Development and Calibration of Seasonal Probabilistic Forecasts of Ice-Free Dates and Freeze-Up Dates

ARLAN DIRKSON,^a BERTRAND DENIS,^b MICHAEL SIGMOND,^c AND WILLIAM J. MERRYFIELD^c

^a *Département des sciences de la Terre et de l'atmosphère, Université du Québec à Montréal, Montreal, Quebec, Canada*

^b *Environment and Climate Change Canada, Meteorological Services of Canada, Montreal, Quebec, Canada*

^c *Environment and Climate Change Canada, Canadian Center for Climate Modeling and Analysis, Victoria, British Columbia, Canada*

(Manuscript received 11 May 2020, in final form 6 October 2020)

ABSTRACT: Dynamical forecasting systems are being used to skillfully predict deterministic ice-free and freeze-up date events in the Arctic. This paper extends such forecasts to a probabilistic framework and tests two calibration models to correct systematic biases and improve the statistical reliability of the event dates: trend-adjusted quantile mapping (TAQM) and nonhomogeneous censored Gaussian regression (NCGR). TAQM is a probability distribution mapping method that corrects the forecast for climatological biases, whereas NCGR relates the calibrated parametric forecast distribution to the raw ensemble forecast through a regression model framework. For NCGR, the observed event trend and ensemble-mean event date are used to predict the central tendency of the predictive distribution. For modeling forecast uncertainty, we find that the ensemble-mean event date, which is related to forecast lead time, performs better than the ensemble variance itself. Using a multidecadal hindcast record from the Canadian Seasonal to Interannual Prediction System (CanSIPS), TAQM and NCGR are applied to produce categorical forecasts quantifying the probabilities for early, normal, and late ice retreat and advance. While TAQM performs better than adjusting the raw forecast for mean and linear trend bias, NCGR is shown to outperform TAQM in terms of reliability, skill, and an improved tendency for forecast probabilities to be no worse than climatology. Testing various cross-validation setups, we find that NCGR remains useful when shorter hindcast records (~20 years) are available. By applying NCGR to operational forecasts, stakeholders can be more confident in using seasonal forecasts of sea ice event timing for planning purposes.

SIGNIFICANCE STATEMENT: As Earth warms, the Arctic is shifting toward a longer open water season. With maritime access consequently increasing, stakeholders are valuing trustworthy information on the timing of transitional sea ice cover provided by seasonal forecasting models. In this study we advance seasonal predictions of the timing of local ice retreat and advance by extending these predictions to include critical information on forecast uncertainty. To do this, we tailor the established “ensemble model output statistics” calibration framework to sea ice retreat and advance dates, and construct probabilistic forecasts of early, normal, and late sea ice timing. Evaluating these predictions over a historical period indicates that stakeholders can place trust in forecast probabilities of sea ice timing for planning purposes.

KEYWORDS: Arctic; Sea ice; Statistical techniques; Ensembles; Probability forecasts/models/distribution; Seasonal forecasting

1. Introduction

Significant warming of the high northern latitudes in recent decades has been accompanied by substantial changes to Arctic sea ice (e.g., [IPCC 2019](#)). These include reductions in both ice coverage and ice thickness ([Maslanik et al. 2007](#); [Serreze et al. 2007](#); [Comiso et al. 2008](#); [Kwok et al. 2009](#); [Stroeve et al. 2012](#)), and a rapid transition from mainly multiyear to seasonal sea ice ([Maslanik et al. 2007, 2011](#); [Polyakov et al. 2012](#)). Additionally, shifts in sea ice melt onset in spring and summer and freeze onset in fall and winter have contributed to strong trends in the length of the open water season ([Comiso 2003](#); [Markus et al. 2009](#); [Stammerjohn et al. 2012](#); [Stroeve et al. 2014](#); [Parkinson 2014](#)). Despite these clear trends in the number of days a given location is ice free, the precise

timing of local sea ice retreat during the melt season and sea ice advance during the freeze season can exhibit large interannual fluctuations. The ability to predict these fluctuations is therefore potentially valuable, as the timing of varying sea ice conditions can inform end users for decision-making purposes (e.g., [Eicken 2013](#); [Stephenson and Pincus 2018](#)).

The Sea Ice Prediction Network has been collecting for the Sea Ice Outlook forecasts of the dates that locations across the Arctic become mainly free of sea ice, referred to as ice-free date (IFD) forecasts (see <http://www.arcus.org/sipn/sea-ice-outlook>). These IFD forecasts, and the analogous freeze-up date (FUD) forecasts made for the sea ice advance season, can be derived either statistically (e.g., [Stroeve et al. 2016](#)), or by using a dynamical forecasting system. [Sigmond et al. \(2016\)](#) provided the first formal assessment of seasonal forecast skill for IFDs and FUDs in such a forecasting system, the Canadian Seasonal to Interannual Prediction System (CanSIPS) ([Merryfield et al. 2013](#)), over a multidecadal period. They determined that

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

DOI: 10.1175/WAF-D-20-0066.1

© 2021 American Meteorological Society. For information regarding reuse of this content and general copyright information, consult the [AMS Copyright Policy](#) (www.ametsoc.org/PUBSReuseLicenses).

FUDs are better predicted than IFDs, with statistically significant interannual predictive skill for FUDs (IFDs) identified at lead times of 3.3 (2.2) months on average in the Arctic, whereas locally skill at longer lead times can be found. In CanSIPS, the predictability of ice retreat can be explained mostly by sea ice concentration (SIC) persistence, whereas the memory of the surface ocean temperatures is most relevant during the freeze season.

To date, IFD and FUD forecasts have only been considered deterministically; however, probabilistic forecasts would provide further value to users. In particular, owing to the intrinsic uncertainty in seasonal forecasts of sea ice from the chaotic nature of the climate system, uncertainty should be communicated by providing the full forecast probability distribution for the IFD or FUD at each location. Probabilistic forecasts are particularly relevant for risk-based decision-making (e.g., Richardson 2000; Buizza 2008; Weisheimer and Palmer 2014). The present study extends dynamical model-based IFD and FUD forecasts to include probabilistic information.

Before a seasonal forecast is issued, it is routine to adjust the forecast statistically for systematic model biases. For deterministic forecasts, this may be as simple as removing the model's mean climatology based on past forecasts, and issuing the forecast either as an anomaly or an anomaly added to the observed mean climatology (e.g., Boer 2009). More sophisticated regression-based methods also exist, generally referred to as model output statistics (MOS) (Wilks 2011). MOS can be tailored to specific applications, such as when model biases are nonstationary because of trends (Kharin et al. 2012), or when biases depend on the initial forecast state (Fućkar et al. 2014).

In addition to being unbiased, probabilistic forecasts should be as sharp (i.e., most certain) as possible, without compromising reliability (Gneiting et al. 2007) (i.e., the statistical consistency between forecast probabilities and observed frequencies). Calibration methods such as quantile mapping (e.g., Yuan and Wood 2013), Bayesian model averaging (Raftery et al. 2005), or ensemble MOS (EMOS) (Gneiting et al. 2005) may be applied to improve forecast reliability and reduce biases. Whereas quantile mapping considers biases between modeled and observed climatological distributions, Bayesian model averaging and EMOS take forecast–observation pairs and corresponding errors into account. Zhao et al. (2017) showed that doing so improves reliability compared to quantile mapping, and helps to ensure that skill is at least as good as climatology, a property known as coherence (Krzysztofowicz 1999).

Probabilistic forecasts of monthly sea ice coverage at the basin scale were calibrated in Krikken et al. (2016) using both extended logistic regression (ELR) (Wilks 2009) and heteroscedastic ELR (HELRL) (Messner et al. 2014b), the latter of which allows for the forecast ensemble spread to be incorporated into the calibration model. In both Dirkson et al. (2019b) (hereafter D19) and Director et al. (2019), methods were developed for calibrating probabilistic monthly SIC forecasts. D19 proposed trend-adjusted quantile mapping (TAQM), which first involves a nonstationary adjustment of the model and observed climatologies to account for trends. TAQM then

applies a parametric-based quantile mapping of the ensemble forecast from the model climatological distribution to the observed climatological distribution. Director et al. (2019) developed a “spatial calibration” procedure for probabilistic forecasts of the sea ice edge (i.e., the 15% SIC contour) called mixture contour forecasting (MCF). MCF adjusts the forecast ensemble-mean ice edge according to spatial discrepancies between the modeled and observed ice edges using a time-dependent linear regression model (Director et al. 2017). MCF then applies a form of Bayesian model averaging (Raftery et al. 2005) to generate calibrated ensemble ice-edge contours.

While it is possible to apply either TAQM or MCF to calibrate daily SIC forecasts, from which the IFD and FUD predictive distribution could be determined, the direct postprocessing of IFDs and FUDs has important advantages. First, the latter approach only requires a single correction per grid cell over the range of the forecast period to obtain the IFD or FUD, considerably lowering computational demands compared to correcting SIC fields for each day of the forecast range. In turn, the single correction reduces model complexity by reducing the number of parameters that need to be estimated during calibration. Furthermore, the calibration of daily SIC would require adaptations of the existing methods for calibrating monthly SIC, such as determining the number of days surrounding the forecast date to include for training, as well as requiring additional steps needed to obtain a parametric IFD or FUD probability distribution from a sequence of daily SIC forecast distributions. Finally, building a calibration model around the final predictand simplifies the model building process and the interpretability of the model itself. To calibrate probabilistic IFD and FUD forecasts, we therefore postprocess these dates after having computed them from uncorrected daily SIC forecasts.

In this study, we develop and compare two methods for calibrating probabilistic IFD and FUD forecasts. The first of these is a simple adaptation of TAQM, whereas the second is a new EMOS-based method. As will be shown, the EMOS-based method has several advantages over TAQM. In section 2, we describe the seasonal forecasting system used to make IFD and FUD forecasts, and provide some definitions regarding these dates. The two calibration methods are then described in section 3. In section 4, we propose a format for these probabilistic forecasts, and assess the impacts of postprocessing on forecast performance. Conclusions are provided in section 5.

2. Data and definitions

a. Seasonal forecasting system

Ensemble IFD and FUD forecasts are made with CanSIPS (Merryfield et al. 2013). CanSIPS is a multimodel system, employing two fully coupled global models: the Canadian Climate Model, version 3 (CanCM3) and version 4 (CanCM4). The land, ocean, and sea ice components of CanCM3 and CanCM4 are identical, whereas the atmospheric components in CanCM3 and CanCM4, respectively, are third- and fourth-generation atmospheric general circulation models (CanAM3 and CanAM4). CanAM3 and CanAM4 have horizontal grid

FIG. 1. Illustration of the relationship between forecast initialization date (the start of each arrow), the melt season (1 Apr–30 Sep), and the freeze season (1 Oct–31 Mar). The blue part of each arrow represents the periods from the initialization date to the start of the melt (freeze) season within which ice-free dates (freeze-up dates) are defined. The red part of each arrow represents the part of the melt or freeze season that is covered by the forecasts. Bounding event dates (minimum a values and maximum b values) are indicated for each initialization date.

spacings of roughly 2.8° , but differ in their vertical resolutions (31 levels for CanAM3 and 35 levels for CanAM4). The ocean in CanCM3/CanCM4 is simulated using the fourth generation ocean model (CanOM4), which has a 100-km nominal horizontal grid resolution with 40 vertical levels. Sea ice in CanCM3/CanCM4 is modeled as a cavitating fluid with a mean-thickness category (Flato and Hibler 1992) on the same horizontal grid as CanAM3/4.

In contrast to Sigmond et al. (2016), where IFDs and FUDs were assessed using retrospective forecasts (i.e., hindcasts) of standard CanSIPS over 1979–2010 [i.e., those described in Merryfield et al. (2011) and Merryfield et al. (2013)], we use a longer 1979–2018 hindcast period based on Mod-CanSIPS. In Mod-CanSIPS, sea ice initial conditions are derived from different sources than those used in standard CanSIPS. In particular, SIC is nudged toward a blend of SIC values from the Hadley Center Ice and Sea Surface Temperature, version 2 (HadISST2) (Titchner and Rayner 2014), and, where available, digitized Canadian Ice Service ice charts (Tivy et al. 2011) over 1979–2012. From 2013 to 2018, SIC is nudged toward the analysis employed for real-time predictions (as is done in standard CanSIPS), which assimilates several data sources including satellite data and ice charts (Buehner et al. 2013a,b, 2016). The change to the SIC initial conditions in Mod-CanSIPS increases temporal consistency with the analysis used for real-time predictions, and incorporates the various improvements in HadISST2 compared to a previous version of this analysis used in standard CanSIPS. Mean grid cell thickness is nudged toward values obtained from the SMv3 statistical model of Dirkson et al. (2017), which improves upon the model-based climatological thickness initialization used in standard CanSIPS.

Mod-CanSIPS hindcasts are initialized on the 1st day of each calendar month and integrated for 12 months. For each start date, an ensemble of 20 hindcasts (10 for CanCM3 and 10 for CanCM4) are run from slightly different initial states obtained from separate assimilation runs that start from different initial conditions, but are nudged toward the same observed states (Merryfield et al. 2013). For computing IFD and FUD values, we rely on daily outputs of SIC for each ensemble member.

b. Observations

Observed IFDs and FUDs are obtained from daily SIC data from the NOAA/NSIDC Climate Data Record of Passive Microwave Sea Ice Concentration, version 3 (Peng et al. 2013). When daily SIC are missing, linear interpolation is used to fill in values based on neighboring dates.

c. Ice-free date and freeze-up date definitions

We use the definitions for IFDs and FUDs employed in Sigmond et al. (2016). To aid in the description of these definitions, we refer the reader to the diagram in Fig. 1. In particular, IFDs and FUDs for both forecasts and observations are defined, respectively, as the first day during the melt season (1 April–30 September), or freeze season (1 October–31 March), that daily SIC falls below, or exceeds, 50%, and remains beyond this threshold for 10 consecutive days thereafter. For hindcasts initialized between 1 January and 1 June (1 July and 1 December), IFDs for the melt season of the current (following) year are considered. For initializations between 1 October and 1 December (1 January and 1 September), FUDs are defined for the current (following) freeze season.

In the event that IFD or FUD conditions are met at the time of initialization, but before the start of the defined melt and freeze seasons, we set IFD and FUD values to the start-of-season dates (i.e., 1 April for IFDs and 1 October for FUDs). Similarly, if retreat or advance conditions are not met by the end of the season, IFDs and FUDs are set to end-of-season values (i.e., 30 September for IFDs and 31 March for FUDs). If the advance or retreat has occurred by the initialization date, and the forecast is initialized within the melt or freeze seasons, values are set to the day prior to initialization. Finally, as IFD (FUD) forecasts initialized on 1 July and 1 August (1 January and 1 February) are for the following season in this study, but the 12-month forecast does not span the full length of that season, the event date is set to the last day of the forecast period if ice retreat (advance) conditions are not met.

The reason for imposing these conventions is to enable the inclusion of all hindcast years for both verification and statistical postprocessing. For example, in the event that an IFD is not forecasted to occur over the melt season, the forecast will be penalized less severely in verification if the observed date is closer to the end of the season compared to if the observed date is earlier in the melt season. To ensure that such cases do not overly influence verification scores, we only consider grid cells where the IFD or FUD occurred within the respective melt or freeze season in at least 75% of the verifying years, for a given initialization date, as in Sigmond et al. (2016). As will be described in section 3c, 1999–2018 is used for verifying probabilistic IFD and FUD hindcasts, so the event must be observed in at least 15 of those 20 years considered.

Because of these conventions, the minimum and maximum IFD and FUD allowed for a particular forecast are functions of the initialization date only. These values are given in Fig. 1, and are denoted hereafter as a for the minimum date, and b for the maximum date.

d. Systematic forecast biases

To motivate the need to calibrate IFD and FUD forecasts, Fig. 2 shows that climatological mean and linear-trend biases for Mod-CanSIPS are widespread for IFD and FUD hindcasts initialized on 1 April and 1 October, respectively, over the period 1999–2018. The mean absolute bias for IFDs and FUDs are, respectively, 19 and 28 days, with an overwhelming tendency for late freeze-up, as in the standard CanSIPS (Sigmond et al. 2016). Trends toward earlier ice break-up in recent years are underrepresented in Mod-CanSIP by, on average, 3.5 days decade⁻¹. During the freeze season, trends toward later advance are too strong in most regions, except for in the North Pacific and the North Atlantic. As will be shown in section 4c, after correcting IFD and FUD hindcasts for mean and trend bias, forecast probabilities remain highly overconfident. Thus, calibration is needed to offset biases and improve statistical reliability.

3. Calibration methods

We now describe the two methods developed to calibrate IFD and FUD forecasts. The first method, TAQM, is similar to that used in D19, but employs a new parametric probability distribution for modeling the ensemble forecast and climatological

distributions. As a second calibration procedure, we develop an EMOS-based method termed nonhomogeneous censored Gaussian regression (NCGR). In the following, section 3a introduces the parametric probability distributions used in TAQM and NCGR, section 3b provides an overview of each calibration method, and section 3c describes the cross-validation strategies employed.

a. Parametric probability distributions

Both TAQM and NCGR employ parametric distributions to model the forecast probability distribution. In addition to the forecast distribution, the modeled and observed climatological distributions are modeled parametrically for TAQM. As described in section 2c, IFDs and FUDs are numerically bounded by the end-dates a and b . Furthermore, as the probability for the preoccurrence (i.e., end-date a) and nonoccurrence (i.e., end-date b) of an IFD or FUD must be quantified explicitly, a suitable parametric distribution should represent:

- 1) the continuous range of dates on the open interval (a, b) ;
- 2) the double boundedness by a and b ;
- 3) the discrete probability at a or b .

Excluding the end-dates a and b , inspection of several IFD and FUD ensemble forecasts from both CanCM3 and CanCM4, as well as observed and modeled climatological distributions, reveals that these dates may be adequately modeled with a Gaussian/normal distribution (not shown). Since the normal distribution is defined on the entire real line, however, it does not satisfy points 2 or 3 above. In the following two sections we describe two variations of the normal distribution that satisfy all three requirements, the choice of which distribution to use depends on the calibration method under consideration.

1) DOUBLY INFLATED TRUNCATED NORMAL DISTRIBUTION

To apply TAQM to IFD and FUD forecasts in a similar manner as applied to monthly SIC forecasts in D19, we consider a parametric probability distribution analogous to the zero- and one-inflated beta distribution (Ospina and Ferrari 2010) used in that study. In particular, the forecast and climatological distributions for IFDs and FUDs are modeled here using a doubly inflated truncated normal (TNINF) distribution. Here, inflation refers to a mixture model that allows for quantifying the probability of end-dates a and b explicitly.

The probability density function (PDF) for a random variable Y with TNINF distribution is defined:

$$f_{\text{tninf}}(y; \mu, \sigma, \rho, \gamma) = \rho(1 - \gamma)\delta(y - a) + (1 - \rho)f_{\text{tnorm}}(y; \mu, \sigma) + \rho\gamma\delta(y - b), \quad (1)$$

where the parameter $0 \leq \rho \leq 1$ quantifies the probability of either a or b occurring. The individual end-dates are modeled with a discrete two-point distribution, such that, given an end-date has occurred, b is realized with probability γ , and a is realized with probability $1 - \gamma$. Point masses are converted to densities through multiplication by a dirac delta function $\delta(\cdot)$, shifted horizontally by a (first term) or b (third term).

FIG. 2. (top) The time-mean bias in days and (bottom) linear-trend bias in days per decade in (left) Mod-CanSIPS IFD and (right) FUD hindcasts over 1999–2018 (bias = model – observations). The example shown is for IFDs predicted from 1 Apr and FUDs predicted from 1 Oct. White and cyan correspond to regions where the observed date is defined in fewer than 15 out of 20 verification years.

The remaining $1 - \rho$ fraction of the TNINF distribution is described by a normal distribution, truncated on the interval $[a, b]$, with PDF:

$$f_{\text{tnorm}}(y; \mu, \sigma) = \frac{\phi\left(\frac{y - \mu}{\sigma}\right)}{\sigma \left[\Phi\left(\frac{b - \mu}{\sigma}\right) - \Phi\left(\frac{a - \mu}{\sigma}\right) \right]}, \quad a \leq y \leq b. \quad (2)$$

In Eq. (2), μ and σ are, respectively, the mean and standard deviation for a nontruncated normal distribution, and

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-(1/2)z^2}, \quad (3)$$

$$\Phi(z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-(1/2)z'^2} dz', \quad (4)$$

are, respectively, the PDF and cumulative distribution function (CDF) for the standard normal distribution, centered on zero with unit standard deviation. For reference, the expected value and standard deviation for the TNINF distribution are given in appendix A.

The PDF and CDF for a hypothetical IFD forecast initialized on 1 May are shown in Figs. 3a and 3b. For 1 May initialization,

$a = 120$ (30 April; day prior to initialization) and $b = 273$ (30 September; last day of the melt season) are, respectively, the minimum and maximum allowed IFDs. The TNINF distribution shown has parameter values $(\mu, \sigma, \rho, \gamma) = (240, 15, 0.2, 1)$, corresponding to a probability for the preoccurrence of the IFD of $\Pr(Y = a) = \rho(1 - \gamma) = 0\%$, and a probability for the non-occurrence of the IFD over the melt season of $\Pr(Y = b) = \rho\gamma = 20\%$. This results in a point mass on the PDF at $y = b$ of magnitude 0.2, and a jump discontinuity on the CDF at $y = b$ from $1 - \rho\gamma$ (i.e., 0.8) to 1. For the remaining IFDs on the interval (a, b) , the parameters μ and σ describe a gradual increase in cumulative probability from zero after approximately day-of-year (DOY) 215 (3 August).

2) DOUBLY CENSORED NORMAL DISTRIBUTION

EMOS-based methods require estimating regression coefficients that generally scale in number with the number of predictive distribution parameters. To keep NCGR simple, rather than use the relatively complex four-parameter TNINF distribution, we model the predictive distribution using a doubly censored normal (DCNORM) distribution (Cohen 1950).

The DCNORM distribution upholds the assumption that IFDs and FUDs on the interval (a, b) can be modeled effectively with a normal distribution, and explicitly represents the end-dates a and b using censoring (hence the name). For the DCNORM distribution, censoring refers to the conversion of probability density for the normal distribution for values outside the range (a, b) to point masses at those values. Censoring has been used when applying EMOS-based methods to other bounded variables, such as wind speed and precipitation (Thorarinsdottir and Gneiting 2010; Bentzien and Friederichs 2012; Messner et al. 2014a; Scheuerer and Hamill 2015).

For random variable Y' with normal distribution described by mean μ and standard deviation σ , censoring Y' below a and above b (with $b > a$) results in the random variable Y with DCNORM distribution. Its PDF is

$$f_{\text{denorm}}(y; \mu, \sigma) = \Phi\left(\frac{a - \mu}{\sigma}\right) \delta(y - a) + \frac{1}{\sigma} \phi\left(\frac{y - \mu}{\sigma}\right) 1_{(a,b)} + \left[1 - \Phi\left(\frac{b - \mu}{\sigma}\right)\right] \delta(y - b). \quad (5)$$

The first term in Eq. (5) quantifies the probability density at $y = a$, which is determined by the point mass of magnitude $\sigma^{-1} \int_{-\infty}^a \phi[(y' - \mu)/\sigma] dy' = \Phi[(a - \mu)/\sigma]$ and distribution $\delta(y - a)$. The second term in Eq. (5) quantifies the probability density for $a < y < b$, according to the PDF for a normal distribution, multiplied by the indicator function, such that $1_{(a,b)} = 1$ when $a < y < b$, and $1_{(a,b)} = 0$ otherwise. The third term in Eq. (5) quantifies the probability density at $y = b$, determined by the point mass of magnitude $\sigma^{-1} \int_b^{\infty} \phi[(y' - \mu)/\sigma] dy' = 1 - \Phi[(b - \mu)/\sigma]$ and distribution $\delta(y - b)$. When censoring leads to finite probability at either a or b , the expected value and standard deviation for the DCNORM distribution are no longer μ and σ , respectively. For reference, these statistics are provided in appendix B.

The DCNORM distribution for a hypothetical IFD forecast initialized on 1 May is illustrated in Figs. 3c and 3d.

FIG. 3. (a) PDF and (b) CDF for a TNINF probability distribution with parameters $(\mu, \sigma, \rho, \gamma) = (240, 15, 0.2, 1)$ for a synthetic ice-free date forecast initialized on 1 May (DOY 120); (c) PDF and (d) CDF for a DCNORM probability distribution with parameters $(\mu, \sigma) = (265, 20)$. The arrow in (c) illustrates censoring—the conversion of probability density (area shaded red) to a point mass (red circle). Values a and b are denoted on each panel by the vertical dashed lines. Red circles denote point masses in (a) and (c), and jump discontinuities in (b) and (d) (open = excluding, solid = including).

The distribution shown has parameter values $(\mu, \sigma) = (265, 20)$. To illustrate censoring, we show the PDF for the corresponding uncensored normal distribution, and draw an arrow from the censored portion of this distribution (shaded in red) to the point mass at $y = b$ having probability $\Pr(Y = b) = 34\%$. While technically the distribution is also censored below a , the probability $\Pr(Y = a)$ is essentially zero for this set of parameters.

b. Calibration

Before presenting the calibration methods, it is important to note that discrepancies between the forecast SIC and verifying SIC at the time of initialization can occur from inconsistencies between the verifying and assimilation data sources, or from the assimilation process itself. Therefore, for all forecasts we assess (calibrated and benchmark), in instances when ice break-up or advance has already occurred in the verifying observations on the day *prior* to forecast initialization, the forecast is set to indicate the event has also already occurred. Naturally, we only apply this correction for IFD (FUD) forecasts when initializing within the melt (freeze) season.

Throughout the description of TAQM and NCGR, we make use of the following notation. Let a forecast IFD or FUD be denoted generically as $x_i(t)$, where $i = 1, \dots, n$ is the ensemble member, and t is the forecast initialization year. In the Mod-CanSIPS two-model ensemble, the postprocessing methods are applied to each model separately, as is common procedure for multimodel ensembles (e.g., [Doblas-Reyes et al. 2005](#)); thus,

$n = 10$ here. The observed IFD or FUD is similarly denoted generically as $y(t)$.

We now provide an overview of the two calibration methods developed, TAQM and NCGR, with details provided in [appendix C](#) and [D](#), respectively. As will be shown, NCGR produces consistently better results than TAQM. Consequently, we have made the software to apply NCGR available for community use as a Python package: <https://github.com/adirkson/sea-ice-timing>.

1) TREND-ADJUSTED QUANTILE MAPPING

The TAQM calibration method is carried out analogously as it was applied to monthly SIC forecasts in [D19](#). As a first step of TAQM, the hindcast and observed event dates used for training are linearly de-trended and recentered on the linear regression of the original time series on the verifying year t . This step makes the training data stationary with respect to the long-term mean. Next, the trend-adjusted hindcasts and trend-adjusted observations, as well as the ensemble forecast to be calibrated, are separately fit to the TNINF distribution using maximum likelihood (ML) estimation.

Calibration is carried out on the discrete (end-dates) and continuous portions of the forecast probability distribution separately. Initially, the raw forecast probabilities, $\Pr[X(t) = a]$ and $\Pr[X(t) = b]$, are corrected for their biases in the trend-adjusted hindcasts compared to trend-adjusted observations. Through this correction, the calibrated forecast parameters ρ and γ are obtained. If the bias-corrected probability for either end-date becomes 100% from this step, calibration is complete. Otherwise, the raw forecast ensemble members,

$a < x_i(t) < b$, are quantile mapped from the trend-adjusted hindcast probability distribution to the probability distribution for the trend-adjusted observations. The calibrated forecast parameters μ and σ are then obtained by fitting the quantile-mapped ensemble members to the truncated normal distribution using ML estimation.

Along with the mathematical formulas used in these calculations, the steps taken when either the training observations, training hindcasts, or the raw forecast ensemble always predict the preoccurrence or nonoccurrence of the event with 100% probability are described in [appendix C](#).

2) NONHOMOGENEOUS CENSORED GAUSSIAN REGRESSION

For NCGR, the probability distribution for the observed IFD or FUD, conditioned on the raw forecast ensemble, is assumed to follow a DCNORM distribution; i.e., $Y(t)|x_1(t), x_2(t), \dots, x_n(t) \sim \mathcal{N}_{\text{dcnorm}}[\mu(t), \sigma(t)]$. The parameters $\mu(t)$ and $\sigma(t)$ are related to the raw forecast, as well as the training observations, through the following regression equations:

$$\mu(t) = \alpha_1 \mu_c(t) + \alpha_2 x_{(i)}^d(t), \quad (6a)$$

$$\sigma(t) = \beta_1 \sigma_c + \beta_2 x_{(i)}^{\text{tc}}(t), \quad (6b)$$

where $\alpha_1, \alpha_2, \beta_1$, and β_2 are coefficients to be estimated, and the mathematical expression for each predictor is given in [appendix D](#).

In Eqs. (6a) and (6b), the predictors μ_c and σ_c are intended to represent the corresponding statistics of the observed climatology, which is typically nonstationary. As such, we assume that the observed climatology can be described by a mean that varies linearly in time (i.e., μ_c), and a constant standard deviation (i.e., σ_c). While climate change studies suggest that sea ice variability can change as a function of mean state ([Goosse et al. 2009](#)), detecting those possible changes for IFDs and FUDs and modeling them through the term σ_c is impracticable with the small sample sizes available from satellite observations. Ideally, if the ensemble forecast is indistinguishable from the model climatology, or the remaining predictors in the regression equations are inconsequential, we would like $\mu \rightarrow \mu_c$ and $\sigma \rightarrow \sigma_c$ so that the predictive distribution approaches that of the nonstationary observed climatology. This asymptotic design is intended to promote the coherence property described in the introduction. Coefficients attached to μ_c and σ_c account for the uncertainty associated with estimating these statistics from observations.

Besides μ_c , the regression equation for μ , given by Eq. (6a), also depends on the forecast ensemble mean, as is common in EMOS formulations. In particular, we use $x_{(i)}^d(t)$, the forecast ensemble-mean anomaly relative to the linear trend of the forecast ensemble mean over the training period. By constructing Eq. (6a) in this way, μ is modeled as a linear function of the long-term observed trend, as well as interannual variations around the trend which scale with the de-trended forecast ensemble mean.

The calibrated forecast distribution spread, represented by σ [Eq. (6b)], first depends on σ_c , which represents the sample standard deviation of the de-trended observations (computed from the training sample). To relate σ to the raw forecast, we

use $x_{(i)}^{\text{tc}}(t)$, which represents the trend-corrected ensemble-mean event date. This allows forecast uncertainty to vary from year to year, hence the term “nonhomogeneous” in NCGR. The trend-corrected ensemble mean date is directly related to forecast lead time (i.e., later dates imply a longer time has passed between the forecast start date and the event date), and thus typically greater uncertainty. We find that $x_{(i)}^{\text{tc}}(t)$ predicts forecast error better than the standard deviation of the raw forecast ensemble, and that subsequently greater skill is achieved by using Eq. (6b) compared to two other candidates: one based on only the intercept-like term, σ_c , and one based on the ensemble standard deviation (see [appendix D](#) and [Figs. D1](#) and [D2](#)).

The coefficients in Eqs. (6a) and (6b) are estimated by minimizing the continuous rank probability score (CRPS) over the set of training hindcasts and observations, as is standard practice ([Gneiting et al. 2005](#)). This choice is motivated by the CRPS being sensitive to both forecast reliability and sharpness ([Gneiting and Raftery 2007](#)). The CRPS measures the difference between the forecast cumulative distribution $F(x)$ and the observation y according to

$$\text{crps}(F, y) = \int_{-\infty}^{+\infty} [F(x) - H(x - y)]^2 dx, \quad (7)$$

where $H(x - y)$ is the Heaviside step function. The CRPS ranges on the interval $[0, \infty)$, with zero representing a perfect forecast; hence, it is optimized at its minimum value. An analytic expression for Eq. (7) when $F(x)$ is the CDF for the DCNORM distribution assists with computational expedience [see Eq. (D2)]. [Appendix D](#) contains further details on the software and implementation of the numerical minimization.

c. Cross validation

[Figure 4](#) summarizes the three cross-validation schemes we have tested to train, calibrate, and verify probabilistic IFD and FUD hindcasts from Mod-CanSIPS. For all three setups, a common 20-yr period from 1999 to 2018 is used for verification. For our primary evaluation, TAQM and NCGR are trained using all hindcasts and observations from 1979 through the year prior to the year being calibrated and verified. In this way, we emulate an operational forecasting scenario where only past data are available for postprocessing real-time forecasts. This framework is referred to as “operational-long” and is illustrated in the top panel of [Fig. 4](#).

As shorter hindcast records of 20–30 years exist for other seasonal forecasting systems (e.g., those available on the Copernicus Climate Data Store; <https://cds.climate.copernicus.eu>), typically data from all years excluding the verifying year are used to apply EMOS-type methods like NCGR. However, this leave-one-out cross-validation approach is less justifiable when the data used for training exhibit trends, as is the case here for IFDs and FUDs, since the assumption of independence between the training and testing data is violated. While it is common to estimate relative forecast accuracy (i.e., skill) in such situations using reference forecasts based on the same nonindependent training data ([Kharin et al. 2012](#); [Fućkar et al. 2014](#); [Sansom et al. 2016](#);

FIG. 4. Schematics of the three cross-validation setups used to train, calibrate, and verify IFD and FUD hindcasts. The 1999–2018 period over which hindcasts are calibrated and verified is denoted by the blue line in each diagram. Five verifying years are shown as examples: 1999, 2000, 2007, 2017, and 2018 (blue circles). Each red line corresponds to the training period for the given hindcast year being calibrated and verified.

Krikken et al. 2016), such an approach may still not accurately represent the forecast skill achieved operationally.

To enable forecasting centers using leave-one-out cross-validation to put skill scores in context, we leverage the longer record used here to test the fidelity of TAQM and NCGR using two additional cross-validation frameworks. The first setup, referred to as “operational-short,” is analogous to operational-long, but only uses the previous 19 years of hindcasts/observations for training (Fig. 4, middle panel). The second scheme uses leave-one-out cross-validation also with a 19-yr training record to control for sample size, and will be referred to as “leave-one-out-short” (Fig. 4, bottom panel).

4. Verification results

a. Benchmark forecasts

To evaluate probabilistic IFD and FUD hindcasts calibrated with TAQM and NCGR we consider two benchmark forecasts, whose regression coefficients are computed using the same cross-validation setup as the calibrated hindcasts.

We consider a “trend-corrected forecast” as a first benchmark, which is the raw forecast adjusted for mean and linear trend bias (i.e., no ensemble spread adjustment). The i th trend-corrected ensemble member, x_i^{tc} , is computed according to

$$x_i^{\text{tc}}(t) = x_i(t) + [\bar{y}(t) - \bar{x}_{(i)}(t)], \quad i = 1, \dots, n, \quad (8)$$

where the tilde symbol $\tilde{\cdot}$ denotes the linear regression of the given variable on time (i.e., year), and $\langle i \rangle$ denotes the ensemble mean.

As a second benchmark, we consider a trend-adjusted observed climatology, computed using Eq. (C1a), and referred to hereafter as TA-clim. TA-clim changes as a function of the verifying year in order to represent the long-term trend, and to avoid an overdispersive climatological distribution (D19).

Predictive skill relative to TA-clim implies accurate predictions of interannual variations of the forecast probability distribution, as well as an accurate representation of the observed long-term trend.

b. Categorical forecasts

Seasonal probabilistic forecasts of Arctic sea ice coverage are typically displayed as a map showing probabilities that monthly averaged SIC exceeds 15%, also known as sea ice probability (SIP) (e.g., Stroeve et al. 2015). This type of binary categorical forecast is easily interpreted by users who require knowing the location of the ice edge. For probabilistic IFD and FUD forecasts, however, the timing of varying sea ice conditions can be naturally understood relative to a local climatological timing.

We adopt a similar format to that often used for probabilistic forecasts of temperature and precipitation, in which a map displays probabilities for the most likely of three climatologically equiprobable categories: below normal, near normal, or above normal. For the probabilistic IFD and FUD forecasts displayed here, the three tercile categories are labeled as: early, normal, or late. An example of such a forecast is shown in the left-hand panels of Fig. 5 for an IFD hindcast initialized on 1 May 2016, using the trend-corrected benchmark [i.e., Eq. (8)], TAQM, and NCGR.

The calculation of the probability for each category is illustrated in the right-hand panels of Fig. 5 for a grid cell in the Chukchi Sea, marked by the star in each map. The PDF for the hindcast IFD is plotted along with vertical lines representing the terciles of the observed climatology. Whereas observed climatologies are frequently defined with respect to a 1981–2010 baseline (WMO 2017), we use a shorter period consisting of the 10 years directly preceding the hindcast year of interest, which for this example corresponds to 2006–15. This shorter record helps ensure that a particular category is not overwhelmingly preferred by the predictive distribution due to

FIG. 5. A categorical IFD forecast initialized on 1 May 2016. (left) Maps of the probabilities associated with the most likely of three categories (early, normal, late) with respect to 2006–15 climatology; equal chances (EC); (right) forecast PDFs and the observed climatological 1/3 and 2/3 terciles (vertical dashed lines) used to compute forecast probabilities for each category for a grid cell in the Chukchi Sea (star on map). Colors: red = early; green = normal; blue = late. The probability for each category is given by the value above each bar in the top right.

trends, and better reflects the fact that the observed climatology is nonstationary. In addition, users may prefer this since it is representative of the conditions experienced most recently. The early, normal, and late climatological categories are defined according to the terciles of those ten observed dates, such that each category has a 1/3 probability of occurring. Due to the small sample size, terciles are estimated with the Harrell–Davis quantile estimator (Harrell and Davis 1982). Probabilities for each category are illustrated by the shaded areas under the hindcast PDF that overlap each category determined by the climatological terciles.

c. Dependence of reliability on postprocessing

The correspondence between postprocessing and the reliability of probabilistic IFD and FUD hindcasts is assessed in Fig. 6 with the attributes diagram (e.g., Jolliffe and Stephenson 2012).

The reliability curve on each diagram is plotted by binning hindcast probabilities for early, normal, and late events to the nearest 10%, and computing the observed frequency for each forecast probability. We quantify reliability and resolution through the Brier score (Brier 1950) decomposition: $BS = BS_{rel} - BS_{res} + BS_{unc}$ (Murphy 1973), computed separately for each event. Smaller values of BS_{rel} indicate better reliability (i.e., a closer proximity of the reliability curve to the 1:1 line). Larger values of BS_{res} indicate better resolution, describing the ability of the forecast to distinguish between different observed outcomes. The Brier score uncertainty term, BS_{unc} , is only a function of the observations and therefore does not change between the postprocessed forecasts; thus, we do not consider it here. Results for the melt (freeze) season are summarized by aggregating hindcasts initialized on the 1st of March–June (September–December).

Probabilistic IFD and FUD hindcasts postprocessed with trend correction, TAQM, and NCGR are typically overconfident, as can be seen by the tendency for low and high probabilities to be overpredicted. Compared to trend correction, the better agreement between the reliability curves and the 1:1 line for the TAQM-calibrated hindcasts, and the generally lower BS_{rel} values, indicate that TAQM is able to improve reliability. Calibration with NCGR considerably improves reliability compared to both trend correction and TAQM, as indicated by the reliability curves for each event more closely following the 1:1 line, and the much smaller BS_{rel} values. According to the BS_{res} values, TAQM improves resolution compared to trend correction. NCGR shows poorer resolution than trend correction and TAQM for IFDs, but better resolution for FUDs. Although fewer forecast probabilities near 0% and 100% are issued for the NCGR-calibrated hindcasts (relative size of circles), indicating less sharp and confident predictive distributions, the larger numbers of these probabilities in the trend-corrected and TAQM-calibrated hindcasts are associated with poorer reliability.

d. Dependence of skill on postprocessing

We now investigate the degree to which the different postprocessing methods impact the predictive skill for probabilistic IFD and FUD hindcasts. To simplify this analysis, rather than consider the early, normal, and late categories separately, we evaluate the accuracy of the full predictive distribution. In particular, hindcast skill is quantified with the CRPSS, which is defined as

$$CRPSS(\%) = \left(1 - \frac{CRPS_{fct}}{CRPS_{ref}}\right) \times 100, \quad (9)$$

and quantifies the percent improvement of the CRPS [Eq. (7)] for a target forecast (“fct”) relative to a reference forecast (“ref”). Here, the target forecast is the trend-corrected, TAQM-calibrated, or NCGR-calibrated hindcasts, and the reference forecast is TA-clim.

The spatial distributions of probabilistic skill for IFD and FUD hindcasts initialized on the same start dates used for assessing reliability are shown in Figs. 7 and 8, respectively. Summary

FIG. 6. Attributes diagrams for probabilistic (top) IFD and (bottom) FUD hindcasts of each of the three categorical events: red = early; green = normal; blue = late. Vertical dashed lines are hindcast climatological probabilities, and horizontal dashed lines are observed climatological frequencies. The size of each circle is proportional to the relative number of hindcasts for which that probability was issued for a given event. The BS_{rel} and BS_{res} values in each panel represent the reliability and resolution terms for the Brier score, respectively.

metrics on each map include the spatially averaged CRPSS (black text), and the area percent that $CRPSS > 5\%$ (red text). During the melt season (Fig. 7), all three methods result in somewhat comparable levels of skill, with TAQM showing a small net decrease in skill relative to trend correction, and NCGR a net increase in skill. For the freeze season (Fig. 8), summary metrics indicate that hindcast skill for a given initialization month improves sequentially with trend correction, TAQM, and NCGR, with substantially improved skill using NCGR.

For both IFD and FUD hindcasts, certain regions are consistent in showing some degree of skill, independent of the postprocessing applied; however, nearly always the skill in each of these regions is higher for TAQM compared to trend correction, and higher still for NCGR compared to TAQM. Improvement using NCGR tends to be greatest for start dates for which overall predictive skill is higher (i.e., June for IFDs and all start dates for FUDs compared to IFDs). In regions where the trend-corrected and TAQM-calibrated hindcasts show considerably worse skill than TA-clim (highly negative CRPSS values), reflecting poor coherence, skill is more similar to TA-clim using NCGR (CRPSS values near zero).

As in Zhao et al. (2017), we find that locations with poor coherence tend to correspond to locations where the anomaly

correlation between raw ensemble-mean hindcasts and observations (after linearly de-trending) is not statistically significant, based on a one-sided t test ($p < 0.05$; Fig. 9). However, even when correlations are not statistically significant, NCGR tends to produce more neutrally skillful forecasts. Furthermore, when correlations are statistically significant, probabilistic skill is not guaranteed for the trend-corrected and TAQM-calibrated hindcasts, whereas it becomes much more likely for the NCGR-calibrated hindcasts, particularly for FUDs.

To summarize probabilistic skill across all start dates, Fig. 10 shows maximum average lead times that skill can be obtained for IFD and FUD hindcasts. These lead times are computed as the difference between the climatological date of the event (over 1999–2018), and the first start date that skill is present and is maintained for subsequent start dates. Here, we define $CRPSS > 5\%$ as indicating skill. Averaged over the Arctic, probabilistic skill can be obtained for IFD hindcasts 0.8 (0.8 months) in advance after postprocessing with trend correction (TAQM), whereas skill arises ~ 2 weeks earlier (1.2 months in advance) using NCGR. As in Sigmond et al. (2016), FUDs are found to be predicted skillfully earlier than IFDs, and in particular at lead times of 1.3, 2.1, and 2.2 months, respectively, for

Ice-Free Dates

FIG. 7. Maps of the CRPSS (skill scores) for IFD hindcasts initialized on the date as labeled. Blue CRPSS values indicate forecast accuracy worse than TA-clim (see main text for definition), and red CRPSS values indicate skillful forecasts. Percentages in the top right in each diagram are the mean CRPSS (black) and area percent that CRPSS > 5% (red). White and cyan correspond to regions where the observed date is defined in fewer than 15 out of 20 verification years.

FIG. 8. As in Fig. 7, but for FUD hindcasts.

FIG. 9. Scatterplots of CRPSS values from Figs. 7 and 8 (same color scaling), aggregated across all four start dates, vs the anomaly correlation coefficient between uncorrected ensemble-mean hindcasts and observations (after de-trending). Horizontal dashed line delineates skillful forecasts (positive CRPSS) and unskillful forecasts (negative CRPSS). Vertical dashed line represents the threshold for statistically significant correlation, based on a one-sided t test ($p < 0.05$).

trend correction, TAQM, and NCGR. Skill at relatively long lead times ranging from 2 to 6 months is seen for IFD hindcasts calibrated with NCGR in the Bering Sea, Hudson Bay, and Baffin Bay. For FUDs, skill ≥ 2 months in advance is widespread, particularly in the Siberian Arctic where it is absent for IFDs.

e. Sensitivity to cross validation

Having established that NCGR calibration performs better than TAQM for the probabilistic IFD and FUD forecasts assessed here, we now investigate the sensitivity of the skill scores to cross-validation. Figures 11 and 12 show the CRPSS for IFD and FUD hindcasts, respectively, calibrated with

NCGR using each cross-validation strategy described in section 3c and illustrated in Fig. 4.

First, by comparing the operational-long and operational-short cross-validated hindcasts (top rows of Figs. 11 and 12) we find that probabilistic forecast accuracy is better when using a longer training period versus a 19-yr training window (i.e., local CRPSS values are mainly positive), with spatially averaged improvements ranging from 2% to 4%. While probabilistic accuracy for the operational-long cross-validated hindcasts is higher than operational-short, we find that when NCGR-calibrated hindcasts are assessed relative to the TA-clim reference forecast, skill results are generally consistent for both cross-validation schemes [comparing

FIG. 10. Maximum lead times (in months) that predictive skill can be obtained over the (top) melt season and (bottom) freeze season. Lead times less than zero (in blue) correspond to when probabilistic skill is not identified at the shortest possible lead time. Value at the top right of each panel is the area-averaged lead time. White and cyan correspond to regions where the observed date is defined in fewer than 15 out of 20 verification years.

the middle row of Fig. 11 (Fig. 12) to the bottom row of Fig. 7 (Fig. 8)]. This implies that the accuracy of the reference forecast, TA-clim, also benefits from using a longer historical record.

As stated previously, some forecasting centers employ shorter hindcast records, and may require using leave-one-out cross-validation when applying NCGR to estimate hindcast skill. Comparing skill scores for the NCGR-calibrated hindcasts against TA-clim using each the operational-short and leave-one-out-short cross-validation setups (respectively, the middle and bottom rows in Figs. 11 and 12), we find that averaged over the Arctic, leave-one-out-short skill scores can be inflated by up to $\sim 5\%$ when averaged over the Arctic. The 50%, 75%, and 95% percentiles for this skill inflation for IFDs (FUDs) are,

respectively, 4%, 14%, and 28% (-3% , 8%, and 19%). However, inflation is not ubiquitous since higher skill using operational-short is also sometimes observed, implying that the TA-clim reference forecast using leave-one-out cross-validation is inflated somewhat similarly to NCGR-calibrated hindcasts.

5. Conclusions

Two calibration methods have been developed and tested here for improving seasonal probabilistic forecasts of the timing of local sea ice retreat and advance. Calibration is performed on the event dates themselves, after having computed them from raw ensemble daily SIC forecasts. These ensemble IFD and FUD forecasts were produced using Mod-CanSIPS,

FIG. 11. CRPSS for NCGR-calibrated IFD hindcasts using operational-long, operational-short, and leave-one-out-short cross-validation setups. Each row compares “fst” vs “ref” [as in Eq. (9)]. For example, in the first row positive CRPSS means better performance using operational-long compared to operational-short. Values at the top right of each map are as in Figs. 7 and 8. White and cyan correspond to regions where the observed date is defined in fewer than 15 out of 20 verification years.

FIG. 12. As in Fig. 11, but for FUD hindcasts.

which is an experimental version of CanSIPS with improved sea ice initialization.

The first method is a simple modification of TAQM (D19), which, as for its original application to monthly SIC forecasts, is well suited to handle the doubly bounded and nonstationary IFD and FUD forecasts considered here. The modification involves replacing the probability distribution used in D19 with a doubly inflated truncated normal distribution.

The second method, NCGR, is novel to this study and improves upon TAQM. The predictive distribution in NCGR is assumed to be adequately modeled using the two-parameter DCNORM distribution. The parameter of this distribution that describes central tendency is modeled as a linear combination of the observed trend, and the de-trended forecast ensemble-mean anomaly. The parameter associated with the predictive uncertainty is modeled as a linear combination of the observed climatological standard deviation (after the removal of trends), and the trend-corrected ensemble-mean date. We find that the latter is a stronger predictor of forecast error than the ensemble variance itself. This is due to a relationship between the ensemble-mean IFD/FUD and forecast lead time, such that later event dates relative to the forecast start date are typically associated with larger errors. The coefficients relating the DCNORM distribution parameters to these different predictors are found by minimizing the CRPS.

To assess predictive performance after applying TAQM and NCGR, we compare the calibrated hindcasts against a simple uniform adjustment of the hindcast ensemble based on historical errors in trend and mean bias, following Kharin et al. (2012). While reliability is improved using TAQM compared to such a trend correction, both postprocessing methods result in overconfident probabilistic forecasts, particularly for probabilities near 0% and 100%. NCGR on the other hand is able to produce much more reliable forecasts, although some overconfidence remains.

Probabilistic skill relative to a trend-adjusted observed climatology is also improved using TAQM compared to trend correction; however, the improvement obtained using NCGR is notably better. Part of this improvement comes from the ability of TAQM and NCGR to boost and expand the coverage of predictive skill. Improvements also result from enhanced coherence, i.e., the tendency for probabilistic forecasts to be no worse than climatology. Despite this general improvement using either calibration method, substantially enhanced coherence is achieved with NCGR.

Using various cross-validation strategies, we find that NCGR benefits from using more past data to train and estimate calibration coefficients, but remains useful for forecasting systems limited by shorter (~ 20 year) hindcast records. If by practical necessity NCGR is applied using leave-one-out cross validation, our results suggest that, assuming an appropriate trend-adjusted climatological reference forecast is used, skill estimates may be overestimated by up to 5% (when averaged over the Arctic) relative to an operational scenario. Even so, spatial patterns appear to be largely robust and independent of the cross-validation employed.

To potentially further improve NCGR, future research could incorporate the inclusion of information from neighboring grid cells in order to mitigate parameter overfitting and improve the spatial smoothness and quality of forecast probabilities. Techniques for doing so have proven to be beneficial in other forecast calibration and downscaling contexts (e.g., Cannon 2008; Scheuerer 2014; Kharin et al. 2017).

Acknowledgments. We thank three anonymous reviewers for their constructive feedback and for improving the manuscript. We also thank Reinel Sospedra-Alfonso for helpful

comments on an earlier version of this manuscript, as well as Woosung Lee and Cathy Reader for their assistance with data preparation. Denis and Merryfield acknowledge funding from the Marine Environmental Observation Prediction and Response (MEOPAR) and Polar Knowledge Canada. This is a contribution to the Year of Polar Prediction (YOPP), a flagship activity of the Polar Prediction Project (PPP), initiated by the World Weather Research Programme (WWRP) of the World Meteorological Organization (WMO).

Data availability statement. NCGR is available as a Python package at <https://github.com/adirkson/sea-ice-timing>, which also contains the code used to perform TAQM. Mod-CanSIPS hindcast daily SIC data are available at <ftp://crd-data-donnees-rdc.ec.gc.ca/pub/CCCMA/bmerryfield/goapp/goapp/SIPN2/<MM>/<MODEL>/SIC/daily/> in subdirectories <MM>/<MODEL>/SIC/daily, where <MODEL> can be either CanCM3 or CanCM4, and <YYYY> and <MM> are the hindcast initialization year and month, respectively. Filenames for 1979–2012 are given by `sic_daily_CCCma-<MODEL>_NEW_original_grid_i<YYYY><MM>01.nc`, and for 2013–18 by `sic_daily_CCCma-<MODEL>_NEW_OPER_original_grid_i<YYYY><MM>01.nc`.

APPENDIX A

Doubly Inflated Truncated Normal Distribution

a. Maximum likelihood estimation

Following the derivation of the ML equations for the BEINF distribution provided in Ospina and Ferrari (2010), the likelihood function for the TNINF distribution with parameter vector $\theta = (\mu, \sigma, \rho, \gamma)$ given the sample $y_1, \dots, y_n$ is

$$L(\theta) = \prod_{i=1}^n f_{\text{tninf}}(y_i; \mu, \sigma, \rho, \gamma) = L_1(\rho) L_2(\gamma) L_3(\mu, \sigma),$$

where

$$L_1(\rho) = \prod_{i=1}^n \rho^{1_{\{0,1\}}(z_i)} (1 - \rho)^{1 - 1_{\{0,1\}}(z_i)} = \rho^{T_1} (1 - \rho)^{n - T_1},$$

$$L_2(\gamma) = \prod_{i: z_i \in \{0,1\}} \gamma^{z_i} (1 - \gamma)^{(1 - z_i)} = \gamma^{T_2} (1 - \gamma)^{(T_1 - T_2)},$$

$$L_3(\mu, \sigma) = \prod_{i: y_i \in (a,b)} f_{\text{tnorm}}(y_i; \mu, \sigma).$$

In the expressions above, $z_i = (y_i - a)/(b - a)$ is a transformed variable such that $z_i = 0$ when $y_i = a$ and $z_i = 1$ when $y_i = b$, $T_1 = \sum_{i=1}^n 1_{\{0,1\}}(z_i)$ is the number of y_i that are either a or b , and of those values, $T_2 = \sum_{i=1}^n z_i 1_{\{0,1\}}(z_i)$ is the number that equal b . The log-likelihood is

$$\log[L(\theta)] = l_1(\rho) + l_2(\gamma) + l_3(\mu, \sigma),$$

where

$$l_1(\rho) = T_1 \log \rho + (n - T_1) \log(1 - \rho),$$

$$l_2(\gamma) = T_2 \log \gamma + (T_1 - T_2) \log(1 - \gamma),$$

$$l_3(\mu, \sigma) = (n - T_1) \log[f_{\text{tnorm}}(y_i, \mu, \sigma)].$$

Setting $\partial l_1/\partial \rho = 0$ and $\partial l_2/\partial \gamma = 0$, the ML estimates for ρ and γ are found to be $\hat{\rho} = T_1/n$ and $\hat{\gamma} = T_2/T_1$. Analytic expressions for $\hat{\mu}$ and $\hat{\sigma}$ cannot be found when setting $\partial l_3/\partial \mu = 0$ and $\partial l_3/\partial \sigma = 0$; instead these estimates are found numerically using the `stats.truncnorm.fit` function in Python's Scipy library (Jones et al. 2001).

b. Expected value

The expected value is

$$E(Y) = \int_{-\infty}^{\infty} y f_{\text{tnorm}}(y; \mu, \sigma, \rho, \gamma) dy$$

$$= a\rho(1 - \gamma) + (1 - \rho)E(Y') + b\rho\gamma, \quad (\text{A1})$$

where the random variable Y' has truncated normal distribution with expected value

$$E(Y') = \mu + \frac{\phi(a^*) - \phi(b^*)}{\Phi(b^*) - \Phi(a^*)} \sigma \quad (\text{A2})$$

(Johnson et al. 1994), and $*$ denotes a standardized variable [e.g., $z^* = \sigma^{-1}(z - \mu)$].

c. Standard deviation

The standard deviation is given by

$$\text{SD}(Y) = \sqrt{\int_{-\infty}^{\infty} [y - E(Y)]^2 f_{\text{tnorm}}(y; \mu, \sigma, \rho, \gamma) dy}$$

$$= \sqrt{Z_1 - E(Y)^2}, \quad (\text{A3})$$

where $E(Y)$ is given by (A1), and

$$Z_1 = a^2\rho(1 - \gamma) + (1 - \rho)[E(Y') + \text{Var}(Y')]^2 + b^2\rho\gamma. \quad (\text{A4})$$

In Eq. (A4), $E(Y')$ is given by Eq. (A2) and

$$\text{Var}(Y') = \sigma^2 \left\{ 1 + \frac{a^*\phi(a^*) - b^*\phi(b^*)}{\Phi(b^*) - \Phi(a^*)} - \left[\frac{\phi(a^*) - \phi(b^*)}{\Phi(b^*) - \Phi(a^*)} \right]^2 \right\} \quad (\text{A5})$$

is the variance of random variable Y' with truncated normal distribution (Johnson et al. 1994).

APPENDIX B

Doubly Censored Normal Distribution

a. Expected value

The expected value is

$$E(Y) = \int_{-\infty}^{\infty} y f_{\text{dcnorm}}(y; \mu, \sigma) dy$$

$$= a\Phi(a^*) + \mu[\Phi(b^*) - \Phi(a^*)]$$

$$+ \sigma[\phi(a^*) - \phi(b^*)] + b[1 - \Phi(b^*)]. \quad (\text{B1})$$

b. Standard deviation

The standard deviation is given by

$$\text{SD}(Y) = \sqrt{\int_{-\infty}^{\infty} [y - E(Y)]^2 f_{\text{dcnorm}}(y; \mu, \sigma) dy}$$

$$= \sqrt{Z_2 - E(Y)^2}, \quad (\text{B2})$$

where $E(Y)$ is given by Eq. (B1), and

$$Z_2 = a^2\Phi(a^*) + \sigma^2[\Phi(b^*) - b^*\phi(b^*) - \Phi(a^*) + a^*\phi(a^*)]$$

$$+ 2\mu\sigma[\phi(a^*) - \phi(b^*)] + \mu^2[\Phi(b^*) - \Phi(a^*)]$$

$$+ b^2[1 - \Phi(b^*)].$$

APPENDIX C

Trend-Adjusted Quantile Mapping

a. Trend adjustment and distribution fitting

Observed and hindcast climatological distributions over the training period are adjusted for their long-term linear trends according to

$$y'(t) = \bar{y}(t) + [y(\tau_t) - \bar{y}(\tau_t)], \quad (\text{C1a})$$

$$x'_i(\tau_t) = \bar{x}_{(i)}(t) + [x_i(\tau_t) - \bar{x}_{(i)}(\tau_t)], \quad i = 1, \dots, n, \quad (\text{C1b})$$

where $\langle i \rangle$ denotes the ensemble mean, and $\bar{\cdot}$ denotes the linear least squares fit of the corresponding variable over the set of years in the training period τ_t . Trend-adjusted data that fall outside the interval $[a, b]$ are clipped to the appropriate lower or upper bound. Trend adjustment is only performed when linear trends are statistically significant at the 95% confidence level, based on a two-sided t test.

Next, trend-adjusted values $y'(\tau_t)$ and $x'_1(\tau_t), \dots, x'_n(\tau_t)$, as well as the ensemble forecast $x_1(t), \dots, x_n(t)$, are each fit to a TNINF distribution using ML estimation (as in appendix A).

b. Calibrated forecast distribution

For the calibration of the forecast probability distribution, we first compute calibrated parameters $\rho(t)$ and $\gamma(t)$ according to

$$\rho = \rho_x + \rho_{y'} - \rho_{x'}, \quad (\text{C2a})$$

$$\gamma = \frac{\rho_x \gamma_x + \rho_{y'} \gamma_{y'} - \rho_{x'} \gamma_{x'}}{\rho}, \quad (\text{C2b})$$

where the subscripts denote which data the distribution parameters have been fit to from the previous step. These equations are equivalent to adjusting the raw forecast probabilities $\text{Pr}[X(t) = a]$ and $\text{Pr}[X(t) = b]$ according to the climatological hindcast bias over the training period (see D19). If through this step, $\rho = 1$, calibration is considered complete as this implies a 100% probability for the end-date a or b .

If $\rho \neq 1$, the raw ensemble members $a < x_i(t) < b$ are quantile mapped as

$$\hat{x}_i(t) = F_{\text{tnorm}}^{-1(\text{obs})}\{F_{\text{tnorm}}^{(\text{mod})}[x_i(t)]\}, \quad i = 1, \dots, n - m. \quad (\text{C3})$$

Here, $F_{\text{tnorm}}^{(\text{mod})}[x_i(t)] = r$ is the CDF of the truncated normal distribution for the trend-adjusted hindcast climatology, and evaluated at the i th raw ensemble member; $F_{\text{tnorm}}^{-1(\text{obs})}(r)$ is the inverse CDF of the truncated normal distribution for the trend-adjusted observed climatology, evaluated at the same cumulative probability r . The quantile-mapped ensemble members $\hat{x}_1(t), \dots, \hat{x}_{n-m}(t)$ are then fit to a truncated normal distribution to derive the calibrated parameters μ and σ .

If either all observations or trend-adjusted observations over τ_i are equal to a or b , all forecast ensemble members are set to the corresponding end-date. In another case, if all ensemble members predict the end-dates a or b with 100% probability [i.e., $n - m = 0$ in Eq. (C3)], quantile mapping cannot be performed. In such cases, the training hindcasts may also either always predict that end-date, or they may show variability in the forecast event date. If the former, the calibrated forecast is set to the trend-adjusted observed climatology [i.e., Eq. (C1a)]. If the latter, the raw forecast is replaced with the conditional distribution of observed dates corresponding to when that same end-date was forecast to occur with 100% probability over the training period, as suggested in Dirkson et al. (2019a). When such an event appears only rarely in the training hindcasts (in <5 of the years in τ_i), we simply trust the raw ensemble forecast and do not apply any postprocessing.

APPENDIX D

Nonhomogeneous Censored Gaussian Regression

a. Regression equations

The following describes the definitions of the various terms tested in the regression equations explored for the NCGR model:

- $\mu_c(t) = \tilde{y}(t) :=$ linear regression of the observed date onto year t ;
- $\sigma_c(t) = \text{SD}[y(\tau_i) - \tilde{y}(\tau_i)] :=$ standard deviation of the linearly de-trended observations;
- $x_{(i)}^d(t) = x_{(i)}(t) - \tilde{x}_{(i)}(t) :=$ linearly de-trended forecast ensemble mean;
- $x_{(i)}^{\text{tc}}(t) = \tilde{y}(t) + [x_{(i)}(t) - \tilde{x}_{(i)}(t)] :=$ trend-corrected forecast ensemble mean [following Kharin et al. (2012)];
- $s_x(t) = \text{SD}[x_i(t)] :=$ forecast ensemble standard deviation.

Linear least squares fits are only applied in the calculations above when the linear trend for the variable of interest is statistically significant over τ_i ($p < 0.05$). Otherwise, the climatological mean is used instead. In the computation of $x_{(i)}^{\text{tc}}(t)$, if the resultant value falls outside the interval $[a, b]$, it is set to the appropriate end-date.

We tested the following three regression equations for the parameter σ , denoted σ_I , σ_{II} , and σ_{III} before determining the final NCGR model presented in the main text:

$$\sigma_I(t) = \beta_1 \sigma_c, \quad (\text{D1a})$$

$$\sigma_{II}(t) = \beta_1 \sigma_c + \beta_2 s_x(t), \quad (\text{D1b})$$

$$\sigma_{III}(t) = \beta_1 \sigma_c + \beta_2 x_{(i)}^{\text{tc}}(t). \quad (\text{D1c})$$

The NCGR models given by Eq. (6a) in the main text, coupled to each of Eqs. (D1a), (D1b), and (D1c) will be referred to below as NCGR_I, NCGR_{II}, and NCGR_{III}, respectively.

In the most basic model, NCGR_I, $\sigma = \sigma_I$ is proportional to the interannual standard deviation of the observed IFD/FUD [Eq. (D1a)]. In this formulation, predictive uncertainty is assumed to be constant from one year to the next, and contains no information from the ensemble forecast. A similarly simple representation of predictive uncertainty was considered in both Kharin et al. (2009) and Kharin et al. (2017) for seasonal precipitation forecasts, albeit based on the climatological variance of model hindcasts, rather than on observations.

Typically a measure of ensemble spread is included as a predictor for the parameter that describes predictive uncertainty in EMOS formulations (e.g., Gneiting et al. 2005; Thorarinsdottir and Gneiting 2010; Scheuerer 2014; Messner et al. 2014b). To determine if additional probabilistic skill can be achieved by incorporating an estimate of ensemble dispersion, we consider a second model, NCGR_{II}, in which $\sigma = \sigma_{II}$ depends also on the ensemble standard deviation s_x [Eq. (D1b)]. Ideally, even if the ensemble forecasts are systematically over or underdispersive, any spread-error correlation will contribute positively to the predictive uncertainty through the rescaling factor β_2 . We only consider s_x as a predictor if it is a significant predictor of the forecast absolute error over the training period. Otherwise, σ is modeled according to Eq. (D1a). We define the absolute error to be $|x_{(i)}^{\text{tc}}(\tau_i) - y(\tau_i)|$, where the trend-corrected ensemble mean is used (rather than the raw ensemble mean) so that any dependence of the long-term model bias on the absolute error does not affect the statistical test. Statistical significance is determined with a two-sided t test ($p < 0.05$).

IFD and FUD forecasts tend to exhibit a strong positive correlation between $x_{(i)}^{\text{tc}}$ and s_x . This is to be expected since later IFDs and FUDs correspond to longer lead times, and thus typically greater forecast uncertainty. The exception to this is when trends place the event systematically late in the season at one point in the historical record with relatively small ensemble spread, and in the middle of the season at another point with relatively large ensemble spread; in such situations, negative correlations can occur. Similarly, forecast errors are generally larger when $x_{(i)}^{\text{tc}}$ (and thus also lead time) is large. Again the opposite can be true when trends place the climatological date late in the season during one point of the historical record and not the other. Consequently, as a third model, NCGR_{III}, $\sigma = \sigma_{III}$ is posited to depend on $x_{(i)}^{\text{tc}}$ [Eq. (D1c)]. As before, we only model σ in this way when $x_{(i)}^{\text{tc}}(\tau_i)$ is a significant predictor of $|x_{(i)}^{\text{tc}}(\tau_i) - y(\tau_i)|$.

Figure D1 shows box-and-whisker diagrams of the correlation between s_x and forecast error, $|x_{(i)}^{\text{tc}} - y|$ (blue), and between $x_{(i)}^{\text{tc}}$ and $|x_{(i)}^{\text{tc}} - y|$ (orange) for CanCM4 hindcasts over a testing period 1979–98 (i.e., prior to the verification period). Although both relationships are usually weak for the overall fraction of hindcasts (i.e., the median correlation tends to be below the significance threshold), there remain numerous

FIG. D1. Box-and-whisker diagrams of the correlation between the ensemble standard deviation s_x and the absolute error $|x_{(i)}^{tc} - y|$ (blue diagrams) and between the trend-corrected ensemble mean $x_{(i)}^{tc}$ and the absolute error $|x_{(i)}^{tc} - y|$ (orange diagrams). (Note the absolute value of the correlation is taken.) Replicates for spread are the correlations at individual grid points. Correlations are calculated using CanCM4 hindcasts, although qualitatively consistent results are obtained with CanCM3 hindcasts. The dashed black line represents the threshold for significant correlation according to a two-sided t test ($p < 0.05$).

instances for which the relationships are significant. Importantly, the relationship between $x_{(i)}^{tc}$ and $|x_{(i)}^{tc} - y|$ tends to be more robust than that between s_x and $|x_{(i)}^{tc} - y|$.

b. Estimating NCGR coefficients

As described in the main text, NCGR regression coefficients are estimated through minimization of the CRPS. The CRPS must be evaluated repeatedly during optimization, so it is beneficial to express it analytically. For the DCNORM distribution we find that Eq. (7) can be written as

$$\begin{aligned} \text{crps}(F, y) = & \sigma \{ b^* \Phi(b^*)^2 - a^* \Phi(a^*)^2 \\ & + 2[\Phi(b^*)\phi(b^*) - \Phi(a^*)\phi(a^*)] \\ & + \frac{1}{\sqrt{\pi}} [\Phi(\sqrt{2}a^*) - \Phi(\sqrt{2}b^*)] \\ & + 2[y^* \Phi(y^*) - b^* \Phi(b^*)] \\ & + 2[\phi(y^*) - \phi(b^*)] + (b^* - y^*) \}, \end{aligned} \quad (\text{D2})$$

where $*$ denotes a standardized variable [e.g., $z^* = \sigma^{-1}(z - \mu)$]. The derivation of Eq. (D2) is given in appendix E.

For a forecast in year t , the coefficients for each NCGR_I , NCGR_{II} , and NCGR_{III} are estimated by minimizing the time-mean CRPS for the N pairs of training forecasts and observations over τ_t . Minimization uses the sequential least squares programming algorithm (Kraft 1988), which is the default algorithm in Python's Scipy library for minimization when constraints are needed (i.e., using the `scipy.optimize.minimize` function) (Jones et al. 2001). While μ can be any real value for the DCNORM distribution, the constraint $a - 1 \leq \mu \leq b + 1$ is imposed for faster convergence. Note that we allow μ to vary outside the interval $[a, b]$ by ± 1 day so that the probabilities $\text{Pr}[Y(t) = a] \approx 100\%$ and $\text{Pr}[Y(t) = b] \approx 100\%$ remain possible. Additionally, we require $\sigma > 0$. Initial parameter guesses are $\alpha_1 = \alpha_2 = 1$ and $\beta_1 = 1$. For β_2 , we set initial guesses, respectively, to $\beta_2 = \sigma_c / \langle s_x \rangle_{\tau_t}$ and $\beta_2 = \sigma_c / \langle x_{(i)}^{tc} \rangle_{\tau_t}$ so that the second term in each equation has a

magnitude similar to σ_c ; $\langle \cdot \rangle_{\tau_t}$ denotes the time average over τ_t .

To limit parameter overfitting and reduce computational time, the “tol” parameter in the `scipy.optimize.minimize` function, which controls the convergence of the objective, is changed from its default value of 10^{-6} to 5×10^{-2} . Smaller values than 5×10^{-2} are found to result in both poorer accuracy (larger cross-validated CRPS) and reduced reliability, indicative of overfitting (not shown).

c. Final NCGR model

We now investigate whether probabilistic skill is improved through the additional predictors for the uncertainty parameter σ (the ensemble standard deviation or the trend-corrected ensemble mean) compared to modeling σ as a constant.

Figure D2 shows distributions of the CRPSS, calculated by setting the time-mean CRPS for NCGR_{II} (blue bars) and NCGR_{III} (orange bars) as the target forecasts in the CRPSS, and setting the time-mean CRPS for NCGR_I as the reference forecast. Positive CRPSS indicates skill is improved through the additional predictor in the respective model. We only consider grid cells where either s_x or $x_{(i)}^{tc}$ are statistically significant in at least 15 of the 20 verifying years using the operational-long cross-validation scheme. As in Fig. D1, results are shown for the CanCM4 model, although similar results are obtained with CanCM3. Stacking at CRPSS = 0 when comparing NCGR_{II} against NCGR_I corresponds to locations where s_x is not statistically significant but $x_{(i)}^{tc}$ is; the converse is true when comparing NCGR_{III} against NCGR_I .

For IFD hindcasts initialized from 1 March through 1 June, improvements using NCGR_{II} and NCGR_{III} are overall trivial compared NCGR_I . On the other hand, for FUD hindcasts initialized from 1 September through 1 December, improvements are seen predicting σ using $x_{(i)}^{tc}$. While the improvements with NCGR_{III} are not unanimous, the skill improvements during the freeze season suggest that $x_{(i)}^{tc}$ contributes positively to the model; thus, we present

FIG. D2. Histograms of CRPSS (i.e., skill scores) across individual grid cells, quantifying the percent improvement of the CRPS for (top) IFD and (bottom) FUD hindcasts calibrated using NCGR_{II} compared to NCGR_I (blue bars), and using NCGR_{III} compared to NCGR_I (orange bars). Positive CRPSS values indicate improved probabilistic accuracy relative to NCGR_I, and thus added value from the respective additional predictor in NCGR_{II} or NCGR_{III}. Vertical dashed lines are spatially averaged CRPSS values, which are calculated from the area-weighted mean CRPS for each NCGR model. For comparison with Fig. D1, results are shown for CanCM4.

NCGR_{III} in the main text as the final NCGR model, and refer to it simply as NCGR.

APPENDIX E

Derivation of the CRPS for the DCNORM Distribution

The CDF for the DCNORM probability distribution is defined as

$$F(x; \mu, \sigma) = \begin{cases} 0, & -\infty < x < a \\ \Phi\left(\frac{x - \mu}{\sigma}\right), & a \leq x < b \\ 1 & b \leq x < \infty \end{cases} \quad (\text{E1})$$

and has jump discontinuities at $x = a$ from 0 to $\Phi[(a - \mu)/\sigma]$, and at $x = b$ from $1 - \Phi[(b - \mu)/\sigma]$ to 1.

The CRPS [Eq. (7)] with F given by Eq. (E1) thus has discontinuities at $x = a$, $x = y$, and $x = b$, and can be written as

$$\begin{aligned} \text{crps}(F, y) = & \lim_{l_1 \rightarrow a^+} \int_{l_1}^{l_2} \Phi\left(\frac{x - \mu}{\sigma}\right)^2 dx \\ & + \lim_{l_3 \rightarrow y^+} \int_{l_3}^{l_4} \Phi\left(\frac{x - \mu}{\sigma}\right)^2 - 2\Phi\left(\frac{x - \mu}{\sigma}\right) + 1 dx, \end{aligned}$$

where we have used that $F(x) = H(x - y) = 0$ over $(-\infty, a)$, $F(x) = H(x - y) = 1$ over $[b, \infty)$, $H(x - y) = 0$ over $[a, y)$, and $H(x - y) = 1$ over $[y, b]$. Making the substitution $dx^* = dx/\sigma$ leads to

$$\text{crps}(F, y) = \sigma \left[\lim_{l_1 \rightarrow (a^*)^+} \int_{l_1}^{l_2} \Phi(x^*)^2 dx^* + \lim_{l_3 \rightarrow (y^*)^+} \int_{l_3}^{l_4} \Phi(x^*)^2 - 2\Phi(x^*) + 1 dx^* \right].$$

Applying integration by parts, one can show that

$$\int \Phi(x^*) dx^* = x^* \Phi(x^*) + \phi(x^*) + C,$$

and

$$\begin{aligned} \int \Phi(x^*)^2 dx^* &= x^* \Phi(x^*)^2 - 2 \int x^* \phi(x^*) \Phi(x^*) dx^* \\ &= x^* \Phi(x^*)^2 + 2\Phi(x^*) \phi(x^*) - \int \phi(x^*)^2 dx^* \\ &= x^* \Phi(x^*)^2 + 2\Phi(x^*) \phi(x^*) - \frac{1}{\sqrt{\pi}} \Phi(\sqrt{2} x^*) + C, \end{aligned}$$

where C is a constant of integration. Using this result we obtain the following:

$$\begin{aligned} \text{crps}(F, y) = & \sigma \left[x^* \Phi(x^*)^2 + 2\Phi(x^*)\phi(x^*) - \frac{1}{\sqrt{\pi}} \Phi(\sqrt{2}x^*) \right] \Big|_{l_1}^{l_2} \\ & + \sigma \left[x^* \Phi(x^*)^2 + 2\Phi(x^*)\phi(x^*) \right. \\ & \left. - \frac{1}{\sqrt{\pi}} \Phi(\sqrt{2}x^*) - 2x^* \Phi(x^*) - 2\phi(x^*) + x^* \right] \Big|_{l_3}^{l_4}. \end{aligned}$$

Evaluating the limits of integration and rearranging we obtain the following:

$$\begin{aligned} \text{crps}(F, y) = & \sigma \{ b^* \Phi(b^*)^2 - a^* \Phi(a^*)^2 \\ & + 2[\Phi(b^*)\phi(b^*) - \Phi(a^*)\phi(a^*)] \\ & + \frac{1}{\sqrt{\pi}} [\Phi(\sqrt{2}a^*) - \Phi(\sqrt{2}b^*)] \\ & + 2[y^* \Phi(y^*) - b^* \Phi(b^*)] + 2[\phi(y^*) - \phi(b^*)] \\ & + (b^* - y^*) \}, \end{aligned}$$

which is the solution given by Eq. (D2).

REFERENCES

- Bentzen, S., and P. Friederichs, 2012: Generating and calibrating probabilistic quantitative precipitation forecasts from the high-resolution NWP model COSMO-DE. *Wea. Forecasting*, **27**, 988–1002, <https://doi.org/10.1175/WAF-D-11-00101.1>.
- Boer, G., 2009: Climate trends in a seasonal forecasting system. *Atmos.–Ocean*, **47**, 123–138, <https://doi.org/10.3137/AO1002.2009>.
- 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, T. Carrieres, L. Pogson, and M. Lajoie, 2013a: Overview of sea ice data assimilation activities at Environment Canada. *Proc. ECMWF-WWRP/THORPEX Polar Prediction Workshop*, Reading, United Kingdom, ECMWF, 24–27.
- , —, L. Pogson, T. Carrieres, and P. Pestieau, 2013b: A new Environment Canada regional ice analysis system. *Atmos.–Ocean*, **51**, 18–34, <https://doi.org/10.1080/07055900.2012.747171>.
- , —, T. Carrieres, and L. Pogson, 2016: 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>.
- Buizza, R., 2008: The value of probabilistic prediction. *Atmos. Sci. Lett.*, **9**, 36–42, <https://doi.org/10.1002/asl.170>.
- Cannon, A. J., 2008: Probabilistic multisite precipitation downscaling by an expanded Bernoulli–Gamma density network. *J. Hydrometeorol.*, **9**, 1284–1300, <https://doi.org/10.1175/2008JHM960.1>.
- Cohen, A. C., Jr., 1950: Estimating the mean and variance of normal populations from singly truncated and doubly truncated samples. *Ann. Math. Stat.*, **21**, 557–569, <https://doi.org/10.1214/aoms/1177729751>.
- Comiso, J. C., 2003: Warming trends in the Arctic from clear sky satellite observations. *J. Climate*, **16**, 3498–3510, [https://doi.org/10.1175/1520-0442\(2003\)016<3498:WTITAF>2.0.CO;2](https://doi.org/10.1175/1520-0442(2003)016<3498:WTITAF>2.0.CO;2).
- , C. L. Parkinson, R. Gersten, and L. Stock, 2008: Accelerated decline in the Arctic sea ice cover. *Geophys. Res. Lett.*, **35**, L01703, <https://doi.org/10.1029/2007GL031972>.
- 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>.
- , —, and —, 2019: Probabilistic forecasting of the Arctic sea ice edge with contour modeling. *arXiv preprint arXiv:1908.09377*.
- Dirkson, A., W. J. Merryfield, and A. Monahan, 2017: Impacts of sea ice thickness initialization on seasonal arctic sea ice predictions. *J. Climate*, **30**, 1001–1017, <https://doi.org/10.1175/JCLI-D-16-0437.1>.
- , B. Denis, and W. Merryfield, 2019a: A multimodel approach for improving seasonal probabilistic forecasts of regional Arctic sea ice. *Geophys. Res. Lett.*, **46**, 10 844–10 853, <https://doi.org/10.1029/2019GL083831>.
- , W. J. Merryfield, and A. H. Monahan, 2019b: Calibrated probabilistic forecasts of Arctic sea ice concentration. *J. Climate*, **32**, 1251–1271, <https://doi.org/10.1175/JCLI-D-18-0224.1>.
- Doblas-Reyes, F. J., R. Hagedorn, and T. Palmer, 2005: The rationale behind the success of multi-model ensembles in seasonal forecasting—II. Calibration and combination. *Tellus*, **57A**, 234–252, <https://doi.org/10.1111/j.1600-0870.2005.00104.x>.
- Eicken, H., 2013: Arctic sea ice needs better forecasts. *Nature*, **497**, 431–433, <https://doi.org/10.1038/497431a>.
- 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., and A. E. Raftery, 2007: Strictly proper scoring rules, prediction, and estimation. *J. Amer. Stat. Assoc.*, **102**, 359–378, <https://doi.org/10.1198/016214506000001437>.
- , —, 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>.
- 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>.
- Harrell, F. E., and C. Davis, 1982: A new distribution-free quantile estimator. *Biometrika*, **69**, 635–640, <https://doi.org/10.1093/biomet/69.3.635>.
- IPCC, 2019: *IPCC Special Report on the Ocean and Cryosphere in a Changing Climate*. Cambridge University Press, in press, <https://www.ipcc.ch/srocc/>.
- Johnson, N. L., S. Kotz, and N. Balakrishnan, 1994: *Continuous Univariate Distributions*. Vol. 1, Wiley, 158 pp.
- Jolliffe, I. T., and D. B. Stephenson, 2012: *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. John Wiley & Sons, 372 pp.
- Jones, E., T. Oliphant, and P. Peterson, 2001: SciPy: Open source scientific tools for Python. Accessed 28 January 2020, <https://www.scipy.org/about.html>.

- Kharin, V., G. Boer, W. Merryfield, J. 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. Merryfield, G. Boer, and W.-S. Lee, 2017: A post-processing 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>.
- , 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>.
- Kraft, D., 1988: A software package for sequential quadratic programming. Tech. Rep. DFLR-FB 88-28, DLR German Aerospace Center–Institute for Flight Mechanics, Köln, Germany, 33 pp.
- 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>.
- Krzysztofowicz, R., 1999: Bayesian theory of probabilistic forecasting via deterministic hydrologic model. *Water Resour. Res.*, **35**, 2739–2750, <https://doi.org/10.1029/1999WR900099>.
- Kwok, R., G. Cunningham, M. Wensnahan, I. Rigor, H. Zwally, and D. Yi, 2009: Thinning and volume loss of the Arctic Ocean sea ice cover: 2003–2008. *J. Geophys. Res.*, **114**, C07005, <https://doi.org/10.1029/2009JC005312>.
- Markus, T., J. C. Stroeve, and J. Miller, 2009: Recent changes in Arctic sea ice melt onset, freezeup, and melt season length. *J. Geophys. Res.*, **114**, C12024, <https://doi.org/10.1029/2009JC005436>.
- Maslanik, J., C. Fowler, J. Stroeve, S. Drobot, J. Zwally, D. Yi, and W. Emery, 2007: A younger, thinner Arctic ice cover: Increased potential for rapid, extensive sea-ice loss. *Geophys. Res. Lett.*, **34**, L24501, <https://doi.org/10.1029/2007GL032043>.
- , 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>.
- Merryfield, W., B. Denis, J. Fontecilla, W. Lee, V. Kharin, J. Hodgson, and B. Archambault, 2011: The Canadian Seasonal to Interannual Prediction System (CanSIPS). Environment and Climate Change Canada, 51 pp., https://collaboration.cmc.ec.gc.ca/cmc/cmci/product_guide/docs/lib/op_systems/doc_opchanges/technote_cansips_20111124_e.pdf.
- , and Coauthors, 2013: 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, D. S. Wilks, and A. Zeileis, 2014a: Extending extended logistic regression: Extended versus separate versus ordered versus censored. *Mon. Wea. Rev.*, **142**, 3003–3014, <https://doi.org/10.1175/MWR-D-13-00355.1>.
- , —, A. Zeileis, and D. S. Wilks, 2014b: 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>.
- Murphy, A. H., 1973: A new vector partition of the probability score. *J. Appl. Meteor.*, **12**, 595–600, [https://doi.org/10.1175/1520-0450\(1973\)012<0595:ANVPOT>2.0.CO;2](https://doi.org/10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2).
- Ospina, R., and S. L. Ferrari, 2010: Inflated beta distributions. *Stat. Hefte*, **51**, 111.
- Parkinson, C. L., 2014: Spatially mapped reductions in the length of the Arctic sea ice season. *Geophys. Res. Lett.*, **41**, 4316–4322, <https://doi.org/10.1002/2014GL060434>.
- Peng, G., W. Meier, D. Scott, and M. 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>.
- Polyakov, I. V., J. E. Walsh, and R. Kwok, 2012: Recent changes of Arctic multiyear sea ice coverage and the likely causes. *Bull. Amer. Meteor. Soc.*, **93**, 145–151, <https://doi.org/10.1175/BAMS-D-11-00070.1>.
- Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev.*, **133**, 1155–1174, <https://doi.org/10.1175/MWR2906.1>.
- Richardson, D. S., 2000: Skill and relative economic value of the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **126**, 649–667, <https://doi.org/10.1002/qj.49712656313>.
- Sansom, P. G., C. A. Ferro, D. B. Stephenson, L. Goddard, and S. J. Mason, 2016: Best practices for postprocessing ensemble climate forecasts. Part I: Selecting appropriate recalibration methods. *J. Climate*, **29**, 7247–7264, <https://doi.org/10.1175/JCLI-D-15-0868.1>.
- Scheuerer, M., 2014: Probabilistic quantitative precipitation forecasting using ensemble model output statistics. *Quart. J. Roy. Meteor. Soc.*, **140**, 1086–1096, <https://doi.org/10.1002/qj.2183>.
- , and T. M. Hamill, 2015: Statistical postprocessing of ensemble precipitation forecasts by fitting censored, shifted gamma distributions. *Mon. Wea. Rev.*, **143**, 4578–4596, <https://doi.org/10.1175/MWR-D-15-0061.1>.
- Serreze, M. C., M. M. Holland, and J. Stroeve, 2007: Perspectives on the Arctic's shrinking sea-ice cover. *Science*, **315**, 1533–1536, <https://doi.org/10.1126/science.1139426>.
- Sigmond, M., M. Reader, G. Flato, W. Merryfield, and A. Tivy, 2016: Skillful seasonal forecasts of Arctic sea ice retreat and advance dates in a dynamical forecast system. *Geophys. Res. Lett.*, **43**, 12 457–12 465, <https://doi.org/10.1002/2016GL071396>.
- Stammerjohn, S., R. Massom, D. Rind, and D. Martinson, 2012: Regions of rapid sea ice change: An inter-hemispheric seasonal comparison. *Geophys. Res. Lett.*, **39**, L06501, <https://doi.org/10.1029/2012GL050874>.
- Stephenson, S. R., and R. Pincus, 2018: Challenges of sea-ice prediction for Arctic marine policy and planning. *J. Borderl. Stud.*, **33**, 255–272, <https://doi.org/10.1080/08865655.2017.1294494>.
- Stroeve, J., T. Markus, L. Boisvert, J. Miller, and A. Barrett, 2014: Changes in Arctic melt season and implications for sea ice loss. *Geophys. Res. Lett.*, **41**, 1216–1225, <https://doi.org/10.1002/2013GL058951>.
- Stroeve, J. C., E. Blanchard-Wigglesworth, V. Guemas, S. Howell, F. Massonnet, and S. Tietsche, 2015: Improving predictions of Arctic sea ice extent. *Eos, Trans. Amer. Geophys. Union*, **96**, 11, <https://doi.org/10.1029/2015EO031431>.
- , V. Kattsov, A. Barrett, M. Serreze, T. Pavlova, M. Holland, and W. N. Meier, 2012: Trends in Arctic sea ice extent from CMIP5, CMIP3 and observations. *Geophys. Res. Lett.*, **39**, L16502, <https://doi.org/10.1029/2012GL052676>.
- , A. D. Crawford, and S. Stammerjohn, 2016: Using timing of ice retreat to predict timing of fall freeze-up in the Arctic. *Geophys. Res. Lett.*, **43**, 6332–6340, <https://doi.org/10.1002/2016GL069314>.
- Thorarindottir, T. L., and T. Gneiting, 2010: Probabilistic forecasts of wind speed: Ensemble model output statistics by using

- heteroscedastic censored regression. *J. Roy. Stat. Soc.*, **173A**, 371–388, <https://doi.org/10.1111/j.1467-985X.2009.00616.x>.
- 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. 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>.
- Weisheimer, A., and T. Palmer, 2014: On the reliability of seasonal climate forecasts. *J. Roy. Soc. Interface*, **11**, 20131162, <https://doi.org/10.1098/rsif.2013.1162>.
- Wilks, D. S., 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.
- WMO, 2017: WMO guidelines on the calculation of climate normals. World Meteorological Organization Switzerland, WMO-1203, 29 pp.
- Yuan, X., and E. F. Wood, 2013: Multimodel seasonal forecasting of global drought onset. *Geophys. Res. Lett.*, **40**, 4900–4905, <https://doi.org/10.1002/grl.50949>.
- Zhao, T., J. C. Bennett, Q. 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>.