A Novel Set of Geometric Verification Test Fields with Application to Distance Measures

ERIC GILLELAND,[a] GREGOR SKOK,[b] BARBARA G. BROWN,[a] BARBARA CASATI,[c] MANFRED DORNINGER,[d] MARION P. MITTERMAIER,[e] NIGEL ROBERTS,[e] AND LAURENCE J. WILSON[c]

[a] *National Center for Atmospheric Research, Boulder, Colorado*
[b] *University of Ljubljana, Faculty of Mathematics and Physics, Ljubljana, Slovenia*
[c] *Environment and Climate Change Canada, Montreal, Canada*
[d] *University of Vienna, Vienna, Austria*
[e] *MetOffice, Exeter, United Kingdom*

(Manuscript received 6 August 2019, in final form 26 December 2019)

ABSTRACT

As part of the second phase of the spatial forecast verification intercomparison project (ICP), dubbed the Mesoscale Verification Intercomparison in Complex Terrain (MesoVICT) project, a new set of idealized test fields is prepared. This paper describes these new fields and their rationale and uses them to analyze a number of summary measures associated with distance and geometric-based approaches. The results provide guidance about how they inform about performance under various scenarios. The new case comparisons are grouped into four categories: (i) pathological situations such as when a variable is zero valued at all grid points; (ii) circular events aimed at evaluating how different methods handle contrived situations, such as equal but opposite translations, the presence of multiple events of same/different size, boundary effects, and the influence of the positioning of events in the domain; (iii) elliptical events representing simplified scenarios that mimic commonly encountered weather phenomena in complex terrain; and (iv) cases aimed at analyzing how the verification methods handle small-scale scattered events, very large events with holes (e.g., a small portion of clear sky on a cloudy overcast day), and the presence of noise in one or both fields. Results show that all analyzed measures perform poorly in the pathological setting. They are either not able to provide a result at all or they instigate a special rule to prescribe a value resulting in erratic results. The analysis also showed that methods provide similar information in many situations, but that each has its positive properties along with certain unique limitations.

1. Introduction

A common problem encountered in many forecasts is an offset in the predicted position of an event relative to where it actually occurred. When using a nonspatial verification metric (i.e., a metric that only compares values of the observed and forecasted field at collocated grid points), such a forecast can incur a double penalty if the event position is sufficiently displaced in the forecast (cf. Mass et al. 2002; Rossa et al. 2008). Another problem is that the nonspatial measures do not distinguish between a small displacement and a much larger one. These combined issues tend to provide results that are inconsistent with a subjective evaluation of the forecast made by a human analyst. The need to reconcile subjective evaluations of forecast quality with objective measures of forecast accuracy has been the main motivation in researching and developing diagnostic methods for spatial verification that more adequately reflect their worth (Brown et al. 2011).

In the last decade and a half, a range of new spatial verification methods have been developed. The first spatial forecast verification intercomparison project (ICP; Ahijevych et al. 2009; Gilleland et al. 2009, 2010) employed both real case studies and idealized test fields. The latter proved to be highly useful in understanding how many of the methods informed

Denotes content that is immediately available upon publication as open access.

Corresponding author: Eric Gilleland, ericg@ucar.edu

DOI: 10.1175/MWR-D-19-0256.1

about spatial location errors. The main advantage of using idealized test fields is that it is often much clearer what the correct verification result should be in subjective terms, which is generally not the case for real fields that might include a large number of events with different sizes/shapes/positions inside a single field.

The analysis of spatial methods continues in the second phase of the ICP, called the Mesoscale Verification Intercomparison over Complex Terrain (MesoVICT, https://ral.ucar.edu/projects/icp/; Dorninger et al. 2018) project. The goals of this second phase are to investigate how the various spatial verification methods behave when faced with a richer, more realistic, set of test cases and particularly over mountainous terrain; whereas the ICP focused on the central plains of the United States, which is relatively flat. The project includes test cases representing meteorological events that span not only over space but over time, and include multiple variables (e.g., precipitation, wind), as well as ensembles of forecasts and even ensembles of observations.

Here, a new set of idealized binary fields to be included within MesoVICT is presented. A one-valued grid point represents an event. An event might represent an underlying field value that exceeds a threshold, such as cloud amount and sea ice concentration. Contiguous events, or clusters of discontiguous events, is termed an event area.

A total of 50 new individual fields with 55 proposed comparisons are included in the study, which can be grouped into a few broad types, with similar characteristics. Each group of fields is aimed at testing different aspects of the behavior of the verification measures. These categories include: (i) common but pathological situations (e.g., fields that are completely empty, nearly empty or those entirely covered by events), (ii) circular, (iii) elliptical event areas (aimed at analyzing the influence of displacement length, orientation, border effect, and frequency bias), and (iv) event areas aimed at analyzing the influence of noise, such as small-scale scattered rain areas or clutter in radar.

These new idealized test fields are mainly conceived for comparing the diagnostic ability of different distance measures, which is a subclass of spatial verification methods (Dorninger et al. 2018). Distance measures give the spatial distance between forecast and observed event areas, which is usually expressed as a number of grid points that can be converted to physical distance (e.g., kilometers). Some of these measures are designed to quantify the spatial distance that would reflect a subjective estimate of the spatial displacement of features in one field compared to those in the other field. An estimate of spatial displacement is very appealing for forecast interpretation because it is easy to understand and mimics how humans tend to judge fields by eye. At the same time, it is important to note that some measures are not designed to reflect the subjective analysis of displacement.

The use of distance measures as stand-alone methods for verification purposes is relatively new. Examples of such measures being used for precipitation verification are Gilleland (2011, 2017) and Skok and Roberts (2018). At the same time, their behavior has, so far, not been studied or intercompared in a comprehensive manner. Therefore, alongside the primary goal of introducing these new idealized test fields, an analysis of the behavior of some of these measures is also conducted. The fields are also analyzed using geometric indices that provide information about the dispersiveness of an individual field.

2. Description of tested verification methods

To demonstrate the path for other researchers to follow in utilizing these fields to test the behavior of new methods, a comparison of several methods is made here. The R software package SpatialVx is used to obtain all the results presented here (Gilleland 2018).

a. Geometric indices

The geometric indices are those proposed in AghaKouchak et al. (2011), and are aimed at describing the spatial pattern of individual binary fields, rather than a direct comparison between two fields. While they propose applying the indices to the binary field as a whole, they are also useful to apply to individual features; one of which, the area index, has been in use in the MODE method since its inception (Davis et al. 2006a,b, 2009). The three geometric indices are: area index (A_{index}), shape index (S_{index}), and connectivity index (C_{index}).

Table 1 displays the equations used to calculate these indices; note that each is a value between zero and one inclusive. The area index (A_{index}) measures how convex versus concave a field is. For example, a circle is convex where a C-shaped object is concave. When applied to an entire field, A_{index} describes how structured or dispersive the field is overall. Values of A_{index} that are closer to one indicate a more structured pattern, and those closer to zero indicate a more dispersive one.

The shape index, S_{index}, is very similar to the area index, and typically provides redundant information. Nevertheless, as some of the fields will illustrate, it does inform about different properties. Values close to zero indicate more dispersive fields while values closer to one

TABLE 1. Equations for the geometric indices. Each index falls between zero and one inclusive. N_P is the number of nonzero grid points in the event area and N_C is the number of isolated clusters of events (called connected components).

Area index	$A_{\text{index}} = \frac{N_P}{A_{\text{ch}}} = \frac{\text{Area of event space}}{\text{Area of convex hull around event space}}$
Shape index	$S_{\text{index}} = \frac{P_{\min}}{P} = \frac{\text{Minimum possible perimeter (circle)}}{\text{Perimeter of event area}}$
	$P_{\min} = \begin{cases} 4\sqrt{N_P} & \text{if } \sqrt{N_P} = \lfloor \sqrt{N_P} \rfloor \\ 2\lfloor \sqrt{2N_P + 1} \rfloor & \text{otherwise} \end{cases}$
Connectivity index	$C_{\text{index}} = 1 - \frac{N_C - 1}{\sqrt{N_P} + N_C}$

indicate a more compact area of nonzero grid points that is close to circular in shape.

The connectivity index, C_{index}, measures how connected a field is. Values near one indicate a highly connected field and those near zero indicate a less connected field (e.g., with numerous smaller features). C_{index} is identically one for a single feature, so it is perhaps most useful when applied to an entire field.

If no events occur in the field $A_{\text{convex hull}} = 0$, $P = 0$, $N_P = 0$, and $N_C = 0$ and thus the values of all three indices are not defined because of the divisions by zero. Finally, note that a single circular object should have $S_{\text{index}} = 1$, but small errors relating to the computations on the grids accumulate to where the estimated values will be less than unity.

b. *Distance measures*

Different distance measures might measure forecast quality in a variety of ways, but their common property is that they all provide results in terms of spatial distance. One interesting point to keep in mind is that while the test fields represent perfectly known situations, for example two identical circles where one is displaced, different methods give different information about the similarity between two fields. One distance measure, for example, might provide information about the average or total amount of transformation of one field to another where another might provide some kind of average summary of actual distances between events, which may differ from the precise transformation invoked. Moreover, some methods might be designed to provide a measure of spatial distance that would reflect a subjective estimate of spatial displacement of events in one field compared to events in the other field. It is hoped that these comparisons will help illuminate how each method informs about the similarity and dissimilarity between two fields for potential users.

The distance measures and geometric indices investigated in this paper are applied to the entire field as a single entity, rather than on individual features within a field (as is done, e.g., in MODE, Davis et al. 2006a,b, 2009). The measures considered in this work could, however, be applied to individual features within a field. In fact, some of the case comparisons proposed could prove challenging for a method, such as MODE, in terms of identifying merges and/or matches of features.

One attribute for a distance measure comparing a field with event sets A and B that can be important for some purposes is for it to be a true mathematical metric, $m(A, B) \geq 0$. The criteria requires that three properties be met: (i) identity $m(A, B) = 0$ if and only if $A = B$, (ii) symmetry $m(A, B) = m(B, A)$, and (iii) the triangle inequality $m(A, C) \leq m(A, B) + m(B, C)$. The first is an obviously good property to have. The second property, however, may or may not be ideal in a forecast verification setting (Gilleland 2017) as the asymmetry of a measure may inform about false alarms as opposed to misses. Finally, the third property is important because it establishes a coherence in the measure. Namely, it ensures that if forecast A is closer to observation C than B in terms of the metric, then $m(A, C) < m(B, C)$.

A general description of each measure is given below. For completeness, the equations defining each measure is provided in Table 2.

(i) Centroid distance: The centroid distance (CDST) is the distance between the centers of mass of two fields. It is often used in feature-based methods as a measure of spatial closeness between two features (cf. Davis et al. 2009). It is a quick and easy measure to calculate, and is readily interpretable. Some of the new fields presented here will also highlight some of its shortcomings. Recall that, here, the distance is calculated between the centroids from the two *entire* field domains rather than a single feature within the domain. Some of the shortcomings might be mitigated by considering only individual features.

(ii) Baddeley's Δ: Baddeley's Δ metric (Baddeley 1992) is easily described using Fig. 1, which shows two binary fields, A and B in the top row. The next row

TABLE 2. Equations for the distance measures. Here, $||\cdot||$ is any distance norm (e.g., great-circle or, here, Euclidean), $\mathscr{D}$ is the spatial domain with coordinates $\mathbf{s} = (x, y) \in \mathscr{D}$, and $A \subset \mathscr{D}$ and $B \subset \mathscr{D}$ are binary event sets for two different fields $Z(\mathbf{s})$. $I_A(\mathbf{s})$ and $I_B(\mathbf{s})$ represent the binary fields containing the sets A and B, respectively. $d(\mathbf{s}, A)$ is the distance map for $I_A(\mathbf{s})$, N is the size of the domain, and n_A is the number of points in the set A. $w(\cdot)$ is any concave function [i.e., $w(t + u) \le w(t) + w(u)$], and is usually taken to be $w(x) = \min\{x, c\}$ for some user-chosen constant c. See the text for details on p.

Centroid distance	$\mathrm{CDST}(A, B) = \|\mathbf{m}_A - \mathbf{m}_B\|$, where $\mathbf{m}_A = \frac{1}{n_A}\sum_{\mathbf{s}\in A} Z(\mathbf{s}) \times \mathbf{s}$
Baddeley's Δ	$\Delta^p_{w(\cdot)}(A, B) = \mathrm{BDEL}(A, B) = \left\{\frac{1}{N}\sum_{x\in\mathscr{D}} \lvert w[d(\mathbf{s}, A)] - w[d(\mathbf{s}, B)]\rvert^p\right\}^{1/p}$
Hausdorff distance	$\Delta^{p=\infty}_{w(x)=x}(A, B) = H(A, B) = \max\{\lvert d(\mathbf{s}, A) - d(\mathbf{s}, B)\rvert\}$
Mean-error distance	$\mathrm{MED}(A, B) = \frac{1}{n_B}\sum_{\mathbf{s}\in B} d(\mathbf{s}, A)$
Zhu's measure	$Z(A, B) = \lambda\sqrt{\frac{1}{N}\sum_{\mathbf{s}\in\mathscr{D}}[I_A(\mathbf{s}) - I_B(\mathbf{s})]^2} + (1 - \lambda)\mathrm{MED}(A, B)$

shows their distance maps, which give the shortest distance from every point in the domain to the nearest point in the event sets. Baddeley's Δ is a type of average of panel g) in the figure, which shows the absolute differences between the two distance maps for A and B. The result is a type of average distance between the two binary event sets. It is a modification of the well-known Hausdorff distance metric. Both are sensitive to the size and shape of the domain of the verification set.

Denote the nonzero grid points in one field by A and those in the other by B. The shortest distance from every point $\mathbf{s} \in \mathscr{D}$ to the nearest nonzero grid point in A (and similarly for B) are calculated and denoted by $d(\mathbf{s}, A)$ and $d(\mathbf{s}, B)$, respectively. These resulting fields defined by $d(\mathbf{s}, A)$ and $d(\mathbf{s}, B)$ are called distance maps, and can be calculated in a computationally efficient manner (e.g., Borgefors 1986; Meijster et al. 2000; Fabbri et al. 2008).

The metric is an L_p norm of the magnitude difference between the distance maps for A and B where p is a user-selected parameter where $p = \infty$ recovers the Hausdorff distance. The cutoff transformation [i.e., $w(x) = \min\{x, c\}$ for a user-chosen constant c] has been found to remove some of the sensitivity to outliers and domain effects (e.g., Schwedler and Baldwin 2011). However, Gilleland (2011) found that results can be sensitive to the choice of constant c. Results reported here are without any transformation because it was found that applying the cutoff transformation did not change any of the pertinent findings related to these geometric comparisons.

For the pathological case where $Z(\mathbf{s}) = 0$ everywhere, a special rule is invoked with the distance map defined to be ∞ everywhere, and in practical applications, this value is reduced to be N; the size of the domain. Different definitions for N are possible. If the grid is $n \times m$, then the size might be taken to be $n + m$, $n \times m$ (used here), $\sqrt{n^2 + m^2}$, etc.

(iii) Squared and centered Baddeley Δ: One way to alleviate the issue of domain placement is to reposition the sets so that they are centered on a new square domain; but where they remain in their same relative positions to each other. Call the resulting metric the squared and centered Baddeley Δ, and denote it by $\eth$ (eth). If this approach is performed for all subsequent comparisons of different sets A and B, then $\eth$ will be consistent across the comparisons where Δ would not. In practice the new domain should be at least as large as the largest verification set to be compared in order to contain both event sets, A and B, so in general the new domain will be larger than the original $\mathscr{D}$. $\eth$ has the same definition as Δ when $Z(\mathbf{s}) = 0$ everywhere, except that $|N|$ will usually be larger. Of course, $\eth$ is still sensitive to $|N|$.

(iv) Mean-error distance: Again using Fig. 1 to illustrate, the mean-error distance (MED; Baddeley 1992; Gilleland 2017) measures the average distance of one event set to the nearest points in another event set. MED (B, A) is the average value of the nonzero values shown in Fig. 1e of the figure, and MED (A, B) is the average of the nonzero values in Fig. 1f. Note that the domain is not shown in these panels to emphasize that the distances are only averaged over the events defined by the respective binary sets.

The MED measures the average shortest distance of nonzero grid points in one field to the nearest nonzero grid points in another. It is an asymmetric measure because $\mathrm{MED}(A, B) \neq \mathrm{MED}(B, A)$. This asymmetry has a useful interpretation for forecast verification because MED(forecast, observation) gives a spatial distance measure of *misses* and MED(observation, forecast) gives a spatial distance measure of *false alarms*. While any measure for $d(\mathbf{s}, \cdot)$ can be used, the computationally

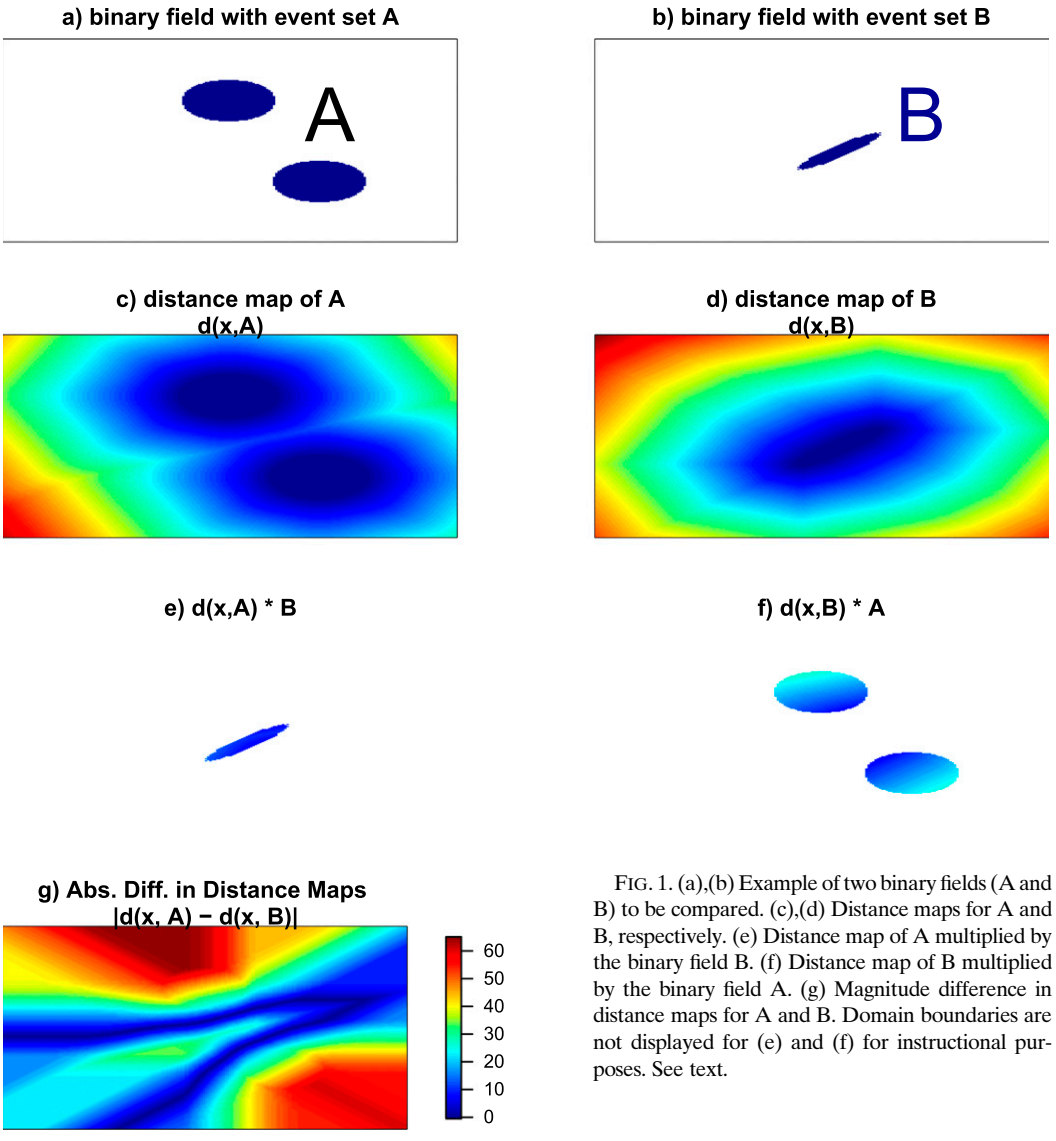

FIG. 1. (a),(b) Example of two binary fields (A and B) to be compared. (c),(d) Distance maps for A and B, respectively. (e) Distance map of A multiplied by the binary field B. (f) Distance map of B multiplied by the binary field A. (g) Magnitude difference in distance maps for A and B. Domain boundaries are not displayed for (e) and (f) for instructional purposes. See text.

fast distance transform used to calculate Baddeley's Δ can be exploited here as well so that the computations can be made in an operational setting. Unlike Δ, however, the MED is insensitive to the specific domain's size or shape; meaning that its value doesn't change if, for example, the domain size were increased. A few modifications of the MED, that are symmetric, and therefore true mathematical metrics include:

$$\begin{aligned}\text{avg MED}(A,B) &= \frac{1}{2}[\text{MED}(A,B) + \text{MED}(B,A)],\\ \text{minMED}(A,B) &= \min\{\text{MED}(A,B), \text{MED}(B,A)\}\\ \text{maxMED}(A,B) &= \max\{\text{MED}(A,B), \text{MED}(B,A)\}.\end{aligned} \tag{1}$$

Again, because of the rule that the distance map is ∞, taken to be $|N|$, at every grid point when $Z(\mathbf{s}) = 0$ everywhere, MED(A, B) is still defined when A is empty (zero everywhere), if $B > 0$ somewhere. However, MED(B, A) would be undefined.

(v) Zhu's measure: Zhu's measure (Zhu et al. 2011) is designed to be a true mathematical metric in the case where two competing forecast fields are to be compared against the same observation field. However, in the simplified version that compares a single forecast field to an observed field, it is not a true metric, and is therefore referred to as Zhu's measure here. It is defined by a weighted average between the root-mean square deviation between the two binary fields and MED (A, B). The reliance on MED in the second term demonstrates that $Z(A, B) \neq Z(B, A)$

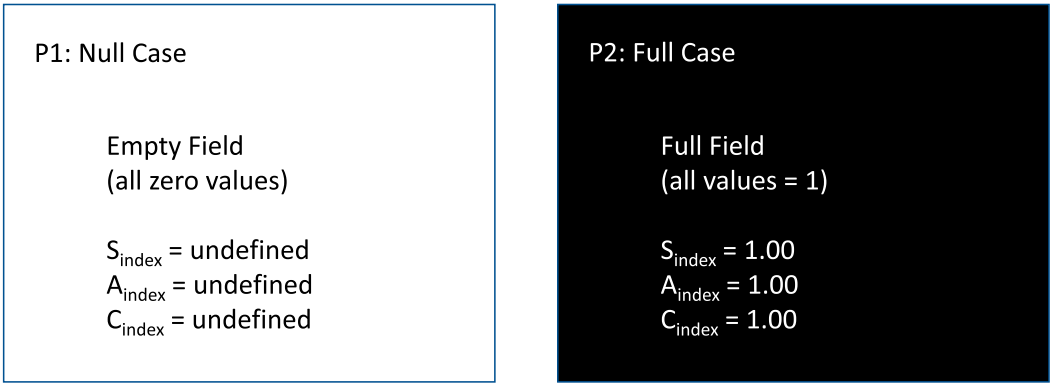

FIG. 2. Diagrams of pathological test fields P1 and P2 with values of the results of the geometrical indices. P1 has only zero values, P2 has value one at every grid cell. Black squares indicate one-valued grid cells and white space indicates zero-valued cells. Domain is 200×200. The numbers represent the values of geometrical indices (S_{index}, A_{index}, and C_{index}), which are described in section 4, and their values are rounded to two decimal places.

in general, and is subsequently not symmetric unless $\lambda_2 = 0$. Here, $\lambda_1 = \lambda_2 = 1/2$ so that each term is weighted equally.

(vi) Fractions skill score displacement: The fractions skill score displacement (dFSS; Skok and Roberts 2018) uses the fractions skill score (FSS; Roberts and Lean 2008; Roberts 2008) to provide information about the spatial displacement between events in one field compared to the events in another field. The FSS gives a measure of forecast skill at different spatial scales and the dFSS value is obtained via determining the scale at which FSS exceeds 1/2 while separately treating the regions with and without overlap.

The dFSS was designed with the aim of providing a measure of spatial distance that would reflect a subjective estimate of spatial displacement of events in the two fields. In Skok and Roberts (2018) the behavior of dFSS was analyzed for various idealized and real setups and the results showed that the dFSS can indeed be used to determine spatial displacement in a meaningful way. The results also showed that the larger events have a larger influence on the resulting score value. Further, dFSS exhibits some dependence of the resulting value on the orientation of the displacement and the presence of the nearby domain border; however, the influence of these two factors on the score value tends to be small.

The dFSS requires that the frequency bias be small (the number of nonzero points in the two fields should be similar) and Skok and Roberts (2018) recommend that it should not be used in cases when the frequency bias is outside of the interval between 1/2 and 2; an appropriate value of bias can usually be achieved by using a percentile threshold.

3. Description of the test fields and case comparisons

As mentioned, the new fields and the associated comparisons are designed to test specific scenarios that the test fields of the original ICP do not address, and includes four broad types: 1) pathological, 2) circular, 3) elliptical, and 4) scattered, holes and noisy. In total, 50 individual fields are examined. The fields are freely available from the MesoVICT web page (http://www.ral.ucar.edu/projects/icp/). The number of cases is large and the number of potential comparisons much larger still. Nevertheless, in creating the various cases, it became clear that the distance measures yield similar results as one another for most of them. However, each has a case comparison where its value diverges wildly from the others, or is undefined. Therefore, it is hoped that these cases will serve as a means for researchers and practitioners to test other, possibly new, methods. Deviations in expectations about their behavior can then be discerned. They have also proven useful in testing and validating software that computes verification summaries.

All fields and case comparisons are shown in Figs. 2–10 and Tables 3–5. These figures also show the results from applying geometric indices and distance measures, which will be discussed in section 4. The fields are binary and constructed on a 200×200 regular grid. Many methods, especially those that primarily analyze location errors, require a binary field. In verification practices, such fields are often obtained through a thresholding

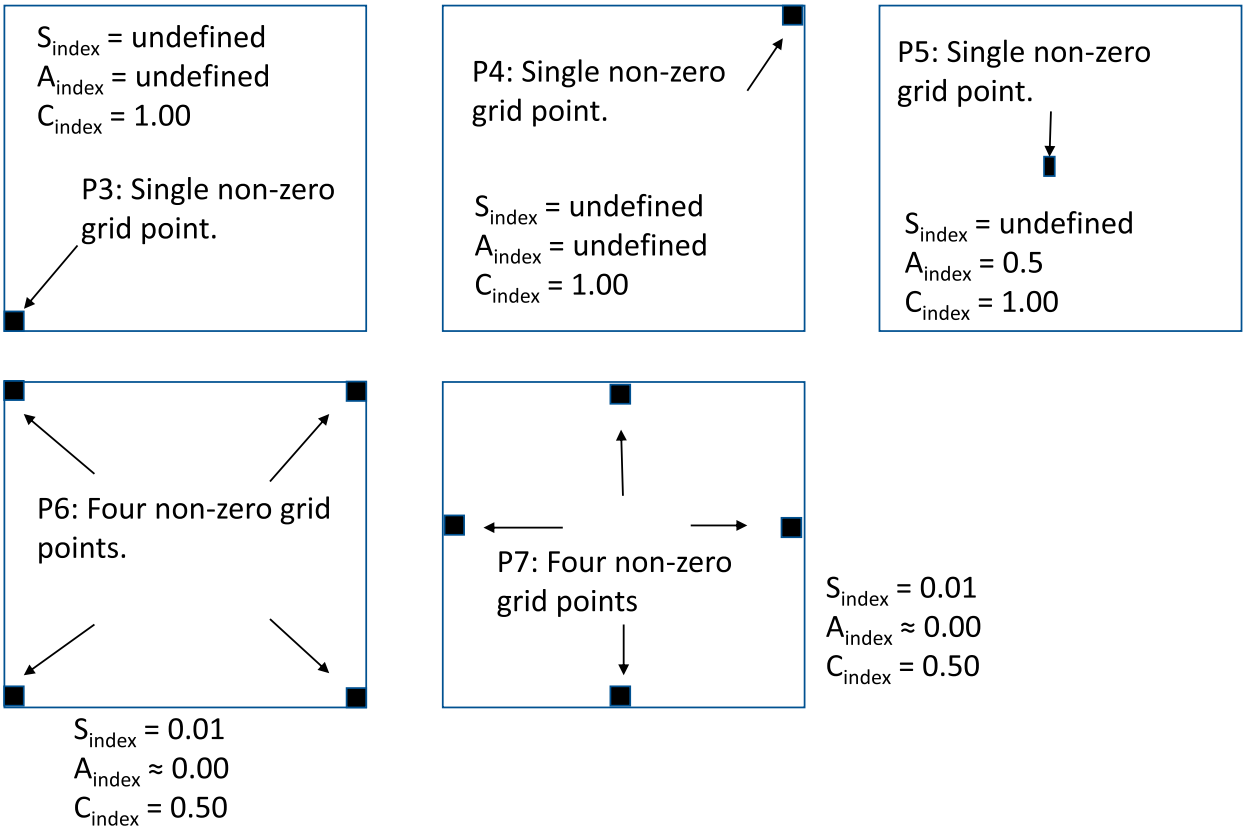

FIG. 3. As in Fig. 2, but for fields P3–P7. Please note that in this figure the size of isolated nonzero grid points is intentionally exaggerated to make them more visible.

process of the original field (e.g., by setting all values below a threshold to zero, and all values above the threshold to one). For some variables, the threshold choice defines the type of event analyzed. For example, for precipitation, a low threshold of 0.2 mm is usually used to identify large-scale pfeatures, whereas a higher threshold, such as 10 mm, might be used to identify intense convective cells. On the other hand, some meteorological variables are quasi-binary by nature. Cloud amount or sea ice concentration, for example, can assume continuous values between zero and one, but most values cluster either near zero (indicating clear-sky or open water) or one (overcast or sea ice).

In addition to the new test fields, a list of proposed comparisons is given in Table 6. The list does not include all the possible comparisons between all the test fields (i.e., every test field compared with all the other fields). However, additional comparisons could be made to evaluate a method. Each group of cases is described next.

(i) Pathological: The pathological fields are described in Table 3 with fields and comparisons shown in Figs. 2–4. Figures 2 and 3 depict the individual pathological fields in cartoon format where the size of isolated nonzero points is greatly exaggerated because a true depiction of the fields makes the isolated points difficult to see. Figure 4 shows the proposed comparisons for these fields using the actual fields without the exaggeration of isolated points.
A simple goal is to determine whether or not each method can handle the situation with no events in one or both fields. It is also desired to test the methods for their behavior when the spatial area of events in one or both of the forecast and observation goes to the limit of 0 and, the opposite, a complete coverage of the domain. That is, if both fields are empty of any event, one would hope that the measure would yield a perfect score. Most of the methods are not defined in this case, but they might have a special rule for the situation. However, in this case, suppose that each field has a single grid point with an event, but in a very different part of the domain. It would be hoped that the summary measure would still give a nearly perfect score in many situations where the additional point does not, for example, represent an extreme event.
(ii) Circular: The circular fields are similar to the test fields of the original ICP but heir number and

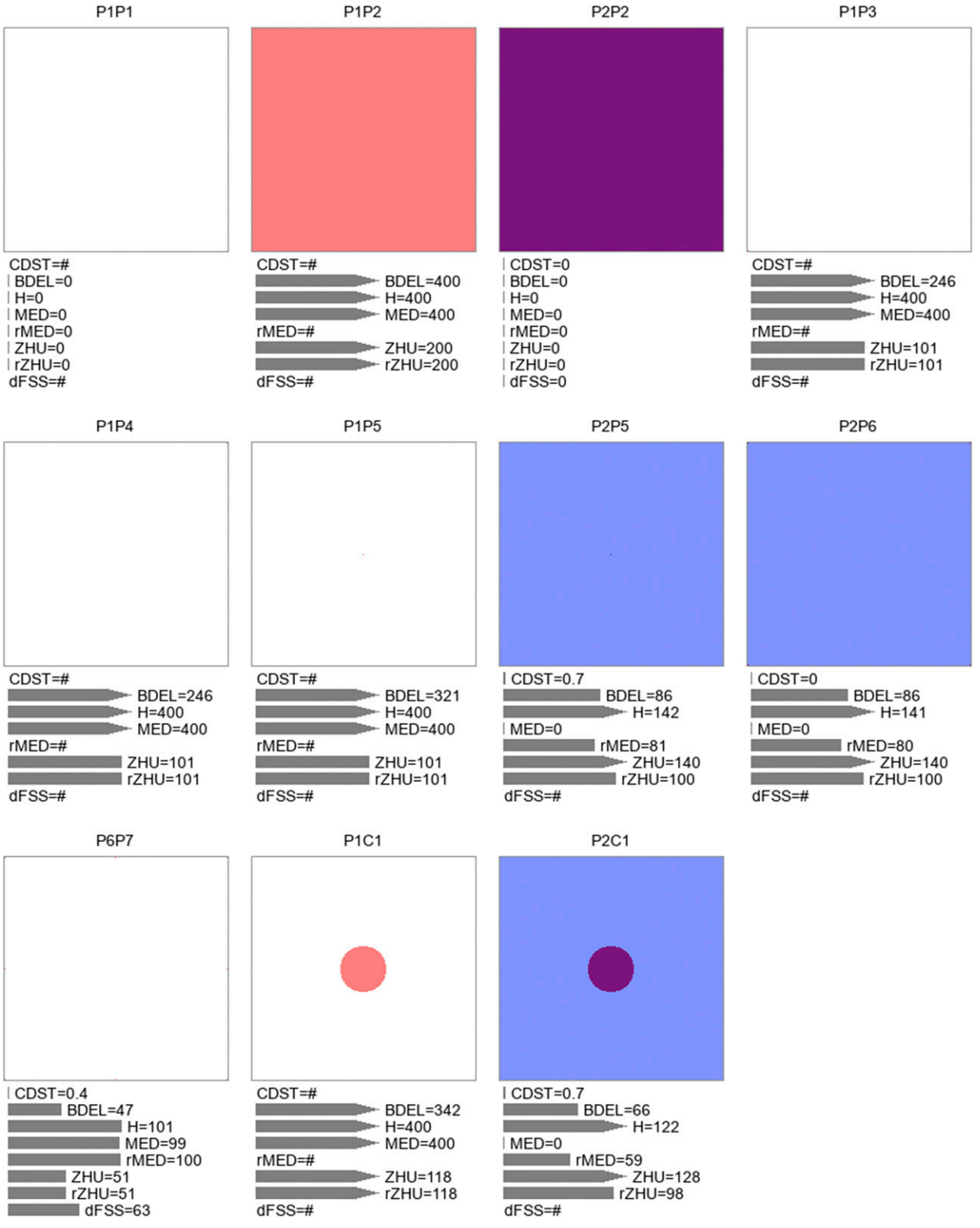

FIG. 4. Pathological comparisons with results of the distance measures. Colors: light blue (only the first field is nonzero), light red (only the second field is nonzero), violet (both fields are nonzero), and white (both fields are zero). The distance provided by the distance measures is shown with the length of gray bars beneath each comparison along with a numerical value. A pointed bar indicated a distance larger than 110 points. A value of ''#'' represents situations where the measure fails to produce a result. The nonsymmetrical metrics have the prefix ''r'' for the reversed comparison to differentiate between the first fields being compared to the second field and vice versa. Acronym BDEL represents Baddeley's Δ, ZHU the Zhu's measure, and CDST the CDST; dFSS stands for fractions skill score displacement, MED is mean-error distance. These measures are discussed in section.

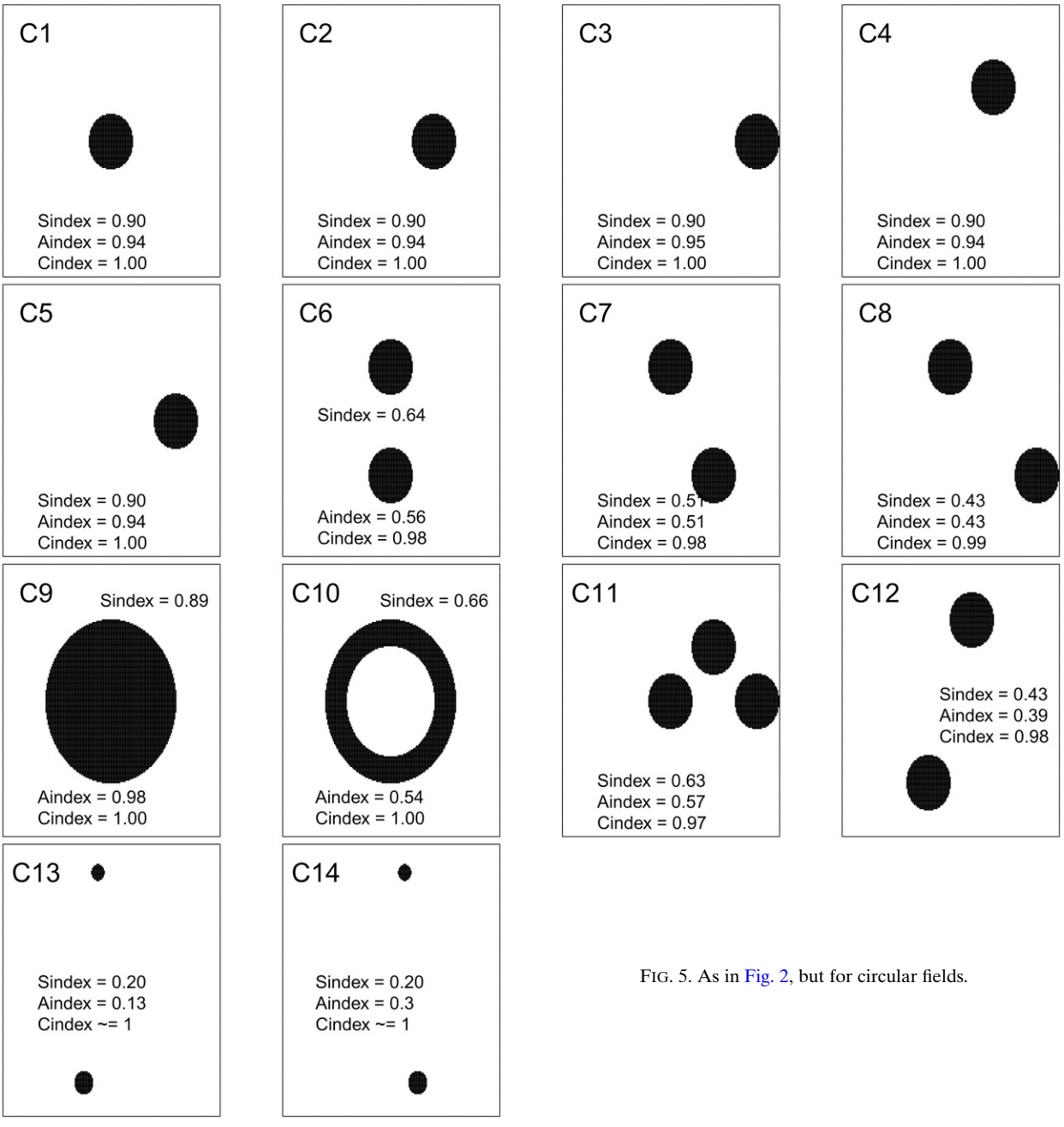

FIG. 5. As in Fig. 2, but for circular fields.

variability is greatly increased to test situations that were not considered before. The circular fields are described in Table 3 with fields and comparisons shown in Figs. 5 and 6. Circular shapes are generally representative of convective systems, and although they are similar to the original ICP shapes, they represent an expansion of realistic scenarios that pose difficult questions in terms of verifying them. In terms of ranking comparisons, for example, a user-specific ranking will depend on the user so that one method might be preferred by one user and not another. The situations represent challenges for verification methods such as, positional and boundary effects (cf. C1 and C2 vs C2 and C3, where C3 touches the edge of the domain; see Fig. 6), directional effects (cf. C2 and C4 vs C1 and C2), issues when one field has more event areas than the other, especially if they are positioned so that it is ambiguous as to whether or not they are misses/false alarms or location errors (cf. C1 and C6).

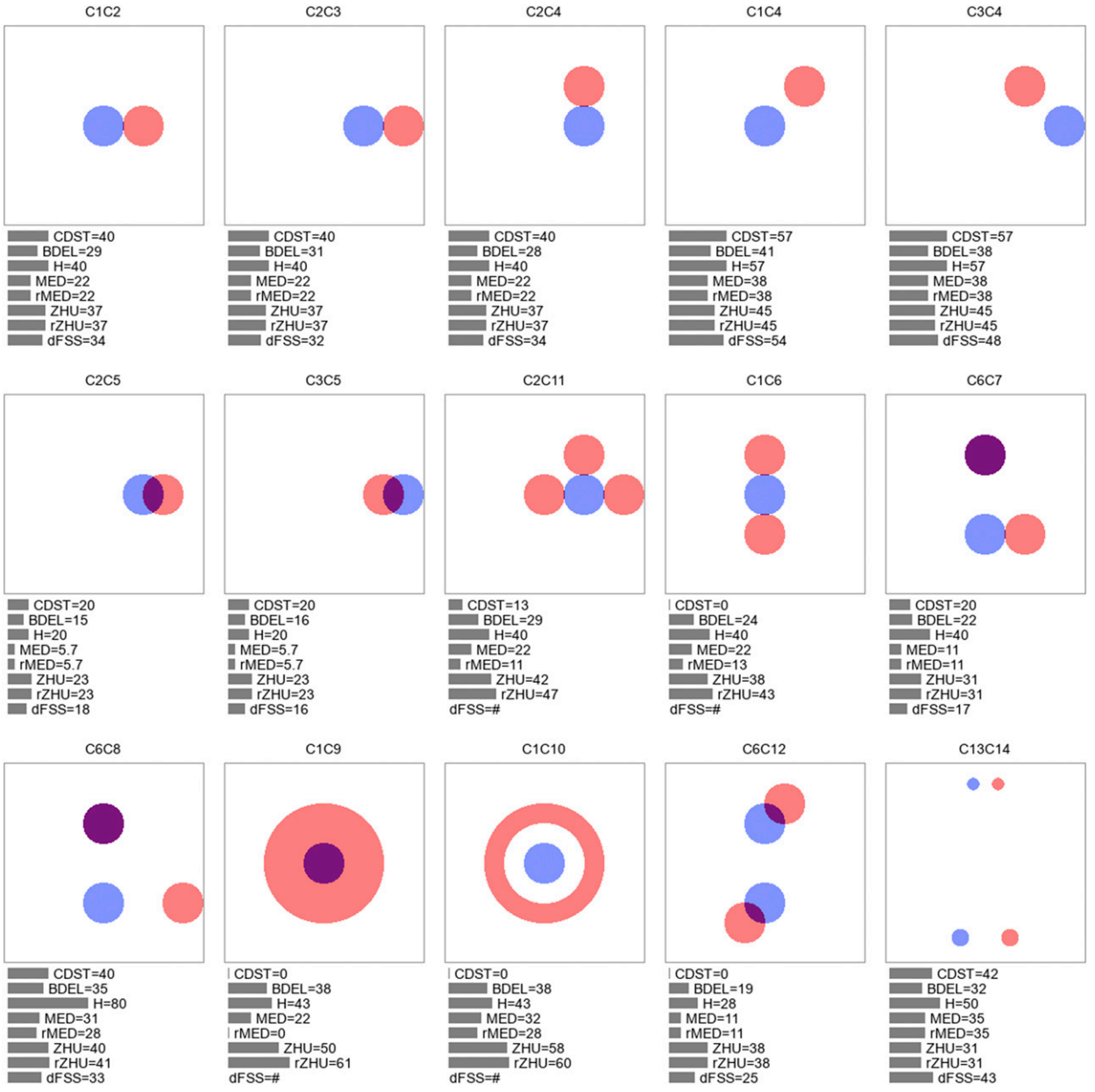

FIG. 6. As in Fig. 4, but for circular comparisons.

(iii) Elliptical: The elliptical fields represent features that are more realistic and typical of frontal events or complex terrain environments. The fields are described in Table 4 with fields and comparisons shown in Figs. 7 and 8. The configurations are designed to allow for a systematic analysis of how the methods deal with three types of error: (i) translation, (ii) rotation, and (iii) size of event areas. The test proceeds with each type of error isolated (e.g., only translation error), two types of errors and all three. For most purposes, it is not necessary to test every pair of these fields E1 to E16 (120 possible pairs), so a subset is suggested here (cf. Table 6). Fields E17 to E20 are simplified, realistic scenarios that may present challenges for certain methods. They mimic phenomena such as clouds or precipitation that appear fragmented in frontal situations or because of the orography.

(iv) Scattered, holes, and noisy: The last group of fields consists of three subgroups that are designed to test aspects of the methods not yet examined. The fields are described in Table 5 with fields and comparisons shown in Figs. 9 and 10.

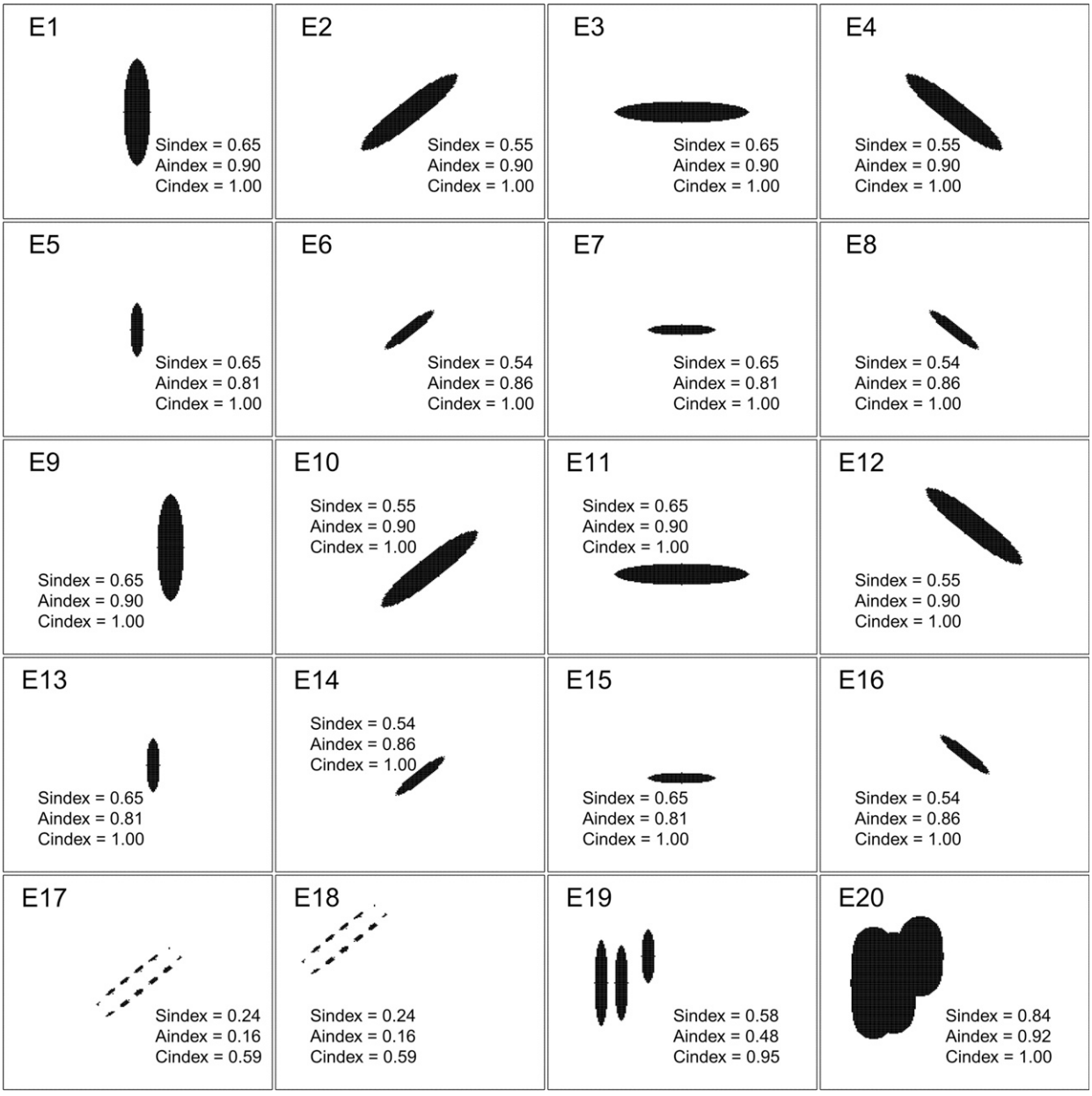

FIG. 7. As in Fig. 2, but for elliptical fields.

The fields S1 to S3 are designed to test how the methods handle small-scale events scattered inside an envelope (these cases are similar to the rectangular envelopes in Skok and Roberts 2016). The envelopes in S1 and S2 are identical and events within the two envelopes can overlap so that their comparison relates to small-scale convective systems where the forecast correctly identifies the general region of convection but the placement of individual convective cells may be less accurate. The S3 envelope is displaced by 100 grid points relating to the situation where the general area of convection is misplaced.

Fields H1 and H2 are the same as C1 and C2 but with inverted values (zeros are replaced with ones and vice versa). The root-mean-square error (RMSE) would yield identical results for the comparisons H1H2 and C1C2 because it treats the zero and non-zero areas in an equal manner. Other methods might score them differently, however. For example, it is feasible that one method might ignore the empty areas and focus solely on the event areas, which could lead to a much better score value for the H1H2 comparison than for the C1C2 comparison. The point

FIG. 8. As in Fig. 4, but for elliptical comparisons.

is to test a method's sensitivity to the spatial size of event areas.

Because different methods operate on different principles, it is reasonable to expect that their sensitivity to noise will differ. Fields N1 to N4 demonstrate a situation in which noise is present in the fields. The presence of noise in measurements can be caused, for example, by uncertainties in remote sensing techniques. Fields N1 and N2 are copies of C1 and C4 but also contain a small amount of noise. The noise in N1 and N2 is simulated by randomly generating nonzero grid points with a frequency 0.1%. The total number of randomly generated points is considerably smaller than the number of nonzero grid points inside the dominant circular

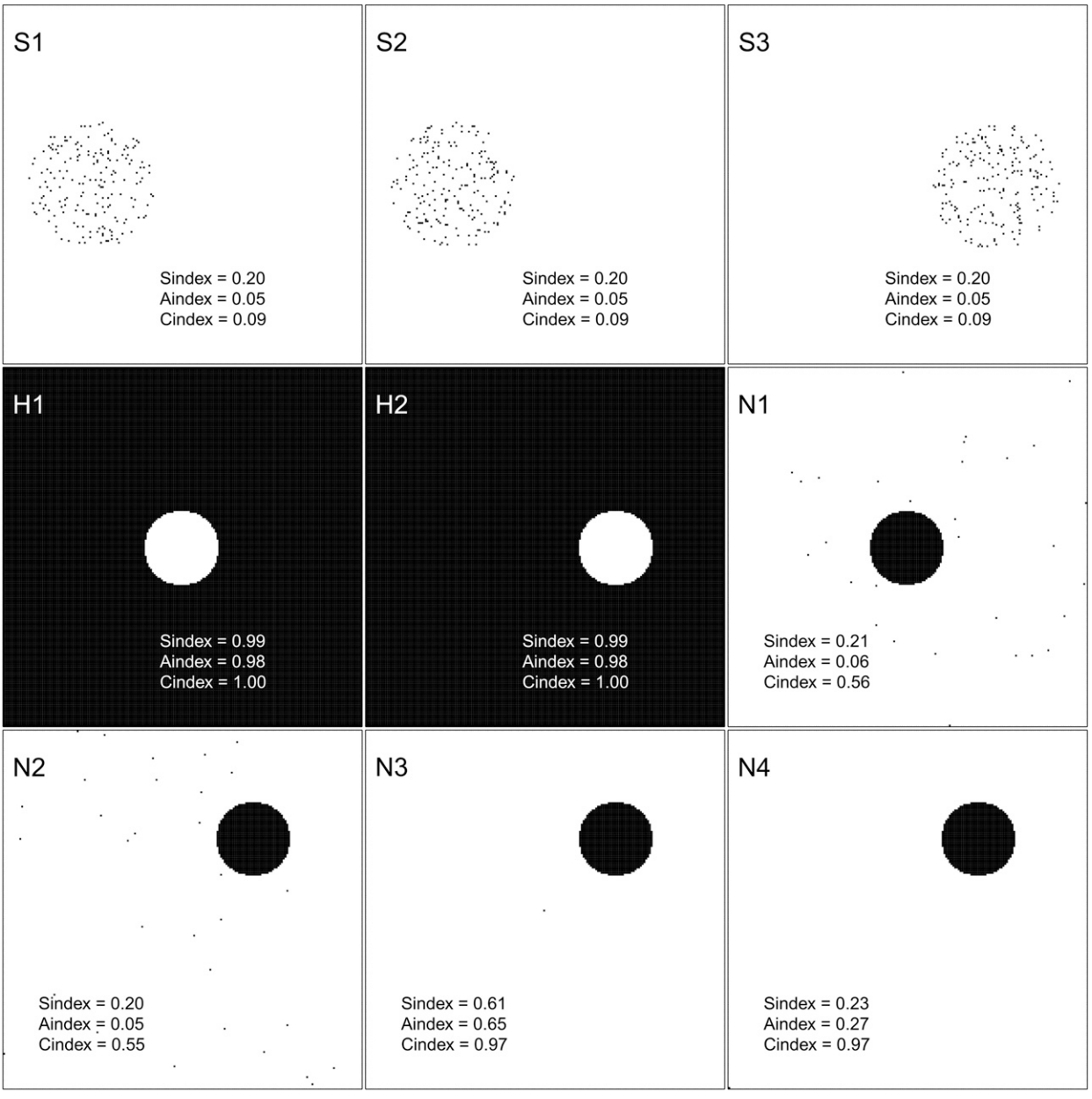

FIG. 9. As in Fig. 2, but for fields wind small-scale scattered events, holes and noise. N3 has a single point near the center of the field, and N4 has one in the lower-left corner, but both are difficult for a human observer to see (cf. Fig. 3 where N3 is a union of P5 with C4 and N4 a union of P3 with C4).

area (approximately 1300 versus 40 grid points). To determine the magnitude of sensitivity to noise the results of C1C4 and N1N2 can be compared. Fields N3 and N4 can be used to test the sensitivity of a method when a single isolated nonzero point is added to the field. While the above S1 to S3 cases also involve randomly scattered small-scale events, the metaverification goal for these fields differs. For N1 to N4, the assessment is on how noise might affect results for cases where obvious large-scale phenomena dominates, and S1 to S3 are designed to assess how a more homogenous isolation of scattered event areas are analyzed.

4. Results

a. Geometric indices

Figures 2 and 3 show the geometric indices introduced by AghaKouchak et al. (2011) for the pathological

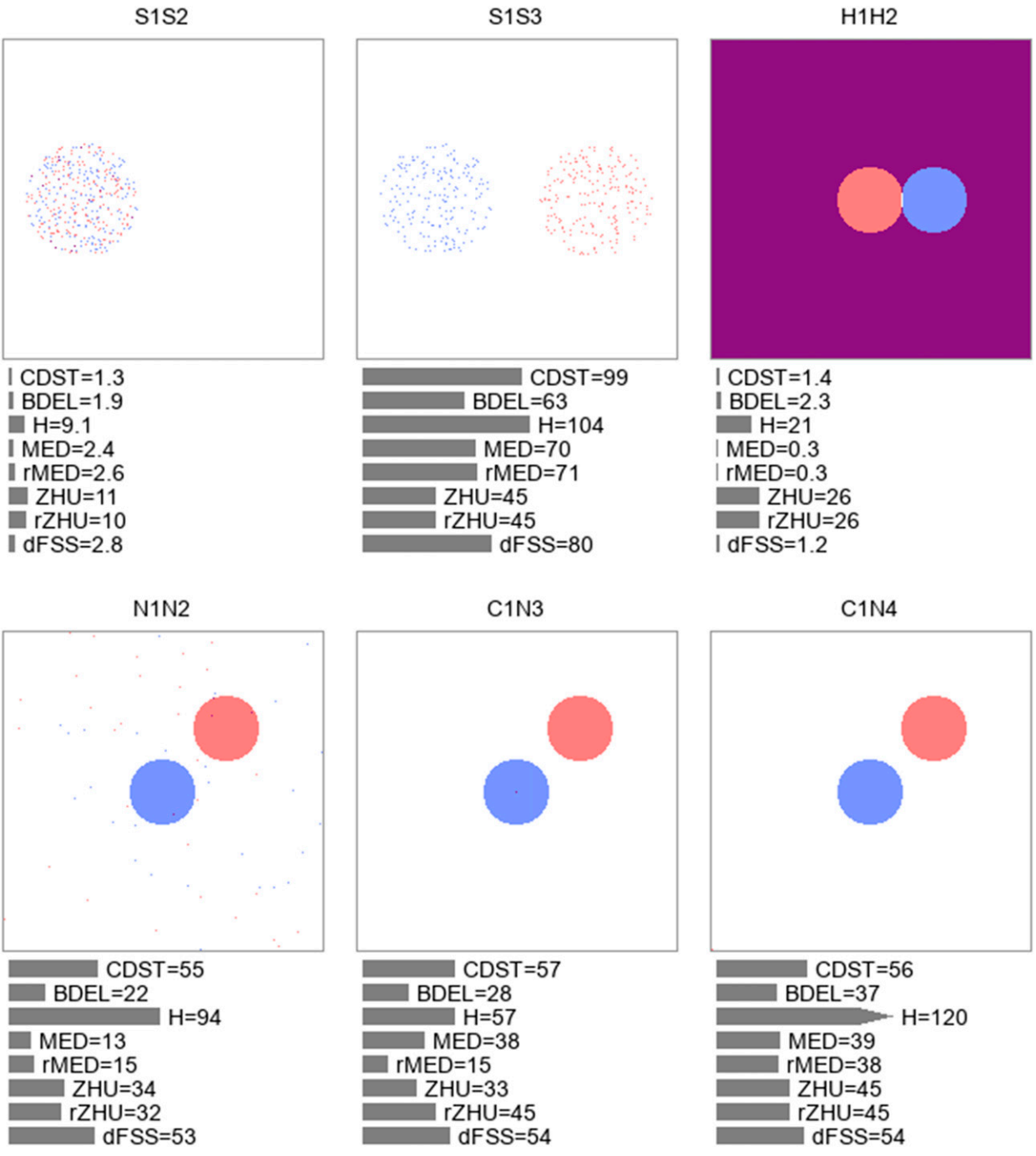

FIG. 10. As in Fig. 4, but for comparisons using small-scale scattered events, holes, and noise.

fields. These indices are not comparative by themselves as they are only applied to individual fields from the verification set, rather than evaluating how closely one field matches the other. Comparing their values across fields, however, tells about the similarity/difference in terms of the overall spatial pattern, or texture of the fields. None of the indices is defined when no events occur in the field because of division by zero.

By its definition, A_{index} should be 1.00 for any single point in $\mathscr{D}$. However, the software used to calculate the value for this index (R package SpatialVx; Gilleland 2018) returns a missing value when there is only a single nonzero value placed in either corner. Further, the

TABLE 3. General description of pathological and circular fields. The domain size is 200 × 200 grid points. The coordinate of the point in the lower-left corner is (1, 1). The coordinates are given as (x, y), where x is the horizontal and y the vertical axis.

	Pathological fields
P1	Null field (all values are zero)
P2	Full field (all values are one)
P3	One nonzero point in lower-left corner
P4	One nonzero point in upper-right corner
P5	One nonzero point near the center of the field at (100, 100)
P6	Four nonzero points in the four corners
P6	Four nonzero points near the middle of each boundary at (1, 100), (100, 1), (200, 100) and (100, 200)
	Circular fields
C1	Circle with radius 20 centered at (100, 100)
C2	Circle with radius 20 centered at (140, 100)
C3	Circle with radius 20 centered at (180, 100) (touches edge of domain)
C4	Circle with radius 20 centered at (140, 140)
C5	Same as C1 to C4, but centered at (160, 100) in order to overlap some with C2 and C3
C6	Two circles centered at (100, 140) and (100, 60) with radii 20
C7	Same as C6, but lower circle translated 40 points to the right
C8	Same as C6, but lower circle translated 80 points to the right
C9	Large circle (radius = 60) centered at (100, 100)
C10	Ring centered at (100, 100) with inner radius 40 and outer radius 60 (i.e., width = 20)
C11	Union of C1, C3 and C4
C12	Two circles with radius 20 centered at (120, 160) and (80, 40)
C13	Two circles: one with radius 8 centered at (75, 25) and one with radius $8/\sqrt{2}$ centered at (88, 180)
C14	Two circles: one with radius 8 centered at (125, 25) and one with radius $8/\sqrt{2}$ centered at (113, 180)

TABLE 4. General description of elliptical fields. The ratio of the major vs minor axis of all ellipses is 5:1 with dimensions of the ellipses either 100 by 20 (large ellipse) or 25 by 5 (small ellipse). All ellipses are centered at (100, 100) unless translated.

	Elliptical fields
E1	Vertical large ellipse
E2	Large ellipse at 45° angle
E3	Large horizontal ellipse
E4	Large ellipse at 135° angle
E5	Small vertical ellipse
E6	Small ellipse at 45° angle
E7	Small horizontal ellipse
E8	Smallellipse at 135° angle
E9	E1 translated 25 grid points east
E10	E2 translated 15 grid points east and 20 south
E11	E3 tranlated 25 grid points south
E12	E4 translated 15° east and 20° north
E13	E5 translated 12 grid points east
E14	E6 translated 8 grid points east and 10 south
E15	E7 translated 12 grid points south
E16	E8 translated 8 grid points east and 10 north
E17	Several slanty to the right ovals within the oval in 2
E18	Ovals from E17 shifted to the upper-left corner
E19	Three vertical ovals centered at (100, 40), (100, 55), and (125, 75), respectively, scaled by (40, 5), (35, 5), and (25, 5), respectively
E20	E19 smoothed using a disk kernel with radius 12

computation returns 0.50 for the single event in the middle of the field because the function returns the value 2 for the area of the convex hull. In theory, however, for each single-point case, the area should be one. Results for S_{index} for these fields are very similar, but the software returns NA for single event points. C_{index} is not defined for P1, and is not very interesting for the rest of these fields as they are mostly single contiguous features so that C_{index} is identically one for all fields. For the four-point fields, its value is 0.50.

Figure 5 shows the geometric indices for each of the circular fields. The A_{index} is found to have a slight edge effect, as is evidenced by the small difference in its value for C3 compared to those of C1 and C2. This effect may simply be the result of computational idiosyncracies of the software used to calculate them; the difference is only on the order of one hundredth. The similar S_{index} does not suffer from the edge effect problem.

To illustrate how the area and shape indices inform about dispersiveness, compare C6 with C7. The two identical circles are slightly farther apart for C7 than for C6, yielding slightly lower values of these indices, and even lower for C8 where they are separated even more. The connectivity index, on the other hand, is the same for all three fields. The C_{index} is again less interesting for these fields as they are all composed of one or a couple continuous event areas. They do demonstrate how having only two sets of isolated clusters only lowers the value by a negligible amount.

The elliptical fields (Fig. 7) demonstrate the primary difference between A_{index} and S_{index}. The former is nearly one as the ellipse is fully concave, but it is far from a perfect circle so that S_{index} is much lower ranging from about 0.54 to about 0.65. E17 and E18 have a low A_{index} of about 0.16 despite having a tight concave area of activity, but nevertheless consist of small "storm" cells. The S_{index} is slightly higher but still very low at about 0.24, which is not surprising as the nonzero grid cells form an elliptical region. The reason why S_{index} is not identical for fields E1 to E4 is that E1 and E3 both have 1553 events where E2 and E4 have slightly more at 1573, which is a result of the way in which the fields are created.

For the small-scale scattered events (S1 to S3 in Fig. 9), the area index simply reflects the frequency of

TABLE 5. General description of scattered, hole, and noisy fields.

	Small-scale scattered events inside an envelope
S1	Small-scale scattered events (with frequency 5%) inside a circular envelope with radius 35 centered at (50, 100)
S2	As in S1, but a different realization of the scattered events
S3	As in S1 and S2, but with the envelope translated 100 grid points and a different realization of the scattered events
	Holes
H1	Inverted C1 (i.e., 1 if C1 = 0, and 0 if C1 = 1)
H2	Inverted C2
	Noisy
N1	As in C1, but with a small amount of randomly generated noise with frequency 0.1%
N2	As in C4, but with a small amount of randomly generated noise with frequency 0.1%
N3	As in C4, but with a single nonzero point added near the center of domain at (100, 100); union of C4 and P5
N4	As in C4, but with a single nonzero point added in the lower-left corner; union of C4 and P3

occurrence of the rain areas (5%) because it is essentially the area of the circular envelope divided by the total area of rain for each of the fields. The shape index is similarly defined to the area index, but is related to the square root of frequency because perimeters instead of areas are used to calculate the index value. The square root of 0.05 is approximately 0.22 which is similar to 0.2 given by the index. The small difference occurs because of the way $P_{\min}$ is estimated. One might consider a more precise calculation for the perimeter, but it is unclear what the added value of the S_{index} is over the A_{index} in this setting; if considering a single object within a field, then the S_{index} can provide useful additional information. The C_{index} for these fields describes the texture of the field as having numerous discontiguous rain areas with very low values; all of which are rounded to about 0.09.

The fields with holes (H1 and H2 in Fig. 9) are not very dispersive, which is captured by these indices. Because all one-valued grid cells are connected, C_{index} is one for both fields as all one-valued grid points are connected.

For the noisy fields (N1 to N4 in Fig. 9), the shape and area indices are heavily influenced by the noise; giving very low values suggestive of high dispersiveness. All the values for C4 alone without P5 added to the field give shape, area and connectivity indices at or near 1, indicating undispersive fields. After adding P5 to C4 (i.e., N3) or P3 (i.e., N4), both shape and area indices drop drastically to around 60%; showing a high sensitiviy of these indices to noise because a single added event can cause a substantial difference in the index value. The connectivity index perhaps gives more reasonable information under this situation because its relatively low values of 0.56 and 0.55 for the cases with multiple scattered events is suggestive of low, but not very low connectivity. On the other hand, the addition of the single point does not reduce its value much; going from 1.00 to 0.97 with the additional event. It should be noted that AghaKouchak et al. (2011) were aware of the high sensitivity of these indices to isolated outliers and recommended removing noise before applying the measure.

b. Distance measures

With so many cases, it is difficult to navigate the numerous possible results. In what follows, a summary of each measure's propoerties for each set of comparison cases is given with an aim of only highlighting interesting properties. An overall summary is given at the end.

(i) Pathological comparisons: Figure 4 shows the pathological comparisons with results for the distance measures. The centroid for the null fields is undefined. A single event (one grid point), however, is all that is necessary for it to be defined; and a single event is its own centroid. For example, the centroid for P3 is (1, 1), while for P4 it is (200, 200) and P5 it is (100, 100). Therefore, the CDST for P1P1, P1P3, P1P4 and P1P5 is undefined. As expected all the measures give a perfect value (zero distance) for the comparison P2P2, which consist of two identical full fields.

For the P2P5 comparison the point is displaced by $1/\sqrt{2} = 0.71$ from the center of the domain, because the domain does not have an exact center, so is also the result given by the CDST. The P6P7 comparison gives a CDST of about 0.35. The P2P6 CDST is calculated to be identically zero because P6 has a single point in each of the four corners so that the centroids of both P2 and P6 are the same. Finally, the CDST for (r)P1C1 is the same as for (r)P1P5 at about 141.42. The CDST for (r)P2C1 is about 0.71.

As previously described, the distance map used to calculate Δ is defined to be c everywhere when no one-valued grid points are in the domain, and for practical purposes, when this value is ∞, it is reduced to the size of the domain; in this case, it is reduced to 400. When the entire domain is one-valued, it is simply zero. Therefore, the P1P1 (and P2P2) comparison is identically zero for Δ, a perfect score. (r)P1P2 yields the max value of c in each case, which for δ is just slightly larger because of the new domain on which it is constructed.

It is interesting that Δ gives the same answer, as would be desired, for the comparisons (r)P1P3 and P1P4, which both compare the null field against the

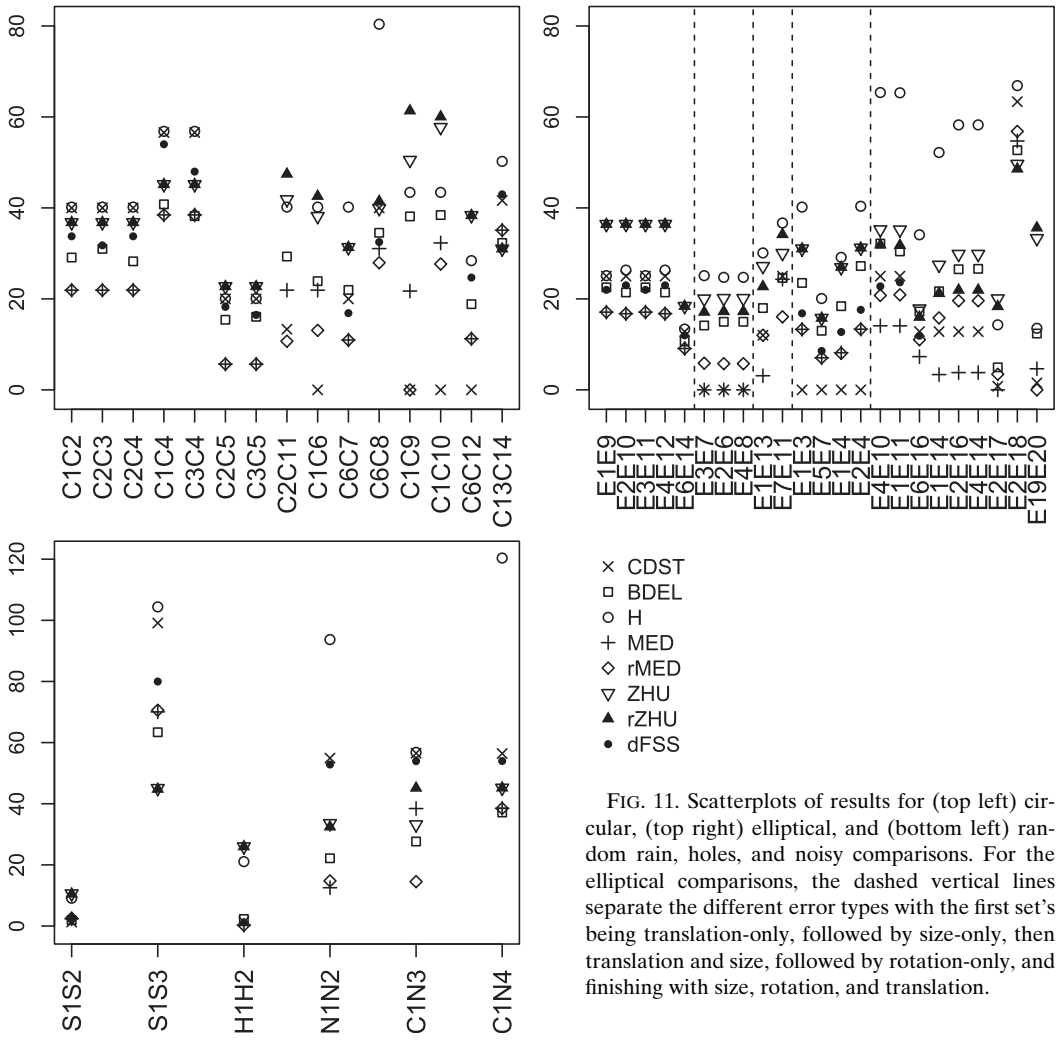

FIG. 11. Scatterplots of results for (top left) circular, (top right) elliptical, and (bottom left) random rain, holes, and noisy comparisons. For the elliptical comparisons, the dashed vertical lines separate the different error types with the first set's being translation-only, followed by size-only, then translation and size, followed by rotation-only, and finishing with size, rotation, and translation.

field with only one nonzero grid point, but where the one-valued grid point is in two opposite corners from P3 to P4. However, Δ gives a different answer for P1P5, which is the same as the other two except that the single one-valued grid point is now in the center of the field instead of a corner. For Zhu's measure, the MED component is undefined for P1P1 because the set over which the average is to be taken is empty. If $\lambda_2 = 0$, then Zhu's measure is also zero. One might consider defining MED to be zero for this case, as is considered in the results here, because both sets A and B are empty.

Because the majority of pathological comparisons include either one or two fully empty fields or exhibit very large frequency bias, the dFSS value is not defined for most of these comparisons. The only exceptions are P2P2, where the dFSS gives a perfect result of 0 and P6P7 where the dFSS gives a result of 63 points, which is an underestimation compared to the true displacements of 100 points. This underestimation is probably related to the fact that P6P7 exhibits an extreme example of the border effect because all the isolated points are located on the opposite sides of the domain at the borders.

Generally, none of the measures provides satisfactory results for the limiting case of no event areas, and each is fairly sensitive to the addition of only a few events. A possible solution for a user needing to summarize over numerous comparisons (e.g., over many time points) would be to report such cases as not applicable for these measures, and complement every case with the frequency bias and the size of event areas in each field of the verification set. Determining a lower bound of event area size is a question best left to specific users to decide depending on the variable being studied and their particular needs.

(ii) Circle test-case comparisons: Figure 6 shows the results of the circle test-case comparisons for the

TABLE 6. List of proposed comparisons.

Pathological	P1P1, P1P2, P2P2, P1P3, P1P4, P1P5, P2P5, P2P6, P6P7, P1C1, P2C1
Circular	C1C2, C2C3, C2C4, C1C4, C3C4, C2C5, C3C5, C2C11, C1C6, C6C7, C6C8, C1C9, C1C10, C6C12, C13C14
Elliptical	SE1E9, E2E10, E3E11, E4E12, E6E14, E3E7, E2E6, E4E8, E1E13, E7E11, E1E3, E5E7, E1E4, E2E4, E4E10, E1E11, E6E16, E1E14, E2E16, E4E14, E2E17, E2E18, E19E20
Scattered, holes, and noisy	S1S2, S1S3, H1H2, N1N2, C1N3, C1N4

various distance measures. Baddeley's Δ metric is using $p = 2$ and the constant, $c = \infty$. Because this metric is a true mathematical one, it is symmetric. The sensitivity of the metric to the location of the pairs of objects in the field can be seen as the value changes depending on the location of the two objects. For example, C1C2 is identical to C2C3 in every way except their position within the domain, where the latter is near the boundary. Additionally, the comparison C2C4 is also identical to C1C2 and C2C3 except that they are positioned vertically instead of horizontally to one another, but the values for Δ changes, although only slightly, for each comparison at 29, 31, and 28, respectively. The Hausdorff distance does not suffer from this issue and results in a value of about 40 for all three comparisons. Interestingly, CDST also equals 40 meaning that the average translation error is equal to the maximum translation error for this set of test-case comparisons. Δ penalizes C2C11 more than C1C2, C2C3 or C2C4, which are effectively subset cases of C2C11. That is, it does penalize for having over or under forecast the areal extent. The Hausdorff distance does not detect this behavior; having the same value for each of these comparisons.

The squared and centered version of Baddeley's Δ, $\eth$, is not shown in the figure but yields more consistent results as expected; its value being the same for C1C4 and C3C4. Moreover, it penalizes C2C11 more than it does the first three comparisons, although only by about half a grid point when $c = \infty$. It also yields a major difference between comparisons C6C7 and C6C8, which involve a similar error, but where C6C8 has a much larger displacement for the southern circle than C6C7. $\eth$ slighlty favors comparison C1C9 over C1C10, but by only about 1/2 and 1/3 of a grid point for $c = 50$ and $c = \infty$, respectively. C1C9 has some overlapping events while C1C10 has no areas of overlap. On the other hand, C1C10 has more correct negative area, which is the reason for the closeness in the two $\eth$ values. Both Δ and $\eth$ give sensible values for the comparisons where the circles are the same size and shape and have a slight enough translation error that they overlap (C2C5, C3C5, and C6C12).

The dFSS gives a reasonable estimate for the translational displacement for the comparisons with a single displaced circle (C1C2, C2C3, C2C4, C1C4, C3C4, C2C5, C3C5) albeit somewhat underestimating the true size of the displacement. Similarly to Δ the dFSS is also somewhat sensitive to the positioning of the circle pairs as its value also changes slightly depending on their relative positions within the domain (34, 32, and 34, for C1C2, C2C3, and C2C4, respectively), mainly changing depending on how close to the border the circles are located. It also gives close to the true average displacement for the C6C7/C6C8 comparisons, which are composed of two sets of circles where one is displaced and the other is not displaced. For example, the true displacement of the bottom set of circles in C6C7 is 40 points for an average displacement of 20, whereas the dFSS gives 17 points. A somewhat similar comparison is C13C14 where the smaller upper set is displaced by 25 points while the larger bottom set is displaced by 50 points. Because larger events have a larger influence on the dFSS, the resulting value of 43 is closer to 50 than to 25 points. dFSS also does a good job for comparison C6C12 where the two sets of circles are displaced by $20\sqrt{2} \approx 28$ points while the dFSS gives 25 points. For comparisons C2C11, C1C6, C1C9, and C1C10 the dFSS value is not defined because the frequency bias is too large.

For the comparisons with a single displaced circle the CDST gives a perfect estimate of translational displacement, and similarly for C6C7, C6C8, and C13C14. The limitations of the CDST become very apparent in comparisons C1C6, C1C9, C1C10, and C6C12 where the resulting CDST is 0 despite that the fields are very different from one another; but each pair has the same centroids as each other. Of course, if the intent is to compare the detailed structure of different fields, it would not be recommended to use a summary measure such as the CDST in the same way as the mean should not be used to measure variability, and these results illustrate the need to understand what sort of information each method provides.

Because Zhu's measure is a linear combination of RMSE and MED, it is not surprising that the values are higher, but analogous as the MED. In each case, these values give a reasonable summary of how closely one field matches the other, where the asymmetry, again, indicates whether or not one event set is a subset of the other or not; Zhu's method

emphasizes the asymmetry less because of the symmetric RMSE component.

(iii) Elliptical comparisons: Elliptical comparisons with the distance measure results are shown in Fig. 8. It has already been seen that Baddeley's Δ and dFSS are somewhat sensitive to the relative positions of event areas in the domain so it is not surprising that these values fluctuate a bit for comparisons E1E9, E2E10, E3E11, and E4E12, which all have the same two ellipses separated by the same translation error, but rotated in different directions. Similarly, as mentioned earlier, the dFSS value is not defined in the presence of large bias. Finally, the deficiencies with CDST have already been addressed as well. Otherwise, the various distance measures provide similar results for the elliptical scenarios.

Comparisons E1E3 and E1E4 both have zero CDSTs with a rotational displacement. Similar comparisons were analyzed using dFSS in (Skok and Roberts 2018). Which of these cases is "best" will, as usual, depend on the specific user. For E1E3 and E1E4 the dFSS-derived distances are about 17% and 13% of the ellipse's major axis length, which is consistent with the fact that about a quarter of the ellipse is overlapping (24% for E1E3 and 33% for E1E4). Similar results are given by Δ, MED and Zhu's measure; noting that MED gives lower values and Zhu's measure gives higher values, but they yield similar information on their own scales.

Comparisons E3E11 and E7E11 represent almost the same situation, but where one ellipse is much smaller (E7E11) than in the other (E3E11). Δ is slightly lower (indicating a better agreement between the two fields) for the more similar objects in E3E11 than E7E11. Similarly for avg MED(A, B), which is 17 for E3E11 and 20 for E7E11. On the other hand, min MED(A, B) = 17 for E3E11 and only 16 for E7E11, suggesting that E7E11 would be preferred while max MED(A, B) = 17 for E3E11 and 24 for E7E11, so that it more emphatically prefers E3E11 over E7E11. Again, which of E3E11 and E7E11 is "better" will depend on a specific user's needs. An asymmetric measure such as MED, applied in both directions, provides a more thorough summary in this situation.

The most general conclusion from the elliptical comparisons by looking at Fig. 8 for each method is that greater overlap is the most important property to reduce each measure's value closer to perfect. Reducing the size of one feature also seems to reduce the values of the measures; implying that a forecast could be hedged to improve performance by underforecasting event areas.

(iv) Scattered, holes, and noisy comparisons: Fig. 10 shows the results for the scattered, holes and noisy comparisons. The comparisons S1S2 and S1S3 are designed to test how the methods handle small-scale events scattered inside an envelope. In S1S2 the envelopes overlap. Because the positioning of nonzero grid points inside the envelope is random the displacement can be interpreted in terms of average nearest neighbor distance from each nonzero grid point to the closest nonzero grid point in the other field, which is about 2.35 grid points. The Δ, MED and dFSS all give very similar results that reflect the average nearest neighbor distance well. Zhu's measure gives a much larger value while, as could be expected, the CDST gives a smaller value because the centers of mass of the two envelopes are near each other. In S1S3 the envelopes are displaced by 100 grid points and is reasonably well represented by values of CDST, dFSS, and MED; while Δ and especially Zhu's measure give smaller values.

The comparison H1H2 is designed to test how the methods behave when the majority of the domain is covered by events with holes present in the fields. The results of H1H2 can be compared to C1C2 because these comparisons use identical fields with inverted values. As already mentioned, the result for H1H2 and C1C2 would be the same for the RMSE metric because RMSE treats zero and nonzero areas in an equal manner. However, all the tested measures give a better (smaller distance) value for the H1H2 than for C1C2. With the exception of Zhu's measure, they all provide a very small distance; ranging between 0.3 and 2.3 grid points for the H1H2, while the distances for C1C2 tend to be about 10 times larger. This result shows that the tested distance measures tend to focus more on analyzing the spatial matching of event areas (i.e., region of nonzero values, which have a very good overlap in H1H2) as opposed to analyzing the nonevent areas (which have a very good overlap in C1C2). This property is also reflected in the fact that all of the tested methods have problems dealing with empty fields.

The comparisons N1N2, C1N3, and C1N4 are designed to test how sensitive the methods are to the presence of noise in the fields. The results of N1N2 can be compared to the result of C1C4. Compared to C1C4 the MED and Δ values are reduced by about 50% for N1N2, which indicates a strong sensitivity of these methods to noise. At the same time, the value of Zhu's measure is reduced by about 20% while CDST and dFSS do not exhibit much sensitivity to noise. The position of the added noise is also a factor as noise close to existing features has less impact than

noise that is far away from existing events in the same field and closer to the events in the other field. This effect is shown in C1N3 where a single nonzero value is added into the middle of the event in the other field; thus considerably reducing the distances provided by MED and Δ. On the other hand, in C1N4 the single added point is placed too far from the event in the other field to cause any major impact on the results of any measure.

(v) Overall comparison: As mentioned previously, so many cases can make it difficult to synthesize all of the information. It is suggested that these comparisons be investigated for any new measure that primarily, or in part, accounts for location errors. While many cases may not shed any light on a particular measure's properties, it is highly likely that at least one case will uncover important information. Such a case can then be emphasized when reporting on the measure's properties. Additionally, in order to avoid cherry-picking results, it is important to try to display the results for all of the cases.

Figure 11 shows scatterplots for each measure for all but the pathological comparisons. In general, except for centroid distance, the measures rank the comparisons similarly, suggesting that each of these measures typically provides similar information regarding forecast performance. However, a closer look at the figure suggests some variations and differences in the information they provide; for example, Hausdorff distance is an outlier in the C6C8 comparison; moreover, the ordering of distance measures varies by comparison. In terms of general sensitivities, the most obvious conclusion is that greater overlap of event areas is the most important factor in reducing each measure's value toward the perfect score, which is zero for each of these measures.

5. Conclusions and discussion

This article proposes a new set of idealized test cases for spatial verification methods that serve as a useful common test set that can be employed for new methods as they come to the fore. The test cases represent some very challenging, but realistic verification scenarios to summarize with only a few measures. It is hoped that these cases will be used by others who introduce new methods so that they can be compared on a standard set of cases.

The usefulness of the new test set is also highlighted by the analysis of geometric indices and several distance measures. For example, the analysis showed that all the methods have problems dealing with empty fields devoid of any events. In these situations some methods are not able to provide a result at all while others use a special rule to prescribe some value that is more or less arbitrarily defined. They also demonstrate how distance measures are capable of assessing distances between features, and are sensitive to spatial overlap of features.

For the pathological cases, it is clear that a plan must be implemented by operational centers and other users in order to handle empty cases, noise and cases with only a few events. As suggested in Gilleland (2017), the traditional frequency bias is a useful complementary measure for all of these distance measures and geometric indices. It might be a small matter of identifying cases of only a few event points, flagging them, and perhaps reporting on their frequency bias.

Many of the measures discussed in this paper can be, or are already, used within a framework such as the method for object-based diagnostic evaluation (MODE; Davis et al. 2009). The results from this treatment can provide important information about how these measures can be affected when MODE objects are small, or matched objects have largely different sizes, are translated in equal but opposite directions, and so on. The CDST, for example, has often been used to merge and match objects, and it is clear from this work that inappropriate mergings or matchings can occur if using only this measure for this purpose.

It is also found that various methods provide duplicate information in many situations, but that each method has its positive properties and limitations. For example, the CDST gives an average of true translation errors, but suffers from giving a perfect result for some situations when the fields are wildly different from each other.

Depending on a user's need, the min MED(A, B) or max MED(A, B) from Eq. (1) may also be useful. A perfect value of min MED(A, B) = 0 would imply that either A or B is perfectly aligned spatially with the other. Having a perfect score, however, does not mean that both fields are identical. For example, if min MED(A, B) = 0, then one of the sets A or B is contained in the other. However, it is possible that MED(A, B) is very large when taken as the average distance from the larger set. A perfect (zero) value for max MED(A, B), however, implies that the two fields are identical, and a high value would mean that at least one of the two fields has nonzero grid values where the other field does not.

Zhu's measure is a linear combination of RMSE applied to the binary fields and the MED in one direction, and so is very similar to MED. The addition of the RMSE term allows it to give a closer approximation to true translation errors. However, the user must choose the weights, which were taken to be 0.5 for both in this study. This measure performs erratically in some situations in that it ranks as best or worst when not expected (e.g., comparison C2C5 vs C13C14 or S1S2 vs H1H2).

The CDST gives the average translation error by definition, which will be exactly the translation of a single

object that is translated uniformly in one direction. However, the average translation distance is not very informative under more complicated situations, such as high-frequency bias and multiple translations of several objects in differing directions. The dFSS exhibits some dependence on the orientation of the displacement and the positioning of the events inside the domain; however, the influence of these two factors on its value tends to be small. The biggest drawback of the dFSS is that it can only be used if the frequency bias is small.

In terms of their sensitivity to noise, the measures based on the distance map, such as Δ, Hausdorff, MED and Zhu's measure, all have a high sensitivity to noise; even a single strategically placed nonzero grid point can greatly change their values in most situations, whereas the dFSS and CDSTs, by contrast, are relatively insensitive.

Acknowledgments. Support for the first author on this project was provided by the Developmental Testbed Center (DTC). The DTC Visitor Program is funded by the National Oceanic and Atmospheric Administration, the National Center for Atmospheric Research, and the National Science Foundation. The second author acknowledges the financial support from the Slovenian Research Agency (Research Core Funding P1-0188). The authors thank two anonymous reviewers and Caren Marzban for many detailed and thoughtful comments that helped to improve the presentation of the results.

REFERENCES

AghaKouchak, A., N. Nasrollahi, J. Li, B. Imam, and S. Sorooshian, 2011: Geometrical characterization of precipitation patterns. *J. Hydrometeor.*, **12**, 274–285, https://doi.org/10.1175/2010JHM1298.1.

Ahijevych, D., E. Gilleland, B. G. Brown, and E. E. Ebert, 2009: Application of spatial verification methods to idealized and NWP-gridded precipitation forecasts. *Wea. Forecasting*, **24**, 1485–1497, https://doi.org/10.1175/2009WAF2222298.1.

Baddeley, A. J., 1992: An error metric for binary images. *Robust Computer Vision Algorithms*, W. Forstner and S. Ruwiedel, Eds., Wichmann, 59–78.

Borgefors, G., 1986: Distance transformations in digital images. *Comput. Vis. Graph. Image Process.*, **34**, 344–371, https://doi.org/10.1016/S0734-189X(86)80047-0.

Brown, B. G., E. Gilleland, and E. E. Ebert, 2011: Forecasts of spatial fields. *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, 2nd ed. I. T. Jolliffe and D. B. Stephenson, Eds., John Wiley & Sons, Ltd., 95–117.

Davis, C., B. Brown, and R. Bullock, 2006a: Object-based verification of precipitation forecasts. Part I: Methodology and application to mesoscale rain areas. *Mon. Wea. Rev.*, **134**, 1772–1784, https://doi.org/10.1175/MWR3145.1.

——, ——, and ——, 2006b: Object-based verification of precipitation forecasts. Part II: Application to convective rain systems. *Mon. Wea. Rev.*, **134**, 1785–1795, https://doi.org/10.1175/MWR3146.1.

——, ——, ——, and J. Halley-Gotway, 2009: The Method for Object-Based Diagnostic Evaluation (MODE) applied to numerical forecasts from the 2005 NSSL/SPC spring program. *Wea. Forecasting*, **24**, 1252–1267, https://doi.org/10.1175/2009WAF2222241.1.

Dorninger, M., E. Gilleland, B. Casati, M. P. Mittermaier, E. E. Ebert, B. G. Brown, and L. J. Wilson, 2018: The setup of the MesoVICT project. *Bull. Amer. Meteor. Soc.*, **99**, 1887–1906, https://doi.org/10.1175/BAMS-D-17-0164.1.

Fabbri, R., L. D. F. Costa, J. C. Torelli, and O. M. Bruno, 2008: 2D Euclidean distance transform algorithms: A comparative survey. *ACM Comput. Surv.*, **40**, 2:1–2:44, https://doi.org/10.1145/1322432.1322434.

Gilleland, E., 2011: Spatial forecast verification: Baddeley's delta metric applied to the ICP test cases. *Wea. Forecasting*, **26**, 409–415, https://doi.org/10.1175/WAF-D-10-05061.1.

——, 2017: A new characterization within the spatial verification framework for false alarms, misses, and overall patterns. *Wea. Forecasting*, **32**, 187–198, https://doi.org/10.1175/WAF-D-16-0134.1.

——, 2018: SpatialVx: Spatial Forecast Verification, version 0.6-3. R package, accessed 9 March 2020, https://CRAN.R-project.org/package=SpatialVx.

——, D. Ahijevych, B. G. Brown, B. Casati, and E. E. Ebert, 2009: Intercomparison of spatial forecast verification methods. *Wea. Forecasting*, **24**, 1416–1430, https://doi.org/10.1175/2009WAF2222269.1.

——, ——, ——, and E. E. Ebert, 2010: Verifying forecasts spatially. *Bull. Amer. Meteor. Soc.*, **91**, 1365–1376, https://doi.org/10.1175/2010BAMS2819.1.

Mass, C. F., D. Ovens, K. Westrick, and B. A. Colle, 2002: Does increasing horizontal resolution produce more skillful forecasts? *Bull. Amer. Meteor. Soc.*, **83**, 407–430, https://doi.org/10.1175/1520-0477(2002)083<0407:DIHRPM>2.3.CO;2.

Meijster, A., J. B. T. M. Roerdink, and W. H. Hesselink, 2000: A general algorithm for computing distance transforms in linear time. *Mathematical Morphology and Its Applications to Image and Signal Processing*, J. Goutsias et al., Eds., Kluwer Academic, 331–340.

Roberts, N. M., 2008: Assessing the spatial and temporal variation in the skill of precipitation forecasts from an NWP model. *Meteor. Appl.*, **15**, 163–169, https://doi.org/10.1002/met.57.

——, and H. W. Lean, 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Mon. Wea. Rev.*, **136**, 78–97, https://doi.org/10.1175/2007MWR2123.1.

Rossa, A. M., P. Nurmi, and E. E. Ebert, 2008: Overview of methods for the verification of quantitative precipitation forecasts. *Precipitation: Advances in Measurement, Estimation and Prediction*, S. C. Michaelides, Ed., Springer, 418–450.

Schwedler, B. R. J., and M. E. Baldwin, 2011: Diagnosing the sensitivity of binary image measures to bias, location, and event frequency within a forecast verification framework. *Wea. Forecasting*, **26**, 1032–1044, https://doi.org/10.1175/WAF-D-11-00032.1.

Skok, G., and N. Roberts, 2016: Analysis of Fractions skill score properties for random precipitation fields and ECMWF forecasts. *Quart. J. Roy. Meteor. Soc.*, **142**, 2599–2610, https://doi.org/10.1002/qj.2849.

——, and ——, 2018: Estimating the displacement in precipitation forecasts using the Fractions skill score. *Quart. J. Roy. Meteor. Soc.*, **144**, 414–425, https://doi.org/10.1002/QJ.3212.

Zhu, M., V. Lakshmanan, P. Zhang, Y. Hong, K. Cheng, and S. Chen, 2011: Spatial verification using a true metric. *Atmos. Res.*, **102**, 408–419, https://doi.org/10.1016/j.atmosres.2011.09.004.