Metrics for evaluating linear features

Gad Levy,[1] Max Coon,[1] Giang Nguyen,[2,3] and Deborah Sulsky[2]

Received 19 June 2008; revised 10 September 2008; accepted 22 September 2008; published 15 November 2008.

[1] Novel metrics designed to evaluate the skill of geophysical models in simulating discontinuities and linear features are introduced and tested using a sea-ice model and remotely sensed observations of leads. The metrics are formulated as frequency-based indices of agreement, thus maintaining the simplicity desired in model development and operational applications, while remedying known shortcomings of common existing skill metrics. User-selectable spatial scales and features of significance allow for the general use of the metrics in variable applications, scales, and observation types. **Citation:** Levy, G., M. Coon, G. Nguyen, and D. Sulsky (2008), Metrics for evaluating linear features, *Geophys. Res. Lett.*, *35*, L21705, doi:10.1029/2008GL035086.

1. Introduction

[2] This paper is concerned with developing metrics to measure the accuracy of numerical simulations where predictions of the location and extent of sharp gradients and discontinuities is of importance. Sharp gradients and discontinuities are lower dimensional features contained within a bulk simulation. For example, a discontinuity might be represented by a one-dimensional curve in a two-dimensional simulation. We call such features 'linear features.'

[3] Linear features occur frequently in geophysical applications and often represent discontinuities that are associated with important physical and dynamic processes. For example, a boundary between two land–surface types, atmospheric and oceanic fronts, and the orientation, location, and amount of opening of leads in pack ice and polynyas, all fundamentally affect interfacial fluxes and thus impact climate [e.g., *Koster and Suarez*, 1992; *Levy and Vickers*, 1999; *Lüpkes et al.*, 2008]. Linear features are also important on smaller scales. Leads, for example, are important for wildlife. Seals, whales, penguins, and other animals rely on leads for access to oxygen. *Grumbine* [1998] notes that the sea-ice literature includes relatively little in the way of quantitative model verification, giving the explanation that visual inspection of model output has been sufficiently unambiguous to determine which model or parameterization is better. *Coon et al.* [2007] call for a sea-ice model that would explicitly resolve discontinuities representing leads, and argue that such a model would require metrics for model verification and validation against observations.

[4] The most common metrics used to verify geophysical model skill are adaptations of a least-square metric used in operational numerical weather prediction data assimilation systems. They score a forecast skill based on the mean square error of a model variable with respect to observed values, summed over all grid points in a discretization. This basic measure of accuracy is then normalized to form an agreement index. In operational use, the agreement index is often incorporated into a skill-score by comparison to the agreement index of a reference state, commonly climatology or persistence. These metrics are most appropriate for continuous fields where the observed and model variables are commensurate (i.e., measured with the same units). They, along with correlation-based indices are appraised as measures of model accuracy by several authors [e.g., *Willmott et al.*, 1985; *Murphy and Epstein*, 1989; *Potts et al.*, 1996; *Mason*, 2004].

[5] These least-square metrics are flawed when used in the presence of sharp gradients and discontinuities. Their use to score a model's skill in resolving these linear features is practiced but is problematic. Typically, only an indirect measure is provided of the model's ability to resolve the linear features through proxy variables that are often poorly or indirectly observed or measured. Because the linear features occupy only a small area of the domain, models that fail to predict/diagnose these important features may actually score better by these metrics than those that simulate the dynamically important features albeit at a somewhat different location or orientation than observations indicate [e.g., *Mass et al.*, 2002]. Moreover, when the scored continuous variables are validated against non-synoptic and/or satellite observations, discrepancies in the location of linear features between model prediction/analysis and observations may result in the rejection of good observations as outliers in a quality controlled data assimilation system. *Mass et al.* [2002] discuss in detail a number of examples, in the context of Numerical Weather Prediction, where current verification metrics fail to adequately measure model skill and conclude with a call for additional verification tools.

[6] Here we consider two novel metrics proposed for verification of a sea-ice model that explicitly resolves discontinuities [*Coon et al.*, 2007]. We then demonstrate their use by scoring a limited number of model simulations, where visual inspection does not provide clear guidance as to what model run is better. The simulations are measured against satellite SAR observations from RADARSAT 1 (Figure 1) that exhibit Linear Kinematic Features (LKFs) [*Kwok*, 2001]. We also consider properties of the metrics and their suitability for broader general use as spatial

[1] NorthWest Research Associates, Redmond, Washington, USA.
[2] Department of Mathematics and Statistics, University of New Mexico, Albuquerque, New Mexico, USA.
[3] Now at the School of Civil Engineering, University of Sydney, Sydney, New South Wales, Australia.

0094-8276/08/2008GL035086$05.00

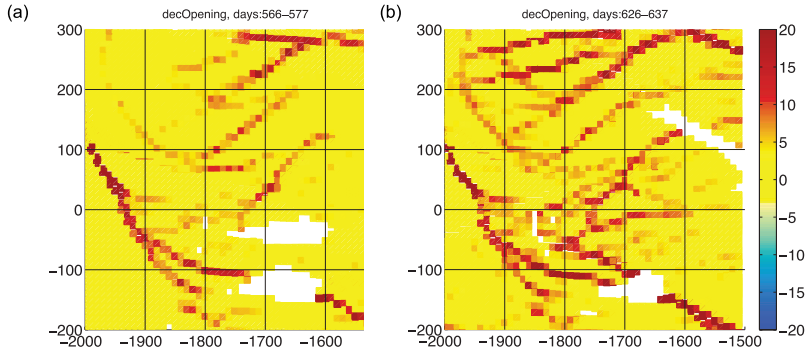

Figure 1. LKFs interpreted from the RADARSAT Geophysical Processor System (RGPS) data on (a) 26 February 2006 and (b) 4 March 2006 corresponding to Day 57 and Day 63 shown in Figures 2 and 3, respectively. The RGPS data were processed assuming that all deformation should be accounted for by shearing, opening and closing of a LKF, which passes through the cell center [*Coon et al.*, 2007].

metrics for linear features. The new metrics are expected to be especially useful for model development and in situations where patterns of linear features are complex and important, such as with modeling sea ice.

2. Metrics for Linear Features

2.1. Formulation

[7] Our main focus is on metrics that can be used for model development and testing in geophysical situations where linear features as defined above are important. Consider the simulation (Figure 2) of small-scale features seen in high-resolution (10 km) synthetic aperture radar imagery (Figure 1). The ability of a model to simulate some of the details and characteristics (e.g., direction, exact location) may change in importance depending on the process or application. Thus, a general metric would ideally be able to measure model accuracy in representing linear features while weighing more heavily those characteristics and spatial scales considered of importance to the physical process being studied. Used consistently, such a metric will quantify improvement or deterioration in model representation of the linear features as the model is modified and refined. For ease of implementation in model development, inter-comparison, and operational applications, it is also desired that the metric be simple and be easily incorporated into an established standard skill score.

[8] We describe two metrics specifically designed to measure model accuracy in representing linear features. While bearing similarity to metrics and indices commonly used in meteorology, the proposed indices circumvent the requirement that the observations and model variables be commensurate by considering the frequencies of features of interest or importance. Also, the user can select the significant features and spatial scales included in the metrics and weigh them appropriately. These properties allow for the general use of these metrics in different and varied applications, with varying scales and observation types.

[9] For both metrics we consider L features of interest and N spatial regions within the domain (sub-domains). The numbers L and N are user or application selectable and are determined based on the processes, resolution, and spatial scales considered important. We define the fractional index of agreement as:

$$I_f = \left(\sum_{i=1}^{L} w_i \sum_{j=1}^{N} \omega_j F_{ij}\right) \Big/ \left(\sum_{i=1}^{L} w_i \sum_{j=1}^{N} \omega_j\right) \tag{1}$$

where F_{ij} is the fractional agreement in terms of grid-cell count of the ith feature in the jth sub-domain. For example, if a feature i is simulated in two grid cells in sub-domain j but is observed in four cells in the same sub-domain, $F_{ij} = .5$. w_i and ω_j are the weights given to the features and the spatial segments, respectively.

[10] We now define a second metric, the RMS index of agreement to evaluate model success in representing features. It, too, treats features through a frequency distribution at predetermined spatial regions of the domain. It contains a term with the familiar format of standard error common in routine distance or root mean square error measures used for continuous variables:

$$I_R = 1 - \sqrt{\left(\sum_{i=1}^{L} w_i \sum_{j=1}^{N} \omega_j D_{ij}\right) \Big/ \left(\sum_{i=1}^{L} w_i \sum_{j=1}^{N} \omega_j\right)} \tag{2}$$

where w_i and ω_j are the same as for the fractional index of agreement, and $D_{i,j}$ is a normalized frequency difference function: $D_{i,j} = \frac{(p_{i,j} - o_{i,j})^2}{(p_{i,j} + o_{i,j})^2}$, where $p_{i,j}$ and $o_{i,j}$ are the predicted (simulated) and observed feature frequencies (cell count) of feature i in sub-domain j.

[11] It can be easily verified that both metrics are non-dimensional and yield scores that range in values between 0 (no agreement) and 1 (perfect agreement).

2.2. An Illustrative Example

[12] In this subsection we test the metrics by scoring six simulations generated through sensitivity tests of a sea-ice model that explicitly models the LKFs in sea ice as discontinuities using an elastic-decohesive constitutive model [*Schreyer et al.*, 2006; *Sulsky et al.*, 2007]. The simulations are of ice behavior in the Beaufort Sea over time intervals of up to 16 days. In the examples shown in

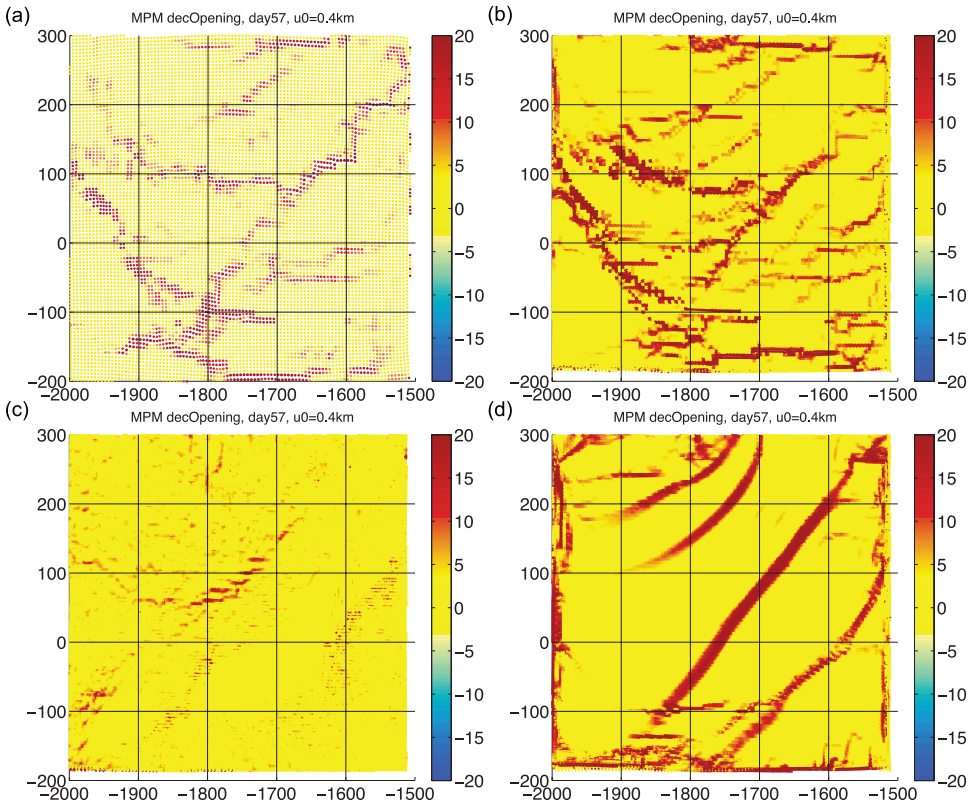

Figure 2. Four different simulation results of a 500 km x 500 km region in the Beaufort Sea scored on a 100 km grid three days after initialization (day 57). Using both the fractional index of agreement and the RMS index of agreement, the simulations score: (a) $I_f = 0.72$, $I_R = 0.79$; (b) $I_f = 0.66$, $I_R = 0.73$; (c) $I_f = 0.54$, $I_R = 0.61$; and (d) $I_f = 0.5$, $I_R = 0.58$. Figure 2a is lower resolution than the others. The different simulations were generated as part of sensitivity tests in which initial conditions, boundary conditions, and material strength parameters were varied. LKFs for the different simulations are scored against RADARSAT observation shown in Figure 2a.

Figures 2 and 3, we consider model output at two different times, six days apart, on day 57 (26 Feb.) and day 63 (4 March) of 2006, for which RADARSAT SAR observations processed through the RADARSAT Geophysical Processor System (RGPS) [*Kwok*, 1998] at 10 km resolution are available for verification (Figure 1).

[13] Figure 2a shows a 250,000 km^2 region on day 3 of the simulation (day 57), cut from a larger 831,600 km^2 simulation of the Beaufort Sea. This simulation uses information from RGPS at 10 km resolution to input an initial pattern of leads and is run on a 10 km^2 grid. The top right panel (Figure 2b) uses the same initialization but performs the simulation directly only on a 500 km $\times$ 500 km domain, using a 5 km^2 grid, where boundary values of the displacement are taken from RGPS data. The simulations presented on the bottom of Figure 2 are similarly performed directly only on a 500 km $\times$ 500 km domain using a 5 km^2 grid. On the left (Figure 2c) there is no initialization and the boundaries are treated as stress-free. On the right (Figure 2d) there is no initialization, and boundary conditions from RGPS data are applied, but the strength has been reduced to 30% of the value in the other simulations. The direct 500 km $\times$ 500 km simulations with specified boundary conditions (Figures 2b and 2d) suffer from boundary effects.

[14] We score these different simulations using both indices against the LKFs interpreted from the RGPS observations of the same section of Beaufort Sea. We consider the agreement in (1) the existence of 100 km $\times$ 100 km grid cells containing LKFs in the domain (subscript e below); and (2) the existence of 100 km $\times$ 100 km grid cells containing LKFs at the observed orientation using four cardinal orientations in the domain (subscripts |, _,/, and \ below). We thus score $L = 5$ features (e, |, _,/, and \) at $N = 1$, the entire simulated/observed domain (in this case the 500 km $\times$ 500 km section of the Beaufort Sea). Here we assume that all weights, w, $\omega = 1$, and the basic spatial scale of interest is of order 10,000 km^2. Frequencies are defined in terms of cell count, and cells with missing data are excluded from the statistics. Equations (1) and (2) respectively can then be written for this case as:

$$I_f = \frac{1}{5}\left(F_e + F_| + F_- + F_/ + F_\right) \text{ and}$$

$$I_R = 1 - \sqrt{\frac{1}{5}\left(D_e + D_| + D_- + D_/ + D_\right)}$$

[15] The resulting scores of this implementation range between 0.5 (0.58) and 0.72 (0.79) for the fractional (RMS) index. The scores are summarized in Figures 2 and 3, where they can also be assessed qualitatively for visual agreement with the observations in Figure 1. We note that the scores in

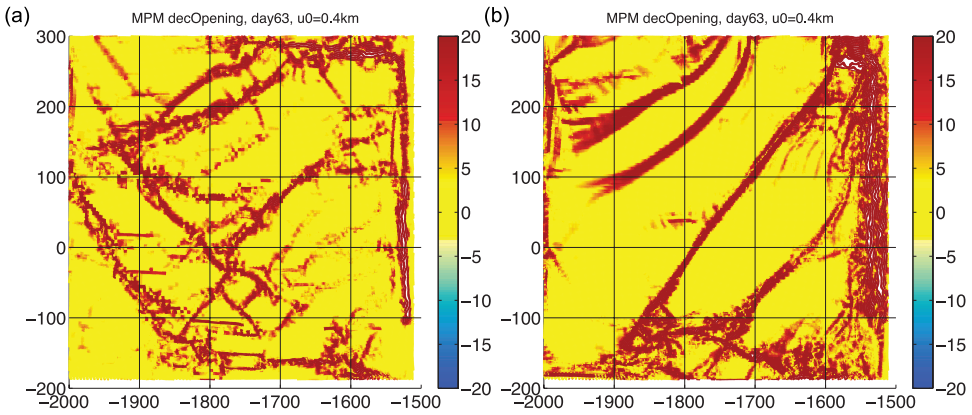

Figure 3. As in Figures 2b and 2d but for day 63. The different simulations were generated with different initial conditions and material strength values. LKFs for these different simulations are scored against RGPS observation shown in Figure 1b. Using both the fractional index of agreement (FI) and the RMS index of agreement (RI), the simulations score: (a) $I_f = 0.63$, $I_R = .71$ (left); and (b) $I_f = 0.51$, $I_R = 0.58$ (right).

this implementation are insensitive to the spatial arrangement of the basic cells or the resolution of the simulation. Other implementations can include increasing N and nesting the sub-domains and applying different spatial filters. Although differently scaled, both indices are capable of quantifying the gradual change in agreement between the model and the observations even when such changes are not immediately apparent through visual inspection. Using the same criteria, the two indices produce remarkably consistent scores that are highly correlated at a high significance level (R and $R^2 = 0.99$). The strong correlation between the two measures is maintained for additional cases and implementations (not shown).

3. Concluding Remarks

[16] Novel metrics designed to evaluate the skill of geophysical models in simulating discontinuities and linear features are introduced and discussed. They are successfully tested using a sea-ice model and remotely sensed observations of LKFs in examples containing a complex pattern of LKFs that pose a challenge to common existing metrics. The metrics are formulated as frequency-based indices of agreement, thus maintaining the simplicity desired in model development and operational applications while remedying known shortcomings of common existing skill metrics. The tested form weighs the different features equally and considers the entire domain, thus yielding scores that are insensitive to the spatial arrangement of the basic cells. However, in the general case, the weights and the subdomains could be adjusted according to the relative importance assigned or desired of a specific feature or scale, as well as for known errors or biases. For example, one could set w_i and ω_j to be different than 1 when accuracy is considered more important for some features, for deformed or uneven grid cells or when observations and simulations are temporally or spatially separated. These user-selectable spatial scales and features of significance along with the ease of incorporating both metrics in generic skill score measures allow for the general use of the metrics in a variety of applications, and with different scales and observation types.

[17] **Acknowledgments.** This work was supported by Minerals Management Service and the National Aeronautics and Space Administration (under contract NNH04 CC 45C), and the National Science Foundation (under grants ARC-0621173 and ATM- 0741832).

References

Coon, M., R. Kwok, G. Levy, M. Pruis, H. Schreyer, and D. Sulsky (2007), Arctic Ice Dynamics Joint Experiment (AIDJEX) assumptions revisited and found inadequate, *J. Geophys. Res.*, *112*, C11S90, doi:10.1029/2005JC003393.

Grumbine, R. W. (1998), Virtual floe ice drift forecast model intercomparison, *Weather Forecast.*, *13*, 886–890.

Koster, R. D., and M. J. Suarez (1992), A comparative analysis of two land surface heterogeneity representations, *J. Clim.*, *5*, 1379–1390.

Kwok, R. (1998), The RADARSAT geophysical processing system, in *Analysis of SAR Data of the Polar Oceans: Recent Advances*, edited by C. Tsatsoulis, and R. Kwok, pp. 235–257, Springer, Berlin.

Kwok, R. (2001), Deformation of the Arctic Ocean sea ice cover between November 1996 and April 1997: A qualitative survey, in *IUTAM Symposium on Scaling Laws in Ice Mechanics and Ice Dynamics*, edited by J. P. Dempsey, and H. H. Shen, pp. 315–322, Kluwer Acad., Dordrecht, Netherlands.

Levy, G., and D. Vickers (1999), Surface fluxes from satellite winds: Modeling air-sea flux enhancement from spatial and temporal observations, *J. Geophys. Res.*, *104*, 20,639–20,650.

Lüpkes, C., T. Vihma, G. Birnbaum, and U. Wacker (2008), Influence of leads in sea ice on the temperature of the atmospheric boundary layer during polar night, *Geophys. Res. Lett.*, *35*, L03805, doi:10.1029/2007GL032461.

Mason, S. J. (2004), On using "climatology" as a reference strategy in the Brier and ranked probability skill scores, *Mon. Weather Rev.*, *132*, 1891–1895.

Mass, C. F., D. Ovens, K. Westrick, and B. A. Colle (2002), Does increasing horizontal resolution produce more skillful forecasts?, *Bull. Am. Meteorol. Soc.*, *83*, 407–430.

Murphy, A. H., and E. Epstein (1989), Skill scores and correlation coefficients in model verification, *Mon. Weather Rev.*, *117*, 572–581.

Potts, J. M., C. K. Folland, I. T. Jolliffe, and D. Sexton (1996), Revised "LEPS" scores for assessing climate model simulations and long-range forecasts, *J. Clim.*, *9*, 34–53.

Schreyer, H. L., D. L. Sulsky, L. B. Munday, M. D. Coon, and R. Kwok (1985), Elastic-decohesive constitutive model for sea ice, *J. Geophys. Res.*, *111*, C11S26, doi:10.1029/2005JC003334.

Sulsky, D., H. Schreyer, K. Peterson, R. Kwok, and M. Coon (2007), Using the material-point method to model sea ice dynamics, *J. Geophys. Res.*, *112*, C02S90, doi:10.1029/2005JC003329.

Willmott, C. J., S. G. Ackleson, R. E. Davis, J. J. Feddema, K. M. Klink, D. R. Legates, J. O'Donnell and C. M. Rowe (1985), Statistics for the evaluation and comparison of models, *J. Geophys. Res.*, *90*, 8995–9005.

M. Coon and G. Levy, NorthWest Research Associates, 4118 148th Ave NE, Redmond, WA 98052, USA. (gad@nwra.com)

G. Nguyen, School of Civil Engineering, University of Sydney, Sydney NSW 2006, Australia.

D. Sulsky, Department of Mathematics and Statistics, University of New Mexico, Albuquerque, NM 87131-0001, USA.