Performance Measures for Validation of Oil Spill Dispersion Models Based on Satellite and Coastal Data

Chris Dearden, Tim Culmer, and Richard Brooke

***Abstract*—This article presents a set of performance metrics, whose purpose is to provide a quantitative measure of the ability of oil spill dispersion models to simulate real-world oil spills. The metrics are described in detail and are applied to the output from an existing oil spill model for two specific case studies. The metrics in question make use of both satellite imagery and coastal impact reports as the basis of the validation. Specifically, we recommend the 2-D measure of effectiveness as a means of quantifying model performance based on the extent of overlap between the observations and the model output. Additionally, we show that it is advantageous to supplement the 2-D measure of effectiveness with a newly proposed set of skill scores, based on the geometric area and centroid of a given oil spill. We also demonstrate how the metrics can be used to assess the sensitivity of a model to its input parameters and the impact this has on the accuracy of the resultant forecast. Finally, we offer a real-world interpretation for each metric introduced and suggest ways that they can be used to assist in cleanup operations of actual oil spills.**

***Index Terms*—Model checking, numerical simulation, oil pollution.**

I. Introduction

OIL spills at sea can have profound impacts on ecosystems, the environment, public health, the economy, and communities. Oil spill dispersion models aim to help reduce the environmental impact by predicting the trajectory of oil spilled at sea. With this information, assets (vessels and aircraft) can be directed to the likely location of the oil, and any mitigation strategies can be deployed effectively, e.g., oil booms and dispersant. Predicting the trajectory and fate of the oil is especially important in the initial stages of the spill before regular surveillance operations have been set up. As an incident progresses, modeling can also show the potential for shoreline impact, helping optimize cleanup operations. With the impact and cost of oil spills being so high, any assistance that can be provided by oil spill modeling can be significant. Oil spill modeling also has a big part to play in risk assessment and readiness planning. By simulating potential scenarios to see the impact they might have, governments, oil exploration and production companies, insurance companies, and other stakeholders can assess whether they have the necessary funds, equipment, and processes to respond.

In order for the output from oil spill models to be relied upon, it is important to be able to demonstrate that their predictions are accurate, but also that the limitations of a model are understood. Central to this is the process of model evaluation. Oil spill models are typically validated based on historical data from three main sources: drifter measurements, satellite observations, and coastal reports. Drifters provide frequent and precise location tracking data over a sustained period, usually several weeks. They are essentially simple floating GPS devices traditionally used to monitor ocean currents, but they can also be used to simulate the 2-D spatial evolution of an oil slick. However, a limitation of the drifter method for validation is that strictly it only provides an evaluation of the accuracy of advection processes in the model (e.g., the movement of oil due to wind and wave motions). Advection is one of numerous physical processes that collectively determine the fate of simulated oil. Although drifter data can be useful to help characterize forecast uncertainties within models, it is preferable to validate models against real-world data from actual oil spills.

Satellites equipped with synthetic aperture radar (SAR) are able to detect oil spills floating on the sea surface [1], [2]. SAR instruments are available on board a number of satellite constellations, e.g., COSMO-SkyMed (CSK), ERS-2, ENVISAT, RADARSAT, and, more recently, Sentinel-1. The main advantage of SAR imagery over drifter data is that it allows the output from numerical models to be compared directly against observations of real oil spills. However, because of their polar orbits, such satellites have relatively long revisit times of a day or more and so can only provide two or three images at best for a typical oil spill before they dissipate to a level where they can no longer be detected in this way. In some cases, coastal report data are also available for model validation purposes, providing a record of specific stretches of coastlines affected by beaching as a result of a particular oil spill event. However, no consensus currently exists with regard to the best way to make use of such data for model validation purposes.

Manuscript received January 6, 2021; revised June 18, 2021; accepted July 15, 2021. Date of publication September 22, 2021; date of current version January 13, 2022. This work was supported by Innovate U.K. through A4I Project 37003 and delivered by the Science and Technology Facilities Council (STFC) Hartree Centre and Riskaware Ltd. STFC and Innovate U.K. are part of U.K. Research and Innovation.

Associate Editor: T. Ross.

Chris Dearden is with the STFC Hartree Centre, WA4 4AD Warrington, U.K. (e-mail: chris.dearden@stfc.ac.uk).

Tim Culmer and Richard Brooke are with Riskaware Ltd., BS1 2NT Bristol, U.K. (e-mail: tim.culmer@riskaware.co.uk; richard.brooke@riskaware.co.uk).

Digital Object Identifier 10.1109/JOE.2021.3099562

With this in mind, the aim of this article is to identify and promote a set of unbiased objective performance metrics suitable for the validation of oil spill models based on satellite observations and coastal report data. By promoting a specific set of methods and metrics, along with a discussion of their strengths and limitations, the intention is to help standardize the validation of oil spill models across the community, leading to a more consistent evaluation of models, and to encourage a more quantitative approach to model assessment and verification.

It is important to note that models that simulate the transport and dispersion of oil can produce different types of prediction. The metrics discussed in this article focus on assessing a model's ability to predict the areas of the sea surface and coastline, which an oil spill could affect. The metrics do not attempt to assess other types of prediction directly, such as the state of the oil, although how well these factors are modeled will ultimately impact the ability to predict the areas affected. Oil spill models may predict the concentration (or thickness) of an oil slick as a function of location and time, while other models may instead assess the likelihood of oil reaching different locations, using methods like Monte Carlo analysis to take into account the uncertainties in a model and its inputs (see, e.g., [3]). In this article, we will consider the ability of the metrics to evaluate both these types of model output, which will be referred to from this point on as deterministic and probabilistic results. Deterministic results give a single estimate of the areas of the sea surface and coastline that the oil will reach, and probabilistic results provide a set of predictions, where areas are grouped by the probability that they will be affected.

The remainder of this article is arranged into the following sections. Section II comprises a literature review providing details of the validation methods currently employed within the oil spill modeling and wider modeling community and their main weaknesses. The outcomes of the literature review are then used to inform a set of newly proposed skill scores, which are introduced in Section III, followed by a summary of our recommendations in Section IV. Section V contains the results of the validation for two selected case studies, together with a brief illustration of how the metrics can be used to investigate the sensitivity of a model to its input parameters. A discussion of the real-world interpretation of the metrics is included in Section VI. Finally, Section VII concludes this article.

II. Existing Validation Methods and Techniques

We begin with a review of the literature to ascertain existing validation methods and their limitations in the context of oil spill modeling. Validation techniques are found to vary depending on the source of the observational data being considered. It should be noted, however, that although models tend to simulate oil at depth as well as oil floating on the ocean surface, current validation techniques are limited to a consideration of surface oil only, due to the lack of observational data relating to oil beneath the surface.

A. Performance Measures Based on Drifter Data

The frequency and precision of drifter measurements has provided arguably the most rigorous means to date for the assessment of oil spill dispersion models. The principal method of validation using drifter data is based on a dimensionless skill score first introduced by Liu and Weisberg [4]. The skill score SS is calculated from the separation index, S defined as

$$S = \frac{\sum_{i=1}^{N} \mathrm{SepDistance_i}}{\sum_{i=1}^{N} \mathrm{ObservedPath_i}} \tag{1}$$

where $\mathrm{SepDistance_i}$ is the Lagrangian separation distance between endpoints of simulated and observed drifters, and $\mathrm{ObservedPath_i}$ is the cumulative length of the observed trajectory at time step i. N is the number of time steps since the beginning of the simulation. The skill score SS is then defined as

$$\mathrm{SS} = 1 - \frac{S}{T}, \qquad \text{for } S < T \text{ and } \mathrm{SS} = 0, \qquad \text{for } S > T \tag{2}$$

where T is a user-selected tolerance threshold. Typically, T is set to a value of 1, meaning that the error should not exceed the magnitude of the cumulative movement.

A number of studies have since used this skill score to evaluate oil spill models for various case studies, where drifter data have been available (see, e.g., [5]–[8]), and as such, it is an established and trusted method within the community. However, drifters do not provide data from real oil spills; they are only used as a surrogate for the movement of oil at sea due to wind and ocean currents and do not account for real-world processes such as deposition, emulsification, and evaporation.

B. Satellite-Based Methods

Numerous studies have utilized satellite imagery to evaluate model simulations of real-world case studies of actual spills, albeit in a largely qualitative sense. This is likely to be attributable to the fact that satellite observations are not yet able to detect variations in the thickness (or concentration) of oil floating on the sea surface, merely the outline and spatial extent. For example, SAR imagery was used as part of the validation of the MEDSLIK-II model [6] for a case study in the Mediterranean Sea on August 6, 2008. Only two images were available: one from the ASAR sensor, which was used to initialize the model, and the other 25 h later from MODIS, which was used for validation. The model was assessed via a visual inspection of the modeled slick with the MODIS imagery superimposed, revealing that the model captured the modified shape of the slick but likely underestimated the northward movement. A similar approach has also been used in several case studies, where simulations have been performed with the General National Oceanic and Atmospheric Administration (NOAA) Operational Modeling Environment (GNOME) model. For example, Cheng *et al.* [9] used GNOME to simulate the trajectory of oil released from a leaking pipeline in the Gulf of Mexico in July 2009. The model output was validated against observations from European and Japanese satellites, both with SAR capability. Differences in location between the observed and simulated oil were attributed to uncertainties associated with ocean currents and in the diffusion coefficient. Satellite data were also used to initialize and validate output from GNOME for an oil spill accident which occurred in the Bohai Sea, China, in June 2011 [10]. In this case, the model was assessed in terms of the movement of the

simulated oil slicks (both direction and distance travelled) compared to the observed oil movement obtained from the available satellite imagery. A similar approach was adopted in the study by Cheng *et al.* [11], who used data from the CSK constellation to validate output from GNOME. Several examples were also seen in the qualitative evaluation of the Deepwater Horizon oil spill trajectory modeling using satellite imagery (see, e.g., [12]).

Huntley *et al.* [13] were the first to introduce a quantitative measure of oil spill model performance based on the evolution of 2-D area rather than drifter trajectories. They demonstrated a validation approach applied to a deterministic model simulation of the Deepwater Horizon oil spill based on the percentage of the predicted spill area contained in the observations and the percentage of the observed spill area contained in the forecast.

C. Coastal Reports

The ability of oil spill forecast models to accurately predict oil–shoreline interactions (also known as "beaching") is an important consideration due to the severe environmental, societal, and economic impacts. However, few studies have focused on the ability of numerical models to predict the risk of beaching in vulnerable coastal zones. Weisberg *et al.* [14] supplemented observations with results from numerical modeling simulations to identify the key mechanisms responsible for the beaching of oil associated with the Deepwater Horizon oil spill. Like SAR imagery, coastal report data do not contain information relating to oil concentration and are simply used to highlight those stretches of coastline affected by beaching.

The Lebanese oil pollution crisis of July 2006 resulted in significant beaching along the Lebanese and Syrian coasts, as reported by several sources. In the hindcast study of Coppini *et al.* [15], these coastal reports were collated and combined with satellite observations to validate the representation of oil–shoreline interactions as simulated by the original MEDSLIK model. The results indicate that MEDSLIK was able to reproduce the general timing and transport of oil northwards along the Lebanese and Syrian coasts, with the model predicting almost 80% of oil would be permanently landed along the shoreline. However, the true quantities of oil that reached the shoreline were not clearly reported, which restricted the validation to a qualitative assessment based on the spatial extent of the coastal impact. Later, Samaras *et al.* [16] demonstrated how the representation of beaching in models could be improved through consideration of an approach based on the Oil Holding Capacity to estimate coastal oil concentrations. As in [15], the validation was based on a visual inspection of 2-D maps comparing the extent of the simulated beaching against the coastal reports. Following on from the numerical simulations of the Bohai Sea oil incident performed by Xu *et al.* [10], an additional study was later conducted by Xu *et al.* [17] focusing on the prediction of oil spill beaching along the Bohai coast. The results showed that ocean currents were most likely to have been responsible for carrying the oil northeast along the coastal region. The areas associated with a high risk of beaching as predicted by the model were also verified against *in situ* coastal reports in the form of photographic evidence obtained from the State Oceanic Administration of China.

Based on the existing literature, it is apparent that the skill score associated with drifter data has provided the most robust and quantitative metric for the assessment of oil spill models to date. Satellite and coastal data have so far mainly been used to provide a qualitative and somewhat subjective assessment of model performance. A more quantitative and objective validation method, building on the approach taken by Huntley *et al.* [13], would represent an important step forward.

D. Methods Used in the Atmospheric Dispersion Modeling Community

In this section, we extend the literature review to consider the metrics employed in the atmospheric dispersion modeling community, and how such metrics could potentially be adapted for use in oil spill validation studies, where satellite observations and/or coastal report data are available. For reasons previously stated, we limit our review to a consideration of threshold-based measures rather than concentration-based measures.

1) Figure of Merit in Space (FMS): The FMS is a statistical coefficient defined as the ratio of the intersection of the observed and predicted areas (A_{OB} and A_{PR}, respectively) to the union of the observed and predicted areas at a fixed time instance and above a defined threshold level. Mathematically, this is written as

$$\mathrm{FMS} = \frac{A_{\mathrm{PR}} \cap A_{\mathrm{OB}}}{A_{\mathrm{PR}} \cup A_{\mathrm{OB}}}. \tag{3}$$

The FMS has been used to compare the predictions from several atmospheric dispersion models (see, e.g., [18]). The higher the FMS value, the better the agreement between the model and observations. One of the drawbacks of the FMS is that in taking the union, both the observed and predicted areas are weighted equally. This means that the FMS cannot distinguish between regions of under- and overprediction relative to the observations.

2) Two-Dimensional Measure of Effectiveness (MOE): The 2-D MOE, introduced by Warner *et al.* [19], attempts to address this limitation of the FMS by calculating the area of overlap with respect to both the observed and predicted areas as separate components. Huntley *et al.* [13] used a similar method to evaluate the Deepwater Horizon oil spill modeling.

The MOE is defined as

$$\mathrm{MOE} = (x, y) = \left(\frac{A_{\mathrm{OV}}}{A_{\mathrm{OB}}}, \frac{A_{\mathrm{OV}}}{A_{\mathrm{PR}}} \right) \tag{4}$$

where A_{OV} is the area of overlap between observations and model prediction.

Alternatively, the MOE can be written as

$$\begin{aligned} \mathrm{MOE} = (x, y) &= \left(\frac{A_{\mathrm{OV}}}{A_{\mathrm{OB}}}, \frac{A_{\mathrm{OV}}}{A_{\mathrm{PR}}} \right) \\ &= \left(\frac{A_{\mathrm{OB}} - A_{\mathrm{FN}}}{A_{\mathrm{OB}}}, \frac{A_{\mathrm{PR}} - A_{\mathrm{FP}}}{A_{\mathrm{PR}}} \right) \\ &= \left(1 - \frac{A_{\mathrm{FN}}}{A_{\mathrm{OB}}}, 1 - \frac{A_{\mathrm{FP}}}{A_{\mathrm{PR}}} \right) \end{aligned} \tag{5}$$

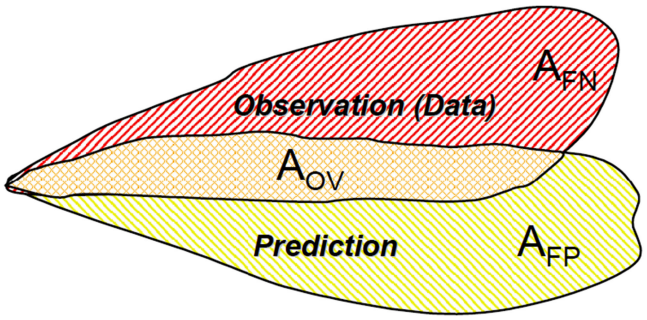


Fig. 1. Conceptual view of the area of overlap (A_{OV}), false negative (A_{FN}), and false positive (A_{FP}) that are used to construct the user-oriented MOE. Adapted from [19].

where A_{FN} is the area of false negative (the region where a hazard is observed but not predicted) and A_{FP} is the area of false positive (the region where a hazard is predicted but not observed).

These regions are illustrated schematically in Fig. 1. For many applications, false positives are more acceptable to the user than false negatives. For example, a forecast that overpredicts the length of coastline impacted by a particular oil spill is preferable to one that fails to predict any coastal impact at all.

By weighting the area of overlap against both the predicted area and observed areas separately, the 2-D MOE takes into account the predicted location of a hazard as well as the size and shape. A perfect MOE score of $(x, y) = (1, 1)$ would indicate complete overlap between the model and observations. A value of x close to 1 and y close to 0 would indicate that the model has a significant false positive region (i.e., the model would be overestimating the extent of the oil spill area). Conversely, a low x value together with a high y value would indicate the dominance of the false negative region, caused by underestimation of the oil spill area. These characteristics of the 2-D MOE space are neatly captured via a simple scatter plot, such as the one shown in Fig. 2. From the MOE equation, it can be seen that cases where $x = y$ imply equal areas of the predicted and observed regions (i.e., $A_{\mathrm{PR}} = A_{\mathrm{OB}}$), even if the locations differ. This is represented by the purple diagonal line in Fig. 2. As one traverses this diagonal line from (0,0) toward $(1, 1)$, the fraction of overlap region between the predicted and observed areas increases.

Numerous studies have applied the 2-D MOE to quantify the performance of atmospheric dispersion models against observations. Warner *et al.* [19] applied the 2-D MOE to both deterministic and probabilistic model results from an atmospheric dispersion model. In the probabilistic case, this was achieved by calculating the MOE separately for each model probability contour. Rolph *et al.* [20] and Stein *et al.* [21] applied the 2-D MOE in their evaluation of the NOAA's operational smoke forecasting system, validating the model output using satellite detections of smoke plumes for selected case studies. Warner *et al.* [22]–[24] used the 2-D MOE to evaluate dispersion models against observations collected during several field experiments. Furthermore, it has also been used to compare simulated plume extents from two different models in the absence of observational data. For example, the study of Pullen *et al.* [25] performed numerical simulations of a hypothetical airborne agent release in major urban areas and used the 2-D MOE to compare the results from a Gaussian puff model relative to those from a building-resolving computational fluid dynamics model. However, to our knowledge, the 2-D MOE has not been used for oil spill model validation before now.

3) Fractions Skill Score (FSS): The FSS described in [26] and [27] is an established method used in the atmospheric sciences to verify the spatial accuracy of high-resolution precipitation forecasts relative to radar observations on a common 2-D grid. Rather than evaluating the model at specific point locations, the FSS considers different-sized sampling areas called neighborhoods, within which both the forecast and radar rainfall fractions are computed. In doing so, the FSS can provide valuable insight into how the skill of a model varies with spatial scale. Indeed, this method has been recently adopted and applied to oil spill forecast assessment by Simecek-Beatty and Lehr [28].

The FSS was originally designed to identify the spatial scales at which high-resolution deterministic models perform most reliably. In comparison with the 2-D MOE, the FSS by its nature provides more insight into the spatial accuracy of a model, and this can be particularly useful for the validation of relatively small oil spills. However, the 2-D MOE can be more readily extended to include validation of probabilistic model output as well as coastal validation and is, therefore, a more appropriate choice for the needs of the present study.

III. Details of Newly Proposed Metrics

The principle aim of this article is to identify suitable performance measures for the validation of oil spill dispersion models, based on satellite imagery and *in situ* coastal reports. The literature review presented in the previous section revealed that the main quantitative performance measure currently applied to oil spill models is based on drifter measurements, and that validation studies based on satellite imagery and coastal reports have to date been primarily qualitative in nature. By also considering performance measures used by the atmospheric dispersion modeling community, a suitable metric based on the region of overlap between observations and model prediction has been identified (the 2-D MOE). However, for simulations that exhibit little or no overlap with the observations, it is necessary to introduce new metrics that can complement the 2-D MOE as part of the overall model evaluation.

With this in mind, we have taken the concept of the Skill Score [4] from (2), used ostensibly in relation to drifter measurements, and adapted it to work with quantities that are readily determined from satellite imagery. Specifically, we propose the introduction of two new skill scores based on the centroid and area of the observed and simulated oil spill geometries, which are thus described.

We begin by defining a centroid displacement index C_I as

$$C_I = \frac{\Delta x}{L_{\mathrm{OBS}}} \tag{6}$$

where Δx is the distance between the geometric centers (centroids) of the observed oil spill shape and the predicted oil spill shape at a given time instance, and L_{OBS} is the length scale of

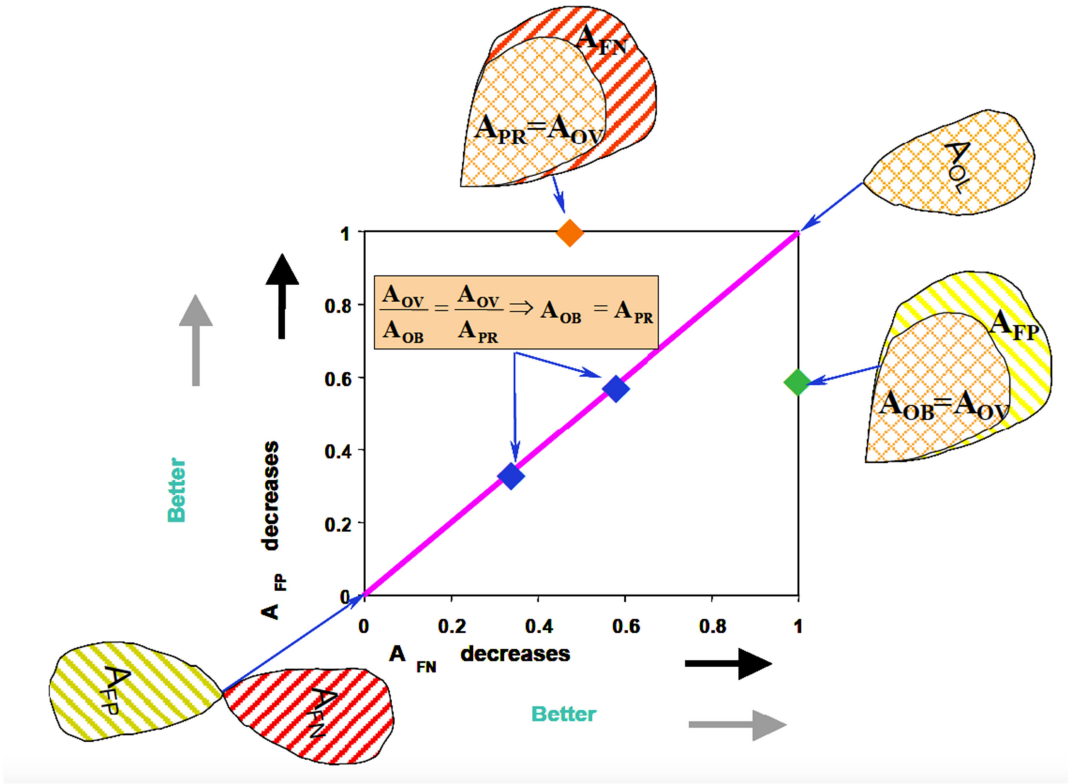


Fig. 2. Key characteristics of the 2-D MOE space. The purple diagonal line where $x = y$ corresponds to situations where the predicted and observed areas are equal, regardless of location. The MOE space above the diagonal corresponds to situations where $A_{\text{FN}} > A_{\text{FP}}$, and below the diagonal corresponds to $A_{\text{FP}} > A_{\text{FN}}$. Adapted from [19].

the observed oil spill area. L_{OBS} is defined as the distance along the diagonal of a bounding box enclosing the observed oil spill region. C_I is then simply a measure of the absolute error in the predicted centroid location, normalized by the length scale of the observed oil spill.

The centroid skill score C_{SS} can then be defined in a manner analogous to the skill score SS from (2)

$$C_{\text{SS}} = 1 - \frac{C_I}{C_{\text{thr}}}, \qquad \text{for } C_I < C_{\text{thr}}$$
$$C_{\text{SS}} = 0, \qquad \text{for } C_I > C_{\text{thr}} \tag{7}$$

where C_{thr} is a user-selected tolerance threshold. A C_{thr} value of 1 would mean that, for the model to have any skill, the distance between the locations of the observed and predicted centroids must not exceed the magnitude of the observed length scale. Fig. 3 provides an illustration of the quantities involved in the calculation of C_{SS}.

Next, based on the comparison of predicted and observed oil spill areas as in [13], we introduce the area index A_I as

$$A_I = \frac{|A_{\text{PR}} - A_{\text{OB}}|}{A_{\text{OB}}} \tag{8}$$

which is simply the magnitude of the difference between the predicted oil spill area and the observed oil spill area at a given time instance, normalized by the observed area.

The area skill score A_{SS} is then defined as

$$A_{\text{SS}} = 1 - \frac{A_I}{A_{\text{thr}}}, \qquad \text{for } A_I < A_{\text{thr}}$$
$$A_{\text{SS}} = 0, \qquad \text{for } A_I > A_{\text{thr}} \tag{9}$$

where A_{thr} is a user-selected tolerance threshold. An A_{thr} value of 1 would mean that the error in predicted area must not exceed the magnitude of the observed oil spill area; otherwise, $A_{\text{SS}} = 0$.

IV. Summary of Recommended Metrics

Moving forward, we recommend three specific performance measures to be used in the validation of oil spill dispersion models: the existing 2-D MOE, and the newly proposed centroid skill score and area skill score. Each metric is summarized below, along with a consideration of their strengths and weaknesses, and the scenarios in which they are most likely to be best suited.

A. Two-Dimensional MOE

The 2-D MOE provides a measure of performance based on the extent to which the predicted area overlaps with the observations at a given time instance, and is an established metric within the atmospheric dispersion modeling community. When applied to the output from oil spill dispersion models, it is likely to be most informative for large spills, where the chances of significant overlap between the model prediction and the observations are greatest. A particular strength of the

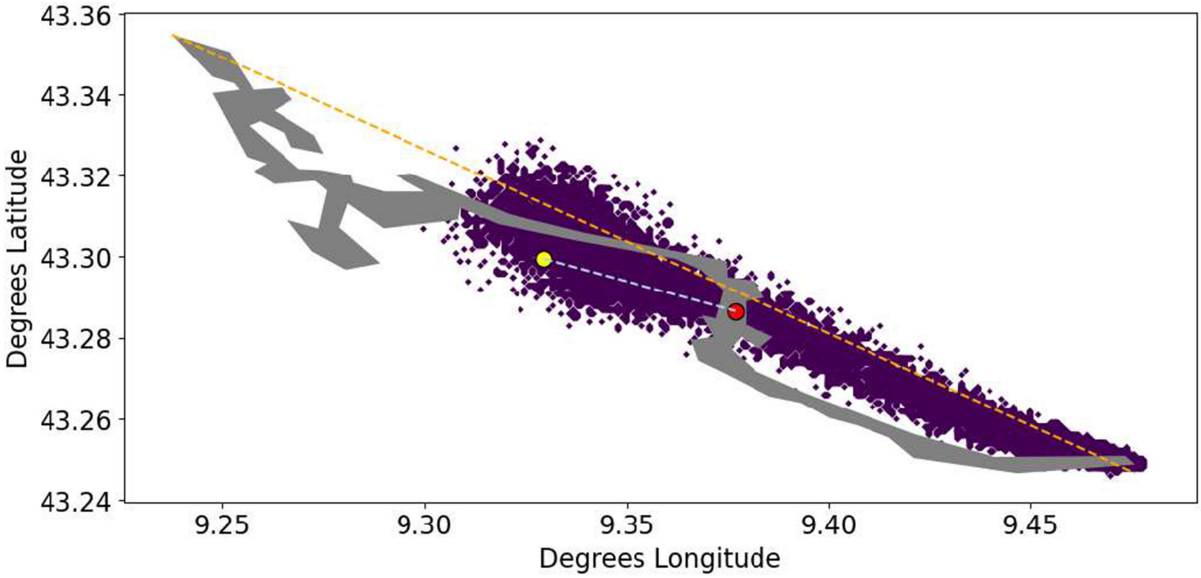


Fig. 3. Example plot showing the quantities involved in the calculation of the proposed centroid skill score, for an observed oil spill region (gray) and deterministic model prediction region (dark purple). The distance between the observed centroid (yellow dot) and predicted centroid (red dot) is denoted by the dashed light blue line, and the length scale of the observations is denoted by the dashed orange line. In this example, the error in predicted centroid location is less than that of the length scale of the observations, and therefore, the model would be deemed to have some "skill," i.e., a C_{SS} value > 0.

2-D MOE is that it provides a quantitative measure of both false negative and false positive regions. As well as satellite imagery, the 2-D MOE also lends itself nicely to coastal report data, where the area of overlap would simply be replaced by the length of overlap. A weakness of the 2-D MOE is that it is based solely on the extent of spatial overlap between model and observations and, therefore, does not provide direct information on the size and shape of simulated spills.

There is also the risk with the 2-D MOE that the area/length of overlap could be small, particularly for deterministic model output, and hence lead to low or zero MOE values. This on its own would not necessarily indicate a poor forecast, since cases with very little or no overlap could be caused by a minor spatial offset despite the forecast accurately capturing the shape and size of the observed spill. Thus, in such cases, we recommend complementing the 2-D MOE with the following additional skill scores.

B. Centroid Skill Score

The centroid skill score is designed to provide an indication of how close the predicted oil spill is to the observed oil spill region. This metric is suitable for validation of model output against satellite imagery. By taking into account the distance between the predicted and observed centroid locations, it provides a measure of the proximity of the predicted region to the observed region, within some user-defined tolerance threshold. A centroid skill score of 1 would correspond to perfect colocation of the observed and predicted oil spill centroids, whereas 0 would indicate no "skill" at all (i.e., the error in predicted centroid location would exceed the user-defined tolerance threshold). For consistency with other studies that make use of a tolerance threshold (see, e.g., [7]), we set C_{thr} to a value of 1, such that if the distance between the locations of the observed and predicted centroids exceeds the magnitude of the observed length scale, the model is deemed to have no "skill," i.e., $C_{\mathrm{SS}} = 0$.

A weakness of the centroid skill score is that it does not provide information on the direction of the predicted centroid with reference to the observed centroid location. Furthermore, it should only be used for cases, where the complete extent of the oil slick is detectable within the swath of the SAR imagery. Otherwise, it could produce misleading results due to the centroid of the observed slick being offset by the cropped imagery.

C. Area Skill Score

The area skill score is designed to address a weakness of the centroid skill score by providing a measure of the size of the predicted oil spill region relative to the observed region. An area skill score of 1 would indicate perfect agreement, i.e., the predicted and observed areas are exactly the same, whereas a score of 0 would indicate that the error in predicted area exceeds some user-defined tolerance value (typically equal to the magnitude of the observed oil spill area). In line with the centroid skill score, we use a threshold A_{thr} value of 1, so that in order for a model to have any "skill," the error in the predicted area must not exceed the magnitude of the observed oil spill area.

It is recommended that the centroid and area skill scores are evaluated together, as they provide information that is complementary to the other. A weakness of the area skill score is that it does not take into account any differences in the shape of the predicted spill relative to the observed spill. Like the centroid skill score, its use should be restricted to cases, where the full extent of the oil slick lies within the detection range of the observations.

V. Model Validation

To demonstrate the use of our recommended metrics and provide an illustration of their strengths and limitations, we apply them to a specific oil spill model developed by Riskaware Ltd. The model treats oil slicks as a set of particles using a method known as Lagrangian modeling [6], [29], common in fluid dynamics, before converting the particle locations to a series of contours based on concentration thresholds. The model is configured to produce both a deterministic and a probabilistic forecast. In the case of the latter, uncertainty in the ocean current forecast is taken into account using a Monte-Carlo-based method (as in [3]) to provide the likelihood of oil affecting different locations. Since the focus of this article is on the metrics themselves, we reserve the full technical details of the Riskaware model for a separate publication.

The metrics are applied to the model for two separate test cases. First, we use satellite data to validate the model, making use of images acquired by the European Space Agency's Sentinel-1 mission of an oil spill off the French Island of Corsica in October 2018, which occurred when two ships collided nearby. Second, we use coastal reports from the oil spill disaster resulting from the grounding of the Sea Empress oil tanker near Milford Haven in Wales, U.K. in February 1996. Both case studies were simulated using metocean data from the E.U. Copernicus Marine Service Information [30]–[32].

A. Corsica Case Study

We begin with an assessment of the probabilistic model results for the Corsica test case using the 2-D MOE performance measure. The plots in the left-hand side of Fig. 4 show maps of the detected oil spill at three different validation times, together with the corresponding probabilistic modeling results, simulated as a continuous release of bunker fuel from the ship's location starting from October 7, 2018, 0503 UTC. The results are also expressed in the form of 2-D MOE scatter diagrams (shown in the right-hand side column of Fig. 4) to quantify the areas of false negative and false positive as a function of probability level. The maps reveal that the lowest probability contour ($>1\%$) covers a region that encompasses very nearly the full extent of the observed oil spill areas, such that there is almost no false negative at this probability level. The significant false positive region is not surprising given the probabilistic nature of the forecast. The degree of false negative also tends to increase with increasing probability level, but again this is to be expected since the higher probability contours do not necessarily indicate that the observations will be confined to these specific regions. Each of the 2-D scatter plots reveal that the region of false positive tends to reach a minimum somewhere above the $>35\%$ probability contour.

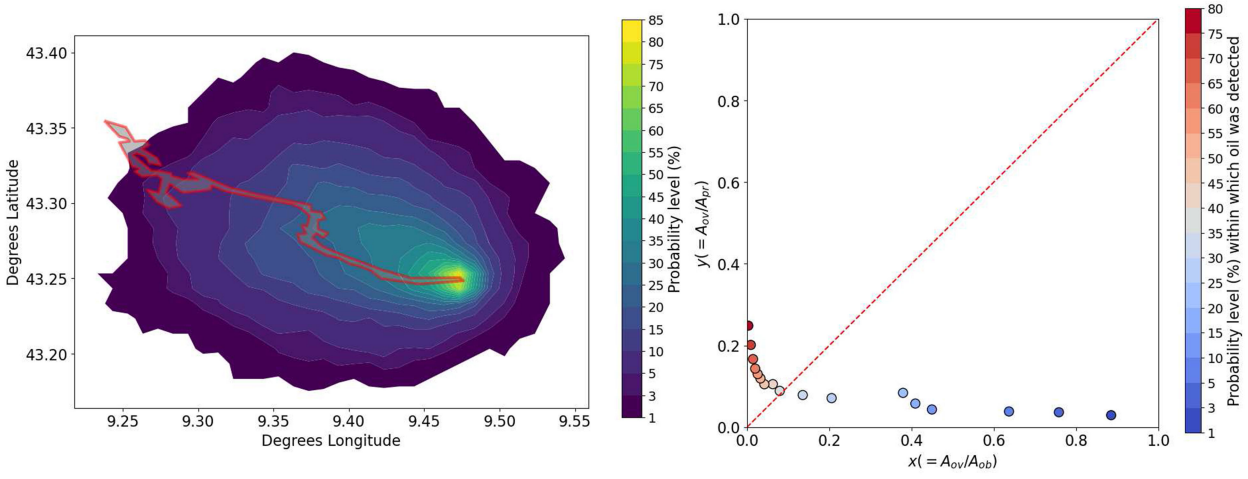
The 2-D MOE is also applied to the deterministic model output, and the results are summarized in Fig. 5. For each of the spatial maps shown in the left-hand side column of Fig. 5, the region of overlap between the observed and predicted areas is highlighted in red. In each of these plots, the predicted spill compares favorably with the observed oil with regard to the general direction of spread and the growth in size. However, subtle differences in shape and/or orientation result in limited overlap between the two; this is particularly evident in Fig. 5(c). Consequently, the values of the 2-D MOE components for the deterministic Corsica case are relatively low (as shown in the right-hand side column of Fig. 5), consistent with the fact that all three comparisons exhibit significant regions of false negative and false positive.

The close proximity of the deterministic spill and observed spills, despite the fact that the overlap regions are small, suggests that the 2-D MOE on its own does not provide a particularly informative assessment of the deterministic model performance for this case. Thus, it is prudent to also consider results from the skill scores based on centroid location and area magnitude, the results of which are shown in Fig. 6. In each case, the distance between the centroid locations (as indicated by the blue dashed lines in the left-hand side column of plots) is comfortably less than the length scale of the observations (dashed orange lines), resulting in relatively high centroid skill score values in the range 0.7–0.85. To put this into context, a centroid skill score of 0.5 for a circular shaped spill would indicate that the centroid of the modeled slick is located at the edge of the detected oil. In real-world terms, this gives a good indication that the locations of the modeled oil spills are accurate enough that they would have helped accident responders locate the slicks during this particular incident. The area skill score for the comparison at 0527 UTC on October 8, 2018 has a low value of 0.07 [see Fig. 6(a)] since the predicted spill is just under a factor of 2 larger than the actual area calculated from the satellite detection. For the other time comparisons shown in Fig. 6, the area skill scores are very high [0.94 and 0.98 in Fig. 6(b) and (c), respectively], indicating that the predicted and observed areas in these instances were almost the same. There could be a number of reasons for the initial low value of the area skill score, but perhaps given the much improved scores in the later comparisons, the release rate may not be constant, with the assumed rate being an overestimate of the initial rate.

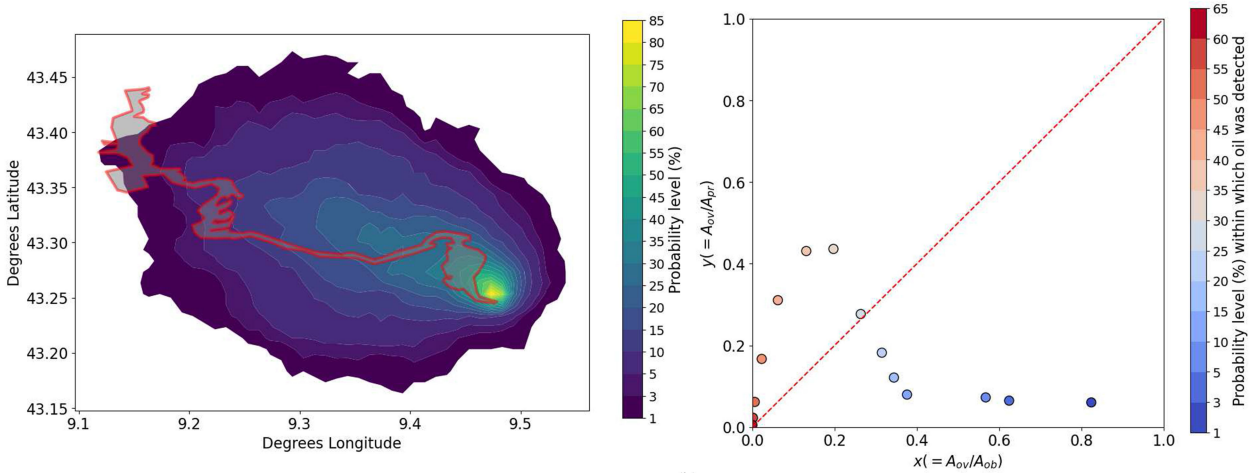
B. Sea Empress Case Study

Fig. 7 shows both the coastal report data for the Sea Empress test case, alongside the deterministic and probabilistic model results to facilitate a visual comparison. In the deterministic case [see Fig. 7(a)], it is apparent that the model underestimates the extent of beaching. Where the model does predict the deposition of oil onto the coast, this is largely consistent with the observations, although there also appears to be a stretch of coastline to the east, where oil was predicted but not recorded in the local reports. The probabilistic output [see Fig. 7(b)] appears to match the coastal report data well, with the $>50\%$ probability regions largely confined to those sections of coastline where beaching was reported, while the lowest probabilities ($<20\%$) typically occur in the regions unaffected by beaching.

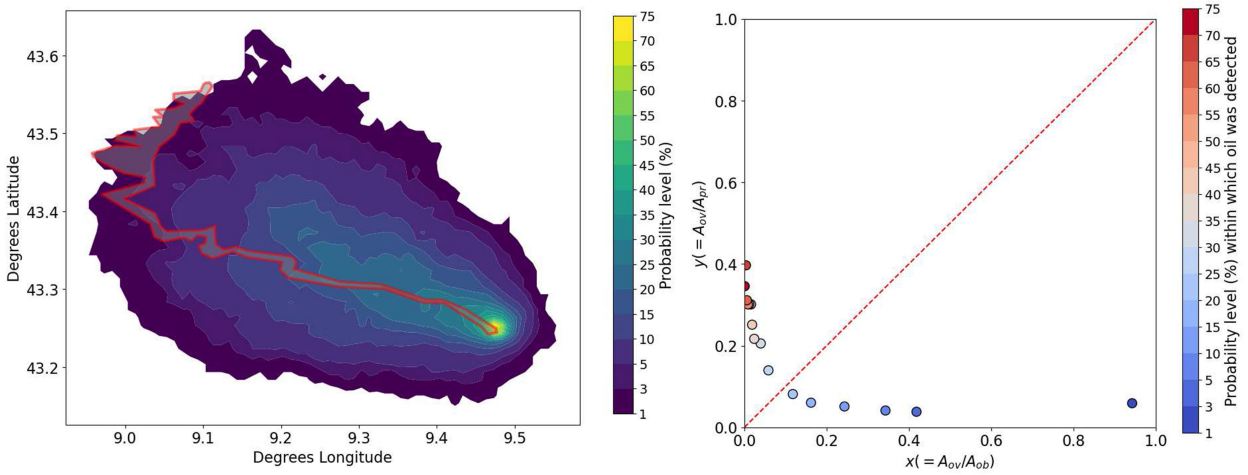
The 2-D MOE diagram in Fig. 8(a) confirms the considerable false negative extent in the deterministic case, with an x-component value of 0.35. However, the y-component is perhaps larger than one might expect from simply examining Fig. 7(a). This can be explained by the fact that the stretch of coastline,



(a)

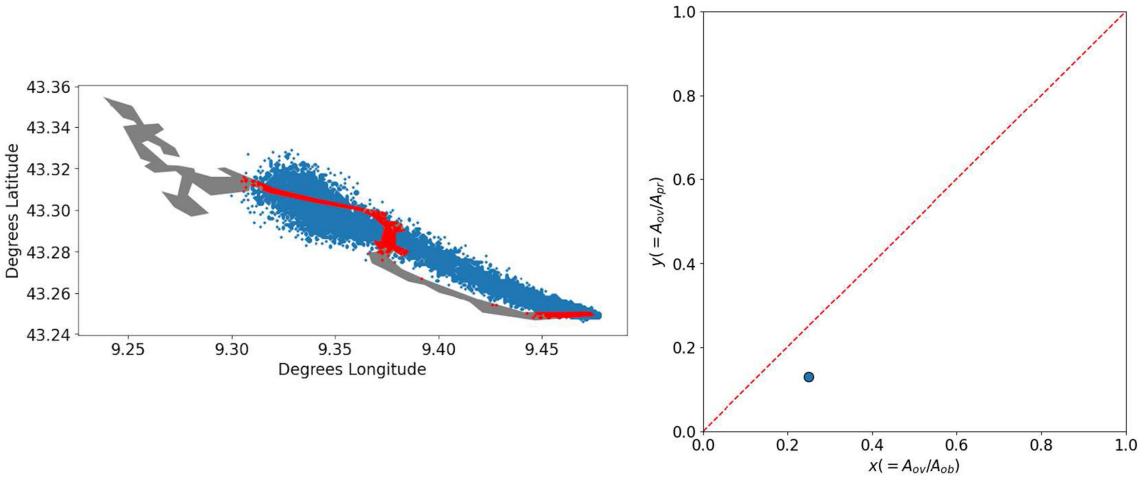


(b)

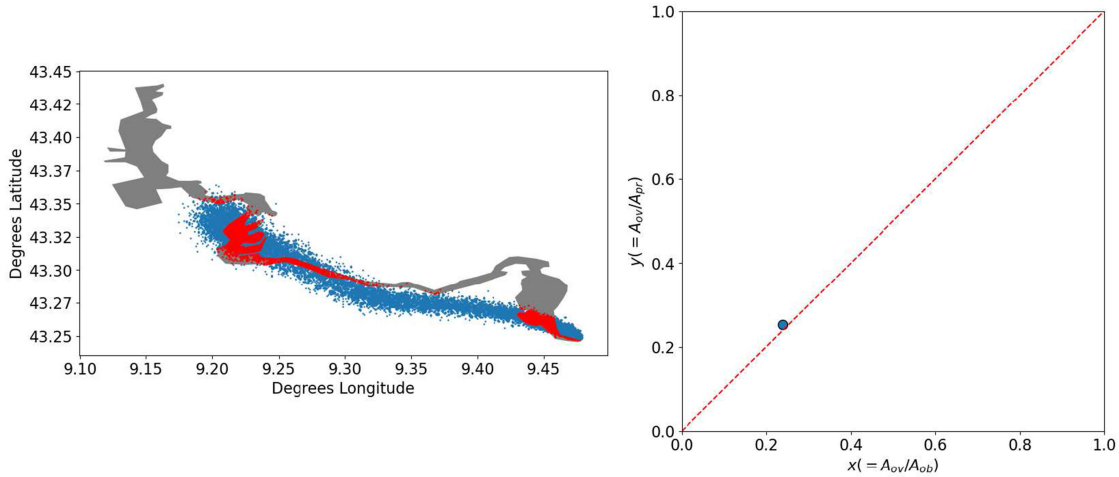


(c)

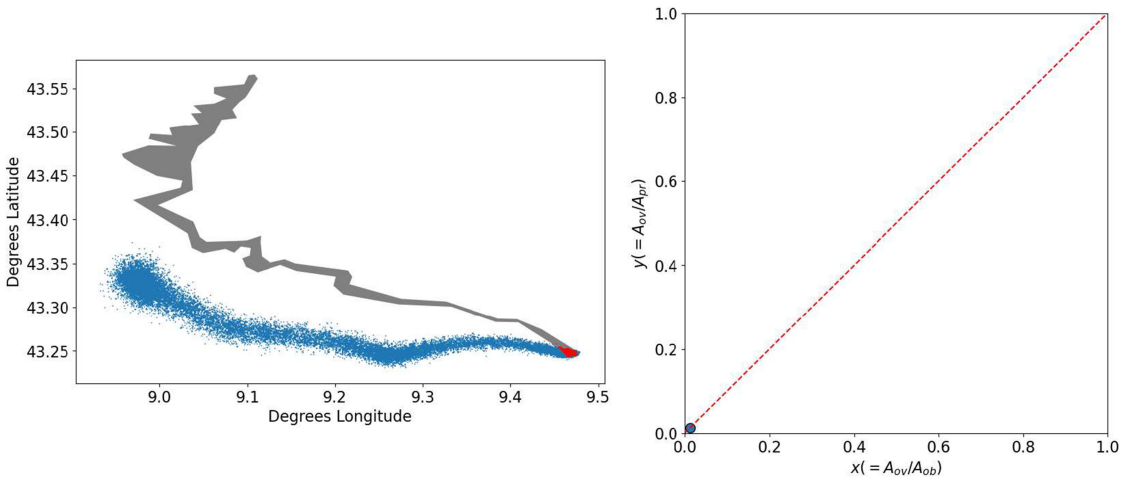
Fig. 4. Probabilistic results from the 2018 Corsica test case, valid at (a) 05:27 UTC October 8; (b) 17:22 UTC October 8; and (c) 17:14 UTC October 9. Maps on the left show the probabilistic model output (blue to yellow contours) with the observed slick overlaid in gray with a red outline. Corresponding 2-D MOE results are shown on the right, where the colors represent the different probability levels from the model for which there was overlap with the observations.



(a)



(b)



(c)

Fig. 5. Deterministic results from the simulation of the Corsica test case, for the same validation times as in Fig. 4. The maps on the left show the extent of the oil spill in the observations (gray) and model prediction (blue), with overlapping regions highlighted in red. The corresponding 2-D MOE results are shown on the right. (a) October 8, 2018, 0527 UTC. (b) October 8, 2018, 1722 UTC. (c) October 9, 2018, 1714 UTC.

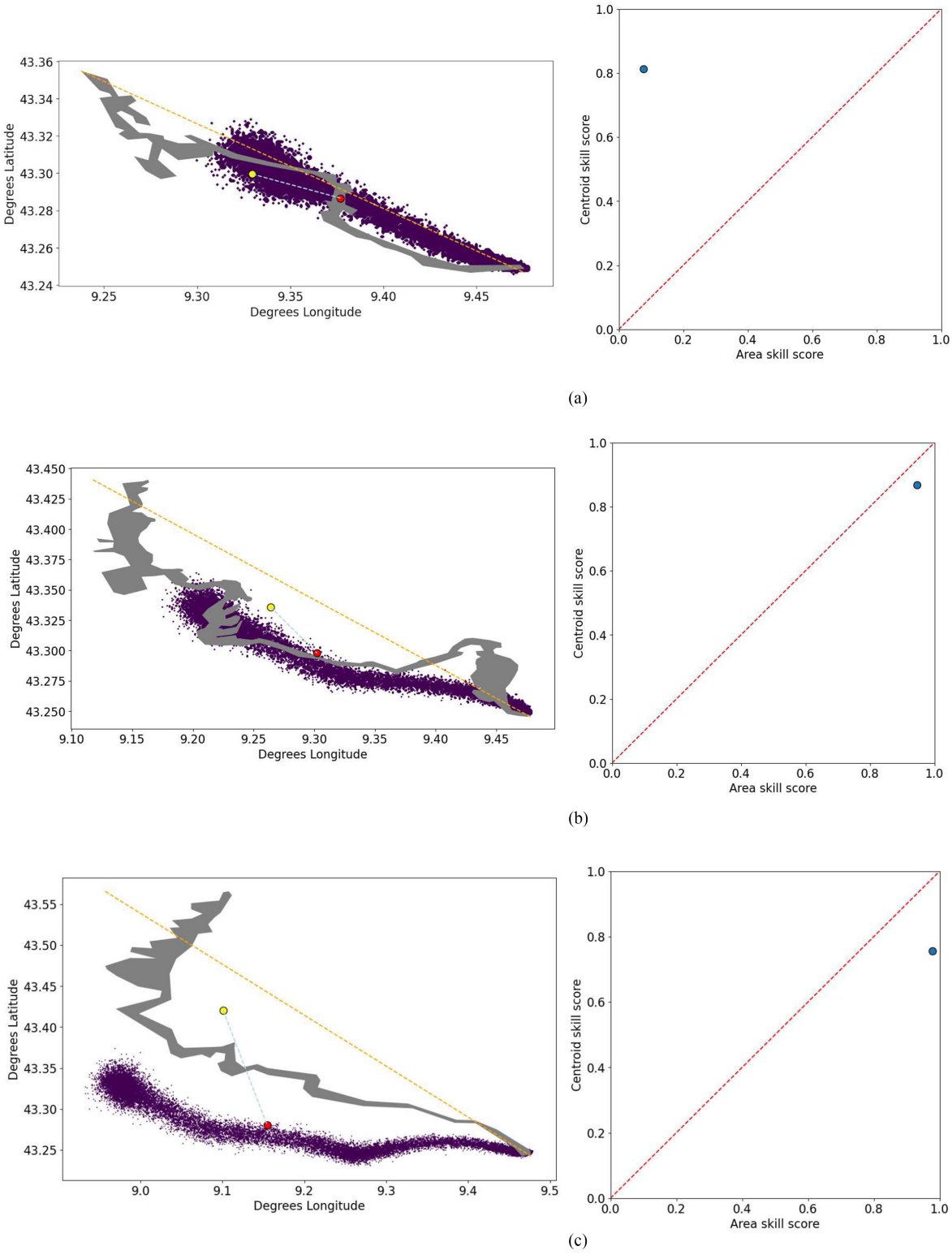


Fig. 6. Area and centroid skill score results based on the deterministic model output from the Corsica test case, for the same validation times as in Fig. 4. The maps on the left show the extent of the observed (gray) and predicted (purple) oil spill areas. The length scales of the observed and predicted oil spills, along with their corresponding centroid locations, are shown in a manner consistent with Fig. 3. The scatter plots on the right reveal the corresponding area and centroid skill scores. (a) October 8, 2018, 0527 UTC. (b) October 8, 2018, 1722 UTC. (c) October 9, 2018, 1714 UTC.

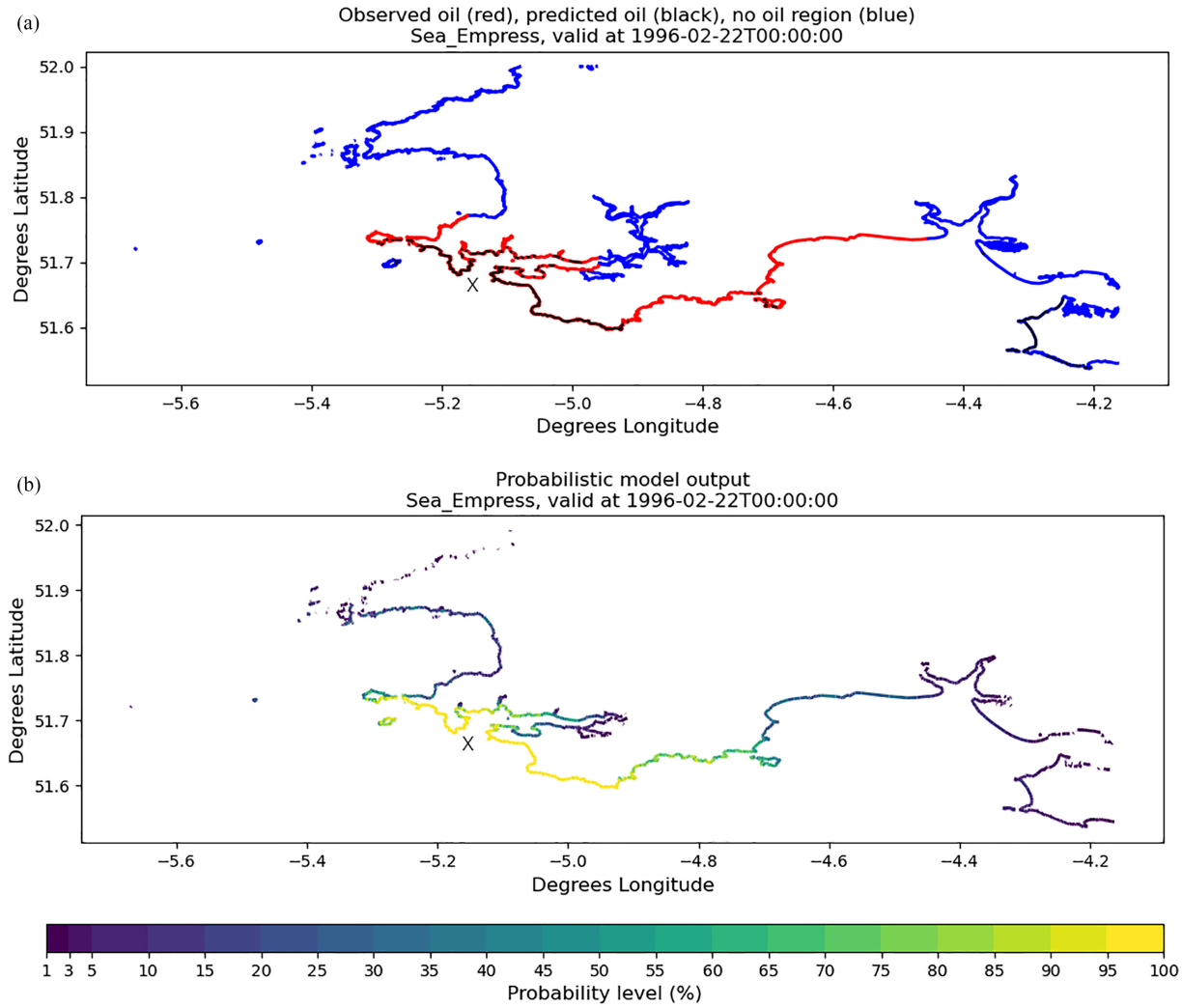


Fig. 7. Maps showing the extent of beaching in relation to the Sea Empress oil spill from 1996. (a) Confirmed beaching from coastal report data (red line), unaffected coastlines from coastal report data (blue line), deterministic model prediction of beached oil (black line). (b) Probabilistic model output, shaded according to probability level. In each image, the "X" marks the release location of the oil.

along which much of the simulated oil is confined, is longer than it looks at first glance due to the complex local topography of this region. The 2-D MOE diagram for the probabilistic output [see Fig. 8(b)] reveals that the x-component values are high between the 1% and 20% probability level, not dropping below 0.8 in this range. The y-component values improve with increasing probability, with no false positive at the highest probability level. Arguably, the model performance peaks around the 20% probability level, where a balance between the false negative and false positive extents is achieved.

It is interesting to note that the 2-D MOE scores are considerably higher for the Sea Empress case compared to the Corsica case. This may be a reflection of the fact that oil washed up on shore has fewer degrees of freedom available to it than oil at sea, where the oil is free to move in three dimensions rather than being constrained to the profile of the coastline.

C. Calibration of Model Parameters

Here, we demonstrate the use of the validation metrics as a tool to investigate the sensitivity of a model to its input parameters and determine optimal values for these. For illustrative purposes, we focus our attention on the Corsica test case. Oil spill models typically include the effects of small-scale turbulent ocean processes such as eddy currents, which are not resolved in the input data, by adding a stochastic element to the oil motion. The horizontal component of this is controlled by the horizontal diffusivity parameter. Fig. 9 shows how changing the horizontal diffusivity parameter affects the model performance for the Corsica test case. As expected, the horizontal diffusivity has a strong influence on the area skill score. For the Corsica test case, which has been simulated as a continuous release of oil, the skill score is maximized when the horizontal diffusivity equals 1.2 $\mathrm{m}^2 \cdot \mathrm{s}^{-1}$.

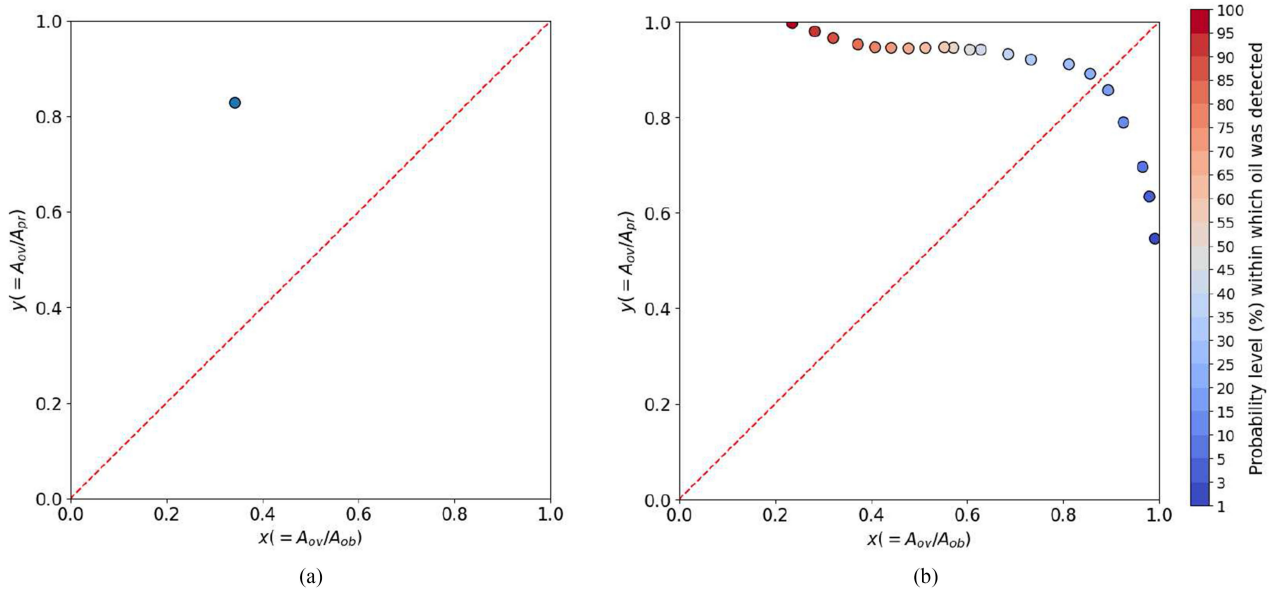


Fig. 8. (a) Two-dimensional MOE space diagram for the Sea Empress deterministic model prediction, valid on February 22, 1996, 00:00 UTC. (b) Two-dimensional MOE space diagram for the Sea Empress probabilistic model prediction, also valid on February 22, 1996, 00:00 UTC.

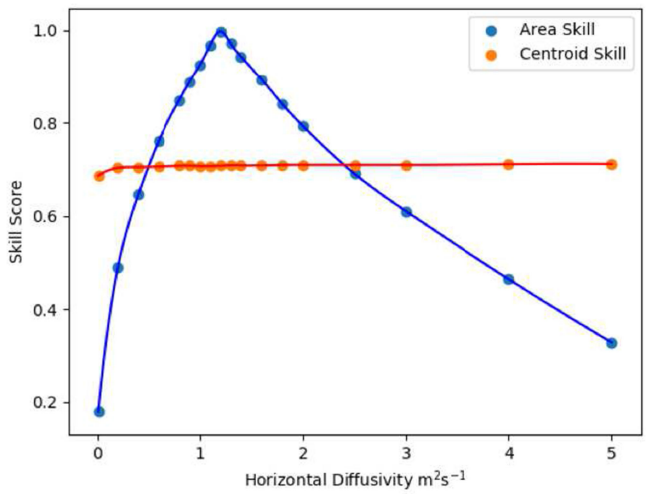


Fig. 9. Plot showing how the area skill score and centroid skill score vary as a function of the horizontal diffusivity parameter for the Corsica test case scenario.

VI. Discussion

A. Analysis of Metric Performance

The validation metrics have performed broadly as expected based on the recommended use cases outlined in Section IV. The centroid and area skill scores provide a good measure of the performance for the deterministic model output. However, the 2-D MOE method based on the degree of overlap was shown to be unforgiving to small deviations in the direction of the spill, and consequently, it did not provide an informative assessment of the deterministic Corsica simulation. However, the 2-D MOE did provide a more useful assessment of the probabilistic model output, where it was used to help identify the most informative probability levels to output from the model.

One possible weakness of the area skill score is its inability to distinguish between an over- or underestimation of the oil spill area; however, this can still be determined from a manual assessment of the spill outlines. Arguably, the centroid skill score provides the most useful insight into the model performance since the predicted center of the slick is likely to be where responders will head toward when trying to initially locate an oil spill. With this in mind, we now discuss in more detail how each of the metrics can be more generally interpreted beyond the two cases considered thus far.

B. Real-World Interpretation of the Metrics

The values produced by the metrics presented in this article allow the performance of different models and model parameters to be objectively compared. They can also help to show whether a model is fit for its intended purpose. In this section, we relate the metric scores to a model's ability to benefit an oil spill response. In reality, this will also depend on factors such as timeliness and effective communication of results; however, here, we only attempt to use the metrics to assess the accuracy of the model output. The purpose is to provide an assessment of how the metrics could be used to assist responders who are responsible for deploying a cleanup operation following an oil spill detection. With a sufficiently large number of retrospective case studies, the metrics provide an appraisal of how well the model should be expected to perform in a future spill. This information may be useful in an active spill response when planning the deployment of resources. We consider common use cases and determine what insight the metrics can give about the model's effectiveness in assisting with each task.

TABLE I
SUMMARY OF THE USES FOR THE DIFFERENT METRICS IN ASSESSING A MODEL'S ABILITY TO ASSIST WITH AN OIL SPILL RESPONSE

Metric	Contours	Coastlines
2-D MOE x-value	Assessment of the predicted extent of the oil (and probability of locating oil at sea)	Amount of impacted coastline reached by clean-up teams
2-D MOE y-value	Efficiency of finding oil at sea	Fraction of clean-up teams who find oil where they are deployed
Area Skill Score	Estimating the area of a spill	N/A
Centroid Skill Score	Probability of locating oil at sea	N/A

1) Locating the Oil at Sea Using the Deterministic Model: Often, the first task in responding to an oil spill will be to locate the oil in the water, which can be particularly difficult when the spill happens in a remote area a long way from land. Models can help to locate oil by forecasting the movement of a slick forward from its last known location (for example, a satellite or aerial observation) to where it will be when the responders reach it. Assuming that the responders head to the centroid of the modeled oil in order to find it, then we can look for correlation between the metric scores and them successfully finding oil. In this context, we can define the success of a model according to its ability to produce a predicted centroid within 500 m of an actual oil spill (where 500 m is an estimate for the maximum distance oil can be spotted from a ship). We applied these criteria to a sample set of nine test cases, where satellite data were available for validation, and found that the model was successful in locating oil in six out of the nine test cases. For each of the six successful locations, the centroid skill score never fell below a value of 0.75, and the corresponding 2-D MOE x-values were all above 0.1. Despite the small sample size, this suggests a correlation between these metrics and finding oil in real-world situations.

If responders look for oil at a randomized location within a contour in the model results, instead of at the centroid, then the 2-D MOE y-value is also significant. The y-value is a measure of the degree of false positive in the results, and for this use case, it represents the probability that oil would be found at any given location. Therefore, it can be thought of as a measure of the efficiency with which responders could locate oil. A higher y-value would mean that responders would theoretically need less attempts to initially locate the spill.

2) Calculating the Area of an Oil Spill: Knowing the area of a slick on the surface of the water resulting from a spill of oil is useful to responders for several reasons. By combining it with knowledge gathered separately about the thickness of the oil (perhaps taken from samples or from observations using the Bonn Agreement Oil Appearance Code [33]), predictions of the slick area can help with the following tasks:

1) assessing the required size of the response, e.g., personnel, vessels, and equipment;
2) informing mitigation strategies, e.g., amounts of dispersant required;
3) knowing if all the oil has been accounted for. To do this, it may be necessary to use estimates from the modeling to understand the fraction of the overall slick that still remains on the surface, i.e., that which has not yet evaporated or dispersed.

The area skill score metric is ideally suited to this use case. This score gives a very clear assessment of a model's ability to forecast the size of an oil spill on the surface of the water, with a score of 1.0 indicating a perfect prediction of the slick's area.

3) Deploying Coastline Cleanup Teams: If a model forecasts that an oil spill will reach the coastline, then its results can be used to the help indicate which shorelines are most likely to be impacted. Cleanup teams can then be sent to those shorelines to remove the oil and help any affected wildlife. 2-D MOE values for affected coastlines give an indication of how well the model can inform this deployment of cleanup teams.

If we assume that a response has sufficient resource to deploy cleanup teams to every coastline that the model predicts will be affected, then the 2-D MOE x-value will be the fraction of the total area affected by oil which the cleanup teams will reach.

Obviously, the more affected areas the teams can reach the better, but often resources are limited and there may not be enough teams to reach all of the areas that the model predicts will be impacted. In this instance, the MOE y-value is also significant as it gives the fraction of the cleanup teams, which would find oil on the beach where they were posted. The value can be thought of as an assessment of how efficiently the resources would be used if they were deployed based on a particular set of modeling results.

4) Knowing the Maximum Possible Extent of an Oil Spill: Statistical modeling such as the probabilistic output produced by Riskaware Ltd.'s model is sometimes used to calculate the maximum possible extent of an oil spill. This information can then be used to place local authorities on standby or prepare equipment in case it is needed in the response. The 2-D MOE x-value can be applied to model predictions for the maximum extent of an oil spill to assess its ability to capture the actual outcome. A successful prediction for the extent should completely encompass the outlines of the actual spill and, therefore, should have an x-value of 1.0. Effectively, the x-value gives us the fraction of the actual oil spill that was encompassed by the predicted extent.

Through this brief discussion, we have shown that each metric provides its own insight into how effective a set of modeling results would be in a response. Table I summarizes the findings.

VII. Conclusion

This article has introduced a set of novel validation metrics for assessing the accuracy of oil spill models based on satellite observations and coastal report data. We have applied the metrics to simulations of real-world oil spill test cases and demonstrated how the results can be used to quantify the performance of a model. In doing so, our intention is to encourage the adoption of these metrics within the wider oil spill modeling community, to improve oil spill modeling capabilities and allow unbiased comparison between different models. We have also shown how the metrics can be used to investigate the sensitivity of models to different input parameters, in order to identify the key strengths and weaknesses of a given model, and where work on further improvements should be focused.

In our assessment, the validation metrics were found to have the following key strengths and weaknesses.

1) While the 2-D MOE method can be applied to both deterministic and probabilistic output, it is most useful for assessing probabilistic modeling, where the degree of false negative and positive for different probability levels can be easily assessed. However, this metric is less suitable for deterministic simulations due to its reliance on the need for an overlap in order to produce a nonzero score.
2) The centroid skill score provides a better indication of a prediction's accuracy in terms of the oil's location in a deterministic simulation, while the area skill score provides a useful assessment of the spreading and weathering processes.
3) A modified version of the 2-D MOE method, based on the lengths of affected coastline, showed that the method can also be used to assess the accuracy of the coastal deposition predictions.
4) We reiterate that the metrics are only designed to assess oil floating on the ocean surface and beached on the coastline, but not any oil below the surface of the water, whether suspended in the water column or on the seabed.
5) We note that there is scope for the metrics to include a temporal dimension, by plotting the evolution of the skill scores as a function of time since model initialization. We currently lack sufficient snapshots of the observed oil to enable such a comparison, but, in principle, this could be incorporated into the analysis as part of future work.
6) In addition to their use for model validation and sensitivity analysis, we highlight the potential for the metrics presented in this study to be used as the basis of a standardized model intercomparison study, whereby the performance of multiple models can be compared alongside each other.

The code used to calculate the metrics and generate the plots shown in this article is publicly available under an LGPL license from https://github.com/riskaware-ltd/omen

References

[1] M. Fingas and C. E. Brown, "A review of oil spill remote sensing," *Sensors*, vol. 18, no. 1, Jan. 2018, Art. no. 91.

[2] S. K. Chaturvedi, "Study of synthetic aperture radar and automatic identification system for ship target detection," *J. Ocean Eng. Sci.*, vol. 4, no. 2, pp. 173–182, Jan. 2019.

[3] A. J. Abascal, S. Castanedo, R. Minguez, R. Medina, Y. Liu, and R. H. Weisberg, "Stochastic Lagrangian trajectory modeling of surface drifters deployed during the Deepwater Horizon oil spill," in *Proc. 38th AMOP Tech. Seminar Environ. Contamination Response*, 2015, pp. 71–99.

[4] Y. Liu and R. H. Weisberg, "Evaluation of trajectory modeling in different dynamic regions using normalized cumulative Lagrangian separation," *J. Geophys. Res., Oceans*, vol. 116, 2011, Art. no. C09013.

[5] I. Ivichev, L. R. Hole, L. Karlin, C. Wettre, and J. Röhrs, "Comparison of operational oil spill trajectory forecasts with surface drifter trajectories in the Barents sea," *J. Geol. Geosci.*, vol. 1, 2012, Art. no. 105.

[6] M. De Dominicis, N. Pinardi, G. Zodiatis, and R. Archetti, "MEDSLIK-II, a Lagrangian marine surface oil spill model for short-term forecasting—Part 2: Numerical simulations and validations," *Geosci. Model Develop.*, vol. 6, no. 6, pp. 1871–1888, 2013.

[7] D. P. French-McCay, T. Tajalli-Bakhsh, K. Jayko, M. L. Spaulding, and Z. Li, "Validation of oil spill transport and fate modeling in Arctic ice," *Arctic Sci.*, vol. 4, no. 1, pp. 71–97, Mar. 2018.

[8] A. Ribotti *et al.*, "An operational marine oil spill forecasting tool for the management of emergencies in the Italian seas," *J. Mar. Sci. Eng.*, vol. 7, no. 1, Jan. 2019, Art. no. 1.

[9] Y. Cheng, X. Li, Q. Xu, O. Garcia-Pineda, O. B. Andersen, and W. G. Pichel, "SAR observation and model tracking of an oil spill event in coastal waters," *Mar. Pollut. Bull.*, vol. 62, no. 2, pp. 350–363, Feb. 2011.

[10] Q. Xu, X. Li, Y. Wei, Z. Tang, Y. Cheng, and W. G. Pichel, "Satellite observations and modeling of oil spill trajectories in the Bohai sea," *Mar. Pollut. Bull.*, vol. 71, nos. 1/2, pp. 107–116, Jun. 2013.

[11] Y. Cheng *et al.*, "Monitoring of oil spill trajectories with COSMO-SkyMed X-band SAR images and model simulation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 7, pp. 2895–2901, Jul. 2014.

[12] Y. Liu, A. MacFadyen, Z.-G. Ji, and R. H. Weisberg, *Monitoring and Modeling the Deepwater Horizon Oil Spill: A Record-Breaking Enterprise* (Geophysical Monograph Series), vol. 195. Washington, DC, USA: AGU/Geopress, 2011, p. 271.

[13] H. S. Huntley, B. L. Lipphardt, Jr., and A. D. Kirwan, Jr., "Surface drift predictions of the Deepwater Horizon spill: The Lagrangian perspective," in *Monitoring and Modeling the Deepwater Horizon Oil Spill: A Record-Breaking Enterprise* (Geophysical Monograph Series), vol. 195. Washington, DC, USA: Amer. Geophys. Union, 2011, pp. 179–195.

[14] R. H. Weisberg, L. Zheng, and Y. Liu, "On the movement of Deepwater Horizon Oil to northern Gulf beaches," *Ocean Model.*, vol. 111, pp. 81–97, Mar. 2017.

[15] G. Coppini *et al.*, "Hindcast of oil-spill pollution during the Lebanon crisis in the Eastern Mediterranean, July–August 2006," *Mar. Pollut. Bull.*, vol. 62, no. 1, pp. 140–153, Jan. 2011.

[16] A. G. Samaras, M. De Dominicis, R. Archetti, A. Lamberti, and N. Pinardi, "Towards improving the representation of beaching in oil spill models: A case study," *Mar. Pollut. Bull.*, vol. 88, nos. 1/2, pp. 91–101, Nov. 2014.

[17] Q. Xu, Y. Cheng, B. Liu, and Y. Wei, "Modeling of oil spill beaching along the coast of the Bohai sea, China," *Front. Earth Sci.*, vol. 9, no. 4, pp. 637–641, Dec. 2015.

[18] S. Mosca, G. Graziani, W. Klug, R. Bellasio, and R. Bianconi, "A statistical methodology for the evaluation of long-range dispersion models: An application to the ETEX exercise," *Atmos. Environ.*, vol. 32, no. 24, pp. 4307–4324, Dec. 1998.

[19] S. Warner, N. Platt, and J. F. Heagy, "User-oriented two-dimensional measure of effectiveness for the evaluation of transport and dispersion models," *J. Appl. Meteorol.*, vol. 43, no. 1, pp. 58–73, Jan. 2004.

[20] G. D. Rolph *et al.*, "Description and verification of the NOAA smoke forecasting system: The 2007 fire season," *Weather Forecasting*, vol. 24, no. 2, pp. 361–378, Apr. 2009.

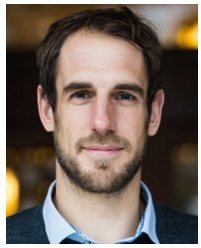
[21] A. F. Stein, G. D. Rolph, R. R. Draxler, B. Stunder, and M. Ruminski, "Verification of the NOAA smoke forecasting system: Model sensitivity to the injection height," *Weather Forecasting*, vol. 24, no. 2, pp. 379–394, Apr. 2009.

[22] S. Warner, N. Platt, and J. F. Heagy, "Comparisons of transport and dispersion model predictions of the URBAN 2000 field experiment," *J. Appl. Meteorol.*, vol. 43, no. 6, pp. 829–846, Jan. 2004.

[23] S. Warner, N. Platt, and J. F. Heagy, "Application of user-oriented measure of effectiveness to transport and dispersion model predictions of the European tracer experiment," *Atmos. Environ.*, vol. 38, no. 39, pp. 6789–6801, Dec. 2004.
[24] S. Warner, N. Platt, J. F. Heagy, J. E. Jordan, and G. Bieberbach, "Comparisons of transport and dispersion model predictions of the mock urban setting test field experiment," *J. Appl. Meteorol. Climatol.*, vol. 45, no. 10, pp. 1414–1428, Oct. 2006.
[25] J. Pullen, J. P. Boris, T. Young, G. Patnaik, and J. Iselin, "A comparison of contaminant plume statistics from a Gaussian puff and urban CFD model for two large cities," *Atmos. Environ.*, vol. 39, no. 6, pp. 1049–1068, Feb. 2005.
[26] N. Roberts, "Assessing the spatial and temporal variation in the skill of precipitation forecasts from an NWP model," *Meteorol. Appl.*, vol. 15, no. 1, pp. 163–169, Mar. 2008.
[27] N. M. Roberts and H. W. Lean, "Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events," *Monthly Weather Rev.*, vol. 136, no. 1, pp. 78–97, Jan. 2008.
[28] D. Simecek-Beatty and W. J. Lehr, "Oil spill forecast assessment using fractions skill score," *Mar. Pollut. Bull.*, vol. 164, 2021, Art. no. 112041.
[29] M. De Dominicis, N. Pinardi, G. Zodiatis, and R. Lardner, "MEDSLIK-II, a Lagrangian marine surface oil spill model for short-term forecasting—Part 1: Theory," *Geosci. Model Develop.*, vol. 6, no. 6, pp. 1851–1869, 2013.
[30] *Global Monitoring and Forecasting Center, GLOBAL_ANALYSIS_FORECAST_PHY_001_024—Global Ocean Analysis and Forecast System, E.U. Copernicus Marine Service Information [Data set]*. Accessed: Jun. 14, 2021. [Online]. Available: https://resources.marine.copernicus.eu/?option=com_csw&view=details&product_id=GLOBAL_ANALYSIS_FORECAST_PHY_001_024
[31] *Global Monitoring and Forecasting Center, WIND_GLO_WIND_L4_REP_OBSERVATIONS_012_016—Global—Blended Mean Wind Fields, E.U. Copernicus Marine Service Information [Data set*]. Accessed: Jun. 14, 2021. [Online]. Available: https://resources.marine.copernicus.eu/?option=com_csw&view=details&product_id=WIND_GLO_WIND_L4_REP_OBSERVATIONS_012_006
[32] *Global Monitoring and Forecasting Center, IBI_MULTIYEAR_PHY_005_002—Iberian Biscay Irish Ocean Reanalysis System, E.U. Copernicus Marine Service Information [Data set*]. Accessed: Jun. 14, 2021. [Online]. Available: https://resources.marine.copernicus.eu/?option=com_csw&view=details&product_id=IBI_MULTIYEAR_PHY_005_002
[33] *Bonn Agreement Oil Appearance Code*. Accessed: Jun. 14, 2021. [Online]. Available: https://www.bonnagreement.org/site/assets/files/1081/special_on_volume_calculation_20160607.docx

Chris Dearden received the Ph.D. degree in atmospheric physics from the University of Manchester, Manchester, U.K., in 2011, for research into numerical modeling of warm and mixed-phase cloud microphysics.

From 2003 to 2007, he was a Climate Model Development and Evaluation Scientist with the U.K. Met Office, Exeter, U.K. Upon completion of his Ph.D., he remained with the University of Manchester as a Research Fellow, before joining the University of Leeds, Leeds, U.K., as a Software Development Scientist in January 2018. Since 2019, he has been a Research Software Engineer with the STFC Hartree Centre, Warrington, U.K., working on a range of scientific computing projects.



Tim Culmer received the M.Eng. (first-class Hons.) degree in mechatronic engineering from the University of Lancaster, Lancaster, U.K., in 2005.

He is an experienced software developer and a project manager, with more than 14 years' involvement with the IT sector. He was a Software and Electronics Engineer with ITDev, Southampton, U.K. Since 2015, he has been with Riskaware Ltd., Bristol, U.K., where he is currently a Consultant. His experience covers a range of technical domains including digital forensics; geographic information systems; chemical, biological, radiological, and nuclear hazard modeling; and oil spill dispersion. He often presents talks on his work, with the most recent being at the Acceleration and Innovation through Satellite Applications Sarawak Conference in Malaysia.

Richard Brooke received the Ph.D. degree in physics from the University of Bristol, Bristol, U.K., in 2016, for his work on investigating the use of ferromagnetic contacts for single-molecule devices.

Since 2019, he has been a Software Engineer with Riskaware Ltd., Bristol. He was a Research Associate in Molecular Electronics with the School of Physics, University of Bristol. His current research interests include the MarineAware modeling platform, particularly the underlying oil spill dispersion model.