Received: 05 September 2017 | Revised: 01 December 2017 | Accepted: 02 January 2018

DOI: 10.1002/qj.3242

RESEARCH ARTICLE

Quarterly Journal of the Royal Meteorological Society RMetS

A probabilistic verification score for contours: Methodology and application to Arctic ice-edge forecasts

H. F. Goessling[1] | T. Jung[1,2]

[1]Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research, Bremerhaven, Germany
[2]Department of Physics and Electrical Engineering, University of Bremen, Germany

Correspondence
H. F. Goessling, Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research, Am Handelshafen 12, 27570 Bremerhaven, Germany.
Email: helge.goessling@awi.de

We introduce a verification score for probabilistic forecasts of contours – the Spatial Probability Score (SPS). Defined as the spatial integral of local (Half) Brier Scores, the SPS can be considered the spatial analogue of the Continuous Ranked Probability Score (CRPS). Applying the SPS to idealized ensemble forecasts of the Arctic sea-ice edge in a global coupled climate model, we demonstrate that the metric responds in a meaningful way to ensemble size, spread, and bias. When applied to individual forecasts or ensemble means (or quantiles), the SPS is reduced to the 'volume' of mismatch, which in the case of the ice edge corresponds to the Integrated Ice Edge Error (IIEE). By comparing initialized forecasts with climatological and persistence forecasts, we confirm earlier findings on the potential predictability of the Arctic sea-ice edge from a probabilistic viewpoint. We conclude that the SPS is a promising probabilistic verification metric, for contour forecasts in general and for ice-edge forecasts in particular.

KEYWORDS

Arctic sea-ice edge, contour forecast, ensemble prediction, probabilistic score, spatial verification

1 | INTRODUCTION

Forecast verification is an important part of the workflow in environmental prediction. Verification is critical for monitoring and comparing forecast system performance, it guides model development, and it informs forecast users about the quality of predictions (Casati *et al.*, 2008). Predictions for different quantities require different types of verification metrics (or scores), ranging from summary metrics like the widely used 500 hPa geopotential height anomaly correlation to highly specific metrics like the Flight Time Error which is tailored to aviation needs (Rickard *et al.*, 2001). With the advent of ensemble-based probabilistic forecast systems, which have become the standard in medium-range weather prediction (Palmer, 2000; Gneiting and Raftery, 2005), came an increased need for probabilistic verification metrics. These take the full probabilistic forecast information into account instead of just evaluating the ensemble-mean or individual ensemble members in a deterministic fashion.

Prominent examples for probabilistic metrics are the (Half) Brier Score (Brier, 1950) for dichotomous (i.e. binary) events and the Continuous Ranked Probability Score (CRPS; Matheson and Winkler, 1976) for continuous quantities. The Brier Score is simply defined as the squared difference between the forecast probability of an event and its observed probability, the latter being either 1 or 0 (i.e. binary, assuming perfect observability). The CRPS is closely related as it integrates the Brier Score for the probability of a continuous variable surpassing a threshold over the range of possible thresholds. In contrast to the Brier Score, which has no units, the CRPS inherits the units of the continuous variable considered through the integration, allowing the interpretation of the CRPS as a meaningful distance of the forecast probability density function from the observed value. When applied

to deterministic forecasts, the Brier Score yields 1 for a false forecast and 0 for a true forecast, and the CRPS is reduced to the absolute error. This implies that both can be used to assess deterministic and probabilistic forecasts in the same framework. Another appealing property is that these metrics are strictly proper and thus resistant to hedging (e.g. Wilks, 2011).

Another element of interest when it comes to assessing the quality of a forecast is how well certain *spatial* features are forecast. For example, a rainband may have been forecast accurately in terms of timing, intensity, and shape but there might be a shift in location. A point-based evaluation will indicate poor skill and will not acknowledge the correct timing, intensity, and shape of such a forecast. To overcome this limitation, a number of spatial verification metrics that can be classified as neighbourhood, scale-separation, feature-based, and field-deformation approaches have been devised (Gilleland *et al.*, 2009). For example, the neighbourhood-type Fractions Skill Score does not require an exact alignment of forecast and observed patterns, but evaluates the statistics over a certain neighbourhood (e.g. Mittermaier and Roberts, 2010). An example for a useful feature-based verification metric is the partial or modified Hausdorff distance (e.g. Venugopal *et al.*, 2005; Dukhovskoy *et al.*, 2015) which measures essentially a mean distance between forecast and observed features. Importantly, however, most exisiting spatial verification metrics are not well suited to evaluate probabilistic forecasts. Notably, one of the aims of the ongoing community project MesoVICT (Mesoscale Verification Intercomparison over Complex Terrain) is to test the suitability of existing spatial verification methods for ensemble forecasts (Dorninger *et al.*, 2018).

Here we suggest a probabilistic verification score for contours that combines probabilistic and spatial forecast verification – the Spatial Probability Score (SPS). Being the spatial integral of local Brier Scores, the SPS can be considered the spatial analogue of the CRPS, inheriting numerous desirable properties of these classic metrics. Building on the introduction of the Integrated Ice Edge Error and the simulation experiments used in Goessling *et al.* (2016 hereafter G16), we demonstrate the behaviour of the SPS based on idealized ensemble forecasts of the Arctic sea-ice edge – a contour that receives increasing attention given the rapid Arctic warming and associated opportunities for shipping and other human activities (Emmerson and Lahn, 2012; Smith and Stephenson, 2013).

The article is structured as follows. In section 2 we introduce the Spatial Probability Score (SPS). In section 3 we apply the SPS to idealized ensemble forecasts. Firstly, we explain the simulation set-up (section 3.1); secondly we describe how we manipulate the perfect ensemble forecasts to mimic changes in ensemble size, spread, and bias (section 3.2); and thirdly we analyse how the SPS responds to these changes (section 3.3). In section 4 we discuss limitations and outline possible applications and extensions of the SPS, followed by a brief summary and conclusions in section 5.

2 | THE SPATIAL PROBABILITY SCORE

Consider an observed (o) and a forecast (f) spatial probability field $P(\mathbf{x})$ of a dichotomous event. We define the Spatial Probability Score (SPS) as

$$\mathrm{SPS} = \int_V \{P_\mathrm{f}(\mathbf{x}) - P_\mathrm{o}(\mathbf{x})\}^2 \,\mathrm{d}V\,, \tag{1}$$

where V is the 'volume' of interest spanned by an arbitrary number N of spatial dimensions, and $\mathbf{x} \in V$. The SPS inherits units length^N from V through $\mathrm{d}V$, which allows for an intuitive interpretation. Obviously, $\mathrm{SPS} = 0$ if and only if $P_\mathrm{f} = P_\mathrm{o}$ everywhere, which implies that the SPS is strictly proper and thus resistant to hedging (compare Wilks, 2011). Note that $P_\mathrm{f}, P_\mathrm{o} \in [0, 1]$ with P_o being binary (a field of zeros and ones) for perfect observations.

The squared term in Equation 1 is the local (Half) Brier Score (Brier, 1950). The SPS is thus mathematically equivalent to the Continuous Ranked Probability Score (CRPS) (Matheson and Winkler, 1976), with the dimension of a continuous physical quantity being replaced by one or more spatial dimensions. Accordingly, the cumulative distribution function used for the CRPS is replaced by the probability of being 'enclosed' by a contour.

In the case of probabilistic forecasts, P_f is generally a field of numbers in the continuous range $[0, 1]$. Typically, P_f will be derived from forecast ensembles, either directly from the relative frequencies of outcomes (raw probabilities), or with additional spatial 'dressing' of the raw probabilities (e.g. by Gaussian smoothing) and/or other types of adjustments such as bias correction.

If the dichotomous event under consideration is perfectly observed, P_o is a binary field, that is either 1 or 0. However, the definition of the SPS allows us to account for observational uncertainties in a meaningful way by assigning values between 0 and 1 to P_o; for example, this could be based on discrepancies between different observational products or between analyses of different operational centres (compare, e.g. Bauer *et al.*, 2016 figure 6). Note, however, that it is an open question how best to account for observational errors in probabilistic forecast verification, and that other approaches exist (e.g. Saetra *et al.*, 2004). In the present perfect-model study, we focus on situations where the truth is known perfectly.

While spatial dressing could also be used to generate quasi-probabilistic forecasts from single forecasts, the SPS is also well defined for raw single (i.e. deterministic) forecasts. In this case (assuming that the observations are also non-probabilistic) the SPS is reduced simply to the total volume of mismatch, that is, of all locations where the forecast and the observations disagree on the location being enclosed

by the contour. This property allows a meaningful evaluation and comparison of probabilistic and deterministic forecasts in the same framework.

Note that in forecast applications a dichotomous event is often defined by a continuous variable being above or below a particular threshold. In section 3 we consider the two-dimensional case of Arctic sea-ice concentration being above or below 15%.

3 | IDEALIZED ARCTIC ICE-EDGE FORECASTS

In the following we demonstrate the SPS with the two-dimensional case of ice-edge position forecasts. Using the most common definition for the ice edge – the 15% sea-ice concentration (sic) contour – Equation 1 becomes

$$\mathrm{SPS} = \int_x \int_y \{P[\mathrm{sic} > 0.15]_\mathrm{f}(x,y) - P[\mathrm{sic} > 0.15]_\mathrm{o}(x,y)\}^2 \,\mathrm{d}y\mathrm{d}x\,. \quad (2)$$

$P[\mathrm{sic} > 0.15]_\mathrm{f}$ is the so-called sea-ice probability, that is, the forecast probability of being enclosed by the ice edge. When applied to deterministic sea-ice edge forecasts, the SPS is reduced to the Integrated Ice Edge Error (IIEE) introduced in G16, allowing us to assess deterministic and probabilistic sea-ice edge forecasts in the same framework.

Before describing the simulation set-up, forecast types, and results based on Equation 2, we exemplify the analogy between the SPS and the CRPS with a simplified one-dimensional case: the location of the summer sea-ice edge on a line from the northern coast of Siberia (e.g. at 120°E) across the North Pole towards Greenland. Assuming that the summer ice edge always crosses this transect exactly once, which is mostly valid, the probability of being 'enclosed' by the contour increases monotonically from zero to one along the transect, and the probability can be interpreted as the cumulative distribution function of the location where the contour crosses the transect. In this case the SPS is reduced to one spatial dimension and equates to the CRPS with the continuous quantity being the distance from an arbitrary reference point on the transect. Limitations to this analogy are discussed in section 4, and in the following we return to the two-dimensional case given by Equation 2.

3.1 | Simulation set-up

Our analysis is based on simulations with the Alfred Wegener Institute Climate Model (AWI-CM; Sidorenko *et al.*, 2015; Rackow *et al.*, 2016). Originally, the experiments were conducted as a contribution to the Arctic Predictability and Prediction On Seasonal to Interannual TimEscales (APPOSITE; Tietsche *et al.*, 2014; Day *et al.*, 2016) project. Following the APPOSITE protocol, we conducted 'perfect model'-type simulations. Ensembles with small initial-condition perturbations (white noise with standard deviation 10^{-4} K added to the sea surface temperatures) were branched off at different points in time from a quasi-equilibrium, multi-centennial control integration with greenhouse gas concentrations held constant at 1990 levels. The results reported here are based on 18 nine-member ensembles initialized on 1 July from different years of the control run, integrated for three years. Details of the APPOSITE experimental set-up and data access are given elsewhere (Day *et al.*, 2016).

3.2 | Forecast types

We determine the behaviour of the SPS by analysing a range of forecast types, i.e. different probabilistic and deterministic forecasts derived from the forecast ensembles and from the control integration (Table 1). The different forecast types allow us to investigate how the SPS responds to ensemble size, spread, and bias, and how the ensemble forecasts relate to common reference forecasts.

TABLE 1 Forecast types

Description	ID	Ensemble size	Spread	Bias
Full ensemble	FULL	8	Reliable	Unbiased
Random subsample	RAND	4	Reliable	Unbiased
Single forecast	SINGLE	1	—	Unbiased
Climatological probabilities	CLIM	(200)	Reliable	Unbiased
Climatological median	CMED	(200→1)	—	Unbiased
Persistence	PER	(1)	—	Seasonally biased
Quantile range 0–0.5 inflated	HIGH	8*	Underdispersive	High-biased
Quantile range 0.5–1 inflated	LOW	8*	Underdispersive	Low-biased
Quantile ranges 0–0.25 & 0.75–1 inflated	OVER	8*	Overdispersive	Unbiased
Quantile range 0.25–0.75 inflated	UNDER	8*	Underdispersive	Unbiased
Median of full ensemble	MED	8→1	—	Unbiased

The arrows indicate that a single (median) contour is derived from a larger number of contours. Similarly, the stars denote types where the ensemble size is ill-defined because the probability fields based on the eight-member ensembles are manipulated by inflating certain quantile ranges.

The 'perfect model' approach followed here is particularly well suited to test the characteristics of a verification metric because the approach warrants reliability and precludes biases of the original forecast ensembles. Therefore, our premise for a proper verification metric (e.g. Wilks, 2011) is that deteriorative manipulations of the original forecasts based on the full ensembles must lead to lower skill according to the metric.

The eleven forecast types listed in Table 1 are illustrated in Figure 1b–l for an arbitrary forecast for 14 September (i.e. with 2.5 months lead time). One of the nine sea-ice edges comprising the forecast ensemble (Figure 1a) has been selected randomly as 'truth' (red contour in all panels). The remaining eight ice edges (or subsets thereof) are used to construct the ensemble-based forecast types.

3.3 | Results

In the following, we discuss the different forecast types, grouped such that they reveal insights into (i) how the full ensemble-based forecasts perform relative to common reference forecasts based on climatology and persistence (FULL, CLIM, CMED, PER), (ii) the influence of ensemble size (RAND4, SINGLE), (iii) the influence of spread (OVER, UNDER, MED), and (iv) the influence of bias (HIGH, LOW). The results for the SPS shown in Figure 2 and discussed below are averaged over 18 forecast cases (start dates), with each of the nine ensemble members selected once as truth to reduce sampling uncertainty.

3.3.1 | Full ensembles versus reference forecasts

The full ensemble-based forecasts (FULL; Figure 1b) are constructed from all ensemble members except the one selected as truth. Given the limited ensemble size, the sea-ice probabilities in the FULL forecast vary with steps of $1/8$. In contrast, the climatological forecast based on the 200 states of the control run (on the corresponding day of the year; Figure 1c) is not only much smoother, but also broader in terms of the ice-edge location. This hints at some potential predictability remaining at the depicted lead time and time of the year, as confirmed quantitatively below.

As additional reference forecasts, we consider the climatological median ice edge, i.e. the 50% sea-ice probability contour based on the control run (CMED; Figure 1e), and a simple persistence forecast where the ice edge is kept at its initial (1 July) location (PER; Figure 1f). Both these forecast types are deterministic in the sense that the forecast ice edge is exactly localized, meaning that the sea-ice probability is either 0 or 1 everywhere. In these cases the SPS is reduced to the Integrated Ice Edge Error (G16).

The SPS of the climatological forecasts (CLIM, solid grey curve in Figure 2) has a distinct seasonal dependence, with values around $0.6 \times 10^6\,\mathrm{km}^2$ in winter and spring, and around $0.9 \times 10^6\,\mathrm{km}^2$ in summer and autumn. This seasonal cycle is tied to strong seasonal variations of the mean ice-edge location and in particular associated with changes of the ice-edge length (G16). The shorter ice edge in winter and spring, when the Arctic ice cover is laterally bounded to a large extent by coastlines rather than an ice edge, implies a smaller area where forecast errors contribute to the SPS.

A coherent seasonal cycle is imprinted on all forecast types except the persistence forecasts (PER; dotted grey curve in Figure 2) which follow their own seasonality because they reflect a fixed time of the year (1 July) throughout the forecasts. While accurate at the very beginning, the average error of the persistence forecasts grows rapidly and exceeds the average SPS of CLIM after ~14 days. The persistence forecasts remain more skilful than forecasts based on the climatological median ice edge (CMED; dashed grey curve) out to day ~19, which has implications for how the ice edge should be prescribed in atmospheric forecast systems without interactive sea ice. For longer lead times, persistence is, not surprisingly, the least skilful forecast type, with errors peaking at $\sim 4 \times 10^6\,\mathrm{km}^2$ in September and $\sim 5 \times 10^6\,\mathrm{km}^2$ in March (exceeding the visible range in Figure 2). Errors drop every year to $\sim 1.5\text{–}2 \times 10^6\,\mathrm{km}^2$ not only when the forecasts pass through the initial time of the year in early July, but also around November when the ice cover undergoes a similar spatial distribution in the course of the freezing period.

The reference forecasts reduced to the climatological median ice edge (CMED) score worse than the probabilistic climatological forecasts (CLIM), with the SPS offset by $\sim 0.2\text{–}0.4 \times 10^6\,\mathrm{km}^2$. This supports the notion that the SPS is a meaningful probabilistic verification metric in the sense that it rewards the provision of reliable forecast uncertainties.

The SPS of the full ensemble-based forecasts (FULL; solid black curve in Figure 2) grows much slower than the persistence forecast error, reaching values of about $0.6 \times 10^6\,\mathrm{km}^2$ in September. The leveling-off at that lead time does not imply the loss of potential predictability, given that the climatological error (CLIM) is still significantly higher with values close to $1.0 \times 10^6\,\mathrm{km}^2$. Rather, a more continuous loss of potential predictability is counteracted by the seasonal cycle of the (climatological) error.

The full ensembles remain skilful compared to climatology until ~10 months into the forecasts (compare solid black and grey curves in Figure 2). Thereafter the climatological forecasts mostly beat the eight-member ensembles, which is probably due to the limited ensemble size (see next section). An exception are the months around the maximum Arctic sea-ice extent in March, where the eight-member ensembles remain marginally skilful throughout the 3-year forecast lead time. This annual 're-emergence' of skill probably occurs because the Arctic sea-ice edge extends into parts of the northern North Atlantic where pronounced decadal modes of SST variability influence the ice-edge position (e.g. G16, Rackow *et al.*, 2016).

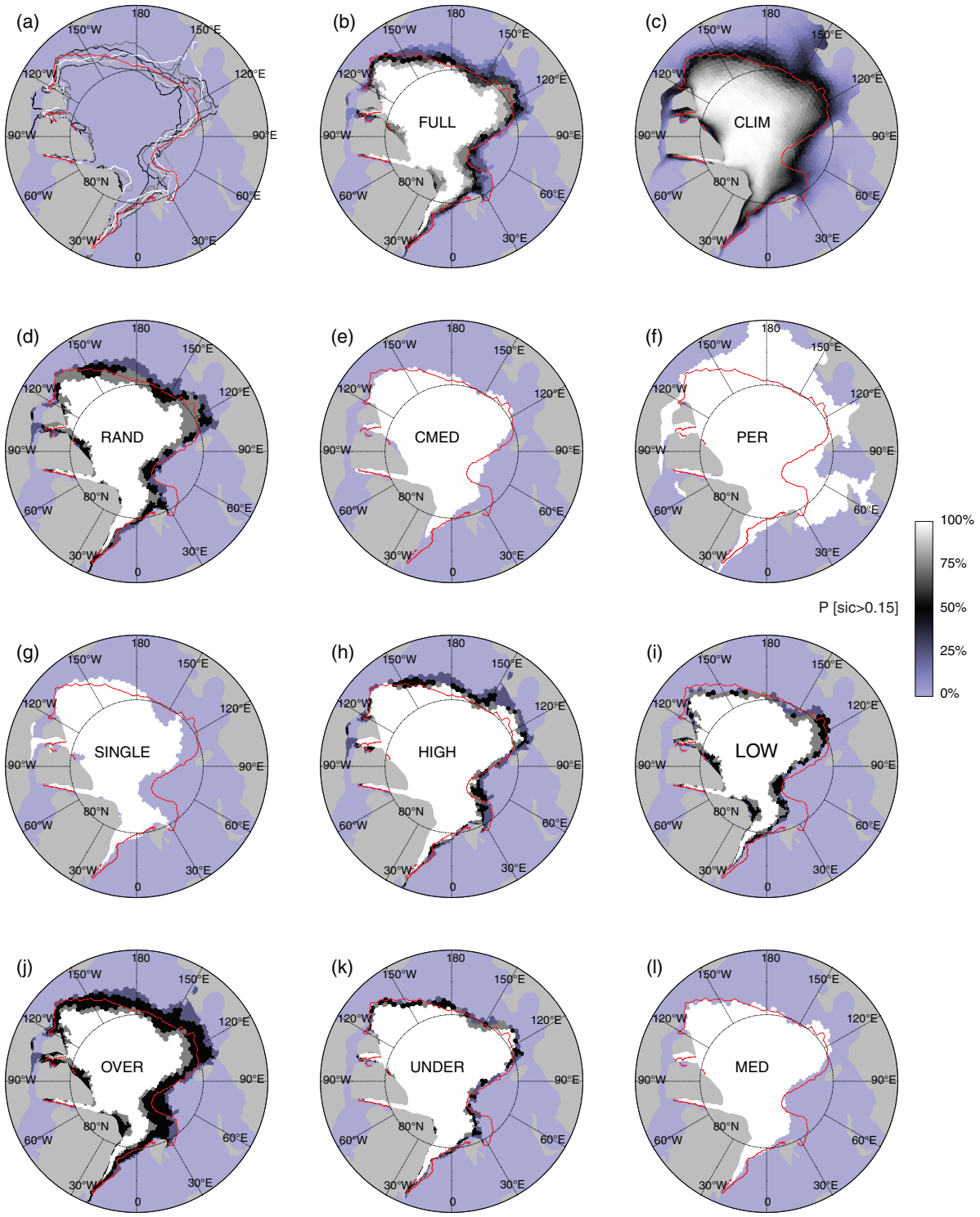

FIGURE 1 (a) Ice-edge locations (15% sea-ice concentration contours) on 14 September in a nine-member 'perfect-model' ensemble initialized on 1 July of the same arbitrary year; (b)–(l) The eleven ice-edge forecast types listed in Table 1, most of which are based on the ensemble shown in (a) except for one member which was randomly chosen as 'truth' (red curve in all panels). (c) and (e) showing CLIM and CMED, respectively, are based on the 200 year control run. The colour scale in (b)–(l) denotes the so-called sea-ice probability, i.e. the forecast probability of being enclosed by the ice edge. The land–sea geometry and the polygonal structure correspond to the ocean mesh employed in the AWI-CM

3.3.2 | Ensemble size

Comparing the full eight-member ensembles (FULL) with randomly drawn four-member ensembles (RAND; Figure 1d) reveals that the decreased ensemble size inflates the SPS by a fairly constant factor of ∼1.11 (compare solid and dashed black curves in Figure 2). Reducing the ensemble size more extremely to single members (SINGLE; Figure 1g and dotted black curve in Figure 2) results in 'deterministic' forecasts. The corresponding SPS is increased by a much larger, again fairly constant, factor of ∼1.78 relative to the full ensembles.

Measured in terms of the SPS, the single-member forecasts (SINGLE) are outperformed by climatology (CLIM)

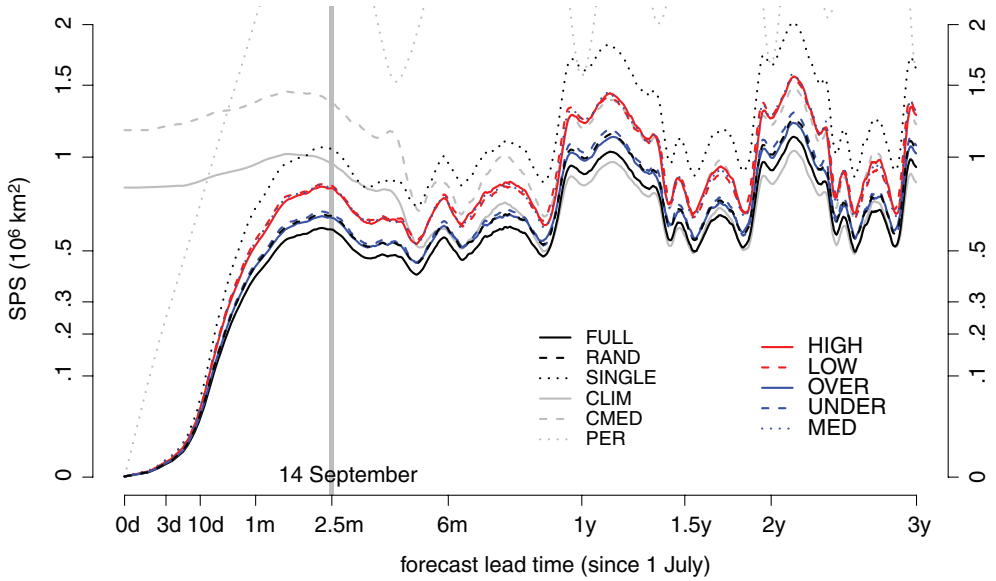

FIGURE 2 The Spatial Probability Score (SPS) for all ice-edge forecast types described in Table 1 as a function of forecast lead time, averaged over all 18 start dates with each ensemble member selected once as 'truth'. The grey vertical bar denotes 14 September (2.5 months into the forecasts), corresponding to the situation depicted in Figure 1. Both axes are plotted on a square-root scale to emphasise short lead times and associated smaller error values. Note that the dotted blue curve (MED) coincides with the red curves (HIGH and LOW), and that the persistence forecast (PER; dotted grey curve) mostly attains values outside the plot range after about 1 month into the forecasts

already after ∼2 months, and by the climatological median ice edge (CMED) after ∼5 months (Figure 2). At 1, 2, and 3 years lead time, that is, around 1 July, the single-member forecasts are exactly as skilful as the persistence forecasts. This is consistent with the paradigm that the predictability of Arctic sea ice arises mostly from simple first-order memory of the system (carried by the sea ice itself and by the ocean) rather than from higher-order dynamics which could cause quasi-oscillatory behaviour as known, for example, from the El Niño Southern Oscillation (ENSO).

The increase of skill with increasing ensemble size complies with the necessity implied by the 'perfect model' set-up that the average SPS of the forecast ensembles would eventually not exceed the SPS of the climatological forecasts at any time (as is the case for the eight-member ensembles outside March beyond 11 months into the forecasts), but asymptotically approach the climatological error level. Our results are consistent with earlier findings for the effects of ensemble size on related probabilistic scores including the CRPS (Ferro *et al.*, 2008).

3.3.3 | Spread

To mimic probabilistic forecasts which are based on overdispersive or underdispersive ensembles, we manipulate the sea-ice probabilities of the unbiased eight-member ensembles P_{FULL} as follows:

$$P_{\mathrm{OVER}} = \begin{cases} 2P_{\mathrm{FULL}} & P_{\mathrm{FULL}} \le 0.25, \\ 0.5 & 0.25 < P_{\mathrm{FULL}} < 0.75, \\ 2P_{\mathrm{FULL}} - 1 & P_{\mathrm{FULL}} \ge 0.75, \end{cases} \quad (3)$$

and

$$P_{\mathrm{UNDER}} = \begin{cases} 0 & P_{\mathrm{FULL}} \le 0.25, \\ 2P_{\mathrm{FULL}} - 0.5 & 0.25 < P_{\mathrm{FULL}} < 0.75, \\ 1 & P_{\mathrm{FULL}} \ge 0.75\,. \end{cases} \quad (4)$$

The resulting spatial probability fields (OVER and UNDER; Figure 1 j,k) are slightly unrealistic, in particular in the overdispersive case with a large area occupied by $P_{\mathrm{OVER}} = 0.5$. However, the simple manipulations appear suitable for a first-order test of how the SPS responds to unreliable forecast probabilities.

To supplement the forecast types OVER and UNDER, we consider forecasts defined by the ensemble median ice edge (MED; Figure 1l). This forecast type, like CMED, PER, and SINGLE, can be considered a 'deterministic' forecast for which the SPS is reduced to the IIEE. In terms of reliability, the ensemble median ice edge can be interpreted as the most extreme case of an underdispersive (but unbiased) ensemble-based forecast. Note that for MED we have spatially smoothed P_{FULL} slightly using a Gaussian filter with a length-scale of ∼40 km to render the $P_{\mathrm{FULL}} = 0.5$ contours well-defined (which is not the case for the raw probabilities of the evenly-sized (and small) eight-member ensembles).

The forecast types OVER and UNDER score worse than the full ensembles, with the SPS increased by similar fairly constant factors of ∼1.11 and ∼1.13, respectively (solid and dashed blue curves in Figure 2). This demonstrates that the metric penalises unreliability in a similar way in both directions. The more extremely underdispersive ensemble-median forecasts (MED; dotted blue curve in Figure 2) yield an SPS increased by a factor of ∼1.36 compared to the full ensembles.

Measured in terms of SPS, the median ice-edge forecasts (MED) outperform the single forecasts (SINGLE) by a factor of ∼0.76 . Moreover, the mean SPS of MED relates to the error of the climatological median ice-edge forecasts (CMED) in a way that mirrors the relation between the full ensemble forecasts (FULL) and the unreduced climatological forecasts (CLIM). The mean error of MED surpasses the error of CMED after ∼11 months and drops again slightly below only around March of the subsequent years.

3.3.4 | Bias

To mimic probabilistic forecasts that are based on ensembles biased towards higher or lower sea-ice extent, we manipulate the sea-ice probabilities of the unbiased eight-member ensembles P_{FULL} as follows:

$$P_{\mathrm{HIGH}} = \begin{cases} 2P_{\mathrm{FULL}} & P_{\mathrm{FULL}} < 0.5, \\ 1 & P_{\mathrm{FULL}} \geq 0.5, \end{cases} \tag{5}$$

and

$$P_{\mathrm{LOW}} = \begin{cases} 0 & P_{\mathrm{FULL}} < 0.5, \\ 2P_{\mathrm{FULL}} - 1 & P_{\mathrm{FULL}} \geq 0.5\,. \end{cases} \tag{6}$$

Defined this way, P_{HIGH} (P_{LOW}) corresponds to an inflation of the range $[0, 0.5]$ ($[0.5, 1]$) of the full ensemble probabilities to the full probability range $[0, 1]$ (Figure 1h,i). This implies that the introduced bias develops only with the development of spread of the full ensembles, with no bias at initial time. Moreover, the spread of these forecasts is reduced compared to the full-ensemble forecasts, which one would not necessarily expect for actual biased forecast ensembles.

The mean SPS of the biased forecast types (HIGH and LOW; solid and dashed red curves in Figure 2) is increased by the same, relatively constant factor ∼1.36 . This increase could partly be due to the simultaneously reduced spread of the biased forecasts. An estimate for the contribution from the reduced spread can be inferred by comparison with the underdispersive forecasts (UNDER) as these feature a similar, if not slightly stronger, spread reduction (compare spread in (h) and (i) versus (k) in Figure 1). For the underdispersive forecasts, the SPS is increased by the factor ∼1.13, leaving at least the factor ∼1.20 caused by the biases in HIGH and LOW. This attests that the SPS penalises biases.

4 | DISCUSSION

In the previous section we have demonstrated some appealing properties of the Spatial Probability Score (SPS) when applied to sea-ice edge forecasts. However, we anticipate that the SPS will be useful also when applied to a variety of other contours. Whether or not this metric can be considered user-relevant depends strongly on the choice of the contour. Beside the sea-ice edge, possible user-relevant applications of the SPS include the verification of probabilistic forecasts of extreme events, considering for example the probability that precipitation or wind speed exceeds a certain (extreme) threshold of particular interest to one or more user groups. However, the application to rare events may necessitate considerably larger ensembles than used here, and/or spatial dressing (e.g. Gaussian smoothing) of the raw forecast probabilities, to generate sufficiently continuous fields. We also hypothesize that appropriate spatial dressing applied to the relatively small ensembles used in section 3 would reduce the SPS compared to the raw ensemble probabilities, which would reveal the adequacy of such a posteriori adjustments. In a similar way, the perfect-model approach set forth here could be used to exploit whether the SPS (or any other verification metric) is suitable for different predictands.

When the sea-ice edge is considered, a meaningful variant of the SPS can be obtained through division by the total length of the contour(s). The resulting normalized SPS corresponds to an effective mean distance between the forecast and the observed ice edges, where *effective* implies that the correspondence holds exactly only in the limit of straight ice edges. Such a normalization has the advantage that the metric becomes even more user-relevant because it provides a directly interpretable displacement distance. However, this advantage comes at the price of (a) the non-trivial task of determining the length of contours and (b) an ambiguity with respect to which contour(s) should be used for the normalization. In case of the ice edge, which typically varies within a zone of uncertainty separating a central region with (close to) 100% climatological sea-ice probability and an outer region with 0% probability, it appears reasonable to normalize by the length of the climatological median ice edge (the 50% sea-ice probability contour) as it is smoother and thus closer to the limit of straight ice edges; the resulting normalized SPS could be compared with other metrics which aim to quantify an effective mean distance of contours, such as the Modified Hausdorff Distance (e.g. Dukhovskoy *et al.*, 2015). However, for other contours, for example associated with extreme events, a median contour may not exist at all because climatological probabilities do not cross the 50% level anywhere; in such cases the suggested normalization of the SPS does not appear meaningful.

Moreover, one can think of different ways to decompose the SPS in order to retrieve additional information on probabilistic forecast quality. One possibility is to decompose the SPS into reliability, resolution, and uncertainty components, by analogy with the well-known decompositions of the Brier Score (Murphy, 1973) and the CRPS (Hersbach, 2000). In the simplest case, one would just spatially integrate the corresponding local components of the Brier Score. This type of decomposition would consider large numbers of individual forecasts simultaneously and relate them to climatological forecast probabilities. An alternative approach considers individual forecasts only; the decomposition of the IIEE into Mean Extent Error and Misplacement Error components for the ice-edge location, as set forth in G16, can be adopted. For

the SPS, determining the former two components based on the forecast median ice edge would leave an additional sharpness component that reflects the spatial uncertainty associated with an individual forecast (compare, e.g. Gneiting *et al.*, 2007). These or other types of decomposition of the SPS merit additional scrutiny.

Regarding the interpretation of the SPS as the spatial analogue of the CRPS, in section 3 we provide a one-dimensional example where the SPS is equal to the CRPS of the distance of the ice edge from a certain reference point. From a mathematical point of view, considering the SPS as the spatial analogue of the CRPS appears generally indisputable. However, the replacement of the cumulative distribution function used in the CRPS with a general spatial probability field is associated with certain conceptual changes. In contrast to the above example, the spatial field associated with the probability of being 'enclosed' by a contour may not cover the whole interval $[0\%, 100\%]$ (e.g. for extreme-event probabilities as mentioned above) and probabilities may vary along transects non-monotonically. In such cases the probabilities along transects cannot be interpreted as cumulative distribution functions and the analogy is less stringent.

Furthermore, one may wonder how the SPS differs from alternative ways to aggregate probabilistic forecast skill spatially. For example, one could spatially integrate or average the local CRPS of a continuous quantity (instead of spatially integrating the local Brier Score of a binary version of that quantity based on a fixed threshold, as done for the SPS). The spatially integrated or averaged CRPS would reflect different aspects of a forecast; it would penalise forecast error also in regions where the continuous quantity is consistently away from the threshold used for the SPS. This might be favourable in some situations, but not in others; in the case of sea-ice concentration, for example, one may not be interested in small deviations in the central ice pack which would receive a large weight given the large associated area, but rather in errors along the marginal ice zone. Moreover, if the CRPS is integrated over two spatial dimensions, the resulting unit is $u \times$ area, where u is the unit of the continuous quantity. (Note that sea-ice concentration would be a special case here as it has no unit.) Such a unit is not easy to interpret, but this could be overcome by averaging instead of integrating, giving simply the unit u (as for the CRPS). However, averaging would exacerbate the problem that large regions with low errors could dominate the metric; in case of sea ice, the central ice pack as well as open-ocean areas off the marginal ice zone with trivial predictability could easily dominate. One could try to circumvent this issue by defining meaningful subdomains over which the metric is computed, such as the marginal ice zone, but this would add another level of complexity and ambiguity. In conclusion, it appears that the SPS is a more elegant way of aggregating probabilistic forecast skill spatially, in particular if a threshold of specific interest exists.

5 | SUMMARY AND CONCLUSIONS

We have introduced a probabilistic verification score for ensemble-based forecasts of contours. The Spatial Probability Score (SPS) is defined as the spatial integral of local (Half) Brier Scores. Applying the SPS to idealised ensemble forecasts of the Arctic sea-ice edge with a global coupled climate model, we have demonstrated that the SPS increases in response to dedicated attempts to degrade the original forecasts, namely by decreasing ensemble size, by overdispersion or underdispersion, and by bias. We conclude that the SPS is a meaningful verification score which penalises typical kinds of probabilistic forecast deficiencies.

Moreover, we have argued that the SPS can be applied to other contours and discussed how its interpretation may depend on the properties of the associated probability field. We have delineated how the SPS can be normalized and/or decomposed to reveal additional information on probabilistic forecast quality, and pointed out limits to the interpretation of the SPS as the spatial analogue of the CRPS.

Given that the SPS reflects forecast skill in a meaningful way, we can draw some specific conclusions with respect to Arctic sea-ice edge prediction. The eight-member ensembles remain skilful compared to climatology until ~10 months into the forecasts, with a 're-emergence' of marginal skill in late winter of subsequent years presumably due to slow modes of variability in the North Atlantic. This confirms earlier findings on the limits of predictability for the Arctic ice edge (G16) from a probabilistic viewpoint. Furthermore, in our set-up, persistence forecasts remain more skilful than forecasts based on the climatological median ice edge out to day ~19, suggesting that a similar time-scale should be used to merge persistence and climatology in atmospheric forecast systems without interactive sea ice.

Finally, we suggest that the SPS could serve as a headline verification score to document progress in sea-ice forecasting, in particular in the context of current efforts associated with the Year of Polar Prediction (Jung *et al.*, 2016) and beyond. Resources like the recently implemented Subseasonal-to-seasonal database (Vitart *et al.*, 2017), where several operational forecast systems with dynamical sea-ice model components contribute hindcasts and forecasts, offer an unprecedented opportunity to document and advance our ability to forecast sea ice, on the basis of meaningful verification techniques. We expect that the SPS will prove to be a useful probabilistic verification metric, for ice-edge forecasts as well as for contour forecasts in general.

ACKNOWLEDGEMENTS

The simulations were performed with resources provided by the North German Supercomputing Alliance (HLRN). The model data are openly available from the British Atmospheric Data Centre: http://dx.doi.org/10.5285/

45814db8-56cd-44f2-b3a4-92e41eaaff3f. Helge Goessling acknowledges the financial support by the Federal Ministry of Education and Research of Germany in the framework of SSIP (grant 01LN1701A). Thomas Jung acknowledges the funding from the European Union's Horizon 2020 Research & Innovation programme through grant agreement No. 727862 APPLICATE. We thank Barbara Casati, Chris Ferro, Cecilia Bitz, and Stephan Juricke, with whom we have had very helpful discussions, as well as three anonymous reviewers who provided very thoughtful and constructive comments. We are also grateful to the Joint Working Group on Forecast Verification Research (JWGFVR) of WMO's World Weather Research Programme for rewarding our work within their 'Challenge to Develop and Demonstrate the Best New User-Oriented Forecast Verification Metric.'

ORCID

H. F. Goessling http://orcid.org/0000-0001-9018-1383
T. Jung http://orcid.org/0000-0002-2651-1293

REFERENCES

Bauer, P., Magnusson, L., Thépaut, J.-N. and Hamill, T.M. (2016) Aspects of ECMWF model performance in polar areas. *Quarterly Journal of the Royal Meteorological Society*, 142(695), 583–596. https://doi.org/10.1002/qj.2449.

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

Casati, B., Wilson, L., Stephenson, D., Nurmi, P., Ghelli, A., Pocernich, M., Damrath, U., Ebert, E., Brown, B. and Mason, S. (2008) Forecast verification: current status and future directions. *Meteorological Applications*, 15(1), 3–18.

Day, J.J., Tietsche, S., Collins, M., Goessling, H.F., Guemas, V., Guillory, A., Hurlin, W.J., Ishii, M., Keeley, S.P.E., Matei, D., Msadek, R., Sigmond, M., Tatebe, H. and Hawkins, E. (2016) The Arctic predictability and prediction on seasonal-to-interannual timescales (APPOSITE) data set version 1. *Geoscientific Model Development*, 9(6), 2255. https://doi.org/10.5194/gmd-9-2255-2016.

Dorninger, M., Gilleland, E., Casati, B., Mittermaier, M., Ebert, E., Brown, B. and Wilson, L. (2018) The set-up of the Mesoscale Verification Inter-Comparison over Complex Terrain (MesoVICT) project. *Bulletin of the American Meteorological Society*. https://doi.org/10.1175/BAMS-D-17-0164.1. in press.

Dukhovskoy, D.S., Ubnoske, J., Blanchard-Wrigglesworth, E., Hiester, H.R. and Proshutinsky, A. (2015) Skill metrics for evaluation and comparison of sea ice models. *Journal of Geophysical Research: Oceans*, 120(9), 5910–5931.

Emmerson, C. and Lahn, G. (2012) *Arctic Opening: Opportunity and Risk in the High North*. London: Lloyd's and Chatham House.

Ferro, C.A., Richardson, D.S. and Weigel, A.P. (2008) On the effect of ensemble size on the discrete and continuous ranked probability scores. *Meteorological Applications*, 15(1), 19–24. https://doi.org/10.1002/met.45.

Gilleland, E., Ahijevych, D., Brown, B.G., Casati, B. and Ebert, E.E. (2009) Intercomparison of spatial forecast verification methods. *Weather and Forecasting*, 24(5), 1416–1430.

Gneiting, T., Balabdaoui, F. and Raftery, A.E. (2007) Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology*, 69(2), 243–268. https://doi.org/10.1111/j.1467-9868.2007.00587.x.

Gneiting, T. and Raftery, A.E. (2005) Weather forecasting with ensemble methods. *Science*, 310(5746), 248–249.

Goessling, H.F., Tietsche, S., Day, J.J., Hawkins, E. and Jung, T. (2016) Predictability of the Arctic sea ice edge. *Geophysical Research Letters*, 43(4), 1642–1650. https://doi.org/10.1002/2015GL067232.

Hersbach, H. (2000) Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 15(5), 559–570. https://doi.org/10.1175/1520-0434(2000)015<0559:dotcrp>2.0.CO;2.

Jung, T., Gordon, N.D., Bauer, P., Bromwich, D.H., Chevallier, M., Day, J.J., Dawson, J., Doblas-Reyes, F., Fairall, C. and Goessling, H.F. (2016) Advancing polar prediction capabilities on daily to seasonal time scales. *Bulletin of the American Meteorological Society*, 97(9), 1631–1647. https://doi.org/10.1175/BAMS-D-14-00246.1.

Matheson, J. and Winkler, R. (1976) Scoring rules for continuous probability distributions. *Management Science*, 22(10), 1087–1096.

Mittermaier, M. and Roberts, N. (2010) Intercomparison of spatial forecast verification methods: identifying skilful spatial scales using the fractions skill score. *Weather and Forecasting*, 25(1), 343–354.

Murphy, A.H. (1973) A new vector partition of the probability score. *Journal of Applied Meteorology*, 12(4), 595–600.

Palmer, T.N. (2000) Predicting uncertainty in forecasts of weather and climate. *Reports on Progress in Physics*, 63(2), 71. https://doi.org/10.1088/0034-4885/63/2/201.

Rackow, T., Goessling, H.F., Jung, T., Sidorenko, D., Semmler, T., Barbi, D. and Handorf, D. (2016) Towards multi-resolution global climate modeling with ECHAM6-FESOM. Part II: climate variability. *Climate Dynamics*, 50(7–8), 2369–2394. https://doi.org/10.1007/s00382-016-3192-6.

Rickard, G.J., Lunnon, R.W. and Tenenbaum, J. (2001) The Met Office upper air winds: prediction and verification in the context of commercial aviation data. *Meteorological Applications*, 8(3), 351–360.

Saetra, Ø., Hersbach, H., Bidlot, J.R. and Richardson, D.S. (2004) Effects of observation errors on the statistics for ensemble spread and reliability. *Monthly Weather Review*, 132(6), 1487–1501. https://doi.org/10.1175/1520-0493(2004)132<1487:eooeot>2.0.CO;2.

Sidorenko, D., Rackow, T., Jung, T., Semmler, T., Barbi, D., Danilov, S., Dethloff, K., Dorn, W., Fieg, K., Gössling, H.F., Handorf, D., Harig, S., Hiller, W., Juricke, S., Losch, M., Schröter, J., Sein, D.V. and Wang, Q. (2015) Towards multi-resolution global climate modeling with ECHAM6–FESOM. Part I: model formulation and mean climate. *Climate Dynamics*, 44(3-4), 757–780. https://doi.org/10.1007/s00382-014-2290-6.

Smith, L.C. and Stephenson, S.R. (2013) New Trans-Arctic shipping routes navigable by midcentury. *Proceedings of the National Academy of Sciences of United States of America*, 110(13), 4871–4872.

Tietsche, S., Day, J.J., Guemas, V., Hurlin, W.J., Keeley, S.P.E., Matei, D., Msadek, R., Collins, M. and Hawkins, E. (2014) Seasonal to interannual Arctic sea ice predictability in current global climate models. *Geophysical Research Letters*, 41(3), 1035–1043. https://doi.org/10.1002/2013GL058755.

Venugopal, V., Basu, S. and Foufoula-Georgiou, E. (2005) A new metric for comparing precipitation patterns with an application to ensemble forecasts. *Journal of Geophysical Research: Atmospheres*, 110(D8). https://doi.org/10.1029/2004JD005395.

Vitart, F., Ardilouze, C., Bonet, A., Brookshaw, A., Chen, M., Codorean, C., Déqué, M., Ferranti, L., Fucile, E. and Fuentes, M. (2017) The Subseasonal to Seasonal (S2S) Prediction Project Database. *Bulletin of the American Meteorological Society*, 98(1), 163–173. https://doi.org/10.1175/BAMS-D-16-0017.1.

Wilks, D. (2011) *Statistical Methods in the Atmospheric Sciences*, Vol. 100. Cambridge, MA: Academic Press.

How to cite this article: Goessling HF, Jung T. A probabilistic verification score for contours: Methodology and application to Arctic ice-edge forecasts. *Q J R Meteorol Soc.* 2018;144:735–743. https://doi.org/10.1002/qj.3242