Initializing Sea Ice Thickness and Quantifying Uncertainty in Seasonal Forecasts of Arctic Sea Ice

by

Arlan Dirkson
B.A. University of Montana, 2013

A Dissertation Submitted in Partial Fulfillment of the
Requirements for the Degree of

DOCTOR OF PHILOSOPHY

in the School of Earth and Ocean Sciences

Initializing Sea Ice Thickness and Quantifying Uncertainty in Seasonal Forecasts of Arctic Sea Ice

by

Arlan Dirkson
B.A. University of Montana, 2013

Supervisory Committee

Dr. William J. Merryfield, Co-Supervisor
(School of Earth and Ocean Sciences)

Dr. Adam Monahan, Co-Supervisor
(School of Earth and Ocean Sciences)

Dr. Michael Sigmond, Departmental Member
(School of Earth and Ocean Sciences)

Dr. Slava Kharin, Outside Member
(Canadian Centre for Climate Modelling and Analysis)

Supervisory Committee

Dr. William J. Merryfield, Co-Supervisor
(School of Earth and Ocean Sciences)

Dr. Adam Monahan, Co-Supervisor
(School of Earth and Ocean Sciences)

Dr. Michael Sigmond, Departmental Member
(School of Earth and Ocean Sciences)

Dr. Slava Kharin, Outside Member
(Canadian Centre for Climate Modelling and Analysis)

ABSTRACT

Arctic sea ice has undergone a dramatic transformation in recent decades, including a substantial reduction in sea ice extent in summer months. Such changes, combined with relatively recent advancements in seasonal (1-12 months) to decadal forecasting, have prompted a rapidly-growing body of research on forecasting Arctic sea ice on seasonal timescales. These forecasts are anticipated to benefit a vast array of end-users whose activities are dependent on Arctic sea ice conditions. The research goal of this thesis is to address fundamental challenges pertaining to seasonal forecasts of Arcitc sea ice, with a particular focus placed on improving operational sea ice forecasts in the Canadian Seasonal to Interannual Prediction System (CanSIPS).

Seasonal forecasts are strongly dependent on the accuracy of observations used as initial condition inputs. A key challenge initializing Arctic sea ice is the sparse availability of Arctic sea ice thickness (SIT) observations. I present on the development of three statistical models that can be used for estimating Arctic SIT in real time for sea ice forecast initialization. The three statistical models are shown to vary in their ability to capture the recent thinning of sea ice, as well as their ability to capture interannual variations in SIT anomalies; however, each of the models is shown to dramatically improve the representation of SIT compared to the climatological SIT estimates used to initialize CanSIPS.

I conduct a thorough assessment of sea ice hindcast skill using the Canadian Climate Model, version 3 (one of two models used in CanSIPS), in which the dependence of hindcast skill on SIT initialization is investigated. From this assessment, it can be concluded that all three statistical models are able to estimate SIT sufficiently to improve hindcast skill relative to the climatological initialization. However, the accuracy with which the initialization fields represent both the thinning of the ice pack over time and interannual variability impacts predictive skill for pan-Arctic sea ice area (SIA) and regional sea ice concentration (SIC), with the most robust improvements obtained with two statistical models that adequately represent both processes.

The final goal of this thesis is to improve the quantification of uncertainty in seasonal forecasts of regional Arctic sea ice coverage. Information regarding forecast uncertainty is crucial for end-users who want to quantify the risk associated with trusting a particular forecast. I develop statistical post-processing methodology for improving probabilistic forecasts of Arctic SIC. The first of these improvements is intended to reduce sampling uncertainty by fitting ensemble SIC forecasts to a para-

metric probability distribution, namely the zero- and one- inflated beta (BEINF) distribution. It is shown that overall, probabilistic forecast skill is improved using the parametric distribution relative to a simpler count-based approach; however, model biases can degrade this skill improvement. The second of these improvements is the introduction of a novel calibration method, called trend-adjusted quantile mapping (TAQM), that explicitly accounts for SIC trends and is specifically designed for the BEINF distribution. It is shown that applying TAQM greatly reduces model errors, and results in probabilistic forecast skill that generally surpasses that of a climatological reference forecast, and to some degree that of a trend-adjusted climatological reference forecast, particularly at shorter lead times.

Contents

List of Tables

List of Figures

ACKNOWLEDGEMENTS

I thank my family for their unwavering support throughout my academic career. I especially thank my parents, Douglas and Sarah, and step-mom, Evelyn, for instilling in me the confidence to pursue my passion. I thank my Grandpa, Tom, for taking an interest in my academic success and cultivating my love for science. I thank my sister, Sasha, for continually looking out for my well-being and always believing in me. I thank my younger brothers, Sam and Sage, for their love and support.

There are several friends I owe gratitude. I thank SEOS alumni Ben Johnson, Bennit Mueller, Duncan Mackay, Mitchell Wolf, Jess Shaw, and Kassandra Del Greco for sharing many o' beers, laughs, and lasting memories. I also appreciate the entertainment of my cubical neighbour, Alicia Lew. I would like to thank Ale Perez for all of her support, as well as that of my long-time friend and brother, Jon Piepenbrink.

I'd also like to thank Allison in the SEOS main office for her help in keeping me organized.

I thank a number of professors and mentors from my past, including Don Hickethier and Johnathan Bardsely for sharing their passion for mathematics with me. I thank Joel Harper, my undergraduate mentor, for providing me with countless opportunities to develop as a young scientist.

Finally, I would like to thank my supervisors, Adam Monahan and Bill Merryfield, for all of their guidance throughout this process. I also thank my two committee members, Michael Sigmond and Slava Kharin, for their thoughtful involvement with all of my work. I thank Woosung Lee for producing the hindcasts used in this dissertation, as well as for her help with our contribution to the 2017 Sea Ice Outlook.

"If you're walking down the right path and you're willing to keep walking, eventually you'll make progress."

–Barack Obama

DEDICATION

I dedicate this dissertation to my family for always encouraging me to strive for my dreams.

Chapter 1

Introduction

1.1 Background and Motivation

Arctic sea ice is a defining feature of the Earth's surface and a key component of the global climate system. For millenia it has played an intimate cultural role in the lives of those who reside in the Arctic region. More recently, dramatic changes in sea ice conditions associated with a transition to a warmer climate (e.g., Meier et al., 2014b) have prompted the economic and social considerations of a much broader reach of society (Ellis and Brigham, 2009).

Over the last four decades, Arctic sea ice has transformed considerably. Satellite observations reveal that total Arctic sea ice extent (SIE) has decreased in all calendar months (e.g., Parkinson et al., 1999; Meier et al., 2007; Serreze et al., 2007; Comiso et al., 2008; Stroeve et al., 2012b; Comiso et al., 2017), with the largest magnitude of decline of $-13.3\pm2.6\%$ per decade (1979-2016) observed in September (National Snow and Ice Data Center, Sea Ice Index Version 2). These large reductions in SIE have led to an overall younger ice pack. Indeed, the extent of multi-year ice (i.e. ice that has survived at least one melt season) has declined by 33% in March and by 50% in September from 1980-2011 (Maslanik et al., 2011). Although sea ice thickness (SIT) has been only sparsely observed compared to SIE, a range of observations reveal that Arctic sea ice has thinned substantially (Rothrock et al., 1999; Kwok and Rothrock, 2009), with recent analyses showing a declining rate of -0.58 ± 0.07 m per decade from 1975-2012, and a decrease in mean sea ice thickness (SIT) from 3.9 m to 1.5 m (Lindsay and Schweiger, 2015).

Negative trends in Arctic sea ice coverage are expected to continue into the future

under continued increases in greenhouse gas concentrations (e.g. Stroeve et al., 2012a); however, substantial uncertainties in sea ice projections remain due to model biases, inter-model uncertainty, internal climate variability (e.g. Kay et al., 2011; Swart et al., 2015), and uncertainty in emission scenario (Stroeve et al., 2012a; Liu et al., 2013). These projected changes, in addition to those that have already taken place, have drawn the attention of a wide range of socially– and economically– impacted stakeholders, including indigenous populations, fishing communities, the shipping industry, various resource exploitation industries, the ecotourism industry, and decision makers (Ellis and Brigham, 2009). The potential to forecast sea ice conditions months in advance, and to provide such end-users with sea ice information for planning efforts, has prompted a rapidly-growing area of research on sea ice prediction on seasonal time scales (Guemas et al., 2016).

Both statistical and dynamical models have been explored for forecasting sea ice on seasonal time scales. Currently, statistical models are nearly as skilful as fully-coupled atmosphere-ocean global climate models (AOGCMs) at forecasting end of summer sea ice conditions (Guemas et al., 2016). However, whereas statistical predictions will be limited by non-stationarities between the various predictors and sea ice predictands used in such models (Lindsay et al., 2008; Holland et al., 2011), AOGCMs have the ability to capture such relationships. This suggests that in a continued non-stationary climate, AOGCMs have a greater potential to outperform such statistical models. Furthermore, perfect-model studies suggest that initialized forecasts have not yet reached their potential. In particular, additional untapped skill in AOGCM-based forecasts is expected from improving sea ice thickness (SIT) initialization (e.g. Lindsay et al., 2012; Tietsche et al., 2013; Day et al., 2014), improving model physics (Blanchard-Wrigglesworth et al., 2015), resolution, and correcting model biases *a posteriori* (Krikken et al., 2016; Blanchard-Wrigglesworth et al., 2016; Director et al., 2017).

The sparse availability of pan-Arctic SIT observations for initializing both real-time forecasts and hindcasts likely inhibits Arctic sea ice prediction skill. Unlike sea ice concentration (SIC), which has been well-observed by satellites since the late 1970's, no continuous and spatially extensive record of SIT exists over this period. Submarine-based measurements of SIT are archived at the National Snow and Ice Data Center (NSIDC) over a relatively long record from 1960-2005. In addition, there exists a suite of point measurements of SIT, made by a collection of moorings and buoys. Satellite missions from 2003-2009 (ICESat) (Kwok and Cunning-

ham, 2008) and from 2010-present (CryoSat-2) (Wingham et al., 2006) have taken spatially-extensive measurements using laser and radar altimetry which can be used to estimate sea ice freeboard and infer SIT. NASA's Operation IceBridge (2009-present) was launched to "bridge the gap" in SIT measurements between ICESat and ICESat-2 using airborne laser altimetry (Kurtz et al., 2012). However, none of these datasets provide "continuous" observations over both space and time covering the most recent 30-year time period needed for initializing hindcasts for validation, nor are they available in real time throughout the year.

An early approach for handling the SIT initialization problem in AOGCMs allowed the model to determine the SIT distribution while assimilating surface forcing from reanalysis (e.g. Zhang and Rothrock, 2003; Chevallier et al., 2013; Guemas et al., 2013). More recent analyses include the assimilation of well-observed SIC (e.g. Peterson et al., 2015). Another approach updates SIT in the assimilation procedure based on analysis updates of SIC through a simple hypothesized proportionality relationship between these quantities (Tietsche et al., 2013). The Canadian Seasonal to Interannual Prediction System (CanSIPS) (Merryfield et al., 2013b), whose sea ice predictions are considered in this thesis, initializes SIT by assimilating surface forcings and SIC observations, while relaxing SIT toward a model-based seasonally varying climatology.

Since 2014, the Sea Ice Outlook (SIO) has been accepting forecasts of spatially-distributed *sea ice probability* (SIP), which describes the probability that sea ice will be present locally within a model grid cell. This communication of uncertainty is a fundamental step in the seasonal forecasting process (Troccoli et al., 2008), as uncertainties in predicting the climate system on seasonal time scales can be substantial. One source of uncertainty arises from the sensitivity of the climate system to minute changes in initial conditions. Ensemble forecasting methods, in which multiple forecasts are generated from slightly different initial conditions, offer a means of sampling this uncertainty (e.g. Reynolds et al., 1994). Forecasting a categorical quantity like SIP can be done directly from the finite ensemble generated. However, it is well established that ensemble forecasts benefit from post-processing methods, as these provide only a small sample of the underlying forecast distribution (Wilks, 2002). Furthermore, model errors form another basis of uncertainty in seasonal forecasts. Methods for correcting for such errors in probabilistic forecasts of Arctic sea ice coverage have only just begun to be explored (Krikken et al., 2016).

1.2 Research Goals and Outline

The goal of this dissertation is to address fundamental challenges pertaining to seasonal forecasts of Arctic sea ice in AOGCMs. In particular, I develop methods for addressing these challenges, and test them using one of the two models used in CanSIPS.

The first of these challenges is the initialization of sparsely-observed SIT. This problem is addressed by developing and testing three statistical models for improving upon the current climatological initialization technique used in CanSIPS. It is shown how these statistical models, which vary in complexity and predictor information, differ in their ability to capture the thinning of sea ice and interannual variations in SIT. This work is described in chapters 2 and 3, and is published in *Geophysical Research Letters* as Dirkson et al. (2015) and the *Journal of Climate* as Dirkson et al. (2017).

Each of these statistical models is then used to initialize SIT in hindcasts using the Canadian Center for Climate Modelling and Analysis Canadian Climate Model, version 3 (CanCM3) (one of two models used in CanSIPS) over a 32-year period from 1981-2012. The dependence of hindcast skill on SIT initialization is assessed by comparing hindcasts initialized with each statistical model against hindcasts initialized with the method employed in CanSIPS. In this assessment, I focus on pan-Arctic sea ice area, as well as regional sea ice coverage. The impact that SIT initialization has on model biases is also explored. This work is described in chapter 3 and is published in the *Journal of Climate* as Dirkson et al. (2017).

The second challenge I address involves the quantification of uncertainty in seasonal forecasts of Arctic sea ice. In particular, I consider the uncertainty in spatially distributed sea ice concentration (SIC). I first apply a parametric distribution to ensemble forecasts of SIC in order to infer the underlying forecast distribution, and assess the impact this has on probabilistic forecast skill. I then introduce a parametric calibration method specifically designed for SIC forecasts, and assess the efficacy of this method at reducing model biases. This work is presented in chapter 4 and will be submitted to the *Journal of Climate*.

The methods developed in this dissertation have recently been applied in CanSIPS to produce real-time forecasts submitted to the 2017 SIO. A brief discussion of these SIO contributions is given in chapter 5, and conclusions are presented in chapter 6.

Chapter 2

Real-time estimation of Arctic sea ice thickness through maximum covariance analysis

The following chapter is a manuscript published as:

Dirkson, A., Merryfield, W.J. and Monahan, A., 2015. Realtime estimation of Arctic sea ice thickness through maximum covariance analysis. *Geophysical Research Letters*, 42(12), pp.4869-4877.

The manuscript is repeated here with small modifications to fit the format of this dissertation.

2.1 Abstract

A challenge for model-based seasonal predictions of sea ice is an accurate representation of sea ice initial conditions, particularly sparsely-observed sea ice thickness (SIT). The Canadian Seasonal to Interannual Prediction System (CanSIPS) currently initializes SIT by nudging simulated values toward a model-based climatology. To improve on this, we use sea ice data from Pan-Arctic Ice Ocean Modeling and Assimilation System (PIOMAS) to investigate how accurately SIT can be estimated in real time using better-observed and physically-relevant predictors. We: 1) test the skill of several predictors using maximum covariance analysis (MCA), 2) apply an approach which blends sea ice concentration and lagged (4-month averaged) sea level pres-

sure, and 3) compare this method against the current CanSIPS initialization scheme over 1981-2012. The MCA-based statistical model reduces SIT areal- and temporal-mean absolute errors by 48% relative to the current CanSIPS initialization and shows consistent skill estimating ice volume in all months ($r = 0.95$).

2.2 Introduction

Reductions in Arctic sea ice extent and volume have provided a new impetus for modelling and forecasting high-latitude climate (NRC, 2010; Eicken, 2013). A particular motivation for these efforts is an increasing interest in the maritime access of the Arctic region by stakeholders ranging from fishing communities and marine tourism operators to commercial shippers and national security organizations (Ellis and Brigham, 2009). Beyond influencing marine accessibility, Arctic sea ice plays an important role in the global climate through the ice-albedo feedback, contributing to recent Arctic amplification (Screen and Simmonds, 2010). These reductions in sea ice have been shown to affect local atmospheric variability (e.g., Simmonds and Keay, 2009; Porter et al., 2012) and polar climate sensitivity (Holland et al., 2006), although the degree to which they might impact remote weather (e.g. mid-latitude temperature or cyclone intensity) and dominant modes of variability (e.g. the North Atlantic Oscillation), is an ongoing area of research (e.g., Francis et al., 2009; Overland and Wang, 2010; Cohen et al., 2014; Mori et al., 2014; Barnes et al., 2014).

Atmosphere-ocean global climate models (AOGCMs) initialized using observational data are increasingly being applied to predict sea ice on seasonal timescales (1-12 months). The skill of such predictions hinges on the ability to accurately predict climate variability and associated coupled processes, as well as to adequately represent climate system initial conditions. Recent work suggests that there is inherent initial-value predictability in Arctic sea ice out to two years (Holland et al., 2011; Blanchard-Wrigglesworth et al., 2011b), partly owing to the relatively long memory of its thickness anomalies (Blanchard-Wrigglesworth and Bitz, 2014). Sea ice thickness is therefore a variable of particular interest for initializing an AOGCM (Doblas-Reyes et al., 2013). Within a "perfect model" framework, Day et al. (2014) showed that by accurately initializing thickness in a July forecast, skill in forecasting Arctic sea ice extent could be improved out to September of the following year. When initialized in January, errors in thickness were found to have less of an impact on Arctic ice extent forecasts; however, associated errors in conductive heat flux through the sea ice result

in biases in the overlying atmosphere, specifically 2 m temperature.

Unlike sea ice concentration, which has been continuously observed by satellites over the past few decades, thickness observations typically have been sparse (Lindsay and Schweiger, 2015). Although altimeter-based satellite measurements have begun to fill this gap, these do not span the multidecadal period needed for initializing hindcasts, and in any case are not available in real time in all calendar months (in particular June-September), as required by operational forecasting systems. This lack of observations poses a challenge for initializing sea ice hindcasts and real-time forecasts and potentially limits forecast quality. One proposed method for initializing thickness in seasonal predictions using AOGCMs relaxes ice thickness at a rate proportional to the difference between the model background and the observed values of ice concentration (Tietsche et al., 2013). Attempts have also been made to model ice distribution and thickness using coupled ice-ocean models driven by reanalysis-based surface forcing (e.g., Zhang and Rothrock, 2003; Chevallier et al., 2013; Guemas et al., 2013; Msadek et al., 2014), while others include the assimilation of observed sea ice concentration (Peterson et al., 2015). These efforts represent first approaches to the sea ice thickness initialization problem and additional approaches continue to be explored.

Our aim is to establish a computationally inexpensive, statistically-based method for estimating Arctic sea ice thickness in real time using better-observed and physically relevant predictors. This method will ultimately be assessed within the Canadian Seasonal to Interannual Prediction System (CanSIPS), the details of which are outlined in Merryfield et al. (2013b). Currently, CanSIPS initializes sea ice using an ensemble of assimilation runs, in which sea ice concentration is relaxed toward HadISST (Rayner et al., 2003) values. In grid cells where concentration exceeds a threshold value, sea ice thickness is relaxed to values from a seasonally varying, model-based climatology. However, because sea ice thickness has been subject to an accelerating decreasing trend, such a climatological initialization is inadequate. This is likely responsible, at least in part, for weaker than observed trends in Arctic sea ice predictions (Merryfield et al., 2013a), which in turn undermines the predictive skill obtained from capturing the trend (Sigmond et al., 2013). Apart from variations in the ice edge as represented in the concentration field, neither interannual variability nor the long-term trend are accounted for in the current thickness initialization. Significant potential therefore exists for improving sea ice thickness initialization in CanSIPS, although the sparse observational record poses a substantial challenge as

discussed previously.

In the remainder of this study, we first describe the data and statistical predictors considered (section 2.3.1). We briefly describe the statistical model framework in section 2.3.2, and assess the application of different predictors in section 2.4.1. In section 2.4.2 we identify an optimal predictor, by which initialization fields are constructed, and in section 2.4.3 these are compared against the current thickness fields used to initialize CanSIPS over the period 1981-2012.

2.3 Methods

2.3.1 Data and Predictors

As a means of constructing and validating statistical models, we use reconstructions of monthly sea ice concentration (SIC) and sea ice thickness (SIT) from the Pan-Arctic Ice Ocean Modeling and Assimilation System (PIOMAS) over the period 1979-2012 (Zhang and Rothrock, 2003). PIOMAS is a regional coupled sea ice/ocean model which assimilates National Snow and Ice Data Center (NSIDC) SIC, and is forced with atmospheric fields and sea surface temperatures from the National Centers for Environmental Prediction (NCEP) / National Center for Atmospheric Research (NCAR) reanalysis (Kalnay et al., 1996). Sea ice fields from PIOMAS are regularly updated, but are not available in real time. PIOMAS SIT fields have been assessed using CryoSat-2 over the winters of 2010/11 and 2011/12, and were found to underestimate the rate of volume decline in autumn, whereas losses in winter are overestimated (Laxon et al., 2013). Compared against ICESat derived values over a longer record, Schweiger et al. (2011) found PIOMAS to agree reasonably well with respect to both errors (< 0.1 m mean difference) and pattern correlation ($r > 0.8$). Finally, Stroeve et al. (2014a) found that compared with a range of observational data sets spanning different time periods, PIOMAS overestimates thin ice and underestimates thick ice. We also use monthly sea level pressure (SLP) fields from the ERA-Interim (for 1979-2012) (Dee et al., 2011) and ERA-40 (for 1978) (Uppala et al., 2005) reanalyses.

Five predictors that are available in near-real time and can hence be applied in an operational setting are considered for testing in this study: three single-field predictors, SIC, SLP, and lagged SLP (SLPlag); and two combined-field predictors, SIC and SLP (SIC&SLP), and SIC and SLPlag (SIC&SLPlag). SLPlag is a 4-month av-

erage of SLP over the previous three months and including the month of interest, and each combined predictor (SIC&SLP and SIC&SLPlag) is, by virtue of the statistical technique, treated as a single field.

In addition to their availability in near-real time, these predictors have been chosen for their physical relevance. SIC is expected to be an effective predictor of thickness changes near the ice boundary where both quantities are variable, although less skilful in areas where SIC is regularly 100% but SIT varies. In addition, SIC is expected to represent the thermally-driven decline in Arctic sea ice. Although near-surface (2 m) air temperature poleward of 60°N from ERA-Interim was also considered as a predictor, it did not perform as well as SIC. Other studies have considered the relationships between sea ice and thermodynamically important variables such as surface energy fluxes, humidity, and melt ponds (Drobot et al., 2006; Kapsch et al., 2013; Schröder et al., 2014). Thermodynamically-related predictive information contained in these quantities should also be contained in near-surface air temperatures, and some of these quantities (such as melt pond fraction) are available and relevant to SIT only in certain seasons. Therefore, we focus on SIC as a thermodynamically-based predictor of SIT. Sea ice thickness variability resulting from wind-driven transport though convergence, ridging, and divergence (Bitz et al., 2002; Rigor et al., 2002) is not expected to be well captured by the SIC field. To represent such variability, we consider SLP north of 48°N (the southernmost latitude included in PIOMAS). In addition to the current-month SLP, SLPlag has been considered to account for cumulative effects associated with slowly varying SLP anomalies. No advantage was found in using averaging periods longer than four months, or in using an exponentially decaying weighting rather than a simple average over the four months. Together, these predictors are expected to capture both thermodynamic and mechanical aspects of SIT evolution.

2.3.2 Statistical Model

The statistical prediction technique adopted for this study, maximum covariance analysis (MCA), identifies pairs of vectors from spatiotemporal fields which maximize covariance between their associated expansion coefficients, under a constraint of orthonormality between vectors within each field (Bretherton et al., 1992; Von Storch and Zwiers, 2001). In the case of a single predictor field, each pair of vectors (or mode) can be interpreted as representing two fixed spatial patterns - one for the

predictor and one for the predictand. When combined predictor fields (SIC&SLP or SIC&SLPlag) are used, three maps result - two maps for the predictors and one for the predictand. These patterns can be applied to a regression-based model which minimizes the mean sum of squared errors between the regression prediction and the predictand projection time series. It should be noted that if one were to use the entire set of patterns in the regression model, MCA would be no different than multiple linear regression utilizing the whole field; but by choosing a reduced number of patterns to represent only the dominant modes of co-variability, we minimize the number of regression parameters to be estimated and therefore the effects of overfitting.

To identify patterns of co-variability to make real-time estimates of the predictand $\tilde{\mathbf{d}}_m(t_e)$ (where the tilde distinguishes the estimated predictand from the true predictand $\mathbf{d}_m(t_e)$), from the predictor $\mathbf{r}_m(t_e)$, for month m and year t_e, a moving training window τ over the fifteen previous years (i.e. $\tau = t_e - 15, t_e - 14, \cdots, t_e - 1$) is considered. This particular length of training window is adopted to balance the non-stationarity of recent sea ice variability in the Arctic with an adequate time series to effectively identify coherent patterns of co-variability.

To increase computational efficiency, an empirical orthogonal function (EOF) prefiltering is performed on each temporally centred field prior to executing MCA. This consists of taking the leading I principal components (PCs), $\mathbf{x}_m(\tau) = \mathbf{E}_r^{\mathrm{T}}\mathbf{r}'_m(\tau)$ of the predictor field anomalies, and the leading J PCs, $\mathbf{y}_m(\tau) = \mathbf{E}_d^{\mathrm{T}}\mathbf{d}'_m(\tau)$ of the predictand field anomalies (where the prime symbol denotes anomalies), constructed using the non-square projection matrices $\mathbf{E}_r$ and $\mathbf{E}_d$, whose columns contain the corresponding fields' leading I or J eigenvectors. For every field considered, we use $I = J = 15$, which is the maximum number of non-zero eigenvalues produced from each field constrained by the length of the training period; thus, no temporal variance is lost during this step. When combined predictor fields (SIC&SLP and SIC&SLPlag) are used, $\mathbf{x}(t)$ is the concatenation of PCs calculated for each individual field, with each field normalized by their spatially averaged temporal standard deviation.

Here, MCA is thus based on the singular value decomposition of the cross covariance matrix of the predictor and predictand fields' leading PCs,

$$\mathbf{C}_{xy} = \langle \mathbf{x}_m \mathbf{y}_m^{\mathrm{T}} \rangle_\tau = \sum_{j=1}^{N} \sigma_j \mathbf{u}_j \mathbf{v}_j^{\mathrm{T}}, \tag{2.1}$$

where angle brackets $\langle\rangle_\tau$ denote the expected value over the training period τ, $\mathbf{u}_j$

represents the j^{th} projected predictor co-variability pattern, $\mathbf{v}_j$ is the j^{th} projected predictand co-variability pattern, and σ_j is their corresponding singular value.

The regression model uses a reduced number ($K < N$) of mode pairs, in addition to the predictor PCs for month m and year t_e, to make a least-squares estimate of the predictand PCs,

$$\tilde{\mathbf{y}}_m(t_e) = \sum_{j=1}^{K} \frac{\sigma_j}{\text{var}(\mathbf{x}_m^{\text{T}}\mathbf{u}_j)} [\mathbf{x}_m(t_e)^{\text{T}}\mathbf{u}_j]\mathbf{v}_j. \tag{2.2}$$

The regression-based predictand field is constructed as

$$\tilde{\mathbf{d}}_m(t_e) = \langle \mathbf{d}_m \rangle_\tau + \mathbf{E}_d \tilde{\mathbf{y}}_m(t_e). \tag{2.3}$$

2.3.3 Monte Carlo Significance Tests

In applying MCA, the number of singular vector modes to be retained must be determined. To inform this choice, estimates of SIT for the months of September and March over the period 1995-2012 are considered for each predictor, with the number of modes varied from $K = 1$ to $K = 5$. The fraction of squared covariance explained beyond the fifth mode is small (generally $< 10\%$) depending on the predictor and particular training window. The time period 1995-2012 is chosen to span both low-trend and trend dominated subperiods in order to assess the performance of different predictors in these distinct periods.

To distinguish between robust predictive skill and apparent skill due to overfitting, an ensemble of Monte Carlo tests is performed. Each ensemble member is constructed by randomly and independently shuffling the time series of the predictor and the predictand over the training period prior to building the statistical model, while continuing to use the real-time predictor, $\mathbf{r}_m(t_e)$ to estimate $\tilde{\mathbf{d}}_m(t_e)$. For each combination of predictor and number of modes retained (CPM), 100 such estimates of SIT are made. These are compared against estimates using unshuffled predictor-predictand time series.

2.4 Results

2.4.1 Evaluation of Predictors

As a first metric for predictor selection, we use the *areal- and temporal-mean absolute errors* (ATMAEs) for each CPM (shuffled and unshuffled), displayed in Fig. 2.1. To assess the overall level and statistical significance of predictive skill for each CPM, we consider: the width of the distribution of shuffled ATMAEs, the value of the unshuffled ATMAE, and the ranking of the unshuffled ATMAE relative to the shuffled distribution. The width of each shuffled distribution indicates the range of apparent predictive skill in a given CPM that is attributable to random sampling. While the value of the unshuffled ATMAE describes the skill of that CPM, this must be considered in the context of the distribution of shuffled estimates' skills. For example, if an unshuffled ATMAE is smaller than all but 5% of the shuffled values, that CPM has statistically significant skill with 95% confidence based on a one-tailed test.

From Fig. 2.1, we see that the width of the distribution for each CPM decreases as the number of modes retained increases. In addition, the distributions for the first two modes are wider in September than in March, reflecting the larger sea ice variance in September. When the distribution is relatively narrow, the predictors tend to perform similarly with respect to both their shuffled and unshuffled ATMAEs, indicating an increased degree of overfitting. With respect to different CPMs, we see that any estimate of SIT that uses only the first mode of SIC - either alone or in combination with SLP or SLPlag - has comparable or greater skill and statistical significance than SIC CPMs using a larger number of modes. This is mainly due to the large decreasing trend toward lower SIT over 2004-2012 (Fig. A1c,d), which is shared by the first mode of SIC. Comparing SLP and SLPlag over 1995-2012, SLPlag is found to have superior skill and statistical significance based on optimal CPMs. In March, this is achieved by retaining the first three modes of SLPlag, whereas in September, the first mode of SLPlag is optimal. Combined with SIC, SLP and SLPlag perform quite similarly in both March and September over the whole period. However, during the earlier interval (1995-2003), the first mode of SIC&SLPlag is more skillful with a higher degree of statistical significance in September (Fig. A1b), whereas the first mode of SIC&SLP is more significant in March (although not as significant as SLP alone) (Fig. A1a).

The spatial distribution of the statistical significance of the CPMs unshuffled

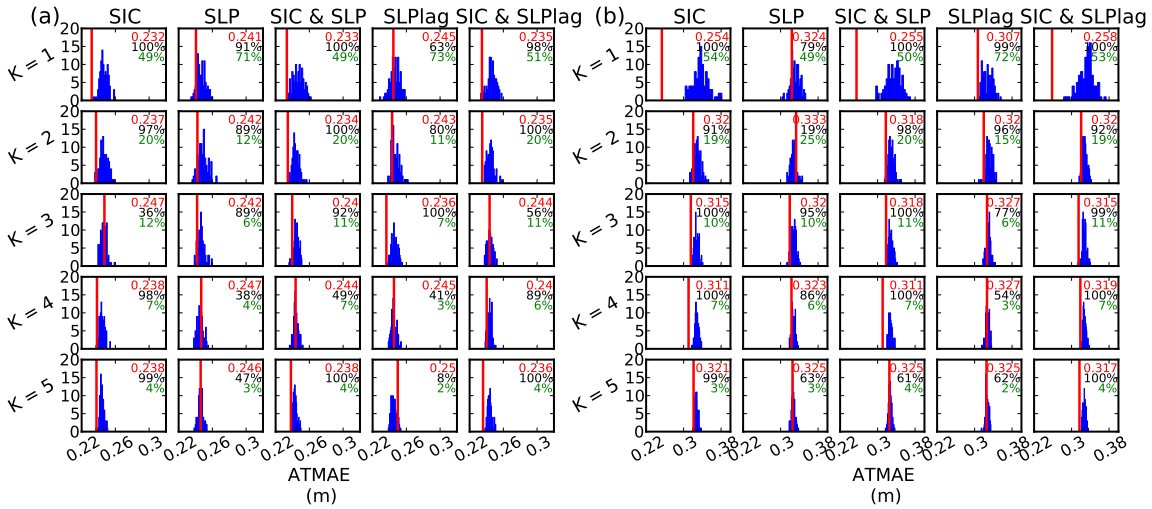

Figure 2.1: Summary of Monte Carlo tests, showing the distribution of the 100 ATMAE estimates based on shuffled predictor-predictand pairs (blue histograms with 20 bins), and the unshuffled estimate ATMAE (red line) for each predictor and K number of modes retained (as labelled). These are given as averages over the 1995-2012 testing period for the months of March (a) and September (b). Unshuffled ATMAEs and the percentage of times the unshuffled error outperforms the shuffled errors are indicated in red and black text, respectively. A percentage $\geq 95\%$ is taken to be statistically significant ($p < 0.05$). The temporally averaged fraction of squared covariance explained by each unshuffled CPM is given as a percentage in green.

estimates is illustrated in Fig. 2.2, with the number of modes retained for each predictor identified using the results from Fig. 2.1. The colour scale represents the probability that the unshuffled estimate will have a lower *temporal mean absolute error* (TMAE) than a randomly drawn shuffled estimate at that location. Note that these estimates of statistical significance were not carried out with separate shufflings at each location, but are based on the pointwise fraction of the full shuffled fields that have a higher TMAE (poorer skill) than the unshuffled field. Similar to the ATMAE results (Fig. 2.1), in September the statistical significance of an estimate of SIT involving only the first mode of SIC (alone or in combination with SLP or SLPlag) shows a high degree of confidence over a large majority of the domain, with the largest area of significance ($p < 0.05$) shown by SIC&SLPlag. Consistent with the areal-mean results shown in Fig. 2.1, SLPlag outperforms SLP in both March and September. Because the skill associated with capturing the trend is less important in March, we see a wider range in the spatial distribution of significance between predictors in this month; yet, the first mode of SIC continues to perform best. The

significance level of all predictors involving SIC is consistently lowest in the East Siberian Sea extending into the Arctic Basin.

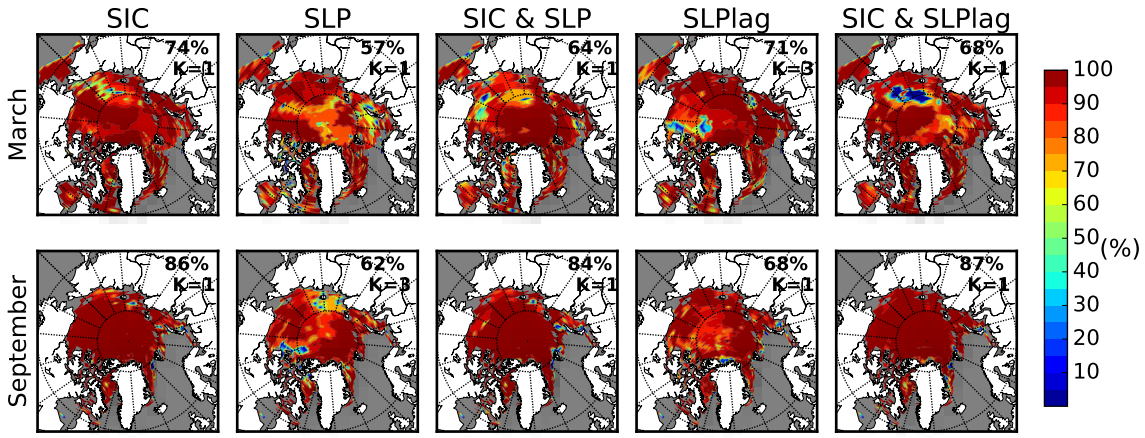

Figure 2.2: The spatial distribution of statistical significance, represented by the probability for an unshuffled estimate's TMAE to be smaller (have greater skill) than a randomly drawn shuffled estimate. This is shown for each predictor (as labelled) for the months of March (top row) and September (bottom row), with the K number of modes retained and the spatial extent (given as a percent) of values significant with $p < 0.05$, indicated by hatched areas on each map.

From these results, it is apparent that over the entire period (1995-2012) in both March and September, the first mode of SIC is the best and most robust predictor (Fig. 2.1). However, within the first or second of the two sub-periods considered separately (Fig. A1), a choice of an optimal predictor (specifically over 1995-2003) is ambiguous. During the earlier subinterval (1995-2003) in March, when interannual variability dominates the long-term trend, sea level pressure (lagged or simultaneous and alone or in combination with SIC) shows greater skill than SIC by itself, whereas in September SIC&SLPlag is optimal. Consideration of the spatial distribution of predictive skill statistical significance (Fig. 2.2) suggests a slight preference for SLPlag over the first period. In the second subinterval (2004-2012), the long-term trend toward lower SIT dominates interannual variability and the first mode of SIC is the optimal predictor.

2.4.2 Real-time SIT Statistical Model

The fact that different predictors are optimal in the two subperiods, dominated by different types of variability, reflects the pronounced non-stationarity of sea ice statis-

tics over the period considered. To account for this, we blend SIT estimates made using the first modes of SIC and SLPlag by means of a weighted average. The relative weighting of these two fields depends on the relative strength of interannual variability and the trend in the years preceding the estimate time. In particular, we use sea ice volume (SIV) over the training period τ, to weight the importance of SIC and SLPlag by separating its temporal variance into trend (T) and interannual (I) components, $\sigma^2 = \sigma_{\mathrm{T}}^2 + \sigma_{\mathrm{I}}^2$. The trend is found by a linear fit and the interannual component is the residual around it; the fact that these time series are uncorrelated allows for the above decomposition of variance. Estimates of $\tilde{\mathbf{d}}_m(t_e)$ made separately with the first mode of SIC and the first mode of SLPlag are then weighted and combined as

$$\tilde{\mathbf{d}}_m(t_e) = \frac{\sigma_{\mathrm{T}}^2}{\sigma^2}\tilde{\mathbf{d}}_m^{\mathrm{SIC}}(t_e) + \frac{\sigma_{\mathrm{I}}^2}{\sigma^2}\tilde{\mathbf{d}}_m^{\mathrm{SLPlag}}(t_e). \tag{2.4}$$

This weighting has the effect of emphasizing the predictive qualities of the first mode of SIC when the trend dominates variability, and the characteristics of the first mode of SLPlag when interannual variability dominates (Fig. A2).

The performance of the blended estimator has been assessed (in terms of the AT-MAE) over the period 1994-2012 and is found to out-perform several reference predictions (RPs) including climatology (temporal mean SIT over τ), persistence (previous year's SIT), a linear trend calculated over τ and extrapolated to t_e, and a combination of persistence and linear trend extrapolation (see Appendix A). However, over 1981-1993 (before a full 15-year construction period is possible), climatology (based on the years 1979 through $t_e - 1$) is slightly superior. Over this period, we therefore simply use climatology. Equation (2.4), despite capturing the decreasing trend in SIT with greatest skill, consistently overestimates areal mean thickness during the trend period, with a generally linear increase in errors over time (not shown). We therefore introduce a bias correction, removing the time-mean error over the previous five prediction years from each grid point. The final statistical model (SM) thus consists of using climatology during 1981-1993 and equation (2.4) during 1994-2012 along with the bias correction. Because SIC is known for prediction year t_e, we impose SIT values of zero in grid cells where SIC is also zero. A comparison between the SM and the RPs is summarized in Table A.1 and Fig. A.3 in Appendix A and supports the motivation for the SM.

2.4.3 Statistical Model vs. CanSIPS

We now assess the improvement in SIT estimation using the SM relative to the current CanSIPS initialization method. Figure 2.3 shows the *areal mean absolute error* (AMAE) of the SM and CanSIPS initialization fields as a function of month and year, as well as averages for each year and calendar month (line plots to the right and above respectively). Averaged over the 1981-2012 record, the ATMAE is reduced by 48% using the SM (relative to CanSIPS), with particularly large improvement seen in 2007-2012 from June through December when CanSIPS absolute error exceeds that of the SM by an average of 40 cm.

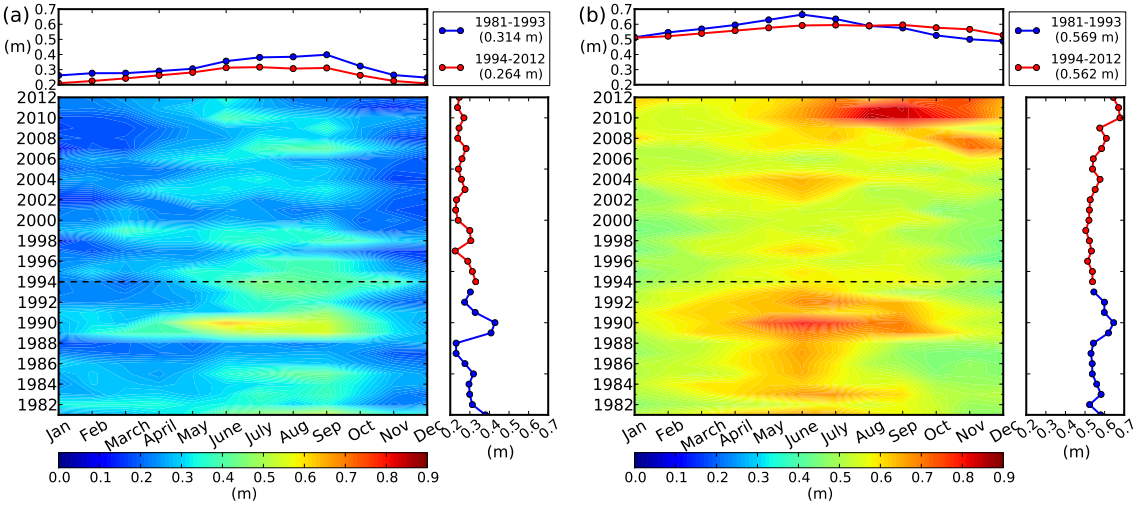

Figure 2.3: Comparison of AMAEs for each month and year during 1981-2012 for (a) the SM, and (b) CanSIPS. The marginal means are shown above (across years) and to the right (across months), in blue for the period when climatology is used as a predictor (1981-1993), and in red when equation (2.4) is used as the predictor (1994-2012). This separation is indicated in the contour plot by the black dashed line. The ATMAE for each period is shown in the upper right box.

The annually averaged AMAE is consistently lower for the SM (Fig. 2.3a, right line plot), averaging 25 cm after the year 2000. Before this, AMAE values vary considerably from 23 cm in 1987-1988, up to nearly 43 cm in 1989-1990 during a time of unusually anomalous SIT distribution (high in the Canadian Archipelago and low in the Chukchi Sea). This episode coincides with an extreme positive AO event during the winters of 1989 and 1990, which helps to advect ice away from the Eurasian and Alaskan coasts (e.g. Rigor and Wallace, 2004). The marginal mean AMAEs across given months (Fig. 2.3a, upper line plot) scale with the standard deviation of spatially

averaged SIT. Normalizing these errors by this standard deviation reveals a relatively weak seasonal cycle, with a maximum in winter and a minimum in summer, with a peak-to-peak difference in the normalized errors of 14% (not shown).

The *root mean squared error* (RMSE) in CanSIPS SIV is 3115 km^3, while for the SM it is 1117 km^3, 64% lower. A comparison between three-month averaged ice volume amounts given by CanSIPS and the SM relative to PIOMAS over the period 1981-2012 are displayed in Fig. 2.4. We see that PIOMAS shows a downward trend toward lower SIV in all months, captured accurately only by the SM. CanSIPS shows a slower rate of decline in the summer and autumn, most pronounced at times of the year when the ice is most rapidly retreating. The trend in winter and spring is missed by CanSIPS entirely. Correlations between time series for individual months show that the SM consistently captures SIV variability (with nearly constant $r = 0.95$), whereas CanSIPS correlations display a strong seasonal cycle with skill similar to that of the SM only in the peak melt months (not shown).

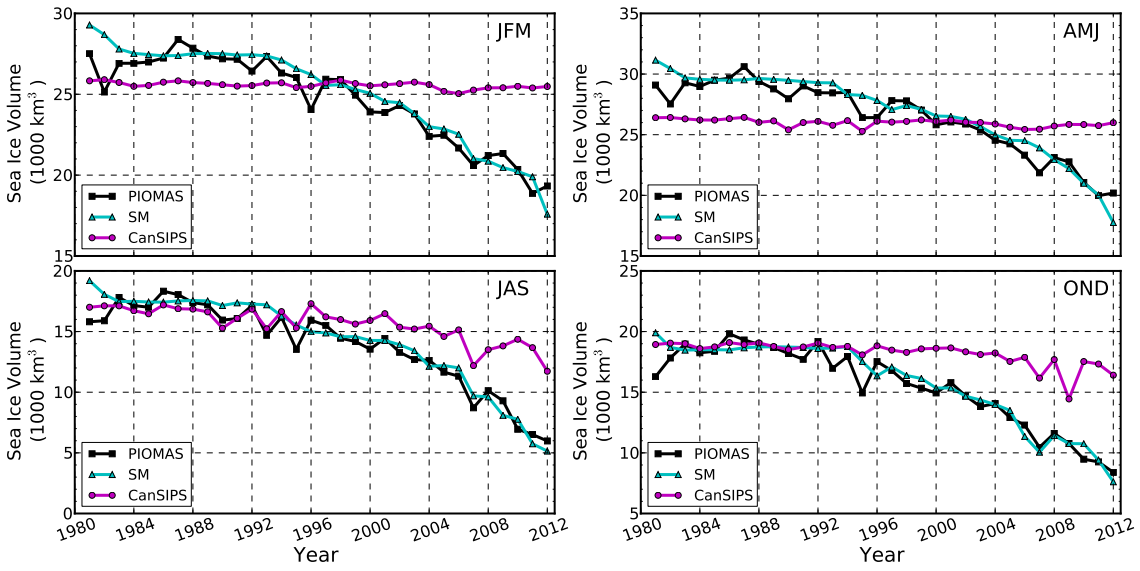

Figure 2.4: Time series of three-month averaged SIV over the period 1981-2012, given in units of 10^3 km^3 for PIOMAS (black), the SM (cyan), and CanSIPS (magenta).

2.5 Conclusion

In this study, we investigate the skill of a MCA-based statistical model to predict Arctic sea ice thickness fields using regularly-observed and physically relevant pre-

dictors. The skill of the MCA prediction model was found to be highly dependent on both predictor choice and the number of modes retained. Based on Monte Carlo tests, an optimal prediction model was chosen which blends skill from SLPlag in the low-trend period with SIC in the trend-dominated period.

The SM greatly improves upon the current scheme used to initialize SIT fields in CanSIPS. This is evidenced by a reduction in ATMAE (relative to CanSIPS fields) over the period 1981-2012 of 48%, and a reduction in ice volume RMSE of 64%. The SM most notably improves SIT estimates over the trend-dominated part of the record, with greatest skill during the summer and autumn (in terms of standardized ATMAEs). While the SM gains much of its skill from capturing the trend, future work could focus on improving its prediction of interannual variability.

The improvement in constructing SIT initial conditions has the potential to increase predictive skill for Arctic sea ice in an operational seasonal forecasting system. This improvement could potentially translate to improved skill predicting other components in the climate system, both within and beyond the Arctic. The SM is being tested in CanSIPS, and impacts on hindcast skill will be reported in a future study. This relatively simple and inexpensive initialization method could potentially be implemented in other seasonal prediction systems.

Chapter 3

Impacts of sea ice thickness initialization on seasonal Arctic sea ice predictions

The following chapter is a manuscript published as:

Dirkson, A., Merryfield, W.J. and Monahan, A., 2017. Impacts of Sea Ice Thickness Initialization on Seasonal Arctic Sea Ice Predictions. *Journal of Climate*, 30(3), pp.1001-1017.

The manuscript is repeated here with small modifications to fit the format of this dissertation.

3.1 Abstract

A promising means for increasing skill of seasonal predictions of Arctic sea ice is improving sea ice thickness (SIT) initial conditions; however, sparse SIT observations limit this potential. Using the Canadian Climate Model, version 3 (CanCM3), three statistical models designed to estimate SIT fields for initialization in a real time forecasting system are applied to initialize sea ice hindcasts over 1981-2012. Hindcast skill is assessed relative to two benchmark SIT initialization methods (SIT-IMs): a climatological initialization currently used operationally and SIT values from the Pan-Arctic Ice Ocean Modeling and Assimilation System (PIOMAS). Based on several

measures of skill, sea ice predictions are generally improved relative to a climatological initialization. The accuracy with which the initialization fields represent both the thinning of the ice pack over time and interannual variability impacts predictive skill for pan-Arctic sea ice area (SIA) and regional sea ice concentration (SIC), with the most robust improvements obtained with SIT-IMs that adequately represent both processes. Similar skill to that achieved by initializing with PIOMAS, including skilful predictions of detrended September SIA from May, is obtained by initializing with two of the statistical models. Regional skill for September SIC is also enhanced using improved SIT-IMs, with an increase in the spatial coverage of statistically significant skill from 10% to 60-70% of the appreciably varying ice pack. Reduced skill is seen however in the Nordic Seas using the improved SIT-IMs, resulting from an inherent cold sea surface temperature bias in CanCM3 that is amplified by a thicker initial ice cover.

3.2 Introduction

Seasonal forecasting of Arctic sea ice has received increased attention in recent years due to a growing demand for forecasts from an array of stakeholders. This demand has grown largely as a result of the increased access to Arctic waterways (Ellis and Brigham, 2009), owing to the reduction in sea ice coverage which is most prominent in the summer months (Serreze et al., 2007). The overall negative trend in pan-Arctic sea ice extent (SIE) is consistent with climate projections that show these reductions continuing into the future under increased greenhouse gas emission scenarios (Stroeve et al., 2012a). Methodologies that have been used for seasonal sea ice forecasts include statistical regression-based methods, fully coupled atmosphere-ocean global climate models (AOGCMs), as well as heuristic approaches (Stroeve et al., 2014b; Guemas et al., 2016). However, few centers currently produce these forecasts operationally.

Arctic sea ice has been shown to be predictable on seasonal to interannual time scales using AOGCMs in both "perfect model" experiments and initialized hindcasts. By construction, perfect model experiments are unaffected by either systematic model errors or by imperfect initial conditions. However, the inherent predictability in the model itself may be biased. By contrast, initialized hindcasts provide an estimate of practical forecast skill subject to the availability of observations, observational uncertainties, and model biases.

Using the perfect model approach, studies indicate that pan-Arctic sea ice area

(SIA) and SIE are inherently predictable for up to 1-2 years, with seasonal reemergence of skill occurring out to 4 years (Blanchard-Wrigglesworth et al., 2011a; Day et al., 2014; Tietsche et al., 2014). Through sensitivity studies, it has been shown that the predictability of September SIE and sea ice concentration (SIC) – the fractional ice coverage in a local grid cell – up to 8 months in advance relies on an accurate representation of sea ice thickness (SIT) initial conditions (Day et al., 2014). For forecasts initialized in winter, pre-conditioning of SIT anomalies contributes to predictive skill for ice area in the summer months, particularly from the memory of ice at least 1.5 m thick (Chevallier and Salas-Mélia, 2012).

Initialized hindcasts (Chevallier et al., 2013; Sigmond et al., 2013; Merryfield et al., 2013a; Wang et al., 2013; Msadek et al., 2014; Collow et al., 2015) suggest that September SIE and SIA anomalies are predictable for long lead times (up to a year). A large component of this skill is attributable to the large negative trend in sea ice coverage. Variations around the trend have been found to be much less predictable, limited to maximum lead times between 2-6 months for September SIE or SIA. Sigmond et al. (2013) found that the Canadian Seasonal to Interannual Prediction System (CanSIPS), employing a crude SIT initialization (Merryfield et al., 2013b), produced hindcasts of detrended September SIA anomalies that are statistically skilful only when initialized in June or later. With more realistic SIT initialization procedures and using other models, statistically significant hindcast skill of detrended SIE/SIA in September has been obtained from as early as May (Chevallier et al., 2013; Wang et al., 2013) or even March (Msadek et al., 2014; Collow et al., 2015).

Most sea ice predictability studies have focused primarily on integral measures such as SIE or SIA; the predictability of regional sea ice coverage in initial condition-based hindcasts has received less attention. One such analysis by Collow et al. (2015) showed that skill in predicting regional September SIC is sensitive to the SIT initialization used, and also to model physics changes which reduce model biases in sea surface temperatures (SSTs).

Both perfect model and initial condition based hindcasts suggest that forecast skill depends strongly on the SIT initialization used (Day et al., 2014; Tietsche et al., 2014; Collow et al., 2015). However, the ability to initialize SIT both in hindcasts and in real time is hampered by the limited observational record of SIT and further complicated by inconsistent SIT observing systems (Lindsay and Schweiger, 2015). These difficulties pose a challenge to initializing SIT accurately in real time in a

manner that is consistent with the 20-30 years of hindcasts that enable real-time bias correction and calibration.

The goal of the present study is to evaluate sea ice hindcast skill using a range of SIT initialization methods (SIT-IMs) that include statistical models designed to be applicable in both hindcast and real-time forecast settings. We examine sea ice hindcast skill over a 32-year period spanning 1981-2012. Hindcasts are generated using the Canadian Centre for Climate Modelling and Analysis (CCCma) Canadian Climate Model, version 3 (CanCM3) (Merryfield et al., 2013b). The details of these hindcast simulations are described in section 3.3. A summary of the five SIT-IMs considered is given in section 3.4. Hindcast skill is evaluated in section 3.5, wherein predictive skill for both integrated Arctic SIA and spatially varying SIC is examined. The dependence of SIC skill on differences in hindcast SIT and SST is also considered. Finally, a discussion and conclusions are presented in section 3.6.

3.3 Sea Ice Hindcasts

The Canadian Seasonal to Interannual Prediction System (CanSIPS) produces operational seasonal forecasts based on CanCM3 and the Canadian Climate Model, version 4 (CanCM4) (Merryfield et al., 2013b). CanCM3 uses the third-generation Canadian atmospheric general circulation model (CanAM3), whereas CanCM4 uses the fourth-generation model (CanAM4). Both CanAM3 and CanAM4 have horizontal grid spacings of approximately 2.8°, but differ in their vertical resolutions (31 levels for CanAM3 and 35 levels for CanAM4). CanCM3 and CanCM4 share the same land, ocean, and sea ice components. The ocean model used is the CCCma fourth-generation ocean model (CanOM4). CanOM4 has an approximately 100 km horizontal grid resolution with 40 vertical levels with spacing of 10 m near the surface and increasing with depth. Sea ice is modelled as a cavitating fluid with a one-category ice thickness following Flato and Hibler (1992).

Hindcasts using CanCM3 are considered in this study because of its lower computational expense compared to CanCM4. As multi-model ensembles are generally more skilful than single-model forecasts (e.g. Kharin et al., 2009), our results likely provide a lower-end estimate of Arctic sea ice skill relative to that which could be achieved by the two-model combination employed by CanSIPS.

Merryfield et al. (2013b) show that freely-running historical simulations of CanCM3 yield Arctic sea ice biases relative to the Hadley Center Sea Ice and Sea Surface

Temperatures, version 1.1 (HadISST1.1) observational dataset (Rayner et al., 2003). Specifically, CanCM3 overestimates SIE in all calendar months and incorrectly simulates the minimum seasonal extent to occur in August rather than September. Additionally, CanCM3 shows biases in its climatological SIC distribution. Positive SIC biases of 15-50% occur in September in the Greenland, Barents, Kara, Laptev, and East Siberian Seas, while the western Arctic near the Canadian Archipelago shows negative biases of up to 15%. In March, large positive biases in SIC are seen in the Labrador, Greenland and Barents Seas, and negative biases of 25-50% in the Bering Sea.

3.3.1 Hindcast Configuration

The hindcasts considered in this study extend six target months, initialized on the 1st day of each of March, May, June, and September. The March, May, and June initializations are chosen to include the spring and summer melt months within the forecast range. September initializations are chosen to include the autumn freeze and winter growth seasons. Each hindcast set consists of ten ensemble members which have slightly different initial conditions intended to represent observational uncertainties. As described in Merryfield et al. (2013b), initial conditions are obtained from a set of assimilation runs (one for each ensemble member), in which SSTs, SIC, and atmospheric variables are constrained near observation-based values with relaxation time scales of three days for SST and SIC, and 24 hours for atmospheric fields. SIC is nudged toward HadISST1.1, whereas SIT is nudged toward values predicted by each SIT-IM, also with a relaxation time scale of three days. Daily sea ice fields are obtained by interpolating between monthly-mean fields.

3.3.2 Defining Interannual Variability

Throughout this study, the distinction between sea ice prediction skill associated with the trend and prediction skill associated with interannual variability will be made. In previous studies, the interannual predictive skill for SIA or SIE has been assessed by evaluating the skill of their residual values relative to a best-fit linear trend (Chevallier et al., 2013; Sigmond et al., 2013; Merryfield et al., 2013a; Wang et al., 2013; Collow et al., 2015). To date, no study has assessed the prediction skill for interannual variability of spatially distributed variables like SIC. Here, we offer a cautionary note when defining interannual variability that should be considered for

both spatially-integrated and spatially-varying sea ice quantities.

In quantifying skill for quantities with a non-stationary mean, like most measures of Arctic sea ice (i.e. SIE, SIT, SIC), skill metrics like the anomaly correlation coefficient (ACC) can be substantially affected by the presence of a trend. Thus, it is often useful to separate skill associated with the trend from that associated with interannual variability. By construction, the definition of interannual variability is determined by how one chooses to define the long-term trend. The choice of trend definition should thus be made carefully depending on the variable and time period considered, as a detrending method suited for one sea ice quantity may not be well-suited for another sea ice quantity. In particular, if the trend accelerates with time then a choice other than linear trend removal may be appropriate. Throughout this work, we therefore consider skill metrics after having detrended the time series of interest using both linear and quadratic fits, and assess the sensitivity of skill to the use of each trend fitting method. When differences in skill are evident, an assessment of the time series under consideration is done to inform the choice of the most appropriate fit, bearing in mind the potential for overfitting when applying multi-parameter fits to short records.

3.4 SIT Initialization Methods

3.4.1 Original

The SIT-IM currently used in CanSIPS will be referred to as *Original*, and consists of nudging SIT values toward a model-based monthly SIT climatology (Merryfield et al., 2013b), the CCCma Synthetic Sea Ice Thickness Climatology, which was developed for use under the Atmospheric Model Intercomparison Project (AMIP). These climatological ice thicknesses were obtained through a sea ice growth relationship similar to that described by Anderson (1961), using prescribed seasonally varying climatological sea ice concentrations and near-surface temperatures as input. Due to the late 20th century epoch of this input data, the simulated thicknesses are more reflective of conditions before 2000 than of the more recent period which has seen a substantial decline in ice volume. Hence, this method does not account for the negative trend in SIT, nor does it represent SIT interannual variability. Consequently, the use of Original leads to an underestimation of the negative SIA trend in hindcasts, and potentially limits skill in forecasting SIA interannual variability (Sigmond et al.,

2013).

3.4.2 PIOMAS

The Pan-Arctic Ice and Ocean Modelling and Assimilation System (PIOMAS) is a high-resolution (averaging 4/5°) coupled sea ice/ocean model, in which sea ice evolves based on a multicategory ice thickness and enthalpy distribution (TED) model (Zhang and Rothrock, 2003). Several fields are assimilated in PIOMAS through a flow-dependent nudging, including SIC data from the National Snow and Ice Data Center (NSIDC), as well as SSTs and atmospheric variables from the National Centers for Environmental Prediction (NCEP) / National Center for Atmospheric Research (NCAR) reanalysis (Kalnay et al., 1996). We refer the reader to other studies for a detailed assessment of the skill of PIOMAS SIT reconstructions relative to observations (Schweiger et al., 2011; Laxon et al., 2013; Stroeve et al., 2014a) and other reanalyses (Chevallier et al., 2016). In general SIT from PIOMAS compares reasonably well with a range of satellite and in-situ observations.

Monthly SIT fields from PIOMAS are regularly updated online, but are not available in real time. Hindcasts initialized by relaxing SIT to PIOMAS are performed because PIOMAS is used to train the statistical models considered in this study for SIT initialization. Hindcasts initialized with PIOMAS are therefore expected to represent an upper limit for predictability relative to the statistical models developed here.

Comparisons of March and September SIT climatologies – defined over the period 1981-2010 to be consistent with hindcasts – for PIOMAS and Original are presented in Fig. 3.1. Both PIOMAS and Original show larger ice thickness in the western Arctic compared to the eastern Arctic. However, PIOMAS generally has thinner ice in the central Arctic, extending south into the Laptev and Kara Seas, whereas PIOMAS has thicker ice in the western Kara Sea, Barents Sea, Greenland Sea, and along the Greenland, Canadian, and Alaskan coastlines into the Chukchi Sea. These greater ice thickness values for PIOMAS are more widespread in March compared to September. Differences in SIT between PIOMAS and Original during the spring months resemble those in winter (not shown).

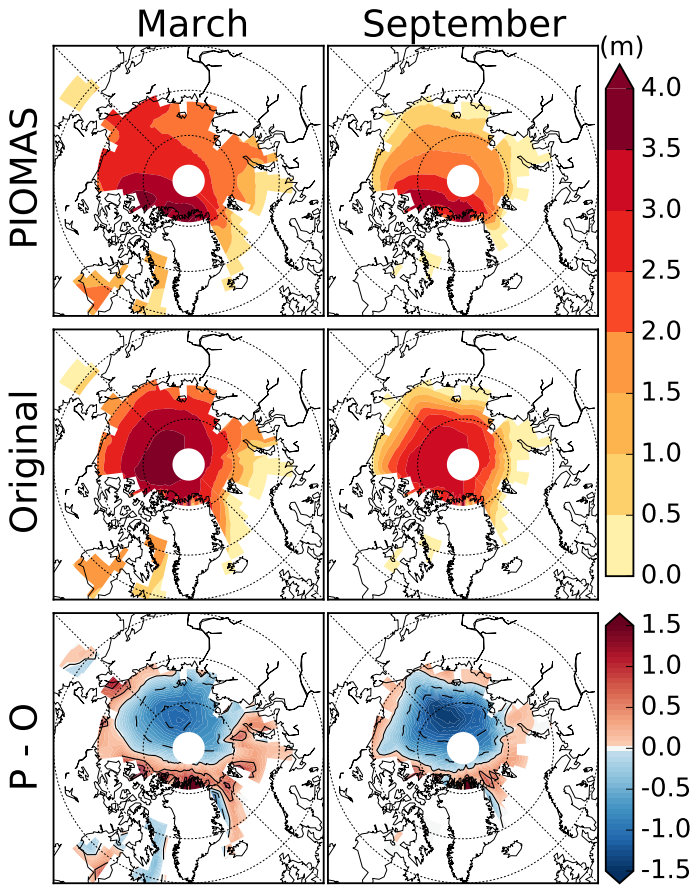

Figure 3.1: Climatological SIT fields (1981-2010) for the months of March and September. Original (top row), PIOMAS (middle row), and their difference (PIOMAS - Original) (bottom row).

3.4.3 Statistical Models

Motivated by the potential for improving upon the Original initialization scheme, three statistical models, denoted SMv1 through SMv3, have been developed to estimate monthly-mean SIT in real time. Although these three models rely on PIOMAS SIC and SIT data to estimate model parameters, the statistical models only require that PIOMAS data be available up to one year prior to the month in which SIT is to be estimated. Because these statistical models do not require PIOMAS SIT fields in real time, they can be applied in an operational setting.

The statistical models make use of (either or both of) two predictor fields: SIC from PIOMAS and sea level pressure (SLP) from the ERA-Interim (Dee et al., 2011)

and ERA-40 (Uppala et al., 2005) reanalyses. The physical basis for these predictors lies in their thermal and dynamical relationships with SIT. SIC is considered because it is correlated with mean ice thickness in most months (Lisæter et al., 2003), as well as locally with SIT in the marginal ice zone (e.g. Tietsche et al., 2013). Furthermore, both SIC and SIT show a strong negative trend over most of the Arctic. SLP is chosen because of the coupling between atmospheric motion and sea ice motion on monthly to multi-monthly time scales (Thorndike and Colony, 1982), potentially influencing SIC through convergent or divergent ice motion (Rigor et al., 2002). Other near-surface atmospheric dynamical predictors such as winds are expected to be well-correlated with SLP on the time scales considered.

For years 1994 onward, the statistical models use a 15-year training period $\tau = \{t_{e-15}, t_{e-14}, ..., t_{e-1}\}$ preceding the target year for initialization t_e. Over τ the predictor(s) and predictand are both known. The predictor and predictand fields for month m and year t are denoted respectively by $\mathbf{x}_m(t)$ and $\mathbf{y}_m(t)$. Statistical model parameters are first estimated over τ using the predictor(s) $\mathbf{x}_m(\tau)$ and predictand $\mathbf{y}_m(\tau)$. The parameters are then applied with the real-time predictor(s) $\mathbf{x}_m(t_e)$, to make an estimate of $\mathbf{y}_m(t_e)$, denoted $\tilde{\mathbf{y}}_m(t_e)$. For the predictand $\mathbf{y}_m(t_e)$, we use PIOMAS SIT, and skill measures for SIT estimated by the statistical models are computed by treating PIOMAS as truth. For years through 1993 (prior to having a 15-year training period), the statistical models use simpler approaches for estimating SIT (to be described) based on training data spanning a shorter period $\tau_s = \{1979, ..., t_{e-1}\}$. The three statistical models are described below and summarized in Table 3.1.

SMv1

The SMv1 statistical model for initializing SIT is described in detail in chapter 2. In brief, SMv1 employs maximum covariance analysis (MCA) over τ to identify patterns of covariability between PIOMAS SIT and each of two predictors: PIOMAS SIC and lagged (4-month averaged) sea level pressure (SLPlag). These predictors, denoted as $\mathbf{x}_m^{\mathrm{SIC}}(\tau)$ and $\mathbf{x}_m^{\mathrm{SLPlag}}(\tau)$, are used separately to construct distinct MCA models. In the formulation of SMv1, separate SIT estimates are made using the leading mode of covariability between SIT and each predictor. This is motivated by the finding in chapter 2 that SLPlag is the more skillful predictor over the first nine years of the verifying period (1995-2003) when the negative trend in SIT is smaller, whereas SIC performs best over the second nine year period (2004-2012) when the negative trend

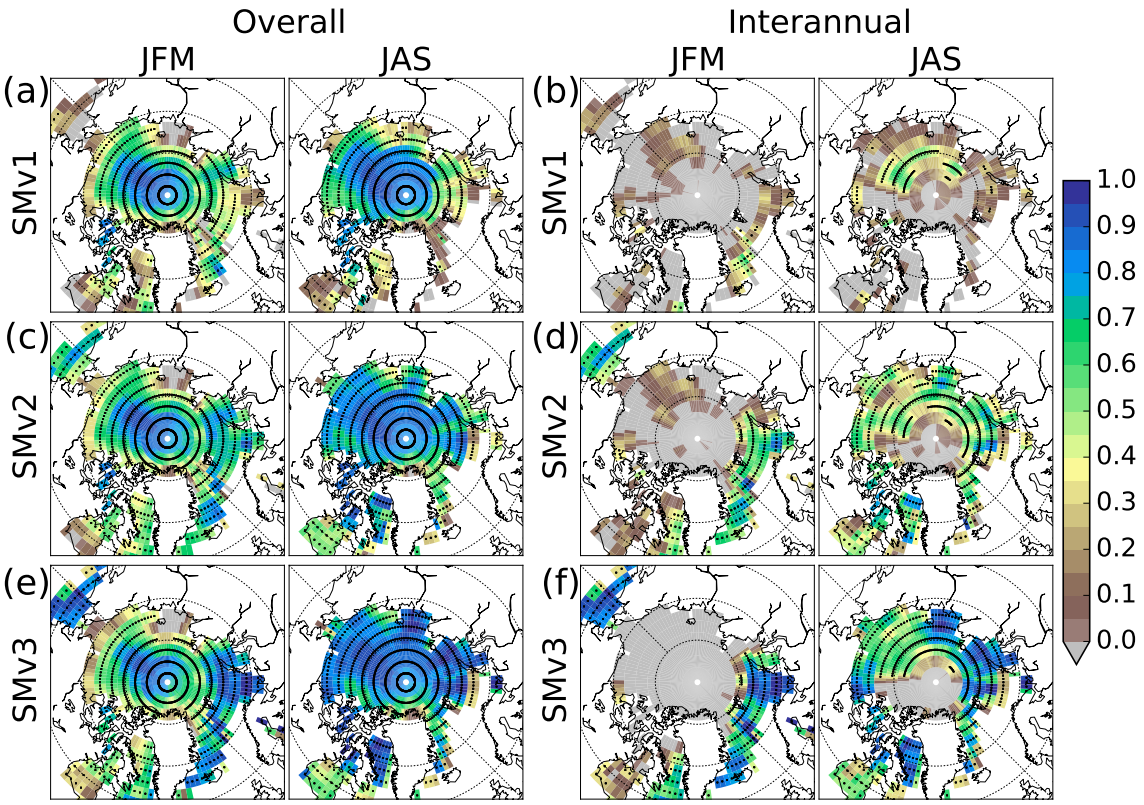

Figure 3.2: ACCs of SIT initialization fields produced by the statistical models assessed relative to PIOMAS: SMv1 (a,b), SMv2 (c,d), and SMv3 (e,f). ACCs are calculated over the period 1994-2012. The ACC skill is considered using the full SIT time series (a,c,e) and using linearly detrended time series (b,d,f). Significant correlations at the 95% confidence level are indicated by stippling.

is larger. It is worthwhile noting the first mode of covariability between SIC and SIT has a strong negative trend circa $t_e = 2000$ onward, and that this mode is uncorrelated with the first mode of covariability between SLPlag and SIT on an interannual basis. These estimates of SIT are combined to produce the final estimates, weighted by the relative importance of interannual variability and the trend,

$$\tilde{\mathbf{y}}_m(t_e) = \frac{\sigma_I^2}{\sigma^2}\tilde{\mathbf{y}}_m^{\mathrm{SLPlag}}(t_e) + \frac{\sigma_T^2}{\sigma^2}\tilde{\mathbf{y}}_m^{\mathrm{SIC}}(t_e), \tag{3.1}$$

where σ^2 represents the total variance of sea ice volume in the interval τ, and is the sum of the variances associated with interannual variability and the trend, respectively denoted σ_I^2 and σ_T^2. Prior to 1994, SIT is set to its average value for the month considered over the shorter training period τ_s. Finally, from 1994 onward a 5-year

mean bias correction is applied to SIT estimated by Eq. (3.1) to reduce a positive SIT bias.

Over 1994-2012 and across all calendar months, SMv1 reduces the areal and temporal mean absolute error (ATMAE) of estimated SIT by 50% compared to Original (Table 3.2). The primary contributor to this improvement is the more accurate representation of the negative trend in SIT. Although SMv1 skill in modelling SIT interannual variability is improved through the inclusion of $\tilde{\mathbf{y}}_m^{\mathrm{SLPlag}}(t_e)$ in Eq. (3.1), such skill is still limited. We illustrate these differences in predictability using maps of the ACC for SIT for two cases: with and without the long-term trend in SIT included. Here, we define interannual variability through linear detrending, which was chosen after inspecting local SIT time series at several locations and concluding that a non-linear component in the trend is not important. The ACCs are calculated separately for each calendar month over the period 1994-2012 (over which time SIT is not obtained from climatology) and then averaged over two three-month periods: JFM and JAS. Statistical significance at the 95% confidence level is estimated by resampling detrended SIT time series using bootstrapping, with a sample size of 10,000 at each grid point. When the trend is included (Fig. 3.2a), SMv1 correlations are significant over most of the domain in both three-month periods. However, areas of statistically significant skill are localized in the central Arctic and Beaufort Sea in JAS when the trend is removed (Fig. 3.2b).

SMv2

SMv2 improves upon SMv1 through an additional step. After computing SMv1 SIT, the sign of SIC and predicted SIT anomalies relative to climatology over τ are compared at each grid location. In cases where the SIC and SIT anomalies disagree in sign, the SIT anomaly is set to a value proportional to the local SIC anomaly using,

$$\tilde{\mathbf{y}}_m(t_e) = \langle \mathbf{y}_m \rangle_\tau + \alpha[\mathbf{x}_m^{\mathrm{SIC}}(t_e) - \langle \mathbf{x}_m^{\mathrm{SIC}} \rangle_\tau], \tag{3.2}$$

where α is a proportionality constant found through sensitivity testing, and the angled brackets $\langle\rangle_\tau$ denote the mean over the training period. The parameter α has been tested at values in the range [0.5,3] at increments of 0.5 m. The ATMAE for Eq. (3.2) varies only slightly for different α values in this range, with a minimum in the ATMAE occurring at $\alpha = 2$ m. Eq. (3.2) resembles that employed by Tietsche et al. (2013), although their method nudges SIT in an assimilation cycle based on

differences between modelled and observed SIC. Applying Eq. (3.2) for 1981-1993 with $\tau = \tau_s$ instead of the SIT climatology (as in SMv1) further improves skill based on the ATMAE (Table 3.2). As in SMv1, from 1994 onward a 5-year mean bias correction is applied to SIT estimates to reduce a positive SIT bias.

The additional step in SMv2 given by Eq. (3.2) relies on the assumption that SIC and SIT are positively correlated on a year-to-year basis. This assumption is more robust in the marginal ice zone, but is less valid in locations where SIC is consistently near 100%. As measured by the ATMAE, SMv2 outperforms SMv1 (Table 3.2). Furthermore, SMv2 improves over SMv1 in terms of the ACC for SIT. This improvement is seen when the trend is included in the Bering Sea and Sea of Okhotsk in JFM, and in Fram Strait and Davis Strait in JAS (Fig. 3.2c). SMv2 is seen to greatly improve upon SMv1 in terms of its representation of interannual variability (Fig. 3.2d), in that the use of SMv2 results in larger values of skill and a larger spatial extent of significant correlation in both JFM and JAS.

SMv3

The improvement of SMv2 relative to SMv1, particularly with respect to interannual variability, further demonstrates that a large fraction of skill using the MCA-based approach (SMv1) is a result of capturing the negative trend in SIT. To assess the skill of a model which represents this trend simply through extrapolation, we introduce the third statistical model given by:

$$\tilde{\mathbf{y}}_m(t_e) = \hat{\mathbf{y}}_m(t_e) + \alpha[\mathbf{x}_m^{\mathrm{SIC}}(t_e) - \hat{\mathbf{x}}_m^{\mathrm{SIC}}(t_e)]. \tag{3.3}$$

In Eq. (3.3), $\hat{\mathbf{y}}_m(t_e)$ and $\hat{\mathbf{x}}_m^{\mathrm{SIC}}(t_e)$ respectively represent the extrapolation of the local linear SIT and SIC trends calculated over τ, and the quantity $\mathbf{x}_m^{\mathrm{SIC}}(t_e) - \hat{\mathbf{x}}_m^{\mathrm{SIC}}(t_e)$ represents detrended SIC anomalies. Like Eq. (3.2), Eq. (3.3) assumes that SIC and SIT are positively correlated on an interannual basis and is therefore subject to the same limitations stated previously.

While SMv3 shows the same skill values as SMv2 over 1994-2012 based on the ATMAE (Table 3.2), Eq. (3.2) remains the model with the lowest ATMAE over 1981-1993. Skill based on the ACC is generally better for SMv3 in regions where both SMv2 and SMv3 have positive skill (Fig. 3.2e,f). However, detrended SIT fields produced by SMv3 show poorer skill relative to SMv2 in the near-polar region of the ice pack where SIC varies relatively little. In the marginal ice zone, SMv3 generally performs

better than SMv2. The largest improvements relative to SMv2 in the detrended case are in JAS.

Table 3.1: The algorithms for all statistical models used to initialize SIT: SMv1, SMv2, and SMv3.

Period	Step	SMv1	SMv2	SMv3
1981-1993	1.	Set SIT to $\langle \mathbf{y}_m \rangle_{\tau_s}$	Set SIT to Eq. (3.2) with $\tau = \tau_s$	Set SIT to Eq. (3.2) with $\tau = \tau_s$
1994-2012	2.	Set SIT to Eq. (3.1)	Set SIT to Eq. (3.1)	Set SIT to Eq. (3.3)
	3.	Subtract preceding 5-year mean error	Where SIC and SIT anomalies disagree in sign, set SIT to Eq. (3.2)	N/A
	4.	N/A	Subtract preceding 5-year mean error	N/A

Table 3.2: Areal and temporal mean absolute error (ATMAE) across calendar months for two periods: 1981-1993 and 1994-2012. The percentage that the ATMAE improves relative to Original (IRO) is displayed for SMv1, SMv2, and SMv3.

	1981-1993		1994-2012	
	ATMAE (m)	IRO	ATMAE (m)	IRO
Original	0.57	n/a	0.56	n/a
SMv1	0.32	44%	0.28	50%
SMv2	0.29	49%	0.26	54%
SMv3	0.29	49%	0.26	54%

3.5 Hindcast Results

3.5.1 Verification Data

To assess the sea ice hindcasts performed in this study, we use the NSIDC merged SIC dataset (Meier et al., 2014a). This product uses the Climate Data Record (CDR) SIC merging algorithm, which combines estimates of SIC derived from NASA Bootstrap (Comiso, 1986) and NASA Team (Gloersen and Cavalieri, 1986) retrieval algorithms. The CDR algorithm uses the more accurate estimate of the sea-ice edge from NASA Bootstrap based on a 10% SIC coverage threshold. Within this region, the merging algorithm assigns SIC on a grid point by grid point basis, according to the larger value between NASA Bootstrap and NASA Team. This is done because both NASA Bootstrap and NASA Team tend to underestimate SIC, but from different sources of bias (Meier et al., 2014a). Prior to all calculations, this product is interpolated onto the CanCM3 model grid. Afterwards, both the verification product and hindcast SIC values below 10% are set to 0% to be consistent with the original verification dataset.

3.5.2 Sea Ice Area

The dependence of Arctic sea ice prediction skill on the five SIT-IMs is first assessed by considering hindcast SIA. SIA is defined as the area integral of SIC in the NH. SIA anomalies for individual hindcast ensemble members are calculated relative to the 1981-2010 baseline climatology for ensemble mean SIA. The final deterministic SIA anomaly hindcast is then defined as the mean of the SIA anomalies across all ensemble members.

We first consider hindcasts of September SIA anomalies initialized in May. The

hindcast SIA anomalies (both the ensemble mean and ensemble spread, defined as ± 2 standard deviations) are indicated along with observed SIA anomalies in Fig. 3.3. A second order polynomial fit is shown for both the observed and hindcast SIA anomalies to provide a visual comparison of trends. The accelerating decline of SIA motivates the use of a quadratic fit. Compared against the fitting statistics of the widely-used linear fit, the quadratic fit reduces the root mean square error (RMSE) between the fit and observed SIA from 0.49×10^6 km^2 to 0.39×10^6 km^2, and increases r^2 from 0.73 to 0.83. The RMSE between the ensemble mean hindcast anomalies and the observed anomalies in SIA is indicated on each panel, in addition to the ACC with the long-term trend included (r), linearly detrended (r_l), and quadratically detrended (r_q).

The RMSE for September hindcasts of SIA decreases from 0.89×10^6 km^2 using Original, to between 0.55×10^6 km^2 and 0.61×10^6 km^2 when using the improved SIT-IMs. Additionally, skill as measured by the ACC is improved considerably using all SIT-IMs other than Original when the trend is included, with ACC values increasing from 0.38 to 0.87-0.91. When the linear trend is removed, SMv1-SMv3 and PIOMAS hindcasts show similar skill (ACCs between 0.52-0.65); however, skill is sensitive to the detrending method chosen, as can be seen by the further reduction in skill using quadratic detrending, with ACC values ranging from 0.24-0.49.

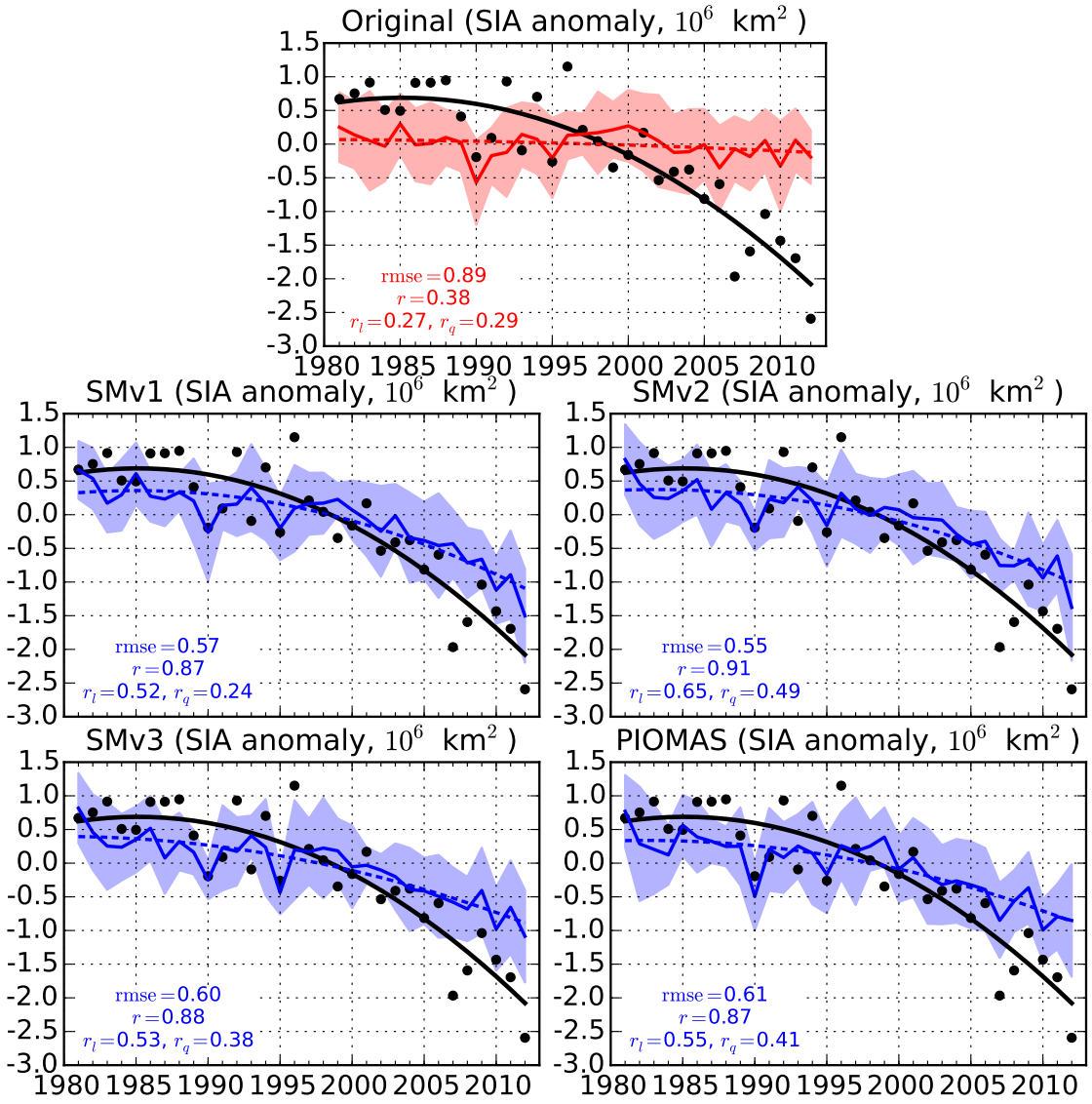

Figure 3.3: Predicted September SIA anomalies over the period 1981-2012 for hindcasts initialized in May. Each panel is for a different SIT initialization method: Original, SMv1, SMv2, SMv3, and PIOMAS. The ensemble spread is indicated by the color-shaded area and the ensemble mean is indicated by the solid color line. Dashed colored lines are second-degree polynomial fits for the ensemble mean SIA anomalies. Observed SIA anomalies are presented as black circles and a second-order polynomial fit for the observed anomalies is indicated by the solid black line. The root mean square error (RMSE) in units of 10^6 km^2 is shown on each panel, in addition to the anomaly correlation coefficient (ACC) for hindcasts which include the trend (r), that have been linearly detrended (r_l), and quadratically detrended (r_q)

Although the negative trend in SIA is represented in hindcasts initialized with SIT-IMs other than Original, the long-term trend is still underestimated relative to the observed trend. Beyond an inherently small SIA/SIE trend in uninitialized historical forecasts using CanCM3 (Merryfield et al., 2013b), one of the possible reasons for the underestimation of the predicted negative trend in summer posited in Sigmond et al. (2013) is the fact that the SIC dataset HadISST1.1 used to initialize hindcasts has smaller trends in SIA compared to the NSIDC product, particularly in winter and spring and including May when these hindcasts are initialized. We test this hypothesis directly with a separate set of hindcasts initialized with SIC from NSIDC and SIT from SMv3. A comparison with hindcasts initialized with HadISST1.1 SIC and SMv3 SIT reveals that although hindcasts initialized in May start out with a more negative linear trend in SIA when initializing with NSIDC SIC (by 0.23×10^6 km^{2}dec^{-1}) over the month of April prior to a May initialization, the difference reduces to only 0.07×10^6 km^{2}dec^{-1} for a lead time of zero months (averaged over May). By September, the initialized SIC field has no statistically significant influence on the trend. Other factors must therefore be inhibiting sea ice decline in CanCM3, and the exact source of bias is beyond the scope of this study.

To assess the skill of each of the 6-month hindcasts initialized in the four months considered, we consider the ACC for SIA over the period 1981-2012 (Fig. 3.4). Skills which are statistically significant at the 95% confidence level are determined by re-sampling detrended SIA time series using bootstrapping with a sample size of 10,000. It should be noted however that statistically significant skill may not translate directly to forecast accuracy (which is a subjective quantity), often considered to require an ACC > 0.6 (Collins et al., 2006). The ACCs shown in Fig. 3.4a are for the full SIA anomaly time series, including the long-term trend. We see that hindcast skills for each target month are statistically significant for nearly all SIT-IMs, excluding August SIA when initialized in March using Original. Hindcast SIA skills obtained using all other SIT-IMs improve substantially over Original. This improvement is especially evident for hindcasts initialized in winter or spring. The improvements in skill for hindcasts initialized with SMv1-SMv3 and PIOMAS result primarily from their improved representation of the SIA trend.

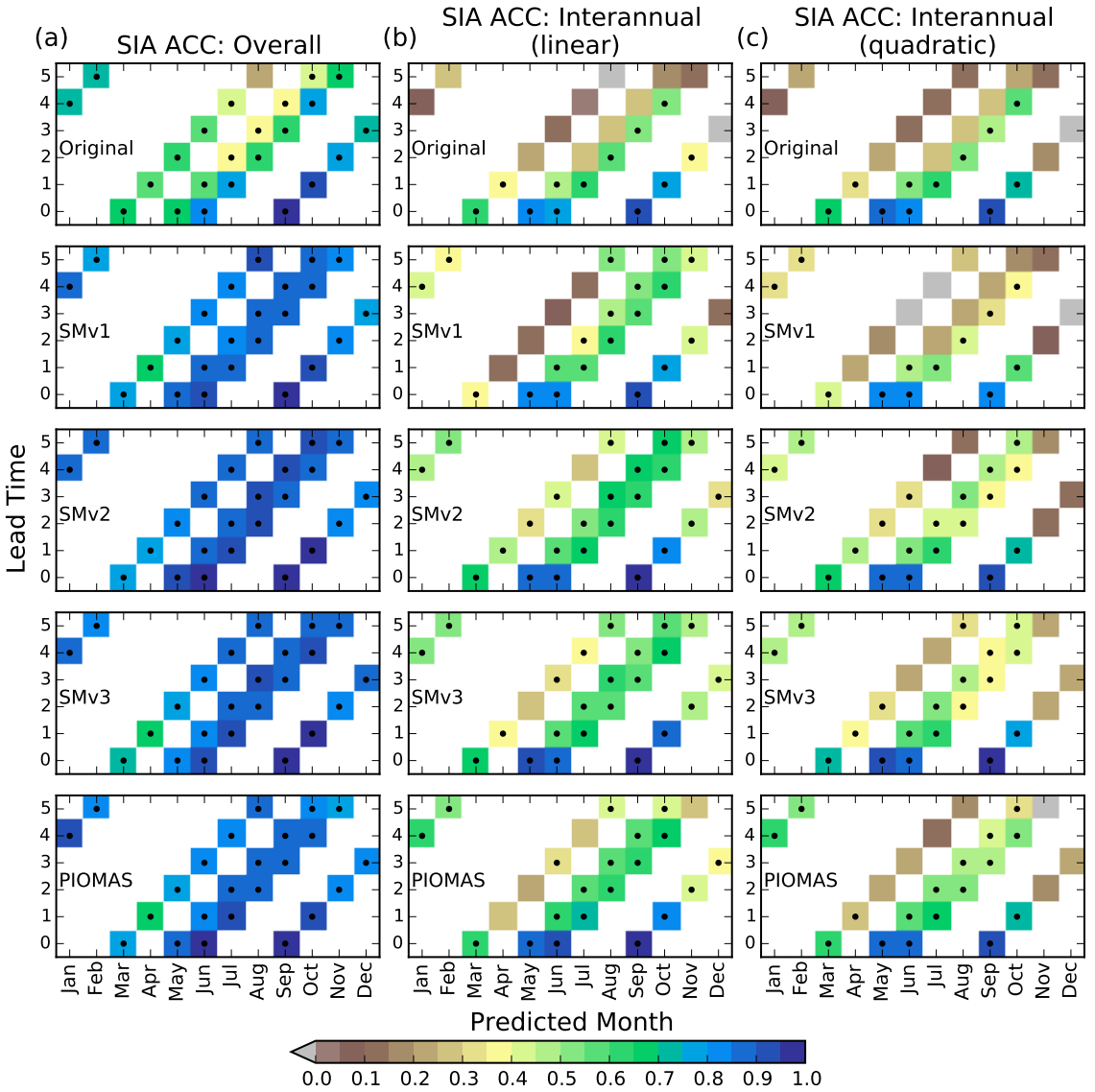

Figure 3.4: ACCs for SIA over the period 1981-2012, shown as a function of target month (horizontal axis) and lead month (vertical axis). The ACCs measure (a) overall skill based on the original SIA time series, (b) interannual skill based on linearly detrended SIA time series, and (c) interannual skill based on quadratically detrended SIA time series. Stippling indicates statistical significance at the 95% confidence level.

The long-term trend is a prominent aspect of the SIA record, and it is important that a forecast system should capture it. However, trend extrapolation could be captured by a simple regression-based statistical model and without the need for a

comprehensive AOGCM. Of arguably greater importance in the assessment of the skill of a seasonal forecast system is quantifying how well interannual variability is captured. To estimate interannual skill for each set of hindcasts, we calculate the ACC using both linearly detrended SIA anomalies (Fig. 3.4b) and quadratically detrended SIA anomalies (Fig. 3.4c). Linear detrending is shown so that these results may be compared against other studies that have also represented interannual skill using linear detrending. However, we argue that quadratic detrending is the more appropriate technique for SIA over the time period considered here because linear detrending leaves a residual long-term quadratic signal in the observed time series, particularly in target months during the late spring and summer when the trend and its curvature are largest, as evidenced by the lower RMSE and higher r^2 for the quadratic fit to September SIA discussed above. Others have used different approaches for detrending SIA/SIE time series such as linear regression onto observed CO2 concentrations (Germe et al., 2014). This approach has also been attempted here, but the results are very similar to those obtained using linear detrending and so are not shown.

The skill of the detrended hindcasts is lower than that of hindcasts including the trend, as has been found in other studies (e.g. Chevallier et al., 2013; Sigmond et al., 2013). Furthermore, skill is generally higher when using linear detrending as compared to quadratic detrending. Hindcasts generated with SMv1 show relatively minor improvements relative to Original with linear detrending, whereas there is a reduction in skill when using quadratic detrending. Hindcasts initialized using SMv2, SMv3, and PIOMAS on the other hand show generally greater skill both with the trend included and with the two types of detrending. These results demonstrate that a better representation of interannual variability in the initialized SIT field improves interannual hindcast skill of SIA.

Focusing on hindcasts initialized in March, detrended SIA skill falls below significant values within 1-2 months after initialization for Original, SMv1, and PIOMAS (Fig. 3.4b,c). Skill using SMv2 and SMv3 is larger than that for SMv1, Original, and PIOMAS. However, the relatively small differences in skill for target months April through August between these SIT-IMs is likely attributable to sampling rather than robust differences in skill, particularly when ACC values lie near the threshold for statistical significance. The reemergence of skill in August when using linear detrending for SMv1-SMv3 and PIOMAS is less obvious with quadratic detrending and is likely associated with residual skill resulting from the incomplete removal of the trend.

For hindcasts initialized in May, a barrier in predictive skill is seen when Original is used to initialize SIT compared to hindcasts initialized in June, as was found in Sigmond et al. (2013) based on both CanSIPS models. Hindcasts initialized with SMv1 also see this barrier in predictive skill when using quadratic detrending. However, skilful hindcasts initialized in May are produced using SMv2, SMv3, and PIOMAS for all target months with the use of either linear or quadratic detrending. The skill of hindcasts initialized in May in these cases is similar to the skill for those initialized in June, when statistically significant skill predicting detrended SIA through October is produced using all SIT-IMs.

Hindcasts initialized in September show significant skill for interannual variability only through October for all SIT-IMs when quadratic detrending is used, whereas hindcasts initialized with SMv2, SMv3 and PIOMAS show significant skill throughout all six target months using linear detrending. Skill decreases with increasing lead time through December regardless of the detrending method used, and increases again in January for all SIT-IMs. This reemergence of skill is similar to that found in previous predictability studies (Holland et al., 2011; Day et al., 2014; Tietsche et al., 2014), in which perfect model experiments showed reemergence of skill peaking in winter. Furthermore, Sigmond et al. (2013) found that significant detrended skill using CanSIPS can be achieved in January-February at a lead time of 11 months. This high level of skill in winter has been attributed to the predictability of the location of the ice edge (Holland et al., 2011), resulting from heat transport variations and/or the persistence of SST anomalies (Bitz et al., 2005; Blanchard-Wrigglesworth et al., 2011a).

3.5.3 Regional skill

Total Arctic SIA is an integrated measure of sea ice cover that is often useful to describe the conditions of the ice pack as a whole and is commonly used when estimating sea ice prediction skill. While useful for comparing skill between models, integrated measures like SIA and SIE provide little information regarding regional sea ice conditions of interest to potential forecast users. The practical utility of sea ice forecasts is increased by considering spatially distributed quantities like regional SIC.

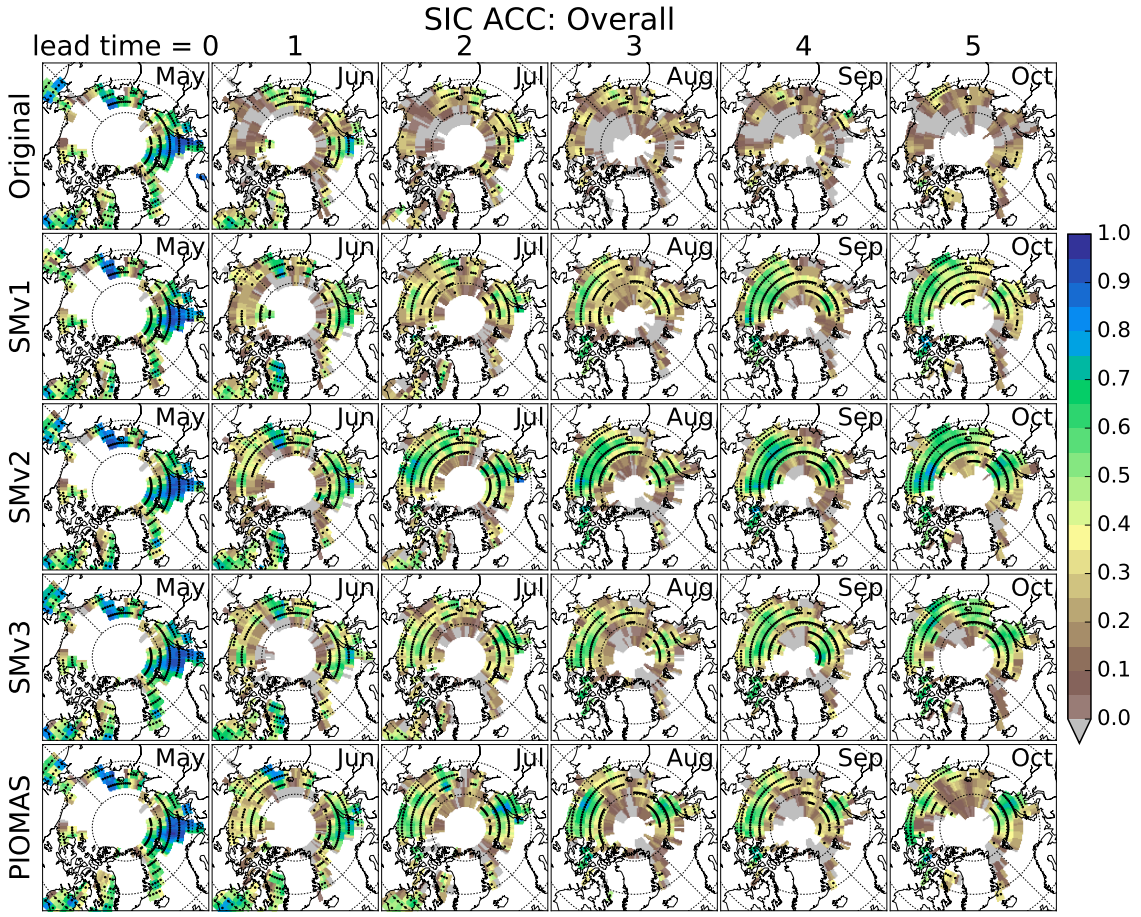

Figure 3.5: Overall skill based on the ACC for SIC over the period 1981-2012 for each SIT-IM. Hindcast skill is shown for the initialization month of May, and the lead time for each target month is indicated above each panel. Areas where the standard deviation for observed SIC is less than 1% are masked to white, and stippling signifies statistical significance with 95% confidence.

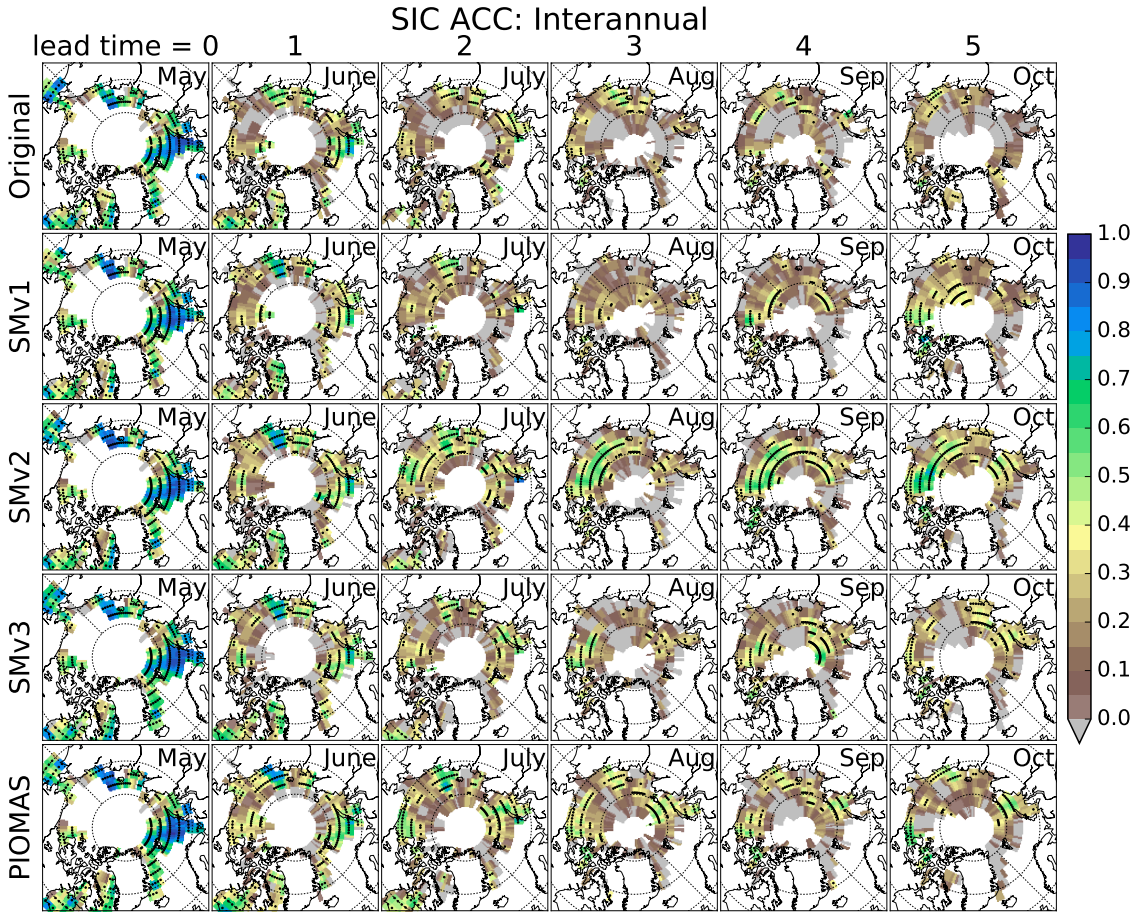

Figure 3.6: As in Fig. 3.5, but for interannual skill based on the ACC of linearly detrended SIC timeseries.

SIC skill: ACC

Maps of ACC for predicted SIC initialized in May are shown in Fig. 3.5 for time series which include the trend (overall skill), and in Fig. 3.6 for time series which have been linearly detrended (interannual skill). Detrending using a second-order polynomial fit was also performed, but differences are generally small between the two trend definitions. In regions where skill differences are larger, examination of the SIC time series reveals that a quadratic fit misrepresents the trend due to overfitting resulting from a relatively small signal to noise ratio compared to SIA. At each model grid point, the ACC 95% confidence level is computed by re-sampling detrended SIC anomalies using bootstrapping with a sample size of 10,000.

For hindcast anomalies which include the trend (Fig. 3.5), we see large improvements in skill in both the western and eastern Arctic when SMv1-SMv3 and PIOMAS are used to initialize ice thickness compared to when Original is used. Skill is greater in the western Arctic in target months June and July for hindcasts with improved thickness initializations compared to Original hindcasts, whereas hindcasts show significant skill through July in a large portion of the eastern Arctic, regardless of the SIT-IM used. In August through October, significant skill is reduced to small localized areas for Original, whereas SMv1-SMv3 and PIOMAS hindcasts show significant skill over a large and coherent portion of both the western and eastern Arctic.

Skill for predicting interannual SIC variability in hindcasts initialized in May is shown in Fig. 3.6. In general, we see lower skill for the improved SIT-IMs than when including the trend, whereas skill in Original hindcasts is similar to when the long-term trend is included. Overall, some significant regional skill is evident for all initialization methods throughout all six target months, and ACC values throughout the Arctic are generally positive independent of the initialization method used. Similar skill is seen for all SIT-IMs in the first two target months (through June). Improved skill relative to Original is seen in the Beaufort Sea in August and September using SMv2, SMv3 and PIOMAS. In October, significant skill is found in the Kara Sea when SIT is initialized with SMv2, SMv3, and PIOMAS. For target months July through October, hindcasts initialized with PIOMAS and SMv2 are more skilful than with other SIT-IMs. However, apparently confounding results such as improved skill using SMv2 relative to PIOMAS in August through October, or significant skill in the Laptev Sea in October only seen in SMv3 hindcasts, suggest that some skill features may be the artifacts of sampling.

As a scalar measure of regional skill for all initialization months and target months, we calculate the fraction of area within the *relevant domain* which contains statistically significant SIC ACC values (Fig. 3.7). The relevant domain is defined as the region where observed SIC standard deviation is at least 1% over the time period considered. This restriction excludes locations that are almost always ice covered or ice free. The advantage over considering skill relative to the total domain is that this fractional quantity is less dependent on the proportional area of the marginal ice zone to the total ice extent, which changes considerably from month to month.

We first consider the areal fraction of significant ACC skill (AFSS) when the trend is included (solid lines in Fig. 3.7). In all initialization months, SMv1-SMv3 and PIOMAS hindcasts show a larger AFSS than Original throughout nearly all target

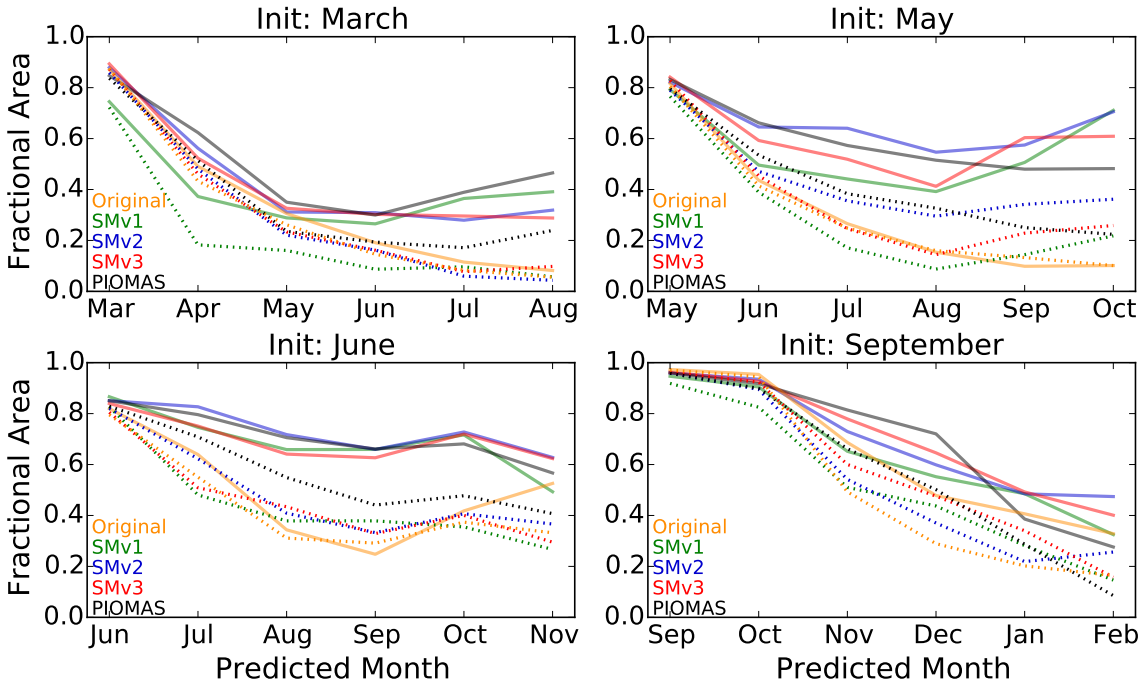

Figure 3.7: The areal fraction of the relevant domain that the ACC for SIC is significant (AFSS in text) at the 95% confidence level. Each panel is for a given initialization month. This metric calculated for ACCs when the trend is included is indicated by solid lines, and for ACCs when local SIC trends are removed using linear detrending by dashed lines.

months. Differences between Original and other SIT-IMs become more pronounced 3 months after initialization for hindcasts initialized in March, and between 1-2 months after initialization for hindcasts initialized in May, June, or September.

When initialized in March, the AFSS decreases relatively rapidly through May for all SIT-IMs. From May through August, skill remains significant across approximately 30-50% of the relevant domain for hindcasts initialized with the improved SIT-IMs, but continues to decrease down to 10% of the domain by August for hindcasts initialized with Original. Hindcasts initialized in May with the improved SIT-IMs show an AFSS averaging near 60% for target months June through October, while Original hindcast skill gradually decreases to near 10% of the relevant domain. For the target month of September, hindcasts initialized with SMv1-SMv3 and PIOMAS have an AFSS 5-6 times greater than that for Original hindcasts. When initialized in June, the AFSS is near or exceeds 60% in all target months using the improved SIT-IMs, whereas it decreases to below 30% in September using Original. When ini-

tialized in September, the AFSS decreases similarly for all SIT-IMs down to 35-40% in February.

The AFSS values after the trend has been removed (indicated by dashed lines in Fig. 3.7) generally vary less between different SIT-IMs than when the AFSS values are calculated with the trend. When initialized in March, the AFSS decreases nearly identically for hindcasts other than those initialized with SMv1 through all target months. Over this time, the AFSS exceeds 20% through May, but decreases below 20% for later target months. For May initialized hindcasts, the largest AFSS differences between different SIT-IMs is seen from June through August, with greatest skill seen using SMv2 and PIOMAS. For September and October target months, the AFSS grows again for SIT-IMs other than Original and covers over 20% of the relevant domain. When initialized in June, all SIT-IMs yield predictions with an AFSS exceeding 30% through all six target months. Skill using PIOMAS is higher than for other SIT-IMs through all target months, with the other SIT-IMs performing similarly to each other. For hindcasts initialized in September, the AFSS remains above 50% through November using all SIT-IMs, at which time a larger separation in skill emerges between SIT-IMs with greatest skill seen using PIOMAS and SMv3. Skills converge again between initialization methods in January and February.

SIC skill: RMSE

Skill measures based on correlation coefficients are insensitive to errors in the magnitude of the hindcast anomalies. To provide a more complete assessment of hindcast performance in predicting SIC, we consider the spatial distribution of RMSE for hindcasts initialized in May. We focus on hindcasts initialized in May to target summer months. Maps of the difference between the RMSE of SIC anomalies for SMv3 and Original hindcasts (SMv3 minus Original) are shown in Fig. 3.8. We focus on SMv3 because differences in this skill metric are small between SMv2, SMv3, and PIOMAS. We do not show SMv1 because of its poor interannual skill in SIA and SIC, as measured by the ACC. RMSE differences are defined such that better skill for SMv3 is indicated by negative values. The RMSE is calculated separately over two 16-year periods from 1981-1996 and from 1997-2012, to highlight differences in skill during periods when the magnitude of the negative trend is small compared to when it is larger.

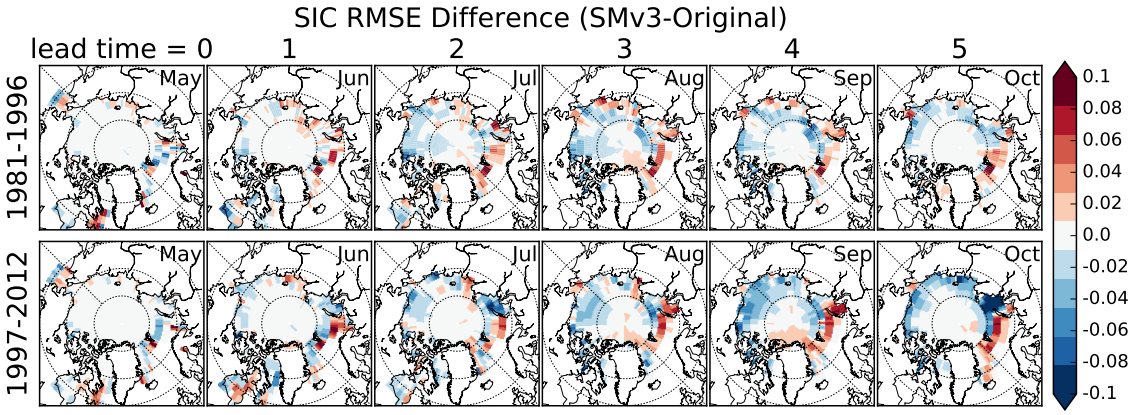

Figure 3.8: Differences in the RMSE of SIC anomalies between SMv3 and Original hindcasts (SMv3 minus Original). Improved skill using SMv3 is represented by negative values (i.e. reduced RMSE). The RMSEs are calculated for May-initialized hindcasts over two periods: 1981-1996 (first row) and 1997-2012 (second row)

Differences in RMSE between SMv3 and Original in May are generally small over both periods. From June through September, we see a broad pattern of higher skill for SMv3 hindcasts in the western Arctic and lower skill in the eastern Arctic. The lower skill in the eastern Arctic in both periods is mainly confined to the northern Greenland and Barents Seas. Higher skill using SMv3 is evident over 1997-2012 throughout much of the Arctic basin from July through October. This improvement is most noticeable from August through September over the Beaufort, northern Chukchi, East Siberian, Laptev, and northern Kara Seas. In September and October, improvements relative to Original span the Alaskan and Russian coasts, and the Canadian Archipelago. The improvement in predicting SIC in the Kara Sea is evident in June-July and again in October when the ice edge advances south into this region. The improvement in fall in the Kara Sea was seen previously using the ACC of detrended SIC in Fig. 3.6, indicating a better representation of interannual refreezing in this region using SMv3.

The next subsection examines reasons for the markedly greater skill of SMv3 hindcasts compared to Original hindcasts in the Kara Sea, as well as the lower skill in the adjacent northern Greenland and Barents Seas, in order to better understand how the SIT initialization together with model biases impacts skill in forecasting SIC.

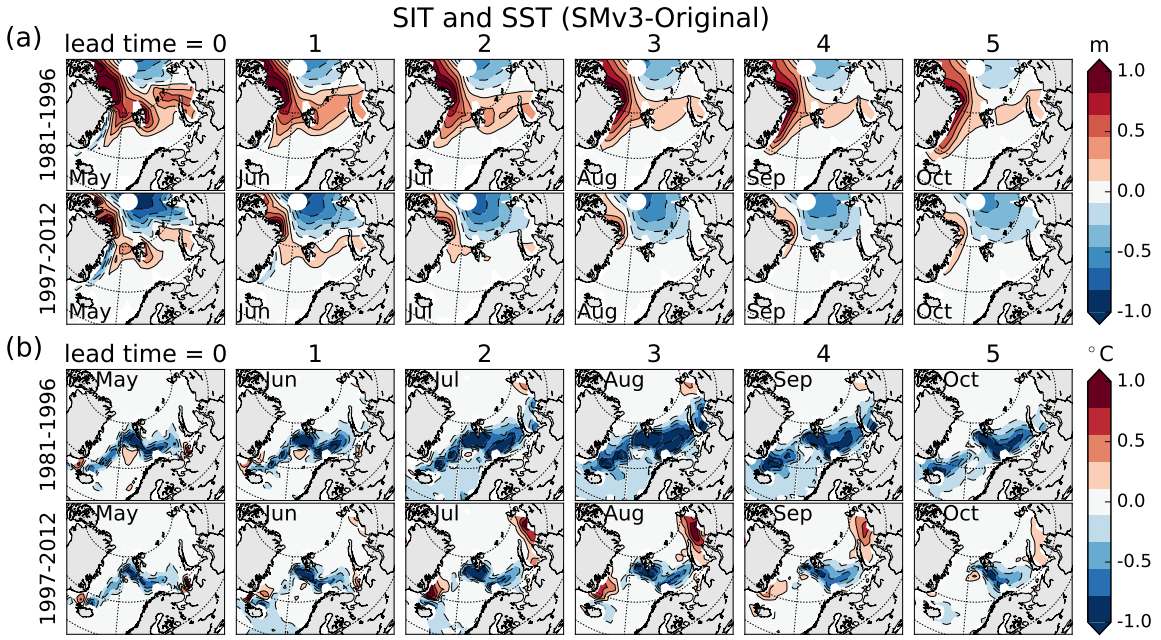

Figure 3.9: Differences in (a) SIT and (b) SST between SMv3 and Original hindcasts (SMv3 minus Original) in a focused region including the Greenland Sea, Barents Sea, and Kara Sea. The differences are calculated over two periods: 1981-1996 (first row) and 1997-2012 (second row). Solid contours are positive differences and dashed contours are negative differences.

SIT influence on hindcast errors

The Nordic Seas region extending from the northern Greenland Sea into the Barents and Kara Sea shows a cold SST bias and positive SIC bias in freely running simulations of CanCM3 (Merryfield et al., 2013b). To understand two features in Fig. 3.8 – the poorer skill in SMv3 hindcasts in the northern Greenland and Barents Seas from June through October, and the improved skill in the Kara Sea in spring and again in fall – we consider mean differences between SMv3 and Original hindcasts of SIT and SST in this region, presented in Fig. 3.9. To compare against the differences in skill using the RMSE, we consider hindcasts initialized in May and mean differences calculated separately over the same two periods, 1981-1996 and 1997-2012. The reader should bear in mind that as lead time increases, differences in SIT and SST fields diminish as the model drifts closer to its own climatology (i.e. as the memory of initial conditions is lost).

We first focus on the region of poorer skill in SMv3 hindcasts throughout northern

Greenland Sea and Barents Sea (Fig. 3.8), and consider the corresponding differences in SIT and SST in Fig. 3.9. We see that in this region, SIT is greater in the SMv3 hindcasts throughout all 6 target months in the first period, and throughout the first 2-3 target months in the second period. Additionally, we see cooler SSTs in this region (to the south of the SIT differences) which are coldest and most expansive in the first period. Because SIT initialization is the only difference in these hindcasts, these cooler SSTs are evidently due to the thicker SITs in this region. The SST bias and corresponding SIC bias in this region in freely running simulations is thus amplified by the thicker SITs at initialization, and therefore degrades skill predicting SIC. Furthermore, the fact that SIC skill in the second period is as poor as in the first period, despite the SIT differences being smaller, implies that the sensitivity of SIC skill to these SIT and SST differences is high.

We now focus on the region of improved skill for SIC RMSE in SMv3 hindcasts in the Kara Sea (Fig. 3.8). Improved skill in June through July and again in October is largest in the second period. During this time, SIT is thinner in the eastern Kara Sea in the first three target months. By July, the ice has retreated north, and warmer SSTs emerge throughout all of the Kara Sea, peak in August, and wane through October. The timing of the improvement in skill implies that the thinner SITs at the time of initialization improve skill in the Kara Sea in June and July (during melting) and that the subsequent warmer SSTs lead to enhanced skill (during freezing) in October.

3.6 Discussion and Conclusions

In this study, we employed hindcasts from CanCM3 to assess the influence of different methods for initializing SIT on the skill of Arctic sea ice area and concentration. The climatological ice thickness initialization employed in CanSIPS, denoted Original, and several more accurate methods were considered, including statistical models SMv1-SMv3 designed to be usable in a real time operational setting as well as PIOMAS. These statistical models differ in their ability to capture interannual variability in the PIOMAS SIT time series used here as a benchmark, but are similar with respect to being able to capture the long-term trend. Using only one predictor field (SIC) and a single model parameter, the statistical model SMv3 performs comparably to the more complicated model SMv2, with both outperforming SMv1.

Evaluating hindcast skill over 1981-2012, improvement in Arctic sea ice prediction

was found using all improved SIT-IMs for nearly all measures of skill. When skill is assessed including the long-term trend, both integrated SIA and regional SIC hindcast skills were shown to strongly rely on the accuracy of the SIT initialization. We found a large sensitivity to interannual SIA skill to the detrending method used, and we argue that detrending using a quadratic fit is more appropriate than using a linear fit over the time period considered. Arctic SIA was found to be more sensitive to the ice thickness initialization than regional SIC when considering the skill of interannual SIA variations around the trend based on linear detrending. However, assessing interannual skill of SIA based on quadratic detrending produced skill for SIA that is more qualitatively similar to regional SIC; i.e. greater interannual skill in SIA generally corresponded to greater interannual skill in the areal fraction of significant skill in SIC. This result provides further support for the choice of assessing interannual skill for SIA based on quadratic detrending.

Hindcast SIA anomalies including the trend were found to have statistically significant skill out to 6 months using all improved SIT-IMs, with largest improvements relative to Original seen in the summer months when initialized in winter and early spring. Using SMv2, SMv3, or PIOMAS, significant skill of quadratically detrended September SIA anomalies can be achieved initializing in May, extending the time that significant skill can be obtained by at least one month relative to using Original. Additionally, unlike hindcasts using Original, linearly detrended August SIA anomalies were found to be predictable with significant skill from March using all improved SIT-IMs. Although this latter result is likely a consequence of the incomplete removal of the trend, it is relevant for comparison against other studies that have used linear detrending to show that statistically significant predictions of September SIA/SIE can be made from March. With these improvements in skill during the summer months, CanCM3 skill for detrended SIA anomalies becomes more comparable to that found in other studies (Chevallier et al., 2013; Wang et al., 2013; Msadek et al., 2014; Collow et al., 2015). Although a direct comparison to these studies is not possible due to differing temporal coverage, the ACC values 0.52-0.65 obtained here for linearly detrended September SIA anomalies when initializing in May are comparable to those ranging from 0.4-0.6 found in Chevallier et al. (2013), Wang et al. (2013), and Msadek et al. (2014) for the same initialization and target month. Skill reemergence for detrended SIA anomalies was found for winter target months for hindcasts initialized in late summer, independent of the SIT-IM used (as in Day et al., 2014) as well as the detrending technique used.

ACC-based regional skill of SIC (with the trend included) was found to be substantially improved using the more accurate SIT-IMs. Large improvements across both the eastern and western Arctic in the summer months result from initializing with the improved SIT-IMs compared to Original. Although the predictive skill of SIC for hindcasts initialized by the improved SIT-IMs is reduced substantially when the trend is removed, hindcasts of September SIC are more skilful than Original using SMv2, SMv3, and PIOMAS, specifically in the Beaufort Sea (where some skill is also present using Original) and eastern central Arctic extending into the Kara Sea. In October, the region of improved skill grows to include all of the Kara Sea.

Given these results, it is clear that similar hindcast skill to that which is obtained initializing with PIOMAS can also be achieved initializing with the statistical models SMv2 or SMv3. Although the SIC ACC-based skill results for May initialization show that SMv2 hindcasts outperform SMv3 hindcasts, these results should not be overinterpreted since skill for SMv2 initialized hindcasts is also greater than PIOMAS initialized hindcasts. Because the statistical models are trained with PIOMAS as the predictand, any improvement in hindcast skill relative to PIOMAS hindcasts seems likely to be the result of sampling. Additionally, SMv3 is substantially simpler than SMv2. Not only does this simplicity hold practical value in terms of parsimony, the simpler statistical model reduces the chance of gaining skill from overfitting. Further, because the only information that SMv3 requires from PIOMAS is the estimate of the local SIT trend over the training period, if PIOMAS were to become unavailable at some point in the future, SMv3 could potentially be constructed with another reanalysis product that adequately represents these trends. For operational purposes, we therefore recommend the use of SMv3 for SIT initialization.

Considering the RMSE as an error metric of SIC brought attention to an unexpected result of using a more accurate SIT initialization within CanCM3. In the region extending from the northern Greenland Sea into the Barents Sea, negative SST biases in CanCM3 were found to be amplified by the thicker initial ice cover specified by the improved SIT-IMs in this region. It should be noted that CanCM4 shows a SIC bias of the opposite sign in this region in freely running September hindcasts (Merryfield et al., 2013b), suggesting that the use of the multi-model approach used in CanSIPS could mitigate this source of error. Finally, as more-accurate initial conditions should always be preferred, this result further motivates efforts to reduce such model biases. As another example, Collow et al. (2015) found that considerable improvement in seasonal sea ice predictions could be achieved not only by improving

SIT initial conditions, but also by reducing a SST bias within the Climate Forecast System, version 2 (CFSv2).

Limitations on observational knowledge of sea ice thickness (both in real time and historically) pose a challenge to operational forecasting centres that wish to produce seasonal forecasts of sea ice. However, as demand is growing for such forecasts, efforts are needed to circumvent this limitation. By applying relatively simple statistical model techniques, SIT initialization can be improved substantially compared to a climatological initialization. Robust improvements in predictive skills of SIA and regional SIC result from using improved ice thickness initializations in CanCM3.

Chapter 4

Calibrated Probabilistic Forecasts of Arctic Sea Ice Concentration

4.1 Abstract

Seasonal forecasts of Arctic sea ice using dynamical models are intrinsically uncertain, and so are best communicated in terms of probabilities. Here, we describe novel statistical post-processing methodologies intended to improve ensemble-based probabilistic forecast of local sea ice concentration (SIC). The first of these improvements is the application of the parametric zero- and one- inflated beta (BEINF) probability distribution, suitable for doubly-bounded variables such as SIC, for obtaining a smoothed forecast probability density function. The second improvement is the introduction of a novel calibration technique, called trend-adjusted quantile mapping (TAQM), that explicitly takes into account SIC trends and is applied to the BEINF distribution. We demonstrate these methods by applying them to a set of 10-member ensemble SIC hindcasts generated with the Canadian Climate Model, version 3 (CanCM3), over the period 1981-2012. Though fitting ensemble SIC hindcasts to the BEINF distribution consistently improves probabilistic hindcast skill relative to a simpler "count-based" probability approach in perfect model experiments, model biases have the potential to offset this improvement when verifying against observations. The TAQM calibration technique is effective at correcting for the large SIC biases present in CanCM3. Furthermore, the TAQM-calibrated SIC hindcasts show high levels of skill in certain regions relative to a standard 1981-2010 climatological reference forecast, particularly in the summer. Skill is present in those same general areas when compared against

a more conservative trend-adjusted climatological reference forecast, but skill scores are typically smaller.

4.2 Introduction

Changes in Arctic sea ice conditions observed over the past four decades are widely documented. Substantial reductions in total Arctic sea ice extent (SIE) of $-13.3 \pm 2.6\%$ per decade in September (National Snow and Ice Data Center, Sea Ice Index Version 2), an overall thinning (Kwok and Rothrock, 2009; Rothrock et al., 1999) and youthening (e.g. Maslanik et al., 2011) of the ice pack, and coincident openings of pan-Arctic marine routes in certain summers (Melia et al., 2016), have lead to a surge of interest in Arctic sea ice forecasts on seasonal time scales. Such forecasts are anticipated to benefit those parties involved in Arctic activities that require long lead time planning and are constrained by sea ice conditions (Ellis and Brigham, 2009).

As with other climate system components, forecasts of Arctic sea ice are inherently uncertain on seasonal time scales, and are therefore best communicated probabilistically. This uncertainty arises, in part, from the chaotic nature of the climate system (particularly in the atmosphere), which causes minute differences in initial conditions to amplify over time (e.g. Reynolds et al., 1994). A simple way of sampling initial condition uncertainty in a seasonal forecast using an atmosphere-ocean global climate model (AOGCM) is to generate an ensemble of deterministic forecasts from slightly different initial conditions. Because these ensembles include a finite, and typically small, number of members, post-processing is needed to infer a continuous forecast distribution (Richardson, 2001). One means of doing this is by fitting a continuous probability distribution to the forecast ensemble (Wilks, 2002).

Uncertainty also arises from model errors which stem from the incomplete representation and numerical approximation of the physical laws that drive climate variability. One effect of these model errors is to degrade forecast reliability (Palmer et al., 2004), so that probabilities forecast for categorical events disagree with their observed frequencies. Further, ensemble forecasts are often under-dispersive (i.e. overconfident), so that the mean-squared-error of the ensemble mean grows faster than the ensemble spread (Gneiting et al., 2005). Model errors and under-dispersion are ongoing challenges in seasonal forecasting, but advances in reducing their effects have been made through calibration (e.g. Gneiting et al., 2005; Wilks, 2011; Kharin et al., 2017), the use of multiple models (e.g. Krishnamurti et al., 1999; Weigel et al., 2008;

Merryfield et al., 2013b), the calibration of multi-model ensembles (e.g. Kharin et al., 2009), as a well as through stochastic parametrization of unresolved processes (e.g. Palmer et al., 2009).

Ensemble forecasts can be used to make probabilistic forecasts of categorical events. One such event related to sea ice coverage, known as *sea ice probability* (SIP), describes the probability that local sea ice concentration (SIC) – the fractional area of a grid cell covered by sea ice – will exceed 15% coverage. The definition of SIP was introduced for the annual Sea Ice Outlook (SIO). Beginning in 2014, the SIO has called for contributions of maps showing the spatial distribution of ensemble mean (local) SIE based on the conventional SIC threshold of 15%. While the 15% SIC threshold is commonly used to delineate the sea-ice edge from open water when estimating SIC from passive-microwave satellites, other SIC event thresholds may be relevant for different forecast end-users.

This study introduces new methodology for improving seasonal probability forecasts of local SIC from ensemble forecasts generated with an AOGCM. These procedures aim to improve the approximation of the underlying forecast SIC probability distribution, which can then be used to forecast not only SIP, but any function of the SIC distribution.

The first of these improvements is the application of a suitable parametric probability distribution for fitting SIC ensemble forecasts. The second is the introduction of a novel calibration method based on the well-known quantile mapping technique. This calibration method explicitly accounts for the observed trends in SIC, and is specifically designed to be applied to the aforementioned parametric distribution.

In section 4.3, we briefly describe the model and hindcast experiments used to test this methodology, as well as the metrics by which we evaluate probabilistic hindcast skill. Two methods for computing SIC forecast event probabilities are described in section 4.4: a counting (i.e. discrete frequency) approach and a parametric approach. A skill comparison of these two methods being applied to the hindcasts is presented in section 4.5. The calibration technique is introduced in section 4.6, and in section 4.7 we evaluate probabilistic hindcast skill after calibration. Conclusions are presented in section 4.8.

4.3 Data and Skill Scores

4.3.1 Hindcasts

The methods introduced here to make SIC probability forecasts are tested on a set of hindcasts produced with the Third Generation Canadian Centre for Climate Modelling and Analysis (CCCma) Canadian Climate Model (CanCM3) (Merryfield et al., 2013b). The atmosphere in CanCM3 is simulated using the Third Generation Canadian Atmospheric General Circulation Model (CanAM3), which has a horizontal grid spacing of approximately 2.8° and 31 vertical levels. CanCM3 simulates the ocean using the CCCma Fourth Generation Ocean Model (CanOM4) with a 100-km nominal horizontal grid spacing, and 40 vertical levels with a spacing of 10 m near the surface and increasing with depth. Sea ice is modelled as a cavitating fluid with a single layer thickness category (Flato and Hibler, 1992).

The hindcast experiments considered here are initialized on first day of March, May, June, and September, and extend 6 months over a 32-year period from 1981-2012. Each ensemble forecast is generated with 10 ensemble members, initialized from slightly different initial conditions that are obtained from assimilation runs. These assimilation runs nudge atmospheric variables with a 1-day time constant, and sea surface temperatures (SSTs) and SIC with a 3-day time constant, toward observation-based values. SIC is nudged toward the NSIDC merged SIC dataset (Meier et al., 2014c), and mean grid cell sea ice thickness (SIT) values are nudged (also with a 3-day time constant) toward estimates obtained from the 'SMv3' statistical model described in chapter 3. The NSIDC merged SIC dataset described above is used to assess probabilistic forecast skill.

4.3.2 Skill Scores

Two metrics are considered here for assessing the skill of probabilistic hindcasts: the Brier score (BS) (Brier, 1950) and the continuous rank probability score (CRPS) (e.g. Hersbach, 2000). The BS is appropriate for assessing the skill of probabilistic forecasts of a specific event (e.g. SIP), whereas the CRPS is a more comprehensive assessment of probabilistic forecast skill which compares the entire forecast probability distribution against the observed SIC value.

The BS for the forecast event Ω is defined as

$$\mathrm{BS}(\Omega) = \frac{1}{M}\sum_{j=1}^{M}\left[P_{f_j}(\Omega) - P_{o_j}(\Omega)\right]^2, \tag{4.1}$$

where the average is taken over $j = 1, ..., M$ forecasts, and $P_f(\Omega)$ and $P_o(\Omega)$ respectively denote the forecast probability for the event Ω and observed outcome for the event Ω (i.e. a binary probability of 0 if the event does not occur, and 1 if the event does occur). Note that the BS provides no meaningful information about probabilistic forecast skill if calculated for an individual forecast, and hence must be aggregated over time or space (or both). BS values range on the closed interval $[0, 1]$, and are negatively oriented, i.e. a lower BS indicates greater agreement between the forecast and the observed outcome. Perfect skill is achieved when $\mathrm{BS}(\Omega) = 0$, and can only be realized if the forecast probability is 100% and the event occurs, or the forecast probability is 0% and the event does not occur.

The CRPS evaluates the full forecast probability distribution against the observed outcome, and can be thought of as the integral of the BS over the continuous range of all mutually exclusive events. When applied to a variable that takes values on the interval $[0, 1]$, such as SIC, the CRPS can be written

$$\mathrm{CRPS} = \frac{1}{M}\sum_{j=1}^{M}\int_0^1 \left[F_{f_j}(x) - H_{y_o}(x - y_{o_j})\right]^2 dx. \tag{4.2}$$

In this definition, $F_{f_j}(x)$ is the cumulative distribution function (cdf) for the jth forecast distribution and $H_{y_o}(x - y_{o_j})$ is the Heaviside function for observed SIC y_o, which increases discontinuously from zero to one at the jth observed SIC value y_{o_j}. Like the BS, the CRPS is defined on the unit interval and is negatively oriented. Unlike the BS, the only way for the CRPS to be zero is if the forecast distribution is perfectly sharp (i.e. has zero variance). For probability SIC forecasts, such distributions are only physically reasonable when a grid cell is completely ice-covered or ice-free.

To compare the probabilistic forecast skill of two forecasting methods, we use the Brier skill score, $\mathrm{BSS} = 1 - \mathrm{BS}_{\mathrm{fcst}}/\mathrm{BS}_{\mathrm{ref}}$, and the continuous rank probability skill score, $\mathrm{CRPSS} = 1 - \mathrm{CRPS}_{\mathrm{fcst}}/\mathrm{CRPS}_{\mathrm{ref}}$. The subscript "fcst" simply refers to the forecast being evaluating relative to a reference forecast, denoted by the subscript "ref". The BSS and CRPSS are each defined on the interval $(-\infty, 1]$, are greater than zero when the forecast of interest has greater skill than the reference forecast,

are less than zero when the forecast of interest has poorer skill than the reference forecast, and are zero when both the reference forecast and forecast of interest have identical skill. In the case that $\mathrm{BS}_{\mathrm{fcst}} = \mathrm{BS}_{\mathrm{ref}} = 0$ ($\mathrm{CRPS}_{\mathrm{fcst}} = \mathrm{CRPS}_{\mathrm{ref}} = 0$), the BSS (CRPSS) is set to zero rather than the undefined value $1 - 0/0$. In the case that $\mathrm{BS}_{\mathrm{ref}} = 0$ ($\mathrm{CRPS}_{\mathrm{ref}} = 0$) but $\mathrm{BS}_{\mathrm{fcst}} \neq 0$ ($\mathrm{CRPS}_{\mathrm{fcst}} \neq 0$), the BSS (CRPSS) is set to $-\infty$ (represented numerically by a very large negative value).

4.4 Probability Estimates

In this section we describe two approaches for making probability forecasts from raw ensemble model output of SIC. The two methods, referred to here as the *count method* and *parametric method*, differ only with respect to their representations of the forecast probability distribution. Generally in probabilistic forecasting applications, the count method refers to computing forecast event probabilities by calculating discrete frequencies based on ensembles of raw (delta function) forecast values. In contrast, the parametric method fits a suitable probability distribution to the forecast ensemble, from which forecast event probabilities can be computed.

The aim of the parametric method is to improve the representation of the underlying forecast distribution, and to reduce the ill-effects that sampling may have on the estimation of event probabilities computed using the count method. Even if ensemble forecasts are perfectly calibrated, sampling may result in unreliable probability estimates (Richardson, 2001). The parametric method offers a means by which forecast event probability estimates can be improved, by interpolating and extrapolating probability density over the range of the under-sampled variable (e.g. Wilks, 2011). This in turn can produce more accurate estimates of quantiles, particularly extremes (Wilks, 2002; Roy et al., 2016). Most importantly from a practical perspective, the parametric method is expected to result in enhanced forecast skill relative to the count method, with larger improvements expected for smaller-sized ensembles (Wilks, 2002; Kharin and Zwiers, 2003). Of course, these advantages assume that the parametric distribution is suitable for modelling the underlying forecast distribution, and that the quality of fit to the discrete forecast ensemble is satisfactory.

Throughout the following, we say that $P(\Omega)$ is the probability for the event Ω, defined by the outcome that random variable X (in this case representing forecast SIC) exceeds a lower threshold value x_l. For instance, by choosing the particular lower SIC threshold $x_l = 0.15$, $P(\Omega)$ is equivalent to the SIP quantity described

above. Extension to calculating the probability of falling below an upper threshold, e.g. of sea ice absence for a threshold of 0.15, is trivial.

The count method and parametric method are now described in detail.

4.4.1 Count Method

The probability of the event Ω can be computed very simply using the count method. The count method is in fact the current method invoked in the request for SIP contributions to the SIO. This method does not assume a distribution on X; rather, it consists of counting the number of ensemble members that satisfy the event criteria, and reporting this relative frequency as the event probability. $P(\Omega)$ computed by the count method is thus

$$P(X > x_l) = \frac{\sum_{i=1}^{n} \mathbb{1}_{\{x_i > x_l\}}}{n}, \tag{4.3}$$

where $i \in 1, ..., n$ denotes the ensemble member, and $\mathbb{1}_{\{\Omega\}}$ is the indicator function, equalling one when Ω occurs and zero otherwise.

Equation 4.3 is related to the empirical cumulative distribution function (ecdf) $F_n(x)$,

$$F_n(x) = \begin{cases} 0, & x < x_1 \\ i/n, & x_i \leq x < x_{i+1}, \quad i = 1, ..., n-1 \\ 1, & x_n \leq x \end{cases} \tag{4.4}$$

calculated from the ordered (from smallest to largest) SIC ensemble forecast values $x_1, x_2, ..., x_n$. The probability given by Eq. 4.3 can be calculated using Eq. 4.4 as $P(X > x_l) = 1 - F_n(x_l)$, since $F_n(1) = 1$.

4.4.2 Parametric Method

Alternatively, SIC event probabilities can be computed by fitting an appropriate parametric distribution to the SIC forecast ensemble. For the statistical modelling of doubly-bounded random variables, such as SIC, the beta distribution stands out as a prominent option.

The probability density function (pdf) for the beta distribution is given by

$$f_{\text{beta}}(x; \alpha, \beta) = \frac{1}{\text{B}(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad 0 < x < 1 \tag{4.5}$$

where $\mathrm{B}(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt$ is the beta function, and $\alpha > 0$ and $\beta > 0$ describe the shape of the distribution. Increasing the parameter α has the effect of shifting the beta pdf toward zero, whereas increasing β shifts the distribution toward one. A detailed description of the properties of the beta distribution is presented in Johnson et al. (1995).

The beta distribution has been used in various applications within the fields of hydrology (e.g. Gottschalk and Weingartner, 1998), meteorology (e.g. Yao, 1974; Tompkins, 2002), and climatology (e.g. Henderson-Sellers, 1978; Li and Avissar, 1994). The distribution is particularly appealing because it can take a wide variety of shapes (e.g. exponential, skewed-unimodal, U-shaped), and because it can support variables that take on values between zero and one. However, the beta distribution cannot account for finite probability of a variable taking the value zero and/or one, as is often the case for ensemble SIC forecasts.

As an alternative, we use a modified version of the beta distribution, termed the zero-and-one inflated beta distribution (BEINF) (Ospina and Ferrari, 2010), that allows for finite probability at the endpoints zero and one. The four parameter BEINF distribution mixes the continuous beta distribution with the degenerate Bernoulli distribution.

The random variable $X \sim \mathrm{BEINF}(\alpha, \beta, p, q)$ has the pdf

$$f(x; \alpha, \beta, p, q) = \begin{cases} p(1-q), & x = 0 \\ (1-p)\ f_{\mathrm{beta}}(x; \alpha, \beta), & 0 < x < 1 \\ pq. & x = 1 \end{cases} \tag{4.6}$$

The parameter $0 \leq p \leq 1$ corresponds to the probability of X falling exactly at the end points of 0 or 1. The probability masses at the endpoints are modelled by a Bernoulli distribution (scaled by p), defined by a single parameter $0 \leq q \leq 1$, such that given that the random variable takes an end point value, $X = 1$ with probability q and $X = 0$ with probability $1 - q$. On the interval $(0, 1)$, the probability density is modelled by a beta distribution, as defined by Eq. 4.5, scaled by $1 - p$.

The cumulative distribution function (cdf) for the BEINF distributions is defined

as

$$F(x;\alpha,\beta,p,q) = \begin{cases} 0, & x < 0 \\ p(1-q), & x = 0 \\ p(1-q)+ & \\ \quad (1-p)F_{\text{beta}}(x;\alpha,\beta), & 0 < x < 1 \\ 1, & x \geq 1 \end{cases} \tag{4.7}$$

where $F_{\text{beta}}(x;\alpha,\beta) = \int_0^x f_{\text{beta}}(x';\alpha,\beta)ds'_x$ is the cdf for the beta distribution.

Although not done here, in some applications it may be desired to clip forecast SIC values below 0.15 to zero, as is often done with satellite observations of SIC. Modelling SIC using the BEINF distribution after clipping can be done by transforming those values $x \in (0.15, 1)$ using

$$s_x = g(x) = \frac{x-c}{1-c}, \quad c < x < 1 \tag{4.8}$$

where $c = 0.15$ is the lower value below which SIC is clipped to zero. The resultant variable s_x takes values on the interval $(0, 1)$, which is then used in place of x in Eqs. 4.6–4.7. However, it should be noted that if this is done, both the count method and parametric method will yield identical forecast probabilities for any event corresponding to a SIC threshold $x_l \leq 0.15$

The four parameters that describe the shape of the BEINF distribution are estimated for each SIC ensemble hindcast, comprised of members $x_1, x_2, ..., x_n$, using Maximum Likelihood (ML) estimation. As described in Appendix B.1, the ML estimates of parameters p and q, denoted $\hat{p}$ and $\hat{q}$, are computed analytically from the complete ensemble of size n (and fit the data perfectly). The ML estimates of parameters α and β, denoted $\hat{\alpha}$ and $\hat{\beta}$, must be computed numerically from those $n - m$ ensemble members $x_1, ..., x_{n-m}$ (where the value m denotes the number of zeros and ones in the complete sample) that lie on the interval $(0, 1)$, denoted x_{sub}.

In the infrequent instances where the ML estimation algorithm does not converge, the method of moments is used to estimate α and β. The method of moments for the beta distribution, which is described in Appendix B.1, requires that $\text{var}(x_{sub}) < \bar{x}_{sub}(1-\bar{x}_{sub})$, where $\bar{x}_{sub}$ is the geometric mean of x_{sub}, and $\text{var}(x_{sub})$ is the unbiased estimator of sample variance.

There are special cases to consider when the parametric method cannot be used, because the parameters α and β cannot be estimated by either ML estimation or by

the method of moments. These cases are as follows:

case 1: $n - m = 0$,

case 2: $n - m = 1$,

case 3: $n - m > 1$, but $\text{var}(x_{sub}) = 0$,

case 4: the ML estimation algorithm does not converge and the method-of-moments condition that $\text{var}(x_{sub}) < \bar{x}_{sub}(1 - \bar{x}_{sub})$ is not met.

Of the total 867 072 ensemble hindcasts (32 years × 4 initialization months × 6 forecast months × 1129 ocean grid cells), case 1 occurs 41.9% of the time (either because the ocean grid cell is completely ice-covered or is completely ice-free). Excluding those hindcasts where case 1 occurs (i.e. the remaining 58.1% hindcasts), case 2 occurs 9.03% of the time, case 3 occurs 0.001% of the time, and case 4 occurs 0.025% of the time. This leaves a total of 458 014 hindcasts that can be fit to the BEINF distribution. The different choices for handling cases 1-4 will be described in subsequent sections, as the choices are specific to the analysis being considered.

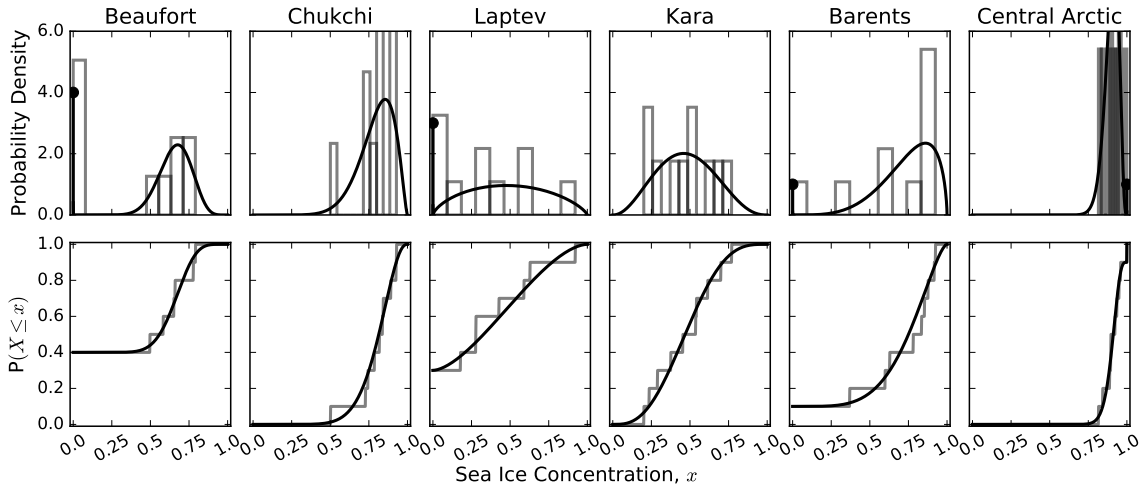

Figure 4.1: SIC ensemble hindcasts for six model grid cells spanning the Arctic Ocean (from regions labelled). Top row: normalized histogram for the hindcast ensemble and corresponding fitted BEINF pdf; the probability masses at the endpoints are scaled by 10 for the purposes of visual comparison. Bottom row: ecdf for the hindcast ensemble and corresponding fitted BEINF cdf.

To illustrate the properties of the BEINF distribution, we present count-based histogram distributions and the corresponding fitted BEINF pdfs for the six SIC

ensemble hindcasts, together with the count-based ecdfs and BEINF-based cdfs, in Fig. 4.1. For reference, these hindcasts are for August, 2002, at a lead time of two months. These particular examples have been chosen to show the wide range of possible distribution shapes that SIC ensemble forecasts can take, and to provide an indication of the suitability of the BEINF distribution for modelling SIC ensemble forecasts.

In Fig. 4.1, the SIC ensemble hindcasts demonstrate how the BEINF distribution extrapolates the probability density away from the region occupied by the empirical distribution. The BEINF distribution is shown to effectively interpolate and smooth probability density between gaps in the empirical distribution. Furthermore, because the BEINF distribution models zeros and ones separately from the SIC values on the interval $(0, 1)$, the parametric distribution is also able to capture bimodality seen for the hindcast ensemble in the Beaufort Sea.

A visual comparison of the BEINF cdf and the ecdf for each of the six cases in Fig. 4.1 provides evidence that the BEINF distribution is a suitable distribution for the application of SIC ensemble forecasts. To quantify the suitability of the BEINF distribution for modelling SIC, we perform goodness-of-fit tests on each of the 458 014 remaining ensemble hindcasts of SIC, excluding grid cells where the count method has been applied (i.e. cases 1–4 described above).

We test the null hypothesis H_0, at significance level α_s, that each SIC ensemble hindcast, comprised of the ensemble members $x_1, x_2, ..., x_n$, is drawn from the population BEINF$(\hat{\alpha}, \hat{\beta}, \hat{p}, \hat{q})$. Because ML estimates $\hat{p}$ and $\hat{q}$ are fit to the data sample exactly, we do not include these parameters in the goodness-of-fit tests, and H_0 reduces to that the transformed sub-sample $x_1, ..., x_{n-m}$, comes from population beta$(\hat{\alpha}, \hat{\beta})$. The alternative hypothesis, H_1, is simply that H_0 is false.

We employ three empirical distribution function (EDF) tests (Stephens, 1986) of varying power to each SIC ensemble hindcast for testing H_0 at significance level α_s. These tests are the Kolmogorov-Smirnov (KS) test, the Cramer-Von Mises (CVM) test, and the Anderson-Darling (AD) test. In order to apply these tests to the beta distribution, we follow the approach recommended in Raschke (2009, 2010). Further details of this approach can be found in Appendix B.2.

Based on the results of the three EDF tests, we conclude that H_0 can be rejected for 11% (AD), 8% (CVM), and 9% (KS) of the SIC ensemble hindcasts at the significance level $\alpha_s = 0.05$. Although we cannot state definitively that the null hypothesis holds (and in fact there is no reason to expect that the distribution of x_{sub} is exactly beta),

the small percentage of rejections suggests that the BEINF distribution is generally appropriate for modelling SIC ensemble forecasts.

4.5 Probabilistic Hindcast Skill: Count vs Parametric

Probabilistic hindcast skill for the count method and parametric method is compared using both *pseudo-perfect model* (PPM) experiments and *observation-verified* (OV) experiments. In the PPM experiments, the initialized hindcasts described in section 4.3.1 are validated against a single ensemble member randomly drawn from the 10-member forecast ensemble. The hindcast probabilities estimated by both the count and parametric methods are then computed from the remaining 9 ensemble members. In the OV experiments, these same initialized hindcasts are verified against observed SIC.

Unlike the OV experiments, the PPM experiments provide a means to compare the count and parametric methods in the absence of model errors, either from initial conditions or from the model itself, including no direct knowledge of true forecast uncertainty (Wilks, 2002). The differences in skill between the PPM experiments and the OV experiments are thus primarily attributable to model errors. To a much lesser extent, a second contributor to differences in skill is the use of all 10 ensemble members in the OV experiments, compared to the use of 9 ensemble members in the PPM experiments.

For this comparison of skill, in instances when any of the cases 1-4 (described in section 4.4.2) are encountered, event probabilities for the parametric method are set equal to the event probabilities computed by the count method.

4.5.1 CRPSS evaluation

The comparison of probabilistic hindcast skill using the count and parametric methods is first assessed with the CRPSS, which is based on the CRPS for each forecast method. As stated earlier, the CRPS given by Eq. 4.2, evaluates the entire SIC distribution estimated by each method against the observed outcome.

The CRPSS for both the PPM and OV experiments is calculated with the parametric method as the forecast being evaluated, and the count method as the reference forecast. To simplify this skill comparison, the area-weighted spatial average of the

integral in the CRPS is computed for each forecast method. In the PPM experiments, this spatial average is taken over that part of the domain where ice is present in the model (per forecast year, initialization month, and lead time) in at least one ensemble member, and in the OV experiments the spatial average is taken over that part of the domain where ice is present either in the model (per forecast year, initialization month, and lead time) in at least one ensemble member, or in the observations. By taking the spatial average prior to the temporal average, we eliminate the effect that the evolving ice edge location over the 1981-2012 period may have on the CRPS. The CRPSS is then computed from these spatially– and temporally– averaged CRPS values.

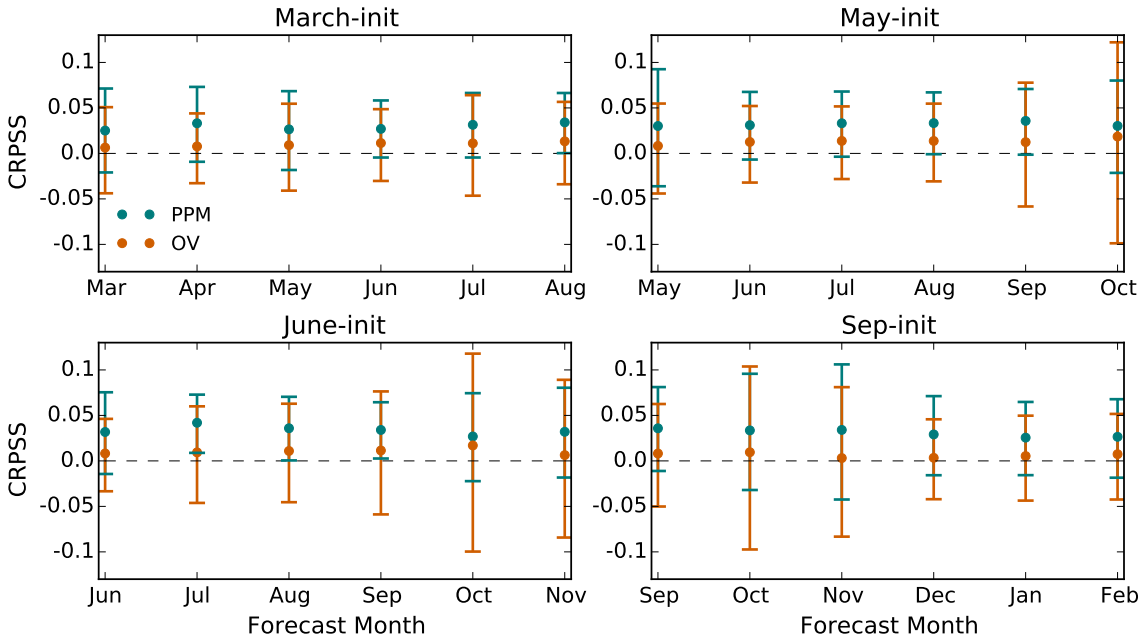

Figure 4.2: The CRPSS for the parametric method (forecast being evaluated) relative to the count method (reference forecast). Blue circles: PPM experiments; red circles: OV experiments. Skill improvement using the parametric method is indicated by CRPSS values greater than zero. Vertical lines are the 5th to 95 percent confidence intervals of the CRPSS values. Each panel is for a different initialization month (as labelled).

In both the PPM and OV experiments, the parametric method outperforms the count method, as indicated by the consistently positive CRPSS values in Fig. 4.2. The improvement in forecast skill using the parametric method is evident in each of the four initialization months, and is approximately constant with increasing lead time in both experiments. A considerably larger improvement in skill using the parametric

method is seen in the PPM experiments compared to the OV experiments. The relatively modest improvement using the parametric method in the OV experiments indicates that in these particular hindcasts, model and observational errors degrade the potential improvement suggested in the PPM experiments.

Uncertainty in these CRPSS values is determined by the 5th and 95th percent confidence intervals presented in Fig. 4.2, computed by the bootstrapping method (Wilks, 2011). Despite uncertainties in these CRPSS values being relatively large, the improvement in skill is statistically significant (5th percentile greater than zero) for some summer forecast months between July-September in the PPM experiments. The improvement in skill is not statistically significant for other individual months in the PPM experiment, or for any individual months in the OV experiments. Nonetheless, the consistently positive CRPSS values in both the PPM and OV experiments provide evidence for the robustness of the improvement using the parametric method.

4.5.2 BSS evaluation

As probabilistic forecasts of specific SIC events are of greatest practical interest, we now use the BSS to compare probabilistic hindcast skill between the count and parametric methods. The BSS is computed for several different events Ω, in which the lower SIC threshold x_l is varied from 0.1 to 0.9, in increments of 0.1. Note that because Ω is binary, inspection of Eq. 4.1 reveals that $\mathrm{BS}(\Omega) = \mathrm{BS}(\Omega^c)$, where Ω^c is the complement of the event Ω; i.e. the BS for the event that $X > x_l$, is the same as that for the event $X \leq x_l$. The BSS is computed from the spatially-averaged squared quantity in the BS (over the relevant grid cells as described for the CRPSS) for each method prior to averaging in time. We present the BSS for the PPM and OV experiments in Fig. 4.3.

Like the CRPSS results described above, the BSS values for the PPM experiments are nearly always positive, which demonstrates that the parametric method results in greater skill than the count method (Fig. 4.3a). Improvements in probabilistic forecast skill are generally higher for mid-SIC and high-SIC event thresholds than for low-SIC event thresholds.

Comparison of probabilistic forecast skill between the count and parametric methods in the OV experiments indicates a more modest, yet overall improvement in forecast skill using the parametric method relative to the count method (Fig. 4.3b). However, for some low-SIC event thresholds, and to a lesser extent some high-SIC

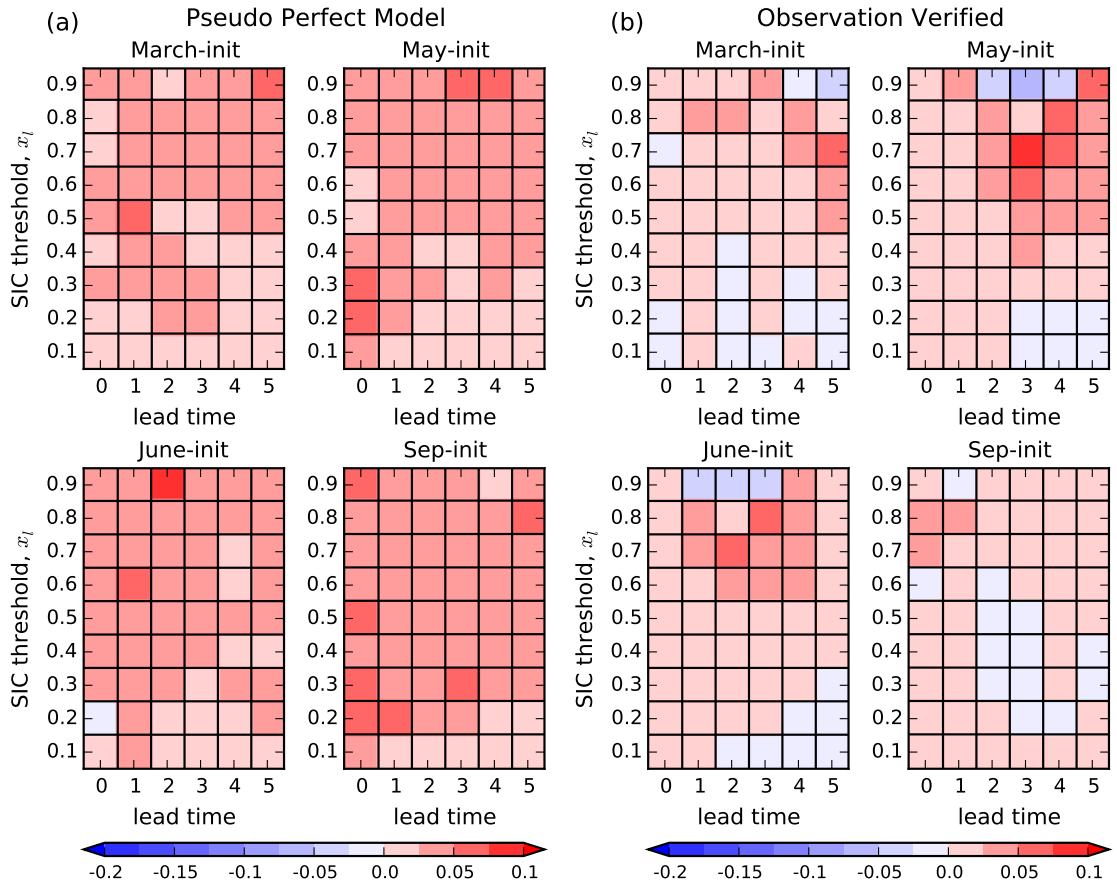

Figure 4.3: The BSS for the parametric method (forecast being evaluated) and the count method (reference forecast) for the (a) PPM experiments, and the (b) OV experiments. Each panel in (a) and (b) is for a different initialization month (as labelled). Skill improvement using the parametric method is indicated by positive (red) BSS values

event thresholds in the initialization months of March, May, and June, the parametric method shows slightly lower forecast skill than the count method. For forecasts of fall and winter sea ice conditions initialized in September, this reduction in skill by the parametric method is seen for low and mid-range SIC event thresholds. The largest improvement using the parametric method in the OV experiments is seen for mid-range event thresholds during the summer forecast months of July-September, for initializations in March, May, and June.

In Appendix B.3, the PPM experiments are used to investigate the dependence of skill improvement using the parametric method on the SIC event threshold. Specifically, we address why low-SIC event thresholds show a more modest skill improvement

compared to mid– and high– SIC event thresholds by focusing on the particular event thresholds $x_l = 0.1$ and $x_l = 0.8$. These events correspond respectively to a relatively modest improvement and a relatively large improvement in hindcast skill using the parametric method (Fig. 4.3a).

First, because the parametric method is expected to particularly outperform the count method in the estimation of extreme quantiles of the underlying forecast distribution (as opposed to less extreme quantiles like the median) (Wilks, 2002), differences in skill forecasting specific events could result from differences in the frequency with which extreme quantiles are sampled for that event. As shown in Appendix B.3, extreme quantiles are indeed sampled more frequently for the high-SIC event threshold compared to low-SIC event threshold. This therefore increases the likelihood for the parametric method to outperform the count method for the high-SIC event threshold relative to the low-SIC event threshold.

Second, improved skill using the parametric method is expected to be most apparent when assessing skill over a large number of hindcasts, purely based on sampling considerations. Given this fact, it is important to bear in mind that each BSS value shown in Fig. 4.3 (for a given initialization month and lead time) is computed from hindcasts for a collection of grid cells that vary according to the SIC event threshold of interest (i.e. those grid cells where $\mathrm{BS}_{\mathrm{fcst}}(\Omega) \neq \mathrm{BS}_{\mathrm{ref}}(\Omega)$). As shown in Appendix B.3, the number of hindcasts over which the BSS is non-zero is substantially larger for the high-SIC event threshold compared to the low-SIC event threshold. This is due to the fact that the majority of the ice pack is dominated by high SIC, which results in more locations where $\mathrm{BS}_{\mathrm{fcst}}(\Omega) \neq \mathrm{BS}_{\mathrm{ref}}(\Omega)$ for the high-SIC event threshold compared to the low-SIC event. In turn, this reduces the likelihood that the improvement in skill using the parametric method is obscured by sampling noise for the high-SIC event.

Differences in hindcast skill between the PPM and OV experiments seen in Fig. 4.3 are almost entirely due to the influence of model biases. For instance, in the central Arctic in summer months, the parametric method estimates forecast probabilities $P(X > x_l)$ for $x_l = 0.9$, that are systematically lower than count method estimates (not shown). However, SIC is biased slightly low in CanCM3 in this region in these months. The systematically lower forecast probabilities for SIC exceeding 0.9 degrades the skill of the parametric method relative to the count method, and largely contributes to the negative BSS values for this event seen in Fig 4.3b. Since model biases are by construction absent in the PPM experiments, the parametric method

outperforms the count method for this event threshold and during summer months.

4.6 Calibration

In both initialized hindcasts and freely-running (i.e. uninitialized) historical experiments, CanCM3 overestimates pan-Arctic SIE in all calendar months, contains widespread (mainly positive) SIC biases, and underestimates the magnitude of the negative trend in pan-Arctic SIE (Merryfield et al., 2013b; Sigmond et al., 2013; Dirkson et al., 2017). To account for model errors in probabilistic SIC forecasts, we employ a novel version of quantile mapping (QM), specifically designed for the SIC variable and the BEINF distribution. We refer to this calibration technique as *trend-adjusted quantile mapping* (TAQM).

Before describing TAQM, we first introduce the standard QM technique, as it would be applied in a forecasting framework. QM can be used to calibrate a forecast value x_t (where t denotes the forecast year of interest), by mapping between quantiles of a historical model (MH) probability distribution and an observed historical (OH) probability distribution, according to

$$\hat{x}_t = F_{\mathrm{o}}^{-1}[F_{\mathrm{m}}(x_t)]. \tag{4.9}$$

In Eq. 4.9, $\hat{x}_t$ denotes the quantile-mapped forecast value, F_{o}^{-1} is the inverse of the cdf for the OH probability distribution, and F_{m} is the cdf for the MH probability distribution. When F_{m} and F_{o} are represented parametrically, Eq. 4.9 can be evaluated either analytically or numerically depending on whether F_{m} and F_{o}^{-1} can be evaluated exactly. When F_{m} and F_{o} are known only as ecdfs, as when the count method is applied, Eq. 4.9 can be evaluated e.g. by interpolating between values on a quantile-quantile plot. In practice, individual forecast ensemble members x_t, are used as inputs to Eq. 4.9.

As an example, consider applications where F_{o} and F_{m} are the cdfs of normally-distributed random variables. In such cases, it can be shown that Eq. 4.9 reduces to $\hat{x}_t = \mu_o + (x_t - \mu_m)\frac{\sigma_o}{\sigma_m}$ (see Appendix B.4), where μ_o and μ_m are the means, and σ_o and σ_m are the standard deviations of the respective distributions F_o and F_m. For such normally-distributed random variables, QM corrects for the mean and spread of the forecast random variable X_t, according to the bias in mean and spread in the historical model distribution. Alternatively, for applications where the distributions

F_{o} and F_{m} have higher-order moments, Eq. 4.9 is effective at correcting, not only for the model bias in mean and spread, but for these higher-order moments as well.

In its standard form given by Eq. 4.9, QM is not a suitable calibration method for seasonal ensemble hindcasts of SIC from 1981 to the present. Like all calibration methods, QM assumes that the statistics of the MH distribution and OH distribution are stationary, and therefore consistent with the statistics of the respective distributions for the forecast values X_t and $\hat{X}_t$. This assumption is commonly violated for SIC forecasts however, as negative trends in SIC over the historical period may be pronounced, particularly in more recent years. Furthermore, QM is not well suited when F_m and F_o are modelled as a discontinuous distribution (like the BEINF distribution), since mapping to or from an endpoint value of zero or one can result in spurious SIC values, such as numerous identical SIC values that are neither zero nor one. TAQM offers a way to resolve both of these complications, enabling SIC ensemble forecasts to be calibrated.

The TAQM calibration technique is now described in detail.

4.6.1 Trend-adjustment

As a first step in TAQM, MH and OH data over the period 1981-2012 are adjusted to account for the non-stationarity in the mean SIC state, which is a function of the forecast year of interest. The non-stationarity of higher-order moments in SIC, such as variance, is not considered in the present study; however, this is potentially an important topic of future research.

Consider the MH SIC time series $x_j(\tau)$ and OH SIC time series $y(\tau)$, where τ denotes all years within the hindcast period (1981-2012), excluding the forecast year t, and j denotes ensemble member. The trend-adjusted values (denoted respectively by TAMH and TAOH) for a particular forecast year t, are computed as

$$x'_j(\tau) = [x_j(\tau) - \tilde{x}(\tau)] + \tilde{x}_t, \tag{4.10a}$$

$$y'(\tau) = [y(\tau) - \tilde{y}(\tau)] + \tilde{y}_t, \tag{4.10b}$$

In Eqs. 4.10a–4.10b, the tilde symbol $\tilde{\cdot}$ denotes the piecewise linear least-squares fit

to the time series of the given variable,

$$\tilde{z}(\tau) = \begin{cases} a_1\tau + b_1, & 1981 \leq \tau < 1999 \\ a_2\tau + b_2, & 1999 \leq \tau \leq 2012 \end{cases} \tag{4.11}$$

where z denotes either x or y. When z corresponds to x, the ensemble mean time series is fit to Eq. 4.11 and used in Eqs. 4.10a–4.10b.

Equations 4.10a–4.10b have the effect of removing the piece-wise linear trends in the OH and the MH data defined by Eq. 4.11, and re-centering the mean of the respective time series about a *non-stationary mean*, which we define as Eq. 4.11 evaluated at the forecast year t.

The purpose of defining the linear least-squares equation for the MH and OH time series in a piecewise fashion, is to represent the observed acceleration of Arctic sea ice decline. The particular break point 1999 was chosen by considering spatial maps of the observed trend acceleration $a_2 - a_1$ for September, when trends are strongest. Among candidate break points ranging from 1995-2004, the occurrence of accelerating negative trends (i.e. $a_2 - a_1 < 0$) is most spatially extensive in 1999.

Eqs. 4.10a–4.10b do not constrain the TAMH and TAOH values to $[0, 1]$, so we set values that fall below zero or above one to the appropriate lower or upper bound. Because of this step, the mean of the trend-adjusted data may no longer equal the non-stationary mean. Therefore, we iteratively adjust

$$x'_j(\tau) \leftarrow [x'_j(\tau) - \langle x'(\tau)\rangle] + \tilde{x}_t, \tag{4.12a}$$

$$y'(\tau) \leftarrow [y'(\tau) - \langle y'(\tau)\rangle] + \tilde{y}_t, \tag{4.12b}$$

truncating to within $[0, 1]$ after each iteration, until $\langle x'(\tau)\rangle = \tilde{x}_t$ and $\langle y'(\tau)\rangle = \tilde{y}_t$ within an absolute tolerance of 10^{-2}. In Eqs. 4.12a–4.12b, the angled brackets denote the temporal mean for $y'(\tau)$ and the temporal– and ensemble– mean for $x'(\tau)$

The trend adjustment technique for a June-initialized hindcast of 2011 September SIC for a grid cell in the Kara Sea is illustrated in Fig. 4.4. Both the MH and OH time series in the left-hand panels show marked negative trends over the 1999-2012 period ($p < 0.05$) prior to trend adjustment. Before 1999, the MH time series shows a slightly negative trend while the OH time series shows a slightly positive trend, neither of which are statistically significant ($p > 0.05$). Following trend adjustment by Eqs. 4.10a–4.10b, the respective means of the TAMH and TAOH time series are

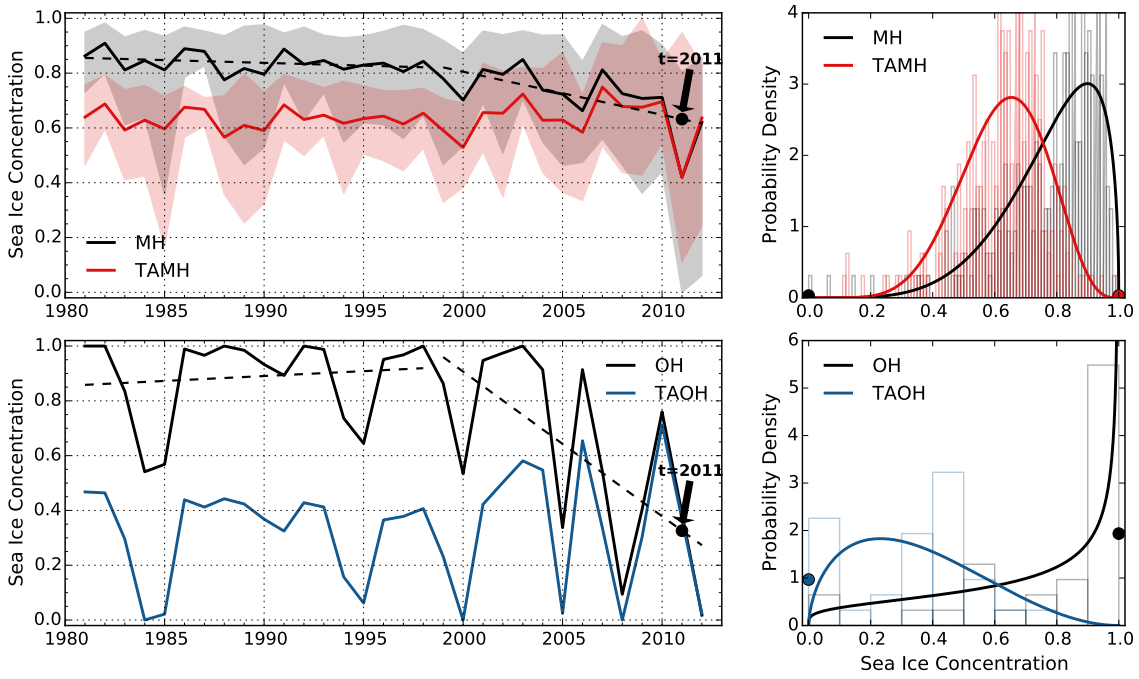

Figure 4.4: Illustration of the trend-adjustment technique employed as a first step in TAQM. Solid black lines are the MH and OH time series (left-hand panels), and the MH and OH histograms and BEINF pdfs (right-hand panels). Dashed black lines are linear-least squares fits to the MH and OH time series over the 1981-1998 and 1999-2012 periods. The red and blue solid lines are respectively the TAMH and TAOH time series (left-hand panels) and TAMH and TAOH histograms and BEINF-fitted pdfs (right-hand panels). The mass points at zero and one for the BEINF pdfs have been multiplied by 10 for comparison with the histogram distributions.

centered about the non-stationary means of the MH and OH time series, defined by Eq. 4.11 evaluated at $t = 2011$. For this particular example, the iteration given by Eqs. 4.12a–4.12b is not required.

The BEINF distributions for these data are shown in the right-hand panels of Fig. 4.4. The TAMH distribution is shifted toward slightly lower SIC values relative to the MH distribution and has become more bell shaped. A more noticeable change relative to the OH distribution is seen for the TAOH distribution. The TAOH distribution is shifted toward much lower SIC values and has changed from quasi-exponential to unimodal. Further, the TAOH distribution no longer shows a probability mass at one, and instead shows an increased probability of equalling zero.

A potential criticism of the TAQM formulation described above is that future information is used in computing the trends that are subtracted to obtain the TAOH

and TAMH time series used as input to TAQM, which in turn might inflate skill scores (even though conventional cross validation has been applied). This fundamental issue of how to apply bias corrections in a non-stationary climate is not unique to forecasting of sea ice, and has for example been encountered in the context of decadal hindcasts for which observed and modeled temperatures have differing long-term trends (Kharin et al., 2012).

Similarly to Kharin et al. (2012), which removed trends based on the entire validation period, we take the view that estimation of trends using the full verification period is justified by the objective of obtaining the best possible forecasts during the subsequent "real time" forecast period. A straightforward extension is that the estimated historical trends can be updated annually using the entire past record for successive real time forecasts; such a procedure has been applied annually to real-time decadal forecasts from the CanCM4 model (V. V. Kharin, personal communication).

4.6.2 Parametric Fitting

As a second step in TAQM, three sets of data are fit to the parametric BEINF distribution: the TAMH time series (32 years $\times$ 10 ensemble members), the TAOH time series (31 years), and the forecast ensemble to be calibrated (10 ensemble members). Throughout the following, we use the notation $X' \sim \text{BEINF}(\alpha_{x'}, \beta_{x'}, p_{x'}, q_{x'})$ for the TAHM random variable, $Y' \sim \text{BEINF}(\alpha_{y'}, \beta_{y'}, p_{y'}, q_{y'})$ for the TAOH random variable, and $X_t \sim \text{BEINF}(\alpha_{x_t}, \beta_{x_t}, p_{x_t}, q_{x_t})$ for the hindcast random variable for year t. The reader is reminded here that the TAMH and TAOH distributions are dependent on the forecast year t, and thus must be fit for each individual year.

As described earlier in section 4.4.2, the parameters α and β cannot be fit for the cases 2–4 outlined therein. These parameters cannot be fit in case 1 either, but how to proceed in this specific case will be discussed later in section 4.6.3. In order to apply the TAQM calibration method for cases 2–4, the three BEINF distributions described above must be defined. As stated previously, the difficulty with defining the BEINF distribution in these cases is that the size $n - m$ sub-sample z_{sub}, where z can represent x', y' or x_t, cannot be fit to the beta-distribution portion of the BEINF distribution using ML estimation or the method of moments. In order to use the BEINF distribution in such cases, an approximate fitting method for estimating α_z and β_z (described in detail in Appendix B.1) is employed, so that the BEINF distribution can be defined.

4.6.3 Calibrating BEINF Parameters

As a final step in TAQM, we calibrate the parameters of the forecast distribution BEINF($\alpha_{x_t}, \beta_{x_t}, p_{x_t}, q_{x_t}$). Even after trend-adjustment, the BEINF distribution as a whole is ill-suited for QM, as it is a discontinuous distribution. However, the beta portion of the BEINF distributions can be used in quantile mapping (assuming non-stationarities have been accounted for). The calibration of α_{x_t} and β_{x_t} is thus done by applying QM (after trend adjustment), whereas the calibration of parameters p_{x_t} and q_{x_t} is done using a simple mean bias correction.

To calibrate α_{x_t} and β_{x_t}, we input those $n-m$ ensemble members $x_{t,1}, ..., x_{t,n-m}$ into

$$\hat{x}_t = F^{-1}_{o,\text{beta}}\left[F_{m,\text{beta}}(x_t)\right], \quad 0 < x_t < 1. \tag{4.13}$$

to be quantile mapped to values $\hat{x}_{t,1}, ..., \hat{x}_{t,n-m}$. In Eq. 4.13, $F^{-1}_{o,\text{beta}}$ is the inverse cdf of the beta portion of the BEINF distribution fit to the TAOH data, and $F_{m,\text{beta}}$ is the beta portion of the BEINF cdf fit to the TAMH data. The parameters $\alpha_{\hat{x}_t}$ and $\beta_{\hat{x}_t}$ are then estimated from the quantile mapped values $\hat{x}_t$ using ML estimation, as it is described in Appendix B.1. In cases 2-4 when the approximate method of Appendix B.1 has been used for the fitting procedure, 1 000 randomly drawn "pseudo ensemble members" from the population beta($\alpha_{x_t}, \beta_{x_t}$) are used as input to Eq. 4.13, so that the quantile mapped values can be fit to the BEINF distribution using ML estimation.

We calibrate parameters p_{x_t} and q_{x_t}, by adding to the forecast probability of equalling zero, given by $p_{x_t}(1-q_{x_t})$, and the forecast probability of equalling one, given by $p_{x_t}q_{x_t}$, the bias in these quantities for the TAMH distribution relative to the TAOH distribution:

$$\begin{aligned} p_{\hat{x}_t}(1-q_{\hat{x}_t}) &= p_{x_t}(1-q_{x_t}) + [p_{y'}(1-q_{y'}) \\ &\quad -p_x(1-q_x)], &(4.14a)\\ p_{\hat{x}_t}q_{\hat{x}_t} &= p_{x_t}q_{x_t} + (p_{y'}q_{y'} - p_xq_x). &(4.14b) \end{aligned}$$

Performing elimination on Eqs. 4.14a–4.14b, it can be easily shown that the calibrated parameters $p_{\hat{x}_t}$ and $q_{\hat{x}_t}$ are given by

$$\begin{aligned} p_{\hat{x}_t} &= p_{x_t} + p_{y'} - p_{x'}, &(4.15a)\\ q_{\hat{x}_t} &= \frac{p_{x_t}q_{x_t} + p_{y'}q_{y'} - p_{x'}q_{x'}}{p_{\hat{x}_t}}, &(4.15b) \end{aligned}$$

where $q_{\hat{x}_t}$ is set to zero when $q_{\hat{x}_t} = 0/0$.

An important situation to consider is when any of $p_{x_t} = 1$, $p_{x'} = 1$, or $p_{y'} = 1$ (i.e. when any of the BEINF distributions involved in calibration are described completely in terms of p and q); note that this is case 1 in section 4.4.2. In such situations, Eq. 4.13 cannot be evaluated because the beta portion of the BEINF distribution cannot be defined. In the case that either the TAMH distribution or TAOH distribution (or both) are always ice-free or always ice-covered (i.e. $p_{x'} = 1$ and/or $p_{y'} = 1$), but the hindcast distribution is not (i.e. $p_{x_t} \neq 1$), the two options are to: (1) trust the hindcast distribution, or (2) trust the TAOH distribution. In the case that all hindcast ensemble members have 0 % SIC or 100% SIC (i.e. $p_{x_t} = 1$), the options remain the same. The preferred choice of how to proceed in these situations is dependent on the degree of bias in the dynamical model being used, which is likely dependent also on lead time and season.

We find that for CanCM3 it is nearly always preferred to trust the TAOH distribution in such cases and not the uncalibrated forecast distribution, with the exception of probabilistic hindcasts of September SIC initialized on September 1$^{\text{st}}$ (not shown). Thus, for September hindcasts at a lead time of zero months we trust the uncalibrated hindcast distribution when $p_{x_t} = 1$. For all other months we simply set $\alpha_{\hat{x}_t} = \alpha_{y'}$, $\beta_{\hat{x}_t} = \beta_{y'}$, $p_{\hat{x}_t} = p_{y'}$, $q_{\hat{x}_t} = q_{y'}$ when any of $p_{x_t} = 1$, $p_{x'} = 1$, or $p_{y'} = 1$ (i.e. we trust the TAOH distribution over the uncalibrated forecast distribution).

4.6.4 Example

We now illustrate the application of the calibration procedure described above for the same case used to illustrate trend adjustment in Fig. 4.4. This example is summarized in Fig. 4.5, where the left-hand panel represents the TAOH and TAMH distributions used to obtain the calibration, and the right-hand panel shows the uncalibrated and calibrated forecast distributions. Because the calibration of α_{x_t} and β_{x_t} is done separately from p_{x_t} and q_{x_t}, we split the BEINF distribution into its Bernoulli portion (circles) and its beta portion (solid curves) to demonstrate the application of Eq. 4.13 and Eqs. 4.15a–4.15b separately. The calibration of both parts of the BEINF distribution for this particular example is now described.

We illustrate TAQM for a particular forecast value, namely $x_t = 0.54$, marked by the dashed orange line in the right-hand panel of Fig. 4.5. The TAMH beta cdf evaluated at this forecast value is given by the quantile $F_{m,\text{beta}}(0.54) = 0.26$,

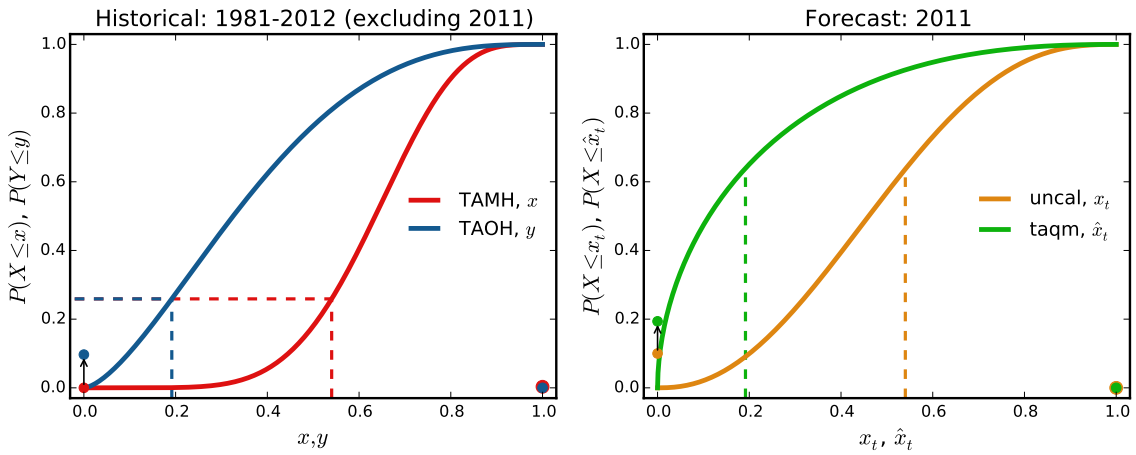

Figure 4.5: Illustration of the BEINF parameter calibration using TAQM for the same hindcast used to illustrate the trend adjustment in Fig. 4.4. Left panel: TAMH and TAOH. Right panel: uncalibrated hindcast and calibrated hindcast for the year 2011. Solid lines are the beta cdfs for the TAMH (red), TAOH (blue), uncalibrated hindcast (orange), and TAQM-calibrated hindcast (green). Circles mark the probabilities of equalling zero and one. Dashed lines and black arrows are described in the main text.

and is marked by the intersection of the dashed red lines in the left-hand panel of Fig. 4.5. The inverse of the TAOH beta cdf evaluated at this same quantile of 0.26, corresponds to an observed SIC value $F^{-1}_{o,\text{beta}}(0.26) = 0.19$, marked by the intersection of the dashed blue lines also in the left-hand panel of Fig. 4.5. The value of the calibrated forecast variable for $x_t = 0.54$ is thus $\hat{x}_t = 0.19$. Eq. 4.13 is solved for all forecast ensemble members, and the calibrated forecast parameters $\alpha_{\hat{x}_t}$ and $\beta_{\hat{x}_t}$ are then computed from the calibrated forecast values $\hat{x}_t$. The resulting best-fit beta portion of the BEINF cdf is shown by the green curve in Fig. 4.5.

Next, the calibration of the Bernoulli-portion of the BEINF distribution, described by p_{x_t} and q_{x_t}, is illustrated. The calibrated parameters, found by solving Eqs. 4.15a–4.15b, yield an increased forecast probability of equalling zero (green circle in the right-hand panel of Fig. 4.5) by the amount given by the length of the black arrow in the left-hand panel of Fig. 4.5, which extends from the blue circle to the red circle (i.e. the historical bias in the probability of SIC being zero). This calibration increases the forecast probability of SIC being zero from 10% to 20%. The probability that SIC is equal to one is 0% for all distributions in Fig. 4.5, and is therefore also 0% for the calibrated forecast.

4.7 TAQM-calibrated Hindcast Skill

We now assess probabilistic skill of TAQM-calibrated hindcasts over the recent 13-year period 2000-2012, so that the results are representative for the current epoch of reduced and declining Arctic sea ice. Specifically we use the CRPSS to compare the skill of TAQM-calibrated hindcasts against that of three reference hindcasts: uncalibrated probabilistic hindcasts (fit to the BEINF distribution), the 1981-2010 OH distribution, and the TAOH distribution. The assessment of skill against the uncalibrated hindcasts determines the degree to which TAQM is able to reduce model errors. The comparison with the 1981-2010 OH distribution specifies hindcast skill against the commonly-used reference hindcast given by the observed climatological distribution. Because of the strong trends in SIC data over the historical record, the comparison against the TAOH distribution provides a more conservative assessment of hindcast skill than the comparison against the raw 1981-2010 OH distribution.

We provide a pan-Arctic quantification of this improvement by computing the percentage of grid points showing positive CRPSS values, relative to the total number of non-zero CRPSS values, abbreviated hereafter as PI for percentage improvement.

4.7.1 TAQM vs Uncalibrated

The skill comparison between the TAQM-calibrated hindcasts and the uncalibrated parametric method hindcasts is shown in spatial maps of the CRPSS in Fig. 4.6. Clearly there is a general improvement in hindcast skill for the TAQM-calibrated hindcasts, as seen by the expansive regions showing positive CRPSS values in most target months. The few areas showing lower probabilistic skill for the TAQM-calibrated hindcasts compared to the uncalibrated hindcasts correspond to regions where skill is already high for the uncalibrated hindcasts (such as during the first hindcast month); however, this reduction in skill is generally quite small, particularly after a lead time of zero months.

Locations where poorer skill is seen for the TAQM-calibrated hindcasts compared to uncalibrated hindcasts tend to correspond to locations that have experienced a rare event at some point in the hindcast record. For instance, the region of negative CRPSS values in the western central Arctic in October corresponds to a region where SIC is nearly always close to 100%, but in 2007 fell to only 30%. This single extreme case results in a TAOH distribution with probability density concentrated between approximately 30% and 100%, thereby having the effect of shifting the forecast dis-

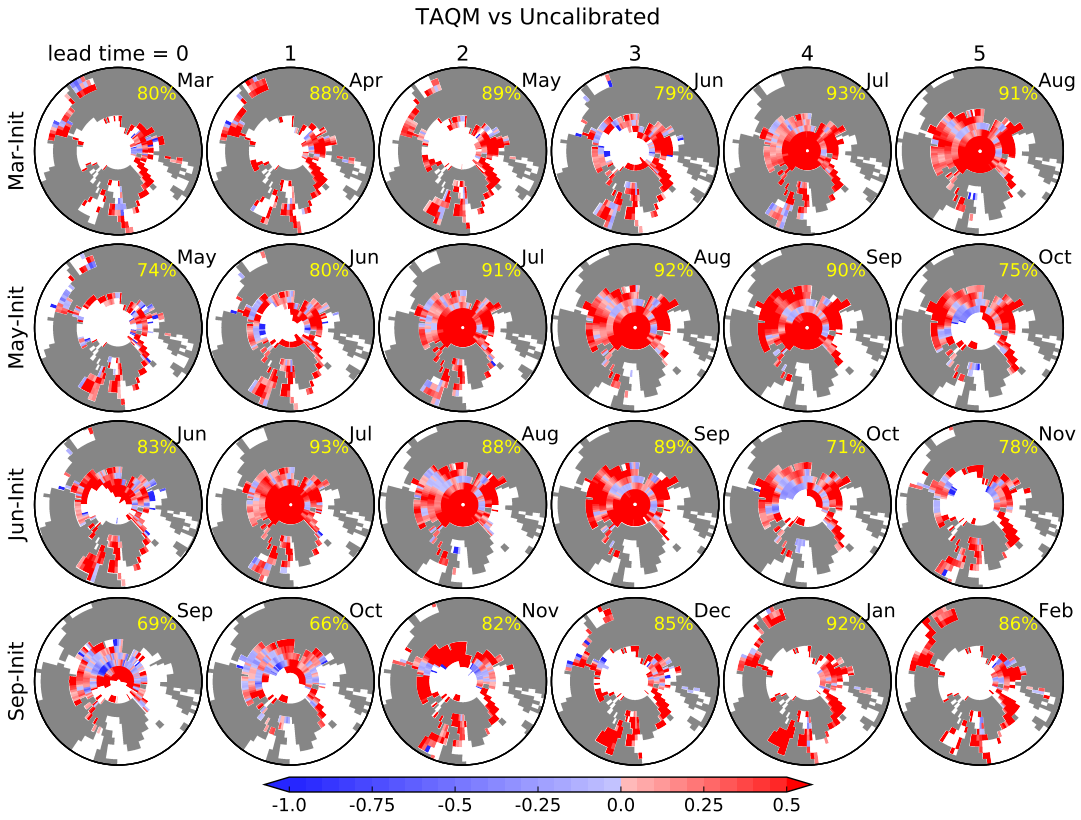

Figure 4.6: Spatial maps of the CRPSS, comparing the TAQM-calibrated hindcasts (forecast being evaluated) against the uncalibrated BEINF-fitted forecast distribution (reference forecast). Each row is for a different initialization month, and each column is for a different lead time increasing from left to right (as labelled). Improvement using the calibration method is indicated by positive (red) CRPSS values. Locations where the TAQM-calibrated hindcasts and the uncalibrated hindcasts have equal skill (i.e. where CRPSS= 0) are masked to white. The "percentage improved" (PI) values given in the top-right corner of each map are described in the main text.

tribution toward low SIC values and degrading skill for the remaining 12/13 hindcast years.

As indicated by the PI values (quoted in the upper part of each map), the improvement in hindcast skill for the TAQM-calibrated hindcasts generally increases after the first hindcast month, as to be expected since biases in the uncalibrated hindcasts also grow with increasing lead time (as the model drifts toward its own biased climatology). The greatest improvements in hindcast skill, as indicated by the largest PI values, are seen in July-September for hindcasts initialized in March, May, and June, and in January and February for hindcasts initialized in September.

Interestingly, the PI values for July and August are similar for all three initialization months, indicating that biases are large in these months. Generally, a lower improvement in skill for the TAQM-calibrated hindcasts relative to the uncalibrated hindcasts is seen during the transition seasons of spring and fall.

The particular regions showing the greatest improvement in probabilistic hindcast skill after applying TAQM are the central Arctic (where skill is now perfect, with CRPSS = 1), and in the eastern Arctic, particularly in the Greenland and Barents Seas where positive SIC biases in CanCM3 are large (Merryfield et al., 2013b; Dirkson et al., 2017). In August-October, substantially greater skill is also seen in the Chukchi and eastern Beaufort Sea, and along the Canadian and Alaskan coastlines.

4.7.2 TAQM vs 1981-2010 Climatology

A comparison of probabilistic hindcast skill between the TAQM-calibrated hindcasts and the 1981-2010 climatological distribution is shown in Fig. 4.7. Note that 1981-2010 climatology distribution does not include the hindcast year of interest so that results are cross-validated. Expansive areas of hindcast skill relative to climatology are seen over most of the Arctic, even for long lead times. As to be expected, the computed PI values show that skill generally decreases with increasing lead time relative to climatology, with the exception of the summer months when an increase in skill is observed.

A broad area of skill is present in the western Arctic in July-October for hindcasts initialized in March and May, and in June-October for hindcasts initialized in June. Skill is particularly high relative to climatology as the sea ice minimum extent is approached in September. High probabilistic hindcast skill relative to climatology is also present in the Nordic Seas in nearly all hindcasts months and for all initialization months, where, prior to TAQM-calibration, large biases are present. The Laptev Sea is also shown to be highly predictable in the transition seasons, particularly in October as sea ice expands southward in this region.

Based on the PI values presented in Fig. 4.7, hindcasts of winter and spring sea ice conditions are generally less skilful (relative to climatology) than during the summer and fall seasons for the Arctic as a whole.

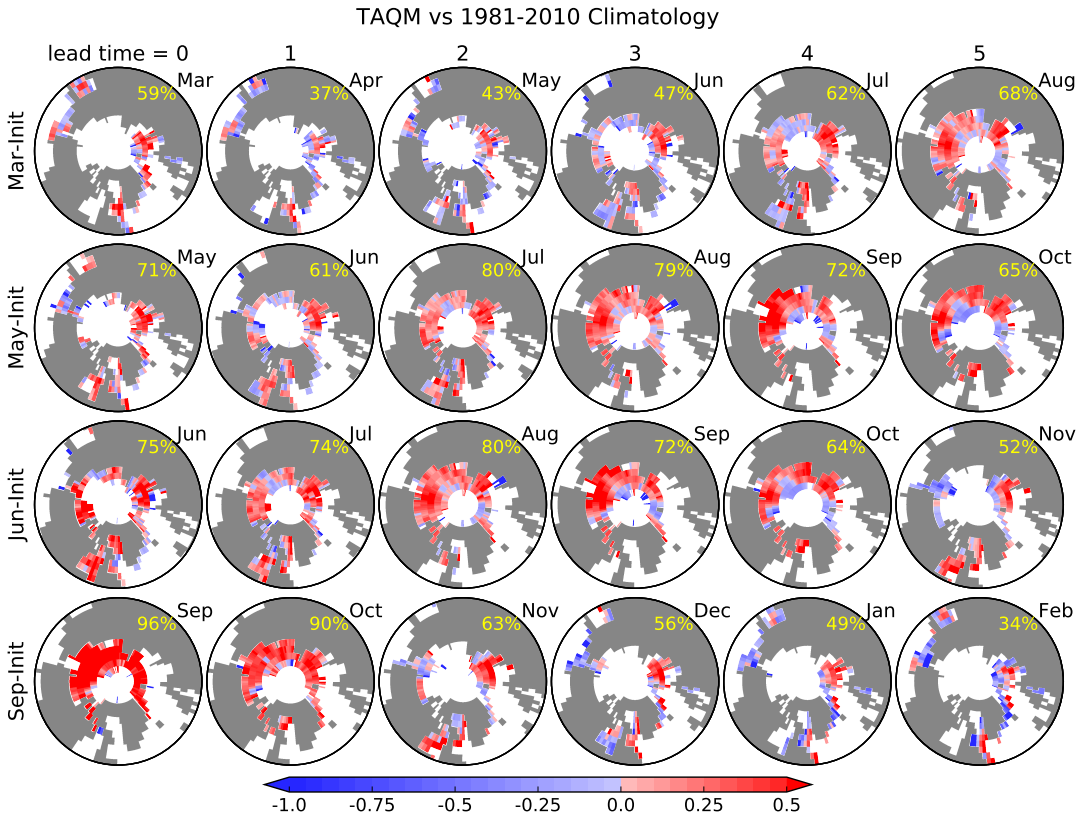

Figure 4.7: Same as in Fig. 4.6, but comparing the TAQM-calibrated hindcasts (forecast being evaluated) against the 1981-2010 climatological distribution (reference forecast).

4.7.3 TAQM vs TAOH Distribution

Comparisons of probabilistic hindcast skill between the TAQM-calibrated hindcasts and the TAOH distribution are presented in Fig. 4.8. To interpret these results, it is important to remember that the TAOH distribution is not climatology with a single overall trend subtracted, but a time-evolving distribution accounting for the trend through Eq. 4.10b. Probabilistic hindcast skill is generally much lower than when compared against the non-trend-adjusted 1981-2010 climatological distribution, except for at short lead times ($<$ 2 months) when skill is quite similar relative to both reference hindcasts. This suggests that a large contribution to hindcast skill for lead times longer than one month, when compared against the non-trend-adjusted climatology, is due to capturing the negative trends in SIC. Although skill decreases rapidly after a lead time of one month, probabilistic hindcast skill is still evident in all target months in some regions. Furthermore, similar regions show positive hindcast

skill relative to both climatological reference hindcasts.

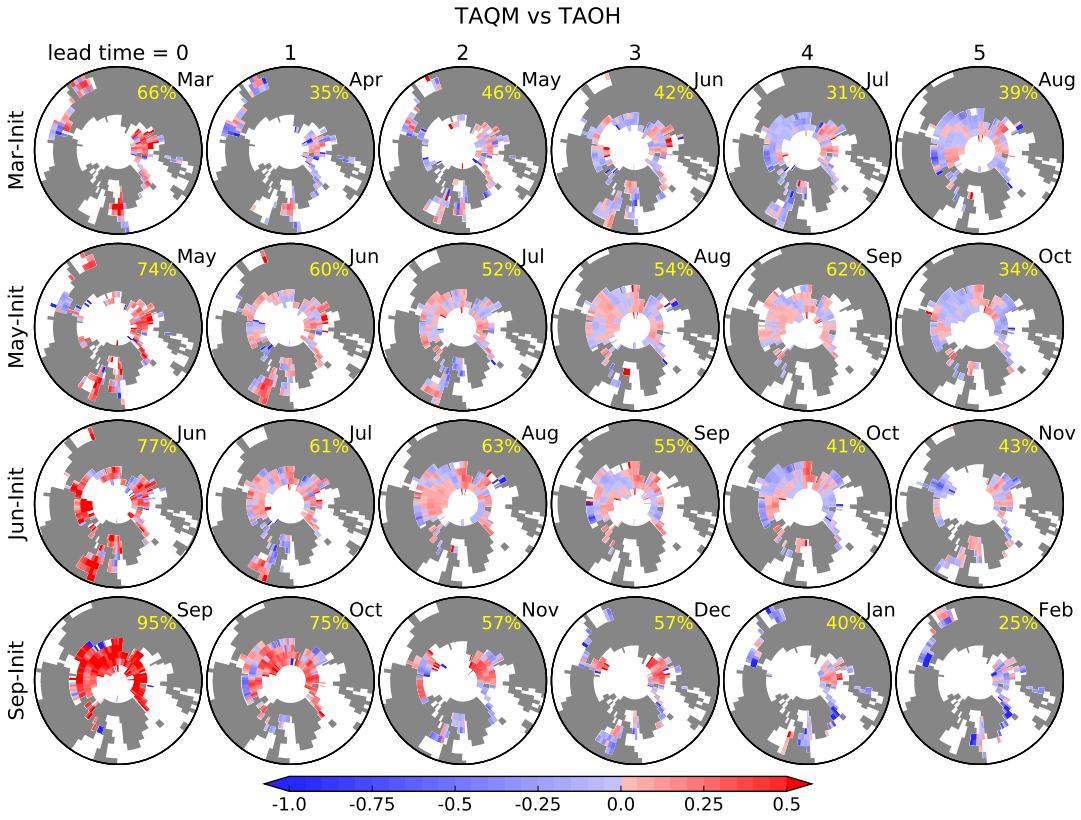

Figure 4.8: Same as in Fig. 4.6, but comparing the TAQM-calibrated hindcasts (forecast being evaluate) against the TAOH distribution (reference forecast).

For probabilistic hindcasts initialized in March, positive hindcast skill is generally confined to the eastern Arctic from May onward, with exceptions in the Labrador Sea and the northern Beaufort Sea. In the eastern Arctic, probabilistic hindcast skill is seen in the Kara and Barents Seas throughout most of the hindcast period. The positive skill in the Beaufort Sea expands northward in June through August in concert with the northward expansion of variable SIC conditions in that region.

The distributions of probabilistic hindcast skill for August and September conditions when initialized in May and June imply that little skill is gained by initializing in the later month. However, hindcast skill is generally higher during the transition seasons of spring and fall when initialized in June, as indicated by the increase in PI values of approximately 10%. Coherent and widespread probabilistic skill is seen in the Laptev Sea for both initialization months through October; however, skill is noticeably larger in magnitude in this region for hindcasts initialized in June.

For hindcasts initialized in September, probabilistic hindcast skill drops by about 30% within the first three hindcast months. Nevertheless, skill remains high in October-December in most regions. By January, hindcast skill is confined nearly entirely to the Barents Sea.

4.8 Conclusions

In this study, we introduced two methods intended to improve seasonal probability hindcasts of Arctic SIC. These methods have been tested in a set of ensemble-based dynamical hindcasts produced using CanCM3.

The first of these improves the representation of the forecast distribution by fitting a parametric distribution, namely the BEINF distribution, to SIC ensemble hindcasts. The BEINF distribution has been shown to be a reasonable model for ensemble SIC hindcasts according to an assessment on quality of fit. Generally, fitting SIC ensemble hindcasts to the BEINF distribution improves probabilistic skill relative to the simpler count method; however, model biases can degrade the skill improvement relative to the potential improvement suggested by pseudo-perfect model experiments.

The second of these improvements is the introduction of a novel calibration technique specifically designed for seasonal hindcasts of SIC (and real-time forecasts). The TAQM calibration method explicitly accounts for non-stationarity in the mean SIC state (i.e. trends), can be used with the BEINF distribution, and is implemented through the following steps applied to each grid location, initialization month and lead time:

1. Trends are removed from observed historical and model (hindcast) historical time series, excluding forecast year t, recentering on a non-stationary mean defined by the trend fit evaluated at year t. A piecewise-linear trend whose break point is 1999 was selected here in order to represent the acceleration of sea ice loss.

2. BEINF fits are applied to the forecast and trend-adjusted observed historical and model historical SIC distributions.

3. The parameters α_{x_t} and β_{x_t} of the beta portion of the forecast BEINF distribution are adjusted according to the quantile mapping of the forecast SIC values, from the beta portion of the trend-adjusted model historical distribution to the beta portion of the trend-adjusted observed historical distribution.

4. The parameters p_{x_t} and q_{x_t} representing the endpoints of the forecast BEINF distribution are adjusted according to a simple bias correction implied by the trend-adjusted model historical and trend-adjusted observed historical distributions.

Additional procedures, detailed in Appendix B.1, must be applied in special cases such as when only one ensemble member has SIC not equal to 0 or 1.

The TAQM calibration method has been shown to dramatically reduce model errors overall when compared against uncalibrated hindcasts, which can be quite large in CanCM3. Compared against the standard 1981-2010 climatological distribution reference hindcast, the TAQM-calibrated hindcasts are highly skilful in specific regions, even at long lead times. When evaluated against the trend-adjusted climatological distribution, the TAQM-calibrated hindcasts show positive skill in primarily the same regions as those for the non-trend-adjusted climatological distribution; however, skill scores are generally smaller in magnitude.

While the TAQM-calibrated hindcasts do not always outperform trend-adjusted climatology, particularly at longer lead times, much room for improvement of sea ice dynamical forecasts remains. These include for example resolution, reduction in model biases, initialization, and multi-model forecasting methodologies. However, the methods presented here provide a path toward maximizing the value of such forecasts in a probabilistic framework to quantify forecast uncertainty. Other statistical post-processing methods, such as skill score optimization and more sophisticated methods for accounting for non-stationarity in higher order statistical moments in the SIC data used in calibration, are potential topics of future research.

Chapter 5

'Modified CanSIPS' contribution to the 2017 Sea Ice Outlook

5.1 Introduction

Organized by the Study of Environmental Arctic Change (SEARCH), the Sea Ice Outlook (SIO) provides an opportunity for both scientists and citizens to contribute forecasts of September sea ice conditions using a range of statistical, heuristic, and dynamical model approaches (Stroeve et al., 2014b). The SIO was started in 2008, inspired in part by the (at-the-time) record low sea ice extent (SIE) of 4.2×10^6 square kilometers that occurred the previous September. The SIO was at first only concerned with forecasts of total Arctic sea ice extent (SIE). Since 2014, however, the SIO has included forecasts of *sea ice probability* (SIP) in its calls for contributions – defined as the probability that forecast local sea ice concentration (SIC) will exceed 15%.

Up until recently, a SIO contribution using the Canadian Seasonal to Interannual Prediction System (CanSIPS) (Merryfield et al., 2013a) had not been made due to large errors in its operational sea ice forecasts. While CanSIPS has shown some skill forecasting SIE in hindcasts (Sigmond et al., 2013), real-time sea ice forecasts issued by CanSIPS have performed poorly, forecasting SIE values that are much too large.

Three deficiencies have been identified that lead to these errors in operational sea ice forecasts using CanSIPS. First, trends in sea ice coverage in the Hadley Center Sea Ice and Sea Surface Temperatures, version 1.1 (HadISST1) observational dataset (Rayner et al., 2003) used to initialize SIC in hindcasts are unrealistically weak in

magnitude (Sigmond et al., 2013). Consequently, forecast SIE anomalies (computed relative to this hindcast climatology) in most recent years are unrealistically positive. Second, the Canadian Meteorological Centre (CMC) SIC product used operationally, which is based on synthetic aperture radar data, shows more expansive ice cover than passive microwave-based products (like HadISST1) used to initialize hindcast SIC. This results in an even larger positive bias in forecast SIE anomalies. Third, the climatological sea ice thickness (SIT) initialization method used in CanSIPS leads to poor forecast skill because it omits influences of the thinning of the ice pack over recent decades and does not capture interannual SIT anomalies (see chapter 3).

To address the first two deficiencies, a new set of CanSIPS hindcasts have has produced, initialized with a SIC product that has more realistic trends and is more consistent with the CMC product used to initialize SIC in real-time forecasts. This SIC product, termed Had2CIS, takes the maximum of the SIC values between the Hadley Center Sea Ice and Sea Surface Temperatures, version 2.2 (HadISST2) passive-microwave based product (Titchner and Rayner, 2014) and the Canadian Ice Service (CIS) product (Tivy et al., 2011). The HadISST2 product has more realistic trends than HadISST1, and the CIS product shows larger SIC values in the Canadian sector (which is consistent with the CMC product). This improved initialization consistency results in more realistic forecast anomalies. To remedy the third deficiency, the SMv3 model described in chapter 3 is used to replace the climatological SIT initialization in CanSIPS.

5.2 Outlook Results

CanSIPS was applied in a modified experimental mode, referred to as 'Modified CanSIPS', to generate sea ice forecasts for the 2017 SIO. CanSIPS combines forecasts from two models, CanCM3 and CanCM4, with a total of 20 ensemble members (10 from CanCM3, 10 from CanCM4). The forecasts for each SIO submission were initialized on the last day of May, June and July. For each submission month, our contribution included the initial condition fields of SIC and SIT, a forecast of mean September total Arctic SIE (with an estimate of uncertainty), and a spatial map of mean September SIP.

In Modified CanSIPS, SIC was initialized by nudging toward the Canadian Meteorological Centre (CMC) daily SIC analysis, exactly as in operational CanSIPS. The initialized SIT field, which was nudged toward climatological values in operations, was

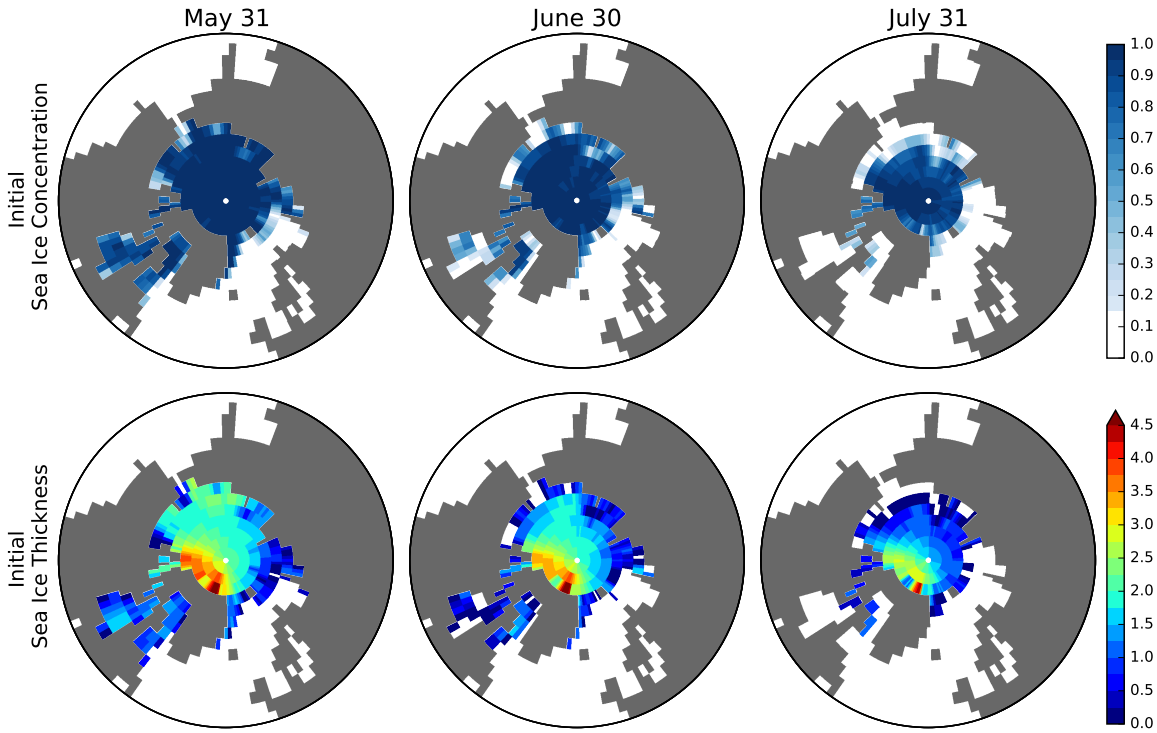

Figure 5.1: Sea ice concentration (top row) and sea ice thickness (bottom row) initialization fields (mean across ensemble members and CanCM3/CanCM4) for modified CanSIPS forecasts initialized on the last day of May, June and July 2017 (from left to right).

replaced by SIT from the SMv3 statistical model described in chapter 3, where the daily CMC SIC field was used as the real-time predictor in SMv3. The initialization fields for each submission month are shown in Fig. 5.1.

For each SIO contribution, the forecast September SIE anomaly was calculated for each individual ensemble member relative to the 1981-2010 climatology for the respective model. The 1981-2010 climatology for each model was estimated using the hindcasts described in section 5.1. These anomalies were then added to the NSIDC September climatological SIE value of 6.5 million square kilometers, and then averaged over all 20 ensemble members to yield the forecast total SIE. The forecast SIE for initializations on the last day of May, June and July are respectively 4.53 ± 0.72, 4.33 ± 0.58, and 4.45 ± 0.37 million square kilometers. The uncertainty in forecast total SIE was found by calculating the standard deviation of the ensemble of 20 forecast SIE anomalies, and multiplying by 1.96 to estimate the 95% range (between the 2.5 to 97.5 percentiles, assuming gaussianity) of the forecast distribution.

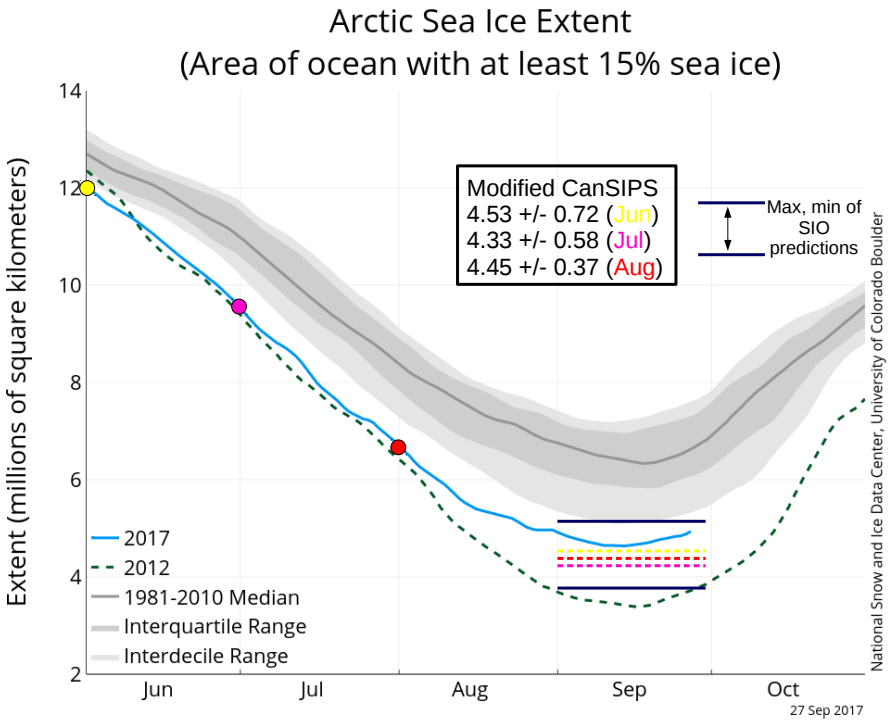

Figure 5.2: Total Arctic SIE: observed (light blue curve), Modified CanSIPS outlooks (as labeled). The circles denote the SIE values on the first of each initialization month. This figure was adapted from the original NSIDC figure (`http://nsidc.org/arcticseaicenews/`). The three dashed horizontal lines show the Modified CanSIPS forecast SIE (multi-model ensemble mean), and the solid dark blue horizontal lines show the minimum and maximum of the 95% confidence intervals of all three forecasts.

These forecast SIE values are plotted in Fig. 5.2., together with the observed SIE through September 27. The collection of all SIO contributions for the initialization month of June are shown in Fig. 5.3, where the Modified CanSIPS outlook of 4.53 million square kilometers is highlighted.

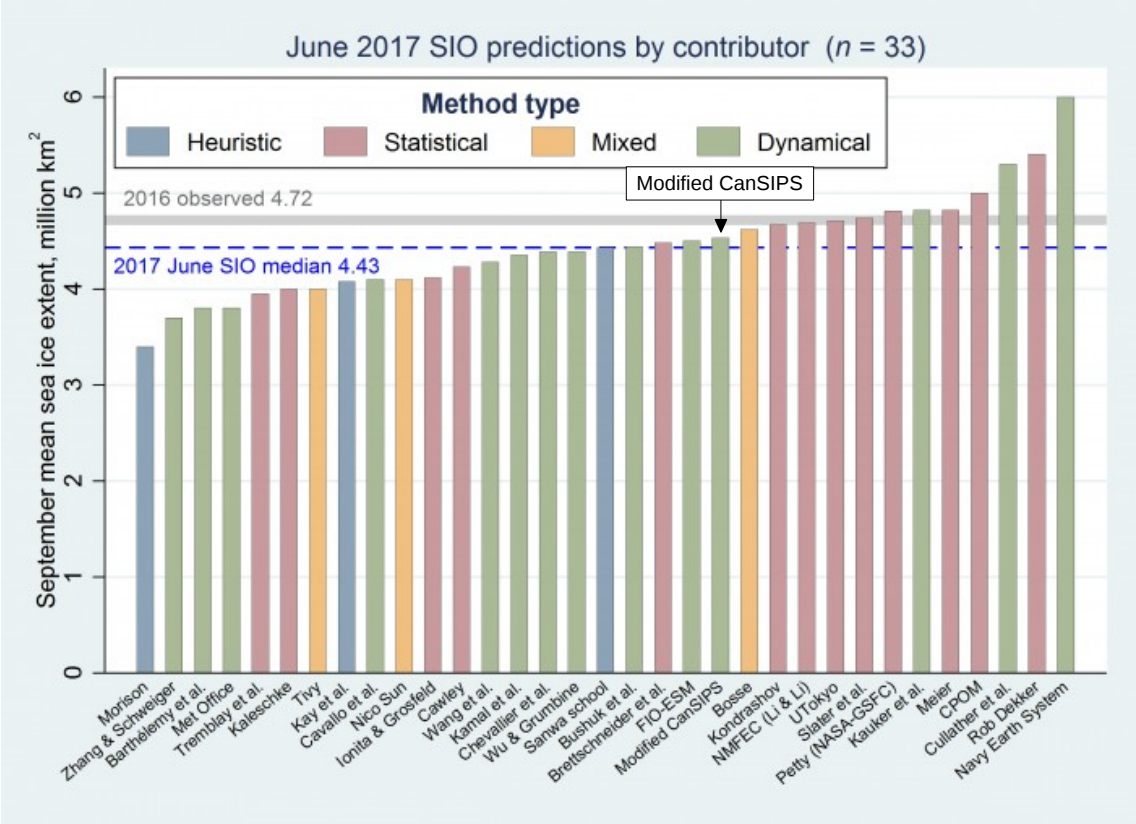

Figure 5.3: Sea ice outlooks for the initialization month of June from all contributors. The Modified CanSIPS outlook is highlighted. The original figure can be found at `https://www.arcus.org/sipn/sea-ice-outlook/2017/june`.

The uncertainty in spatially-distributed (local) SIE is provided in the maps of September SIP for the forecasts initialized in June, July, and August (Fig. 5.4). The observed 15% SIC contour for September is also plotted. To construct the SIP map for each monthly SIO submission, we first fit the 10-member ensemble SIC values from each model (per grid point) to the zero- and one-inflated beta distribution (Ospina and Ferrari, 2010), as described in chapter 4. After calibrating the parametric distribution per grid point and per model using trend-adjusted quantile mapping (TAQM) (as described in chapter 4), the probability that forecast SIC exceeds 15% was calculated from the calibrated parametric distribution. This illustrates the applicability of the methodology of chapter 4 to multi-model ensembles. The tendency in the SIP forecasts for the July (July minus June) and August (August minus July) are also

included in Fig. 5.4.

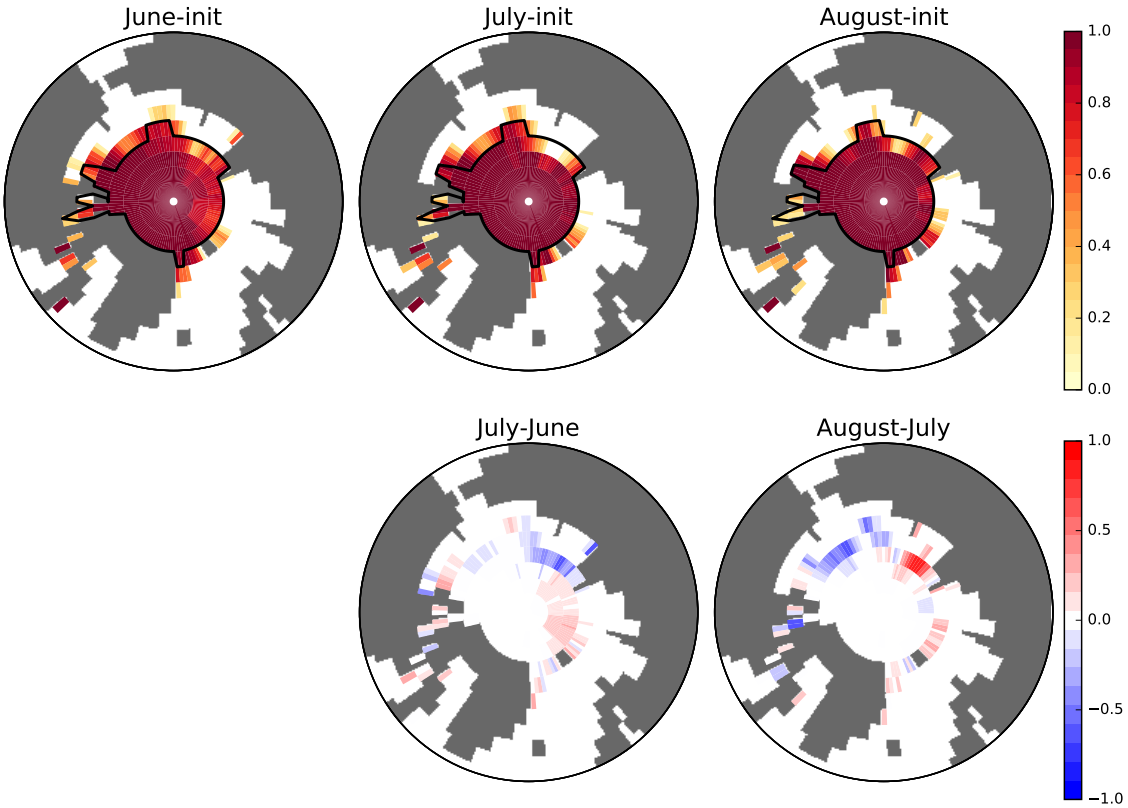

Figure 5.4: Sea ice probability forecasts using Modified CanSIPS (top row) for initializations on the last day of May, June, and July (as labelled). The observed 15% SIC contour is plotted in black. Tendencies in SIP are shown in the bottom row for July (July minus June) and August (August minus July).

An example of how the TAQM-calibration method modifies forecasts of SIP relative to uncalibrated forecasts of SIP is presented in Fig. 5.5. The observed 15% SIC contour for September is also plotted. Whereas the uncalibrated SIP forecasts produced using CanCM3 and CanCM4 are quite different from one another, the calibration method tends to produce a much more consistent forecast between the two models. This illustrates the efficacy of the calibration method to reduce individual model biases.

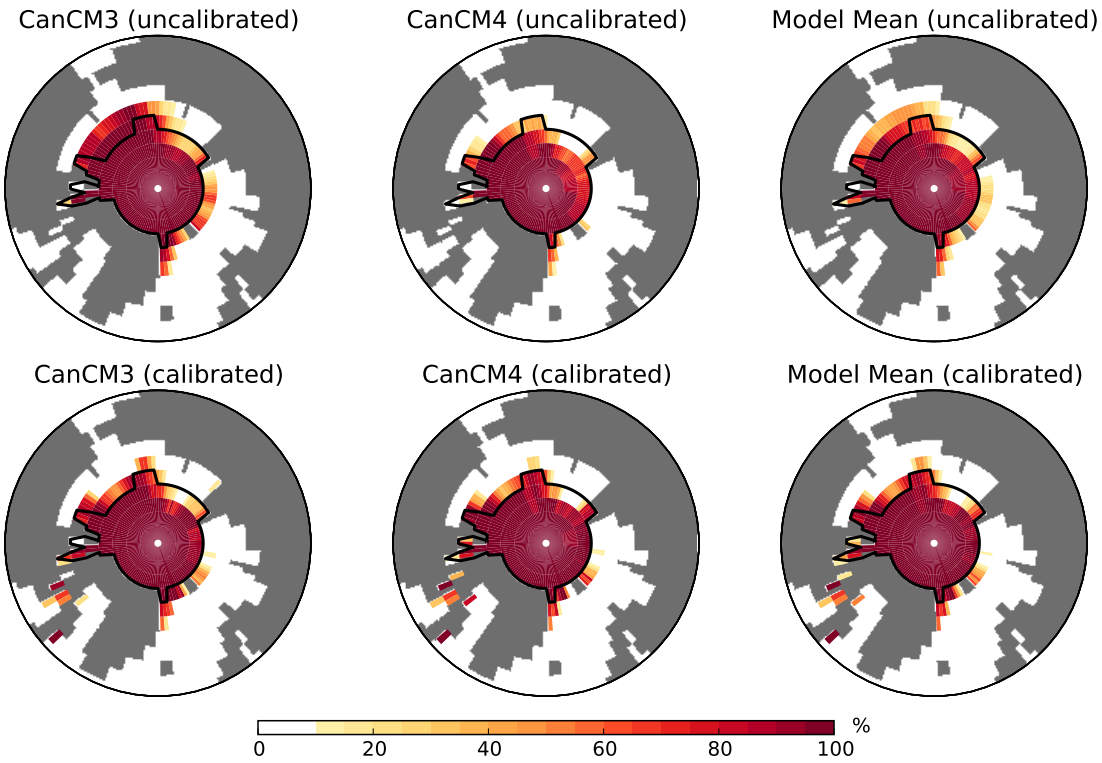

Figure 5.5: Sea ice probability forecasts using Modified CanSIPS initialized on the last day June. The individual model SIP forecasts for CanCM3 and CanCM4 are shown, as well as the multi-model mean SIP forecast (as labelled). The observed 15% SIC contour is plotted in black. The top row shows uncalibrated SIP forecasts and the bottom row shows the TAQM-calibrated SIP forecasts.

Chapter 6

Conclusions

Arctic sea ice stands out as a major component of the climate system, and is currently undergoing dramatic changes associated with a warming climate (e.g., Meier et al., 2014b). Such changes have spurred a rapidly-growing area of research on sea ice forecasting on seasonal time scales, in order to meet the needs of a number of stakeholders (Ellis and Brigham, 2009).

The research conducted in this dissertation addresses key challenges in forecasting Arctic sea ice conditions on seasonal time scales. First, I explored practical options for improving the sea ice thickness (SIT) initialization procedure employed in the Canadian Seasonal to Interannual Prediction System (CanSIPS), which currently consists of nudging SIT toward a model-based monthly-varying climatology. This work included the development of three statistical models for estimating SIT in real time. Second, each of these statistical models was tested in a set of hindcast experiments using the Canadian Climate Model Version 3 (CanCM3), and a thorough assessment of sea ice hindcast skill was carried out to ascertain the advantages of using each method relative to the current initialization method employed in CanSIPS. Third, I developed statistical post-processing methods for making "smoothed" and calibrated probability density functions from which probabilistic forecasts of local sea ice concentration (SIC) can be derived.

Sea ice thickness has been, and continues to be, sparsely observed in the Arctic. This provides a challenge for initializing seasonal forecasts of Arctic sea ice. To address this challenge, I developed three statistical models, SMv1, SMv2, and SMv3, that can be used to estimate SIT in real time from readily-available and physically-relevant SIT predictors. The statistical models were trained using the pan-Arctic Ice and Ocean Modelling and Assimilation System (PIOMAS) SIT (Zhang and Rothrock,

2003) (which is not available in real time) as a predictand, against which each model was evaluated. As shown in chapter 2, the first model, SMv1, which combines SIC and sea level pressure (SLP) data in an optimal way, improves on the current CanSIPS initialization scheme by 48%, in terms of its areal- and temporal- mean absolute error (ATMAE) (relative to PIOMAS). However, much of this improvement is owed to its ability to capture sea ice thinning over the historical record, whereas the statistical model has difficulty predicting interannual SIT anomalies. The SMv2 model described in chapter 3 builds on SMv1, taking SIC anomaly information into account directly, thereby improving the representation of interannual SIT anomalies. The third SIT statistical model, SMv3, was shown to be nearly as skilful as SMv2, but is considerably less complex, and does not require the bias correction needed in SMv1 and SMv2.

To assess the influence that each of these statistical models has on sea ice prediction skill, five sets of sea ice hindcasts were performed using the Canadian Climate Model version 3 (CanCM3), each initialized with a different SIT initialization method: SMv1, SMv2, SMv3, PIOMAS, and the current climatological initialization employed in CanSIPS (referred to as 'Original'). As described in chapter 3, all three statistical models improve sea ice prediction skill overall. However, only by initializing with SMv2 and SMv3 was hindcast skill similar to that found initializing with PIOMAS. When considering sea ice prediction skill of both pan-Arctic sea ice area (SIA) and local SIC, the three statistical models improved sea ice prediction skill in all forecast months. Much of this improvement was found to be due to the long-term trend. When interannual variations in SIA and SIC were evaluated for predictive skill, SMv1-initialized hindcasts were found to be no more skilful than those initialized with Original. For hindcasts initialized with SMv2 and SMv3 however, sea ice prediction skill is similar to hindcasts initialized with PIOMAS. Most notably, statistically significant skill predicting September SIA is now seen for hindcasts initialized on May 1st, whereas before, such skill was limited to predictions made from the 1st of June.

By considering the root mean square error (RMSE) of hindcast SIC anomalies, notably poorer skill was found in the summer months in the Nordic Seas in hindcasts initialized with the statistical models and PIOMAS (only SMv3 was shown), relative to hindcasts initialized with Original. This region is affected by a positive SIC bias and cold sea surface temperature (SST) bias in freely running simulations of CanCM3 (Merryfield et al., 2013a). An analysis of SIC, SIT, and SST revealed that this poorer skill is due thicker sea ice in PIOMAS in this region (relative to Original), which

amplifies both the positive SIC bias and cold SST bias already present in CanCM3.

The quantification of forecast uncertainty is an important step in the seasonal forecasting process (Troccoli et al., 2008) that has only begun to be explored in sea ice prediction studies. To improve probabilistic forecasts of regional sea ice coverage, namely SIC, I investigated the suitability of the zero- and one- inflated beta (BEINF) distribution (Ospina and Ferrari, 2010) for modelling SIC ensemble forecasts in chapter 4. In a pseudo-perfect model framework using CanCM3, fitting hindcast SIC ensembles to the BEINF distribution led to greater skill forecasting categorical SIC events, including sea ice probability (SIP) used in the annual Sea Ice Outlook (SIO). However, model errors were found to reduce this improvement when verifying these probabilistic hindcasts against observations.

Additionally in chapter 4, I developed a calibration method adapted from the well-known quantile mapping technique. This novel parametric-based calibration method, termed trend-adjusted quantile mapping (TAQM), is specifically designed for the BEINF distribution, and explicitly accounts for non-stationarities (i.e. trends) in the SIC historical record used in calibration. An assessment of probabilistic skill after applying TAQM revealed a dramatic reduction in model errors relative to uncalibrated probabilistic hindcasts. Through additional comparisons against the observed climatology and a more conservative trend-adjusted observed climatology, it was shown that high probabilistic skill can be obtained in specific regions, even at long lead times.

For the first time, CanSIPS has been used to produce Arctic sea ice forecasts for the annual SIO for the summer of 2017. As described in chapter 5, the SMv3 model was applied in real time to initialize SIT in these forecasts, and the probabilistic post-processing methods described in chapter 4 were applied to produce forecast maps of SIP.

Although the SMv3 statistical model for estimating SIT for real-time forecast initialization results in similar forecast skill as that initializing with PIOMAS (the SMv3 predictand), this may be the result of a lack of sensitivity in CanCM3 to small differences between SIT initialization methods. Future work could investigate whether similar skill can also be achieved using this method to initialize higher-resolution models with more sophisticated sea ice physics. Additionally, SMv3 performs poorly in the central Arctic in the winter and early spring when SIC is not variable (see chapter 3); the inclusion of a dynamical predictor in this sub-domain in those months could improve on SMv3. Furthermore, more advanced assimilation techniques that

allow for flexible input of data (e.g. point measurements or temporally inconsistent measurements) could be used to assimilate SIT data at times when it is readily available. While the inconsistencies of this approach could raise issues for bias correction, such trade-offs between initialization consistency and utilizing SIT observations could be assessed.

With respect to probabilistic forecasting of sea ice conditions, many possibilities remain to be explored. Further improvements may result from multi-model ensemble approaches, as well as skill score optimization. Additionally, while the TAQM calibration method explicitly accounts for trends in SIC, taking non-stationarities of higher-order moments into account (e.g. interannual variance) may lead to further improvements in calibration. Finally, methods for post-processing probabilistic forecasts of other sea ice quantities, such as SIT and sea ice retreat/advance dates, should also be explored.

Appendix A

Real-time estimation of Arctic sea ice thickness through maximum covariance analysis supplementary material

Figure A.1 shows the *areal- and temporal-mean absolute errors* (ATMAEs) for each combination of predictor and number of modes retained (CPM) (shuffled and unshuffled) referred to in section 2.3.3.

For Fig. A.3 and Table A.1, six reference predictions (RPs) are shown that are briefly noted in the main article in section 2.4.2 while developing the SM. For clarity, these are described in further detail here:

- SM pre-bias - The SM without applying the bias correction. This is equivalent to the use of climatology over 1981-1993 and equation 2.4 from chapter 2 over 1994-2012.
- Climatology - Mean SIT calculated over training period τ for month m.
- Persistence - SIT for month m and year $t_e - 1$.
- Extrapolation - A linear trend of SIT calculated for month m over training period τ, and extrapolated to year t_e.
- Extrap. + Pers. - Same as 'Extrapolation', with the anomalies relative to the trend at year $t_e - 1$ added to the extrapolated trend.

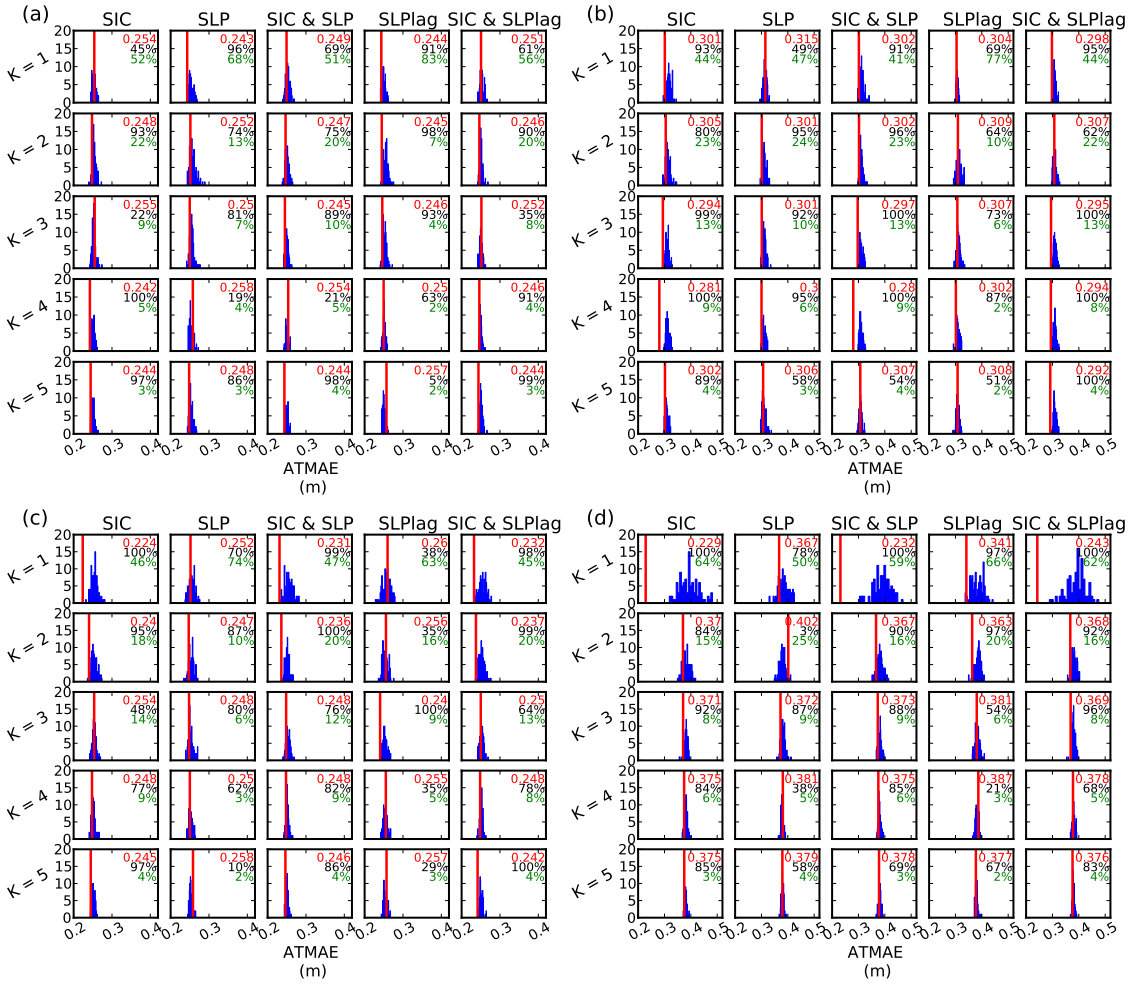

Figure A.1: As in Fig. 2.1 from the main article, but showing averages over the 1995-2003 (a,b) and 2004-2012 (c,d) sub-periods, for the months of March (a,c) and September (b,d).

As was stated in the main article in section 2.4.2, the development of the SM was motivated by a comparison between the blended predictor (equation 2.4) and the RPs, based on the ATMAEs summarized in Table A.1. This led to the use of climatology in the SM prior to 1994, as well as the bias correction. To complement Fig. 2.4 from the main article, we have also included Fig. A.2, which serves to provide an analogous interpretation of the SM against the RPs by considering SIV. The improvement in the SM compared to the RPs is less impressive for SIV relative to the ATMAEs for SIT directly, but because the SM is not designed to model SIV specifically, these results are not inconsistent with our overall assessment of the SM.

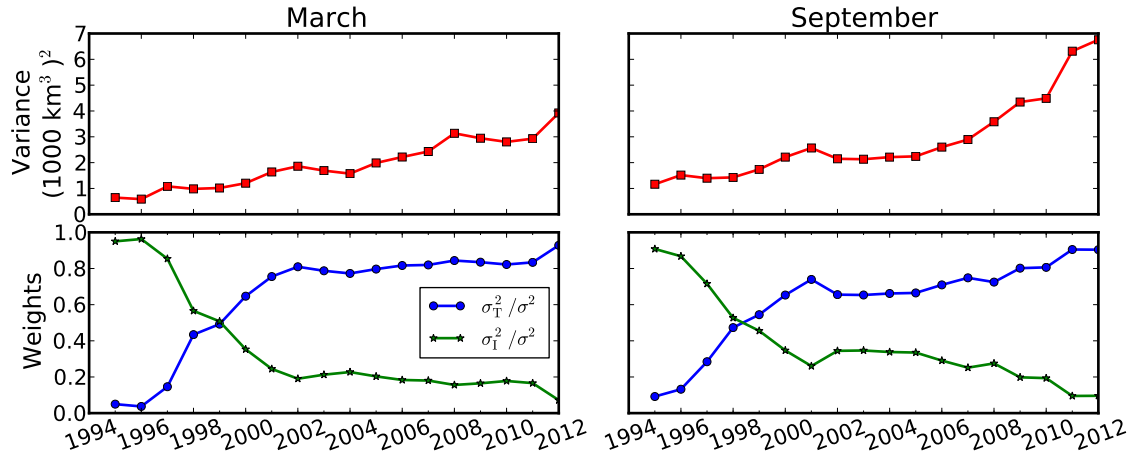

Figure A.2: SIV variance (σ^2) and weighting terms (σ_T^2/σ^2 and σ_I^2/σ^2) used in equation 2.4 in chapter 2 for the months of March (left) and September (right), given over the period 1994-2012. Each value is calculated using the time series of SIV over the training period τ.

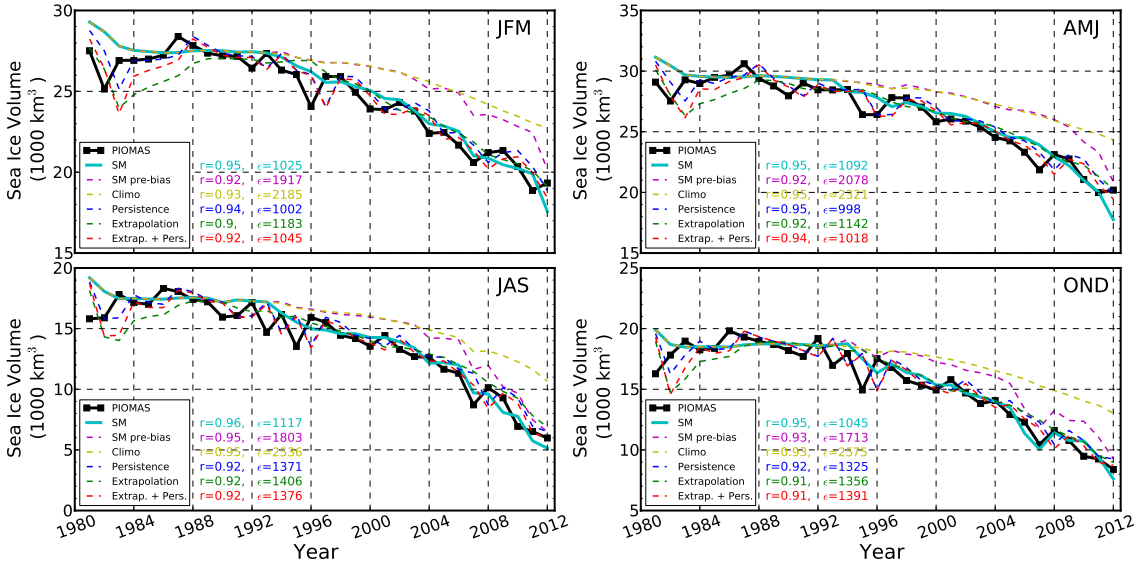

Figure A.3: Time series of three-month averaged SIV over the period 1981-2012, given in units of 10^3 km^3 for PIOMAS and the RPs. The corresponding correlation coefficients (r) and RMSEs (ϵ, in units of 10^3 km^3) are shown for reference.

Table A.1: ATMAEs given in meters, for the periods 1981-1993, 1994-2012, and 1981-2012 for the SM and the reference predictions (RPs).

Prediction	1981-1993	1994-2012	1981-2012
SM	0.314	0.264	0.289
SM pre-bias	0.314	0.278	0.296
Equation (4)	0.319	0.278	0.299
Climatology	0.314	0.324	0.319
Persistence	0.328	0.299	0.314
Extrapolation	0.370	0.280	0.325
Extrap. + Persist.	0.370	0.302	0.336

Appendix B

Calibrated Probabilistic Forecasts of Arctic Sea Ice Concentration Appendices

B.1 Estimating BEINF Parameters

B.1.1 Maximum Likelihood Estimation

The derivation of the ML estimates $\hat{p}$ and $\hat{q}$ for the BEINF distribution can be found in Ospina and Ferrari (2010). The ML estimate $\hat{p}$ is the fraction of zeros and ones in the sample, and of those values in the sample that are either zero or one, $\hat{q}$ is the fraction of ones. Their analytic solutions are given by $\hat{p} = \sum_i \mathbb{1}_{\{0,1\}}/n$ and $\hat{q} = \sum_i x_i \mathbb{1}_{\{0,1\}}/m$, where n is the size of the entire sample, $m = \sum_i \mathbb{1}_{\{0,1\}}$ is the number of zeros and ones in the sample, and $\hat{q} = 0/0$ is set to zero by convention.

The ML estimates $\hat{\alpha}$ and $\hat{\beta}$ for the BEINF distribution, are computed from the $n - m$ size sub-sample x_{sub}. ML estimation is carried out on x_{sub} using the "beta.fit" function in the Python's "scipy.stats" module. In this fitting algorithm, the roots of the gradient of the log-likelihood function for the beta distribution are solved for numerically, with starting values $\hat{\alpha}_0$ and $\hat{\beta}_0$ obtained by the method of moments:

$$\hat{\alpha}_0 = \bar{x}_{sub}\left(\frac{\bar{x}_{sub}(1-\bar{x}_{sub})}{\mathrm{var}(x_{sub})} - 1\right), \tag{B.1a}$$

$$\hat{\beta}_0 = (1-\bar{x}_{sub})\left(\frac{\bar{x}_{sub}(1-\bar{x}_{sub})}{\mathrm{var}(x_{sub})} - 1\right). \tag{B.1b}$$

In Eqs. B.1a–B.1b, $\bar{x}_{sub} = \frac{1}{n-m}\sum_k x_k$ is the geometric mean of x_{sub}, and $\text{var}(x_{sub}) = \frac{1}{n-m-1}\sum_k (x_k - \bar{x}_{sub})^2$ is the unbiased estimator of sample variance.

When the ML algorithm does not converge, $\hat{\alpha}$ and $\hat{\beta}$ are set to the starting values $\hat{\alpha}_0$ and $\hat{\beta}_0$ computed using the method of moments (Eqs. B.1a–B.1b).

B.1.2 Special fitting procedure for cases 2-4

The following describes the special fitting procedure that was introduced in section 4.6.2 in detail. This fitting method is used for making inference on α and β for the purpose of TAQM calibration when any of cases 2–4 (outlined in section 4.4.2) are encountered.

First, parameters $\hat{p}$ and $\hat{q}$ are computed using ML estimation. To estimate α and β, consider the sample mean of the size $n-m$ sub-sample z_{sub}, denoted as $\bar{z}_{sub}$, where z is any of x', y', or x_t. The population mean for the beta distribution, given population parameters α and β, is defined as $E(Z) = \alpha/(\alpha+\beta)$. The estimates $\hat{\alpha}$ and $\hat{\beta}$ are found by minimizing the cost function

$$c(\hat{\alpha}, \hat{\beta}; \bar{z}_{sub}) = \left(\bar{z}_{sub} - \frac{\hat{\alpha}}{\hat{\alpha}+\hat{\beta}} \right)^2, \tag{B.2}$$

subject to the asymmetry constraint $\hat{\alpha} < \hat{\beta}$, when $\hat{q_z} \leq 0.5$, and $\hat{\alpha} > \hat{\beta}$, when $\hat{q_z} > 0.5$. These constraints ensure that the beta distribution is positively (negatively) skewed when the majority of the remaining ensemble members are zero (one), which in turn ensures that the resultant distribution has density primarily concentrated around low (high) SIC values. Typically of course, q_z is either zero or one when the special fitting method is used; however, in some cases both zeros and ones are present in the data and $0 < q_z < 1$.

Initial guesses $\hat{\alpha}_0$ and $\hat{\beta}_0$ are chosen to be consistent with the asymmetry constraint and distribution shape desired:

Cases 2 and 3:

$q = 0$, set $(\hat{\alpha}_0, \hat{\beta}_0) = (1, 4)$

$q = 1$, set $(\hat{\alpha}_0, \hat{\beta}_0) = (4, 1)$

$0 < q \leq 0.5$, set $(\hat{\alpha}_0, \hat{\beta}_0) = (0.2, 0.4)$

$0.5 < q < 1$, set $(\hat{\alpha}_0, \hat{\beta}_0) = (0.4, 0.2)$

Case 4:

$$0 \leq q \leq 0.5, \text{ set } (\hat{\alpha}_0, \hat{\beta}_0) = (0.2, 0.4)$$

$$0.5 < q \leq 1, \text{ set } (\hat{\alpha}_0, \hat{\beta}_0) = (0.4, 0.2).$$

The minimization of Eq. B.2, subject to the constraint on distribution asymmetry and initial guesses $\hat{\alpha}_0$ and $\hat{\beta}_0$, is performed numerically using the "minimize" function in Python's "scipy.optimize" module.

Examples of SIC forecasts for cases 2-4, and the resultant BEINF distributions fitted using this special fitting procedure, are shown in Fig. B.1. As can be seen in all three cases, the resultant BEINF distributions model these data adequately for the calibration procedure. For case 2, the beta portion of the BEINF distribution is concentrated around low SIC values around the single ensemble member, and the population mean for the estimated beta distribution equals the sample mean. The same can be said for case 3, although in this case probability density is mainly confined to high SIC values around the two equal-valued ensemble members. In case 4, the U-shaped distribution clearly fits the SIC forecast well, as can be seen by the comparison of the BEINF cdf with the ecdf; however, the resultant population mean of this distribution is slightly different than the sample mean.

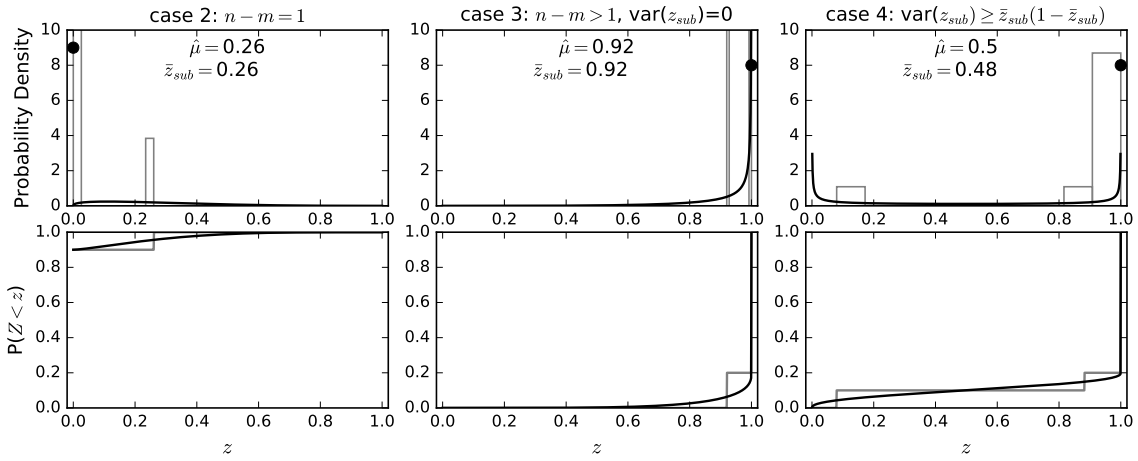

Figure B.1: Illustration of the special fitting method used when any of cases 2-4 (described in the main text) are encountered. Top row: normalized histogram distributions of z_{sub} and corresponding fitted BEINF pdfs (the probability of equalling zero or one is multiplied by 10 for easier comparison); the population mean $\hat{\mu}$ and sample mean $\bar{z}_{sub}$ are given on each panel. Bottom row: ecdfs of z_{sub} and corresponding fitted BEINF cdfs.

B.2 Goodness-of-fit tests

As described in section 4.4.2 in the main text, goodness of fit of the BEINF distribution is assessed per SIC ensemble hindcast. In particular, we test the null hypothesis H_0, that the hindcast SIC ensemble $x_{sub} \in (0, 1)$, comes from the population beta$(\hat{\alpha}, \hat{\beta})$, where $\hat{\alpha}$ and $\hat{\beta}$ are ML estimated parameters.

We implement three empirical distribution function (EDF) tests, as they are given in Stephens (1986): the Kolmogorov-Smirnov (KM) test, the Cramer Von-Mises (CVM) test, and the Anderson-Darling (AD) test. In general, EDF tests compare the ecdf of a given sample against the cdf of a parametric probability distribution, from which the sample used to compute the ecdf is assumed to be drawn from. These particular tests weight the discrepancies between the ecdf and cdf to different levels of scrutiny, and have been chosen here to serve as lower, middle, and higher classifications of "test power", based on the share of rejections these tests detect when H_0 is truly false (Raschke, 2010).

The critical values used here can be found in Table 4.7 (upper tail) in Stephens (1986), where they are stated for tests of normality. For the KM, CVM, and AD tests at the significance level $\alpha_s = 0.05$, the critical values are respectively 0.895, 0.126, and 0.752.

In order to test H_0 using these critical values for normality, we follow the procedure outlined in Raschke (2009, 2010), modified for the problem here:

- Compute ML estimates of $\hat{\alpha}$ and $\hat{\beta}$ from the sub-sample of hindcast values $x_1, ..., x_{n-m}$.
- Derive a new sample $y_1, y_2, ..., y_{n-m}$, where $Y \sim F^{-1}_{\text{snormal}}\left[F_{\text{beta}}(x_{sub}; \hat{\alpha}, \hat{\beta})\right]$; F^{-1}_{snormal} is the inverse cdf for the standard normal distribution, with mean $\mu = 0$ and standard deviation $\sigma = 1$.
- Compute ML estimates $\hat{\mu}$ and $\hat{\sigma}$ from sample $y_1, ...y_{n-m}$, for the distribution $F_{\text{normal}}(y; \hat{\mu}, \hat{\sigma})$; F_{normal} is the cdf for the generalized normal distribution with mean $\hat{\mu}$ and standard deviation $\hat{\sigma}$.
- Apply a particular EDF test for testing H_0: that the sample $y_1, ..., y_{n-m}$, comes from the distribution $F_{\text{normal}}(\hat{\mu}, \hat{\sigma})$.

B.3 Dependence of BSS on SIC-Event Threshold

As described in the main text, greatest improvements using the parametric method over the count method are expected to be seen in forecasting extreme quantiles (Wilks, 2002). To assess the dependence of the BSS on the extremity of quantiles being sampled, we use a quantile extremity index,

$$h(w) = \begin{cases} -2\ w + 1, & 0 \leq w \leq 0.5 \\ 2\ w - 1, & 0.5 < w \leq 1 \end{cases} \tag{B.3}$$

where $w = P(X > x_l)$ is the forecast probability for the event Ω. This index varies linearly from a value of 1 at the 0% quantile to a value of 0 at the median quantile, and then increases linearly again to a value of 1 at the 100% quantile. When the event is either very unlikely to occur or very likely to occur, the associated quantile is considered extreme and $h(w)$ is closer to 1. When the event has similar chances of occurring and not occurring, the associated quantile is less extreme and $h(w)$ is closer to 0. Thus, the parametric method is expected to yield greater improvements over the count method when $h(w)$ is close to 1, whereas it is expected to yield lesser improvements over the count method when $h(w)$ is close to 0. We use the probabilities estimated by the parametric method for the $w = P(X > x_l)$ values in Eq. B.3; however, results are qualitatively similar to using the count method probabilities instead.

We evaluate the dependence of the BSS on quantile extremity by correlating the BSS with temporal-mean $h(w)$, aggregating over all relevant grid locations (where $\mathrm{BS_{fcst}} \neq \mathrm{BS_{ref}}$) and SIC events for a given hindcast month. While the correlation coefficient is not necessarily an ideal metric for quantifying this dependence (since the BSS is asymmetric), it has been chosen to provide a low-level indication for this relationship. This correlation was found to be positive with $p < 0.001$ for 21/24 hindcast months (with correlations ranging from 0.065 to 0.33, with a mean of 0.23), and positive with $p > 0.001$ for the remaining three months. In general, the skill improvement of the parametric method over the count method is therefore higher when quantiles are more extreme (i.e. when events are either highly likely or highly unlikely to occur), as expected.

To illustrate how this result bears on the tendency for greater skill improvement at higher SIC thresholds (as shown in Fig. 4.3a), we present histogram distributions

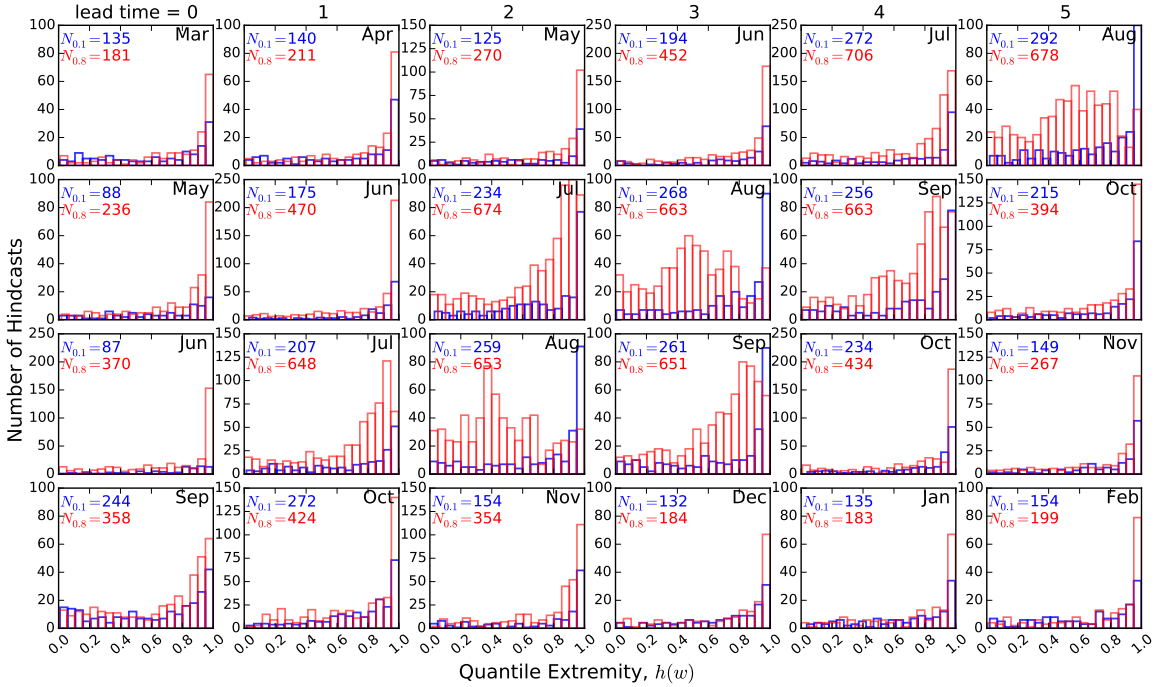

Figure B.2: Histograms of the quantile extremity values given by Eq. B.3 (main text), per initialization month and lead time, for the SIC-event thresholds $x_l = 0.1$ (blue) and $x_l = 0.8$ (red). Quantile extremity scales with increasing values on the horizontal axis. The number of hindcasts that contribute to the BSS values plotted in Fig. 4.3a for the respective event thresholds (per initialization month and lead time) are given by the values $N_{0.1}$ and $N_{0.8}$ in each panel.

for the quantity $h(w)$ for the event thresholds $x_l = 0.1$ and $x_l = 0.8$ in Fig. B.2. As expected, the histogram distributions show that extreme quantiles (where $h(w) \approx 1$) are sampled more frequently for hindcasts of the high-SIC event ($x_l = 0.8$), compared to the low-SIC event ($x_l = 0.1$). This increases the likelihood for the parametric method to outperform the count method for the high-SIC event, and suggests that quantile sampling is a factor in the greater skill improvement using the parametric method that is observed for the high-SIC event.

Additionally, the number of hindcasts sampled for the event threshold $x_l = 0.8$ for each forecast month ($N_{0.8}$ in Fig. B.2) are typically much larger than the number of hindcasts sampled for the event threshold $x_l = 0.1$ ($N_{0.1}$ in Fig. B.2). The fact that overwhelmingly $N_{0.8} > N_{0.1}$ further increases the likelihood that skill improvements relative to the count method are not obscured by sampling noise.

B.4 Quantile Mapping - Normal to Normal

Consider the normally-distributed random variable $X \sim N(\mu_m, \sigma_m)$, with cdf

$$F_m(x) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{x - \mu_m}{\sigma_m\sqrt{2}}\right)\right] \tag{B.4}$$

and the normally-distributed random variable $Y \sim (\mu_o, \sigma_o)$, with inverse cdf

$$F_o^{-1}(\gamma) = \mu_o + \sigma_o\sqrt{2}\ \text{erf}^{-1}(2\gamma - 1). \tag{B.5}$$

In Eqs. B.4–B.5, $\text{erf} = 1/\sqrt{\pi}\int_0^x e^{-t^2}dt$ is the "error function", and in Eq. B.5, γ is the probability associated with quantile $y = F_o^{-1}(\gamma)$. By substituting Eq. B.4 into Eq. B.5 for γ, Eq. 4.9 in the main text reduces to

$$\begin{aligned}
\hat{x}_t &= F_o^{-1}[F_m(x_t)], \\
&= \mu_o + \sigma_o\sqrt{2}\ \text{erf}^{-1}\left\{2\left[\frac{1}{2}\left(1 + \text{erf}\left(\frac{x_t - \mu_m}{\sigma_m\sqrt{2}}\right)\right)\right] - 1\right\}, \\
&= \mu_o + \sigma_o\sqrt{2}\ \text{erf}^{-1}\left[\text{erf}\left(\frac{x_t - \mu_m}{\sigma_m\sqrt{2}}\right)\right], \\
&= \mu_o + \sigma_o\left(\frac{x_t - \mu_m}{\sigma_m}\right).
\end{aligned} \tag{B.6}$$

Bibliography

Anderson, D. L. (1961). Growth rate of sea ice. *Journal of Glaciology*, 3(30):1170–1172.

Barnes, E. A., Dunn-Sigouin, E., Masato, G., and Woollings, T. (2014). Exploring recent trends in Northern Hemisphere blocking. *Geophysical Research Letters*, 41(2):638–644.

Bitz, C., Fyfe, J. C., and Flato, G. M. (2002). Sea ice response to wind forcing from AMIP models. *Journal of Climate*, 15(5):522–536.

Bitz, C. M., Holland, M. M., Hunke, E. C., and Moritz, R. E. (2005). Maintenance of the Sea-Ice Edge. *Journal of Climate*, 18(15):2903–2921.

Blanchard-Wrigglesworth, E., Armour, K. C., Bitz, C. M., and DeWeaver, E. (2011a). Persistence and Inherent Predictability of Arctic Sea Ice in a GCM Ensemble and Observations. *Journal of Climate*, 24(1):231–250.

Blanchard-Wrigglesworth, E., Barthélemy, A., Chevallier, M., Cullather, R., Fučkar, N., Massonnet, F., Posey, P., Wang, W., Zhang, J., Ardilouze, C., et al. (2016). Multi-model seasonal forecast of Arctic sea-ice: forecast uncertainty at pan-Arctic and regional scales. *Climate Dynamics*, pages 1–12.

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

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

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

Bretherton, C. S., Smith, C., and Wallace, J. M. (1992). An intercomparison of methods for finding coupled patterns in climate data. *Journal of climate*, 5(6):541–560.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3.

Chevallier, M. and 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. *Journal of Climate*, 25(8):3025–3038.

Chevallier, M., Smith, G. C., Dupont, F., Lemieux, J.-F., Forget, G., Fujii, Y., Hernandez, F., Msadek, R., Peterson, K. A., Storto, A., Toyoda, T., Valdivieso, M., Vernieres, G., Zuo, H., Balmaseda, M., Chang, Y.-S., Ferry, N., Garric, G., Haines, K., Keeley, S., Kovach, R. M., Kuragano, T., Masina, S., Tang, Y., Tsujino, H., and Wang, X. (2016). Intercomparison of the Arctic sea ice cover in global ocean–sea ice reanalyses from the ORA-IP project. *Climate Dynamics*, pages 1–30.

Chevallier, M., y Mélia, D. S., Voldoire, A., Dqu, M., and Garric, G. (2013). Seasonal Forecasts of the Pan-Arctic Sea Ice Extent Using a GCM-Based Seasonal Prediction System. *Journal of Climate*, 26(16):6092–6104.

Cohen, J., Screen, J. A., Furtado, J. C., Barlow, M., Whittleston, D., Coumou, D., Francis, J., Dethloff, K., Entekhabi, D., Overland, J., et al. (2014). Recent Arctic amplification and extreme mid-latitude weather. *Nature geoscience*, 7(9):627–637.

Collins, M., Botzet, M., Carril, A. F., Drange, H., Jouzeau, A., Latif, M., Masina, S., Otteraa, O. H., Pohlmann, H., Sorteberg, A., Sutton, R., and Terray, L. (2006). Interannual to decadal climate predictability in the north atlantic: A multimodel-ensemble study. *Journal of Climate*, 19(7):1195–1203.

Collow, T. W., Wang, W., Kumar, A., and Zhang, J. (2015). Improving Arctic Sea Ice Prediction Using PIOMAS Initial Sea Ice Thickness in a Coupled OceanAtmosphere Model. *Monthly Weather Review*, 143(11):4618–4630.

Comiso, J. C. (1986). Characteristics of Arctic winter sea ice from satellite multispectral microwave observations. *Journal of Geophysical Research: Oceans*, 91(C1):975–994.

Comiso, J. C., Meier, W. N., and Gersten, R. (2017). Variability and trends in the Arctic Sea ice cover: Results from different techniques. *Journal of Geophysical Research: Oceans*.

Comiso, J. C., Parkinson, C. L., Gersten, R., and Stock, L. (2008). Accelerated decline in the Arctic sea ice cover. *Geophysical research letters*, 35(1).

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

Dee, D., Uppala, S., Simmons, A., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M., Balsamo, G., Bauer, P., et al. (2011). The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, 137(656):553–597.

Director, H. M., Raftery, A. E., and Bitz, C. M. (2017). Improved Sea Ice Forecasting Through Spatiotemporal Bias Correction. *Journal of Climate*, (2017).

Dirkson, A., Merryfield, W. J., and Monahan, A. (2015). Real-time estimation of Arctic sea ice thickness through maximum covariance analysis. *Geophysical Research Letters*, 42(12):4869–4877. 2015GL063930.

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

Doblas-Reyes, F. J., García-Serrano, J., Lienert, F., Biescas, A. P., and Rodrigues, L. R. (2013). Seasonal climate predictability and forecasting: status and prospects. *Wiley Interdisciplinary Reviews: Climate Change*, 4(4):245–268.

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

Eicken, H. (2013). Ocean science: Arctic sea ice needs better forecasts. *Nature*, 497(7450):431–433.

Ellis, B. and Brigham, L. (2009). Arctic marine shipping assessment 2009 report.

Flato, G. M. and Hibler, W. D. (1992). Modeling Pack Ice as a Cavitating Fluid. *Journal of Physical Oceanography*, 22(6):626–651.

Francis, J. A., Chan, W., Leathers, D. J., Miller, J. R., and Veron, D. E. (2009). Winter Northern Hemisphere weather patterns remember summer Arctic sea-ice extent. *Geophysical Research Letters*, 36(7).

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

Gloersen, P. and Cavalieri, D. J. (1986). Reduction of weather effects in the calculation of sea ice concentration from microwave radiances. *Journal of Geophysical Research: Oceans*, 91(C3):3913–3919.

Gneiting, T., Raftery, A. E., Westveld III, A. H., and Goldman, T. (2005). Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, 133(5):1098–1118.

Gottschalk, L. and Weingartner, R. (1998). Distribution of peak flow derived from a distribution of rainfall volume and runoff coefficient, and a unit hydrograph. *Journal of hydrology*, 208(3):148–162.

Guemas, V., Blanchard-Wrigglesworth, E., Chevallier, M., Day, J. J., Dqu, M., Doblas-Reyes, F. J., Fukar, N. S., Germe, A., Hawkins, E., Keeley, S., Koenigk, T., Salas y Mlia, D., and Tietsche, S. (2016). A review on Arctic sea-ice predictability and prediction on seasonal to decadal time-scales. *Quarterly Journal of the Royal Meteorological Society*, 142(695):546–561.

Guemas, V., Doblas-Reyes, F. J., Mogensen, K., Keeley, S., and Tang, Y. (2013). Ensemble of sea ice initial conditions for interannual climate predictions. *Climate Dynamics*, pages 1–17.

Henderson-Sellers, A. (1978). Surface type and its effect upon cloud cover: a climatological investigation. *Journal of Geophysical Research: Oceans*, 83(C10):5057–5062.

Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 15(5):559–570.

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

Holland, M. M., Bitz, C. M., Hunke, E. C., Lipscomb, W. H., and Schramm, J. L. (2006). Influence of the sea ice thickness distribution on polar climate in CCSM3. *Journal of Climate*, 19(11):2398–2414.

Johnson, N. L., Kotz, S., and Balakrishnan, N. (1995). Continuous univariate distributions, vol. 2 of wiley series in probability and mathematical statistics: applied probability and statistics.

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Woollen, J., et al. (1996). The NCEP/NCAR 40-year reanalysis project. *Bulletin of the American meteorological Society*, 77(3):437–471.

Kapsch, M.-L., Graversen, R. G., and Tjernström, M. (2013). Springtime atmospheric energy transport and the control of Arctic summer sea-ice extent. *Nature Climate Change*, 3(8):744–748.

Kay, J. E., Holland, M. M., and Jahn, A. (2011). Inter-annual to multi-decadal Arctic sea ice extent trends in a warming world. *Geophysical Research Letters*, 38(15).

Kharin, V., Boer, G., Merryfield, W., Scinocca, J., and Lee, W.-S. (2012). Statistical adjustment of decadal predictions in a changing climate. *Geophysical Research Letters*, 39(19).

Kharin, V. V., Merryfield, W. J., Boer, G. J., and Lee, W.-S. (2017). A Postprocessing Method for Seasonal Forecasts Using Temporally and Spatially Smoothed Statistics. *Monthly Weather Review*, 145(9):3545–3561.

Kharin, V. V., Teng, Q., Zwiers, F. W., Boer, G. J., Derome, J., and Fontecilla, J. S. (2009). Skill assessment of seasonal hindcasts from the Canadian Historical Forecast Project. *Atmosphere-ocean*, 47(3):204–223.

Kharin, V. V. and Zwiers, F. W. (2003). Improved Seasonal Probability Forecasts. *Journal of Climate*, 16(11):1684–1701.

Krikken, F., Schmeits, M., Vlot, W., Guemas, V., and Hazeleger, W. (2016). Skill improvement of dynamical seasonal Arctic sea ice forecasts. *Geophysical Research Letters*, 43(10):5124–5132.

Krishnamurti, T., Kishtawal, C., LaRow, T. E., Bachiochi, D. R., Zhang, Z., Williford, C. E., Gadgil, S., and Surendran, S. (1999). Improved weather and seasonal climate forecasts from multimodel superensemble. *Science*, 285(5433):1548–1550.

Kurtz, N., Farrell, S., Studinger, M., Galin, N., Harbeck, J., Lindsay, R., Onana, V., Panzer, B., and Sonntag, J. (2012). Sea ice thickness, freeboard, and snow depth products from Operation IceBridge airborne data. *The Cryosphere Discussions*, 6:4771–4827.

Kwok, R. and Cunningham, G. (2008). ICESat over Arctic sea ice: Estimation of snow depth and ice thickness. *Journal of Geophysical Research: Oceans (1978–2012)*, 113(C8).

Kwok, R. and Rothrock, D. (2009). Decline in Arctic sea ice thickness from submarine and ICESat records: 1958–2008. *Geophysical Research Letters*, 36(15).

Laxon, S. W., Giles, K. A., Ridout, A. L., Wingham, D. J., Willatt, R., Cullen, R., Kwok, R., Schweiger, A., Zhang, J., Haas, C., et al. (2013). CryoSat-2 estimates of Arctic sea ice thickness and volume. *Geophysical Research Letters*, 40(4):732–737.

Li, B. and Avissar, R. (1994). The impact of spatial variability of land-surface characteristics on land-surface heat fluxes. *Journal of Climate*, 7(4):527–537.

Lindsay, R., Haas, C., Hendricks, S., Hunkeler, P., Kurtz, N., Paden, J., Panzer, B., Sonntag, J., Yungel, J., and Zhang, J. (2012). Seasonal forecasts of arctic sea ice initialized with observations of ice thickness. *Geophysical research letters*, 39(21).

Lindsay, R. and Schweiger, A. (2015). Arctic sea ice thickness loss determined using subsurface, aircraft, and satellite observations. *The Cryosphere*, 9(1):269–283.

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

Lisæter, A. K., Rosanova, J., and Evensen, G. (2003). Assimilation of ice concentration in a coupled ice–ocean model, using the Ensemble Kalman filter. *Ocean Dynamics*, 53(4):368–388.

Liu, J., Song, M., Horton, R. M., and Hu, Y. (2013). Reducing spread in climate model projections of a September ice-free Arctic. *Proceedings of the National Academy of Sciences*, 110(31):12571–12576.

Maslanik, J., Stroeve, J., Fowler, C., and Emery, W. (2011). Distribution and trends in Arctic sea ice age through spring 2011. *Geophysical Research Letters*, 38(13).

Meier, W., Peng, G., Scott, D., and Savoie, M. (2014a). Verification of a new NOAA/NSIDC passive microwave sea-ice concentration climate record. *Polar Research*, 33(0).

Meier, W. N., Hovelsrud, G. K., van Oort, B. E., Key, J. R., Kovacs, K. M., Michel, C., Haas, C., Granskog, M. A., Gerland, S., Perovich, D. K., Makshtas, A., and Reist, J. D. (2014b). Arctic sea ice in transformation: A review of recent observed changes and impacts on biology and human activity. *Reviews of Geophysics*, 52(3):185–217. 2013RG000431.

Meier, W. N., Peng, G., Scott, D. J., and Savoie, M. H. (2014c). Verification of a new noaa/nsidc passive microwave sea-ice concentration climate record. *Polar Research*, 33.

Meier, W. N., Stroeve, J., and Fetterer, F. (2007). Whither Arctic sea ice? A clear signal of decline regionally, seasonally and extending beyond the satellite record. *Annals of Glaciology*, 46(1):428–434.

Melia, N., Haines, K., and Hawkins, E. (2016). Sea ice decline and 21st century trans-Arctic shipping routes. *Geophysical Research Letters*, 43(18):9720–9728.

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

Merryfield, W. J., Lee, W.-S., Boer, G. J., Kharin, V. V., Scinocca, J. F., Flato, G. M., Ajayamohan, R., Fyfe, J. C., Tang, Y., and Polavarapu, S. (2013b). The Canadian seasonal to Interannual Prediction System. Part I: Models and initialization. *Monthly Weather Review*, 141(8):2910–2945.

Mori, M., Watanabe, M., Shiogama, H., Inoue, J., and Kimoto, M. (2014). Robust Arctic sea-ice influence on the frequent Eurasian cold winters in past decades. *Nature Geoscience*, 7(12):869–873.

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

NRC (2010). *Assessment of Intraseasonal to Interannual Climate Prediction and Predictability*. The National Academies Press.

Ospina, R. and Ferrari, S. L. (2010). Inflated beta distributions. *Statistical Papers*, 51(1):111.

Overland, J. E. and Wang, M. (2010). Large-scale atmospheric circulation changes are associated with the recent loss of Arctic sea ice. *Tellus A*, 62(1):1–9.

Palmer, T., Buizza, R., Doblas-Reyes, F., Jung, T., Leutbecher, M., Shutts, G., Steinheimer, M., and Weisheimer, A. (2009). Stochastic parametrization and model uncertainty. *ECMWF Tech. Memo*, 598:1–42.

Palmer, T., Doblas-Reyes, F., Hagedorn, R., Alessandri, A., Gualdi, S., Andersen, U., Feddersen, H., Cantelaube, P., Terres, J., Davey, M., et al. (2004). Development of a European multimodel ensemble system for seasonal-to-interannual prediction (DEMETER). *Bulletin of the American Meteorological Society*, 85(6):853–872.

Parkinson, C. L., Cavalieri, D. J., Gloersen, P., Zwally, H. J., and Comiso, J. C. (1999). Arctic sea ice extents, areas, and trends, 19781996. *Journal of Geophysical Research: Oceans*, 104(C9):20837–20856.

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

Porter, D. F., Cassano, J. J., and Serreze, M. C. (2012). Local and large-scale atmospheric responses to reduced Arctic sea ice and ocean warming in the WRF model. *Journal of Geophysical Research: Atmospheres (1984–2012)*, 117(D11).

Raschke, M. (2009). The biased transformation and its application in goodness-of-fit tests for the beta and gamma distribution. *Communications in Statistics-Simulation and Computation*, 38(9):1870–1890.

Raschke, M. (2010). Empirical behaviour of tests for the beta distribution and their application in environmental research. *Stochastic Environmental Research and Risk Assessment*, 25(1):79–89.

Rayner, N. A., Parker, D. E., Horton, E. B., Folland, C. K., Alexander, L. V., Rowell, D. P., Kent, E. C., and Kaplan, A. (2003). Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *Journal of Geophysical Research: Atmospheres*, 108(D14). 4407.

Reynolds, C. A., Webster, P. J., and Kalnay, E. (1994). Random error growth in NMC's global forecasts. *Monthly weather review*, 122(6):1281–1305.

Richardson, D. S. (2001). Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size. *Quarterly Journal of the Royal Meteorological Society*, 127(577):2473–2489.

Rigor, I. G. and Wallace, J. M. (2004). Variations in the age of Arctic sea-ice and summer sea-ice extent. *Geophysical Research Letters*, 31(9).

Rigor, I. G., Wallace, J. M., and Colony, R. L. (2002). Response of Sea Ice to the Arctic Oscillation. *Journal of Climate*, 15(18):2648–2663.

Rothrock, D. A., Yu, Y., and Maykut, G. A. (1999). Thinning of the Arctic sea-ice cover. *Geophysical Research Letters*, 26(23):3469–3472.

Roy, P., Laprise, R., and Gachon, P. (2016). Sampling errors of quantile estimations from finite samples of data. *arXiv preprint arXiv:1610.03458*.

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

Schweiger, A., Lindsay, R., Zhang, J., Steele, M., Stern, H., and Kwok, R. (2011). Uncertainty in modeled Arctic sea ice volume. *Journal of Geophysical Research: Oceans (1978–2012)*, 116(C8).

Screen, J. A. and Simmonds, I. (2010). The central role of diminishing sea ice in recent Arctic temperature amplification. *Nature*, 464(7293):1334–1337.

Serreze, M. C., Holland, M. M., and Stroeve, J. (2007). Perspectives on the Arctic's shrinking sea-ice cover. *science*, 315(5818):1533–1536.

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

Simmonds, I. and Keay, K. (2009). Extraordinary September Arctic sea ice reductions and their relationships with storm behavior over 1979–2008. *Geophysical Research Letters*, 36(19).

Stephens, M. (1986). Tests based on EDF statistics. In *Goodness-of-fit Techniques*, volume 68 of *Statistics, textbooks and monographs*. New York: Dekker.

Stroeve, J., Barrett, A., Serreze, M., and Schweiger, A. (2014a). Using records from submarine, aircraft and satellites to evaluate climate model simulations of Arctic sea ice thickness. *The Cryosphere*, 8(5):1839–1854.

Stroeve, J., Hamilton, L. C., Bitz, C. M., and Blanchard-Wrigglesworth, E. (2014b). Predicting September sea ice: Ensemble skill of the SEARCH Sea Ice Outlook 20082013. *Geophysical Research Letters*, 41(7):2411–2418.

Stroeve, J. C., Kattsov, V., Barrett, A., Serreze, M., Pavlova, T., Holland, M., and Meier, W. N. (2012a). Trends in Arctic sea ice extent from CMIP5, CMIP3 and observations. *Geophysical Research Letters*, 39(16). L16502.

Stroeve, J. C., Serreze, M. C., Holland, M. M., Kay, J. E., Malanik, J., and Barrett, A. P. (2012b). The Arctics rapidly shrinking sea ice cover: a research synthesis. *Climatic Change*, 110(3):1005–1027.

Swart, N. C., Fyfe, J. C., Hawkins, E., Kay, J. E., and Jahn, A. (2015). Influence of internal variability on Arctic sea-ice trends. *Nature Climate Change*, 5(2):86.

Thorndike, A. S. and Colony, R. (1982). Sea ice motion in response to geostrophic winds. *Journal of Geophysical Research: Oceans*, 87(C8):5845–5852.

Tietsche, S., Day, J. J., Guemas, V., Hurlin, W. J., Keeley, S. P. E., Matei, D., Msadek, R., Collins, M., and Hawkins, E. (2014). Seasonal to interannual Arctic sea ice predictability in current global climate models. *Geophysical Research Letters*, 41(3):1035–1043.

Tietsche, S., Notz, D., Jungclaus, J. H., and Marotzke, J. (2013). Assimilation of sea-ice concentration in a global climate model physical and statistical aspects. *Ocean Science*, 9(1):19–36.

Titchner, H. A. and Rayner, N. A. (2014). The Met Office Hadley Centre sea ice and sea surface temperature data set, version 2: 1. Sea ice concentrations. *Journal of Geophysical Research: Atmospheres*, 119(6):2864–2889.

Tivy, A., Howell, S. E., Alt, B., McCourt, S., Chagnon, R., Crocker, G., Carrieres, T., and Yackel, J. J. (2011). Trends and variability in summer sea ice cover in the canadian arctic based on the canadian ice service digital archive, 1960–2008 and 1968–2008. *Journal of Geophysical Research: Oceans*, 116(C3).

Tompkins, A. M. (2002). A prognostic parameterization for the subgrid-scale variability of water vapor and clouds in large-scale models and its use to diagnose cloud cover. *Journal of the atmospheric sciences*, 59(12):1917–1942.

Troccoli, A., Harrison, M., Anderson, D. L., and Mason, S. J. (2008). *Seasonal climate: forecasting and managing risk*, volume 82. Springer Science & Business Media.

Uppala, S. M., Kllberg, P. W., Simmons, A. J., Andrae, U., Bechtold, V. D. C., Fiorino, M., Gibson, J. K., Haseler, J., Hernandez, A., Kelly, G. A., Li, X., Onogi, K., Saarinen, S., Sokka, N., Allan, R. P., Andersson, E., Arpe, K., Balmaseda, M. A., Beljaars, A. C. M., Berg, L. V. D., Bidlot, J., Bormann, N., Caires, S., Chevallier, F., Dethof, A., Dragosavac, M., Fisher, M., Fuentes, M., Hagemann, S., Hlm, E., Hoskins, B. J., Isaksen, L., Janssen, P. A. E. M., Jenne, R., Mcnally, A. P., Mahfouf, J.-F., Morcrette, J.-J., Rayner, N. A., Saunders, R. W., Simon, P., Sterl, A., Trenberth, K. E., Untch, A., Vasiljevic, D., Viterbo, P., and Woollen, J. (2005). The ERA-40 re-analysis. *Quarterly Journal of the Royal Meteorological Society*, 131(612):2961–3012.

Von Storch, H. and Zwiers, F. W. (2001). *Statistical analysis in climate research*. Cambridge university press.

Wang, W., Chen, M., and Kumar, A. (2013). Seasonal Prediction of Arctic Sea Ice Extent from a Coupled Dynamical Forecast System. *Monthly Weather Review*, 141(4):1375–1394.

Weigel, A. P., Liniger, M. A., and Appenzeller, C. (2008). Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts? *Quarterly Journal of the Royal Meteorological Society*, 134(630):241–260.

Wilks, D. S. (2002). Smoothing forecast ensembles with fitted probability distributions. *Quarterly Journal of the Royal Meteorological Society*, 128(586):2821–2836.

Wilks, D. S. (2011). *Statistical methods in the atmospheric sciences*, volume 100. Academic press.

Wingham, D., Francis, C., Baker, S., Bouzinac, C., Brockley, D., Cullen, R., de Chateau-Thierry, P., Laxon, S., Mallow, U., Mavrocordatos, C., et al. (2006). CryoSat: A mission to determine the fluctuations in Earth's land and marine ice fields. *Advances in Space Research*, 37(4):841–871.

Yao, A. Y. (1974). A statistical model for the surface relative humidity. *Journal of Applied Meteorology*, 13(1):17–21.

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