# Probabilistic Forecasts of Arctic Sea Ice Thickness

Peter A. GAO, Hannah M. DIRECTOR, Cecilia M. BITZ, and
Adrian E. RAFTERY

In recent decades, warming temperatures have caused sharp reductions in the volume of sea ice in the Arctic Ocean. Predicting changes in Arctic sea ice thickness is vital in a changing Arctic for making decisions about shipping and resource management in the region. We propose a statistical spatio-temporal two-stage model for sea ice thickness and use it to generate probabilistic forecasts up to three months into the future. Our approach combines a contour model to predict the ice-covered region with a Gaussian random field to model ice thickness conditional on the ice-covered region. Using the most complete estimates of sea ice thickness currently available, we apply our method to forecast Arctic sea ice thickness. Point predictions and prediction intervals from our model offer comparable accuracy and improved calibration compared with existing forecasts. We show that existing forecasts produced by ensembles of deterministic dynamic models can have large errors and poor calibration. We also show that our statistical model can generate good forecasts of aggregate quantities such as overall and regional sea ice volume.

Supplementary materials accompanying this paper appear on-line.

**Key Words:** Spatio-temporal forecasting; Gaussian process; Integrated nested Laplace approximation; Stochastic partial differential equation; Spatial statistics.

## 1. INTRODUCTION

Over the past few decades, the sea ice covering much of the Arctic Ocean has retreated dramatically. Large declines are seen in both the annual extent as well as the average thickness of the ice (Notz and Stroeve 2018). Continued declines in sea ice will open more

---

---

P. A. Gao (✉), Department of Statistics, University of Washington, Seattle, WA 98105, USA (E-mail: *peter-gao@uw.edu*). H. M. Director, Department of Applied Mathematics and Statistics, Colorado School of Mines, Golden, CO 80401, USA (E-mail: *hdirector@mines.edu*). C. M. Bitz, Department of Atmospheric Sciences, University of Washington, Seattle, WA 98105, USA (E-mail: *bitz@uw.edu*). A. E. Raftery, Departments of Statistics and Sociology, University of Washington, Seattle, WA 98105, USA (E-mail: *raftery@uw.edu*).
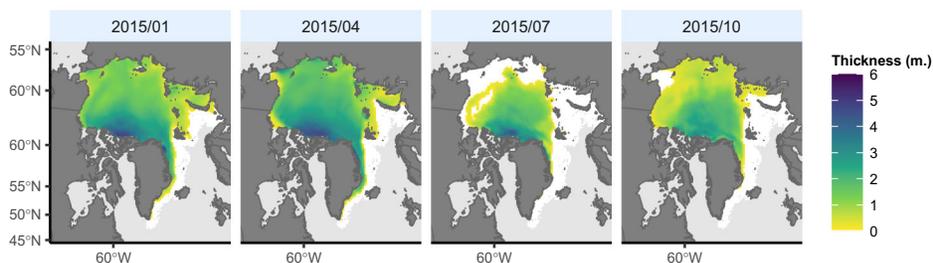
Figure 1.   Maps of monthly average Arctic sea ice thickness estimates for four months in 2015 (reanalysis domain in white). The dynamic model assimilates satellite sea ice thickness measurements in the months when these data are available (mid-October to mid-April). Dark gray areas represent land, white areas represent ocean in the reanalysis domain, and light gray areas represent ocean outside the reanalysis domain .

ice-free shipping routes (Melia et al. 2017), making accurate forecasts of sea ice thickness crucial for predicting when and where the Arctic Ocean can be traversed.

Historical estimates of Arctic sea ice thickness are typically provided by reanalysis products. Reanalysis products provide estimates of sea ice conditions at fixed locations and time intervals in the past. They are made by inputting available historical observations into physics-based dynamic models that use differential equations to estimate the state of the sea ice and surrounding environment, including at times and locations where no observations were made. Models are run repeatedly with varied initial conditions, producing sets of estimates, called ensembles, that provide some information about spread. Many ensembles do not directly assimilate estimates of thickness, meaning that only observations of related quantities like sea ice concentration are assimilated into the model. So generated reanalysis estimates of sea ice thickness may be biased (Chevallier et al. 2017).

Ensemble methods are also commonly used to forecast the evolution of quantities such as sea ice concentration and area. Ensembles also produce forecasts of sea ice thickness, but the use and evaluation of these forecasts have been more limited. These forecasts may also be biased or improperly calibrated (Director et al. 2017, 2021; Dirkson et al. 2019). Statistical post-processing or purely statistical methods can address these issues and have been used in settings such as weather, climate, and renewable energy (Vannitsem et al. 2018). For sea ice forecasting, post-processing techniques (Kharin et al. 2012; Director et al. 2017; Dirkson et al. 2018, 2019) and purely statistical methods (Handcock and Raphael 2020; Zhang and Cressie 2020) have been used for forecasting sea ice quantities other than thickness.

In this article, we confirm that existing ensemble forecasts of sea ice thickness may be biased and poorly calibrated. We then propose a purely statistical model of sea ice thickness. We apply our method to forecast Arctic sea ice thickness and show that a statistical approach provides good forecasts up to seasonal time scales.

Modeling sea ice thickness presents a number of statistical challenges. As seen in Fig. 1, sea ice thickness is nonzero for large, continuous regions of the Arctic. When we sample sea ice thickness fields from our model, we desire realistic presence-of-ice fields, with large, continuous regions where sea ice is present. Using a binary indicator variable to represent presence of sea ice at a given location, any model of sea ice should account for this spatial correlation of the presence-of-ice indicators at nearby locations.

We also seek probabilistic forecasts of thickness that not only are marginally calibrated at individual location/time pairs, but also accurately estimate spatial and temporal correlations in forecast errors. Accounting for these correlations is crucial for making probabilistic forecasts of aggregate quantities like sea ice volume.

Finally, it is important to balance model complexity with computational cost. Forecasting sea ice thickness on a high-resolution grid requires large training datasets. Our modeling approach must also account for seasonality in model parameters, as Arctic sea ice undergoes dramatic seasonal changes annually. More flexible models will be better able to account for this complexity, but in general will demand added computational resources.

To address these challenges, we propose a probabilistic, spatio-temporal two-stage thickness (TST) model. We leverage newly available reanalysis estimates of sea ice thickness. We train the model to make monthly probabilistic forecasts of Arctic sea ice thickness up to three months in advance (lead time). These forecasts can predict changes in sea ice volume and sea ice thickness for all months of the year. Since few ensemble forecasts directly assimilate thickness estimates, we present a stand-alone statistical approach but note that our method could be adapted to work as a post-processing method for use with skillful dynamical forecasts. In the first stage of our model, we propose a spatio-temporal extension of the contour model introduced by Director et al. (2021) to forecast which locations will be covered by sea ice in a given month. Based on this model, we sample ice-covered regions that exhibit the desired spatial and temporal correlations. In the second stage, we use a spatio-temporal hierarchical Gaussian random field model to represent the distribution of sea ice thickness for a set of locations, conditional on sea ice being present at those locations.

For large datasets, working with Gaussian random field models can be computationally expensive. In recent years, much research has focused on approximations based on low-rank or sparse covariance and precision matrices for improving computational efficiency (Banerjee et al. 2008; Cressie and Johannesson 2008; Lindgren et al. 2011; Sang and Huang 2012; Datta et al. 2016; Nychka et al. 2015; Katzfuss et al. 2020). For a review of some of these methods, see Heaton et al. (2019). We carry out approximate Bayesian inference for both stages of the model using integrated nested Laplace approximations (Rue et al. 2009). This yields probabilistic forecasts of both sea ice thickness at individual locations and times as well as aggregate quantities like sea ice volume.

Forecasts from our two-stage thickness model provide useful information for lead times up to three months in the future. In terms of accuracy, point forecasts from our model are similar to existing ensemble forecasts obtained from the European Centre for Medium-Range Weather Forecasts (ECMWF) and improve on a climatological reference forecast. However, our model is able to account for spatial correlation in forecast errors, allowing us to generate prediction intervals for sea ice volume. We show that the statistical model-based intervals are better calibrated than existing interval forecasts, providing useful information about seasonal changes in sea ice volume.

The paper is organized as follows. In Sect. 2, we describe details of our two-stage thickness model. In Sect. 3, we outline our strategy for estimating model parameters for the model and generating forecasts conditional on reanalysis estimates. In Sect. 4, we apply our two-stage thickness model to generate forecasts of Arctic sea ice thickness and compare

our forecasts with the ECMWF forecasts and statistical reference forecasts. Finally, Sect. 5 discusses our results and proposes areas for future research.

## 2. THE TWO-STAGE THICKNESS MODEL

### 2.1. STAGE ONE: SPATIO-TEMPORAL CONTOUR MODEL

In the first stage of our two-stage thickness model, we develop a probabilistic model for generating plausible sea ice-covered regions for each forecast month. Following Director et al. (2021), we use a contour model for forecasting the sea ice edge, the location of the outer-most boundary of the region covered by sea ice.

In this framework, we model the region covered by sea ice for a given month as the set of locations enclosed by a random contour, treating our spatial domain as a two-dimensional grid with Euclidean distance. We define a random contour at time $t$ as an ordered set of random points $\mathbf{c}(t) = \{c(1, t), \ldots, c(n, t)\}$. By carefully choosing a distribution for the points' locations, we can ensure that connecting the points in order yields the boundary of a polygon, which we will refer to as the ice-covered region. We define an ordered set of line segments $\{l(1, t), \ldots l(n, t)\}$, where each $l(i, t)$ extends from some fixed point $b_i$ in a fixed direction, but with random length $y(i, t)$. Thus, the only random, time-varying aspect of the line segment $l(i, t)$ is its length. Each line $l(i, t)$ has a fixed end point $b_i$ and a random end point that is a distance $y(i, t)$ away from $b_i$ along some fixed angle. We connect the random end points of the lines $\{l(1, t), \ldots l(n, t)\}$ to generate random contours.

Figure 2 gives two examples of this in the Arctic, which we split into multiple subregions. Our subregions are based upon definitions used by the National Snow and Ice Data Center (NSIDC) (Cavalieri and Parkinson 2012) and have been modified for convenience when defining the fixed points and line segments. The contour model is trained using satellite images of the Arctic: these images provide gridded observations of which locations are covered by ice. The contour modeling approach developed by Director et al. (2021) requires an initial data processing step in which the grid derived from the satellite images is used to define the set of fixed points and line segments for the contour model. The gridded satellite data can then be converted to observed lengths $y$. The bottom left plot of Fig. 2 depicts the central Arctic region, where we specify just one fixed point $b_i \equiv b$ from which lines $l(i, t)$ extend in all directions at fixed angle intervals. In this region, we connect the random end points of the lines to enclose a polygon.

The bottom right plot of Fig. 2 shows the Greenland Sea, where we place fixed points $b_i$ at fixed intervals along the coast of Greenland and specify $l(i, t)$ to be lines of random length extending out at a fixed angle from the $b_i$. In the Greenland Sea, we construct the ice-covered region by connecting all the random end points of the lines $l(i, t)$ to the fixed points $b_i$. By combining the polygons for all subregions, we obtain an overall ice-covered region for the Arctic. The full specifications of the fixed points and line angles used for each subregion can be found in the IceCast R package (Director et al. 2020). As noted in Director and Raftery (2021), the basic contour model can generate only star-shaped polygons and so does not account for features like polynyas—regions of open water that form within ice sheets—or large pieces of sea ice that float freely apart from the main ice sheet. Director
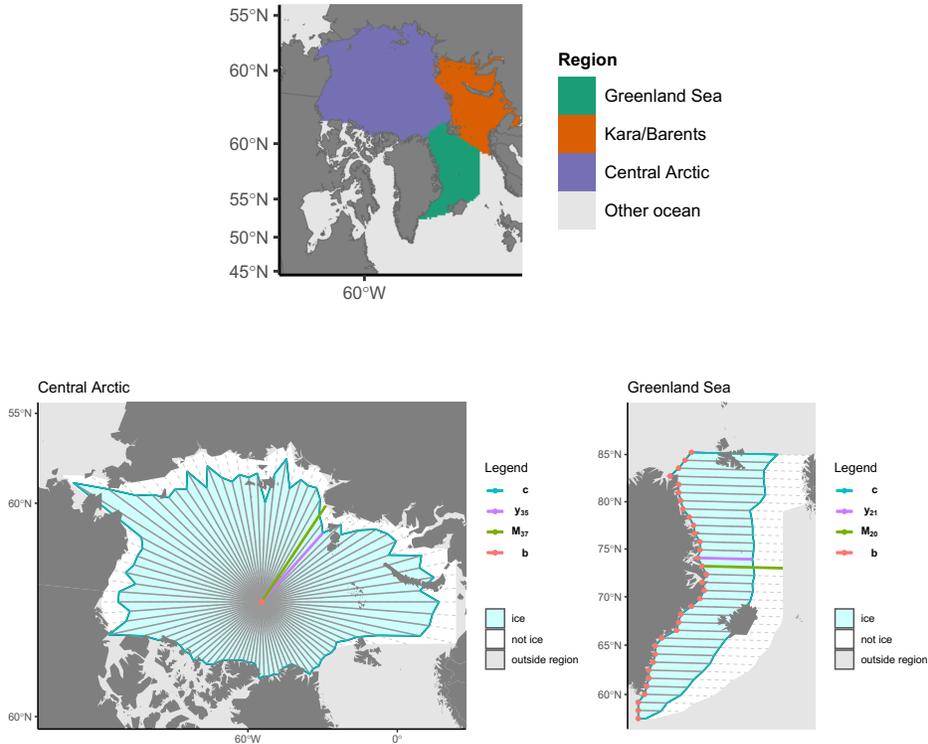
Figure 2.   Top: map of reanalysis domain with NSIDC subregions. Dark gray areas represent land and light gray areas represent ocean beyond the reanalysis domain. Bottom: Schematic of simulated contours (**c**) based on fixed points (**b**) and line segments (**l**) of observed length **y** and maximum length **M** in Central Arctic (bottom left) and Greenland Sea (bottom right). Subregions used are derived from NSIDC subregions, but modified for convenience when defining **b** and **l** .

et al. (2021) address this by using Bayesian model averaging to combine the contour model forecasts with a climatological forecast, producing a mixture model that can better predict such features. For the sake of simplicity, we have used only the basic contour model below, but note that the mixture contour forecasting method could lead to improved results.

In our model, each line $l(i, t)$ has a random length $y(i, t)$ representing the length of ice-covered ocean extending from $b_i$ at a fixed angle. In practice, each $l(i, t)$ has maximal length $M_i$ before reaching a land or a regional boundary. Some lines cross small lengths of land; for these lines, we specify $M_i$ to be the sum of all the lengths of ocean along the corresponding line. To induce a probability model for the contours, we specify a probability model for the real-valued random variables $w(i, t)$ for $t = 1, \ldots, T$ and $i = 1, \ldots, n$, where $w(i, t)$ is a logit transformation of the random proportion $y(i, t)/M_i$, namely:

$$
w(i, t) = \begin{cases} \text{logit}(1 - \epsilon) & y(i, t)/M_i > 1 - \epsilon \\ \text{logit}(y(i, t)/M_i) & \epsilon \leq y(i, t)/M_i \leq 1 - \epsilon \\ \text{logit}(\epsilon) & y(i, t)/M_i < \epsilon. \end{cases} \tag{1}
$$

When $y(i, t)$ is very close to zero or to its maximal length, i.e., when $y(i, t)/M_i$ is close to zero or one, we ensure that the logit transform is applicable by adding or subtracting a small quantity $\epsilon$. We assume a latent Gaussian model for $\mathbf{w}(t) = (w(1, t), \ldots, w(n, t))^T$ that includes a month-specific mean function $\mu$, a spatiotemporal AR(1) term $z$, and a measurement error term $\delta_w$:

$$w(i, t) = \mu_{m(t)}(i) + z(i, t) + \delta_w(i, t),$$

where $m(t)$ is the month of $t$, $\mu_{m(t)}(i)$ is the month- and location-specific mean of $w(i, t)$, and we assume $\delta_w \sim N(0, \sigma_\delta^2)$ are identically and independently distributed at each location/month.

To account for spatial correlations in the observed $w(i, t)$, we assume that the random vector $\mathbf{z}(t) = (z(1, t), \ldots, z(p, t))'$ follows an AR(1) model in the index $i$ representing the order of the line segments, as follows:

$$z(i, t) = \rho_s z(i - 1, t) + \varepsilon(i, t), \qquad i = 2, \ldots, n, \tag{2}$$

where $z(1, t) = \varepsilon(1, t)$. Here $\rho_s$ controls the correlation between spatially adjacent $z(i, t)$, and $\varepsilon(i, t)$ is a zero-mean Gaussian random variable with variance $(1 - \rho_s)\sigma_z^2$ for $i = 1, \ldots, n$. For the Central Arctic region, the lines extend out in all directions and the last line is adjacent to the first line. Some of the lines correspond to portions of the region that are entirely covered by sea ice for all months, so we fix their lengths to have maximal value and fit the model for the remaining line segments, conditional on the fixed line segments.

We can extend this approach to a spatio-temporal model for all times $t = 1, \ldots, m$ by specifying a model for the random vector obtained by stacking the $\mathbf{z}(t)$ vectors, $\mathbf{z} = (\mathbf{z}(1)^T, \ldots, \mathbf{z}(T)^T)^T$. To do this, we assume an AR(1) structure in time as well as space, yielding a spatiotemporal multivariate Gaussian model on $\{z(i, t)\}$ for all $t = 1, \ldots, T$ and $i = 1, \ldots, n$ where

$$E(z(i, t)) = 0, \tag{3}$$

$$\mathrm{Cov}(z(i, t), z(j, u)) = \sigma_z^2 \rho_s^{|i-j|} \rho_t^{|t-u|}, \tag{4}$$

for $i, j \in \{1, \ldots, n\}$ and $t, u \in \{1, \ldots, T\}$. In Equation (4), $\rho_t$ controls the correlation between time-adjacent $z(i, t)$ and $\sigma_z^2$ is a marginal variance parameter that controls typical size of the deviations from the mean process $\mu$. The resulting covariance function is separable in time and space. This assumption facilitates computation and interpretability of parameters, but implies that the time covariance function is constant in space and vice versa. To help address this inflexibility, we train our model on only the most recent training data when making forecasts, as discussed below.

Suppose we have data for times $t = T - k, T - k + 1, \ldots, T - 1, T$. In summary, for each region, we propose the following hierarchical model for $\mathbf{w} = (\mathbf{w}(T - k)', \ldots, \mathbf{w}(T)')'$:

$$(\mathbf{w}(t) \mid \boldsymbol{\mu}_{m(t)}, \mathbf{z}(t), \sigma_\delta^2) \sim \mathbf{N}_{n(t)}(\boldsymbol{\mu}_{m(t)} + \mathbf{z}(t), \sigma_\delta \mathbf{I}_{n(t)}); \quad t = T - k, \ldots, T \tag{5}$$

$$(\mathbf{z} \mid \sigma_z, \rho_s, \rho_t) \sim \mathbf{N}_n(\mathbf{0}, \boldsymbol{\Gamma}_z) \tag{6}$$

$$(\sigma_\delta, \sigma_z, \rho_s, \rho_t) \sim \mathbf{\Pi}_w(\boldsymbol{\theta}_w) \tag{7}$$

In this model, $n(t)$ denotes the number of line segments in the selected region and $n = \sum_{t=T-k}^{T} n(t)$ is the total number of line segment/time pairs. $\mathbf{z} = (\mathbf{z}(T-k)', \ldots, \mathbf{z}(T)')'$ is the vector obtained by concatenating the values of $z(s, t)$ for all line segment/time pairs. $\mathbf{\Gamma}_z$ denotes the covariance matrix defined by (4), while $\mathbf{\Pi}_w$ denotes a prior on hyperparameters controlled by $\boldsymbol{\theta}_w$ (discussed in Sect. 2.3). We assume that the values of $w$ for different line segments/time pairs are conditionally independent given the values of $z$ at those line segments and times.

In this application, $\rho_s$ and $\rho_t$ are assumed to be positive; if the ice edge extends unexpectedly far from a given fixed point in a particular month, that anomaly will tend to persist in neighboring months and for adjacent lines. Thus, this model supposes that if a given variable $w(i, t)$ is large compared to its mean value $\mu_{m(t)}(i)$, then its time-neighbors $w(i, t+1)$ and $w(i, t-1)$ will also tend to be large. The same relationship holds for neighbors in space $w(i-1, t)$ and $w(i+1, t)$.

## 2.2. STAGE TWO: CONDITIONAL THICKNESS MODEL

A sample from the contour model (stage one) yields a polygon representing the ice-covered region of the Arctic. In the second stage, we propose a model for sea ice thickness conditional on the ice-covered locations. For reanalysis grid locations outside the contour, we define sea ice thickness to be zero; for locations within the contour, we model sea ice thickness via a spatio-temporal Gaussian random field. Let $f(s, t)$ denote the thickness of sea ice at location $s$ and time $t$. We let $s \in \mathbf{C}(t)$ denote the event that $s$ is contained in the polygon defined by contour $\mathbf{c}(t)$. We assume that if $s \notin \mathbf{C}(t)$, $f(s, t) = 0$, and if $s \in \mathbf{C}(t)$, then $f(s, t)$ is drawn from a Gaussian random field.

We assume the following latent Gaussian model:

$$f(s, t) = \eta_{m(t)}(s) + h(s, t) + \delta_f(s, t), \tag{8}$$

for all $s \in \mathbf{C}(t)$ where $m(t)$ denotes the month of $t$, $\eta_{m(t)}(s)$ denotes the location- and month-specific conditional mean thickness, $\delta_f \sim N(0, \tau_\delta^2)$ represents measurement error and $h(s, t)$ is a Gaussian random field with mean zero and covariance function

$$\text{Cov}(h(r, t), h(s, u)) = \tau_h^2 \phi_t^{|t-u|} C_{\kappa, \nu}(r, s); \tag{9}$$

for locations $r, s$ and times $t, u$. Here, $\tau_h^2$ is a marginal variance parameter for the latent field $h$, $\phi_t$ is a temporal autoregressive parameter, and $C_{\kappa, \nu}$ is the Matérn correlation function commonly parameterized as follows, with scale parameter $\kappa$ and smoothness $\nu$:

$$C_{\kappa, \nu}(r, s) = \frac{1}{\Gamma(\nu)2^{\nu-1}}(\kappa||r-s||)^\nu K_\nu(\kappa||r-s||),$$

where $K_\nu$ is the modified Bessel function of the second kind Krainski et al. (2018). We have specified a covariance function for $h$ that is separable in space and time, with an AR(1)

structure in time and Matérn correlation in space. This assumption makes computation much simpler but places strong restrictions on the fitted covariance function.

Suppose we have data for times $t = T - k, T - k + 1, \ldots, T - 1, T$. We propose the following hierarchical model for $\mathbf{f} = (\mathbf{f}(T - k)', \ldots, \mathbf{f}(T)')'$:

$$(\mathbf{f}(t) \mid s \in \mathbf{C}(t), \boldsymbol{\eta}_{m(t)}, \mathbf{h}(t), \tau_\delta) \sim \mathbf{N}_{q(t)}(\boldsymbol{\eta}_{m(t)} + \mathbf{h}(t)), \tau_\delta^2 \mathbf{I}_{q(t)}); \quad t = T - k, \ldots, T \tag{10}$$

$$(\mathbf{h} \mid \tau_h, \phi_t, \kappa, \nu) \sim \mathbf{N}_q(\mathbf{0}, \boldsymbol{\Sigma}_h) \tag{11}$$

$$(\tau_\delta, \tau_h, \phi_t, \kappa, \nu) \sim \boldsymbol{\Pi}_f(\boldsymbol{\theta}_f) \tag{12}$$

Here, $q(t)$ denotes the number of locations contained in $\mathbf{C}(t)$ and $q = \sum_{t=T-k}^{T} q(t)$ is the total number of location/time pairs. We let $\mathbf{h} = (\mathbf{h}(T - k)', \ldots, \mathbf{h}(T)')'$ denote the vector obtained by concatenating the values of $h(s, t)$ for all location/time pairs, $\boldsymbol{\Sigma}_h$ denote the covariance matrix defined by (9), and $\boldsymbol{\Pi}_f$ denote a prior on hyperparameters (discussed in Sect. 2.3), which is controlled by $\boldsymbol{\theta}_f$. We assume that the values of $f$ for any two different location/time pairs are conditionally independent given the values of $h$ at those locations and times.

We train this model using only the reanalysis estimates $\{f(s, t)\}$ where $s \in \mathbf{C}(t)$. At locations outside the ice-covered region, sea ice thickness is equal to zero, so including reanalysis estimates at those locations would deflate the estimated variance parameters. When forecasting future thickness values, we sample a forecast ice-covered region from the model specified above. We predict thickness outside the ice-covered region to be zero and thickness within the region to be drawn from the joint multivariate Gaussian distribution. The model in (8) does not restrict thickness to be nonnegative. In practice, however, $\mu_{m(t)}(i)$ is typically located away from zero and the marginal variance of $f(s, t)$ is such that negative predicted values of $f$ are rare; in such cases, we set the predicted value to be zero.

## 2.3. COMPUTATION AND INFERENCE

Since both the contour model and the conditional thickness model are latent Gaussian models we can carry out approximate Bayesian inference for our models via the integrated nested Laplace approximation (INLA) (Rue et al. 2009) using the R-INLA package (www.r-inla.org). The INLA approach facilitates fast approximate inference for latent Gaussian models using Laplace approximations to generate posterior marginal distributions for latent variables and hyperparameters.

For the contour model, we adopt penalized complexity (PC) priors (Simpson et al. 2017). PC priors treat components of the model as proposed extensions from a simplified base model. In particular, PC priors specify a distance from a proposed model to the base model and place a prior on the distance, penalizing large deviations from the base model so that less complicated models are favored. We specify the PC prior for the measurement error precision $1/\sigma_\delta^2$ such that $P(\sigma_\delta > 1) = 0.01$. This ensures that the estimated variance of the latent Gaussian variables is not unrealistically large. We specify PC priors for the precision of the AR(1) term $1/\sigma_z^2$ such that $P(\sigma_z > 1) = 0.01$. We specify PC priors for

the autoregressive parameters $\rho_s$ and $\rho_t$ such that $P(\rho_s > 0) = 0.9$ and $P(\rho_t > 0) = 0.9$, assuming base models where $\rho_s$ and $\rho_t$ are equal to 1 (Sørbye and Rue 2017).

When fitting the conditional thickness model, we would like to work with hundreds of thousands of observations, but naively applying the Matérn correlation function for $h$ leads to dense covariance matrices, making computation intractable. Instead we use the INLA-SPDE approach (Lindgren and Rue 2015). We approximate the continuously indexed Gaussian random field $h$ via a discretely indexed Gaussian Markov random field (GMRF) following the stochastic partial differential equation (SPDE) representation of Lindgren et al. (2011). The SPDE approach approximates a Gaussian random field by combining a set of piecewise linear basis functions defined on a triangular mesh of locations over the spatial domain. Basis function weights are modeled as a GMRF with a sparse precision matrix which facilitates computation. Rue and Held (2005) provide details on algorithms for working with such GMRFs. The precision matrix of the GMRF is specified so that the resulting continuously indexed field has a correlation function that is approximately Matérn. In particular, the resulting field is an approximate solution to an SPDE whose exact stationary solutions are Gaussian random fields with Matérn correlation functions.

For the conditional thickness model, we fix the Matérn smoothness parameter $\nu = 1$ but as with the contour model, we adopt PC priors for the remaining hyperparameters. We specify the PC prior for the measurement error precision $1/\tau_\delta^2$ such that $P(\tau_\delta > 2) = 0.01$. Furthermore, we specify the PC prior for the autoregressive parameter $\phi_t$ such that $P(\phi_t > 0) = 0.9$ (assuming a base model where $\phi_t$ is equal to 1). Finally, following Fuglstad et al. (2019), we also specify a joint PC prior for the practical range $r = \sqrt{8}/\kappa$ and marginal standard deviation $\tau_h$ parameters for the Matérn random field $h$ such that $P(r < 2000) = .5$ and $P(\tau_h > 2) = 0.01$, where $r$ is measured in kilometers and $\tau_h$ is measured in meters. The prior for $r$ was chosen because the reanalysis domain has a radius that is approximately 2000 km and the prior for $\tau_h$ was chosen because sea ice rarely exceeds 10 m in thickness.

## 3. IMPLEMENTATION

### 3.1. OBSERVATIONS

Although recent advances in remote sensing have made estimates of sea ice thickness more accessible, a temporally and spatially complete record of sea ice thickness in the Arctic remains out of reach. Direct measurements based on ice cores, underwater buoys and airborne radar observations provide data only for scattered locations. As summarized by Sallila et al. (2019), satellite-derived ice thickness estimates vary widely and almost none provide measurements in the summer, when melt ponds on the sea ice surface can lead to poor data quality. We therefore turn to thickness data products that provide temporally and spatially comprehensive estimates of sea ice thickness. These products generally use data assimilation methods to generate *reanalysis* datasets that represent the best estimate of the true sea ice thickness at a given time based on a given dynamic model and set of satellite estimates. We use the term data assimilation to refer to methods for using observational data to update estimates of state variables in a numerical model, following Nychka and Anderson

(2010). In the summer, when satellite-based estimates of thickness are unavailable, the estimates rely entirely on model-based estimates of sea ice thickness.

We note that reanalyses differ from retrospective forecasts: a reanalysis assimilates observations between short runs of a dynamical model, while our forecast relies on no observations beyond the initial conditions. The reanalysis estimates can be seen as reconstructions of the historical sea ice thickness, while our forecasts attempt to predict future conditions. Our approach of using reanalysis estimates to generate statistical forecasts is analogous to previous efforts in weather prediction (Weyn et al. 2019).

We train our models using reanalysis estimates of monthly Arctic sea ice thickness for 2007–2018 from the Alfred Wegener Institute Climate Model (AWI-CM) (Mu et al. 2020a). Since the AWI-CM assimilates satellite measurements starting only in 2011, we constrain our analysis to the period 2011–2018. Although this data product is not based solely on observations, we believe that it represents the best available "ground truth" record of sea ice thickness (Mu et al. 2020b). The AWI-CM thickness estimates are provided on an irregular grid of locations. Since the physical behavior of very thin ice is qualitatively different from that of typical sea ice (Haas 2016), we restrict our analysis to sea ice that is at least 0.15 meters thick, setting thickness values below that threshold to zero. We regrid all thickness estimates, modeled and observational, to the AWI-CM model grid for comparison.

When fitting the contour models, we use monthly averaged sea ice concentration estimates from the National Aeronautics and Space Administration Nimbus-6 SMMR and DMSP SSM/I-SSMIS satellites provided by the NSIDC (Comiso 2017). We define the fixed points and angles necessary for the contour model using the NSIDC grid, yielding 429 total line segments used among the various regions to define the boundary of the ice-covered region. We say that sea ice is present in a grid cell when the ice-covered area is at least 15%, as is conventional in sea ice research due to the low reliability of observations at low concentrations. The observed ice-covered region is used to compute the appropriate transformed variables **w**.

### 3.2. TRAINING THE CONTOUR MODEL

We desire monthly sea ice thickness forecasts extending one to three months into the future. For any time $t$, we train the contour model for a given region using a historical training dataset to estimate $\mu_{m(t)}(i)$ and hyperparameters $\boldsymbol{\theta}_w = (\rho_s, \rho_t, \sigma_\delta, \sigma_z)$. For each time $t$ in the historical training period and each line $l(i, t)$ in a region, we compute $y_i(t)$, the length from the fixed start point $b_i$ to the observed ice edge along $l(i, t)$. We use the approaches outlined in Director et al. (2017) to address rare issues with lines crossing land or with contours with self-intersections, using the Douglas–Peucker algorithm (Douglas and Peucker 2006), as implemented in the IceCast R package (Director et al. 2020).

Since we train the model separately for each initialization month and make forecasts extending only a few months into the future, we treat reanalysis estimates for a given month as independent across years. We use INLA to approximate posterior marginal distributions of all hyperparameters. We fit separate models for each region and each initialization month, allowing some variability in hyperparameters across space and time.

### 3.3. Training the Conditional Thickness Model

When training the conditional thickness model for a given forecast initialization date, we estimate the month-specific mean parameters $\eta_{m(t)}(s)$ using data for the desired month from previous years' thickness values, as these parameters represent the historical mean. However, we estimate the remaining model hyperparameters $(\tau_\delta, \tau_h, \phi_t, \kappa, \nu)$ using data from the most recently observed months in order to model changes in sea ice thickness over the recent past. As such, when training and evaluating our model, we specify three disjoint time periods. First, we specify the *testing period* consisting of the months immediately following the initialization month for which we want forecasts of sea ice thickness. Next, we select the *trend training period* including the initialization month and possibly some months immediately before it. Finally, we specify the *climatology training period* including all months in the trend training period and forecast period for some number of years preceding the forecast year. The testing period is for model validation, while the trend training period data is for estimating trends in differences between observed values and the climatological average and the climatology training period is used to estimate the climatological average. We consider only locations for which ice is present for at least one month during our climatological training period.

We split the conditional thickness model training procedure into two steps. First, we use data from the climatology training period to estimate the location- and month-specific conditional thickness means $\eta_m(s)$. For each location $s$ and month $m$, we estimate $\widehat{\eta_m}(s)$ to be the sample mean sea ice thickness for that location and month, for the years in the climatology training period when ice is present. Next, we use data from the trend training period and estimated location- and month-specific conditional thickness means $\widehat{\eta_{m(t)}}(s)$ to estimate the model hyperparameters and the posterior distributions of values of the latent random field $h$ for our trend training period.

To do this, we use the following model for all $s \in \mathbf{C}(t)$:

$$f(s, t) = \widehat{\eta_{m(t)}}(s) + h(s, t) + \delta_f(s, t).$$

We construct an observation vector for all data in the trend training period. We implement our model via the INLA-SPDE approach for approximating spatial Gaussian random fields, first constructing a triangular mesh of basis functions over our reanalysis domain. We then construct a spatio-temporal GMRF indexed spatially by the vertices of the mesh and temporally by month. Using this GMRF to approximate our random field $h$, we approximate the posterior marginal distributions of all conditional thickness model hyperparameters $\boldsymbol{\theta}_f = \{\tau_\delta, \tau_h, \phi_t, \kappa, \nu\}$, fixing the Matérn smoothness parameter $\nu = 1$. Intuitively, this corresponds to using only data from the most recent months prior to initialization to learn about the dynamics of the thickness field.

### 3.4. Generating Forecasts

Based on the estimated posterior distributions for all model hyperparameters, we use our combined model's posterior predictive distribution to obtain thickness forecasts. Assume that we initialize our model at time $t$ and generate forecasts for times $t + 1, \ldots, t + k$. Our

first goal is to generate probabilistic forecasts of thickness at individual locations and times; that is, we desire posterior marginal distributions of the form

$$f(s, t + i) \mid \mathbf{f}(t), \ldots, \mathbf{f}(t - j). \tag{13}$$

Here, $f(s, t + i)$ denotes the thickness at location $s$ and forecast time $t + i$, $i = 1, \ldots, k$, and $\mathbf{f}(t), \ldots, \mathbf{f}(t - j)$ denote the observed thickness fields for our trend training period.

We also seek probabilistic forecasts of aggregate quantities like sea ice volume. The R-INLA package offers the ability to compute approximate posterior marginal distributions for linear combinations of the latent field $h$. Combining our hyperparameter estimates with the area of each cell in our reanalysis grid thus allows us to approximate the posterior marginal distribution of $v(t + i)$, the volume at time $t + i$:

$$v(t + i) \mid \mathbf{f}(t), \ldots, \mathbf{f}(t - j). \tag{14}$$

To approximate distributions (13) and (14), we generate sample forecast contours and compute approximate posteriors *conditional on the sampled forecast contours*. We first sample from the posterior predictive distribution of the contour $\mathbf{c}(t + i)$:

$$\mathbf{c}(t + i) \mid \mathbf{c}(t), \ldots, \mathbf{c}(t - j). \tag{15}$$

We then compute posterior marginal distributions of the desired predictands, conditional on the sampled contour. If $\widetilde{\mathbf{c}}^{(a)}(t + j)$ denotes the $a$-th sample, then we wish to approximate the posterior distributions of the following variables:

$$f(s, t + j) \mid \widetilde{\mathbf{c}}^{(a)}(t + j), \mathbf{f}(t), \ldots, \mathbf{f}(t - j), \tag{16}$$

$$v(t + j) \mid \widetilde{\mathbf{c}}^{(a)}(t + j), \mathbf{f}(t), \ldots, \mathbf{f}(t - j). \tag{17}$$

Approximating these posterior distributions is non-trivial, so we go into further detail below. For now, assuming we have approximated these posteriors for all sampled contours, we integrate numerically over the sampled contours to generate the desired distributions in (13) and (14) for $f(s, t + j)$ and $v(t + j)$. Based on these approximate posterior marginal distributions, we compute point forecasts using the posterior means and interval forecasts by taking appropriate quantiles.

We sample forecast contours for each region by sampling from the posterior predictive distribution $\mathbf{w}(t + 1), \ldots, \mathbf{w}(t + k) \mid \mathbf{w}(t)$ instead of sampling from (15) directly. We combine the sampled $\mathbf{w}$ values for all sub-regions and transform them to obtain the draw $\widetilde{\mathbf{c}}^{(a)}(t + 1), \ldots, \widetilde{\mathbf{c}}^{(a)}(t + k)$. Here, we refer to $\widetilde{\mathbf{c}}^{(a)}(t + j)$ as a *sample contour*, corresponding to a single time point, and to $\widetilde{\mathbf{c}}^{(a)}(t + 1), \ldots, \widetilde{\mathbf{c}}^{(a)}(t + k)$ as a *sample contour trajectory*, corresponding to a prediction across multiple times. Using these sample contour trajectories we can predict which locations will have ice present at times $t + 1, \ldots, t + k$.

For each sample contour trajectory $\widetilde{\mathbf{c}}^{(a)}(t + 1), \ldots, \widetilde{\mathbf{c}}^{(a)}(t + k)$, we identify ice-covered locations and then use R-INLA to approximate posterior marginal distributions of thickness and volume conditional on the sample contour; that is, we approximate the distributions of

(16) and (17). We define a spatio-temporal model for $f(s, t+j) \mid \widetilde{\mathbf{c}}^{(a)}(t+j), \mathbf{f}(t), \ldots, \mathbf{f}(t-j)$ treating the conditioning on the sample contour $\widetilde{\mathbf{c}}^{(a)}(t+j)$ to mean that the sample contour only determines which locations and times to include in our observation vector. We call the resulting model the spatio-temporal model.

Under an alternative interpretation, we could also condition upon sea ice thickness *values* on the edge of the ice-covered region. Assuming that sea ice thickness will be nearly zero at the edge of the ice-covered region, we could add artificial observations at locations along the edge to our observation vector, setting sea ice thickness at these locations along the edge to be very small. Upon implementation, we found that the resulting forecasts generally underperformed when compared with the original spatio-temporal forecasts. We hypothesize that this may be because sea ice thickness does not necessarily decay smoothly to zero moving towards the ice edge; instead, the drop off from ice-covered regions to the ocean can be quite abrupt. As such, conditioning on the values at the edge of the ice-covered region leads to forecasts that underestimate sea ice thickness at the edge.

To generate forecasts, we sample twenty-five contours and approximate the posterior distributions of thickness and volume conditional on each of the sampled contours, which have the forms in (13) and (14). We then combine these conditional distributions to obtain posterior marginal distributions of the thickness at each location and time. Since the distributions conditional on contours in (16) and (17) are Gaussian, we approximate the unconditional posterior marginal distributions (13) and (14) by mixtures of 25 Gaussian distributions. We obtain point forecasts by taking the posterior means of $f(s, t + j)$ and $v(t + j)$ and interval forecasts via numerical computation of the quantiles of the posterior marginal distributions. We set any forecasts of negative thickness to be equal to zero.

## 4. RESULTS: STATISTICAL MODEL-BASED FORECASTS OF ARCTIC SEA ICE THICKNESS AND VOLUME

### 4.1. OVERVIEW OF DATA AND EXISTING FORECASTS

To assess our method, we generate example forecasts of Arctic sea ice for data from initialization months from January 2015 to October 2018. We set the trend training period for our conditional thickness model to be three months and the climatology training period to be four years. We generate forecasts for up to three months into the future from the initialization month. As an example, for initialization month July 2015, our trend training period is May–July 2015, our testing period is August–October 2015, and our climatology training period includes the months May–October for the years 2011–2014.

For each validation month and lead time, we compare our statistical model-based forecasts with the European Centre for Medium-Range Weather Forecasts (ECMWF) ensemble forecast, a climatological forecast, and a forecast based on a non-spatial conditional thickness model. The ECWMF forecasts are based on a 25-member ensemble of a dynamical model that is initialized each month, providing daily forecasts up to 215 days into the future (Johnson et al. 2019). For comparison with the AWI-CM thickness estimates, we regrid the ECMWF output to the AWI-CM model grid using bilinear interpolation (Hijmans and van Etten 2012) and average the daily forecasts for each month to obtain monthly forecasts. All

datasets generated and/or analyzed during the current study are available either upon request from the corresponding agency or from the corresponding author on reasonable request.

The climatological forecast of thickness for a given location and time is equal to the sample location- and month-specific mean over the climatology training period. For instance, the forecasts for August 2015 would be the sample average thickness from the August 2011–2014 reanalysis estimates for each location. The non-spatial version of our two-stage thickness model combines the contour model with a second-stage conditional thickness model that assumes no spatial correlation; we maintain the AR(1) assumption in time but assume all thickness estimates $f(s, t)$ are independently distributed in space.

We evaluate the performance of all forecasts by comparing them with the AWI-CM reanalysis estimates, taken to represent the ground truth, and computing evaluation metrics to assess accuracy and calibration. The AWI-CM reanalysis estimates incorporate information from both satellite-based measurements and an underlying dynamical model. During the summer months when satellite measurements are unavailable (May–September), the AWI-CM estimates are more dependent on the dynamical model, so we provide some separate results for summer months and non-summer months. We also compare our forecasts with airborne measurements of sea ice thickness taken annually in April of 2015–2018 by Operation Ice Bridge Quick Look, a NASA mission that took airborne measurements of sea ice (Kurtz 2015). These measurements are collected along specified flight paths off the coast of Greenland. As such, these estimates provide information about only a small region of the Arctic, and only in April. Nevertheless, we use them to provide some independent validation.

Figure 3 plots example forecasts for October 2015, initialized in September 2015. The top row provides posterior means from the proposed statistical models, which we use as point forecasts. The bottom row provides the ECMWF and climatological forecasts as well as true AWI-CM thickness estimates. The point forecasts from the statistical models are similar and align well with the AWI-CM thickness estimates, whereas the climatological thickness forecasts are too smooth and the ECMWF forecasts predict ice that is too thin for much of the Arctic. The most dramatic differences between the statistical model forecasts and the AWI-CM thickness estimates occur in the Kara/Barents Seas region (see the top right quadrant of each panel), suggesting that forecasts may be less accurate at the edge of the ice sheet. For this example, for the central Arctic region, we obtain marginal posterior medians for the contour model hyperparameters $(\sigma_\delta, \sigma_z, \rho_s, \rho_t) = (0.034, 0.818, 0.783, 0.083)$. This shows strong spatial correlation between nearby values of $w$ and also shows that the bulk of the variance comes from the spatio-temporal $z$ process. The hyperparameters for the conditional thickness model vary for different forecast months but generally indicate strong spatial correlation between values of $z$ at nearby locations. The marginal posterior medians for the conditional thickness model hyperparameters were $(\tau_\delta, \tau_h, \phi_t, r) = (0.049, 1.001, 0.753, 1023.285)$, indicating that the bulk of the variability in the $f$ process comes from the spatio-temporal Gaussian random field $h$. The hyperparameters for the conditional thickness model vary between forecast months but generally indicate strong temporal and spatial correlations between thicknesses at nearby locations and times. For both the contour model and the conditional thickness model, we fit the model with multiple variants of the priors described above, but we did not find the forecasts to be sensitive to the prior specification.
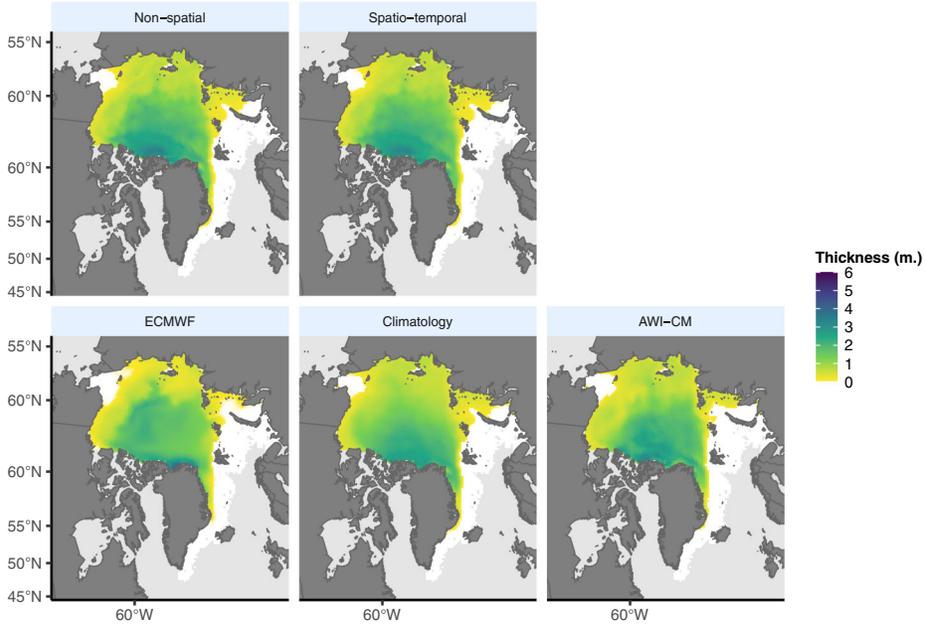
Figure 3.   Example Arctic sea ice thickness forecast means for October 2015 (1-month lead) for all forecast methods with AWI-CM thickness field for comparison .

## 4.2. RESULTS FOR FORECASTS OF SEA ICE THICKNESS

We first consider the point forecasts and interval forecasts generated by our model independently for each location–month pair. To quantify the accuracy of point forecasts, we compute the root-mean-squared error (RMSE):

$$\text{RMSE}(t) = \sqrt{\frac{1}{|S_t|} \sum_{s \in S_t} (f(s, t) - \widehat{f}(s, t))^2},$$

and mean absolute error (MAE):

$$\text{MAE}(t) = \frac{1}{|S_t|} \sum_{s \in S_t} |f(s, t) - \widehat{f}(s, t)|,$$

where $t$ indexes time and $S_t$ represents the set of AWI-CM grid locations at time $t$.

To assess how well uncertainty has been quantified, we compute empirical coverage rates for 90% prediction intervals, i.e., the sample proportion of observed values that are contained in our prediction intervals. An ideal method would exhibit empirical coverage rates that are close to the nominal coverage rate. We also compute the mean interval score

Table 1. Comparison of error metrics for forecasts of sea ice thickness for individual location/time pairs by lead time, for **non-summer** months (October through April) and **summer** months (May through September) between April 2015 and September 2018

|  | Forecast | Lead | RMSE | MAE | 90% Coverage | MIS |
|---|---|---|---|---|---|---|
| Non-summer | Climatology | 1 | 0.375 | 0.241 | NA | NA |
|  |  | 2 | 0.372 | 0.238 | NA | NA |
|  |  | 3 | 0.372 | 0.238 | NA | NA |
|  | ECMWF | 1 | 0.538 | 0.397 | 0.154 | 6.672 |
|  |  | 2 | 0.540 | 0.394 | 0.275 | 5.277 |
|  |  | 3 | 0.560 | 0.411 | 0.333 | 4.714 |
|  | Non-spatial | 1 | 0.241 | 0.166 | 0.824 | 1.189 |
|  |  | 2 | 0.314 | 0.218 | 0.806 | 1.563 |
|  |  | 3 | 0.342 | 0.238 | 0.814 | 1.665 |
|  | Spatio-temporal | 1 | 0.236 | 0.160 | 0.914 | 1.196 |
|  |  | 2 | 0.310 | 0.213 | 0.906 | 1.483 |
|  |  | 3 | 0.338 | 0.232 | 0.904 | 1.578 |
| Summer | Climatology | 1 | 0.420 | 0.253 | NA | NA |
|  |  | 2 | 0.420 | 0.253 | NA | NA |
|  |  | 3 | 0.420 | 0.252 | NA | NA |
|  | ECMWF | 1 | 0.714 | 0.541 | 0.160 | 8.494 |
|  |  | 2 | 0.802 | 0.617 | 0.308 | 6.944 |
|  |  | 3 | 0.881 | 0.682 | 0.379 | 6.334 |
|  | Non-spatial | 1 | 0.289 | 0.195 | 0.754 | 1.605 |
|  |  | 2 | 0.350 | 0.238 | 0.742 | 1.905 |
|  |  | 3 | 0.378 | 0.257 | 0.731 | 2.015 |
|  | Spatio-temporal | 1 | 0.292 | 0.199 | 0.829 | 1.526 |
|  |  | 2 | 0.354 | 0.241 | 0.819 | 1.745 |
|  |  | 3 | 0.379 | 0.256 | 0.819 | 1.803 |

(MIS), following Gneiting and Raftery (2007):

$$\text{MIS}(t) = \frac{1}{|S_t|} \sum_{s \in S_t} \left[ (u(s,t) - l(s,t)) + \frac{2}{\alpha}(l(s,t) - f(s,t))\mathbf{1}(f(s,t) < l(s,t)) \right.$$
$$\left. + \frac{2}{\alpha}(f(s,t) - u(s,t))\mathbf{1}(u(s,t) < f(s,t)) \right]$$

where $u(s,t)$ and $l(s,t)$ represent the upper and lower bounds of our $90\% = (1 - \alpha)\%$ prediction interval for location $s$ at time $t$. Smaller values of the interval score are better: it is large when the truth is located far outside the prediction intervals and when the interval is long.

Table 1 compares the forecasts, taking the AWI-CM thickness reanalysis estimates as ground truth. The top half summarizes the RMSE, MAE, MIS, and empirical 90% prediction interval coverage rates for the four methods for non-summer months (October through April) in the validation period April 2015–September 2018; the bottom half provides the same for the summer months (May through September). The climatology approach produces only point forecasts, so there are no prediction intervals. For all the statistical forecasts, RMSE and MAE increase with lead time, but the statistical forecasts consistently outperform the climatological and ECMWF forecasts. Table 1 also provides empirical 90%

Table 2.  Comparison of error metrics for forecasts of sea ice thickness for individual location/time pairs by lead
time for Operation Ice Bridge observations for April 2015–2018

| Forecast | Lead | RMSE | MAE | 90% coverage | MIS |
|---|---|---|---|---|---|
| Climatology | 1 | 0.781 | 0.575 | NA | NA |
| | 2 | 0.781 | 0.575 | NA | NA |
| | 3 | 0.781 | 0.575 | NA | NA |
| ECMWF | 1 | 0.742 | 0.560 | 0.097 | 11.628 |
| | 2 | 0.744 | 0.556 | 0.176 | 9.944 |
| | 3 | 0.769 | 0.574 | 0.230 | 7.383 |
| Non-spatial | 1 | 0.773 | 0.545 | 0.417 | 7.316 |
| | 2 | 0.777 | 0.556 | 0.434 | 6.886 |
| | 3 | 0.780 | 0.567 | 0.429 | 6.988 |
| Spatio-temporal | 1 | 0.781 | 0.549 | 0.535 | 6.497 |
| | 2 | 0.780 | 0.558 | 0.672 | 5.343 |
| | 3 | 0.772 | 0.562 | 0.692 | 5.210 |

prediction interval coverage rates for prediction intervals for each location–time pair. The
spatial forecasts achieve empirical coverage rates that are closest to the nominal coverage
rate, whereas the non-spatial forecasts exhibit slight undercoverage. The ECMWF inter-
val forecasts derived from the ensemble spread exhibit considerable undercoverage. For all
models, under-coverage is generally more problematic in the summer months than in the
non-summer months. The spatio-temporal model performs best in terms of MIS, indicating
superior quantification of forecast uncertainty. Tables S1 through S3 in the Supplementary
Material provide empirical coverage rates and mean interval scores by region for the three
regions where forecasts were generated. We observe more severe under-coverage outside
the Central Arctic, indicating that the statistical model performs worst at the edge of the sea
ice sheet.

Table 2 summarizes the RMSE, MAE, MIS, and empirical 90% prediction interval cov-
erage rates for each method when treating Operation Ice Bridge measurements as ground
truth for April 2015–2018. In terms of point forecast accuracy, all four forecasts perform
comparably. The ECMWF forecasts perform slightly better than the statistical ones in terms
of RMSE, and slightly worse in terms of MAE. The spatio-temporal model performs much
better than the ECMWF forecasts in terms of uncertainty quantification, as seen from the
MIS and coverage rates, but still exhibits some undercoverage. Operation Ice Bridge obser-
vations include some of the locations with the thickest sea ice throughout the Arctic, which
can be less well represented by physics-based models such as that used for the AWI-CM
reanalysis.

In Figure S1 in the Supplementary Material, we provide quantile–quantile plots for each
month in 2016 comparing the quantiles of the thickness anomalies (obtained by subtracting
the climatological conditional thickness means from the observed AWI-CM thickness esti-
mates) with those of a standard Gaussian distribution. These anomalies represent the data
used to train the conditional thickness model. These plots indicate that during all months,
the left tail of the observed distribution is heavier than that of a Gaussian; during the start of
the melting season in summer, both tails are heavier than that of a Gaussian. In summer, in

Table 3. Comparison of empirical coverage rates for 90% prediction interval estimates of sea ice volume by lead time and region

| Forecast | Lead | All selected regions | Greenland sea | Kara/Barents | Central arctic |
|---|---|---|---|---|---|
| ECMWF | 1 | 0.08 | 0.19 | 0.09 | 0.10 |
| | 2 | 0.06 | 0.51 | 0.13 | 0.09 |
| | 3 | 0.09 | 0.49 | 0.15 | 0.10 |
| Non-spatial | 1 | 0.09 | 0.02 | 0.07 | 0.11 |
| | 2 | 0.05 | 0.05 | 0.09 | 0.05 |
| | 3 | 0.09 | 0.00 | 0.05 | 0.09 |
| Spatio-temporal | 1 | 0.95 | 0.77 | 1.00 | 0.95 |
| | 2 | 0.89 | 0.80 | 1.00 | 0.91 |
| | 3 | 0.89 | 0.73 | 1.00 | 0.86 |

particular, increased variability may be attributed to melting and may not be well modeled by a Gaussian process. This may be a major source of the under-coverage in the prediction intervals. In general, under-coverage is most prominent at locations along the sea ice edge, but occurs throughout the Arctic as well.

### 4.3. RESULTS FOR FORECASTS OF SEA ICE VOLUME

We evaluate the forecasts' ability to account for spatial correlation by computing forecasts of regional sea ice volume based on the thickness forecasts. Table 3 provides empirical coverage rates for 90% prediction intervals of Northern Hemisphere sea ice volume overall validation months from April 2015 to September 2018. The ECMWF prediction intervals display clear undercoverage. Similarly, the non-spatial model forecasts exhibit significant undercoverage, since they do not account for spatial correlation in forecast errors. The spatio-temporal forecasts exhibit empirical coverage rates that are closest to nominal, with undercoverage in the Greenland Sea and overcoverage in the Kara/Barents Seas.

## 5. DISCUSSION

We have proposed a spatio-temporal two-stage statistical model for Arctic sea ice thickness. It generates forecasts that improve upon existing forecasts and provides useful information about short-term changes in Arctic sea ice thickness and volume. Our approach combines a contour model for predicting the ice-covered area of the Arctic with a Gaussian random field model for conditionally generating thickness fields given a specified ice-covered region. In the first stage of our model, we use a contour model to generate plausible ice-covered regions for a given forecast month. Large continuous portions of the Arctic Ocean are always covered by ice throughout the year, so we desire an approach that can generate sample ice-covered regions that include these persistent, continuous regions.

Extant statistical models for sea ice presence may not produce samples with realistic ice-covered regions. Trend-adjusted quantile mapping, as proposed by Dirkson et al. (2018, 2019) can provide only independent forecasts of sea ice presence for each location–time pair.

Similarly, damped anomaly persistence forecasts, as applied to sea ice extent by Blanchard-Wrigglesworth et al. (2015), is typically used to make predictions independently for different time series, meaning that using such an approach to forecast thickness at each location would not yield the desired spatial correlation in forecast errors. Our non-spatial model is similar to the damped anomaly persistence method, except that it lacks the time trend since we train on only a few years' worth of thickness estimates. Zhang and Cressie (2020) proposed a spatio-temporal latent Gaussian model to predict the ice-covered region. Their model assumes observations from latent Gaussian models to be conditionally independent given the latent random field. As long as the sea ice thickness mean is located away from zero and the spatial dependence is captured by the model, their approach should generate realistic ice-covered regions in practice. Our approach similarly relies upon the sea ice thickness mean being typically positive and we find our contour model generates realistic ice-covered regions with the desired continuity. We note, however, that the number of line segments used to define the contour model influences how accurate and realistic the resulting forecast contours appear. If we use too few segments, for example, any predicted ice-covered area may have an unrealistic shape. We find that in practice, the number of segments used here facilitates computational efficiency while producing contours that perform well.

The second stage of our model uses a spatio-temporal Gaussian random field to model thickness conditionally given an ice-covered region. For each month, we assume that the vector of sea ice thickness values is drawn from a spatio-temporal Gaussian random field. Accounting for both temporal and spatial correlation improves our forecasts in two key ways. First, modeling temporal correlation allows us to produce forecasts that incorporate information on recent trends. In addition, modeling spatial correlation improves calibration of interval forecasts of aggregate quantities such as sea ice volume. Accounting for temporal but not spatial correlation yields forecasts that are better for individual locations/times but uncalibrated for aggregate quantities such as sea ice volume.

For other applications, different summaries of sea ice thickness may be important. For example, forecasting locations with thin ice could be of interest when planning shipping routes. Bolin and Lindgren (2015) studied forecasts of excursion sets from latent Gaussian models. Our methods could be applied to the thickness field to study potential routes through the Arctic.

To our knowledge, our two-stage model is the first purely statistical model for generating monthly forecasts of sea ice thickness and the limitations of our method point to several exciting areas for future research. Notably, our Gaussian model for sea ice thickness (8) does not restrict thickness to be nonnegative. We adjust our forecasts to be nonnegative by setting any negative values to be zero; for both the non-spatial and spatial models, only 3% of forecast values are negative. One direction for further research is the use of a likelihood with nonnegative support instead of a Gaussian likelihood. We examined other potential strategies for modeling nonnegative, spatially correlated data, including modeling $f$ as a realization from a truncated multivariate normal distribution as done by Stein (1992). However, we found that our Gaussian random field model was more computationally efficient, while providing good forecast performance. Using the Gaussian likelihood also allows for the use of closed form formulas when computing interval estimates for simple aggregate quantities.

Moreover, as seen in Figure S1 in the Supplemental Material, our data used to train the Gaussian process for the conditional thickness model are not exactly Gaussian. As noted in Sect. 4.2, in many months, the observed distribution of sea ice thickness values has heavier tails than that of a Gaussian. Although we tried various transformations of our response variable to improve performance, as shown by the various distribution shapes represented in the grid of quantile–quantile plots, there may not be a single transformation that can produce approximately Gaussian values across all months. For this reason, more work is needed to improve modeling of potential heavy tails in the distribution of conditional sea ice thickness. In the absence of a clear improved alternative, we favored the untransformed Gaussian model for the sake of computational tractability and simplicity.

In addition, the performance of our approach varies across the Arctic, as indicated by our results in Sect. 4 and Tables S1 through S3 in the Supplementary Material. In particular, under-coverage is more problematic at the edge of the sea ice, though it does appear across the entire region. In fitting our model, we assume constant hyperparameters across all forecast months and locations. However, performance might be improved by allowing for temporally varying hyperparameters. Similarly, thicker ice behaves differently from thinner ice (Haas 2016), so we might see improvements by allowing for spatially non-stationary hyperparameters. Ingebrigtsen et al. (2014) incorporated covariates to allow non-stationarity in variance parameters; such an approach could improve local calibration of our probabilistic forecasts of thickness, reducing systematic under- or overcoverage in different locations. Also, some dynamical models of sea ice model the distribution of sea ice thickness within grid cells, offering more detailed information. Incorporating such information could potentially improve our forecasts but would increase the complexity of our statistical models.

Existing estimates of sea ice thickness from measurements and reanalyses show large variability, especially in the summer months when satellite observations are typically unavailable. Our approach could be extended to incorporate additional data from sources such as Operation Ice Bridge or upward looking sonar observations. Similarly, including measurements or estimates of other physical variables as covariates in our conditional thickness model could also improve the accuracy of our forecasts, at the cost of increasing model complexity.

Sea ice in the Arctic has changed rapidly in recent decades, and continued declines in sea ice volume may complicate forecasts of ice thickness. We show that statistical models of sea ice thickness can provide valuable insights both for sea ice researchers as well as for those working in shipping and resource management. Beyond introducing still more complex statistical models, we also seek to use information from physical models to improve forecast performance. Director et al. (2021) demonstrate the benefits of a post-processing approach that uses statistical modeling to correct biases and improper calibration of ensembles of dynamical models. Although our statistical model provides skillful forecasts at seasonal time scales, our forecasts only crudely approximate the dynamics that govern sea ice evolution.

We anticipate that dynamic sea ice models will continue to improve in the near future, especially as more data from satellite-based sources become available (Petty et al. 2020; Kwok and Cunningham 2015) and more direct assimilation of sea ice thickness data is used. One critical area for future work will therefore be to incorporate information from such physical models into our statistical approach.

# REFERENCES

Banerjee S, Gelfand AE, Finley AO, Sang H (2008) Gaussian predictive process models for large spatial data sets. J R Stat Soc Ser B (Stat Methodol) 70:825–848

Blanchard-Wrigglesworth E, Cullather RI, Wang W, Zhang J, Bitz CM (2015) Model forecast skill and sensitivity to initial conditions in the seasonal Sea Ice Outlook. Geophys Res Lett 42:8042–8048

Bolin D, Lindgren F (2015) Excursion and contour uncertainty regions for latent Gaussian models. J R Stat Soc Ser B (Stat Methodol) 77:85–106

Cavalieri DJ, Parkinson CL (2012) Arctic sea ice variability and trends, 1979–2010. Cryosphere 6:881–889

Chevallier M, Smith GC, Dupont F, Lemieux J-F, Forget G, Fujii Y, Hernandez F, Msadek R, Peterson KA, Storto A, Toyoda T, Valdivieso M, Vernieres G, Zuo H, Balmaseda M, Chang Y-S, Ferry N, Garric G, Haines K, Keeley S, Kovach RM, Kuragano T, Masina S, Tang Y, Tsujino H, Wang X (2017) Intercomparison of the Arctic sea ice cover in global ocean-sea ice reanalyses from the ORA-IP project. Clim Dyn 49:1107–1136

Comiso J (2017) Bootstrap Sea Ice Concentrations from Nimbus-7 SMMR and DMSP SSM/I-SSMIS, Version 3

Cressie N, Johannesson G (2008) Fixed rank kriging for very large spatial data sets. J R Stat Soc Ser B (Stat Methodol) 70:209–226

Datta A, Banerjee S, Finley AO, Gelfand AE (2016) Hierarchical nearest-neighbor gaussian process models for large geostatistical datasets. J Am Stat Assoc 111:800–812

Director HM, Raftery AE (2021+) Contour models for boundaries enclosing star-shaped and approximately star-shaped polygons. arXiv:2007.04386 [stat]. arXiv: 2007.04386

Director HM, Raftery AE, Bitz CM (2017) Improved sea ice forecasting through spatiotemporal bias correction. J Clim 30:9493–9510

Director HM, Raftery AE, Bitz CM (2020) IceCast: apply statistical post-processing to improve sea ice predictions

Director HM, Raftery AE, Bitz CM (2021+) Probabilistic forecasting of the arctic sea ice edge with contour modeling. to appear in Annals of Applied Statistics

Dirkson A, Denis B, Merryfield WJ (2019) A multimodel approach for improving seasonal probabilistic forecasts of regional arctic sea ice. Geophys Res Lett 46:10844–10853

Dirkson A, Merryfield WJ, Monahan AH (2018) Calibrated probabilistic forecasts of arctic sea ice concentration. J Clim 32:1251–1271

Douglas DH, Peucker TK (2006) Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. Cartogr Int J Geogr Inf Geovis 10:112–122

Fuglstad G-A, Simpson D, Lindgren F, Rue H (2019) Constructing priors that penalize the complexity of gaussian random fields. J Am Stat Assoc 114:445–452

Gneiting T, Raftery AE (2007) Strictly proper scoring rules, prediction, and estimation. J Am Stat Assoc 102:359–378

Haas C (2016) Sea ice thickness distribution. In: Thomas DN (ed) Sea Ice, 3rd. Wiley, New York, pp 42–64

Handcock MS, Raphael MN (2020) Modeling the annual cycle of daily Antarctic sea ice extent. Cryosphere 14:2159–2172

Heaton MJ, Datta A, Finley AO, Furrer R, Guinness J, Guhaniyogi R, Gerber F, Gramacy RB, Hammerling D, Katzfuss M, Lindgren F, Nychka DW, Sun F, Zammit-Mangion A (2019) A case study competition among methods for analyzing large spatial data. J Agric Biol Environ Stat 24:398–425

Hijmans RJ, van Etten J (2012) raster: Geographic analysis and modeling with raster data

Ingebrigtsen R, Lindgren F, Steinsland I (2014) Spatial models with explanatory variables in the dependence structure. Spat Stat 8:20–38

Johnson SJ, Stockdale TN, Ferranti L, Balmaseda MA, Molteni F, Magnusson L, Tietsche S, Decremer D, Weisheimer A, Balsamo G, Keeley SPE, Mogensen K, Zuo H, Monge-Sanz BM (2019) SEAS5: the new ECMWF seasonal forecast system. Geosci Model Dev 12:1087–1117

Katzfuss M, Guinness J, Gong W, Zilber D (2020) Vecchia approximations of Gaussian-process predictions. arXiv:1805.03309 [stat]

Kharin VV, Boer GJ, Merryfield WJ, Scinocca JF, Lee W-S (2012) Statistical adjustment of decadal predictions in a changing climate. Geophys Res Lett 39(L19705):1–6

Krainski ET, Gómez-Rubio V, Bakka H, Lenzi A, Castro-Camilo D, Simpson D, Lindgren F, Rue H, Gómez-Rubio V, Bakka H, Lenzi A, Castro-Camilo D, Simpson D, Lindgren F, Rue H (2018) Advanced spatial modeling with stochastic partial differential equations using R and INLA. Chapman and Hall/CRC

Kurtz N (2015) IceBridge sea ice freeboard, snow depth, and thickness quick look, version 1. type: dataset

Kwok R, Cunningham GF (2015) Variability of arctic sea ice thickness and volume from CryoSat-2. Philos Trans R Soc A Math Phys Eng Sci 373:20140157

Lindgren F, Rue H (2015) Bayesian spatial modelling with R-INLA. J Stat Softw 63:1–25

Lindgren F, Rue H, Lindström J (2011) An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. J R Stat Soc Ser B (Stat Methodol) 73:423–498

Melia N, Haines K, Hawkins E, Day JJ (2017) Towards seasonal Arctic shipping route predictions. Environ Res Lett 12:084005

Mu L, Nerger L, Tang Q, Losa SN, Sidorenko D, Wang Q, Semmler T, Zampieri L, Losch M, Goessling HF (2020a) Experiment output data from a seamless sea ice prediction system based on AWI-CM 1.1. type: dataset

Mu L, Nerger L, Tang Q, Losa SN, Sidorenko D, Wang Q, Semmler T, Zampieri L, Losch M, Goessling HF (2020b) Toward a data assimilation system for seamless sea ice prediction based on the AWI climate model. J Adv Model Earth Syst 12:e2019MS001937 eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2019MS001937

Notz D, Stroeve J (2018) The trajectory towards a seasonally ice-free arctic ocean. Curr Clim Change Rep 4:407–416

Nychka D, Bandyopadhyay S, Hammerling D, Lindgren F, Sain S (2015) A multiresolution gaussian process model for the analysis of large spatial datasets. J Comput Graph Stat 24:579–599

Nychka DW, Anderson JL (2010) Data assimilation. In: Gelfand AE, Diggle P, Guttorp P, Fuentes M (eds) Handbook of spatial statistics. CRC Press, Cambridge

Petty AA, Kurtz NT, Kwok R, Markus T, Neumann TA (2020) Winter arctic sea ice thickness from ICESat-2 freeboards. J Geophys Res Oceans 125

Rue H, Held L (2005) Gaussian markov random fields: theory and applications. CRC Press, Cambridge

Rue H, Martino S, Chopin N (2009) Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. J R Stat Soc Ser B (Stat Methodol) 71:319–392

Sallila H, Farrell SL, McCurry J, Rinne E (2019) Assessment of contemporary satellite sea ice thickness products for Arctic sea ice. Cryosphere 13:1187–1213

Sang H, Huang JZ (2012) A full scale approximation of covariance functions for large spatial data sets. J R Stat Soc Ser B (Stat Methodol) 74:111–132

Simpson D, Rue H, Riebler A, Martins TG, Sørbye SH (2017) Penalising model component complexity: a principled, practical approach to constructing priors. Stat Sci 32:1–28

Stein ML (1992) Prediction and inference for truncated spatial data. J Comput Graph Stat 1:91–110

Sørbye SH, Rue H (2017) Penalised complexity priors for stationary autoregressive processes. J Time Ser Anal 38:923–935

Vannitsem S, Wilks DS, Messner JW (Eds.) (2018) Statistical postprocessing of ensemble forecasts. Elsevier, Amsterdam

Weyn JA, Durran DR, Caruana R (2019) Can machines learn to predict weather? Using deep learning to predict gridded 500-hPa geopotential height from historical weather data. J Adv Model Earth Syst 11:2680–2693

Zhang B, Cressie N (2020) Bayesian inference of spatio-temporal changes of arctic sea ice. Bayesian Anal 15:605–631