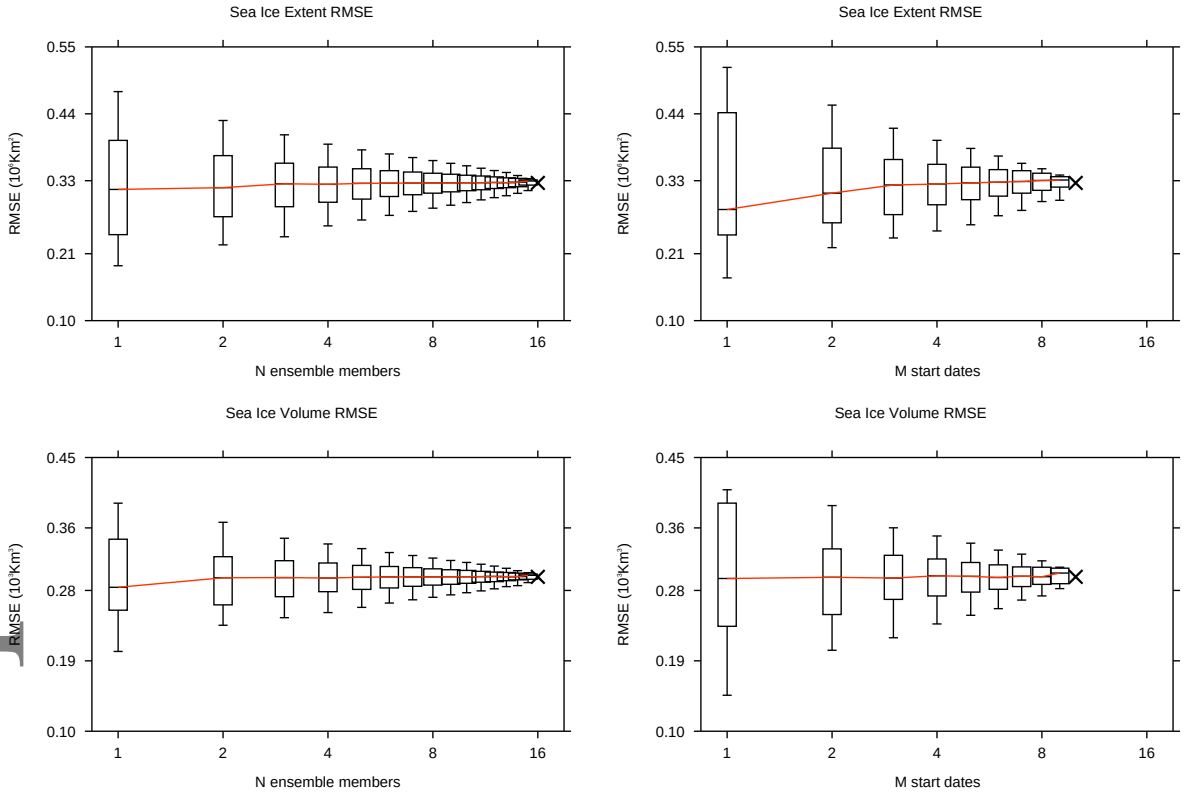

Figure 2. Box and whisker plots showing quantiles (5%, 25%, 50%, 75% and 95%) of RMSE in Arctic sea-ice extent (top) and volume (bottom) for September (forecast month 3), when averaged over distinct subsets. Left: all possible choices of N (out of 16) ensemble members. Right: all possible choices of M (out of 10) start dates.

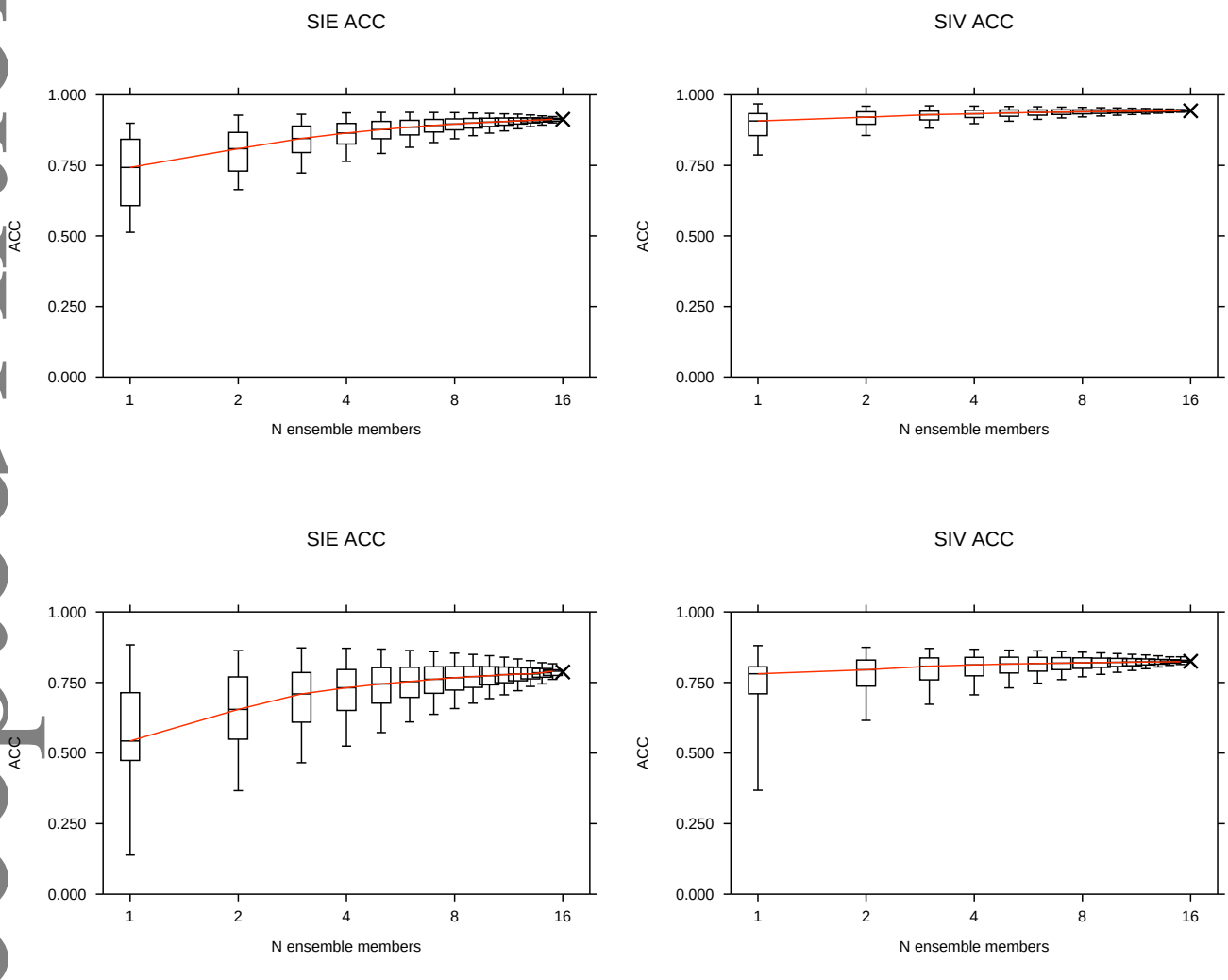

Figure 3. Box and whisker plots showing quantiles (5%, 25%, 50%, 75% and 95%) of anomaly correlation of Arctic sea-ice extent (left) and volume (right) for September, when averaged over all distinct subsets of N ensemble members (out of 16). The rows use different start months: May (top) and January (bottom), so that September is at a lead time of 5 and 9 months respectively.

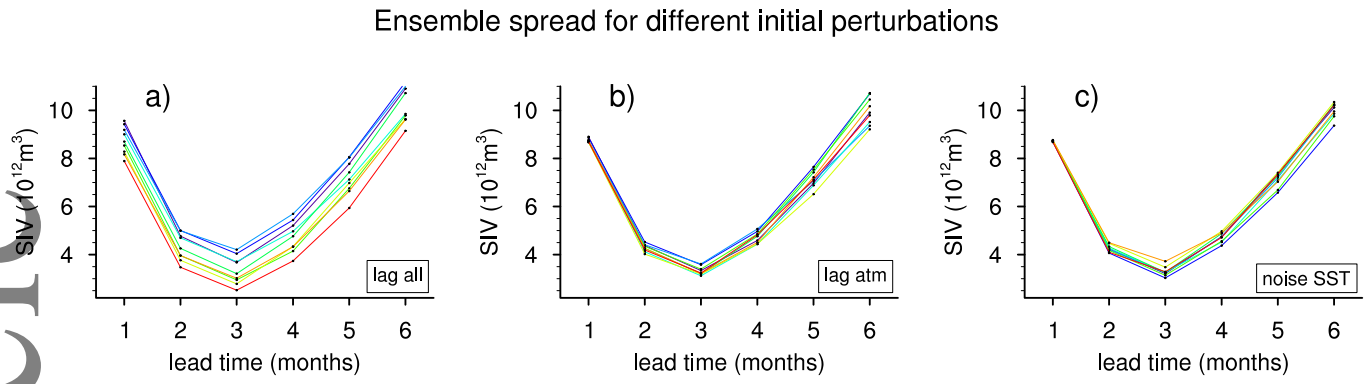

Figure 4. Arctic sea-ice volume ensembles for a case study prediction started on 1st July with MPI-ESM-LR with different types of initial-condition perturbations: (a) atmospheric and oceanic state perturbed by a few days (SL), (b) only atmospheric state perturbed by a few days (AL), (c) SST perturbed by tiny amount of noise (NSST).

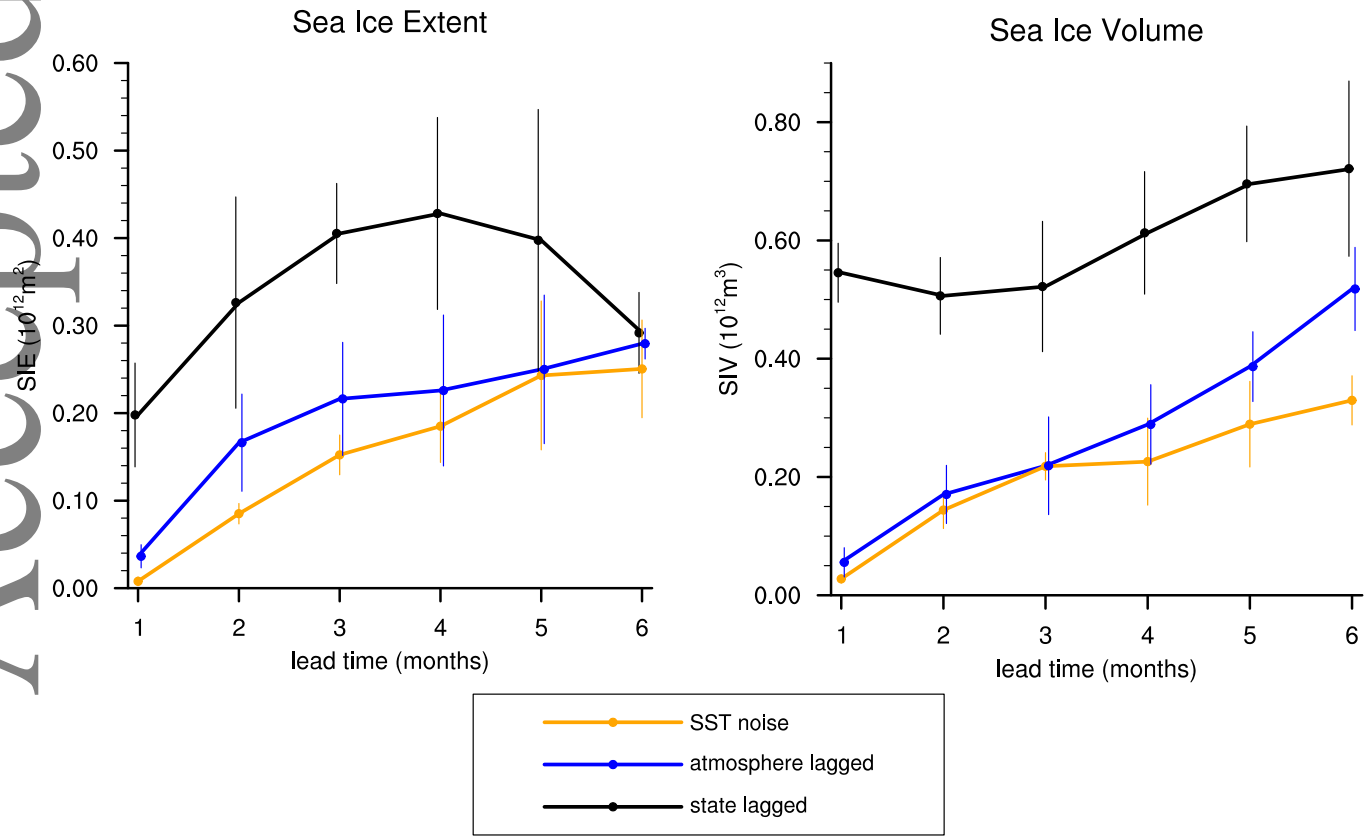

Figure 5. Ensemble spread for different initial perturbations for sea-ice extent (left) and sea-ice volume (right). Ensembles from 4 different start years have been used to calculate the mean spread (thick solid lines with markers) and the standard deviation of the spread (thin error bars) as an indication of how it varies between different years.

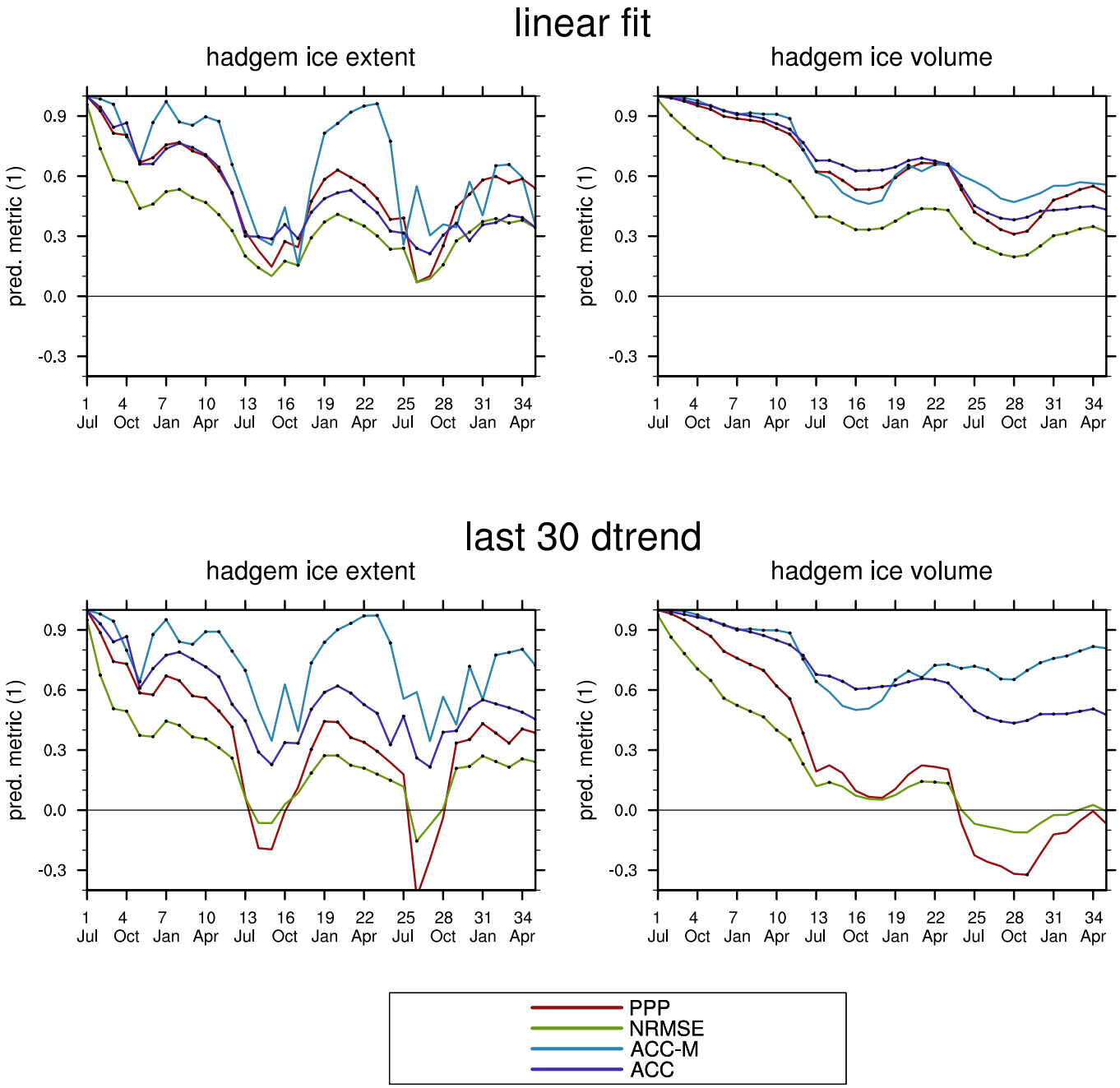

Figure 6. Comparison of different deterministic prediction skill metrics applied to the same data. Upper row: climatological mean and standard deviation defined from linear fit of a long control run - the maximum knowledge (MK) approach. Lower row: climatological mean and standard deviation defined from 30 years prior to prediction start date - the operational approach (OP). Black dots indicate statistical significance of the metric being larger than zero at the 95% confidence level.

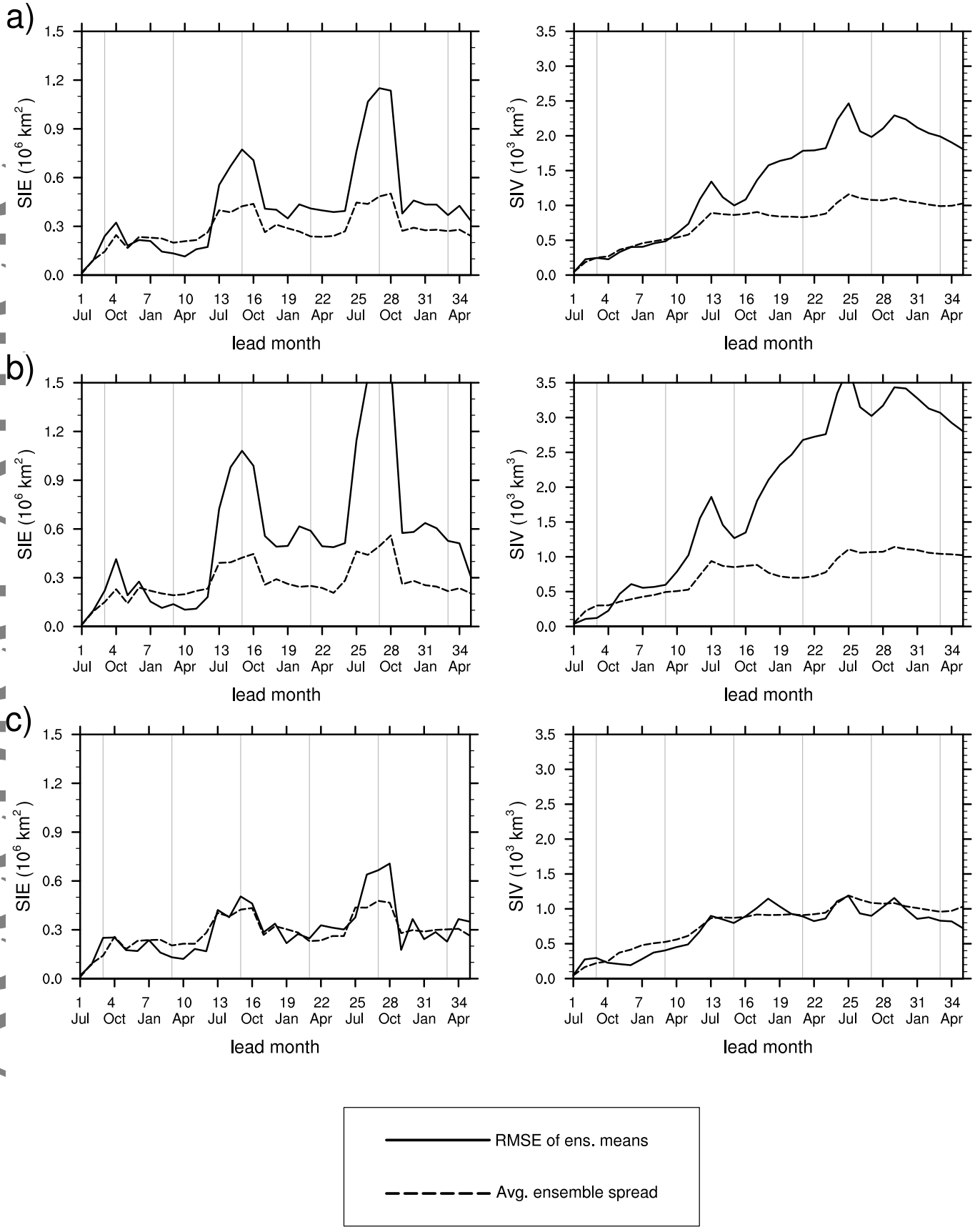

Figure 7. Comparison of ensemble spread and ensemble-mean RMSE for GFDL CM3. (a) Over all 8 start dates, (b) over the three start dates where ensembles are highly under-dispersive, and (c) the remaining five start dates.

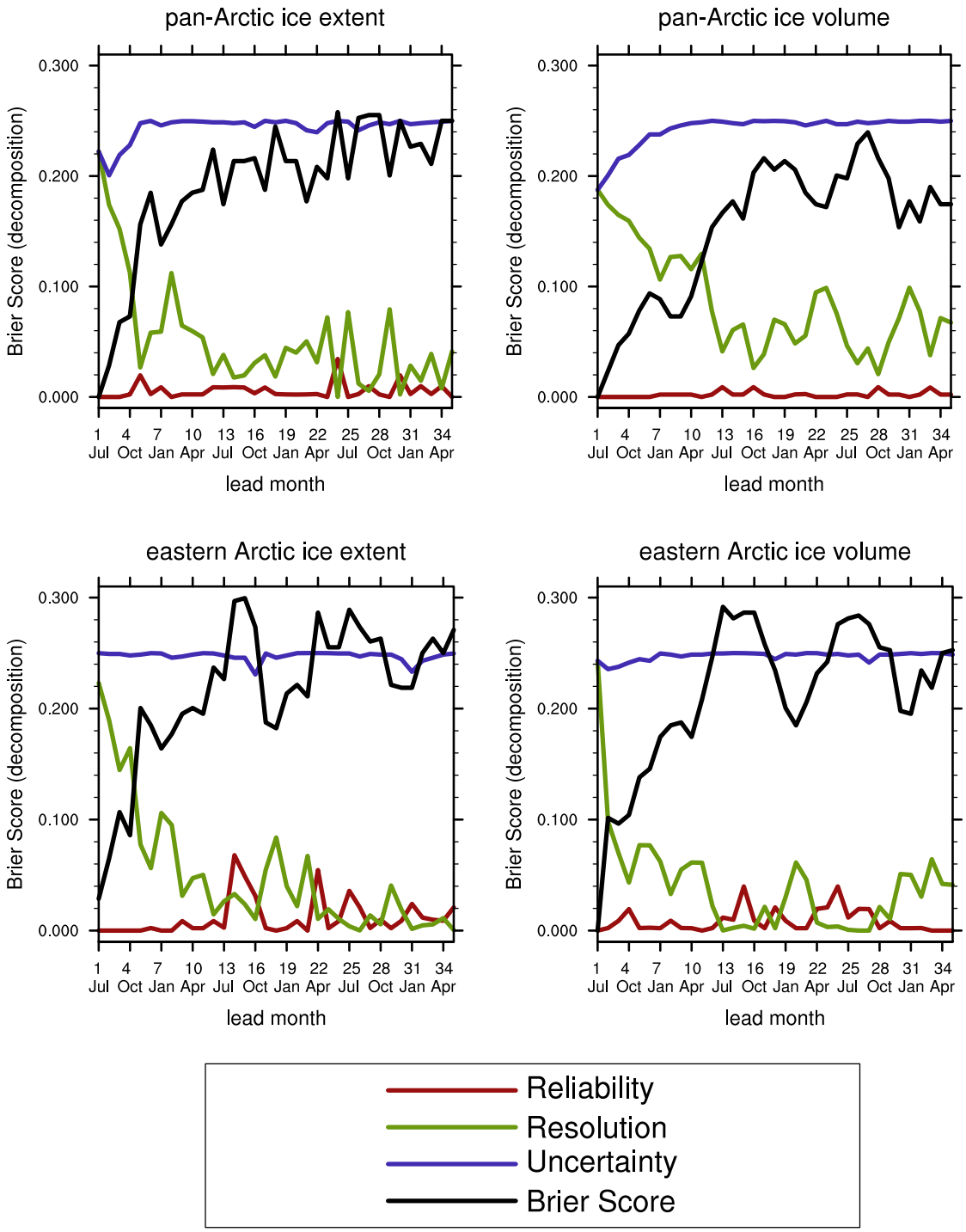

Figure 8. The Brier Score (black) and decomposed into its three components as labelled, for above climatological values of sea-ice extent (left) and volume (right) in the MPI-ESM-LR perfect model predictability experiments for pan-Arctic sea ice extent (top) and the eastern Arctic (bottom).