AGU PUBLICATIONS

JGR

Journal of Geophysical Research: Oceans

RESEARCH ARTICLE

10.1002/2015JC010989

Special Section:
Forum for Arctic Modeling and Observing Synthesis (FAMOS): Results and Synthesis of Coordinated Experiments

Skill metrics for evaluation and comparison of sea ice models

Dmitry S. Dukhovskoy[1], Jonathan Ubnoske[1], Edward Blanchard-Wrigglesworth[2], Hannah R. Hiester[1], and Andrey Proshutinsky[3]

[1]Center for Ocean-Atmospheric Prediction Studies, Florida State University, Tallahassee, Florida, USA, [2]Department of Atmospheric Sciences, University of Washington, Seattle, Washington, USA, [3]Department of Physical Oceanography, Woods Hole Oceanographic Institution, Woods Hole, Massachusetts, USA

Abstract Five quantitative methodologies (metrics) that may be used to assess the skill of sea ice models against a control field are analyzed. The methodologies are Absolute Deviation, Root-Mean-Square Deviation, Mean Displacement, Hausdorff Distance, and Modified Hausdorff Distance. The methodologies are employed to quantify similarity between spatial distribution of the simulated and control scalar fields providing measures of model performance. To analyze their response to dissimilarities in two-dimensional fields (contours), the metrics undergo sensitivity tests (scale, rotation, translation, and noise). Furthermore, in order to assess their ability to quantify resemblance of three-dimensional fields, the metrics are subjected to sensitivity tests where tested fields have continuous random spatial patterns inside the contours. The Modified Hausdorff Distance approach demonstrates the best response to tested differences, with the other methods limited by weak responses to scale and translation. Both Hausdorff Distance and Modified Hausdorff Distance metrics are robust to noise, as opposed to the other methods. The metrics are then employed in realistic cases that validate sea ice concentration fields from numerical models and sea ice mean outlook against control data and observations. The Modified Hausdorff Distance method again exhibits high skill in quantifying similarity between both two-dimensional (ice contour) and three-dimensional (ice concentration) sea ice fields. The study demonstrates that the Modified Hausdorff Distance is a mathematically tractable and efficient method for model skill assessment and comparison providing effective and objective evaluation of both two-dimensional and three-dimensional sea ice characteristics across data sets.

Key Points:
- Quantitative methodologies for skill assessment of sea ice models are analyzed
- A novel approach based on topological metrics is tested for model skill assessment
- The metrics offer quantitative evaluation of spatial distribution of simulated ice

Correspondence to:
D. S. Dukhovskoy,
ddukhovskoy@fsu.edu

Citation:
Dukhovskoy, D. S., J. Ubnoske, E. Blanchard-Wrigglesworth, H. R. Hiester, and A. Proshutinsky (2015), Skill metrics for evaluation and comparison of sea ice models, *J. Geophys. Res. Oceans*, *120*, 5910–5931, doi:10.1002/2015JC010989.

Received 19 MAY 2015
Accepted 6 AUG 2015
Accepted article online 10 AUG 2015
Published online 2 SEP 2015

1. Introduction

Due to its fundamental role in physical, chemical, and biological processes in the polar regions, sea ice is considered to be a key state variable for the Arctic and Antarctic climate environment [*Vaughan et al.*, 2013]. The necessity to understand current and future changes in sea ice and the associated consequences for the climate has stimulated active development of sea ice models. These models have evolved into mathematical systems describing complex physical processes controlling thermodynamics and dynamics of sea ice [e.g., *Kreyscher et al.*, 2000; *Hunke and Holland*, 2007; *Johnson et al.*, 2007; *Lipscomb et al.*, 2007; *Hunke et al.*, 2010; *Germe et al.*, 2014; *Peterson et al.*, 2014]. Vast scatter of simulated sea ice characteristics among sea ice models has been reported [e.g., *Arzel et al.*, 2006; *Zhang and Walsh*, 2006; *Serreze et al.*, 2007; *Blanchard-Wrigglesworth and Bitz*, 2014]. Discrepancies among sea ice numerical forecasts and hindcasts have motivated several model intercomparison projects such as the Sea-Ice Model Intercomparison Project (SIMIP) [*Kreyscher et al.*, 1997, 2000] and Arctic Ocean Model Intercomparison Project (AOMIP) [*Proshutinsky et al.*, 2005, 2011; *Johnson et al.*, 2007, 2012], which aimed to evaluate and improve representation of sea ice in climate and coupled ocean-sea ice models, respectively. Recently, considerable attention has been drawn to representation of polar sea ice in global climate models [e.g., *Parkinson et al.*, 2006; *Connolley and Bracegirdle*, 2007; *Turner et al.*, 2013; *Uotila et al.*, 2012, 2013, 2014].

Quantitative approaches used for sea ice model validation predominantly operate with integrated (or aggregated) scalar fields such as sea ice extent, area, or volume [*Arzel et al.*, 2006; *Hunke and Holland*, 2007; *Schweiger et al.*, 2011; *Turner et al.*, 2013; *Germe et al.*, 2014; *Peterson et al.*, 2014; *Tietsche et al.*, 2014]. These methods provide a quick and easy approach to quantify uncertainty across sea ice models. However, the area-integrated characteristics have very limited application for model skill assessment as models with very different sea ice thickness distribution and configuration of ice extent boundary may have similar area-integrated

quantities [*Connolley and Bracegirdle*, 2007]. More generally, these metrics are limited by a lack of information about spatial patterns of sea ice properties within the region of interest.

Presently available satellite observations of polar regions provide detailed information about spatial distribution of sea ice characteristics [*Kwok et al*., 2004; *Kwok*, 2010; *Laxon et al*., 2013]. These high-resolution and temporally regular observations offer a valuable data set for model evaluation and intercomparison that could be used to evaluate performance of sea ice models more precisely. For example, detailed sea ice edges derived from radiometer and scatterometer observations [*Comiso et al*., 1997; *Meier and Stroeve*, 2008] are available for sea ice edge validation in the sea ice models. However, in order to utilize this information, there is a need to compare the spatial distribution of the sea ice. This requires reliable and robust validation techniques that can assess representation of spatial distribution of sea ice characteristics in the models.

Comparison of spatial patterns across the sea ice models is mostly limited to qualitative methods or visual comparison of spatial patterns of the sea ice characteristics such as concentration, thickness, and draft [e.g., *Parkinson et al*., 2006; *Hunke and Holland*, 2007; *Johnson et al*., 2007; *Laxon et al*., 2013; *Germe et al*., 2014; *Peterson et al*., 2014; *Tietsche et al*., 2014; *Tsamados et al*., 2014]. Quantitative evaluation of sea ice model performance based on spatial patterns (such as shapes or contours of sea ice edge, distribution of ice thickness, or concentration within the contours) of simulated sea ice characteristics has received little attention in the geophysical literature. Several studies have evaluated spatial distribution of sea ice characteristics in sea ice models based on Root-Mean-Square Error (or Deviation) analysis [*Cavalieri*, 1992; *Uotila et al*., 2012; *Karvonen*, 2014] and spatial pattern correlation [*Schweiger et al*., 2011]. The apparent limitation in the assessment metrics specifically designed to account spatial distribution of sea ice can be attributed to the difficulty of developing algorithms to quantify similarity in shape in a way that corresponds with human intuition.

The goal of this study is to investigate several possible methods for objective quantitative evaluation and skill assessment of sea ice models based on spatial patterns of ice characteristics. In general, model validation and skill assessment can be defined as automated, objective quantification of similarity or dissimilarity between a numerical solution and control data. For sea ice applications, the validation can be performed on contours that delineate specified constant values, for instance, comparing the shape of a sea ice edge contour that captures sea ice extent. Furthermore, a validation process should also compare the spatial pattern of sea ice characteristics such as concentration and thickness against the control fields (similar to pattern recognition). In summary, a function is sought that can compare several sea ice models to the control data and rank them according to similarity in contour shape or both in contour shape and a scalar variable within the contour.

Five methodologies are considered: Absolute Deviation (section 2.1), Root-Mean-Square Deviation (section 2.2), Mean Displacement (section 2.3), Hausdorff Distance, and Modified Hausdorff Distance (section 2.4). The first three quantify the error between the numerical solution and the control data. The last two originate in the study of metric spaces and compare the model to the control data via a shape distance metric, which takes into account their similarity in shape.

The metrics are subjected to sensitivity and robustness tests (section 3) in order to demonstrate their response to differences in sea ice contours and patterns (sections 4 and 5). Then the metrics are applied to sea ice concentration fields from coupled ocean-sea ice model experiments (section 6.1) and September 2014 sea ice outlook (section 6.2), and their performance for these realistic cases is evaluated.

2. Validation Metrics

The following validation metrics will be considered for skill assessment of modeled sea ice fields: Absolute Deviation, Root-Mean-Square Deviation, Mean Displacement, Hausdorff Distance, and Modified Hausdorff Distance. The metrics are applied in both 2-D, using only contours of sea ice edge, and 3-D cases, using both the contours and scalar field within.

2.1. Absolute Deviation

Sea ice area and sea ice extent are the simplest and most frequently used skill metrics for evaluation of differences between simulated and observed data [*Arzel et al*., 2006; *Hunke and Holland*, 2007; *Schweiger et al*.,

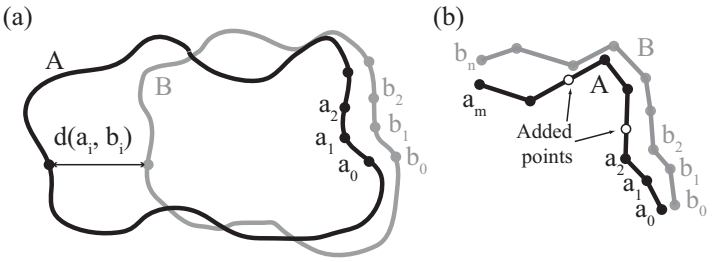

Figure 1. Schematic diagram of the point-to-point matching algorithm. (a) $d(a_i, b_i)$ is the error ("distance") used in calculation of the RMSD score between two contours. The pair-wise correspondence of points is determined as follows. Given two sets of points A and B containing the same number of points, a point a_0 on set A is chosen at random as the initial point, and then the closest point b_0 on set B is selected as its corresponding point (if more than one such point exists, one is selected at random). Point a_1 is chosen to be the point in A\{a_0} (point a_0 is excluded from the set A) closest to a_0. Point b_1 is chosen to be the point in B\{b_0} closest to a_1. The process iterates in this fashion until all pairs have been assigned. (b) Sets A and B have a different number of points. Suppose set A has m points and set B has n points, with $m < n$. Let $k = n - m$. The k longest edges on A's contour are chosen and at each midpoint a point is added. Sets A and B now have the same number of points, and the algorithm from Figure 1a is performed.

2011; *Turner et al.*, 2013; *Peterson et al.*, 2014]. The question is whether an area-integrated statistic can be successfully applied to quantify similarity in shapes of sea ice fields. Area-based metrics have been used as shape description techniques in object recognition applications [*Zhang and Lu*, 2004]. The Absolute Deviation (AD) metric used here is defined as

$$D_{AD}(C_i, C_0) = |C_i - C_0|, \quad (1)$$

where in 2-D, C_i and C_0 are the areas inside the ice edge contour Ω derived from the tested and control data sets, respectively,

$$C = \int_{\Omega} dA. \quad (2)$$

In this case, C_i and C_0 are the sea ice extents, and AD measures an absolute difference in the sea ice extent. For a 3-D application (the metric is referenced as "AD3D"), C_i and C_0 are defined as

$$C_i = \int_{\Omega} g_i(x, y) dA, \quad (3)$$

$$C_0 = \int_{\Omega} g_0(x, y) dA, \quad (4)$$

where g_i and g_0 are the tested and control scalar fields. If these functions describe sea ice concentrations, equations (3) and (4) are sea ice areas in the two data sets bounded by contour Ω.

2.2. Root-Mean-Square Deviation

Root-Mean-Square Deviation (RMSD) is another frequently used quantitative estimate of difference between model output and the control data. In sea ice applications, RMSD (or Root-Mean-Square Error, RMSE) has been used for comparing integrated quantities such as sea ice extent or area [e.g., *Emery et al.*, 1991; *Cavalieri*, 1992; *Tietsche et al.*, 2014]. For 3-D applications, when spatial distribution of sea ice characteristics is assessed, RMSD-based techniques are usually calculated through consideration of point-by-point differences over the domain of interest [*Uotila et al.*, 2012].

Here in 2-D, RMSD is employed to analyze spatial patterns through comparison of contours such as sea ice edge. For two given sets of points A and B, the RMSD score is defined as

$$D_{RMSD}(A, B) = \sqrt{\frac{\sum_{i=1}^{n} [d(a_i, b_i)]^2}{n}}, \quad (5)$$

where $d(a_i, b_i)$ is an error (here, distance) between the ith pair of points $a_i \in A$ and $b_i \in B$. In the case of two contours, RMSD is estimated for n data points along the curves (Figure 1). Note that the definition of RMSD assumes a correspondence between the points in sets A and B ($a_i \in A \leftrightarrow b_i \in B$). This information is generally not known in realistic applications. There are several possible techniques to determine the pair-wise correspondence of points between the two shapes [e.g., *Chui and Rangarajan*, 2003]. Here for two given sets of points, the two closest points from the sets are considered as the matching pairs (see Figure 1 for more detail). Another hurdle related to the RMSD application is when the number of points is different in

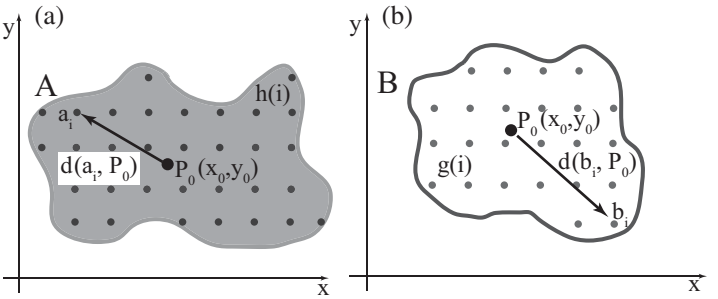

Figure 2. Weighted mean displacements measured for the data sets (a) *A* and (b) *B*. The metric considers the overall dispersion of data points around a reference location (P_0). The arrow shows the distance between P_0 and a data point.

the data sets *A* and *B*. Then extra steps are required before RMSD can be calculated in order to define a point-to-point correspondence between the contours (Figure 1b).

For 3-D applications, when spatial distribution of sea ice characteristics inside a contour is assessed, RMSD-based techniques are usually calculated point-by-point over the domain of interest. In this analysis, a methodology similar to *Uotila et al.* [2012] is employed for 3-D applications (called the "normalized root-mean-square error" in the original paper, here referenced as "RMSD3D")

$$D_{RMSD3D} = \frac{\sqrt{\frac{\sum_{j=1}^{m}\sum_{i=1}^{n}[h(i,j)-g(i,j)]^2}{nm}}}{\max(g)-\min(g)}, \tag{6}$$

where $h(i,j)$ and $g(i,j)$ are tested and control scalar fields, respectively, of the grid cell (i,j). Note that this technique also requires a point-to-point correspondence between the scalar fields (as do most other RMSD-based methods). This means that the methodology does not account for information about position of the scalar fields. This can potentially debilitate sensitivity and robustness of this method to differences in sea ice patterns. For example, if two identical fields are shifted by a few grid points, the metric may give a low skill score (large RMSD3D) despite the similarity in spatial patterns. Another drawback of this method is the requirement that both fields must be on the same grid, which is often not the case in realistic applications such as comparing sea ice fields from different models. In these cases, the scalar fields have to be interpolated onto the same grid before RMSD can be applied. None of the other considered methods have this constraint.

2.3. Mean Displacement

The Mean Displacement (MD) metric originates in spatial statistics as a measure of spread or dispersion of a data set around a reference location (P_0) and has been successfully applied to shape recognition [*Chang et al.*, 1991; *Zhang and Lu*, 2004]. This technique offers a natural metric for sea ice model evaluation. Since MD is a measure of dispersion, the metric is skilled at picking up small fluctuations of sea ice fields along the boundaries where most uncertainties occur. The data set is defined as every point bounded by a given contour (Figure 2). In the 2-D cases here, which consider contour comparison (sea ice edge), the data set is defined as points on the contour. P_0 is taken as the centroid of the data set. The skill is assessed by considering the difference between dispersions of the data points measured relative to P_0 in the data sets. MD can be generalized for a 3-D application (referenced as "MD3D") where each location is "weighted" by functions *h* and *g* that describe distributions of some property (e.g., sea ice thickness) within the data sets *A* and *B*, respectively. The difference in the weighted dispersions gives the MD3D score (D_{MD3D})

$$\bar{D}_A = \frac{1}{n}\sum_{i=1}^{n} h_i\, d(a_i, P_0), \tag{7}$$

$$\bar{D}_B = \frac{1}{m}\sum_{i=1}^{m} g_i\, d(b_i, P_0), \tag{8}$$

$$D_{MD3D}(A,B) = |\bar{D}_A - \bar{D}_B|, \tag{9}$$

where $d(a_i, P_0)$ is the distance between a point a_i and P_0 and similarly for $d(b_i, P_0)$ (Figure 2). In this case, both the field shapes and distribution of the analyzed property within the fields are compared. Note: when $h = g = 1$, equations (7)–(9) give MD.

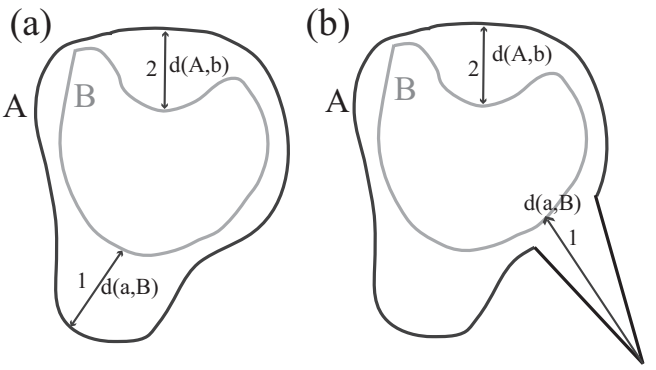

Figure 3. Hausdorff distance between point-sets *A* and *B* represented by two contours. (a) Line 1 represents $\sup_{a\in A} d(a,B)$, and line 2 represents $\sup_{b\in B} d(A,b)$. Hausdorff Distance (D_{HD}) corresponds to line 1, the longer of the two lines. (b) If another point is added to *A* such that this point is placed sufficiently far outside of the original contour *A*, HD is now determined by the distance between this outlier and the closest point on contour *B*.

2.4. Topological Methods: Hausdorff Distance and Modified Hausdorff Distance

Topological shape matching descriptors use a metric distance between the objects as the measure of shape similarity [*Zhang and Lu*, 2004]. The topological approach is a natural means of evaluation when spatial distributions of simulated properties are compared across data sets. The Hausdorff Distance (HD) is a classical, widely used method for both object and shape matching [*Huttenlocher et al.*, 1993; *Dubuisson and Jain*, 1994a; *Rucklidge*, 1997; *Daoudi et al.*, 1999]. Here HD is defined as

$$D_{HD}(A,B)=\max\left\{\sup_{a\in A} d(a,B),\ \sup_{b\in B} d(A,b)\right\}, \tag{10}$$

where $d(a,B)=\inf_{b\in B} d(a,b)$, $d(A,b)=\inf_{a\in A} d(a,b)$, and $d(a, b)$ is the distance between the point *a* on *A* and the point *b* on *B*. The distance $d(a, b)$ can be Euclidean distance or another appropriate distance depending on the application (e.g., great circle distance). Generally, a smaller HD indicates better resemblance between the shapes. However, HD is sensitive to outliers, as illustrated by Figure 3. This is an undesirable property for a model validation metric, since a key requirement is the ability to quantify the overall resemblance of compared data fields.

A modified version of the classic Hausdorff Distance that is resistant to outliers is also considered [*Dubuisson and Jain*, 1994b]. In the present study, this Modified Hausdorff Distance (MHD) between two finite point sets *A* and *B* in the plane is defined as

$$D_{MHD}(A,B)\equiv D(A,B)=\max\left\{\frac{1}{|A|}\sum_{a\in A} d(a,B),\ \frac{1}{|B|}\sum_{b\in B} d(A,b)\right\}, \tag{11}$$

where |A| and |B| are the cardinality of sets *A* and *B* (here the number of points on the contours), $d(a,B)$ and $d(A,b)$ are defined by $d(a,B)=\min_{b\in B} d(a,b)$ and $d(A,b)=\min_{a\in A} d(a,b)$. As with HD, the distance $d(a,b)$ can be Euclidian distance or another appropriate distance (depending on the application). As discussed in section 5, both metrics can be easily extended to a 3-D application (referred as "HD3D" and "MHD3D," respectively).

3. Sensitivity and Robustness Testing of the Skill Metrics

The considered metrics undergo sensitivity and robustness tests in order to evaluate the ability of each metric to quantify similarity in shape between two objects. When quantifying similarity in shape, special consideration should be given to the ability of a skill metric to penalize important (with respect to each particular application) dissimilarities and ignore unimportant ones. For skill assessment of a sea ice model, a validation metric is sought such that can appropriately penalize differences in (1) scale, (2) translation, (3) rotation, and (4) noise.

Although dissimilarity in scale, translation, or rotation may be unimportant in certain applications, such as object recognition [e.g., *Chang et al.*, 1991], for sea ice model validation that is not the case. Displacement of a simulated sea ice field relative to a control sea ice field (either by rotation or translation) is an important dissimilarity to quantify when assessing model skill, since it may indicate biases in sea ice physics. Similarly for scale, two sea ice fields from different models that have a general agreement on the ice edge shape but disagree on the ice extent cannot be considered identical because in this case a different amount of ice is produced in each of the models.

Robustness to noise is also an important property of a skill metric for the sea ice model application. In some cases, compared sea ice shapes may disagree on small details due to random errors stemming from a number of factors (e.g., data interpolation, spatial and temporal resolution of the data sets, differences in remote sensing instruments, and retrieval algorithms). A robust metric should, therefore, rank objects as similar when there are small object deviations (noise) that result in small dissimilarities [*Belongie et al.*, 2002]. As noise increases, the similarity between the objects, as measured by the skill metric, should decay. The particulars of how the metric should penalize each of the four differences are discussed below in the context of testing each of the differences. A more formal description of the suggested approach to robustness and sensitivity testing is given as follows.

Suppose $f_t : X \to Y,\ t \in [0,\ 1]$ is a collection of functions which represent a continuous deformation such that the associated function $F : X \times [0,\ 1] \to Y$ is continuous (more formally, f_t is a homotopy, but this can be envisioned intuitively as a continuous deformation or transformation of an elastic object over time). For a simple example, consider $f_t : R^2 \to R^2$ defined by $f_t(\mathbf{x}) = \mathbf{x} + t\,\mathbf{u}$ for some unit vector $\mathbf{u}$. Next, let $M \subset R^n$ be a sea ice field, $d : R^n \times R^n \to R$ be a proposed validation metric, and $f_t : R^n \to R^n$ be a continuous deformation. Then d is *responsive* to f_t on M if the graph of $d[M,\ f_t(M)]$ is strictly increasing with respect to t.

For example, suppose the sea ice field M is represented as a contour in the plane and subjected to the continuous deformation (in topological sense) $f_t(\mathbf{x}) = \mathbf{x} + t\begin{pmatrix} 1 \\ 0 \end{pmatrix}$, with $t \in [0,\ 1]$. Then f_t performs a translation of M to the right by t as t ranges from 0 to 1; if d (a proposed validation metric) is responsive to f_t, then it is strictly increasing with respect to t. The rate at which d increases characterizes sensitivity of the metric to the tested difference.

The three continuous deformations considered during the tests are scaling, translation, and rotation. Furthermore, the response of each metric to the addition of noise to the boundary is tested. Note, however, that the addition of noise to the boundary is not necessarily a continuous deformation since the continuity requirement in the definition of considered deformation will likely be violated. In this case, a completely different behavior of the metric is expected as noise is being added to an object: a metric needs to be *resistant* (or *robust*) to noise.

In the following section, the validation metrics described in section 2 will be tested through comparison of a control contour (shape) and modified contours with artificially introduced differences. In section 5, the same tests will be performed for the metrics comparing scalar fields randomly distributed within the contours (shape and spatial distribution). Note that for the MD case, the reference point P_0 is kept fixed at the original location (otherwise the metric will not be responsive to the translation or rotation tests). The tests have been applied to 150 randomly generated shapes with similar results and three of these cases (Figures 4a–4c) are presented in the following sections.

4. Sensitivity Tests With a Shape

4.1. Scale Test

Let $X = Y = R^2$. Then, a continuous deformation that performs rescaling is given by $f_t(\mathbf{x}) = (1 + kt)\mathbf{x}$ for some scale $k \in R$ and $t \in [0,\ 1]$. For this test, randomly generated shapes (Figures 4a–4c) undergo a linear increase in size with the maximal scale twice the original size (Figure 4d). At each iteration, scores from the validation metrics are computed between the original control contour and the rescaled shape (Figure 5). For all shapes, each metric exhibits the desired behavior of being strictly increasing. However, the graph of RMSD has sharp discontinuities that exhibit an unstable response of the metric to scaling. This result indicates that although RMSD is generally responsive to the scaling, it is not responsive in a robust way. This may result in biases in skill assessment of model performance when using RMSD. It is noteworthy that the scale test is the simplest of the tests that the metrics are subjected to. A metric that is in any way a measurement of shape is expected to pass this test. All but RMSD have successfully passed the test.

4.2. Rotation Test

Let $X = Y = R^2$. Then, a continuous deformation that performs rotation over angle θ is $f_t(r,\ \theta) = (r, \theta + 2\pi t)$, where r is a distance from the centroid to a data point. The test is performed by rotating the control shape

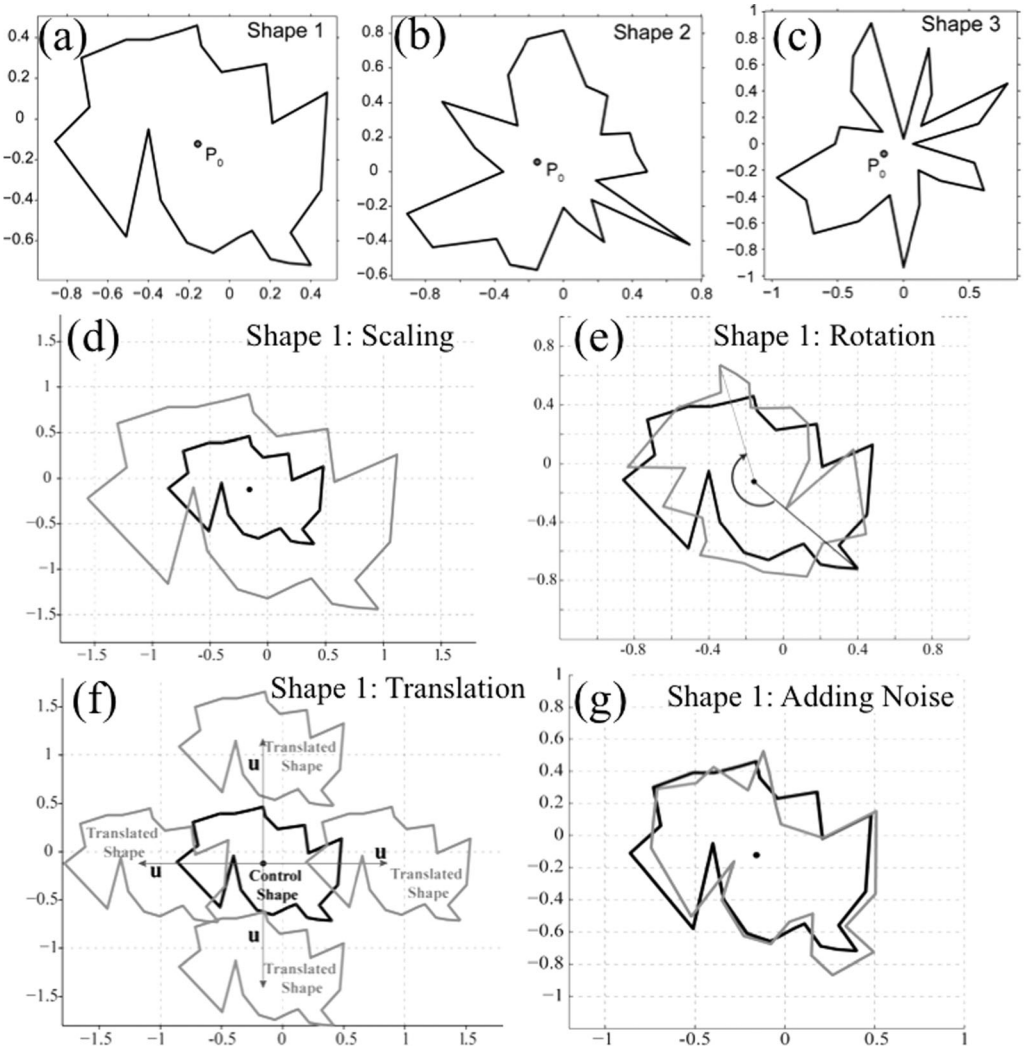

Figure 4. (a–c) Sensitivity tests with three randomly generated shapes. The black dot is the reference point P_0 (centroid) for calculating MD. Illustration of sensitivity and robustness tests are shown for Shape 1: (d) the shape is linearly scaled up to double the original size (grey contour). (e) The contour is rotated clockwise. The grey contour is rotated 210° clockwise relative to the control black contour. The MD score is nearly 0 for the shown grey and black contours (Figure 6a). (f) The contour is linearly translated in all directions along a unit vector **u**. Only four directions are shown. (g) Normally distributed noise is added to the coordinates of the contour. The grey contour is obtained by adding the normally distributed random noise with $\sigma^2 = 9E\text{-}3$.

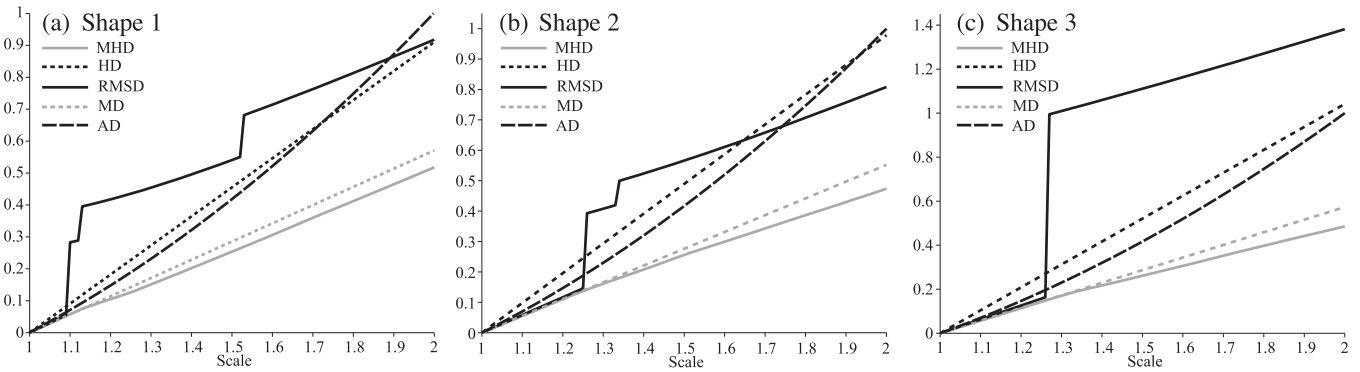

Figure 5. Metric scores versus scale (Figure 4d) for (a–c) three shapes shown in Figures 4a–4c. Different lines represent different skill metrics (see legend). AD score has been normalized by the maximum value for demonstration purposes.

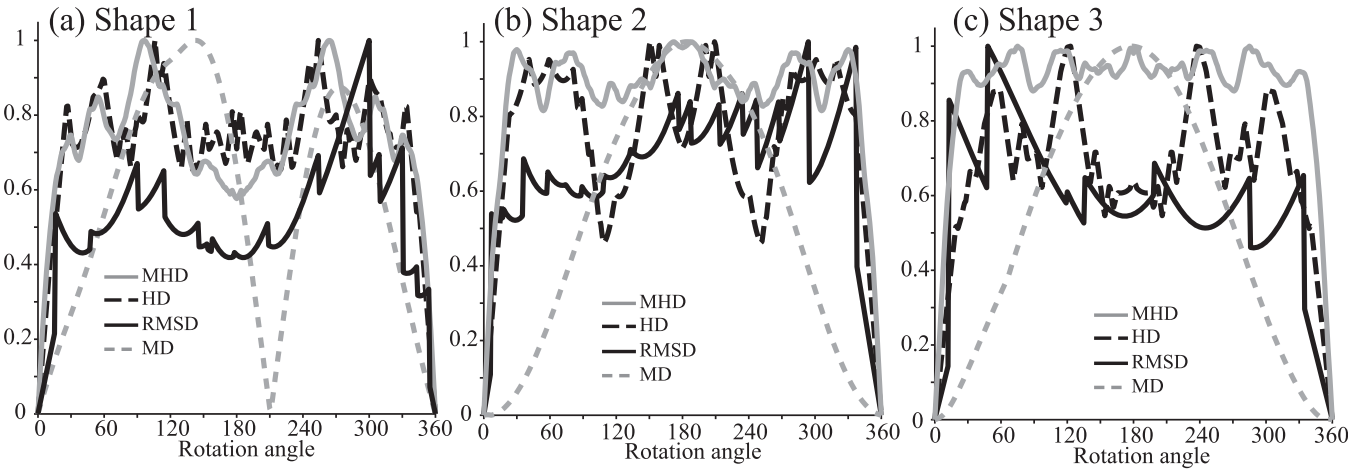

Figure 6. Metric scores versus clockwise rotation angle (degrees) for (a–c) three shapes. The AD score is not shown, as it remains 0 for all rotation angles. All scores have been normalized by their maximum values.

incrementally up to a total rotation of 2π (Figure 4e). At each iteration, the metric distances are computed between the rotated contour and the control (fixed) contour. Note that the AD metric is not able to measure uncertainty in shape orientation, since the total area does not change with rotation.

Ideally, the metric distances are expected to increase over the interval $[0, \pi]$ and then decrease over the interval $[\pi, 2\pi]$ as the contour returns to its initial position. The shape of the curve of the metric scores is expected to be symmetric around π. It is noteworthy that locations of local extrema of the curve may depend on the shape of the object. For example, a square would result in four minima at $[0 + n\pi/2]$ and four maxima $[\pi/4 + n\pi/2]$, $n = 0, 1, 2, 3$.

The RMSD metric demonstrates an unstable response to rotation (Figure 6). MHD and HD have smoother curves that are generally increasing over the interval [0, 30°–60°] and are generally decreasing over the interval [330°, 360°]. The broad local minimum around π in Figure 6a is due to the quasi-symmetric nature of shape 1 which, when rotated over π, resembles the control contour, resulting in smaller scores (Figure 4b). The existence of the minimum around π in this case is a consequence of the symmetry of this specific randomly generated contour. For the other two shapes that are not symmetric, the minimum around π is absent.

MD exhibits a very good response to rotation increasing monotonically over [0, 150°] and decreasing over [∼270°, 360°] (Figure 6a). However, the minimum at 210° is counterintuitive. The score is almost 0, indicating a near-perfect match of the rotated and control contours, which in reality is not the case (see Figure 4b). The false "success" (rotated and control contours match) predicted by the MD algorithm may be attributed to either the shape symmetry or the possibility that a number of sets of points that are not identical can still have the same mean dispersion of points around a reference point (equations (7)–(9)). In other words, MD may give a similar score to dissimilar shapes if dispersions of the points on the contours are the same. For two other shapes (Figures 6b and 6c), MD response behaves as expected by increasing over [0, 180°] and decreasing after that.

The above analysis shows that none of the tested metrics are completely responsive to rotation. However, all the metrics except AD can penalize disorientation of two contours when one contour is rotated over a small angle relative to the other. This is particularly relevant to the practical application considered here because anticipated biases in spatial orientation of sea ice contours have small range; therefore, it is most important that a skill metric is able to penalize rotation over angles much less than $\pi/2$.

4.3. Linear Translation Test

Let $X=Y=R^2$. Then, a continuous deformation that performs a translation is $f_t(\mathbf{x})=\mathbf{x}+t\,\mathbf{u}$, where $t \in [0, 1]$ and $u \in R^2$ is a unit vector pointing in the direction of translation. In this test, the contour is gradually translated by moving the shape away from the original location of the center (0, 0) along straight lines in different directions determined by the unit vector (Figure 4f). The unit vector is incrementally rotated around the

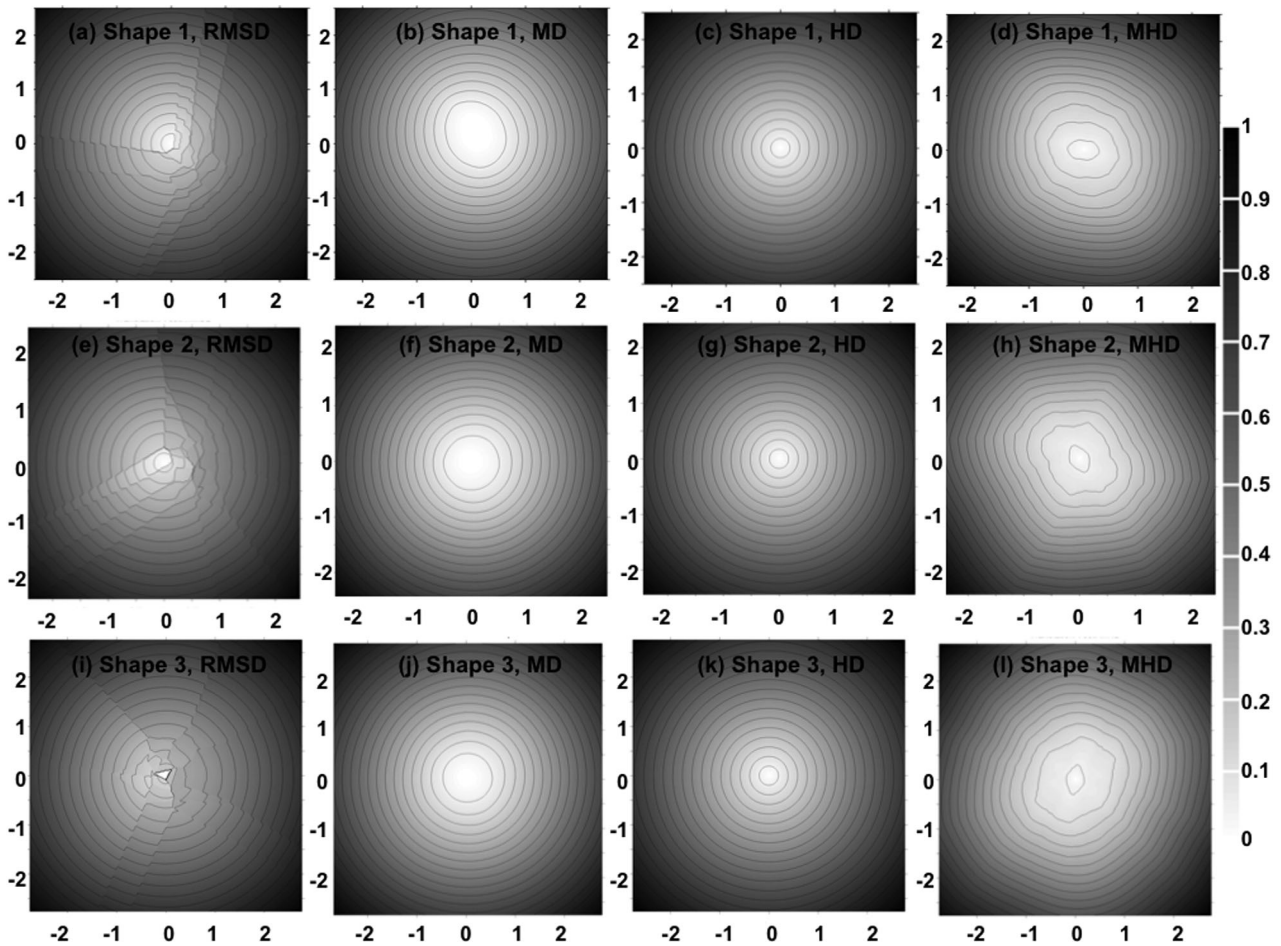

Figure 7. Metric scores for the linear translation test (Figure 4f) that the three shapes (in the rows) are subjected to. Each row corresponds to a different shape and different tests are shown in the columns. The horizontal and vertical axes are distances between the centroids of the shifted and control contours. The zero score corresponds to the initial location of the contour (0,0). (a, e, and i) RMSD. Note discontinuities in the metric field. (b, f, and j) MD. (c, g, and k) HD. (d, h, and l) MHD. The AD metric takes the zero value everywhere and is not shown. The black contours are drawn at intervals of 0.1.

centroid. Again, the AD metric is invariant to translation and thus it is not presented. The resulting metric scores (Figure 7) are represented by a surface with a minimum at (0,0) where the two contours are collocated. With RMSD (Figures 7a, 7e, and 7i), the metric score surface has obvious discontinuities, demonstrating limited sensitivity of this metric to translation. The other metrics all illustrate a good response to the translation test.

4.4. Random Noise Test

In this test, the control shape is perturbed by Gaussian noise with different standard deviations (noise amplitude). Let $X=Y=R^2$. Then, a function that adds increasing amounts of noise is $f_t=\mathbf{p}_i+\xi,\ \forall\mathbf{p}_i\in X$, where ξ is a bivariate normally distributed random vector with zero mean vector and covariance matrix $\Sigma^2=\begin{pmatrix}\sigma^2 & 0\\ 0 & \sigma^2\end{pmatrix}$. To test different levels of noise, σ is continuously increased from 0 to 0.15 (Figure 4g). The procedure is repeated 25 times. The preferred behavior of the metric score function is an overall increasing trend with points representing individual test outcomes distributed around the trend. The smaller the uncertainty range the more robust the metric to noise.

Both the AD and MD approaches (Figures 8a, 8f, and 8k and 8c, 8h, and 8m) completely fail the test. The uncertainty range increases rapidly with noise amplitude. The very large spread of AD and MD scores indicates no robustness of these methods to noise. Most importantly, even for large noise amplitude ($\sigma > 0.05$),

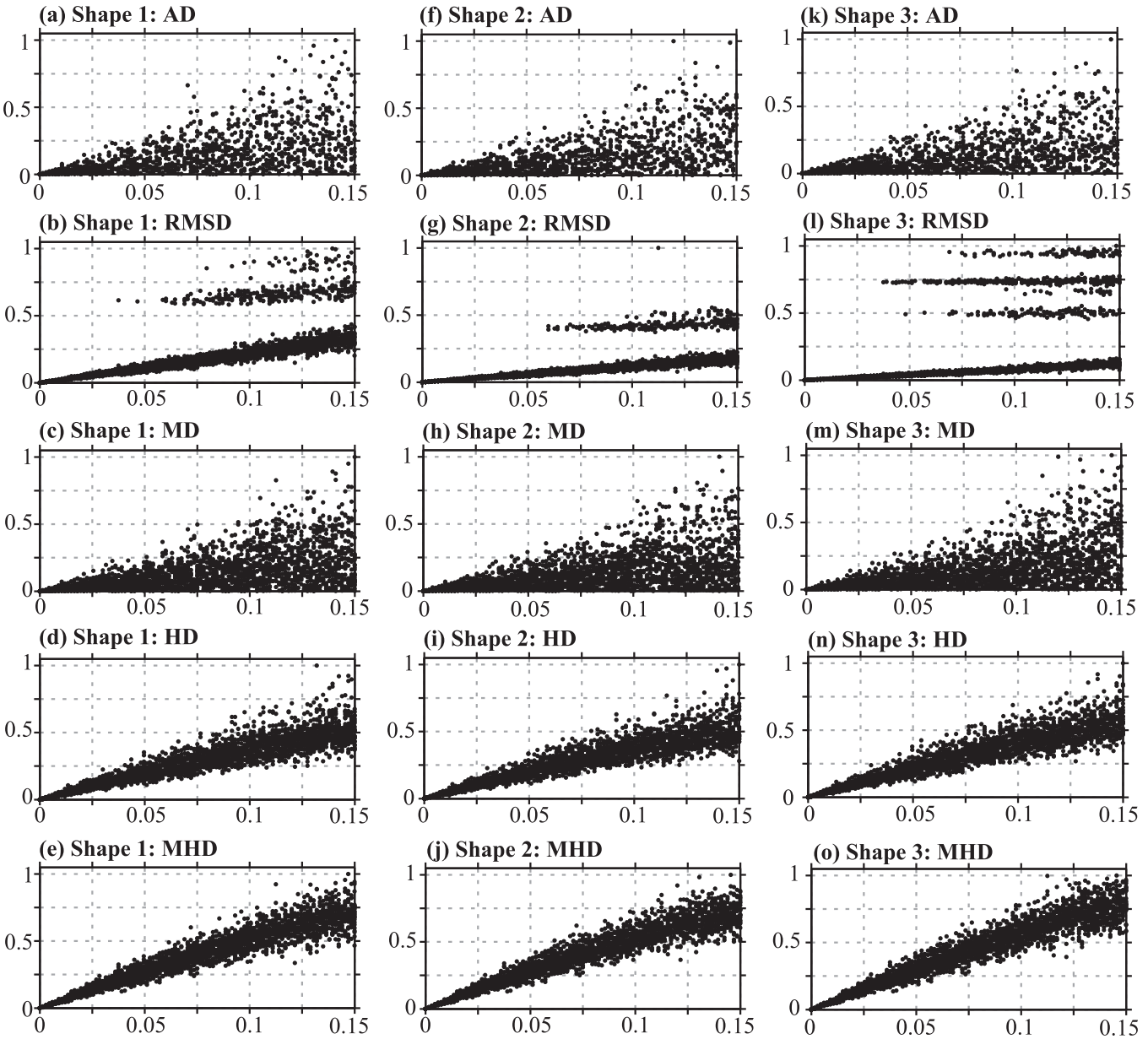

Figure 8. Metric scores (vertical axis) from the noise test for the three shapes (in columns): (a, f, and k) AD, (b, g, and l) RMSD, (c, h, and m) MD, (d, i, and n) HD, and (e, j, and o) MHD. The horizontal axis is the noise amplitude. Note the differences in the uncertainty range (spread) of the scores for different metrics.

the MD and AD scores can be ~0 (mistakenly indicating a perfect match with the control contour). This is another manifestation of the drawback of the MD algorithm discussed in section 4.2. The faulty response to noise of the AD metric stems from the fact that dissimilar shapes may have similar areas.

For RMSD, there is a positive trend in the scores with growing noise amplitude (Figures 8b, 8g, and 8l). However, there is a clear discontinuity and a very large spread of the metric scores for noise with an amplitude greater than $\sim 3 \times 10^{-2}$. RMSD is, therefore, partially responsive to noise and demonstrates good performance for noise with small amplitude. For practical applications, this means that RMSD scores for the sea ice contours with larger perturbations can be inflated and hence are not robust.

Both HD and MHD methods demonstrate a very robust response to added noise (Figure 8, two bottom rows). A slightly larger spread of HD scores for the noise with larger amplitude is due to the sensitivity of HD to outliers (as discussed in section 2.3).

Table 1. Results of the Sensitivity Tests With Contours (2-D) and Spatially Distributed Scalar Fields (3-D)[a]

	AD		RMSD		MD		HD		MHD	
Tests	2-D	3-D	2-D	3-D	2-D	3-D	2-D	3-D	2-D	3-D
Scale	P	P	F	pP[b]	P	P	P	P	P	P
Rotation	F	F	pP[c]	pP[d]	pP[e]	pP[f]	pP[f]	pP[f]	pP[f]	pP[e]
Translation	F	F	F	F	P	P	P	P	P	P
Noise	F	F	pP[g]	P	7F	F	P	pP[5]	P	P

[a]P—passed; F—failed; pP—partially passed.
[b]Partially responsive, may provide inflated scores for small changes in scale.
[c]Partially responsive within a small angle ($<\pi/18$).
[d]Partially responsive within $\pm\pi/2$, may provide inflated scores for small rotation angles.
[e]Partially responsive within approximately $\pm\pi/2$.
[f]Partially responsive within approximately $\pm\pi/6$.
[g]Robust for a small noise amplitude.

In summary, sensitivity tests with 2-D contours demonstrate a better performance of the topological metrics compared to the other metrics (Table 1). The AD metric is the simplest and least reliable skill metric, passing only one test. It is followed by RMSD, which has partially passed two tests but fails the scale and translation tests. MD exhibits good responsiveness to the three continuous deformation tests but fails the noise test. Both topological metrics, HD and MHD, have passed all four tests and although they are not fully responsive to rotation, they are sufficiently responsive for realistic applications.

5. Sensitivity Test With a Shape and Distribution Function

The sensitivity tests described in section 4 compared 2-D shapes (contours) of two objects. In real sea ice applications, differences of the distributions of sea ice characteristics inside the shape are also important and may be greater than the dissimilarity in the shapes of the objects [e.g., *Connolley and Bracegirdle*, 2007; *Katavouta and Myers*, 2014]. Subjected to the same tests, the metrics are used in the context of pattern recognition in order to assess their ability to quantify the resemblance of spatial patterns inside the shapes. In this test, field concentration is considered as the third dimension, and three-dimensional (3-D) Euclidean distance is used as a the distance $d(a,b)$ in RMSD3D, MD3D, HD3D and "weights" in MD3D (g and h in equations (7) and (8)) and AD3D. It should be noted that for the 3-D application, RMSD is calculated using equation (6). Without normalization, equation (6) is a 3-D generalization of equation (5) under the assumption that the grid points on two fields are collocated leading to zero horizontal distances.

The metrics may be sensitive to the scaling of the third dimension (e.g., ice concentration) relative to the spatial dimensions. Here it is assumed that the third dimension should maintain the same range of values as the first two dimensions. Considered ice concentration ranges from [0, 1]. All ice concentration fields have been mapped into an index space of the model grid. Thus, the third dimension is comparable to the spatial dimensions and no scaling is performed in the examples presented in the following sections.

The scalar random field inside the shapes is generated as follows. First, a continuous random field is generated inside a domain (Figure 9a). Then, the random field is subsetted inside the individual random shapes (Figures 9b–9d). It should be mentioned that as the shapes and the scalar field inside the shapes are deformed in the course of the sensitivity testing, it is necessary to interpolate them back onto the original grid for calculation of RMSD3D. Although the other metrics can be applied to scalar fields on different spatial grids (point-to-point correspondence is not required), the scalar fields interpolated back onto the original grid are also used for these cases in order for the deformation tested to be consistent across the metrics.

5.1. Scale Test

For all shapes, all metric scores are strictly increasing as scale increases (Figure 10). However, the graph of RMSD3D has a highly nonlinear change in the response to linearly increasing scale. This suggests an unstable behavior of the metric to scaling with over sensitivity to small changes in scale but weak response to scaling for the remainder of the range of tested values. The other metrics maintain a nearly constant rate of the score change suggesting equal sensitivity to the tested scales.

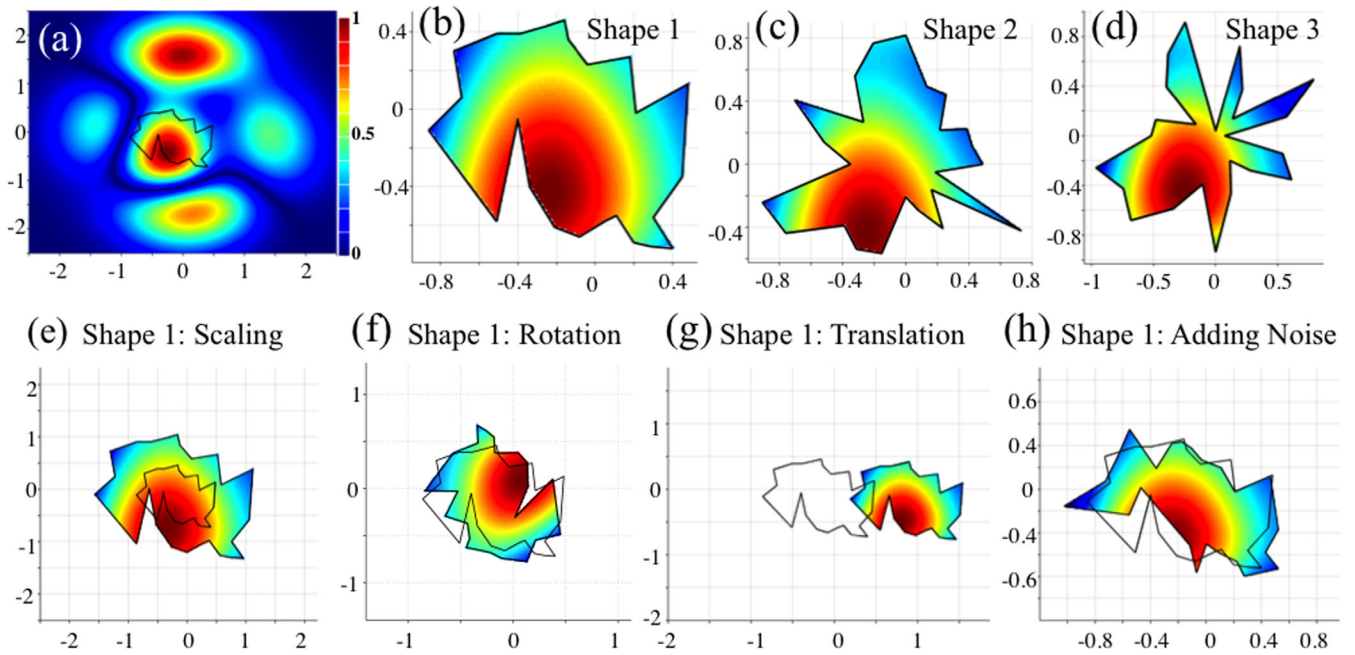

Figure 9. (a) A randomly generated continuous field. (b–d) Concentration fields inside random shapes are obtained by "cutting" the random field with the shapes. For illustration, Shape 1 is shown with the black contour. (b–d) Random shapes (same as in Figures 4a–4c) with concentration fields inside. (e–h) Illustration of sensitivity and robustness tests are shown for Shape 1 similar to Figures 4d–4g. For reference, the empty contour indicates the initial position of the shape.

5.2. Rotation Test

With the exception of MD3D, every metric has an improved response to the rotation test for the 3-D application (Figure 11) compared to the 2-D case (Figure 6). MD3D has two other false "successes" for shapes 1 and 3 (Figures 11a and 11c) that were absent in the 2-D application. The scores for the other metrics are more symmetric about π and monotonically increase over a wider range of the rotation angle ($\pm\pi/3$ and more). Note a nearly perfect response of MHD3D for shape 2 (Figure 11b). The MHD3D score function is symmetric around π and monotonically increasing on [0, π] and monotonically decreasing on [π, 2π]. RMSD3D strongly responds to the initial rotation of the shapes. Rotation by 1° from the original shape results in RMSD3D = 0.3, while all other metrics are $<$0.05. The analysis demonstrates that adding a third dimension has improved performance of the metrics but only MHD3D is close to being completely responsive to rotation.

5.3. Linear Translation Test

All metrics except RMSD3D illustrate a good response to the linear translation test (Figure 12). RMSD3D (Figures 12a, 12e, and 12i) shows an initial large response, with a large score for small translation. However, there is little to no sensitivity in that response, with the metric score varying little as the object translates further from the original position. This behavior stems from the RMSD3D definition (equation (6)) that assumes point-to-point correspondence between the tested and control fields at every (i,j) grid point. After the tested field has been deformed in the course of the sensitivity testing, the point-to-point correspondence is not maintained and the metric calculates the difference between the control field and a random value of the tested field that is now at (i,j) grid point (reflecting the realistic usage of RMSD3D in the context of sea ice validation).

5.4. Random Noise Test

Similar to the 2-D case, both the AD3D and MD3D metrics (Figures 13a, 13f, and 13k and 13c, 13h, and 13m) fail the 3-D noise test. The RMSD3D method demonstrates a very robust response to added noise (Figures 13b, 13g, and 13i) which is a striking improvement from the 2-D noise test with the contours (Figures 8b, 8g, and 8i). HD3D and MHD3D again show a robust response to noise with monotonically increasing scores and a bounded spread through the whole range of noise amplitudes in the test.

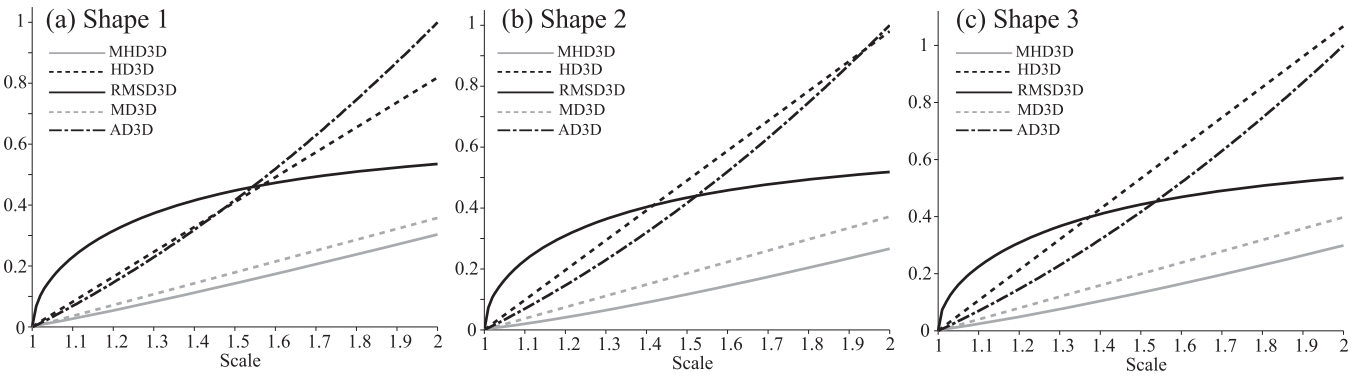

Figure 10. Metric scores versus scale for (a–c) three concentration fields shown in Figures 9a–9c. Different lines represent different skill metrics (see legend). AD3D score has been normalized by the maximum value for demonstration purposes.

In summary, sensitivity tests with spatially distributed random scalar fields inside the contours demonstrate a better performance of the topological metrics compared to the other metrics (Table 1). Both topological methods, HD3D and MHD3D, have passed all tests. Surprisingly, adding the scalar field as a third dimension has degraded performance of HD3D in the noise test, compared to the 2-D case. Again, the AD3D metric is the simplest and least reliable skill metric, passing only the scaling test. RMSD passes the noise test with strikingly improved performance, partially passes the scaling and rotation tests but once again fails the translation test. MD exhibits good responsiveness to the continuous deformations but fails the noise test.

6. Application to Sea Ice Fields

The skill assessment metrics presented in the previous sections are employed to compare sea ice edge contours or shapes (2-D fields) and both sea ice edge contours and distributions of the sea ice concentration within the contours (shape and pattern or 3-D fields). Two sets of the sea ice fields are used to demonstrate the performance of the presented skill metrics in quantifying resemblance of sea ice fields. These include model to model comparison and model to observation comparison. Prior to quantitative comparison, the sea ice fields are ranked based on qualitative analysis. The quantitative ranking of the sea ice fields is compared with that obtained through qualitative analysis.

6.1. Sea Ice Fields From HYCOM-CICE Experiments

The first set of sea ice fields are derived from a suite of numerical experiments with a coupled modeling system that uses the 1/12.5° HYbrid Coordinate Ocean Model (HYCOM) [*Chassignet et al.*, 2006] and Los Alamos Sea Ice

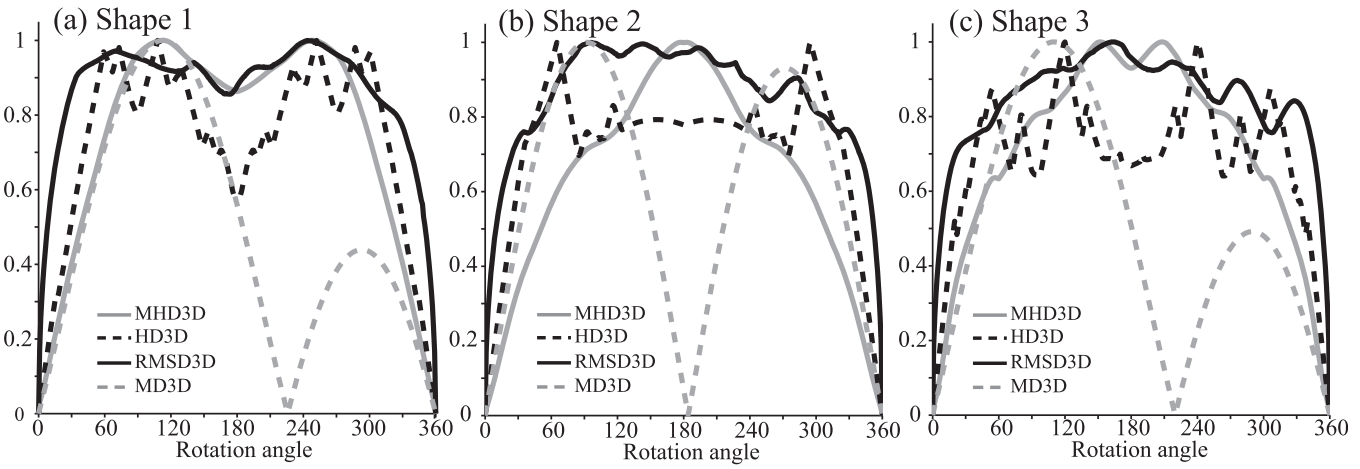

Figure 11. Metric scores versus clockwise rotation angle (degrees) for (a–c) three concentration fields. The AD3D score is not shown, as it remains 0 for all rotation angles. All scores have been normalized by their maximum values.

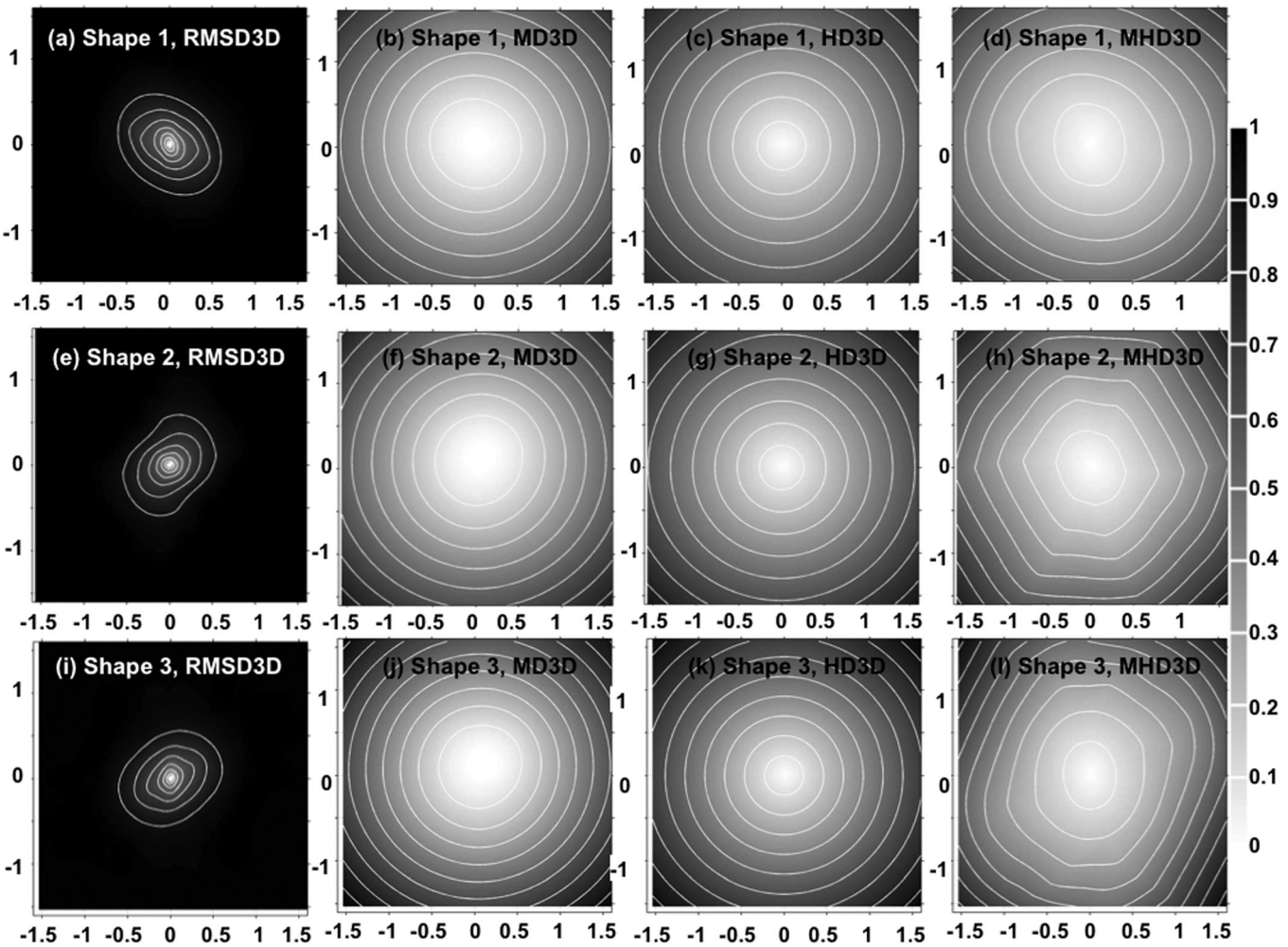

Figure 12. Metric scores for the linear translation test (Figure 9g) that three shapes (in the rows) are subjected to. Each row corresponds to a different shape, and different tests are shown in the columns. The horizontal and vertical axes are distances between the centroids of the shifted and control contours. The zero score corresponds to the initial location of the contour (0,0). (a, e, and i) RMSD3D. (b, f, and j) MD3D. (c, g, and k) HD3D. (d, h, and l) MHD3D. The AD3D metric takes the zero value everywhere and is not shown. The white contours are drawn at intervals of 0.1.

Code (CICE) [*Hunke et al*., 2013] for the Arctic Ocean. Details of the coupled modeling system (hereafter referenced as ARCc0.08) can be found in *Metzger et al*. [2014]. All experiments are initialized from an existing ARCc0.08 data set in October 2005 and run until the end of 2006. Analyzed sea ice fields are derived from identical model configurations forced with different wind fields. The wind stress is a primary driving force of sea ice dynamics and therefore differences in the applied wind fields notably impact the sea ice distribution. The control run is forced with the NCEP Climate Forecast System Reanalysis (CFSR) [*Saha et al*., 2010] winds. Other experiments are driven by winds from (1) experiment 020: National Center for Atmospheric Research (NCAR)—Department of Energy (DOE) reanalysis 2 (NCEPR II) [*Kanamitsu et al*., 2002]; (2) experiment 030: Cross-Calibrated Multi-Platform Ocean Surface Wind Components (CCMP) [*Atlas et al*., 2011]; (3) experiment 040: Arctic System Reanalysis, interim version (ASR) [*Bromwich et al*., 2010]. One more experiment (experiment 050) is forced with winds that are identical to the control run winds (CFSR) but with Greenland river runoff added in this experiment. Experiment 050 is expected to have similar sea ice concentration to the control run, since the changes in the river runoff have a smaller impact on sea ice concentration than discrepancies in the wind fields, at least within the time range of the simulations. Thus, experiment 050 is expected to be the closest to the control run model. The main questions are whether the skill metrics will be able to rank experiment 050 as the best simulation and whether the other experiments will be ranked in accordance with the qualitative ranking described below.

In the numerical experiments, most of the uncertainty in the sea ice fields is related to sea ice concentration in the Nordic and Barents Seas. For the validation analysis, monthly average sea ice concentration fields in

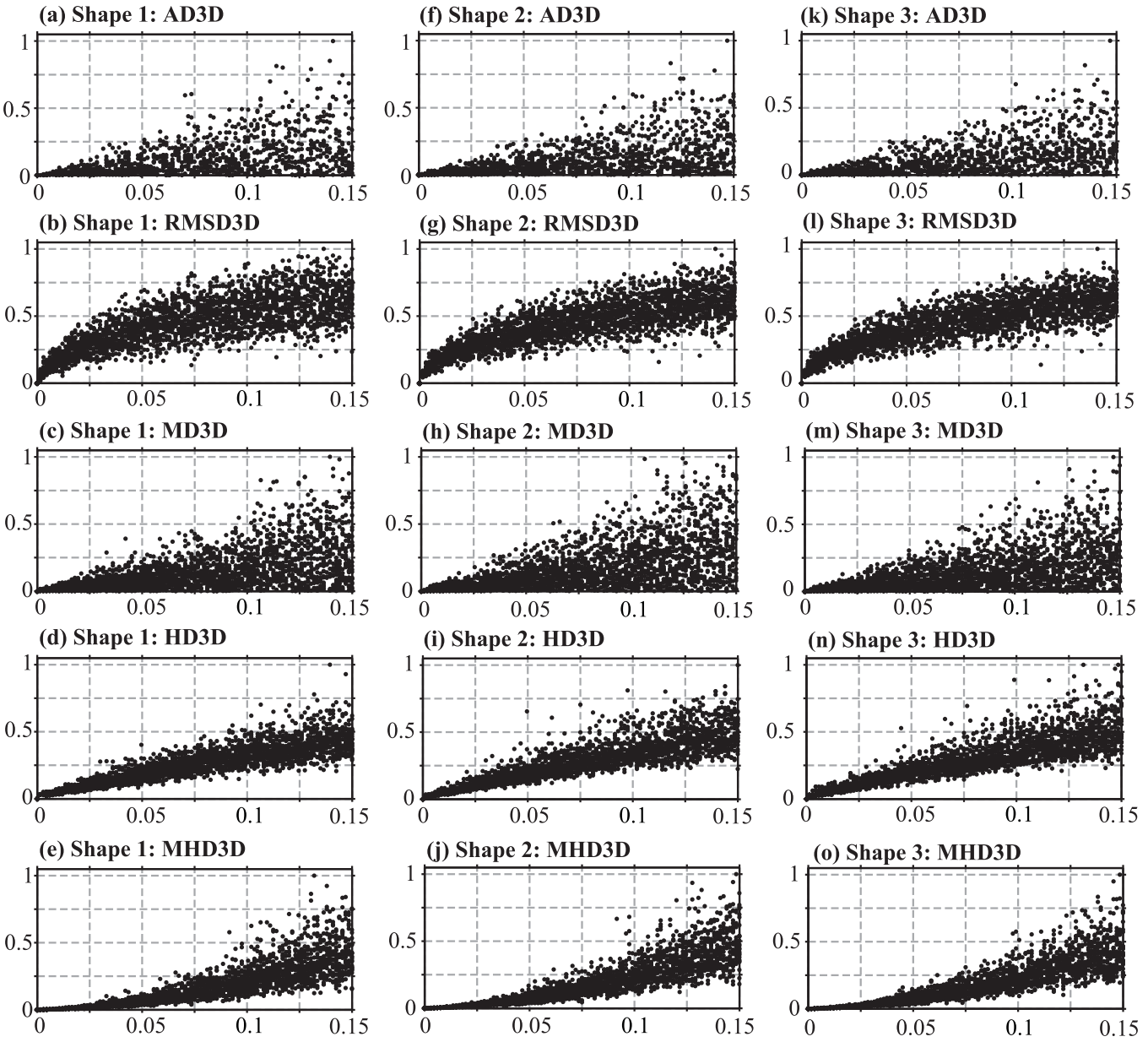

Figure 13. Metric scores (vertical axis) from the noise test for the three concentration field inside the shapes (in columns): (a, f, and k) AD3D, (b, g, and l) RMSD3D, (c, h, and m) MD3D, (d, i, and n) HD3D, and (e, j, and o) MHD3D. The horizontal axis is noise amplitude normalized by the maximum score.

the Nordic and Barents Seas from the model experiments during March 2006 are compared against the control run (Figure 14). The sea ice edge is defined as 0.15 concentration in the open ocean. When the contour intersects land it is forced to follow the coastline. AD is calculated for the region shown in Figure 14 inside the contours excluding land, Baffin Bay, and Canadian Arctic Archipelago.

Visual inspection of the sea ice contours (Figure 14) and sea ice concentration difference (Figure 15d) shows experiment 050 (cyan line in Figure 14) resembles the control run (thick magenta line) almost identically, as expected. From visual inspection of Figure 14, experiment 030 (the green contour) would be ranked, qualitatively, as the second closest to the control run because it follows the control contour closely in the Nordic Seas and western Barents Sea, diverting only in the central and eastern Barents Sea. In contrast to experiment 030, the shapes of the 0.15 concentration contour from experiment 020 (red line in Figure 14) and experiment 040 (blue line) have noticeably larger disagreement with the control run, suggesting lower

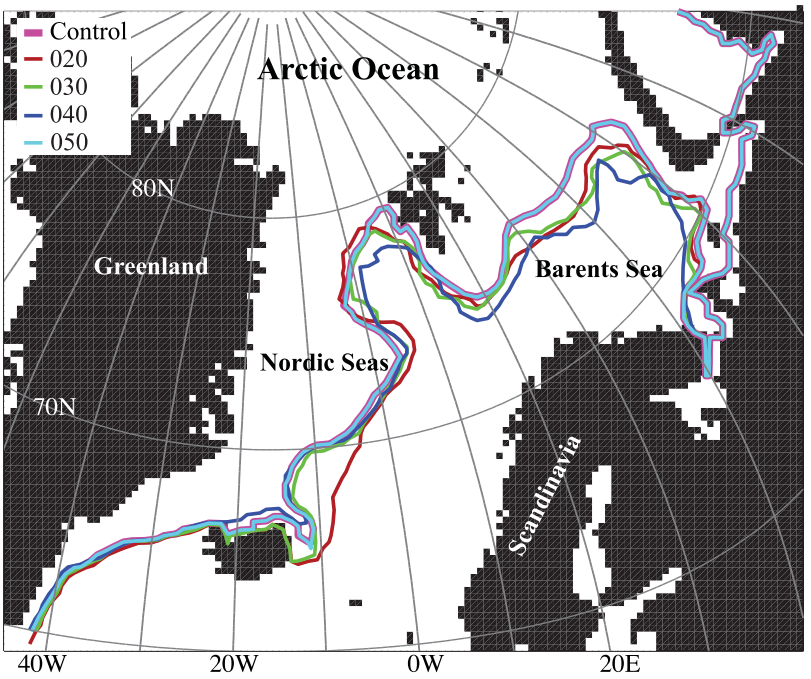

Figure 14. Sea ice edge defined as the 0.15 contour from the time-averaged ice concentration fields simulated in the model experiments in March 2006. Sea ice in Baffin Bay, the Labrador Sea, and the Bering Sea is masked out. The thick magenta line is the 0.15 contour from the control run. Note the cyan and the magenta lines coincide, indicating strong similarity in sea ice concentration between experiment 050 and the control run.

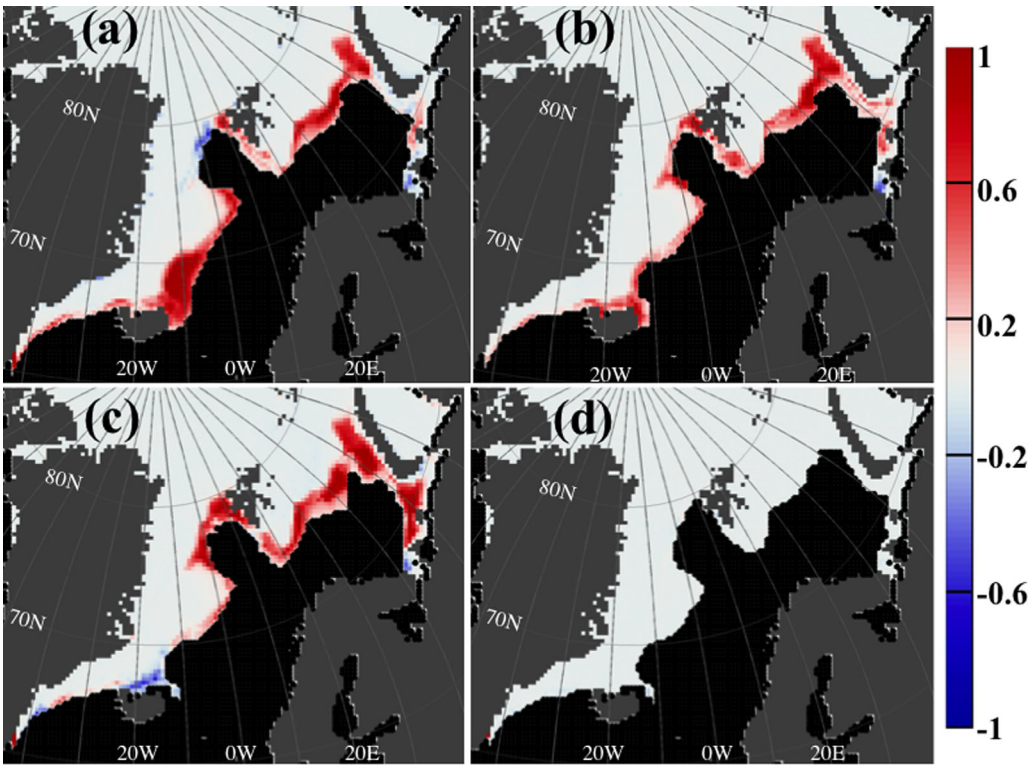

Figure 15. Time-integrated sea ice concentration difference maps between the model experiments for March 2006 (a) 020, (b) 030, (c) 040, (d) 050, and the control run. Grey: land; black: open ocean.

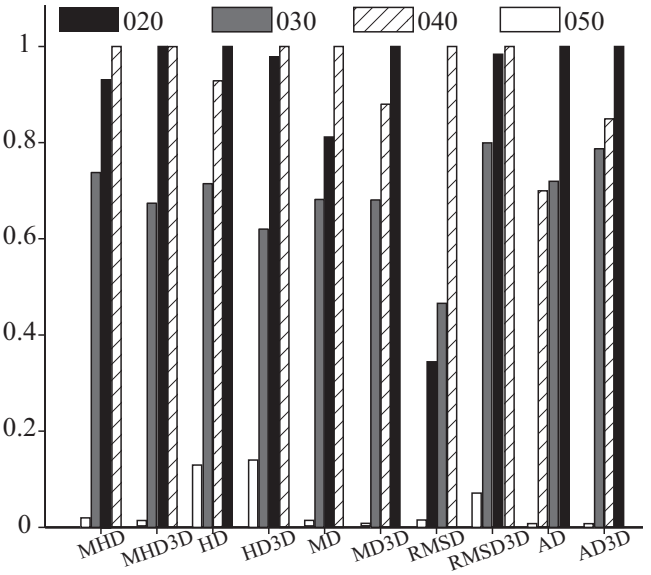

Figure 16. Skill metric scores on the time-averaged ice concentration for March 2006. MHD, HD, MD, RMSD, and AD metrics measure resemblance between the shapes of the sea ice contours. MHD3D, HD3D, MD3D, RMSD3D, and AD3D are metric scores calculated for the sea ice concentration distribution within the 0.15 contour. Within each metric, the model experiments are ordered according to the rankings with the furthest left being the closest to the control run. The models with the most accurate prediction are the furthest left. All metrics identify experiment 050 as the closest simulation to the control run.

ranking of these two experiments. The skill metric scores will be assessed against the qualitative ranking: experiment 050 first, 030 second, and 020 or 040 last.

For 3-D, the biggest disagreement among the models in terms of spatial distributions of the sea ice concentration inside the 0.15 contour is along the ice edge in the Nordic and Barents seas (Figure 15). All experiments but 050 have noticeable discrepancies with the control field, and visual inspection cannot clearly rank the experiments.

Skill metrics are employed to rank the model experiments relative to the control run in order to assess their ranking against the expected ranking (Figure 16). All metrics correctly rank experiment 050 as the closest simulation to the control run in both 2-D and 3-D, in agreement with preliminary visual inspection of the sea ice concentration fields. In 2-D all but the RMSD and AD metrics identify experiment 030 as the second closest to the control run, again in agreement with visual analysis. In contradiction to the visual analysis of Figure 14, both the RMSD and AD metrics give the second highest rank to 020 and 040, respectively. Adding spatial distribution of the sea ice concentration to the skill only has a marginal impact on the scores of MHD3D, HD3D, and MD3D. Both RMSD3D and AD3D return a ranking that agrees with qualitative analysis.

6.2. September 2014 Sea Ice Outlook Experiment

The second set of sea ice data is taken from the Sea Ice Outlook hosted by the Sea Ice Prediction Network (SIPN, http://www.arcus.org/sipn/sea-ice-outlook). Since 2008, the outlook has solicited model predictions of September sea ice extent initialized in the previous months, with the earliest forecasts initialized in May and the latest in early August [*Stroeve et al.*, 2014]. In 2014, forecasts of spatial variables such as sea ice probability (SIP) and Ice-Free Dates were also solicited for the first time. Five models contributed SIP forecasts: NCAR CESM, NASA GMAO (both initialized in May), NOAA CFS (version 2), UW PIOMAS, and SLATER (initialized in August). The first four are dynamical model forecasts, while the SLATER forecast uses a statistical model.

Figures 17a–17e present maps of the sea ice extent probability predicted by different models for September 2014. As outlooks of the actual sea ice concentration were not available, the sea ice edge is approximated by contouring a constant probability (shown with the orange contour). Here for each model, the probabilities that give the best score for the most metrics are selected for contouring (numbers in parentheses in the legend of Figure 17f). The probabilities have been selected iteratively by calculating scores of each metric for every outlook for a range of probability values from [0.25, 0.9] with an increment of 0.05. It is realized that chosen probabilities may not necessarily correspond to a sea ice concentration of 0.15 in the models. While this step may result in the errors for some models being penalized more than other models' errors, the objective here is to demonstrate an application of metrics in ranking real model output, rather than offering a definitive ranking of model performance. Thus, sea ice edges for the metric analysis are derived from these probability maps by contouring predefined probability values.

Only sea ice contours in the interior Arctic Ocean are compared. Isolated patches of ice in the straits of Canadian Arctic Archipelago, Baffin Bay, the Labrador Sea, and Arctic shelf seas are not considered. The skill

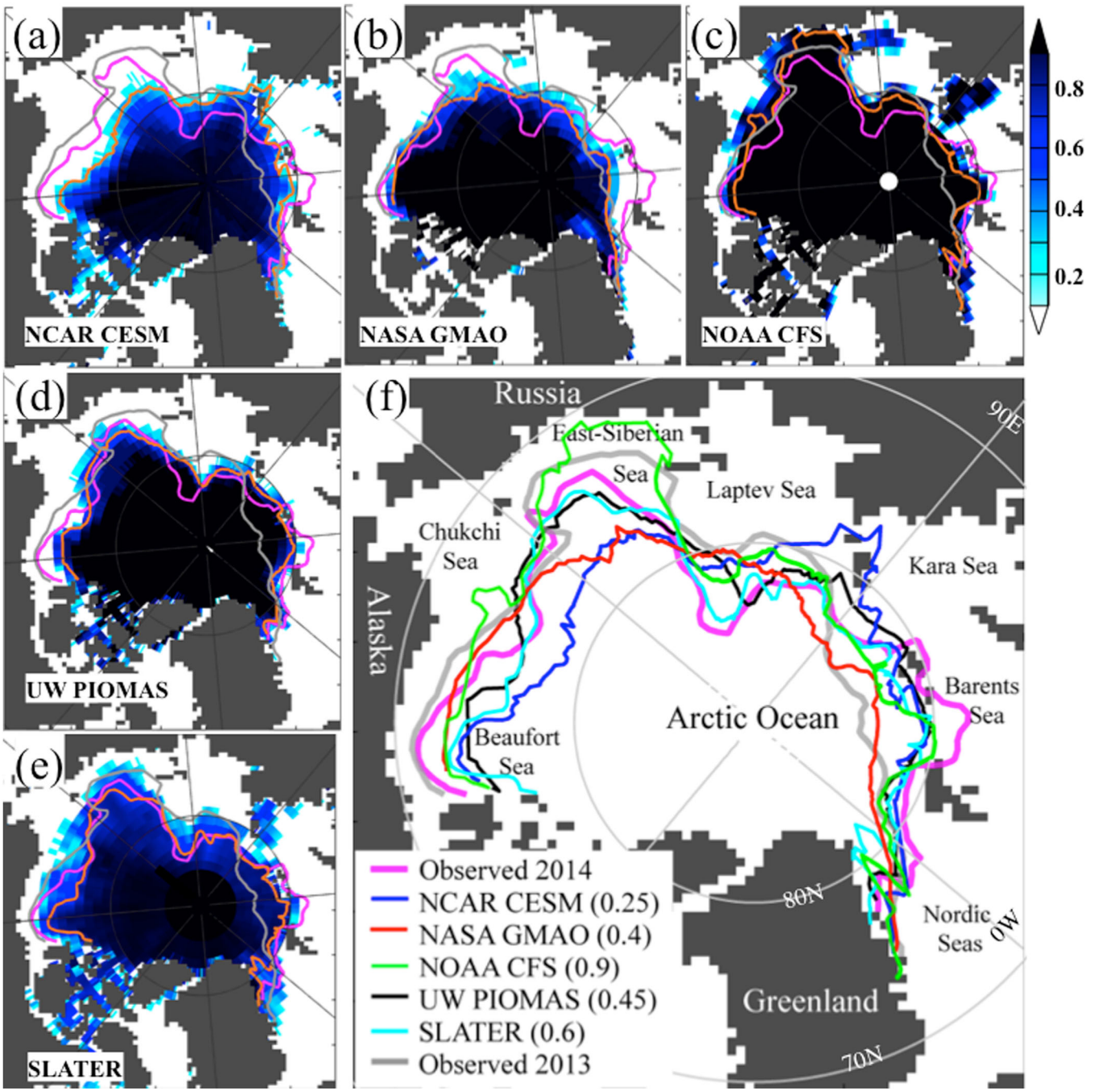

Figure 17. (a–e) Maps of September mean sea ice extent probability predicted by different models for 2014. This is the probability that in each grid cell (1° × 1°), sea ice concentration is above 0.15 (the sea ice edge threshold). Magenta and grey contours are NSIDC sea ice extents for 2014 and 2013, respectively. The orange contour is the sea ice extent estimated from the forecasts used for the comparison. (f) September mean sea ice contours from the forecasts and NSIDC sea ice concentration for 2013 and 2014. Parenthesized numbers in the legend are the probability values used to contour the ice edge from the outlooks.

metrics are employed to evaluate September sea ice mean outlook from the models against the observed September mean sea ice edge derived from the Near-Real-Time daily polar gridded sea ice concentrations at the National Snow and Ice Data Center (NSIDC) for 2014 [*Maslanik and Stroeve*, 1999]. Additionally, metric scores are calculated for the sea ice contour derived from the NSIDC September mean ice concentration field in 2013 (persistence). The outlooks are compared to the persistence scores. It is anticipated that a useful forecast should have smaller errors (higher ranking) than persistence [*Oey et al.*, 2005]. The forecasts that have smaller errors (better resemblance with the control field) than persistence have useful forecasting

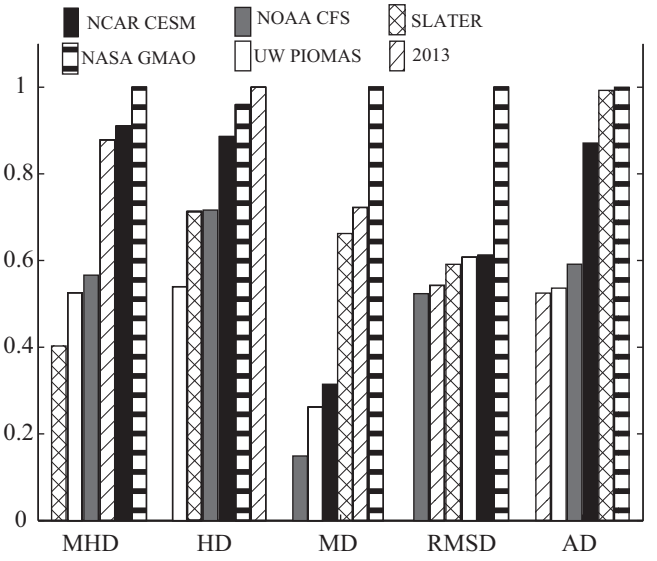

Figure 18. Skill metric scores for September sea ice mean forecasts from the models and from the NSIDC monthly ice concentration in September 2013 (persistence "forecast") grouped by metric. Within each metric group, the models are ordered according to the ranks of the predictions compared to the sea ice edge derived from NSIDC ice concentration field for September 2014. The models with the most accurate prediction are the furthest left. The models to the right of "2013" have failed to outperform the persistence. All metric scores are scaled relative to the maximum score within the group for ease of comparison.

skills. Otherwise, it is better to use the sea ice state from the previous year to predict the next year's sea ice conditions.

Visual inspection of the contours (Figure 15) suggests that predictions by UW PIOMAS and SLATER have the best general resemblance with the NSIDC 2014 (magenta line). The ice contours from both forecasts follow the control contour replicating the distinct protruding part of the ice edge toward the East Siberian Sea noticeable in the observed sea ice. NOAA CFS has some resemblance with the NSIDC September 2014 contour, however the East-Siberian ice tongue extends much farther south all the way to the coast. The persistence contour (NSIDC September 2013, light grey in Figure 15) follows the 2014 ice contour closely in the Canada Basin, with apparent deviation in the Eurasian Basin (north of the Barents and Kara seas). Of all the contours, NASA GMAO and NCAR CESM forecasts look the most different from the observed September ice edge. Based on visual analysis of the contours, the anticipated ranking is SLATER or UW PIOMAS as the closest forecasts, followed by NOAA CFS or persistence (September 2013), and NCAR CESM or NASA GMAO at the end. Note that for this analysis, spatial distribution of sea ice concentration was not available, allowing only 2-D application of the skill metrics.

The MHD metric ranks the contours in the anticipated order (Figure 18) obtained from qualitative analysis of Figure 15f. HD ranks model forecasts in agreement with the visual analysis but persistence receives the lowest skills, meaning that according to this metric all the models outperformed persistence. The ranking from MD is counterintuitive because the first place ranking is given to the forecast from NOAA CFS an obviously inaccurate forecast given the ice edge contour shown in Figure 17c. At the same time, UW PIOMAS and SLATER models receive lower skills. The ranking from RMSD for the closest forecasts contradicts qualitative analysis because SLATER and UW PIOMAS are ranked third and fourth, receiving lower skills than persistence. The AD metric ranks persistence as the best forecast. This result is unsurprising given that there is a very small change in the sea ice area in September 2014 compared to 2013 (~1% decrease) [*Perovich*, 2014]. Nevertheless, despite the very minor change in sea ice area, the shape of the sea ice edge in September 2014 is distinctly different from that in September 2013. Hence, the AD ranking contradicts qualitative analysis. The AD ranking agrees with the expected results on UW PIOMAS (second best) but SLATER unexpectedly receives the second lowest skill, which also contradicts results from the visual inspection of Figure 15.

To conclude, based on ice edge contours derived from the outlooks, all the metrics rank NASA GMAO as the least accurate model outlook but rank the other forecasts differently. MHD once again demonstrates the closet correspondence to the visually derived ranking.

7. Conclusions

As more information about spatial distribution of sea ice characteristics becomes available from remote sensing, there is a potential for improved validation of sea ice models. This demands an automated and objective quantification of similarity or dissimilarity between shapes and patterns in a numerical solution and control data. The potential of five quantitative methodologies for sea ice model validation and

assessment of sea ice model skills in simulating shapes and spatial distribution of sea ice characteristics is discussed.

Three methods commonly used in other applications as quantitative estimates of difference between tested and the control data (AD, RMSD, MD) are evaluated, as well as two topological (HD and MHD) approaches which are well suited to the geometric nature of the data. The corresponding metrics undergo sensitivity tests (scale, rotation, translation, noise, and pattern recognition) in order to assess the ability of the methodologies to quantify similarity between 2-D shapes and spatial distribution of scalar fields (3-D) in sea ice model and control data. The considered metrics are then employed for realistic cases where modeled sea ice extent and concentration fields are compared against control data.

The AD metric is the most basic technique for sea ice skill assessment and has demonstrated the weakest performance for the considered application. Sensitivity tests both for 2-D and 3-D applications have revealed very limited responsiveness of this metric to tested differences. The major drawback of this technique is the likelihood of equally ranking two geometrically different fields or fields with dissimilar spatial distribution of sea ice characteristics but equal area (as also discussed in *Connolley and Bracegirdle* [2007]).

RMSD has been used successfully in the past for estimating uncertainties in simulated sea ice properties. Here the potential of using this method for evaluation of spatial patterns of sea ice characteristics in sea ice models compared to the control field was assessed. The RMSD metric has shown an unsatisfactory performance in the sensitivity tests and realistic application with sea ice data. In general, the metric has poor skills when comparing contours (2-D), particularly for the scaling and translation tests. The metric does have an improved performance for the 3-D case when spatial data are interpolated onto the same grid. However, RMSD3D still has abrupt response to small changes in the contours and shapes in the scale and rotation tests. RMSD does perform well for small noise in the 2-D application and for any noise in the 3-D case. In realistic applications, RMSD (and RMS3D) also underperform.

The poor performance of RMSD is related to the requirement of point-to-point correspondence between the model and the observations, which is an inherent weakness of this method for the discussed application. Specifically, underperformance of RMSD can be in part related to a nonoptimal algorithm of point-to-point matching implemented in this study for 2-D application. The point-to-point matching algorithm could have provided false pair matching between the data sets resulting in abrupt changes in the metric scores during sensitivity testing and realistic application with sea ice data. However, point matching is a serious obstacle in most practical applications due to complex contours and shapes of the sea ice fields.

The MD method is better suited for spatial analysis and it has demonstrated reasonable skills in shape and contour comparison. Compared to AD and RMSD, the metric performed better in the rotation and translation tests. However, this metric can provide unrealistic scores when the sea ice fields differ in small details, and this is related to the faulty response to noise. The major deficiency of this metric is its critical dependence on the reference point (P_0). A choice of P_0 can be dictated by the goals of a study. For instance, if similarity of two shapes is evaluated without the regard to rotation or translation, MD can be calculated relative to the centroids of the contours (P_0 may be different for the contours). When translation and rotation should be penalized, MD will be calculated relative to the same P_0 (a centroid of one of the contours). A more significant problem is that MD can give identical scores to very different contours or patterns when the data points have similar dispersion around the reference point.

Since none of the above metrics have shown high skills in quantification of similarity in spatial shapes and patterns of sea ice fields, two topological approaches have been tested for this purpose. Sensitivity tests and realistic application to sea ice fields have shown that the HD and MHD metrics are reliable methodologies for sea ice application. Both metrics have demonstrated the best sensitivity and robustness to tested differences and good assessment skills in realistic applications. HD is sensitive to outliers, which is undesirable for this application since it tends to inflate dissimilarity between the contours, but this may be a useful property for other applications.

Overall, MHD demonstrates the best response to the tested differences and realistic applications. A further advantage of MHD (and HD) is the ability to operate on contours or surfaces that have a different number of points and with no point-to-point correspondence required, unlike RMSD.

Using the sea ice concentration data in addition to the 2-D spatial data makes the 3-D scores more accurate for most of the tested metrics, since more complete information about verified fields is taken into account.

However, using the third dimension requires special consideration of its appropriate scaling, which is not a trivial task and needs further investigation. Another area that has not been discussed in the study is uncertainty of the model skill estimates.

The study has determined that MHD is a mathematically tractable yet efficient method for model skill assessment and evaluation that has a particular focus on spatial patterns and distribution. It can be effectively applied to objectively evaluate and compare both 2-D and 3-D sea ice characteristics across the models of the Arctic and Antarctic regions. MHD was shown to perform better than the other considered metrics, such as RMSD and AD, and can be easily applied to sea ice model evaluation and assessment. The application of the metric here has been demonstrated on the sea ice concentration fields and can easily be applied to any other sea ice characteristic (e.g., sea ice thickness, distribution of polynyas or melt ponds, ice drift speed). Furthermore, MHD can also be applied to validation of simulated oceanographic fields where both shape and distribution are of importance, for example for river plumes, and oil spill models. It therefore provides an objective and flexible metric that can be utilized both in sea ice and other geophysical applications.

Acknowledgments

Sea ice concentration data used in this study are available at NSIDC. Data set: Near-Real-Time DMSP SSMIS Daily Polar Gridded Sea Ice Concentrations (http://nsidc.org/data/NSIDC-0081). This research was funded by U.S. National Science Foundation (NSF) PLR-0804017, NASA JPL OVWST, and a contract from Bureau of Ocean Energy Management (BOEM) to the FSU (M12PC00003). A. Proshutinsky was funded by NSF projects PLR-0804010, PLR-1313614, and PLR-1203720. J. Ubnoske was funded by a grant from BP/The Gulf of Mexico Research Initiative to the Deep-C Consortium (SA12-12, GoMRI-008). This work was supported by a grant of computer time from the DoD High Performance Computing Modernization Program at NRL SSC. The idea of analysis of objective automated metrics for sea ice model comparison was brought up during discussions at the AOMIP and FAMOS workshops. S. Morey (FSU) was involved in development of presented model validation techniques for other applications. We thank E. Hunke (LANL) and M. Johnson (UAF) for their careful reading of our manuscript and their comments. We acknowledge A. J. Wallcraft (NRL SSC), P. Posey (NRL SSC), and J. Metzger (NRL SSC) for their assistance with the HYCOM-CICE model.

References

Arzel, O., T. Fichefet, and H. Goosse (2006), Sea ice evolution over the 20th and 21st centuries as simulated by current AOGCMs, *Ocean Modell.*, *12*, 401–415.

Atlas, R., R. N. Hoffman, J. Ardizzone, S. M. Leidner, J. C. Jusem, D. K. Smith, and D. Gombos (2011), A cross-calibrated, multiplatform ocean surface wind velocity product for meteorological and oceanographic applications, *Bull. Am. Meteorol. Soc.*, *92*, 157–174, doi:10.1175/2010BAMS2946.1.

Belongie, S., J. Malik, and J. Puzicha (2002), Shape matching and object recognition using shape context, *IEEE Trans. Pattern Anal. Mach. Intel.*, *24*(24), 509–522.

Blanchard-Wrigglesworth, E., and C. M. Bitz (2014), Characteristics of Arctic sea-ice thickness variability in GCMs, *J. Clim.*, *27*(21), 8244–8258.

Bromwich, D., Y.-H. Kuo, M. Serreze, J. Walsh, L. S. Bai, M. Barlage, K. Hines, and A. Slater (2010), Arctic system reanalysis: Call for community involvement, *Eos Trans. AGU*, *91*, 13–14.

Cavalieri, D. J. (1992), The validation of geophysical products using multisensory data, in *Microwave Remote Sensing of Sea Ice, Geophys. Monogr. Ser.*, vol. 68, edited by F. D. Carsey, chap. 11, 462 pp., AGU, Washington, D. C.

Chang, C. C., S. M. Hwang, and D. J. Buehrer (1991), A shape recognition scheme based on relative distances of feature points from the centroid, *Pattern Recognition*, *24*(11), 1053–1063.

Chassignet, E. P., et al. (2006), Generalized vertical coordinates for eddy-resolving global and coastal ocean forecasts, *Oceanography*, *19*, 20–31.

Chui, H., and A. Rangarajan (2003), A new point matching algorithm for non-rigid registration, *Computer Vision and Image Understanding*, *89*, 114–141, doi:10.106/S1077-3142(03)00009-2.

Comiso, J. C., D. J. Cavalieri, C. L. Parkinson, and P. Gloersen (1997), Passive microwave algorithms for sea ice concentration: A comparison of two techniques, *Remote Sens. Environ.*, *60*(3), 357–384.

Connolley, W. M., and T. J. Bracegirdle (2007), An Antarctic assessment of IPCC AR4 coupled models, *Geophys. Res. Lett.*, *34*, L22505, doi:10.1029/2007GL031648.

Daoudi, M., F. Ghorbel, A. Mokadem, O. Avaro, and H. Sanson (1999), Shape distances for contour tracking and motion estimation, *Pattern Recognition*, *32*(7), 1297–1306.

Dubuisson, M.-P., and A. K. Jain (1994a), 2D matching of 3D moving objects in color outdoor scenes, in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 887–891, IEEE, Seattle, Wash.

Dubuisson, M.-P., and A. K. Jain (1994b), A modified Hausdorff distance for object matching, *Pattern Recognition*, *1*, 566–568, doi:10.1109/ICPR.1994.576361.

Emery, W. J., M. Radebaugh, C. W. Folwer, D. Cavallieri, and K. Steffen (1991), A comparison of sea ice parameter computed from advanced very high resolution radiometer and Landsat satellite imagery and from airborne passive microwave radiometry, *J. Geophys. Res.*, *96*(C12), 22,075–22,085.

Germe, A., M. Chevallier, D. S. Y. Melia, E. Sanchez-Gomez, and C. Cassou (2014), Interannual predictability of Arctic sea ice in a global climate model: Regional contrasts and temporal evolution, *Clim. Dyn.*, *43*(9–10), 2519–2538.

Hunke, E. C., and M. Holland (2007), Global atmospheric forcing data for Arctic ice-ocean modeling, *J. Geophys. Res.*, *112*, C04S14, doi:10.1029/2006JC003640.

Hunke, E. C., W. H. Lipscomb, and A. Turner (2010), Sea-ice models for climate study: Retrospective and new directions, *J. Glaciol.*, *56*(200), 1162–1172.

Hunke, E. C., W. H. Lipscomb, A. K. Turner, N. Jeffery, and S. Elliott (2013), CICE: The Los Alamos sea ice model, in *Documentation and Software User's Manual*, 115 pp., Los Alamos Natl. Lab., Los Alamos, N. M.

Huttenlocher, D. P., G. A. Klanderman, and W. J. Rucklidge (1993), Comparing images using the Hausdorff distance, *IEEE Trans. Pattern Anal. Mach. Intel.*, *15*(9), 850–863.

Johnson, M., S. Gaffigan, E. Hunke, and R. Gerdes (2007), A comparison of Arctic Ocean sea concentration among the coordinated AOMIP model experiments, *J. Geophys. Res.*, *112*, C04S11, doi:10.1029/2006JC003690.

Johnson, M., et al. (2012), Evaluation of Arctic sea ice thickness simulated by Arctic Ocean model intercomparison project models, *J. Geophys. Res.*, *117*, C00D13, doi:10.1029/2011JC007257.

Kanamitsu, M., W. Ebisuzaki, J. Woollen, S.-K. Yang, J.J. Hnilo, M. Fiorino, and G. L. Potter (2002), NCEP-DOE AMIP-II Reanalysis (R-2), *Bull. Amer. Meteor. Soc.*, *83*, 1631–1643, doi:10.1175/BAMS-83-11-1631.

Karvonen, J. (2014), A sea ice concentration estimate algorithm utilizing radiometer and SAR data, *Cryosphere*, *8*, 1639–1650, doi:10.5194/tc-8-1639-2014.

Katavouta, A., and P. Myers (2014), Sea-ice concentration multivariate assimilation for the Canadian East coast in a coupled sea ice-ocean model, *Atmos. Ocean*, *52*, 418–433, doi:10.1080/07055900.2014.954096.
Kreyscher, M., M. Harder, and P. Lemke (1997), First results of the sea ice model intercomparison project (SIMIP), *Ann. Glaciol.*, *25*, 8–11.
Kreyscher, M., M. Harder, P. Lemke, and M. Flato (2000), Results of the sea ice model intercomparison project: Evaluation of sea ice rheology schemes for use in climate simulations, *J. Geophys. Res.*, *105*(C5), 11,299–11,320.
Kwok, R. (2010), Satellite remote sensing of sea-ice thickness and kinematics: A review, *J. Glaciol.*, *56*(200), 1129–1140.
Kwok, R., H. J. Zwally, and D. Yi (2004), ICESat observations of Arctic sea ice: A first look, *Geophys. Res. Lett.*, *31*, L16401, doi:10.1029/2004GL020309.
Laxon, S. W., et al. (2013), CyoSat-2 estimates of Arctic sea ice thickness and volume, *Geophys. Res. Lett.*, *40*, 732–737, doi:10.1002/grl.50193.
Lipscomb, W. H., E. C. Hunke, W. Maslowski, and J. Jakacki (2007), Ridging, strength, and stability in high-resolution sea ice models, *J. Geophys. Res.*, *112*, C03S91, doi:10.1029/2005JC003355.
Maslanik, J., and J. Stroeve (1999), Near-Real-Time DMSP SSM/I-SSMIS Daily Polar Gridded Sea Ice Concentrations, NASA DAAC, Natl. Snow and Ice Data Cent., Boulder, Colo., doi:10.5067/U8C09DWVX9LM. [Available at http://nsidc.org/data/nsidc-0081.html.]
Meier, W. N., and J. Stroeve (2008), Comparison of sea-ice extent and ice-edge location estimates from passive microwave and enhanced-resolution scatterometer data, *Ann. Glaciol.*, *48*, 65–70.
Metzger, E. J., et al. (2014), US Navy operational global ocean and Arctic ice prediction systems, *Oceanography*, *27*(3), 32–43.
Oey, L.-Y., T. Ezer, G. Forristall, C. Cooper, S. DiMarco, and S. Fan (2005), An exercise in forecasting loop current and eddy frontal positions in the Gulf of Mexico, *Geophys. Res. Lett.*, *32*, L12611, doi:10.1029/2005GL023253.
Parkinson, C. L., K. Y. Vinnikov, and D. J. Cavalieri (2006), Evaluation of the simulation of the annual cycle of Arctic and Antarctic sea ice coverages by 11 major global climate models, *J. Geophys. Res.*, *111*, C07012, doi:10.1029/2005JC003408.
Perovich, D., S. Gerland, S. Hendricks, W. Meier, M. Nicolaus, and M. Tschudi (2014), Sea Ice, in Arctic Report Card: Update for 2014, NOAA, Md. [Available at http://www.arctic.noaa.gov/reportcard/sea_ice.html.]
Peterson, K. A., A. Arribas, H. T. Hewitt, A. B. Keen, D. J. Lea, and A. J. McLaren (2014), Assessing the forecast skill of Arctic Sea ice extent in the GloSea4 seasonal prediction system, *Clim. Dyn.*, *44*, 147–162.
Proshutinsky, A., et al. (2005), Arctic Ocean study—Synthesis of model results and observations, *Eos Trans. AGU*, *86*(40), 368–371.
Proshutinsky, A., et al. (2011), Recent advances in Arctic Ocean studies employing models from the Arctic Ocean model intercomparison project, *Oceanography*, *24*(3), 102–113.
Rucklidge, W. J. (1997), Efficient locating objects using Hausdorff distance, *Int. J. Comput. Vision*, *24*(3), 251–270.
Saha, S., et al. (2010), The NCEP climate forecast system reanalysis, *Bull. Am. Meteorol. Soc.*, *91*, 1015–1057, doi:10.1175/2010BAMS3001.1.
Schweiger, A., R. Lindsay, J. Zhang, M. Steele, H. Stern, and R. Kwok (2011), Uncertainty in modeled Arctic sea ice volume, *J. Geophys. Res.*, *116*, C00D06, doi:10.1029/2011JC007084.
Serreze, M. C., M. M. Holland, and J. Stroeve (2007), Perspectives on the Arctic's shrinking sea-ice cover, *Science*, *315*(5818), 533–1536, doi:10.1126/science.1139426.
Stroeve, J., L. C. Hamilton, C. M. Bitz, and E. Blanchard-Wrigglesworth (2014), Predicting September sea ice: Ensemble skill of the SEARH sea ice outlook 2008–2013, *Geophys. Res. Lett.*, *41*, 2411–2418, doi:10.1002/2014GL059388.
Tietsche, S., J. J. Day, V. Guemas, W. J. Hurlin, S. P. E. Keeley, D. Matei, R. Msadek, M. Collins, and E. Hawkins (2014), Seasonal to interannual Arctic sea ice predictability in current global climate models, *Geophys. Res. Lett.*, *41*, 1035–1043, doi:10.1002/2013GL058755.
Tsamados, M., D. L. Feltham, D. Schroeder, D. Flocco, S. L. Farrell, N. Kurtz, S. W. Laxon, and S. Bacon (2014), Impact of variable atmospheric and oceanic form drag on simulations of Arctic Sea ice, *J. Phys. Oceanogr.*, *44*, 1329–1353, doi:10.1175/JPO-D-13-0215.1.
Turner, J., T. J. Bracegirdle, T. Phillips, G. J. Marshall, and J. S. Hosking (2013), An initial assessment of Antarctic Sea ice extent in the CMIP5 models, *J. Clim.*, *26*, 1473–1484, doi:10.1175/JCLI-D-12-00068.1.
Uotila, P., S. O'Farrell, S. J. Marsland, and D. Bi (2012), A sea-ice sensitivity study with a global ocean-ice model, *Ocean Modell.*, *51*, 1–18.
Uotila, P., S. O. Farrell, S. J. Marsland, and D. Bi (2013), The sea-ice performance of the Australian climate models participating in the CMIP5, *Aust. Meteorol. Oceanogr. J.*, *63*, 113–136.
Uotila, P., P. R. Holland, T. Vihma, S. J. Marsland, and N. Kimura (2014), Is realistic Antarctic sea-ice extent in climate models the result of excessive ice drift?, *Ocean Modell.*, *79*, 33–42.
Vaughan, D. G., et al. (2013), Observations: Cryosphere, in *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, edited by T. F. Stocker et al., pp. 317–382, Cambridge Univ. Press, Cambridge, U. K.
Zhang, D., and G. Lu (2004), Review of shape representation and description techniques, *Pattern Recognition*, *37*(1), 1–19.
Zhang, X., and J. E. Walsh (2006), Towards a seasonally ice-covered Arctic Ocean: Scenarios from the IPCC AR4 model simulations, *J. Clim.*, *19*, 1730–1747.