Calibrated Probabilistic Forecasts of Arctic Sea Ice Concentration

ARLAN DIRKSON

*Centre pour l'étude et la simulation du climat à l'échelle régionale, Université du Québec à Montréal,
Montreal, Quebec, Canada*

WILLIAM J. MERRYFIELD

*Canadian Centre for Climate Modelling and Analysis, Environment and Climate Change Canada, Victoria,
British Columbia, Canada*

ADAM H. MONAHAN

School of Earth and Ocean Sciences, University of Victoria, Victoria, British Columbia, Canada

(Manuscript received 14 April 2018, in final form 21 November 2018)

ABSTRACT

Seasonal forecasts of Arctic sea ice using dynamical models are inherently uncertain and so are best communicated in terms of probabilities. Here, we describe novel statistical postprocessing methodologies intended to improve ensemble-based probabilistic forecasts 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 distribution. The second improvement is the introduction of a novel calibration technique, termed trend-adjusted quantile mapping (TAQM), that explicitly takes into account SIC trends and is applied using the BEINF distribution. We demonstrate these methods using a set of 10-member ensemble SIC hindcasts from the Third Generation Canadian Climate Coupled Global Climate Model (CanCM3) over the period 1981–2017. 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, it does not itself correct model biases that may reduce this improvement when verifying against observations. The TAQM calibration technique is effective at removing SIC biases present in CanCM3 and improving forecast reliability. Over the recent 2000–17 period, TAQM-calibrated SIC hindcasts show improved skill relative to uncalibrated hindcasts. Compared against a climatological reference forecast adjusted for the trend, TAQM-calibrated hindcasts show widespread skill, particularly in September, even at 3–4-month lead times.

1. 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 $-13.0\% \pm 2.4\% \text{ decade}^{-1}$ in September (Fetterer

et al. 2017), 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 led 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 affected 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 from the chaotic nature of the climate system, which causes minute differences in initial conditions to amplify over time (e.g., Reynolds

Denotes content that is immediately available upon publication as open access.

Supplemental information related to this paper is available at the Journals Online website: <https://doi.org/10.1175/JCLI-D-18-0224.s1>.

Corresponding author: Arlan Dirkson, arlan.dirkson@gmail.com

DOI: 10.1175/JCLI-D-18-0224.1

© 2019 American Meteorological Society. For information regarding reuse of this content and general copyright information, consult the AMS Copyright Policy (www.ametsoc.org/PUBSReuseLicenses).

et al. 1994), and from model errors that stem from the incomplete representation and numerical approximation of the physical laws that drive climate variability.

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, postprocessing 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).

Model errors have the effect of degrading forecast reliability (Palmer et al. 2004), so that forecast probabilities for categorical events disagree with their observed frequencies. Further, ensemble forecasts are often underdispersive (i.e., overconfident), so that the mean-square error of the ensemble mean grows faster than the ensemble spread (Gneiting et al. 2005). Model errors and underdispersion 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. 2017b), the use of multiple models (e.g., Krishnamurti et al. 1999; Weigel et al. 2008; Merryfield et al. 2013a), the calibration of multimodel ensembles (e.g., Kharin et al. 2009), and stochastic parameterization 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; Stroeve et al. 2015; Blanchard-Wrigglesworth et al. 2017). While the 15% SIC threshold is commonly used to delineate the sea ice edge when estimating SIC from passive microwave satellites, other event thresholds may be relevant for different forecast end users.

One recent approach to account for errors in seasonal sea ice forecasts includes a novel nonstationary bias correction method that aims to correct the position of the forecast ensemble-mean ice edge (Director et al. 2017); this methodology has yet to be extended to a probabilistic framework. Krikken et al. (2016) used extended logistic regression (ELR; Wilks 2009) and heteroscedastic ELR (HELRL; Messner et al. 2014) to calibrate probabilistic forecasts of pan-Arctic and regional sea ice area; however, their methodology describes integrated measures of sea ice, which do not characterize sea ice coverage at the local gridcell level.

Extending these approaches to gridcell SIC would require accounting for the bounded nature of the SIC variable, which is not represented in the ELR and HELRL formulations. To reduce errors in decadal forecasts of pan-Arctic SIE, Fučkar et al. (2014) developed a regression method that conditions the posterior adjustment of the forecast on the initial conditions.

This study introduces a new methodology for improving seasonal probability forecasts of local SIC from ensemble forecasts generated with an AOGCM. Its aim is 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 that explicitly accounts for the observed trends in SIC.

In section 2, we briefly describe the model and hindcast experiments used to test this methodology, as well as the metrics used to evaluate probabilistic hindcast skill. Two methods for computing SIC forecast event probabilities are described in section 3: the discrete counting approach and a parametric approach. A skill comparison of these two methods applied to SIC hindcasts is presented in section 4. The calibration technique is introduced in section 5, and in section 6, we evaluate attributes of the probabilistic hindcasts, including skill, after calibration. Conclusions are presented in section 7.

2. Data and skill scores

a. Hindcasts

The methods introduced here are tested on a set of hindcasts from the Canadian Centre for Climate Modelling and Analysis (CCCma) Third Generation Canadian Coupled Global Climate Model (CanCM3; Merryfield et al. 2013a). 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 100-km nominal horizontal grid spacing and 40 vertical levels spaced at 10 m near the surface. Sea ice is modeled as a cavitating fluid with a single layer thickness category (Flato and Hibler 1992).

The hindcast experiments considered here are initialized at the start of March, May, June, and September, extend 6 months, and cover the 37-yr period 1981–2017. Each hindcast has 10 ensemble members, initialized

from slightly different initial conditions obtained from separate assimilation runs. As described in detail in [Merryfield et al. \(2013a\)](#), each assimilation run is nudged toward the same observation-based values but starts from a different initial state, resulting in ensemble spread in the assimilation run states and hence in the forecast initial conditions. Atmospheric variables are nudged with a 1-day time constant and sea surface temperatures and SIC with a 3-day time constant. Over the period 1981–2012, SIC is nudged toward a blend of SIC values from the Hadley Centre Sea Ice and Sea Surface Temperature dataset, version 2 (HadISST2; [Titchner and Rayner 2014](#)) and from digitized Canadian Ice Service (CIS) charts ([Tivy et al. 2011](#)). The purpose of using this blended product is to improve consistency between 1981 and 2012 sea ice initial conditions and those in 2013–17, when SIC is initialized by nudging toward the analysis employed for real-time predictions, which assimilates various data sources including CIS charts ([Buehner et al. 2013a,b, 2015](#)). Over the entire period 1981–2017, mean gridcell sea ice thickness (SIT) values are obtained from the “SMv3” statistical model described in [Dirkson et al. \(2017\)](#). The HadISST2 that employs ice-chart-based bias corrections to SIC values derived from passive microwave measurements is used to assess probabilistic hindcast skill. To show how the results presented might depend on the observational product used, we repeat relevant analyses using a second observational dataset, specifically the NOAA/NSIDC Climate Data Record of Passive Microwave Sea Ice Concentration, version 3 (NOAA/NSIDC CDRv3; [Meier et al. 2017](#); [Peng et al. 2013](#)), the results of which are highlighted in the conclusions ([section 7](#)).

b. Skill scores

We use two metrics to assess 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 binary event (e.g., SIP), whereas the CRPS compares the entire forecast probability distribution against the observed outcome.

Let $P_{f,jk}(\Omega)$ and $P_{o,jk}(\Omega)$ be the respective forecast probability and observed probability for the event Ω for a given year j and gridcell k . The observed probability is defined such that $P_{o,jk}(\Omega) = 1$ if the event is observed and $P_{o,jk}(\Omega) = 0$ if it is not. To simplify notation, it is always implied that we are considering a particular initialization month/forecast month pair (e.g., a July forecast initialized in March), without denoting it explicitly. The BS for a particular year and grid cell is defined by

$$BS_{jk}(\Omega) = \left[P_{f,jk}(\Omega) - P_{o,jk}(\Omega) \right]^2. \quad (1)$$

Similarly, let $F_{f,jk}(x)$ and $F_{o,jk}(x)$ be the respective cumulative distribution functions (cdfs) for forecast and observed SIC. When applied to a variable that takes values on the interval $[0, 1]$, such as SIC, the CRPS can be written

$$CRPS_{jk} = \int_0^1 \left[F_{f,jk}(x) - F_{o,jk}(x) \right]^2 dx, \quad (2)$$

where $F_{o,jk}(x) = H(x - y_{o,jk})$ is the Heaviside function, increasing discontinuously from zero to one at the observed SIC value $y_{o,jk}$.

The BS and CRPS are here defined on the unit interval and are negatively oriented. Perfect skill is achieved when $BS = 0$ or $CRPS = 0$. The only way for the BS to be zero is when the forecast probability is 100% and the event occurs, or the forecast probability is 0% and the event does not occur, whereas the only way for the CRPS to be zero is if the forecast distribution is perfectly sharp and accurate. This is only likely to occur for seasonal ensemble forecasts of SIC if a grid cell is completely ice covered or ice free.

To compare the probabilistic forecast skill of two forecasting methods, we consider the skill score (SS) metric $SS = 1 - S_{fct}/S_{ref}$, where SS represents either the Brier skill score (BSS) or the continuous rank probability skill score (CRPSS). The subscript “fct” refers to the forecast being evaluated relative to a reference forecast (e.g., a climatology), denoted by the subscript “ref.” The reference forecasts used will be defined in subsequent sections. The SS is defined on the interval $(-\infty, 1]$ and is positive when the forecast being evaluated is more skillful than the reference forecast. In the case that $S_{fct} = S_{ref} = 0$, the SS is set to zero rather than the undefined value $1 - 0/0$. In the case that $S_{ref} = 0$ but $S_{fct} \neq 0$, the SS is set to $-\infty$ (as it is represented in the NumPy library of Python).

Throughout this study, we will consider averages of SS over both time and space. Because the SS is asymmetric, all averaging is performed on the individual skill metrics S_{fct} and S_{ref} prior to computing the SS. We denote averaging by expressing the subscript(s) being averaged over in bold. The temporal average of the skill metric S_{jk} (representing either BS_{jk} or $CRPS_{jk}$) is defined $\bar{S}_{jk} = 1/M_j \sum_{j=1}^{M_j} S_{jk}$, $j = 1, \dots, M_j$. The spatio-temporal average, weighted by the gridcell area, is defined

$$S_{\mathbf{jk}} = \frac{1}{M_j} \sum_{j=1}^{M_j} \left(\frac{1}{W} \sum_{k=1}^{M_k} w_k S_{jk} \right), \quad (3)$$

where $W = \sum_{k=1}^{M_k} w_k$, $k = 1, \dots, M_k$, is the sum of the individual ocean gridcell areas poleward of the maximum ice extent for either CanCM3 or the observations over the analysis period (per longitude, per initialization month, per target month). When S represents the BS, the quantity inside the parentheses of (3) is proportional to the discrete approximation of the “spatial probability score” described in Goessling and Jung (2018), only differing by a factor W^{-1} .

3. Probability estimates

This section compares two approaches for computing SIC probabilities from raw ensemble model output of SIC. The *count method* obtains forecast event probabilities from raw forecast values, whereas 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 mitigate the effects of sampling on the estimates 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 probability density over the range of the undersampled variable within the bounds of the lowest and highest ensemble member and extrapolating density outside of the ensemble range (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 modeling the underlying forecast distribution and that the fit to the discrete forecast ensemble is satisfactory.

Throughout the following, we use $P(\Omega)$ to denote the probability for the event Ω , defined as the random variable X (in this case representing forecast SIC) exceeding a threshold value γ . For instance, by choosing the particular SIC threshold $\gamma = 0.15$, $P(\Omega)$ is equivalent to the SIP quantity described above.

a. Count method

The probability of the event Ω can be computed very simply using the count method. This method does not assume a distribution for X ; rather, it consists of counting

the number of ensemble members that satisfy the event criterion and reporting this relative frequency as the event probability. The $P(\Omega)$ computed by the count method is thus

$$P(X > \gamma) = 1 - F_n(\gamma), \quad (4)$$

where $F_n(x)$ is the empirical cumulative distribution function (ecdf) defined by

$$F_n(x) = \begin{cases} 0, & x < x_1 \\ i/n, & x_i \leq x < x_{i+1}, \quad i = 1, \dots, n-1, \\ 1 & x_n \leq x \end{cases} \quad (5)$$

computed from the forecast ensemble members x_i ($i = 1, \dots, n$) in ascending order.

While the count method suffers from large sampling errors, particularly for small ensemble sizes, it continues to be used in some probabilistic forecasting applications. For example, it is the method contributors have been advised to use when submitting SIP forecasts to the annual SIO (<https://www.arcus.org/sipn/sea-ice-outlook>).

b. Parametric method

SIC event probabilities can alternatively be computed by fitting an appropriate parametric distribution to the forecast ensemble. For the statistical modeling of doubly bounded random variables such as SIC, the beta distribution stands out as a natural candidate.

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

$$f_{\text{beta}}(x; \alpha, \beta) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1}, \quad 0 \leq x \leq 1, \quad (6)$$

where $B(\alpha, \beta) = \int_0^1 t^{\alpha-1} (1-t)^{\beta-1} dt$ is the beta function, and $\alpha > 0$ and $\beta > 0$ describe the shape of the distribution. Increasing α has the effect of shifting f_{beta} toward zero, whereas increasing β shifts the distribution toward one. When $\alpha < 1$ ($\beta < 1$), (6) is not defined at $x = 0$ ($x = 1$). 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 bounded by zero and one. However, the beta distribution cannot account for finite probability of a variable equaling exactly zero or one, as is often the case for ensemble SIC forecasts.

An attempt was made to apply the beta distribution to ensemble SIC forecasts despite this limitation. To fit the beta distribution using maximum likelihood (ML) estimation to SIC data that contain zeros or ones, the data were transformed from the interval $[0, 1]$ to $[\varepsilon, 1 - \varepsilon]$ by adding ε to zeros and subtracting ε from ones, where ε was tested on a range of small values. Following this transformation, there were numerous cases in which the ML algorithm failed to converge, particularly for smaller ε values. Additionally, for those cases when convergence did occur, the resultant fits were often unsatisfactory. Consequently, these practical issues restrict the applicability of the beta distribution for SIC forecasts.

As an alternative, we use a modified version of the beta distribution, 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 combines the continuous beta distribution with the degenerate Bernoulli distribution.

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

$$f(x; \alpha, \beta, p, q) = p(1 - q)\delta(x) + (1 - p)f_{\text{beta}}(x; \alpha, \beta) + pq\delta(x - 1), \quad (7)$$

where $\delta(x)$ is the Dirac delta function, and the parameter $0 \leq p \leq 1$ corresponds to the probability of X falling exactly at the end points of 0 or 1. Probability density at the endpoints is modeled by a Bernoulli distribution (scaled by p), defined by a single parameter $0 \leq q \leq 1$, such that the random variable takes an end-point value $X = 1$ with probability q and $X = 0$ with probability $1 - q$. On the open interval $(0, 1)$, the probability density is modeled by the beta distribution, as defined by (6), scaled by $1 - p$.

The cdf for the BEINF distribution is

$$F(x; \alpha, \beta, p, q) = pF_{\text{bern}}(x; q) + (1 - p)F_{\text{beta}}(x; \alpha, \beta), \quad (8)$$

where

$$F_{\text{bern}}(x; q) = \begin{cases} 0, & x < 0 \\ 1 - q, & 0 \leq x < 1 \\ 1, & x \geq 1 \end{cases} \quad (9)$$

is the cdf for the Bernoulli distribution, and $F_{\text{beta}}(x; \alpha, \beta) = \int_0^x f_{\text{beta}}(x'; \alpha, \beta) dx'$ is the cdf for the beta distribution. For the BEINF distribution, the probability of equaling exactly zero is $P(x = 0) = p(1 - q)$, the probability of being between zero and one is

FIG. 1. The cdf for the BEINF distribution. The jump discontinuities (circles) at $x = 0$ and $x = 1$ are expressed in terms of the parameters p and q for an example where $p, q > 0$. The beta portion of the BEINF distribution (black line) is shown simply for illustrative purposes, with parameter values $\alpha = 1.5$ and $\beta = 2.5$.

$P(0 < x < 1) = 1 - p$, and the probability of equaling exactly one is $P(x = 1) = pq$. These probabilities are illustrated graphically in Fig. 1 for an example where $p, q > 0$.

The four parameters that describe the shape of the BEINF distribution are estimated for each SIC ensemble hindcast, made up of members $x_1, x_2, \dots, x_n$ (not necessarily ordered), using ML estimation. As described in the appendix, 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 $x_1, \dots, x_{n-m}$ (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} . Note that if the true population has $p \neq 0$, but no zeros or ones exist in the finite sample, then the estimates of p and q will be biased. Similarly, if $p \neq 1$ for the true population, but the sample contains no values on the interval $(0, 1)$, estimates of α and β will be biased.

In the infrequent instances where the ML estimation algorithm does not converge, the method of moments (MOM) is used to estimate α and β (see the appendix). The MOM for the beta distribution requires that $\text{var}(x_{\text{sub}}) < \bar{x}_{\text{sub}}(1 - \bar{x}_{\text{sub}})$, where $\bar{x}_{\text{sub}}$ is the geometric mean of x_{sub} , and $\text{var}(x_{\text{sub}})$ is the unbiased estimator of sample variance, since $\text{var}(x_{\text{sub}}) \geq \bar{x}_{\text{sub}}(1 - \bar{x}_{\text{sub}})$ would result in $\alpha, \beta < 0$.

There are special cases to consider when α and β cannot be estimated by either ML estimation or by MOM. These cases are as follows:

FIG. 2. SIC ensemble hindcasts for six model grid cells spanning the Arctic Ocean from regions indicated in August at a lead time of two months. (top) 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) The ecdf for the hindcast ensemble and corresponding fitted BEINF cdf.

- Case 1: $n - m = 0$ (all ensemble members take the value of 0 or 1)
- Case 2: $n - m = 1$ (only one ensemble member takes a value other than 0 or 1)
- Case 3: $n - m > 1$, but $\text{var}(x_{\text{sub}}) = 0$
- Case 4: the ML estimation algorithm does not converge, and the MOM condition that $\text{var}(x_{\text{sub}}) < \bar{x}_{\text{sub}}(1 - \bar{x}_{\text{sub}})$ is not met

Of 741 924 instances considered ($37 \text{ years} \times 4 \text{ initialization months} \times 6 \text{ forecast months} \times \text{number of relevant grid cells}$), case 1 occurs 21.6% of the time (either because the grid cell is completely ice covered or is completely ice free). In the remaining 581 826 instances considered, case 2 occurs 8.9% of the time, case 3 occurs 0.003% of the time, and case 4 occurs 0.024% of the time. The 529 788 instances that remain can be fit to the BEINF distribution. When cases 1–4 are encountered, we simply revert to the count method for computing SIC probabilities.

To illustrate the properties of the BEINF distribution, Fig. 2 shows count-based distributions and corresponding fitted BEINF distributions for six ensemble hindcasts in August 2002, at a lead time of two months. These examples have been chosen to illustrate the wide range of distribution shapes that SIC ensemble forecasts can take and to provide an indication of the suitability of the BEINF distribution for modeling SIC ensemble forecasts.

Figure 2 demonstrates how the BEINF distribution fills in probability density across the gaps of 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 the bimodality in the Beaufort and Laptev Seas examples.

A visual comparison of the BEINF cdf and the ecdf for each of the six cases in Fig. 2 provides further evidence that the BEINF distribution is suitable for SIC ensemble forecasts. To quantify the suitability of the BEINF distribution for modeling SIC, we perform goodness-of-fit tests on each of the instances for which it is applicable (i.e., excluding cases 1–4 described above).

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

The Anderson–Darling (AD) test, an empirical distribution function (EDF) test (Stephens 1986) that measures the error between the ecdf of the sample and cdf of the best-fit beta distribution, is applied to each SIC ensemble hindcast to test H_0 at significance level α_s . To apply this test to the beta distribution, we follow the approach recommended in Raschke (2009, 2011). While the sample size of each hindcast ensemble (≤ 10) can decrease the statistical power of the AD test, the AD

FIG. 3. The CRPSS_{jk} values showing the percent increase (positive values) or percent decrease (negative values) in skill of the parametric method (forecast being evaluated) relative to the count method (reference forecast); blue circles: PPM experiments, orange circles: OV experiments. Vertical lines are the 5th–95th percent confidence intervals of the CRPSS_{jk} values. Each panel is for a different initialization month (as labeled).

test has been shown to be the most powerful of several EDF tests for assessing the goodness of fit of the beta distribution (Raschke 2011).

Based on the AD test, we conclude that H_0 can be rejected for 19% of the SIC ensemble hindcasts at the significance level $\alpha_s = 0.05$. This percentage of rejections indicates that while we can conclude that SIC hindcast values are not strictly drawn from the BEINF distribution, the agreement is sufficiently good that this distribution might be useful in practice.

4. Probabilistic hindcast skill: Count versus 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 2a 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 nine ensemble members. In the OV experiments, hindcast probabilities estimated by the count and parametric methods are computed from nine randomly drawn ensemble members (as in the PPM experiments), but 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 and observational uncertainty, either from the initial conditions or from the model itself, and including no direct knowledge of true forecast uncertainty (Wilks 2002).

a. CRPSS evaluation

The comparison of probabilistic hindcast skill using the count and parametric methods is first assessed by the CRPSS_{jk} , with the parametric method as the forecast being evaluated and the count method as the reference forecast. We compute (2) via the trapezoidal approximation for both $\text{CRPS}_{\text{fctst}}$ and CRPS_{ref} , as the cdf for the BEINF distribution cannot be computed exactly.

In both the PPM and OV experiments, the parametric method outperforms the count method, as indicated by the consistently positive CRPSS_{jk} values in Fig. 3. 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. In general, CRPSS_{jk} values are larger for the PPM experiments than the OV experiments, although the values remain consistently positive for the latter. This result implies that while model and observational errors can degrade the potential skill improvement indicated by the PPM experiments, skill is nonetheless improved by fitting to the BEINF distribution in both cases.

Uncertainty in CRPSS_{jk} values is indicated by the 5th and 95th percent confidence intervals in Fig. 3,

FIG. 4. The BSS_{jk} values showing the percent increase (positive values) or percent decrease (negative values) in skill of the parametric method (forecast being evaluated) relative to 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 labeled).

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 late spring and summer forecast months in the PPM experiments. The improvement in skill is not statistically significant for other individual months in the PPM experiments or for any individual months in the OV experiments. Nonetheless, the consistently positive $CRPSS_{jk}$ values in both the PPM and OV experiments provide evidence for the robustness of the improvement using the parametric method.

b. BSS evaluation

As probabilistic forecasts of specific SIC events are of greatest practical interest, we now use the BSS_{jk} to compare probabilistic hindcast skill between the count and parametric methods. The BSS_{jk} values are

computed for several different SIC threshold events Ω , in which the SIC threshold γ is varied from 0.1 to 0.9 in increments of 0.1. Note that because Ω is binary, inspection of (1) reveals that $BS(\Omega) = BS(\Omega^c)$, where Ω^c is the complement of the event Ω ; thus, the BS for the event that $X > \gamma$ is the same as that for the event $X \leq \gamma$. The BSS_{jk} values for the PPM and OV experiments are shown in Fig. 4.

Like the CRPSS results described above, the BSS_{jk} values for the PPM experiments are nearly always positive, which demonstrates that greater skill is achieved using the parametric method over the count method (Fig. 4a). 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 an overall improvement in forecast skill using

the parametric method relative to the count method (Fig. 4b). However, for some low-SIC and high-SIC 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, such a reduction in skill occurs for low- to midrange-SIC event thresholds. The largest improvement using the parametric method in the OV experiments is seen for midrange event thresholds during the summer forecast months of July–September for initializations in March, May, and June.

The dependence of skill improvement using the parametric method on the SIC event threshold was examined in the PPM experiments, as described in Dirkson (2017). Of particular interest is the question of why low-SIC event thresholds show a more modest skill improvement compared to mid- and high-SIC event thresholds. This result was found to occur because the parametric method particularly outperforms the count method in the estimation of extreme quantiles of the underlying forecast distribution (Wilks 2002), which are sampled more frequently for high-SIC events (as events are defined in terms of a threshold exceedance).

Differences in hindcast skill between the PPM and OV experiments seen in Fig. 4 are almost entirely due to the influence of model biases. For example, in the central Arctic in July and August, the parametric method estimates forecast probabilities $P(X > \gamma)$ for $\gamma = 0.9$ that are systematically lower than count method estimates (not shown). However, because SIC is already 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. Since model biases are by construction absent in the PPM experiments, the parametric method outperforms the count method for this particular case.

5. 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. 2013a; Sigmond et al. 2013; Dirkson et al. 2017). To account for model errors in probabilistic SIC forecasts, we employ a novel type 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 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 (oftentimes referred to as the climatological distribution), according to

$$\hat{x}_t = F_o^{-1}[F_m(x_t)]. \quad (10)$$

In (10), $\hat{x}_t$ 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, (10) can be evaluated either analytically or numerically depending on whether F_m and F_o^{-1} are known exactly. When F_m and F_o are only known as ecdfs, (10) can be approximated, for example, by interpolating between values on a quantile–quantile plot. In practice, individual forecast ensemble members x_t are used as inputs to (10).

When F_o and F_m are the cdfs of normally distributed random variables, QM corrects the mean and standard deviation of the forecast random variable X_t through a shifting and rescaling, according to the bias in mean and standard deviation in the historical model distribution. When F_o and F_m are non-Gaussian, (10) is effective at correcting for higher-order moments as well. For comparison, this forecast distribution mapping method differs from the ELR/HELR techniques used to calibrate SIA in Krikken et al. (2016) in that the latter are regression methods that predict the forecast distribution using the forecast ensemble mean and/or standard deviation as predictors. Furthermore, the training data for ELR/HELR are pairs of these forecast statistical moments and corresponding observations, whereas for QM the training data are full MH and OH distributions.

In its standard form, QM is not a suitable calibration method for seasonal ensemble hindcasts of SIC from 1981 to the present for the following reasons. First, 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 variables X_t and $\hat{X}_t$. This assumption is violated for SIC forecasts, however, because of the pronounced negative trends in the historical period, particularly in more recent years. Second, QM is not well suited when F_m and F_o are modeled as a discontinuous distribution (such as the BEINF distribution), since mapping to or from values of zero or one can result in spurious SIC values such as numerous identical SIC values that are neither zero nor one. The TAQM method described

FIG. 5. Illustration of the trend-adjustment technique employed as a first step in TAQM. Solid black lines are the (left) MH and OH time series and (right) MH and OH histograms and BEINF pdfs. The gray and red shaded region in the upper-left panel shows ensemble spread. Dashed black lines are least squares fits to the MH and OH time series over the 1981–2011 period, and the solid black circles denote these trend values extrapolated to the hindcast year 2012. The red and blue solid lines are, respectively, the TAMH and TAOH time series in the left panels and TAMH and TAOH histograms and BEINF-fitted pdfs in the right panels. The mass points at zero and one for the BEINF pdfs have been multiplied by 10 for comparison with the histogram distributions.

here addresses both of these issues, enabling SIC ensemble forecasts to be calibrated.

a. Trend adjustment

As a first step in TAQM, MH and OH data are adjusted to account for nonstationarity in the mean SIC state. Nonstationarity in higher-SIC moments such as interannual variance is not considered in this study; however, this is likely an important topic for future research since sea ice variability can depend on its mean state, as has been shown for pan-Arctic SIE (Goosse et al. 2009) and volume (Massonnet et al. 2018). Although a simple piecewise variance rescaling procedure was attempted, the results were similar to applying the trend adjustment alone; a more sophisticated approach is likely needed to yield the desired effect.

Consider MH SIC time series $x_i(\tau_t)$ and OH SIC value $y(\tau_t)$, where τ_t denotes all years in the hindcast period preceding the forecast year t . When the linear trend over τ_t is statistically significant ($p < 0.05$), we compute the trend-adjusted values based on a particular forecast year t as

$$x'_i(\tau_t) = [x_i(\tau_t) - \tilde{x}_i(\tau_t)] + \tilde{x}_i(t), \quad \text{and} \quad (11a)$$

$$y'(\tau_t) = [y(\tau_t) - \tilde{y}(\tau_t)] + \tilde{y}(t), \quad (11b)$$

where $\tilde{x}_i$ and $\tilde{y}$ denote linear regressions of x_i and y onto τ_t , respectively.

Equations (11a) and (11b) have the effect of removing the linear trends in the OH and the MH data, and re-centering the respective time series about their *nonstationary means*, which we define as their linear trends extrapolated to the forecast year t . Equations (11a) and (11b) do not constrain the trend-adjusted values to $[0, 1]$, so we set values that fall below zero or above one to the appropriate lower or upper bound.

The trend-adjustment technique for a June-initialized hindcast of September 2012 SIC for a grid cell in the Laptev Sea is illustrated in Fig. 5. Both the MH and OH time series in the left-hand panels show marked negative trends over τ_t ($p < 0.05$) prior to trend adjustment. Following trend adjustment by (11a) and (11b), the respective means of the trend-adjusted historical model (TAMH) and trend-adjusted observed historical (TAOH) time series are centered about the nonstationary means of the MH and OH time series evaluated at $t = 2012$.

The BEINF distributions for these data are shown in the right-hand panels of Fig. 5. Relative to the MH

distribution, the TAMH distribution is less skewed with density shifted from high to midrange SIC values. Similarly, the TAOH distribution has density shifted toward midrange SIC values compared to the OH distribution 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 mass at zero.

b. Parametric fitting

Following the trend adjustment described above, the parametric BEINF distribution is fit to each of the TAMH time series (number of years in $\tau_t \times 10$ ensemble members), the TAOH time series (number of years in τ_t), 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 TAMH 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 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 3b, the parameters α and β cannot be fit for the cases 1–4 outlined therein. How to proceed in these cases will be addressed in the following section.

c. Calibrating BEINF parameters

The final step in TAQM is to calibrate the parameters of the forecast distribution $\text{BEINF}(\alpha_{x_t}, \beta_{x_t}, p_{x_t}, q_{x_t})$. QM should not be applied to the BEINF distribution as a whole for calibration, as it is a discontinuous distribution with no true inverse. However, QM may be applied to the beta portion, assuming nonstationarities have been accounted for. The parameters α_{x_t} and β_{x_t} are thus calibrated through QM, whereas parameters p_{x_t} and q_{x_t} are calibrated using a simple mean bias correction.

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

$$\hat{x}_t = F_{o,\text{beta}}^{-1} \left[F_{m,\text{beta}}(x_t) \right], \quad 0 < x_t < 1, \quad (12)$$

to be quantile mapped to values $\hat{x}_{t,1}, \dots, \hat{x}_{t,n-m}$. In (12), $F_{o,\text{beta}}^{-1}$ 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. Parameters $\alpha_{\hat{x}_t}$ and $\beta_{\hat{x}_t}$ are then obtained from the quantile mapped values $\hat{x}_t$ using ML estimation as described in the appendix. When cases 2–4 are encountered for any of the TAOH, TAMH, or the raw forecast values, $F_{m,\text{beta}}(x_t)$ is approximated using (5), and the quantile function $F_{o,\text{beta}}^{-1}$ is approximated by linear interpolation.

We calibrate parameters p_{x_t} and q_{x_t} by adding to the forecast probability of $\text{SIC} = 0$ [given by $p_{x_t}(1 - q_{x_t})$], and to the forecast probability of $\text{SIC} = 1$ (given by $p_{x_t}q_{x_t}$), the bias in these quantities for the TAMH distribution relative to the TAOH distribution:

$$p_{\hat{x}_t}(1 - q_{\hat{x}_t}) = p_{x_t}(1 - q_{x_t}) + [p_{y'}(1 - q_{y'}) - p_{x'}(1 - q_{x'})], \quad \text{and} \quad (13a)$$

$$p_{\hat{x}_t}q_{\hat{x}_t} = p_{x_t}q_{x_t} + (p_{y'}q_{y'} - p_{x'}q_{x'}). \quad (13b)$$

Solving for the calibrated parameters $p_{\hat{x}_t}$ and $q_{\hat{x}_t}$ yields

$$p_{\hat{x}_t} = p_{x_t} + p_{y'} - p_{x'}, \quad \text{and} \quad (14a)$$

$$q_{\hat{x}_t} = \frac{p_{x_t}q_{x_t} + p_{y'}q_{y'} - p_{x'}q_{x'}}{p_{\hat{x}_t}}, \quad (14b)$$

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

Case 1 of section 3b occurs when any of $p_{x_t} = 1$, $p_{x'} = 1$, or $p_{y'} = 1$, corresponding to the location considered being completely ice free or completely ice covered. In such situations, (12) cannot be evaluated because the beta portion of the BEINF distribution cannot be defined. Two options for how to proceed are to not calibrate or to use the TAOH distribution. We find that for CanCM3 the latter is nearly always preferred, so for simplicity we 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 revert to the TAOH distribution).

We illustrate this calibration procedure in Fig. 6 for the same case used to illustrate trend adjustment in Fig. 5. 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 (12) only involves the beta portion of the BEINF distribution (not scaled by $1 - p$), and (14a) and (14b) are based on the probability masses of the Bernoulli distribution, we show the beta distribution and probability masses separately on the same panels in Fig. 6 (rather than the BEINF distributions).

As an example, we consider the ensemble member $x_t = 0.40$, marked by the dashed orange line in the right-hand panel of Fig. 6. The TAMH beta cdf evaluated at this forecast value is given by the quantile $F_{m,\text{beta}}(0.40) = 0.45$, marked where the dashed red lines intersect in the left-hand panel of Fig. 6. The inverse of the TAOH beta cdf evaluated at this same quantile corresponds to an observed SIC value $F_{o,\text{beta}}^{-1}(0.45) = 0.14$, marked where the dashed blue lines intersect. The corresponding calibrated forecast value for $x_t = 0.40$ is thus

FIG. 6. Illustration of the BEINF parameter calibration using TAQM for the same hindcast used to illustrate the trend adjustment in Fig. 5. (left) TAMH and TAOH. (right) Uncalibrated hindcast and calibrated hindcast distributions for the year 2012. Solid lines are the beta portion of the BEINF CDF (not scaled by p) for the TAMH (red), TAOH (blue), uncalibrated hindcast (orange), and TAQM-calibrated hindcast (green). Circles are the probabilities of SIC equaling zero [$p(1 - q)$] and one (pq). Dashed lines and black arrows are described in the main text.

$\hat{x}_t = 0.14$. This procedure is repeated for the remaining 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. 6.

For the Bernoulli portion of the BEINF distribution, described by p_{x_t} and q_{x_t} , the calibrated parameters are found by solving (14a) and (14b). This yields an increase in the forecast probability from 60% to 78% for SIC = 0, indicated by the amount given by the length of the black arrow in the left-hand panel of Fig. 6. The forecast probability of 0% for SIC = 1 remains unchanged.

6. TAQM-calibrated hindcasts

Probabilistic skill assessments and postprocessing methods for seasonal forecasts normally undertake statistical estimation using the entire hindcast sample, employing cross validation to assess out-of-sample performance (e.g., Becker and van den Dool 2016; Kharin et al. 2017a). In doing so, it is assumed that the sample of hindcast years is homogeneous, with no account taken of their temporal ordering so years occurring after the target year can be considered as additional draws from a fixed sample.

In the present case where temporal trends are accounted for explicitly in the postprocessing, the temporal ordering of the hindcast years does enter. Therefore, in TAQM as described above (referred to as TAQM-PAST), the choice to use only that part of the historical record that precedes the hindcast year of interest was made so that the data used to train TAQM are fully out of sample. However, this effectively reduces the size of

the historical sample that is available for training the calibration and hence may lead to underestimation of skill of real-time forecasts, for which the entire hindcast sample is available. Furthermore, estimation of the non-stationary mean in the trend-adjustment step of TAQM may be compromised by the shorter record used to estimate the linear trend.

For these reasons, we consider also for comparative purposes hindcast skill when the full historical record (1981–2017) is employed to train TAQM (referred to as TAQM-FULL), with leave-one-out cross validation applied. Such considerations have guided similar choices in other contexts. For example, Kharin et al. (2012) applied trend corrections to multiannual predictions of temperature that take into account the whole hindcast record. This was justified by applying the same trend-adjustment methodology for the initialized decadal hindcasts to the uninitialized climate simulations that serve as reference forecasts for skill assessment. Krikken et al. (2016) applied similar considerations to seasonal forecasts of regional Arctic sea ice area, noting that model errors in representing the reduction of Arctic sea ice in recent decades will introduce a dependence on forecast year to the model drift and hence to the optimal drift correction. Like Kharin et al. (2012), they applied trend corrections based on the entire hindcast period both to the initialized hindcasts and to a reference forecast consisting of the trend-adjusted climatological distribution.

We now assess bias, reliability, and skill for the calibrated probabilistic hindcasts (TAQM-PAST and TAQM-FULL) over the recent 18-yr period 2000–17. We choose this shorter record on the basis that the results

FIG. 7. Scatterplots of the observed September SIC climatologies as a function of (a) uncalibrated hindcast, (b) TAQM-PAST calibrated hindcast, and (c) TAQM-FULL calibrated hindcast climatologies (for hindcasts initialized on 1 June). Unbiased hindcasts show points along the 1:1 line. Normalized histograms of the PIT values for the same (d) uncalibrated, (e) TAQM-PAST calibrated, and (f) TAQM-FULL calibrated hindcasts.

are representative for the current epoch of reduced and declining Arctic sea ice. To assess bias and reliability, we focus on September hindcasts initialized on 1 June, whereas skill is assessed for all four initialization months and six target months considered.

a. Bias and reliability

Bias is the climatological ensemble-mean hindcast SIC (uncalibrated or TAQM calibrated) minus the observed SIC climatology over the 2000–17 period. If a forecast is reliable, then the observed outcome should behave like a random draw from the forecast distribution—that is, the uncertainty characterized in the forecast distribution should describe the “true” forecast uncertainty that would be obtained in the absence of model errors.

First, we quantify the reliability of each full hindcast distribution (uncalibrated and TAQM calibrated) through the probability integral transform (PIT) adapted for the BEINF distribution,

$$\text{PIT} = \begin{cases} \rho_0, & y_t = 0 \\ F(y_t), & 0 < y_t < 1. \\ \rho_1, & y_t = 1 \end{cases} \quad (15)$$

In (15), y_t is the observed SIC value at time t , F is the BEINF hindcast cdf for time t , ρ_0 is a random

probability value drawn from the uniform distribution $U[0, p(1 - q)]$, and ρ_1 is a random probability drawn from $U(1 - pq, 1)$. If the probabilistic hindcasts are reliable (i.e., y_t has distribution F in the aggregate across all hindcasts) then $\text{PIT} \sim U(0, 1)$ (Dawid 1984; Gneiting et al. 2007).

Second, we assess reliability of hindcast binary events. In order for probabilistic forecasts of specific events (e.g., exceeding a particular SIC threshold) to be reliable, forecast probabilities should agree (in the aggregate) with the observed relative frequency of that event. For instance, in all cases where a forecast probability of 60% is issued for SIC exceeding 0.15, SIC should be observed to be greater than 0.15 in 60% of those cases. Reliability is illustrated using the reliability diagram.

As noted earlier, uncalibrated hindcasts of September SIC suffer from large biases (Fig. 7a) consistent with an overestimation of mean SIC in regions of low and intermediate SIC and an underestimation of mean SIC in the high-SIC central Arctic. Biases are substantially reduced in the TAQM-calibrated hindcasts (Figs. 7b,c). While low and midrange SIC values tend to remain positively biased for both TAQM-PAST and TAQM-FULL, and high SIC values become slightly positively biased for TAQM-PAST, the overall reduction in bias for both is evident. The greatest reduction in bias seen for the TAQM-FULL calibrated hindcasts.

FIG. 8. Reliability diagrams for September hindcasts initialized on 1 June. (top) Uncalibrated hindcasts and (bottom) TAQM-calibrated hindcasts (black squares = TAQM-PAST, white triangles = TAQM-FULL) of the probability of exceeding three threshold SIC values, 0.15, 0.5, and 0.9, as labeled. The horizontal dashed line is the observed relative frequency of the event.

The assessment of reliability according to the PIT histograms shows a substantial improvement in the representation of forecast uncertainty for the TAQM-calibrated hindcasts (Figs. 7e,f) relative to the uncalibrated hindcasts (Fig. 7d). The peak in PIT values between 0 and 0.1 for the TAQM-PAST and TAQM-FULL calibrated hindcasts and the peak in PIT values between 0.9 and 1 for the TAQM-FULL calibrated hindcasts indicates that the hindcasts are too often overconfident. This finding has also been found when using QM to calibrate seasonal precipitation forecasts (Zhao et al. 2017) and is partly because, while QM takes into account spread differences between the modeled and observed historical distributions, QM does not explicitly account for relationships between the ensemble spread of individual forecasts and observed outcomes.

In Fig. 8, we assess the reliability of hindcasts for three separate binary events, specifically the probability of exceeding the SIC thresholds 0.15, 0.50, and 0.90. Here, forecast probabilities have been rounded to the nearest 10%. Uncalibrated probabilistic hindcasts of the event that SIC exceeds 0.15 are underforecast when a 0% chance is forecast, but overforecast for all other forecast probabilities. Similarly, for the event that SIC exceeds 0.5 uncalibrated hindcasts are overforecast, whereas for the event that SIC exceeds 0.9, they are underforecast. These results are consistent with the high bias in SIC in

areas of low coverage and low bias in SIC in areas of high coverage seen in Fig. 7. The TAQM-calibrated hindcasts are substantially more reliable than the uncalibrated hindcasts for all three events; however, the tendency to overforecast (underforecast) the event that SIC exceeds 15% (50%) remains. Calibrated forecasts of exceeding 90% are slightly underforecast for TAQM-FULL but become overforecast for TAQM-PAST. Overall, hindcasts calibrated using TAQM-FULL are the most reliable.

b. Skill

To assess TAQM impacts on hindcast skill, we make use of spatial maps of the CRPSS_{j,k} to compare the skill of TAQM-PAST calibrated hindcasts against that of two reference hindcasts: the uncalibrated probabilistic hindcasts and the TAOH distribution (computed as described in section 5a). Additionally, we compare the skill of TAQM-FULL against TAQM-PAST. The assessment of skill against the uncalibrated hindcasts determines the degree to which TAQM-PAST is able to mitigate skill degradation due to model errors. The comparison of TAQM-PAST with the TAOH distribution benchmarks hindcast skill against the probabilistic analog to a linear trend extrapolation forecast, a common reference forecast used to assess the skill of deterministic forecasts of sea ice. Finally, comparison of

FIG. 9. Spatial maps of CRPSS_{jk} comparing the TAQM-PAST 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 labeled). Improvement using the calibration method is indicated by positive (red) CRPSS values. Locations where SIC standard deviation < 0.025 are masked to white (as described in the text). The percentage of improvement and the CRPSS_{jk} values are given in the top-right corner of each map.

TAQM-FULL with TAQM-PAST gives an indication of the importance of using the full validation period to train TAQM.

Pan-Arctic quantification of skill is provided by computing the total area of grid cells for which $\text{CRPSS}_{jk} > 0$, among the total area of all grid cells for which $\text{CRPSS}_{jk} \neq 0$, referred to as percentage of improvement. We also consider CRPSS_{jk} values, since the percentage improvement is insensitive to the magnitude of skill at individual locations.

1) TAQM-PAST VS UNCALIBRATED

Skill is compared between the TAQM-PAST calibrated hindcasts and the uncalibrated parametric method hindcasts in spatial maps of the CRPSS_{jk} (Fig. 9). In these maps, we have masked out regions where SIC standard deviation is less than 0.025 in either the model or

observations, corresponding to regions where SIC typically varies by 5%. Overall, there is a large improvement in hindcast skill for the TAQM-calibrated hindcasts, as seen by the generally positive CRPSS_{jk} values in most target months. The few areas showing lower skill in the TAQM-calibrated hindcasts than in the uncalibrated hindcasts tend to occur when skill is already high for the uncalibrated hindcasts, such as during the first hindcast month (lead zero).

The TAQM-PAST hindcasts can be less skillful than the uncalibrated hindcasts in locations that have experienced a rare event at some point in the hindcast record. For instance, the region of negative CRPSS_{jk} values in the western central Arctic in October corresponds to a region where observed 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

FIG. 10. As in Fig. 9, but comparing the TAQM-PAST calibrated hindcasts (forecast being evaluated) against the TAOH distribution (reference forecast).

density concentrated between approximately 30% and 100%, effectively shifting the forecast distribution toward low SIC values and degrading skill for the remaining years (2008–17) when SIC is near 100%. Notably, TAQM-PAST hindcasts of September SIC initialized on 1 September are much less skillful than the raw hindcasts. In skill maps comparing TAQM-FULL and the uncalibrated hindcasts (not shown), this degradation in skill is no longer present.

As indicated by the percentage of improvement and CRPSS_{jk} values, the improvement in hindcast skill for the TAQM-PAST hindcasts generally increases in spatial coverage and magnitude after the first or second hindcast month, consistent with the growth of biases in the uncalibrated hindcasts with increasing lead time as the model drifts toward its own biased climatology. After which, the improvement in skill differs depending on the metric considered and is generally uniform. In particular, skill in July and August are similar for all three spring initialization months, reflecting the fact that

biases in the uncalibrated hindcasts are large in these months. Generally, skill improves more for the TAQM-PAST hindcasts relative to the uncalibrated hindcasts during the spring and fall transition seasons.

2) TAQM-PAST VS TREND-ADJUSTED CLIMATOLOGY

A comparison of probabilistic hindcast skill between the TAQM-PAST calibrated hindcasts and the TAOH distribution is shown in Fig. 10. As in the previous section, we have masked out regions where SIC standard deviation less than 0.025 in either the model or observations. Large areas of hindcast skill relative to the trend-adjusted climatology are seen over much of the Arctic, even for long lead times.

For hindcasts initialized in March, skill throughout the melt season is mainly confined to the Barents Sea and Baffin Bay. However, predictions initialized in May and June show considerably greater skill throughout the melt season. In particular, the Barents and Kara Seas

FIG. 11. As in Fig. 9, but comparing the TAQM-FULL calibrated hindcasts (forecast being evaluated) against the TAQM-PAST calibrated hindcasts (reference forecast).

show high skill relative to the TAOH reference forecast throughout the 6-month hindcast period. While skill in the western Arctic is mainly only present in the Beaufort Sea for hindcasts initialized in May, skill is generally more widespread for hindcasts initialized in June. For hindcasts initialized in September, skill is seen over most of the Arctic through October, but becomes limited to the eastern Arctic in the winter.

3) TAQM-FULL vs TAQM-PAST

Calibrated hindcast skill for TAQM-FULL is compared against TAQM-PAST in Fig. 11, which shows that greater skill is achieved using the full historical record (with cross validation) to train TAQM compared to using only past information. Widespread improvement in skill is evident for TAQM-FULL relative to TAQM-PAST for nearly all forecast months. Notably, greater skill is seen in the East Siberian Sea near the Siberian coastline in September for hindcasts initialized in May,

June, and September. This is a region where TAQM-PAST consistently performs worse than the TAOH distribution (Fig. 10). Additionally, areas where TAQM-PAST was shown to have lower skill than the uncalibrated hindcasts (Fig. 9), such as in September at a lead time of zero months, correspond to areas where TAQM-FULL hindcasts are more skillful than the TAQM-PAST hindcasts.

7. Conclusions

In this study, we introduced a methodology to improve seasonal probability hindcasts of Arctic SIC. This methodology has been tested in a set of CanCM3 ensemble hindcasts.

The first component of the methodology improves the representation of the forecast distribution by fitting the BEINF parametric distribution to ensembles of SIC values. The BEINF distribution was shown to be a

reasonable model for this application based 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 pseudoperfect model experiments. Furthermore, we find these results to be only slightly sensitive to the observational dataset used to verify. In particular, there are minimal differences seen for categorical predictions when verifying against NOAA/NSIDC CDRv3 compared with HadISST2, with slightly lower skill predicting the low-SIC event thresholds (Fig. S1 in the online supplemental material), whereas there is little discernible difference in skill for predictions of the complete forecast distribution (not shown).

The second component is a novel calibration technique specifically designed for seasonal hindcasts and real-time forecasts of SIC. This TAQM calibration method explicitly accounts for nonstationarity in the mean SIC state, 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) Linear trends are removed from observed historical and model (hindcast) historical time series, recentering the historical time series on a nonstationary mean defined by the trend fit evaluated at year t .
- 2) BEINF fits are applied to the forecast and trend-adjusted observed historical and model historical SIC distributions. In instances when the BEINF distribution(s) cannot be fit, we revert to the empirical distribution(s).
- 3) The forecast SIC values on the open interval $(0, 1)$ are quantile mapped from the trend-adjusted model historical distribution to the beta portion of the trend-adjusted observed historical distribution. Parameters α_{x_t} and β_{x_t} are then fit to these quantile mapped values, except when they cannot be estimated, in which case we revert to the count method.
- 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 based on the trend-adjusted model historical and trend-adjusted observed historical distributions.

The TAQM calibration method applied to hindcasts over 2000–17 was shown to substantially reduce biases, which can be quite large in CanCM3. The calibrated probabilistic hindcasts also show a substantial improvement in forecast reliability in the summer, particularly for probabilities that SIC exceeds 15% and 50% coverage; however, based on the probability integral

transform, calibrated hindcasts remained slightly overconfident. Very similar results are obtained when verifying against NOAA/NSIDC CDRv3 (not shown). Substantial improvement in TAQM-calibrated hindcast skill is seen relative to uncalibrated hindcasts, particularly at lead times > 1 month, as model biases intensify. Compared against a trend-adjusted climatological distribution reference hindcast, SIC in specific regions is forecast with relatively high levels of skill even at long lead times. Hindcasts are particularly skillful in September for initializations in May and June. As before, these conclusions are the same using NOAA/NSIDC CDRv3 for calibration and verification; however, we see generally smaller improvements relative to those seen using HadISST2 for spring, fall, and winter target months, whereas we see generally larger improvements for summer target months (Figs. S2, S3). Finally, calibration using the full historical record to train TAQM was shown to result in enhanced and widespread skill relative to when only data preceding the hindcast year of interest are used.

Although the TAQM-calibrated CanCM3 hindcasts do not always outperform a trend-adjusted climatology, much room for improvement of sea ice dynamical forecasts remains, including improvements in spatial resolution, initialization, model biases, and multimodel forecasting methodologies. The methods presented here provide a path toward maximizing the value of such forecasts in a probabilistic framework. Additional postprocessing methods remain to be explored, including the adaptation of other calibration methods to SIC and multimodel calibration. Techniques developed for calibrating precipitation forecasts (e.g., Li et al. 2017) could be particularly relevant, as precipitation exhibits similar statistical characteristics as SIC (e.g., boundedness). As noted in section 5, more sophisticated methods for accounting for nonstationarity in higher-order statistical moments in the SIC data used in calibration might be needed (e.g., time-dependent regression-based estimation of the BEINF parameters). Finally, accounting for errors in the calibrated forecast distribution resulting from parameter estimation (Siebert et al. 2016) and a more formal representation of observational uncertainty (Massonnet et al. 2016) could lead to further improvements.

Acknowledgments. The codes used here for applying the BEINF distribution and TAQM calibration method to SIC forecasts were developed in Python v2.7 and are available online (<https://github.com/adirkson/SIC-probability>) with documentation and a tutorial (<https://adirkson.github.io/SIC-probability>). The authors thank Woosung Lee for producing the hindcasts used in this

study, as well as Slava Kharin, Alex Cannon, and Edward Blanchard-Wrigglesworth for their comments on earlier versions of this manuscript. Additionally, we thank three reviewers for their constructive feedback. AD and WM would like to thank the Canadian Sea Ice and Snow Evolution Network (CanSISE) for funding this research. AM acknowledges funding from the Natural Sciences and Engineering Research Council of Canada (NSERC).

APPENDIX

Maximum Likelihood Estimates of BEINF Parameters

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 analytical solutions are given by $\hat{p} = \sum_i \mathbb{1}_{\{0,1\}}(x_i)/n$ and $\hat{q} = \sum_i x_i \mathbb{1}_{\{0,1\}}(x_i)/m$, where n is the size of the entire sample, $m = \sum_i \mathbb{1}_{\{0,1\}}(x_i)$ is the number of zeros and ones in the sample, and $\hat{q} = 0/0$ is set to zero by convention. The expression $\mathbb{1}_{\{0,1\}}(x_i) = 1$ when either $x_i = 0$ or $x_i = 1$, and $\mathbb{1}_{\{0,1\}}(x_i) = 0$ when $x_i \in (0, 1)$.

The ML estimates $\hat{\alpha}$ and $\hat{\beta}$ for the BEINF distribution are computed from the $n - m$ sized subsample x_{sub} . ML estimation is carried out on x_{sub} using the “beta.fit” function of 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 obtained by the MOM:

$$\hat{\alpha} = \bar{x}_{\text{sub}} \left[\frac{\bar{x}_{\text{sub}}(1 - \bar{x}_{\text{sub}})}{\text{var}(x_{\text{sub}})} - 1 \right], \quad \text{and} \quad (\text{A1})$$

$$\hat{\beta} = (1 - \bar{x}_{\text{sub}}) \left[\frac{\bar{x}_{\text{sub}}(1 - \bar{x}_{\text{sub}})}{\text{var}(x_{\text{sub}})} - 1 \right]. \quad (\text{A2})$$

When the ML algorithm does not converge, $\hat{\alpha}$ and $\hat{\beta}$ are set to those obtained by the MOM.

REFERENCES

- Becker, E., and H. van den Dool, 2016: Probabilistic seasonal forecasts in the North American Multimodel Ensemble: A baseline skill assessment. *J. Climate*, **29**, 3015–3026, <https://doi.org/10.1175/JCLI-D-14-00862.1>.
- Blanchard-Wrigglesworth, E., and Coauthors, 2017: Multi-model seasonal forecast of Arctic sea-ice: Forecast uncertainty at pan-Arctic and regional scales. *Climate Dyn.*, **49**, 1399–1410, <https://doi.org/10.1007/s00382-016-3388-9>.
- Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1–3, [https://doi.org/10.1175/1520-0493\(1950\)078<0001:VOFEIT>2.0.CO;2](https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2).
- Buehner, M., A. Caya, L. Pogson, T. Carrieres, and P. Pestieau, 2013a: A new Environment Canada regional ice analysis system. *Atmos.–Ocean*, **51**, 18–34, <https://doi.org/10.1080/07055900.2012.747171>.
- , —, T. Carrieres, L. Pogson, and M. Lajoie, 2013b: Overview of sea ice data assimilation activities at Environment Canada. *Proc. of the ECMWF-WWRP/THORPEX Polar Prediction Workshop*, Reading, United Kingdom, 10 pp.
- , —, —, and —, 2015: Assimilation of SSMIS and ASCAT data and the replacement of highly uncertain estimates in the Environment Canada Regional Ice Prediction System. *Quart. J. Roy. Meteor. Soc.*, **142**, 562–573, <https://doi.org/10.1002/qj.2408>.
- Dawid, A. P., 1984: Present position and potential developments: Some personal views: Statistical theory: The prequential approach. *J. Roy. Stat. Soc.*, **147A**, 278–290, <https://doi.org/10.2307/2981683>.
- Director, H. M., A. E. Raftery, and C. M. Bitz, 2017: Improved sea ice forecasting through spatiotemporal bias correction. *J. Climate*, **30**, 9493–9510, <https://doi.org/10.1175/JCLI-D-17-0185.1>.
- Dirkson, A., 2017: Initializing sea ice thickness and quantifying uncertainty in seasonal forecasts of Arctic sea ice. Ph.D. thesis, University of Victoria, 115 pp.
- , W. J. Merryfield, and A. Monahan, 2017: Impacts of sea ice thickness initialization on seasonal Arctic sea ice predictions. *J. Climate*, **30**, 1001–1017, <https://doi.org/10.1175/JCLI-D-16-0437.1>.
- Ellis, B., and L. Brigham, 2009: Arctic marine shipping assessment 2009 report. Arctic Council Rep., 194 pp.
- Fetterer, F., K. Knowles, W. Meier, M. Savoie, and A. Windnagel, 2017: Sea ice index, version 3. National Snow and Ice Data Center, accessed 1 April 2018, <https://doi.org/10.7265/NSK072F8>.
- Flato, G. M., and W. D. Hibler III, 1992: Modeling pack ice as a cavitating fluid. *J. Phys. Oceanogr.*, **22**, 626–651, [https://doi.org/10.1175/1520-0485\(1992\)022<0626:MPIAAC>2.0.CO;2](https://doi.org/10.1175/1520-0485(1992)022<0626:MPIAAC>2.0.CO;2).
- Fučkar, N. S., D. Volpi, V. Guemas, and F. J. Doblas-Reyes, 2014: A posteriori adjustment of near-term climate predictions: Accounting for the drift dependence on the initial conditions. *Geophys. Res. Lett.*, **41**, 5200–5207, <https://doi.org/10.1002/2014GL060815>.
- Gneiting, T., A. E. Raftery, A. H. Westveld III, and T. Goldman, 2005: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Wea. Rev.*, **133**, 1098–1118, <https://doi.org/10.1175/MWR2904.1>.
- , F. Balabdaoui, and A. E. Raftery, 2007: Probabilistic forecasts, calibration and sharpness. *J. Roy. Stat. Soc.*, **69B**, 243–268, <https://doi.org/10.1111/j.1467-9868.2007.00587.x>.
- Goessling, H. F., and T. Jung, 2018: A probabilistic verification score for contours: Methodology and application to Arctic ice-edge forecasts. *Quart. J. Roy. Meteor. Soc.*, **144**, 735–743, <https://doi.org/10.1002/qj.3242>.
- Goosse, H., O. Arzel, C. M. Bitz, A. de Montety, and M. Vancoppenolle, 2009: Increased variability of the Arctic summer ice extent in a warmer climate. *Geophys. Res. Lett.*, **36**, L23702, <https://doi.org/10.1029/2009GL040546>.
- Gottschalk, L., and R. Weingartner, 1998: Distribution of peak flow derived from a distribution of rainfall volume and runoff coefficient, and a unit hydrograph. *J. Hydrol.*, **208**, 148–162, [https://doi.org/10.1016/S0022-1694\(98\)00152-8](https://doi.org/10.1016/S0022-1694(98)00152-8).
- Henderson-Sellers, A., 1978: Surface type and its effect upon cloud cover: A climatological investigation. *J. Geophys. Res.*, **83**, 5057–5062, <https://doi.org/10.1029/JC083iC10p05057>.

- Hersbach, H., 2000: Decomposition of the continuous ranked probability score for ensemble prediction systems. *Wea. Forecasting*, **15**, 559–570, [https://doi.org/10.1175/1520-0434\(2000\)015<0559:DOTCRP>2.0.CO;2](https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2).
- Johnson, N. L., S. Kotz, and N. Balakrishnan, 1995: *Continuous Univariate Distributions*. Vol. 2. Wiley, 752 pp.
- Kharin, V. V., and F. W. Zwiers, 2003: Improved seasonal probability forecasts. *J. Climate*, **16**, 1684–1701, [https://doi.org/10.1175/1520-0442\(2003\)016<1684:ISPF>2.0.CO;2](https://doi.org/10.1175/1520-0442(2003)016<1684:ISPF>2.0.CO;2).
- , Q. Teng, F. W. Zwiers, G. J. Boer, J. Derome, and J. S. Fontecilla, 2009: Skill assessment of seasonal hindcasts from the Canadian Historical Forecast Project. *Atmos.–Ocean*, **47**, 204–223, <https://doi.org/10.3137/AO1101.2009>.
- , G. J. Boer, W. J. Merryfield, J. F. Scinocca, and W.-S. Lee, 2012: Statistical adjustment of decadal predictions in a changing climate. *Geophys. Res. Lett.*, **39**, L19705, <https://doi.org/10.1029/2012GL052647>.
- , W. J. Merryfield, G. J. Boer, and W.-S. Lee, 2017a: A postprocessing method for seasonal forecasts using temporally and spatially smoothed statistics. *Mon. Wea. Rev.*, **145**, 3545–3561, <https://doi.org/10.1175/MWR-D-16-0337.1>.
- , —, —, and —, 2017b: A postprocessing method for seasonal forecasts using temporally and spatially smoothed statistics. *Mon. Wea. Rev.*, **145**, 3545–3561, <https://doi.org/10.1175/MWR-D-16-0337.1>.
- Krikken, F., M. Schmeits, W. Vlot, V. Guemas, and W. Hazeleger, 2016: Skill improvement of dynamical seasonal Arctic sea ice forecasts. *Geophys. Res. Lett.*, **43**, 5124–5132, <https://doi.org/10.1002/2016GL068462>.
- Krishnamurti, T. N., C. M. Kishtawal, T. E. LaRow, D. R. Bachiochi, Z. Zhang, C. E. Williford, S. Gadgil, and S. Surendran, 1999: Improved weather and seasonal climate forecasts from multimodel superensemble. *Science*, **285**, 1548–1550, <https://doi.org/10.1126/science.285.5433.1548>.
- Kwok, R., and D. A. Rothrock, 2009: Decline in Arctic sea ice thickness from submarine and ICESat records: 1958–2008. *Geophys. Res. Lett.*, **36**, L15501, <https://doi.org/10.1029/2009GL039035>.
- Li, B., and R. Avissar, 1994: The impact of spatial variability of land-surface characteristics on land-surface heat fluxes. *J. Climate*, **7**, 527–537, [https://doi.org/10.1175/1520-0442\(1994\)007<0527:TIOSVO>2.0.CO;2](https://doi.org/10.1175/1520-0442(1994)007<0527:TIOSVO>2.0.CO;2).
- Li, W., Q. Duan, C. Miao, A. Ye, W. Gong, and Z. Di, 2017: A review on statistical postprocessing methods for hydrometeorological ensemble forecasting. *Wiley Interdiscip. Rev.: Water*, **4**, e1246, <https://doi.org/10.1002/wat2.1246>.
- Maslanik, J., J. Stroeve, C. Fowler, and W. Emery, 2011: Distribution and trends in Arctic sea ice age through spring 2011. *Geophys. Res. Lett.*, **38**, L13502, <https://doi.org/10.1029/2011GL047735>.
- Massonnet, F., O. Bellprat, V. Guemas, and F. J. Doblas-Reyes, 2016: Using climate models to estimate the quality of global observational data sets. *Science*, aaf6369, <https://doi.org/10.1126/science.aaf6369>.
- , M. Vancoppenolle, H. Goosse, D. Docquier, T. Fichefet, and E. Blanchard-Wrigglesworth, 2018: Arctic sea-ice change tied to its mean state through thermodynamic processes. *Nat. Climate Change*, **8**, 599–603, <https://doi.org/10.1038/s41558-018-0204-z>.
- Meier, W., F. Fetterer, M. Savoie, S. Mallory, R. Duerr, and J. Stroeve, 2017: NOAA/NSIDC Climate Data Record of Passive Microwave Sea Ice Concentration, version 3. National Snow and Ice Data Center, accessed 1 January 2018, <https://doi.org/10.7265/N59P2ZTG>.
- Melia, N., K. Haines, and E. Hawkins, 2016: Sea ice decline and 21st century trans-Arctic shipping routes. *Geophys. Res. Lett.*, **43**, 9720–9728, <https://doi.org/10.1002/2016GL069315>.
- Merryfield, W. J., W.-S. Lee, G. J. Boer, V. V. Kharin, J. F. Scinocca, G. M. Flato, R. S. Ajayamohan, and J. C. Fyfe, 2013a: The Canadian seasonal to interannual prediction system. Part I: Models and initialization. *Mon. Wea. Rev.*, **141**, 2910–2945, <https://doi.org/10.1175/MWR-D-12-00216.1>.
- Messner, J. W., G. J. Mayr, A. Zeileis, and D. S. Wilks, 2014: Heteroscedastic extended logistic regression for postprocessing of ensemble guidance. *Mon. Wea. Rev.*, **142**, 448–456, <https://doi.org/10.1175/MWR-D-13-00271.1>.
- Ospina, R., and S. L. Ferrari, 2010: Inflated beta distributions. *Stat. Pap.*, **51**, 111, <https://doi.org/10.1007/s00362-008-0125-4>.
- Palmer, T., and Coauthors, 2004: Development of a European multimodel ensemble system for seasonal-to-interannual prediction (DEMETER). *Bull. Amer. Meteor. Soc.*, **85**, 853–872, <https://doi.org/10.1175/BAMS-85-6-853>.
- , R. Buizza, F. Doblas-Reyes, T. Jung, M. Leutbecher, G. J. Shutts, M. Steinheimer, and A. Weisheimer, 2009: Stochastic parametrization and model uncertainty. ECMWF Tech. Memo. 598, 44 pp., <https://doi.org/10.21957/ps8gbwbv>.
- Peng, G., W. N. Meier, D. J. Scott, and M. H. Savoie, 2013: A long-term and reproducible passive microwave sea ice concentration data record for climate studies and monitoring. *Earth Syst. Sci. Data*, **5**, 311–318, <https://doi.org/10.5194/essd-5-311-2013>.
- Raschke, M., 2009: The biased transformation and its application in goodness-of-fit tests for the beta and gamma distribution. *Commun. Stat. Simul. Comput.*, **38**, 1870–1890, <https://doi.org/10.1080/03610910903152631>.
- , 2011: Empirical behaviour of tests for the beta distribution and their application in environmental research. *Stochastic Environ. Res. Risk Assess.*, **25**, 79–89, <https://doi.org/10.1007/s00477-010-0410-3>.
- Reynolds, C. A., P. J. Webster, and E. Kalnay, 1994: Random error growth in NMC's global forecasts. *Mon. Wea. Rev.*, **122**, 1281–1305, [https://doi.org/10.1175/1520-0493\(1994\)122<1281:REGING>2.0.CO;2](https://doi.org/10.1175/1520-0493(1994)122<1281:REGING>2.0.CO;2).
- Richardson, D. S., 2001: Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size. *Quart. J. Roy. Meteor. Soc.*, **127**, 2473–2489, <https://doi.org/10.1002/qj.49712757715>.
- Rothrock, D. A., Y. Yu, and G. A. Maykut, 1999: Thinning of the Arctic sea-ice cover. *Geophys. Res. Lett.*, **26**, 3469–3472, <https://doi.org/10.1029/1999GL010863>.
- Roy, P., R. Laprise, and P. Gachon, 2016: Sampling errors of quantile estimations from finite samples of data. <https://arxiv.org/abs/1610.03458>.
- Siebert, S., P. G. Sansom, and R. M. Williams, 2016: Parameter uncertainty in forecast recalibration. *Quart. J. Roy. Meteor. Soc.*, **142**, 1213–1221, <https://doi.org/10.1002/qj.2716>.
- Sigmond, M., J. C. Fyfe, G. M. Flato, V. V. Kharin, and W. J. Merryfield, 2013: Seasonal forecast skill of Arctic sea ice area in a dynamical forecast system. *Geophys. Res. Lett.*, **40**, 529–534, <https://doi.org/10.1002/grl.50129>.
- Stephens, M., 1986: Tests based on EDF statistics. *Goodness-of-Fit Techniques*, R. B. d'Agostino and M. A. Stephens, Eds., Dekker, 97–194.
- Stroeve, J., E. Blanchard-Wrigglesworth, V. Guemas, S. Howell, F. Massonnet, and S. Tietsche, 2015: Improving predictions of Arctic sea ice extent. *Eos, Trans. Amer. Geophys. Union*, **96**, <https://doi.org/10.1029/2015EO031431>.

- Titchner, H. A., and N. A. Rayner, 2014: The Met Office Hadley Centre sea ice and sea surface temperature data set, version 2: 1. Sea ice concentrations. *J. Geophys. Res. Atmos.*, **119**, 2864–2889, <https://doi.org/10.1002/2013JD020316>.
- Tivy, A., S. E. L. Howell, B. Alt, S. McCourt, R. Chagnon, G. Crocker, T. Carrieres, and J. J. Yackel, 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. *J. Geophys. Res.*, **116**, C03007, <https://doi.org/10.1029/2009JC005855>; Corrigendum, **116**, C06027, <https://doi.org/10.1029/2011JC007248>.
- 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. *J. Atmos. Sci.*, **59**, 1917–1942, [https://doi.org/10.1175/1520-0469\(2002\)059<1917:APPFTS>2.0.CO;2](https://doi.org/10.1175/1520-0469(2002)059<1917:APPFTS>2.0.CO;2).
- Weigel, A. P., M. A. Liniger, and C. Appenzeller, 2008: Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts? *Quart. J. Roy. Meteor. Soc.*, **134**, 241–260, <https://doi.org/10.1002/qj.210>.
- Wilks, D. S., 2002: Smoothing forecast ensembles with fitted probability distributions. *Quart. J. Roy. Meteor. Soc.*, **128**, 2821–2836, <https://doi.org/10.1256/qj.01.215>.
- , 2009: Extending logistic regression to provide full-probability-distribution MOS forecasts. *Meteor. Appl.*, **16**, 361–368, <https://doi.org/10.1002/met.134>.
- , 2011: *Statistical Methods in the Atmospheric Sciences*. 3rd ed. International Geophysics Series, Vol. 100, Academic Press, 704 pp.
- Yao, A. Y. M., 1974: A statistical model for the surface relative humidity. *J. Appl. Meteor.*, **13**, 17–21, [https://doi.org/10.1175/1520-0450\(1974\)013<0017:ASMFTS>2.0.CO;2](https://doi.org/10.1175/1520-0450(1974)013<0017:ASMFTS>2.0.CO;2).
- Zhao, T., J. C. Bennett, Q. J. Wang, A. Schepen, A. W. Wood, D. E. Robertson, and M.-H. Ramos, 2017: How suitable is quantile mapping for postprocessing GCM precipitation forecasts? *J. Climate*, **30**, 3185–3196, <https://doi.org/10.1175/JCLI-D-16-0652.1>.