Marine Pollution Bulletin

Oil spill forecast assessment using Fractions Skill Score

--Manuscript Draft--

Manuscript Number:	MPB-D-20-02060R1
Article Type:	Research Paper
Keywords:	Oil spill, Oil spill model, Model evaluation, Verification, Forecast, Spatial displacement
Corresponding Author:	Debra Simecek-Beatty National Ocean Service Seattle, WA UNITED STATES
First Author:	Debra Simecek-Beatty
Order of Authors:	Debra Simecek-Beatty
	William J Lehr, Phd
Abstract:	In the event of an oil spill, emergency responders must quickly deploy cleanup and protection equipment using guidance provided by a forecast trajectory. Real-time modeling the location of the surface oil over time is standard practice; however, current performance metrics used for assessing the quality of the spill forecast lack consideration of spatial resolution. We show the Fractions Skill Score identifies the scale at which the oil spill forecast demonstrates useful skill. As spatial verification methods are new to oil spill modeling, we found that useful spatial skills for a set of tactile forecasts are consistent with those found in precipitation forecasts.

Highlights

Quantitative performance metrics are evolving for oil spill trajectory forecasts.

Spatial verification methods are new to oil spill forecasting.

Deepwater Horizon spill forecasts are evaluated using remote sensing observations.

Fractions Skill Score provides horizontal scale appropriate for presenting forecasts.

Oil spill forecast assessment using Fractions Skill Score

Debra Simecek-Beatty and William J. Lehr

Debra Simecek-Beatty

NOAA, National Ocean Service, Seattle, Washington, USA

E-mail: debra.simecek-beatty@noaa.gov

William J. Lehr

NOAA, National Ocean Service, Seattle, Washington, USA

Key words: oil spill modelling, verification, fractions skill score, skill scores,

Abstract

In the event of an oil spill, emergency responders must quickly deploy cleanup and protection equipment using guidance provided by a forecast trajectory. Forecasting the location of the surface oil over time is standard practice; however, current performance metrics used for assessing the quality of the spill forecast lack both an appropriate numerical model accuracy score and specification of the expected spatial resolution limit for useful forecast information. This paper adapts the Fractions Skill Score method, commonly used in weather forecasting, to oil forecasting. A subset of satellite images and trajectory forecasts from the Deepwater Horizon oil spill are used as an example of the method.

1. Introduction

Breakthroughs in applied research can consist of developing new techniques designed for a specific field or, conversely, applying techniques developed for other applications to solve challenges in the new field (e.g. Sarrute & Burroni, 2008; Malis, 2004). This paper does the latter. Forecasting the movement and potential landfall of spilled oil is critical to efficient emergency response by providing risk estimates for threatened resources and identifying best locations for cleanup teams. Computer technology has advanced such that spill transport models are capable of extremely high resolution in their forecasts of surface oil distribution, often exceeding the resolution of either the environmental input or the oil observation data. Thus, while increasing model resolution may improve the spill forecast (Janeiro et al., 2014; Pisano, et al., 2016), this outcome is not guaranteed (De Dominicis, et al., 2016).

Weather forecasters are well aware of this fact when doing meteorological predictions. For example, Mittermaier & Csima, (2017) indicate that, for numerical weather estimates, high-resolution models do not necessarily increase accuracy, as errors at small scales may increase due to unmeasured environmental fluctuations not being included. A similar circumstance is likely for oil transport forecasts that often depend critically on local wind and unresolved oceanographic features.

The purpose of this paper is to adapt numerical skill assessments that have proven effective for meteorological predictions to spatial predictions of spilled oil. While assigning a numerical accuracy value to a forecast may seem to be an obvious requirement, traditional oil trajectory models usually do not include this parameter. Instead, forecasts are often qualitatively assessed to their accuracy (Cheng et al., 2011; Cheng, et al., 2014; Le Hénaff, et al., 2012; Özgökmen, et al., 2016; Pisano, et al., 2016). Quantitative metrics primarily involve comparison of the forecast slick area with spill observation area using raw values of 'percent observation in the forecast' and 'percent forecast in the observation', (Huntley, Lipphardt Jr., & Kirwan Jr., 2011; Cheng et al., 2011; Kim et al., 2014; Cheng, et al., 2014; Guo et al., 2018).

By themselves, percent area metrics are of limited value. Consider the following trivial, but illustrative, example. Assume the following two spatial grid systems used to forecast the oil, letting 'O' be an observed patch of surface oil and 'M', an estimated model-forecast location (Figure 1). While the actual oil location and forecasted location are the same in Fig.

1, the finer resolution grid, Fig. 1(a), indicates a miss as the forecasted and oiled grids do not overlap. The coarser grid, Fig. 1(b), shows a 'hit'. Thus resolving appropriate grid scale is an important factor in determining forecast skill.

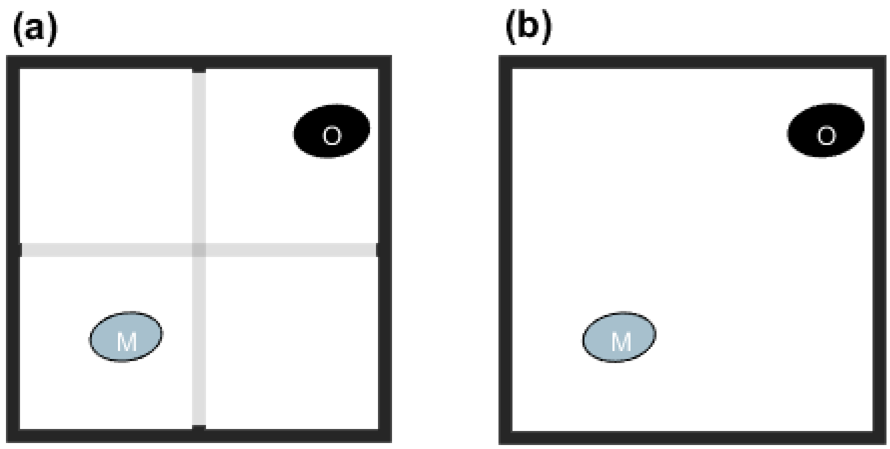

Figure 1. Model-forecast, 'M' is shaded gray and the observed oil, 'O', shaded black with (a) 5 km grid resolution and (b) 10 km grid resolution.

Another important factor to consider is the potential discrepancy between observational area and model-forecast area. The former is usually much larger, meaning that the number of non-oil grid boxes greatly exceeds the number of oiled boxes. Similar concerns are present in weather prediction, the so-called 'rare event' prediction. Consider two forecasts where neither predict the exact oil location but the first forecast misses by a kilometer while the second misses by 10 kilometers.

Obviously, the first forecast is better but both might score the same by the raw metrics mentioned above.

A final consideration in developing a skill metric involves understanding the planned application of the forecast. For any common grid between forecast and observation, Table 1 shows four possible outcomes for any individual grid box; (a) model and observation may agree on oil being present- a hit, (b) model predicts oil but none found- false alarm, (c) oil present but not predicted – a miss, and (d) oil not predicted or observed- a correct rejection. If one is drilling for oil where the cost of the drilling is expensive relative to the value of the potential oil find, then one wants to minimize (b), the false positives, even if this means missing some oil (c). Oil spill response, however, operates under a different standard. Generally, responders adopt a minimum regret strategy (Galt, 1998) to identify all oil possible and minimize misses, (c), even at the expense of increasing the number of false positives, (b).

Table 1. Contingency table to evaluate oil spill forecasts, modified from Jolliffe & Stephenson (2012).

	Oil observation	
Oil Model-forecast	Yes	No
Yes	a (Hit)	b (False alarm)
No	c (Miss)	d (Correct rejection)

There is an important caveat that the reader should be aware when matching spill forecasts to spill observations. On the one hand, the modeler, predicting oil mass or volume distribution, has to carefully simulate very convoluted environmental and oil behavior processes that easily produce oil patches in the same spill that may vary spatially in thickness by orders of magnitude (Spaulding, 2017). On the other hand, visual observation, and even the more sophisticated oil slick remote sensing capabilities, typically show much better accuracy in determining slick surface area than they do in estimating the more useful surface oil volume (Fingas, 2018). Recognizing the greater accuracy in area observation, this paper only looks at comparison of surface area prediction by the models versus the satellite observation of the spill area while recognizing that this situation may change due to new studies.

Researchers are examining non-electromagnetic methods to measure thickness, particularly where there may be interference in direct observation, by using subsurface, upward looking, sonar (Basset et al, 2016), but thus far these remain more experimental than operational. Other researchers are employing alternatives to the more standard radar, visual and near IR frequencies common on many sensor packages (Fingas and Brown, 2018). One older method (Skou, 1986) that is regaining some popularity is passive microwave radiometry that uses the relatively large difference between oil and water emissivity in this band combined with multiple nearby frequencies to estimate oil thickness. However, operational challenges remain including onsite calibration by other means. Similarly, some success has been shown by processing hyperspectral images through advanced neural networks (Yingcheng et al., 2013) but these too require calibration, often site-specific, of the network. Thus, robust, comprehensive and accurate surface, oil volume determination remains to be achieved.

Fortunately, separating the thicker, recoverable, oiled area, of unknown depth but usually containing the preponderance of the surface oil volume (this paper does not consider oil mixed in the water column), from the much thinner sheen is often sufficient for the response. Most widely used spill models, as well as many remote sensing platforms, have this capability. While not applied to the example, the techniques described in this manuscript, by mapping only the thick area, could approximately compare the relative accuracy of the forecast to the observation, even if absolute volume numbers are unknown. One warning to consider is that for, some specific oil products, neglecting sheen volume might not be appropriate.

Lehr et al. (2019) compared oil spill forecasts with satellite observations by overlaying both onto a common grid. They applied categorical skill scores developed for weather forecast verification (Wilks, 2011; Jolliffe & Stephenson, 2012; WWRP/WGNE, 2017) to quantitatively evaluate forecast performance. The study, which involved a small subset of forecasts from an actual spill incident, suggested as good choices the Pierce Skill Score (Peirce, 1884; Jolliffe and Stephenson, 2012) or PSS, and the more modern metric for rare events, Symmetrical Extremal Dependence Index or SEDI (Ferro & Stephenson, 2011). These two metrics have the advantage of considering ‘correct rejections’. However, the drawback of such quantitative methods, as presented in that study, is the performance metrics were dependent on the resolution of the common grid. Alternatively, forecast skill can be evaluated over different spatial skills using fuzzy neighborhood techniques found in the literature, particularly for evaluating precipitation forecasts (Ebert E. , 2008; Ebert E. , 2009). Roberts & Lean (2008) introduced the Fractions Skill Score or FSS, to assess the variation of skill with the spatial scale of either single or aggregated rainfall accumulation forecasts. This approach is different from the previously discussed performance metrics in that an exact match between the forecast and observation, while preferred, is not necessary. This flexibility permits a certain amount of uncertainty in the observation location as well as the forecast. The FSS is potentially useful for oil spill verification by avoiding the double penalty problem associated with other metrics. Hence, one of the advantages of the proposed metric is a complementary strategy for identifying the scale at which the oil spill forecast is most useful.

Section 2 describes the example dataset containing forecasts and satellite observations from an actual spill incident. This section also presents the methodology to derive oiling probabilities for calculating the FSS and the measures to evaluate the forecast quality. Section 3 shows the results of the FSS analysis and discusses the horizontal scale appropriate for presenting the forecast. Section 4 contains the conclusions and suggestions for further work.

2.0 Method

2.1 Example spill forecast and observation data

On April 20, 2010 at 07:45 am local time, an explosion on the *Deepwater Horizon* platform released oil into the Gulf of Mexico for 87 days spilling 4.9 million barrels (USCG, 2011). The incident occurred approximately 65 km offshore over the outer continental slope (Figure 2). Surface oil covered large areas of the eastern Gulf of Mexico in a region well known for complex ocean circulation. Near the well blowout, buoyancy effects from the Mississippi and Atchafalaya River systems and deep ocean circulation influenced the surface circulation (MacFadyen et al., 2011). The dominating deep ocean circulation features in the Gulf of Mexico are the Loop Current (Oey et al., 2005) and the shedding of eddies (Xu et al., 2013). In May 2010, the spill response community was alarmed that deep-water ocean circulation would transport surface oil through the Florida Straits (Liu et al., 2011). As detailed in Liu et al., the shedding of an eddy from the Loop Current prevented the main surface slick from moving further south.

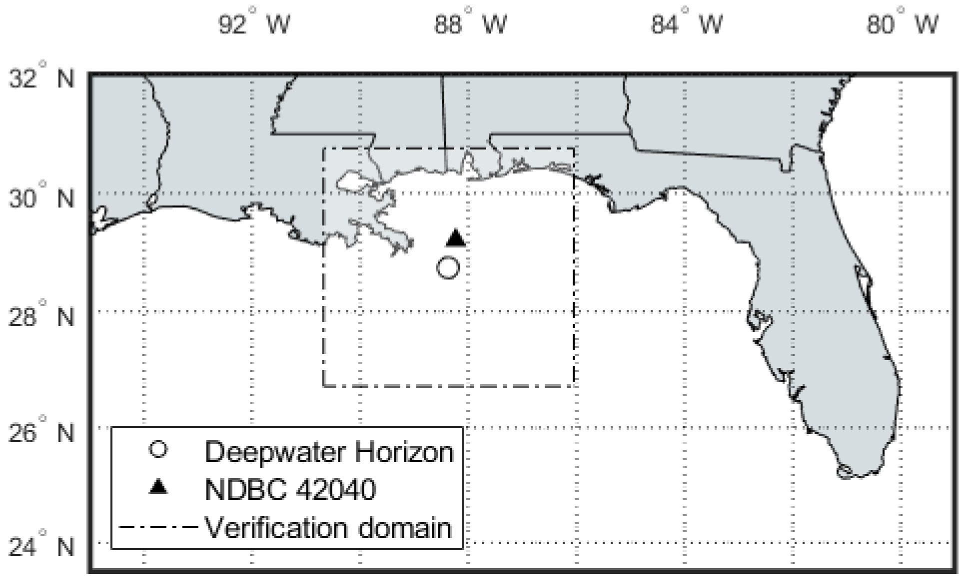

Figure 2. Map showing the *Deepwater Horizon* well site, ‘o’, the National Data Buoy Center (NDBC) Buoy 42040,’▲’ and boundary of the verification domain,’- ‘.

After the spill, the National Oceanic and Atmospheric Administration (NOAA) assembled a collection of oil forecasts and remote sensing products generated during the incident (Deepwater Horizon Natural Resource Damage Assessment Trustees, 2016). From this dataset, we assess a small subset of forecasts and satellite observations examined in detail by Lehr et al., (2019). In their study, the dataset included the Experimental Marine Pollution Surveillance Reports (EMPSR) provided by NOAA’s National Environmental Satellite Data and Information Services (Street, 2011) and the oil trajectory forecasts provided by NOAA’s National Ocean Service (MacFadyen et al., 2011). Figure 2 shows the boundary of the verification domain used in the analysis. Both the EMPSR analysis and the forecasts considered areas that potentially, but not necessarily, contained some oil. This suggests the bounded areas for the observation and for the forecast may contain both oil and non-oiled water.

Lehr et al (2019) determined in the timeframe, 5 May 2010 to 8 May 2010, the surface winds were amenable for oil slick detection with the average local wind speed of ~4 m/s (Figure 3). During this time, the spill release rate was relatively constant with minimal spill mitigation measures in place, including sprayed and injected chemical dispersants that would reduce impact on surface expression of the oil.

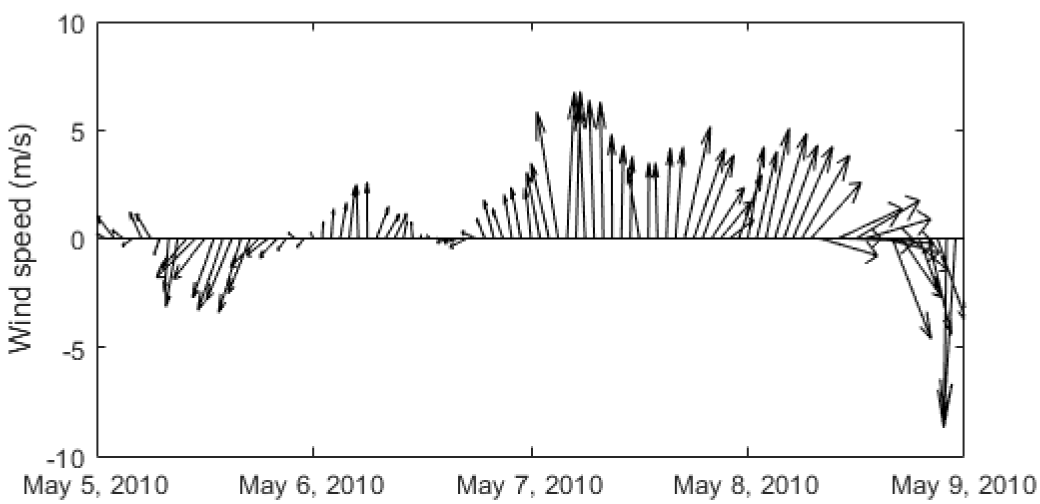

Figure 3. Wind observations for Buoy 42040 (NDBC, 1971).

Hydrodynamic and wind forecast models, from a variety of sources, are typically used in operational forecasting and this was the case during the *Deepwater Horizon*. However, this complicates forecast evaluation in general as depending on a particular spill event different models may be used. For this reason, we have intentionally setup the study using a series of operational forecasts rather than the performance of a particular oil spill model. In this example, the original forecasts and observations are not modified or corrected in any manner and, as originally released during the spill response; date and time are presented in Central Daylight Time (CDT) with the time offset from Coordinated Universal Time (UTC) –5:00.

Beginning on May 5, 2010, NOAA produced twice-daily forecasts of the expected oil location on May 8, 2010 for a total of six forecasts. Table 2 shows the forecasts and the length of time in hours between the issuance of the forecast, ’Prepared’, and the predicted oil location, ‘Estimate’, as the ‘Lead Time’.

Table 2. Description of the six forecasts. Date and time are Central Daylight Time (CDT). Lead-time is the hours between Forecast Prepared and Forecast Estimate.

	Forecast Prepared	Forecast Estimate	Lead Time (h)
1	5 May 2010 at 1300	8 May 2010 at 0600	65
2	5 May 2010 at 2000	8 May 2010 at 1800	70
3	6 May 2010 at 1300	8 May 2010 at 0600	41
4	6 May 2010 at 2000	8 May 2010 at 1800	46
5	7 May 2010 at 1300	8 May 2010 at 0600	17
6	7 May 2010 at 2100	8 May 2010 at 1800	21

All the satellite images (Table 3) selected for this work employed synthetic aperture radar (SAR) detection. SAR senses oil slicks by detecting the Maragoni effect of oil film to dampen the sea surface capillary waves. There is research to estimate oil thickness looking at radar polarization ratios (Garcia-Pineda et al., 2020). However, the images used in this paper only recorded oil surface area. TerraSAR-x and COSMO-Skymed used x-band radar (8–12 GHz) while the RADARSAT satellites used c-band (4-8 GHz).

Table 3. Experimental Marine Pollution Surveillance Reports (EMPSR) used as 'oil observation' for forecast evaluation. Date and time are Central Daylight Time (CDT).

	EMPSR Source	Image Acquisition
1	COSMO-Skymed2	8 May 2010 at 0657
2	RADARSAT-2	8 May 2010 at 0659
3	TerraSAR-X	8 May 2010 at 1823
4	COSMO-Skymed 2	8 May 2010 at 1851
5	RADARSAT -1	8 May 2010 at 1858

The forecasts did not exactly correspond with the image acquisition times on 8 May 2010. We estimated the movement of the surface slicks near the well blowout using simple vector addition of the components due to wind and currents (USCG, 1991). This approximation assumes the oil drifts with the surface current at 100% of the current speed and at 3% of the wind speed (Smith, 1976) providing a single, constant and plausible value for oil movement. Near the spill site, the nominal surface current velocity was about 0.2 m/s (Liu et al., 2011) and 3% of the average wind speed at NDBC Buoy 42040, ~0.1 m/s so the estimated oil transport is roughly 1 km over a one-hour period. Therefore, the EMPSR products are combined over a period of 1-h centered on the acquisition times coinciding with the morning and afternoon forecast of 0600 and 1800 (Table 2).

Combining the satellite images also helped reduce the problem of limited coverage of a particular satellite and provided a composite observation for the entire domain. For two satellites that passed within 2 minutes of each other, we combined areas presented in the EMPSR to represent the observation on 8 May 2010 at 0600 CDT. The remaining three images were within 35 minutes of each other to represent the observation on 8 May 2010 at 1800 CDT.

Unsurprisingly, the simple comparisons of satellite imagery with model forecast are disparate due to the complexity of the spatial distribution of the oil slick and the fundamental difference between oil volume and oil area, as discussed earlier. Figure 4 graphically demonstrates the complexity of the problem by overlaying the lead times of six forecast, 17, 21, 39, 46, 65 and 70 h matched to the oil observations. All satellite-detected oil is black; the forecast is shaded blue and forecast areas that overlap with the observation, dark blue. The forecast coverage of areas likely to contain oil was larger, ranging from approximately 12,000 to 18,000 km^2 with the area believed to be oil based on satellite observations, 6,000 –

9,000 km^2. Satellite observations indicated large oiled areas southeast of the release site, along with oiled areas to the east and west. East of the blowout, the 21-, 46- and 70-h forecasts (Fig. 4(b), (d) and (f)) predicted oil coverage but with no apparent corresponding observed oil in this area. Conversely, the 17-, 41- and 46-h forecasts (Fig. 4(a), (c) and (e)) under represent the observation in the same area. A similar situation occurs for the forecast to the west. Visual inspection shows predictions in these areas are not correct.

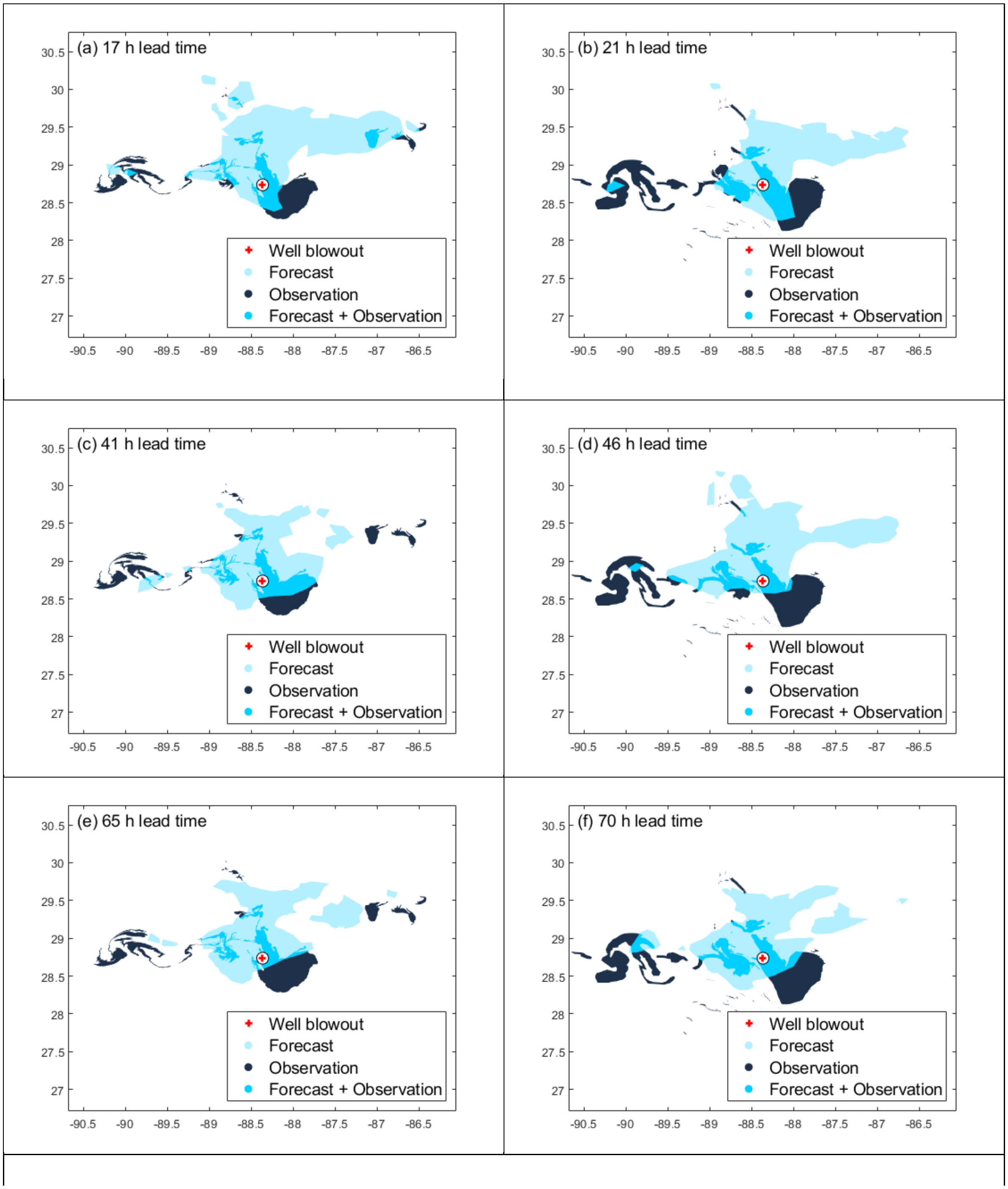

Figure 4. Forecast are shaded blue, the observed oil, black and the overlap of the observed oil and forecast, dark blue. For clarity, the coastlines are not plotted but for reference, the well blowout is marked '+'.

Pooling the experimental dataset into 17 – 21 h, 39-46 h and 65-70 h lead-time improved the coverage (Figure 5). The forecast area likely to contain oil and observational coverage of likely oil increased to approximately 17,000 km^2 and 20,000 km^2, respectively. Forecast coverage of the oil slicks east of the blowout for all lead-times increased significantly but under performed for oil to the west. As previously noted, the forecast is incorrect for the oil southeast of the well blowout.

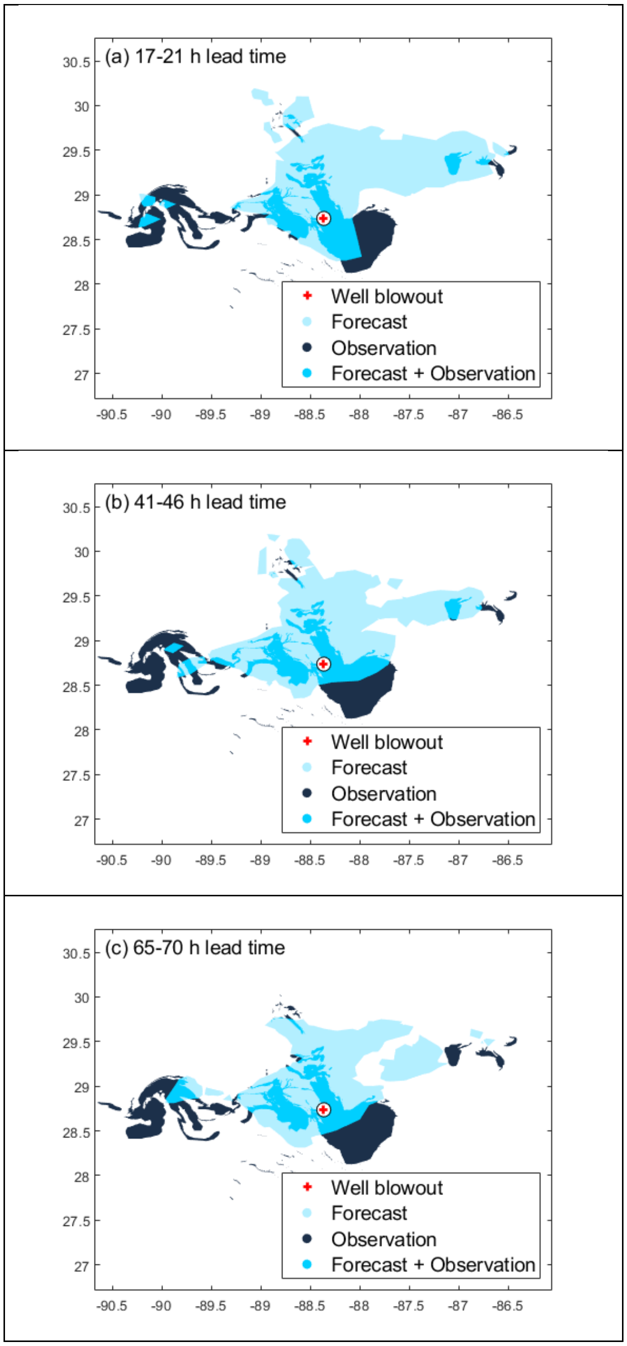

Figure 5. Aggregated forecast are shaded blue, the observed oil, black and the overlap of the observed oil and forecast, dark blue. The coastlines are not plotted but for reference, the well blowout is marked '+'.

Perhaps the most interesting aspect of Figures 4 and 5 is that forecasts consistently under represented the oil coverage southeast of the well blowout. Overtime, satellite imagery indicated oil in this particular area progressed to a long narrow band extending southeastward from the blowout towards the Loop Current. See Huntly et al (2011) and references therein for details regarding this particular feature.

2.2 Fractions Skill Score Method for Oil Spills

The authors recognize that most oil spill experts are unfamiliar with the fractional skill score (FSS) method while meteorologists will not necessarily be knowledgeable of the demands of spill forecasting. Therefore, this section explains FSS and the implementation requirements for application to oil trajectory forecasts.

While not a theoretical requirement, trajectory forecast results based on model simulations (not necessarily the same as the model internal grid) and spill observations used to initiate and validate the model are ideally applied using a common geospatial grid. For spills near the shore, these grids may be nested and restricted to account for shoreline affects. However, for large offshore spills such as the Deepwater Horizon Spill in the Gulf of Mexico (MacFadyen et al., 2011), the key question asked of the modelers is the time and location for significant oil impact to the nearshore and shoreline.

Remote sensing imagery for large offshore oil spills is primarily based on satellite sensors in the visible spectrum and x-band radar. These are useful frequencies for mapping surface oil spatial coverage but typically (see earlier discussion) do not provide information for oil volume coverage. As mentioned, this is a significant limitation as oil thickness can vary over three orders of magnitude and spill and trajectory models forecast oil volume or mass rather than surface area. Given a common grid, spill forecasters must develop guidelines that (1) determine whether a certain grid cell has sufficient oil considered a 'hit', Table 1, and (2) when the model predicts oil amount in the cell above a certain preset threshold. The first guideline can, at present, best be approximated by requiring the cell be more than a set fraction, by 'thick' oil, excluding sheen, to be considered as impacted by oil. Hopefully, this guideline will be improved with development in remote sensing surface volume estimation rather than surface area (Leifer, et al., 2012; Fingas, 2018; Garcia-Pineda et al., 2020). Since the SAR images used in this study did not provide separation of thick and sheen, the authors used oil area fraction in the grid as a guideline. The second guideline models the oil transport using Lagrangian Elements or LEs (Spaulding, 2017): parcels of oil that represent the continuous slick. Then, the model declares a grid 'hit' whenever the number of LEs in a particular grid exceed a set value.

Using the suggested guidelines, one may construct binary matrices for the model forecast, $M_{\pm}(n,m)$ and observed oil, $O_{\pm}(n,m)$ with each spanning a $N\,x\,M$ grid. The two matrices span the grid in a binary fashion with 1 assigned to a grid cell matching the forecast/observation (oil or no oil) while 0 is assigned to a cell that does not match. A model probability of oil detection in the average grid cell is defined as

$$p\langle M_+\rangle = \frac{1}{K}\cdot\sum_{n=1}^{N}\sum_{m=1}^{M} M_+(n,m) \tag{1}$$

with a similar definition for observed oil

$$p\langle O_+\rangle = \frac{1}{K}\cdot\sum_{n=1}^{N}\sum_{m=1}^{M} O_+(n,m) \tag{2}$$

and, $K = NM$

The model marginal probability prediction $p\langle M_+\rangle$ of oiling occurring, also called the marginal frequency or forecast rate, is

$$r = p\langle M_+\rangle = p\langle M_+ | O_+\rangle\, p(O_+) + p\langle M_+ | O_-\rangle\, p(O_-) = \frac{a+b}{a+b+c+d} \tag{3}$$

However, a high does not imply a high correspondence between the forecast and the observation since increases by both hits (a) and by false positives (b). If, $r = 1$ the model forecasts oil over the entire gridded area. In a minimal regret scenario, forecasters lean toward a high since this reduces the risk that the model prediction will fail to forecast oil $\left(p\langle M_- \cap O_+\rangle \neq \varnothing\right)$ for a grid cell with high environmental value (e.g., fish hatchery or turtle nesting site). Oil protection and recovery equipment in an emergency response is often limited. Responders need to know the forecasted 'no oil' areas to pre-stage equipment appropriately to protect threatened high-value resources.

A related descriptive statistic to is the base rate, s, which ignores the forecast and only looks at the marginal probability of observed oiling

$$s=\langle p(O_+)\rangle=\langle p(O_+|M_+)\,p(O_+)+p(O_+|M_-)\,p(O_-)\rangle=\frac{a+c}{a+b+c+d} \quad (4)$$

with s describing the rate of occurrence of the observations, and, for complete oiling, $s=1$. For a performance measure that depends on the base rate, comparison of scores between different oil spill events with different base rates is difficult. For this reason, performance measures independent of the base-rate are preferable when comparing different events or different models.

The ratio of $p(M_+)$ to s defines the frequency bias, B_f, of the forecast

$$B_f=\frac{p(M_+)}{s}=\frac{p(M_+)}{p(O_+)}=\frac{a+b}{a+c} \quad (5)$$

$B_f \geq 0$

is the ratio of the total number of oiled grid cells according to the trajectory forecast compared to the number of cells that were oiled according to observation. This number represents a simple measure that summarizes the tendency of the trajectory to under forecast or over forecast. A trajectory that consistently forecasts more surface oil coverage than observed coverage exhibits a high bias. As mentioned earlier, forecast models that implement a minimum regret strategy intentionally introduce bias into the trajectory to reduce risk to sensitive resources. Meager oil spill observations can also introduced bias into the forecast thru hedging. Although setting of thresholds has proved useful in the weather forecast community for adjusting forecast bias (Mittermaier & Roberts, 2010; Mittermaier, Roberts, & Thompson, 2013), for oil spill response based on minimum regret, setting the thresholds so that the bias is greater is usually advantageous and ensures a performance metric represents the actual forecast.

A very simple metric (PSS) was proposed by Pierce (1884) in the 19th century and is still used today; subtract the fraction of model misses from the number of model hits. The range of the PSS is from -1 (all wrong) to 1 (all correct). Defining

$$p\langle M_+ \cap O_+\rangle = \frac{1}{K}\cdot\sum_n\sum_m M_+(n,m)\cap O_+(n,m)$$

$$PSS = p\langle M_+ \cap O_+\rangle - p\langle M_+ \cap O_-\rangle$$

with forecast skill defined as the improvement over a reference forecast such as climatology, persistence or a random forecast.

There are several drawbacks to using the PSS as an oil spill metric for large offshore spills. Forecasters have a tendency to underestimate the occurrence or extent of rare events. For example, the initial estimate of the oil flow rate for the Deepwater Horizon Spill was underestimated considerably. The scientists, including one the authors, were reluctant to change the flow rate number by an order of magnitude from the original official value even though field observations from, among others, the other paper author, indicated that such a change was justified (States, 2013). This is such a common phenomenon that it has a common label; ‘hedging the forecast’. The PSS metric may actually favor such hedging. Consider that a spill forecast that predicts correctly the amount of surface oil coverage but misses the specific grid location in most cases will have a negative PSS value while the extreme hedge of predicting no oil will get a PSS value of zero. In addition, the raw PSS gives no indication of the proper length scale for the spill forecast. The previous section illustrated why this is important.

Roberts and Lean(2008) introduced an alternative metric, the fractional skill score (FSS) that, unlike traditional categorical skill scores such as PSS, an exact spatial match between forecast and observation is not necessary. While the mathematical notation of FSS is complex, the concept is simple. Beginning with a mesh consisting of a single large common grid cell shared by the prediction and observation. $M_+ \cap O_+ = 1$. For the non-trivial spill case, an oil spill exists, but the resolution is useless for spill response. As the common mesh cell number increases by reducing grid cell size, the forecast and observation will show increasing discrepancy. A 'good' forecast will show improvement over a persistent (assumes oil slick has not moved from last observation) or random forecast at a grid scale that provides optimum practical value to the response. A strength of FSS is that can also provide an estimate of the true spatial resolution of the forecast, which may be larger than the response optimum scale.

From a practical point-of-view creating a verification grid centered about a stationary, point-source release, such as a well blowout or grounded vessel, is straightforward and easily implemented operationally. For the experimental dataset, the nature of the ocean section covered by the oil leant itself to a square grid althought the technique would be similiar for a rectangular grid. The forecasts and the observations in the dataset do not include oil contacting the shoreline; only offshore oil. The model grid resolution during the spill incident ranged from ~3 to 14 km (MacFayden et al., 2011). Unfortunately, we were unable to determine the model resolution used to generate each forecast in the example data set. The resolution of the satellite sensors ranged from 18 to 250 m with 100 being the common pixel size. Although Skok and Roberts (2018) recommend a common grid that closely matches the coarstest resolution of the observation and forecast, we decided to use 5 km for the basis of comparison as this was most frequently used grid resolution during the spill incident. The verification grid consisted of 89 x 89 grid squares with each cell length, 5 .

For convenience, we considered a common square grid mesh of square cells. Some intermediate resolution will group the cells into new larger square cells with integer multiple of

km $N \times N$

the original grid cell length . Because the new grid cells are larger, there are fewer of them to cover the same gridded area or mesh. Let be the number of larger grid cells with . We can define a probability of modeled forecast detection for each of the individual grid cells

$$\langle \quad \rangle \quad — \tag{9}$$

A similar definition holds for observations. Next, define an intermediate skill score, called the Fractions Brier Score (FBS) as

$$FBS_n = \frac{1}{K_n}\sum_{K_n}\left(p_n\langle M_+\rangle - p_n\langle O_+\rangle\right)^2 \tag{10}$$

and the Fractions Skill Score (FSS) as

$$FSS_n = 1 - \frac{K_n \cdot FBS_n}{\left(\sum_{K_n} p_n^2\langle M_+\rangle - \sum_{K_n} p_n^2\langle O_+\rangle\right)} \tag{11}$$

$p_n\langle M_+\rangle = \frac{1}{K_n}\sum_{i=1}^{n}\sum_{j=1}^{n} M_+(i,j)$

$K_n < K$

$n > 1$

Calculations of the FSS are computationally intensive. However, Fagin et al (2015) significantly reduce the computational time by quickly computing “summed area tables”. This approach effectively clips the grid cells extending beyond the domain and avoids the need to pad the matrix with zeros.

As a measure of forecast quality, Roberts and Lean suggest two measures: random and uniform forecasts. The random forecast has the same fractional coverage over the model domain as that of the observed oil, $\langle \quad \rangle$, so that $\langle \quad \rangle$. The FSS has a range of 0 to 1 if there is an equal number of observed and forecast cells containing oil, and therefore, no frequency bias. However, as originally defined by Roberts and Lean (2008) and discussed further in Skok (2015; 2016) and generally recommended in Skok & Roberts (2016; 2018), the forecast indicates useful skill at the smallest neighborhood size at with the following caveat, the frequency bias is less than 1.5 to 2.0. Otherwise, Roberts and Lean indicate a target or useful skill halfway between the random forecast and a perfect skill defined by

$$FSS_{uniform} = \frac{1}{2} \cdot \left(1 + p\langle O_+ \rangle\right) \tag{12}$$

$FSS = 0.5$ $FSS_{random} = p\, O_+$

The FSS procedure inherently contains sampling uncertainties, particularly since the example dataset consists of six coupled forecasts. In addition, Stephenson (2000) suggests an estimate of statistical error can demonstrate that skill does not occur simply by chance. To estimate the FSS uncertainty, we use a similar bootstrapping approached described in Kuell & Bott (2019). The original observation matrix and forecast matrix are each randomly sampled with replacement at the nearest neighbor of each grid point. FSS values are then calculated for 1,000 bootstrap sampled forecast and observation matrixes, ranked in ascending order using the 97.5th and 2.5th percentile of the distribution representing the 95% confidence interval. Figures 6 and 7 demonstrate the use of the confidence intervals.

3.0 Results

The FSS was calculated for each forecast over horizontal scales ranging from a single grid cell, (5 km) to an 89 x 89 (445 km) grid cell domain, Figure 6. Overlaid with a shaded band is the 95% confidence interval. The resulting graphs show all the forecasts improve as the spatial scale increases. The forecasts do not reach perfect skill, 1, as each forecast indicates bias. As discussed in Section 3, modelers may hedge the forecast as part of a minimum regret strategy so that is expected. The for five forecasts ranged from 1.7 and 2.3 , Figure 6(b), (c), (d), (e) and (f). Surprisingly, the 17-h forecast was highest with , Figure 6(a).

Forecast quality is evaluated using two measures: random and uniform forecasts. Figure 6 clearly indicates that at the grid scale, 5 km ( ), all of the forecasts were more skillful than a random forecast with ≃ . The intercept between the curves and indicates the smallest scale that the forecast is considered skillful. Above the line, the forecast displays useful skill. Interestingly of the six forecasts, the 41-h forecast achieved skill at the smallest horizontal scale at ~45 km ( ) with , Figure

6(c). This is consistent with visual examination of the 41-h forecast in Figure 4(c). The predication indicates some oil to the east and slightly more oil southeast of the blowout compared to the other forecasts. Figures 6(b), (e) and (f) indicate the 21-, 65- and 70-h lead-times had similar overall horizontal lengths ranging from ~65 km (n = 13) to 85 km ( ) at . Significantly, the 17- and 46-h forecasts achieved skill at the largest scales ranging from 125 km (n = 24) to 185 km (n = 37). Overall, a compelling feature of Figure 6 is the scores do not correlate well with lead-time. However, within the 95% confidence interval, there is considerable overlap between the 21, 41, 46, 65 and 70 lead-times.

$n = 17$ $FSS_{uniform}$

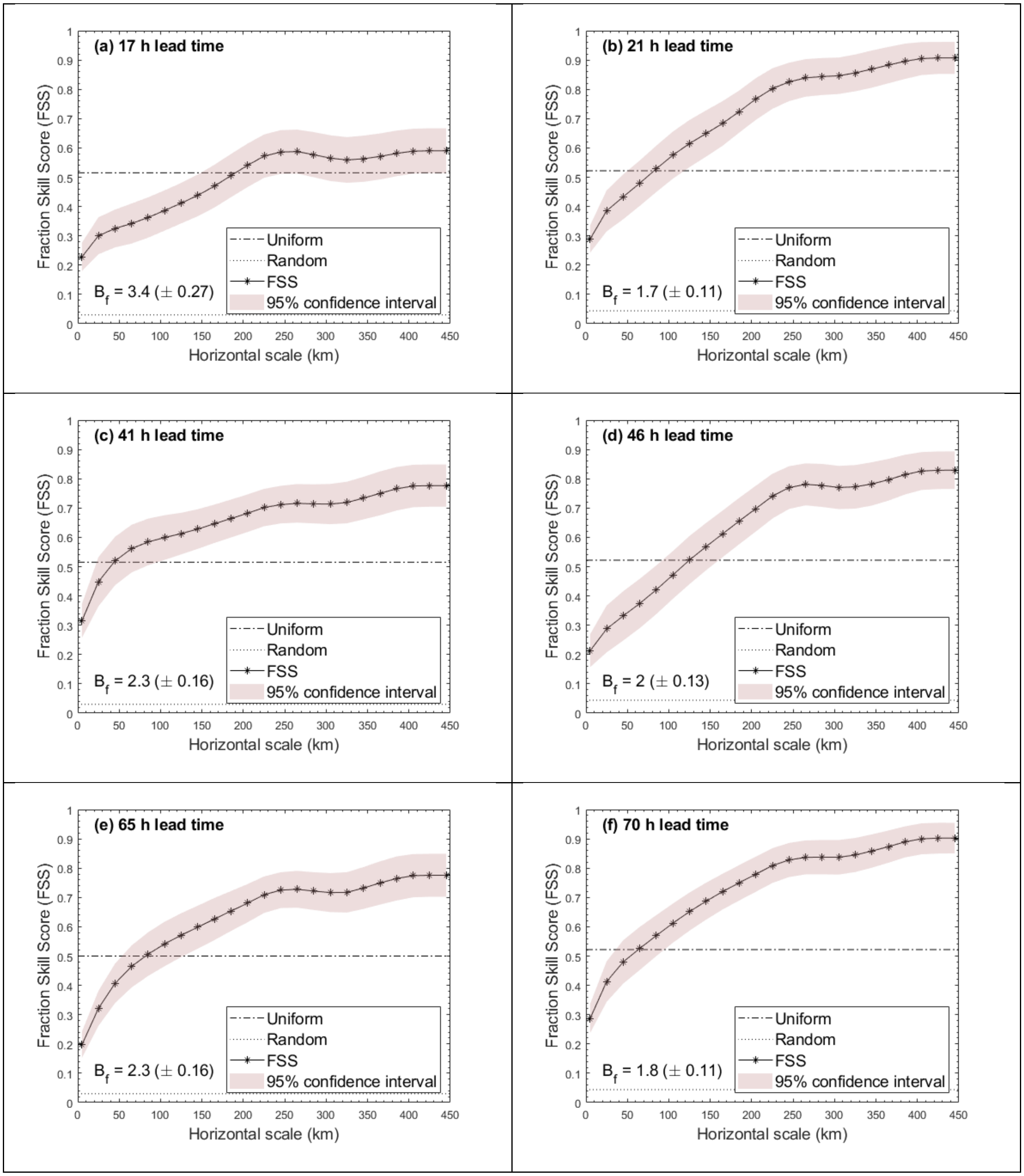

Figure 6. Variation of FSS with horizontal scale for the 17-, 21-, 41-, 65- and 70-h lead times. Dashed and dotted lines are uniform and random FSS, respectively. The shaded band around each curve shows the 95% confidence interval for 1,000 bootstrapped samples.

The matched forecasts and observations are pooled into 17-21, 41-46 and 65-70 h lead-times, Figure 7. As recommended by the (WWRP/WGNE, 2017) for the weather forecast verification community, the binary counts are pooled using the same number of original grid cells. Rather than averaging scores for all forecasts, the aggregated statistics are 'more robust'. Note the shaded band specifies the 95% confidence interval. Aggregating the forecasts lowered the bias such that, . As discussed in Section 3, the forecast indicate skill at the smallest scale when FSS = 0.5, if . That said, the calculated , is consistent with the guidance noted by Skok and Roberts (2018). Again, the pooled forecast exceed the skill of a random forecast at the grid scale 5 km ( ) and ≃ . For the 41-46 and 65-70 lead-times, is reached at scales of ~45 km ( ) and ~75 km ( ), respectively, with considerable overlap as indicated by shaded 95% confidence interval in Figure 7(b) and (c). The 17-21 h lead-time forecast did not perform as well with skill ranging from 185 to 205 km.

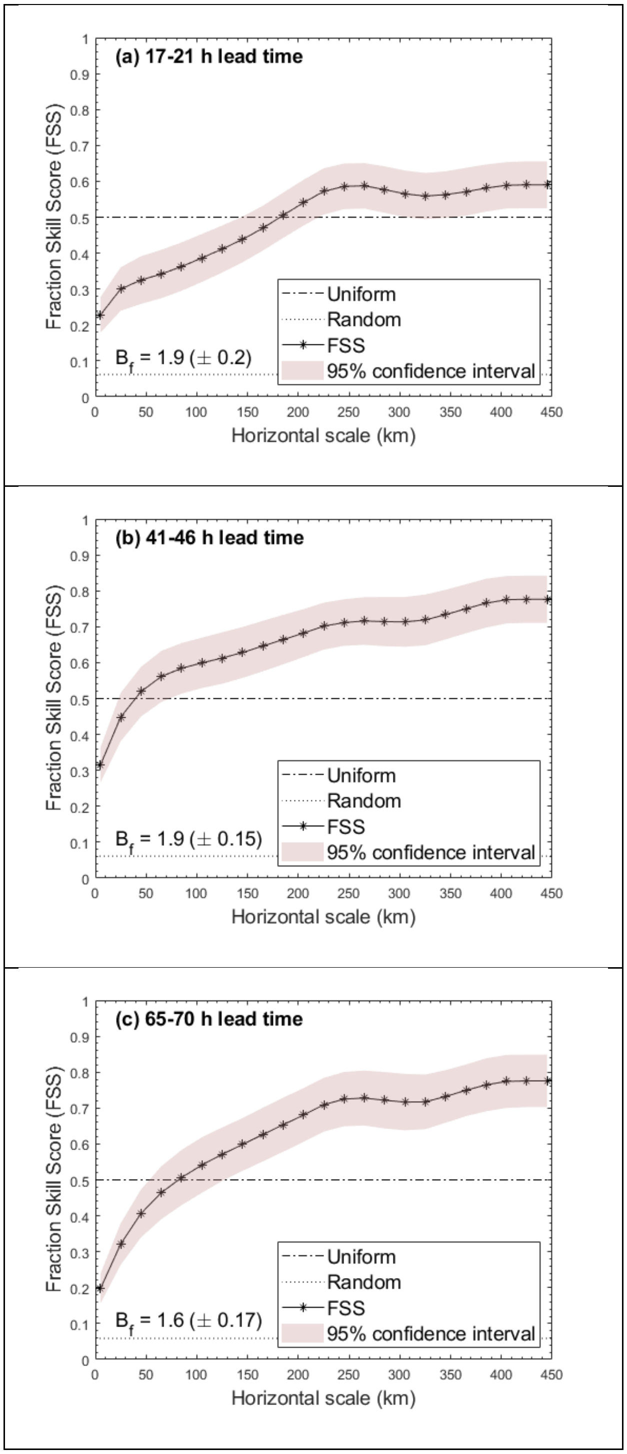

Figure 7. Variation of aggregated FSS with horizontal scale for the 17-, 21-, 41-, 65- and 70-h lead times. Dashed and dotted lines are uniform and random FSS, respectively. The shaded band around each curve shows the 95% confidence interval for 1,000 bootstrapped samples.

4.0 Discussion and conclusion

In this paper, we have presented an approach for identifying the scale at which the oil spill forecast demonstrates useful skill. The experimental dataset used in the study included six forecasts and five satellite products routinely used for day-to-day operations during an actual spill incident. The forecasts, with resolution of ~3 to 14 km, were verified against satellite imagery at 18 to 250 m resolution using a 5 km common grid. We found that temporally compositing the observations centered on 1-h of acquisition times was practical as oil transport for the period of the dataset was ~ over a one-hour period.

As spatial verification methods are new to oil spill forecasting, no other studies exist for comparison with these results. However, the horizontal spatial scales are consistent with FSS greater than found in numerical weather prediction. For precipitation forecasts, Lewis et al. (2015) reported horizontal scales of 30 to 70 km using a 1 km verification grid. In addition, Kuell & Bott (2019) estimated useful scales ranging from 31 to 101 km using a 7 km grid. Here, we used a 5 km verification grid and, by aggregating the forecasts, showed better results than the individual forecasts

$1km$

with the 41-46 h and 65-70 h lead-times achieving useful skill at approximately 45 km and 75 km, respectively. The results in this study are also consistent with other precipitation assessments. Mittermaier (2006) and later, Mittermair et al (2013) noted that precipitation forecast skill is evident at two to three times the coarsest grid resolution. As it turns out, this general guidance appears to hold true for the 41-46 h and 65-70 h lead-times. In the example dataset, the coarsest resolution used to generate the forecast was ~ 14 km, meaning, we would expect apparent forecast skill at a resolution of about 30 - 75 km. In contrast, the 17-21 lead-time reaches between 185 and 205 km. While unexpected and not within the scope of this study, a detailed review of the forecasts, particularly the 17-h lead-time, may reveal a likely cause for this discrepancy.

An important weakness in this analysis is limiting the calculation of the FSS values over the entire verification grid. There is a noteworthy feature observed in the satellite imagery southeast of the blowout. The oil slicks in this area eventual moved into the Loop Current with significant planning repercussions for the emergency response (Liu et al., 2014). MacFadyen et al. (2011) suggested the hydrodynamic models used to develop the forecasts varied in horizontal resolution and were sensitive to the position of the Loop Current and the shedding of eddies. However, we evaluated the FSS for the entire domain, rather than a specific event, resulting in an aggregated skill score. Mittermair and Roberts (2010) suggest assessing a discrete event by examining a smaller domain. Further analysis may provide the hydrodynamic resolution of the Loop Current and eddy movement that compares best with oil observations by reducing the domain to a sub-region containing these features. However, this is not within the scope of this study but should be a consideration in further skill assessments.

Several other factors may have contributed to the poor performance of the 17-21 h results. First, not considered was observational uncertainty. The satellite observations in the study identified the existence of oil film but not thickness, which can vary by three or more orders of magnitude. Spill models, on the other hand, track oil volume. The model was possibly tracking the main oil content properly but not the light sheen recorded along with the thick oil. The image analysis is also susceptible to its own false positives. Waves rupturing the oil film are often mistaken as non-oiled areas and the non-petroleum films interpreted as oiled areas. Therefore, the oil spill remote sensing community is encouraged to report errors and uncertainty in remote sensing products disseminated to both oil spill responders and forecasters.

Operational oil spill modeling should consistently provide helpful forecasts to emergency responders. However, spatially distributed oil spill forecasts present huge challenges regarding verification. A significant finding of this research is that the FSS provides a useful insight into the appropriate scale for presenting the oil spill forecasts. We assume that the primary purpose of the skill scores, which can provide results in real-time, is to allow modelers to adjust the parameters in their models to improve operational forecast accuracy for the next time-period of forecast. An additional benefit accrues to the response team for help in assessing the degree of confidence that placed in any specific forecast or model choice.

Disclaimer

The findings and conclusions in this paper are those of the author(s) and do not necessarily represent the views of the National Oceanic and Atmospheric Administration. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

Bassett, C., A. Lavery, Makym, T., 2016. Broadband acoustic backscatter from crude oil under laboratory-grown sea ice. *J. of the Acoustical Soc. of America*, 140, 2274. https://doi.org/10.1121/1.4963876.

Cheng, Y., Li, X., Garcia-Pineda, O., Andersen, O. B., Pichel, W. G., 2011. SAR observation and model tracking of an oil spill event in coastal waters. *Marine Pollution Bulletin, 62*(2011), 350-363. doi:10.1016/j.marpolbul.2010.10.005

Cheng, Y., Liu, B., Li, X., Xu, Q., Ding, X., Migliaccio, M., 2014. Monitoring of oil spill trajectories with COSMO-SkyMed X-Band SAR images and model simulation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2895 - 2901.doi:10.1109/JSTARS.2014.2341574.

Deepwater Horizon Natural Resource Damage Assessment Trustees, 2016. *Deepwater Horizon oil spill: Final Programmatic Damage Assessment and Restoration Plan and Final Programmatic Environmental Imact Statement.* Retrieved from http://www.gulfspillrestoration.noaa.gov/restoration-planning/gulf-plan.

De Dominicis, M., Bruciaferr, i. D., Gerin, R., Pinardi, N., Poulain, P., Garreau, P., 2016. A multi-model assessment of the impact of currents, waves and wind in modelling surface drifters and oil spill. *Deep Sea Research Part II: Topical Studies in Oceanography*, 21-38. Retrieved from http://dx.doi.org/10.1016/j.dsr2.2016.04.002.

Ebert, E., 2008. Fuzzy verification of high-resolution gridded forecasts: a review and proposed framework. *Meterological Applications*, 51-64. doi:10.1002/met.25

Ebert, E., 2009. Neighborhood Verification: A Strategy for Rewarding Close Forecasts. *Weather and Forecasting*, 1498-1510. doi:10.1175/2009WAF2222251.1.

Faggian, N., Roux, B., Steinle, P., Ebert, B., 2015. Fast calculation of the fractions skill score. *MAUSAM, 66*, 457-466.

Ferro, C. A., Stephenson, D. B., 2011. Extreme dependence indices: Improved verification measures for deterministic forecasts of rare binary events. *Weather and Forecasting, 26*, 699-713. doi:10.1175/WAF-D-10-05030.1

Fingas, M., 2018. The Challenges of Remotely Measuring Oil Slick Thickness. *Remote Sens., 10*, 319. https://doi.org/10.3390/rs10020319.

Fingas M., Brown, C., 2018. A Review of Oil Spill Remote Sensing. *Sensors*, 18, 91.

Galt, J. A., 1998. Uncertainty analysis related to oil spill modeling. *Spill Science & Technology Bulletin*, 231-238.

Garcia-Pineda, O., Staples, G., Jones, C. E., Hu, C., Holt, B., Kourafalou, V., Graettinger, G., DiPinto, L., Ramirez, E., Streett, D., Cho,J., Swayze, G. A., Sun, S., Garcia, D., Haces-Garcia, F., 2020. Classification of oil spill by thicknesses using multiple remote sensors, *Remote Sensing of Environment*, 236. doi:org/10.1016/j.rse.2019.111421.

Guo, W., Jiang, M., Li, X., Ren, B., 2018. Using a genetic algorithm to improve oil spill prediction. *Marine Pollution Bulletin*, 386-396. https://doi.org/10.1016/j.marpolbul.2018.07.026.

Huntley, H. S., Lipphardt Jr., B. L., Kirwan Jr., A. D., 2011. Surface drift predictions of the Deepwater Horizon spill: The Lagrangian perspective. In Y. Liu, A. MacFadyen, Z.-G. Ji, & R. H. Weisberg, *Monitoring and Modeling the Deepwater Horizon Oil Spill: A Record-Breaking Enterprise* (pp. 179-195). Washington D. C.: American Geophysical Union. doi:10.1029/2011GM001146

Janeiro, J., Zacharioudaki, A., Sarhadi, E., Neves, A., Martins, F., 2014. Enhancing the management response to oil spills in the Tuscany Archipelago through operational modelling.

Marine Pollution Bulletin, 574-589. https://doi.org/10.1016/j.marpolbul.2014.03.021.

Jolliffe, I., Stephenson, D., 2012. *Forecast Verification.* Wiley-Blackwell.

Kim, T. H., C.S., Yang, J.H., Oh, Ouchi, K., 2014. Analysis of the contribution of wind drift factor to oil slick movement under strong tidal condition: Hebei Spirit oil spill case. *Plos One, 9*(1), 1-14. Retrieved from www.plosone.org

Kuell, V., & Bott, A. (2019). A physical subgrid-scale information exchange (PSIE) system for parametrization schemes in numerical weather prediction models. *Q. J. R. Meteorol Soc.* , 767-783. https://doi.org/10.1002/qj.3464

Le Hénaff, M., Kourafalou, V. H., Paris, C. B., Helgers, J., Aman, Z. M., Hogan, P., Srinivasan, A., 2012. Surface evolution of the Deepwater Horizon oil spill patch: Combined effects of circulation and wind-induced drift. *Environmental Science & Technology, 46*(2012), 7267-7273. https://doi.org/10.1021/es301570w

Lehr, W., Simecek-Beatty, D., Fingas, M., 2019. Whither oil spill models in the next decade? *Proceedings of the Forty-second AMOP Technical Seminar* (pp. 453-472). Environment and Climate Change Canada.

Leifer, I., Lehr, W. J., Simecek-Beatty, D., Clark, R., P, D., Hu, Y.,Matheson, S., Jones, C. E., Holt, B., Reif, M., Roberts, D. A., Svejkovsky, J., Swayze, G., Wozencraft, J. 2012. State of the art satellite and airborne marine oil spill remote sensing: Application to the BP Deepwater Horizon oil spill. *Remote Sensing of the Environment, 124*(2012), 185-209. https://doi.org/10.1016/j.rse.2012.03.024

Lewis, H., Mittermaier, M., Mylnc, K., Norman, K., Scaife, A., Neal, R., Pierce, C., Harrison, D., Jewell, S., Kendon, M., Saunders, R., Brunet, G., Golding, B., Kitchen, M., Davies, P., Pilling, C., 2015. From months to minutes-exploring the value of high-resolutin rainfall observation and prediction during the UK winter storms of 2013/2014. *Meteorological Applications, 22*, 90-104. https://doi.org/10.1002/met.1493

Liu, Y., Weisberg, R., Hu, C., & Zheng, L., 2011. Trajectory forecast as a rapid response to the Deepwater Horizon oil spill. In Y. Liu, A. Macfadyen, Z.Ji, and R. Weisberg, *Monitoring and Modeling the Deepwater Horizon Oil Spill:* A Record-Breaking Enterprise (153–165). Washington D. C.: American Geophysical Union. https://doi.org/10.1029/2011GM001121.

Liu, Y., R. Weisberg, S. Vignudelli, Mitchum, G. T., 2014. Evaluation of altimetry-derived surface current products using Lagrangian drifter trajectories in the eastern Gulf of Mexico. *J. Geophys. Res. Oceans*, 119, 2827–2842. doi:10.1002/2013JC009710.

MacFadyen, A., Watabayashi, G., Barker, C. H., Beegle-Krause, C. J., 2011. Tactical modeling of surface oil transport during the Deepwater Horizon spill response. In Y. Liu, A. MacFadyen, Z. Ji, & R. Weisberg, *Monitoring and Modeling the Deepwater Horizon Oil Spill:* A Record-Breaking Enterprise (pp. 167-177). Washington D. C.: American Geophysical Union. https://doi.org/10.1029/2011GM001128.

Malis, E., 2004. Improving vision-based control using efficient second-order minimization techniques. *IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA '04. 2004, 2*, pp. 1843-1848. New Orleans. doi:10.1109/ROBOT.2004.1308092.

Mittermaier, M. P., 2006. Using an intensity-scale technique to assess the added benefit of high-resolution model precipitation forecasts. *Atmospheric Science Letters,* 7(2), 35-42. https://doi.org/10.1002/asl.127.

Mittermaier, M. P., Csima, G., 2017. Ensemble versus deterministic performance at the kilometer scale. *Weather and Forecasting, 32*, 1697-1709. https://doi.org/10.1175/WAF-D-16-0164.1.

Mittermaier, M., Roberts, N., 2010. Intercomparison of Spatial Forecast Verification Methods: Identifying Skillful. *Weather and Forecasting, 25*, 343-354. doi:10.1175/2009WAF2222260.1

Mittermaier, M., Roberts, N., Thompson, S., 2013. A long-term assessment of precipitation forecast skill using the Fractions Skill Score. *Meteorological Applications*, 176-186. https://doi.org/10.1002/met.296.

NDBC (National Data Buoy Center), 1971. Meteorological and oceanographic data collected from the National Data Buoy Center Coastal-Marine Automated Network (C-MAN) and moored (weather) buoys. NDBC Buoy 42040. NOAA National Centers for Environmental Information. Dataset. Retrieved from https://www.ndbc.noaa.gov/station_history.php?station=42040

Oey L.-Y., T. Ezer, G. Forristall, C. Cooper, S. DiMarco, Fan, S., 2005. An exercise in forecasting loop current and eddy frontal positions in the Gulf of Mexico, Geophys. Res. Lett., 32, L12611. https://doi.org/10.1029/2005GL023253.

Özgökmen, T. M., Chassignet, E. P., Dawson, C. N., Dukhovskoy, D., Jacobs, G., Ledwell, J., Garcia-Pineda, O., MacDonald, I. R., Morey, S. L., Olascoaga, M. J., Poje, A. C., Reed, M. Skancke, J., 2016. Over what area did the oil and gas spread during the 2010 Deepwater Horizon oil spill? *Oceanography, 29*(3), 97-107. doi:org/10.5670/oceanog.2016.74.

Peirce, C. S., 1884. The numerical measure of the success of predictions. *Science, 4*(93), 453-454.

Pisano, A., De Dominicis, M., Biamino, W., Bignami, F., Gherardi, S., Colao, F., G. Coppini, S. Marullo, M. Sprovieri, P. Trivero, E. Zambianchi, R. Santoleri, 2016. An oceanographic survey for oil spill monitoring and model forecasting validation using remote sensing and in situ data in the Mediterranean Sea. *Deep-Sea Res. II, 133*, 132-145. doi:10.1016/j.dsr2.2016.02.013.

Roberts, N., Lean, H. 2008. Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Monthly Weather Review, 136*. doi:10.1175/2007MWR2123.1

Sarrute, C., Burroni, J., 2008. Using neural networks to improve classical operating system fingerprinting techniques. *Electronic Journal of SADIO, 8*(1), 35-47.

Skok, G., 2015. Analysis of fraction skill score properties for a displaced rainband in a rectangular domain. *Meteorological Applications*, 477-484. doi:10.1002/met.1478

Skok, G., 2016. Analysis of Fraction Skill Score properties for a displaced rainy grid point in a rectangular domain. *Atmospheric Research, 169*, 556-565. https://doi.org/10.1016/j.atmosres.2015.04.012.

Skok, G., Roberts, N., 2016. Analysis of fractions skill score properties for random precipitation fields and ECMWF forecasts. *Q. J. R. Meteorol. Soc.*, 2599-2610. doi:10.1002/qj.2849.

Skok, G., Roberts, N., 2018. Estimating the displacement in precipitation forecasts using the Fractions Skill Score. *Q. J. R. Meteorol. Soc., 144*, 414-425.

Skou, N., 1986. Microwave Radiometry for Oil Pollution Monitoring, Measurements, and Systems. *IEEE Trans. Geosci. Remote Sens.*, 360–367.

Smith C. L., 1976. Determination of the Leeway of Oil Slicks. In: Wolfe DA, Anderson JW, Button DK, Malins DC, Roubal T, Varanasi U, editors. *Fate and Effects of Petroleum Hydrocarbons in Marine Ecosystems and Organisms*, 351. New York: Pergamon Press.

Spaulding, M. L., 2017. State of the art review and future directions in oil spill modeling. *Marine Pollution Bulletin*, 7-19. https://doi.org/10.1016/j.marpolbul.2017.01.001.

States, U., 2013. *The BP oil spill: Accounting for the spilled oil and ensuring the safety of seafood from the Gulf:hearing before the Subcommittee on Energy and*

Environment of the Committee on Energy and Commerce. Washington D.C.: House ofRepresentatives, One Hundred Eleventh Congress, second session, August 19, 2010. Retrieved from http://books.google.com/books?id=0dVIAQAAMAAJ

Stephenson, D., 2000. Use of the ''Odds Ratio'' for Diagnosing Forecast Skill. *Weather and Forecasting, 15*, 221-232. https://doi.org/10.1175/1520-0434(2000)015<0221:UOTORF>2.0.CO;2.

Street, D., 2011. NOAA's satellite monitoring of marine oil. In Y. Liu, A. MacFadyen, Z.-G. Ji, & R. H. Weisberg, *Monitoring and Modeling the Deepwater Horizon Oilspill: A Record-Breaking Enterprise*, 9-18. Washington D. C: American Geophysical Union. doi:10.1029/2011GM001146

USCG (United States Coast Guard), 1991. United States Coast Guard National Search and Rescue Manual, Vol. II: Planning Handbook.

USCG (United States Coast Guard), 2011. On Scene Coordinator Report Deepwater Horizon Oil Spill, submitted to the National Response Team September 2011.

Wilks, D. S., 2011. *Statistical Methods in the Atmospheric Sciences.* Elsivier.

WWRP/WGNE, 2017. 7th International Verification Methods Workshop. Berlin: World Weather Research Programme (WWRP). Retrieved from http://www.cawcr.gov.au/projects/verification/.

Xu, F. H., Chang, Y, L., Oey, L. Y., Hamilton. P., 2013. Loop current growth and eddy shedding using models and observations: Analyses of the July 2011 eddy-shedding event, *J. Phys. Oceanogr.*, 43, 1015- 1027. doi:10.1175/JPO-D-12-0138.1.

Yingcheng, L., Tian, Q., Wang, X., Zheng, G., Li, X., 2013. Determining oil slick thickness using hyperspectral remote sensing in the Bohai Sea of China, *Int. J. of Digital Earth, 6, 76-93.* doi:10.1080/17538947.2012.695404.

Table 1. Contingency table to evaluate oil spill forecasts, modified from Jolliffe & Stephenson (2012).

	Oil observation	
Oil Model-forecast	Yes	No
Yes	a (Hit)	b (False alarm)
No	c (Miss)	d (Correct rejection)

Table 2. Description of the six forecasts. Date and time are Central Daylight Time (CDT). Lead-time is the hours between Forecast Prepared and Forecast Estimate.

	Forecast Prepared	Forecast Estimate	Lead Time (h)
1	5 May 2010 at 1300	8 May 2010 at 0600	65
2	5 May 2010 at 2000	8 May 2010 at 1800	70
3	6 May 2010 at 1300	8 May 2010 at 0600	41
4	6 May 2010 at 2000	8 May 2010 at 1800	46
5	7 May 2010 at 1300	8 May 2010 at 0600	17
6	7 May 2010 at 2100	8 May 2010 at 1800	21

Table 3. Experimental Marine Pollution Surveillance Reports (EMPSR) used as 'oil observation' for forecast evaluation. Date and time are Central Daylight Time (CDT).

	EMPSR Source	Image Acquisition
1	COSMO-Skymed2	8 May 2010 at 0657
2	RADARSAT-2	8 May 2010 at 0659
3	TerraSAR-X	8 May 2010 at 1823
4	COSMO-Skymed 2	8 May 2010 at 1851
5	RADARSAT -1	8 May 2010 at 1858

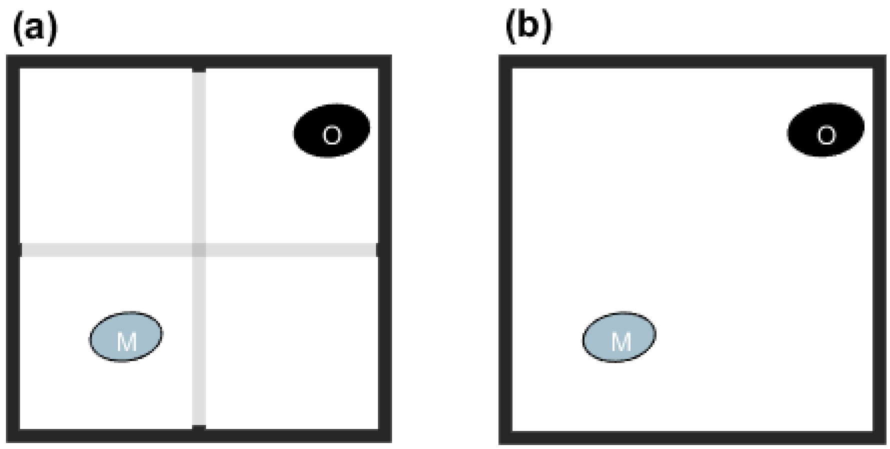

Figure 1. Model-forecast, 'M' is shaded gray and the observed oil, 'O', shaded black with (a) 5 km grid resolution and (b) 10 km grid resolution.

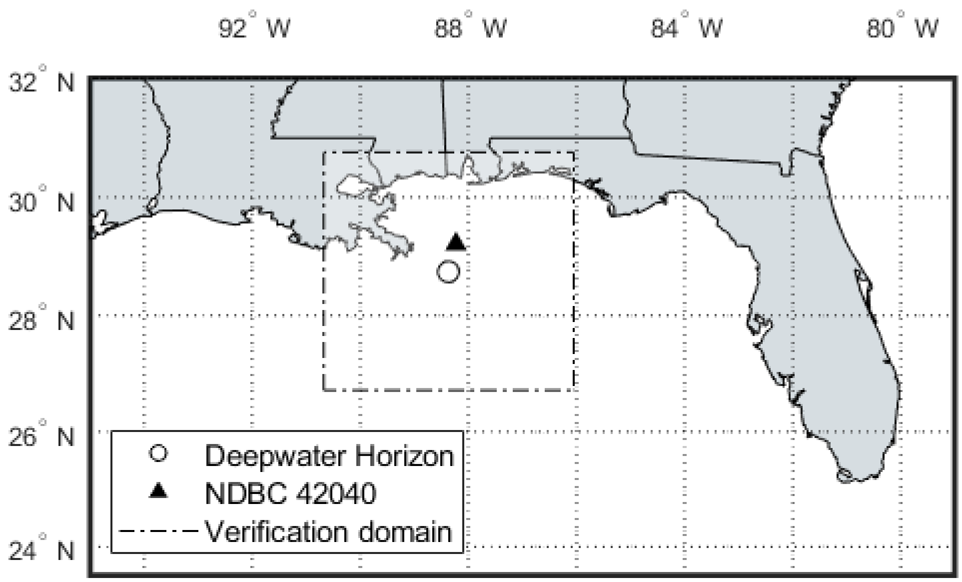

Figure 2. Map showing the *Deepwater Horizon* well site, ‘o’, the National Data Buoy Center (NDBC) Buoy 42040,’▲’ and boundary of the verification domain,’- ‘.

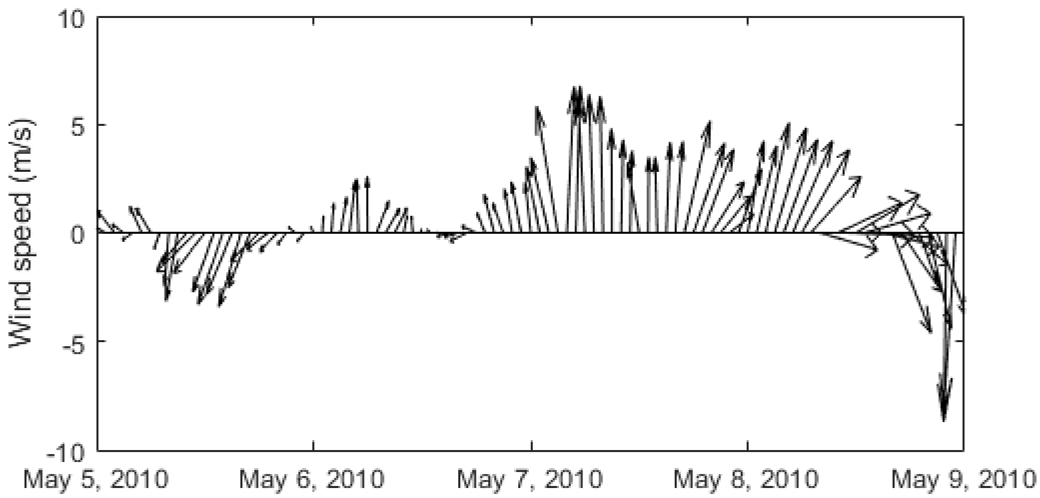

Figure 3. Wind observations for Buoy 42040 (NDBC, 1971).

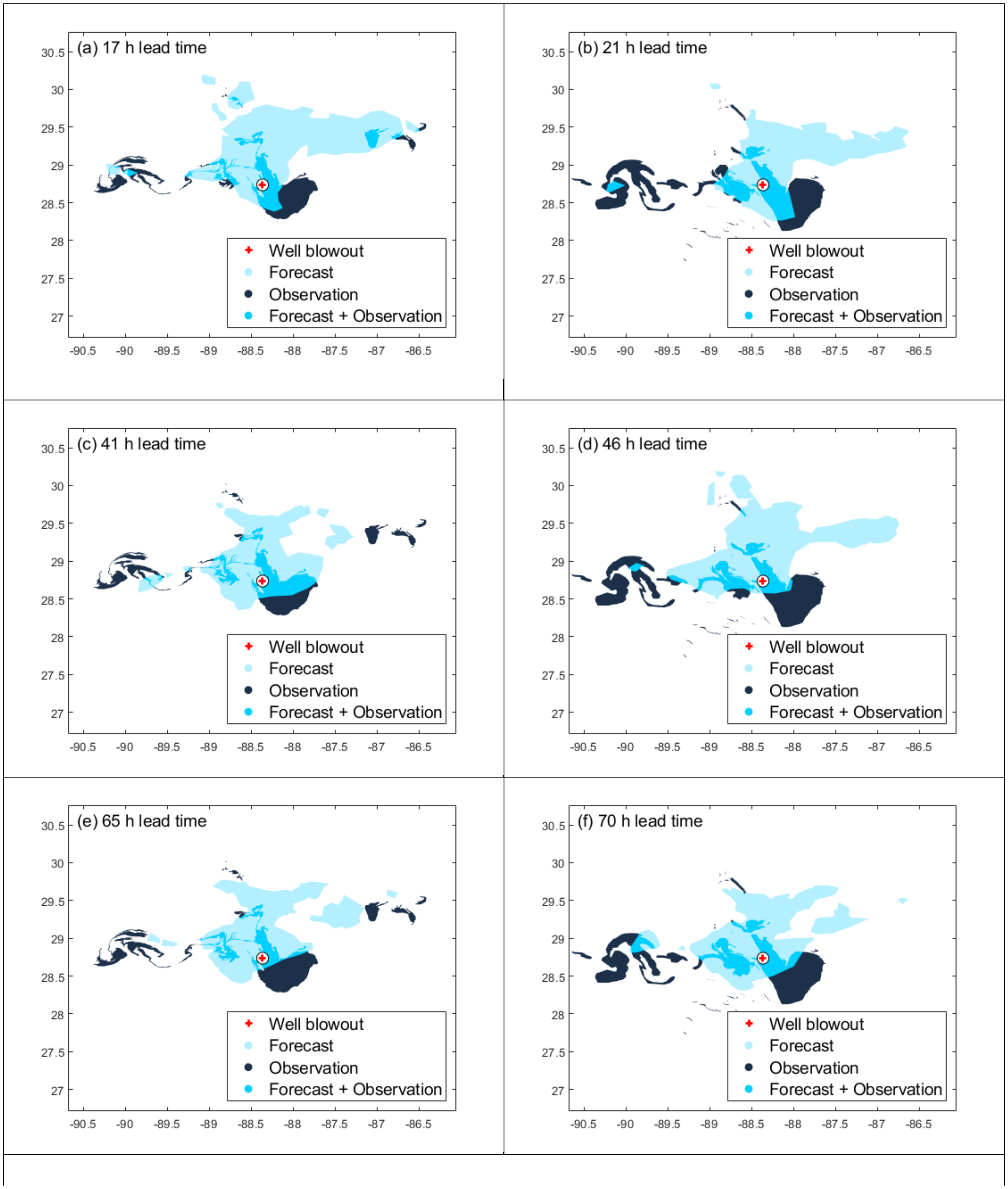

Figure 4. Forecast are shaded blue, the observed oil, black and the overlap of the observed oil and forecast, dark blue. For clarity, the coastlines are not plotted but for reference, the well blowout is marked '+'.

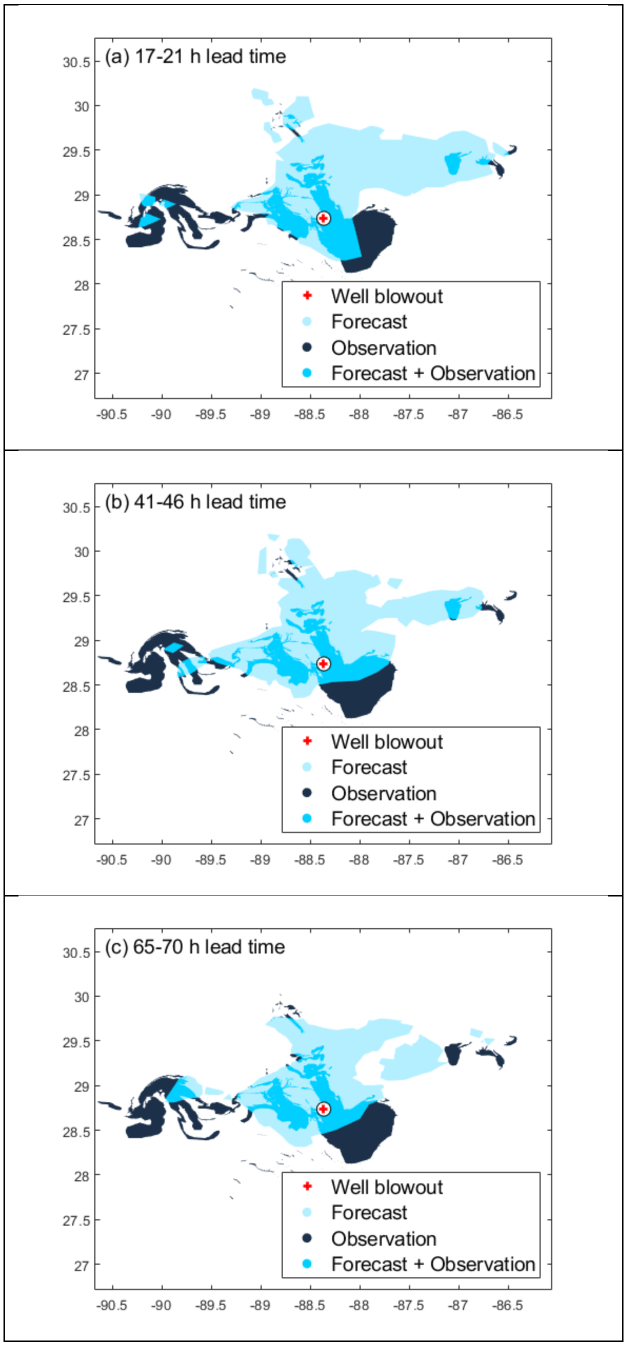

Figure 5. Aggregated forecast are shaded blue, the observed oil, black and the overlap of the observed oil and forecast, dark blue. The coastlines are not plotted but for reference, the well blowout is marked '+'.

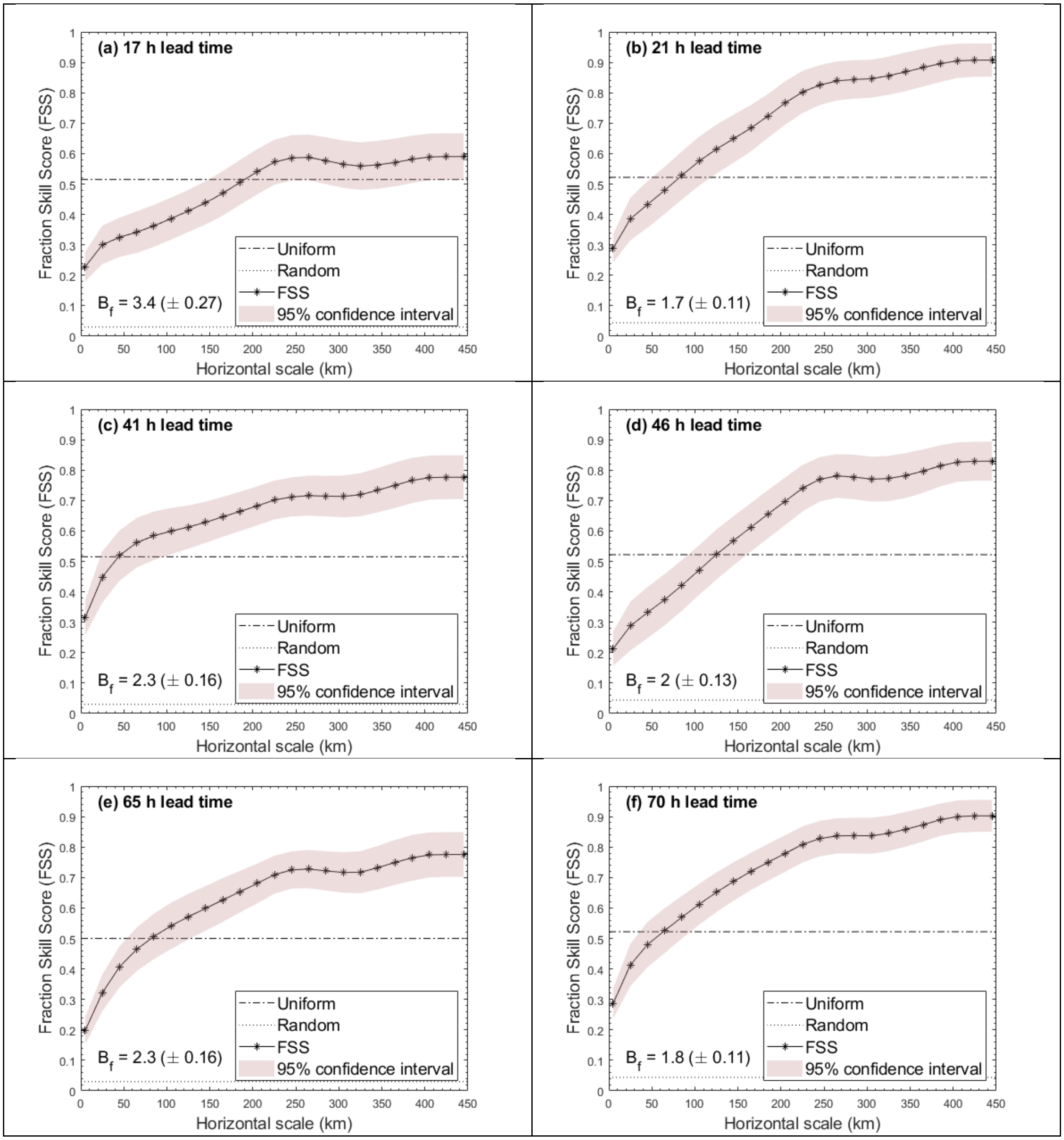

Figure 6. Variation of FSS with horizontal scale for the 17-, 21-, 41-, 65- and 70-h lead times. Dashed and dotted lines are uniform and random FSS, respectively. The shaded band around each curve shows the 95% confidence interval for 1,000 bootstrapped samples.

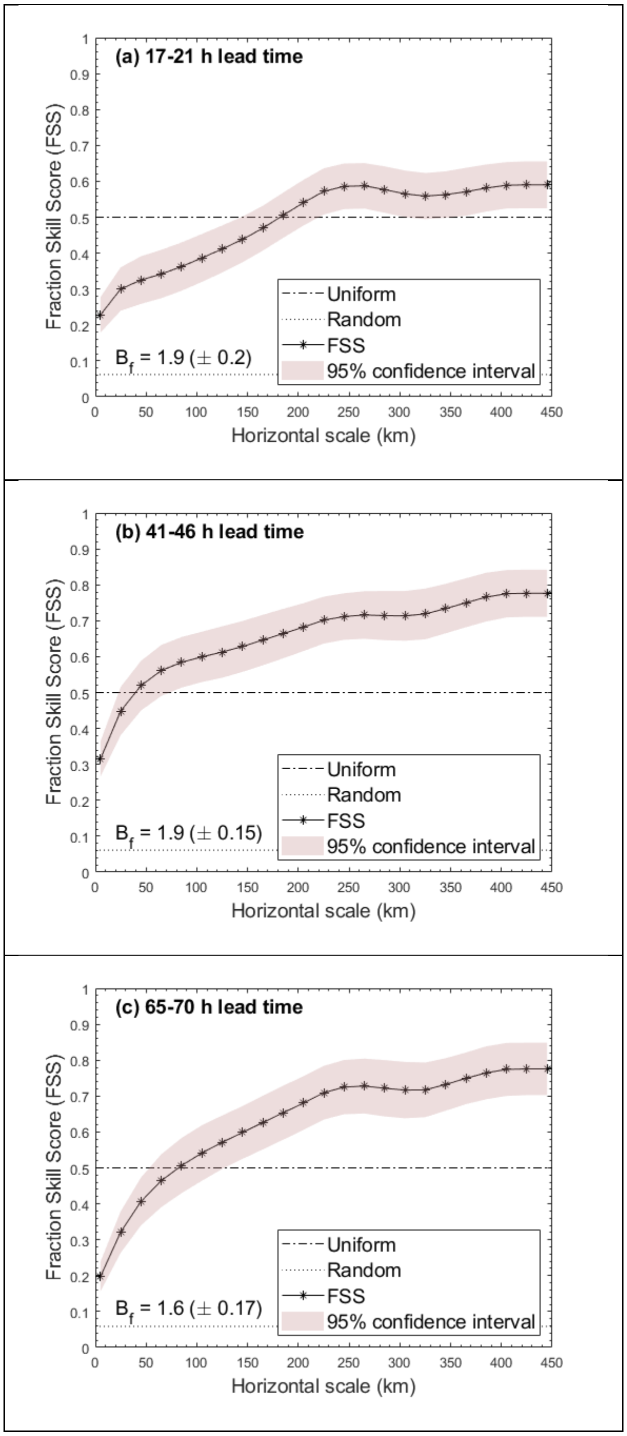

Figure 7. Variation of aggregated FSS with horizontal scale for the 17-, 21-, 41-, 65- and 70-h lead times. Dashed and dotted lines are uniform and random FSS, respectively. The shaded band around each curve shows the 95% confidence interval for 1,000 bootstrapped samples.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Credit Author Statement

Debra Simecek-Beatty, Conceptualization, Methodology, William J. Lehr, Formal analysis