Advanced Distance-based, Field Deformation and Statistical Hypothesis Testing Utilities for MET: Final Report

10/07/2019

Eric Gilleland
Joint Numerical Testbed
Research Applications Laboratory
National Center for Atmospheric Research

Summary

This visit was primarily aimed at adding additional complementary methods to MET for performing spatial forecast verification. In particular, to add components for directly measuring similarities/differences between a forecast model and gridded observation in terms of the spatial location (size, shape, etc.) of event areas determined by exceeding a specified threshold value. Several new summary measures were added to MET, a paper has been submitted to the AMS journal, Monthly Weather Review, and an additional write-up was posted online to provide information about how to perform competing forecast verification using output from MET. The submitted paper introduces many new test cases based on simple geometric shapes designed to challenge verification measures, and they spring off of previous test cases found to have been very useful in ascertaining how different spatial verification methods behave. Additionally, the various methods added to MET were tested on these cases. Mostly, it was found that they all give similar information, but each has its pluses and minuses. It was also determined that certain complementary information, such as the number of "events" should be used in conjunction with these scores to assist interpretations.

Distance-map Based Measures

Many location-based measures utilize distances between objects from one field to another as if they are superimposed onto the same field. In each case, a binary field is first created by setting all values above a user-specified threshold to one and setting the rest to zero. These types of measures would take considerable computational time if not for fast algorithms that allow one to compute a distance map over the entire domain (assumes a rectangular grid). Summary measures that require only a subset of the domain can then simply mask out the part of interest.

A distance map for a binary event set, $A \subset \mathcal{D}$ ($\mathcal{D}$ a spatial domain), gives the shortest distance from every grid point, $\boldsymbol{s} = (x, y)$, in $\mathcal{D}$ to the nearest grid point in A and is denoted by $d(\boldsymbol{s}, A)$. The left panel of Figure 1 illustrates the information provided by the distance map. Most measures make use of only the part of the distance map for one field (say that contains an event area labeled A as in the right part of the figure) where a second object (from another field with the same domain), labeled B, resides. That is, the distance map of A is, for example, averaged over just the values within the event area B.

An early, and popular, measure from image analysis is the Hausdorff distance, which gives the maximum shortest distance between two objects.

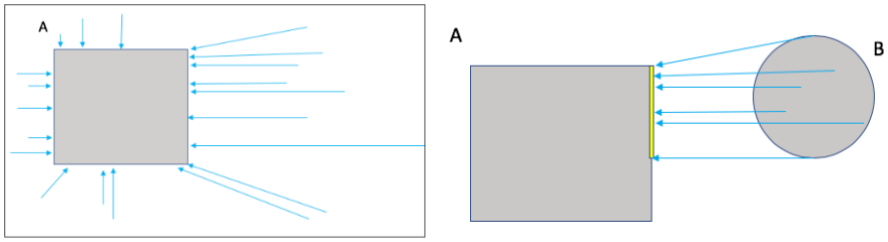

Figure 1: Left is a graphic depicting how a distance map is calculated. It uses every point in the domain (and is therefore sensitive to the domain's size and shape). Each grid point gives the shortest distance from that point to the nearest one-valued grid point in the binary field (the event). Right shows how a subset of a distance map is calculated from one object (event areas defined to be one-valued grid points in the binary field) to another object (hence, the domain boundary is not shown). Note that the distances are still to the nearest point in the other object so that in the figure all of the distances are to the yellow line on the border of the event area labeled A. The Baddeley metric is based on the entire distance map, where others such as mean-error distance, Zhu's measure, and Pratt's Figure of Merit are based on distances between specific objects as in the right panel.

The specific measures added to MET include: Baddeley's Δ (Baddeley 1992; Gilleland et al. 2008; Gilleland 2011), the Hausdorff distance (H, Baddeley 1992; Gilleland et al. 2019), mean-error distance (MED, Peli and Malah 1982; Gilleland 2017), Zhu's measure (for one forecast, henceforth denoted by Z, Zhu et al. 2011; Gilleland et al. 2019), and Pratt's Figure of Merit (F, Pratt 1977; Abdou and Pratt 1979). Hausdorff's distance is one of the oldest, and most popular, metrics from image analysis.

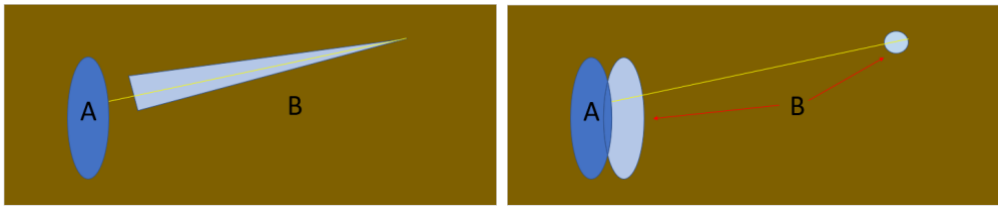

Figure 2: Illustration of Hausdorff's distance, $H(A,B)$. It is the maximum, shortest-distance, between two objects A and B. The depiction in the left panel illustrates why it is useful, but the illustration in the right panel cautions that a small change in the field (e.g. a single event turned on far away from the object) can greatly increase its score.

Figure 2 illustrates $H(A,B)$ for two different scenarios. It is the largest distance among all of the shortest distances between every point in the event set B and the nearest point in A. The left panel shows why it is a reasonable score to consider, because it shows the worst agreement in location between two objects. The right panel, on the other hand, emphasizes its sensitivity to small changes in the field (i.e., the addition of even a single non-zero grid point somewhere in the domain far from the set A). It is precisely this sensitivity that provided the motivation for the metric, Δ.

Hausdorff distance and Baddeley's Δ

The Hausdorff distance is given by

$$H(A,B) = \max\{|d(\boldsymbol{s},A) - d(\boldsymbol{s},B)|\}$$

Baddeley's Δ is given by

$$\Delta_{p,w}(A,B) = \left[\frac{1}{N}\sum_{\boldsymbol{s}\in\mathcal{D}}\left|w\big(d(\boldsymbol{s},A)\big) - w\big(d(\boldsymbol{s},B)\big)\right|^p\right]^{1/p},$$

where N is the size of the domain, $w(\cdot)$ is a concave function, and specifically for MET, $w(x) = \min\{x, c\}$, for a user-chosen constant c, p is a user-chosen parameter that controls the type of averaging. For example, $p = 1$ gives the straight average of $\left|w\big(d(\boldsymbol{s},A)\big) - w\big(d(\boldsymbol{s},B)\big)\right|$. Usually, the Euclidean average is chosen by setting $p = 2$. The limit as $p \to \infty$ gives the Hausdorff distance as a special case; generally the function $w(\cdot)$ is not applied with the Hausdorff distance, although it can be.

Baddeley's Δ is best described through the distance maps. Figure 3 shows a field with a circular event area labeled A and a second binary field with a similar circular event area labeled B. The two circles are identical in every way, except that B is translated 40 grid squares to the right and touches the edge of the domain. The next two panels show the distance maps for each field, and Figure 4 shows the absolute magnitude of the differences between these two distance maps. It can be shown that $H(A,B)$ is the maximum value, and $\Delta(A,B)$ is the L_p norm of the graph in Figure 4. Because of the sensitivity of a distance map to the domain's size and shape, before calculating Δ, the concave function $w(\cdot)$ may be applied to the individual distance maps. While applying this function reduces domain effects, it does not remove them entirely. Of the distance-map based measures described here, only the Baddeley, and to a lesser extent, the Hausdorff distance, are affected by the domain size and shape. The others are not affected because of their conditioning on the event sets.

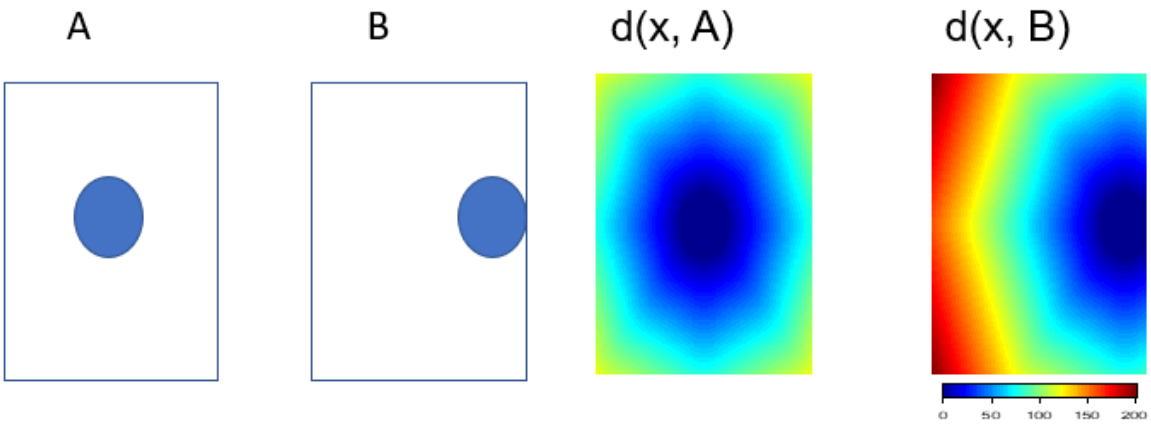

Figure 3: Binary event areas, A and B, in two separate fields (left two panels) and their associated distance maps (middle two panels). The events A and B are from Gilleland (2017) and also utilized in Gilleland et al. (2019).

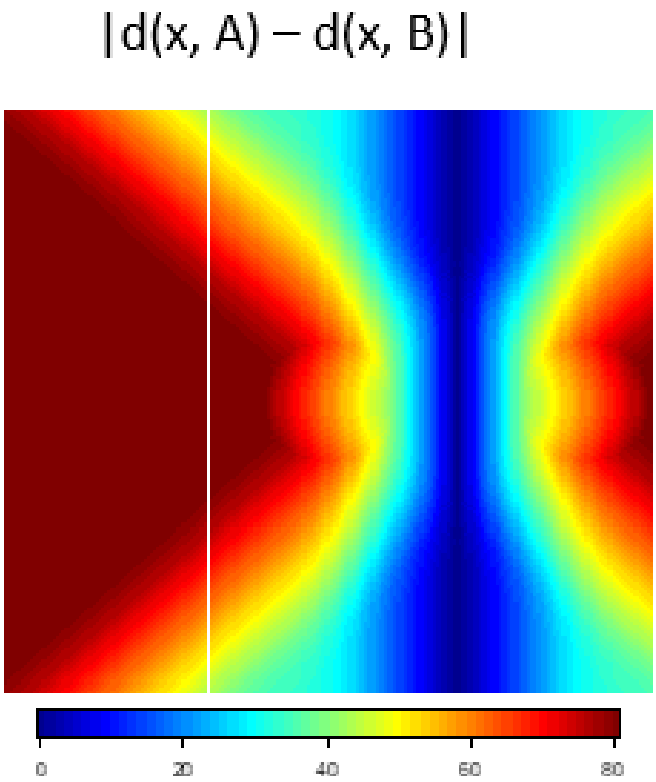

Figure 4: The absolute difference between the two distance maps shown in Figure 3. Baddeley's Δ is the L_p norm of this graph (although sometimes a concave function is first applied to the middle two panels in order to reduce domain effects).

Mean Error Distance and Zhu's Measure

The mean-error distance is the average of shortest distances from one event area to the nearest point in the other event area. Figure 5, for example, demonstrates how the MED(B, A) is calculated using the same example from Figure 3. The MED is given by

$$\mathrm{MED}(A,B) = \frac{1}{n_B}\sum_{\boldsymbol{s}\in B} d(x,A),$$

where n_B is the number of one-valued grid squares in the binary event set B. Note that the sum is only over the one-valued grid squares defined by the event set B rather than the entire domain $\mathcal{D}$. Subsequently, the measure is not sensitive to the size or shape of $\mathcal{D}$ the way H and Δ are.

$Z(A,B)$ is defined as a weighted average between the root-mean-square deviation (RMSD) between the two binary fields, for example the leftmost two panels in Figure 3, and MED. That is,

$$Z(A,B) = \lambda\sqrt{\frac{1}{N}\sum_{\boldsymbol{s}\in\boldsymbol{\mathcal{D}}}\left(I_A(\boldsymbol{s}) - I_B(\boldsymbol{s})\right)^2} + (1-\lambda)\cdot\mathrm{MED}(A,B),$$

where I_A and I_B denote the binary fields consisting of all the zero- and one-valued grid points for the field containing the event sets A and B, respectively, and λ is a user-chosen weight to give more or less importance to one or the other components of RMSD and MED; the default in MET is equal weighting so that $\lambda = \frac{1}{2}$.

Pratt's Figure of Merit

All of the measures discussed so far have a range of zero to infinity (the domain size) with zero as a perfect score and increasing values indicating worse matches. The units are the distance used in their calculation, which in order to use the fast-computational algorithms means that the distances provided by MET are in grid squares. If a grid is such that grid squares are approximately equal and of a specific size, say 4-km, then the grid square units could simply be multiplied by 4-km to obtain the approximate distance in km.

$F(A,B)$, or $F(B,A)$, is similar to MED, but it is defined so that it is unitless and between zero and one, where zero is the worst possible score and one is the best. It is given by

$$F(A,B) = \frac{1}{\max\{n_A, n_B\}} \sum_{\boldsymbol{s} \in B} \frac{1}{1 + \alpha d(\boldsymbol{s}, A)^2},$$

where n_A, n_B are the number of one-valued grid squares in A and B, respectively, and α is a user-defined scaling constant. The default in MET for α is one-ninth, which is the most common choice.

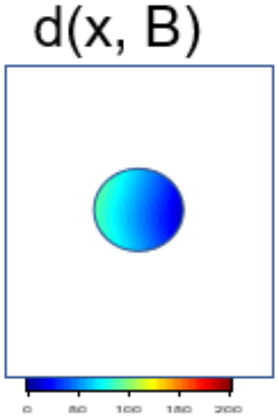

Figure 5: The distance map for the event area B of Figure 3, masked out by the event area A. Averaging over the circle in this figure will give the mean-error distance, MED(B,A).

Properties of Distance-based Measures

Given the numerous possible summary measures for distance-based algorithms, it is valuable to understand their properties. Therefore, a rigorous evaluation process was applied to all of the

measures described above, and included in MET (https://dtcenter.org/community-code/model-evaluation-tools-met), among others (Gilleland et al. 2019).

In image analysis, a true mathematical metric is usually desired. A mathematical metric, $m(A, B) \geq 0$, must have the following three properties:

i. (identity) $m(A, B) = 0$ if and only if $A = B$.
ii. (symmetry) $m(A, B) = m(B, A)$.
iii. (triangle inequality) $m(A, C) \leq m(A, B) + m(B, C)$.

The first establishes that a perfect score is zero and that the only way to obtain a perfect score is if the two sets are identical according to the metric. The second requirement ensures that the order by which the two sets are evaluated will not change the result. The third property ensures that if C is closer to A than B is to A, then $m(A, C) < m(A, B)$.

The measures Δ and H are true metrics, but the rest fail the symmetry property. Gilleland (2017) argued that this lack of symmetry is useful in the forecast verification context because it allows for gleaning information about false alarm type of errors against miss type errors. Nevertheless, MET provides, for convenience, three additional variations of MED, Z and F that are symmetric and therefore are true metrics. For example, variations of MED provided in MET are:

$$\frac{1}{2}(MED(A, B) + MED(B, A)),$$

$$\max\{MED(A, B), MED(B, A)\},$$

$$\min\{MED(A, B), MED(B, A)\}.$$

Figure 6 illustrates how just being a metric does not necessarily mean the information provided is guaranteed to make sense. The centroid distance, for example, is a true mathematical metric. However, it is a measure describing two single points in space, the centroids of objects, and not the whole objects. Therefore, there are many possible ways that two event sets can have identical centroids but not be very similar. Subsequently, a host of test cases (more than 50), including some from Gilleland (2017) were proposed in Gilleland et al. (2019) and these measures were applied to them. Figure 3 and Figure 6 depict two of these cases.

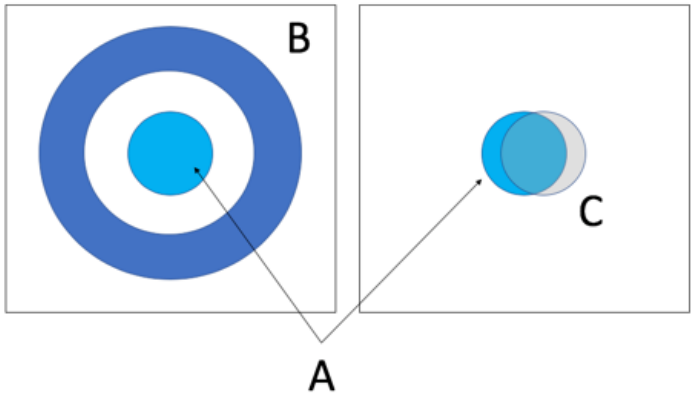

Figure 6: Illustration of how centroid distance, a true mathematical metric, might not give sensible information. On the left the two objects being compared, a circle and a much larger ring, have the same centroid and therefore a zero-valued centroid distance (perfect score!). On the right, the two objects are almost identical, but one is slightly translated to the right, so the centroid distance is positive (i.e., the centroid distance favors the pair on the left over the pair on the right).

Some cases involved pathological but very realistic, and commonly occurring, situations where one or both fields have no events. For example, if both forecast and observed fields have zero at every grid point (i.e., $A = B = \emptyset$), then a good summary measure should result in a perfect score. Some of the measures are technically defined for this case and they do provide a perfect score. However, suppose A contains a single one-valued grid square with zeros at all other points in the domain, and let $B = \emptyset$. In this case, while A and B are not a perfect match, they are nevertheless very similar, so it would be hoped that a summary measure would give a near perfect score. Unfortunately, they tend to give very bad scores in this case. In fact, if both sets have a single one-valued grid square, then they will either give a very good score or a very bad score depending on the relative placements of these singleton event sets. Specific users will need to use their judgement to decide how much event area is needed before these measures make sense. To help guide the user, information about the number of events is also returned from MET.

Figure 7 illustrates a small set of the comparisons proposed in Gilleland et al (2019). Each field is on a 200 by 200 square domain and is labeled the same as in the paper. The first set, labeled C6, has the top circle centered at (100, 140) and the second circle centered at (100, 60). The second event set, denoted C12, has the same two circles from C6 but translated an equal amount ($\sqrt{20^2 + 20^2} = 20\sqrt{2} \approx 28$ grid squares) in opposite directions with centers at (120, 160) and (80, 40).). The next pair of events compares C1, the self-same circle as A in Figure 6 against C9, which is another circle centered in the same location as C1, but with a much larger radius of 60 grid squares instead of 20. The next row shows event sets C1 v. C4 and C1 v. N4. N4 is identical to C4 except for the addition of a single event in the lower left corner of the domain. Event set C4 is the same size circle as C1 but translated 40 grid squares in both directions (i.e., centered at (140, 140) or translated by $\sqrt{40^2 + 40^2} = 40\sqrt{2} \approx 56$ grid squares).

Table 1 shows the results for the distance measures for the comparisons in Figure 7 and Figure 8.

Table 1: Results of distance measures for the comparisons in Figure 7 and Figure 8. Values rounded to the nearest whole number.

Comparison	A=C6, B=C12	A=C1, B=C9	A=C1, B=C4	A=C1, B=N4	A=C2, B=C11
Baddeley's Δ ($c = \infty, p = 2$)	19	38	41	37	29
Hausdorff distance	28	80	57	120	40
MED(A,B)	11	22	38	39	22
MED(B,A)	11	0	38	38	11
Zhu(A,B)	38	61	45	45	47
Zhu(B,A)	38	50	45	45	42
Centroid Distance	0	0	57	56	13

Figure 8 depicts C2 versus C11, where C2 is the same as C1 but centered 40 grid squares to its right, and C11 consists of three circles, also of radius 20, and centered 40 grid squares above and to the left and right of C2.

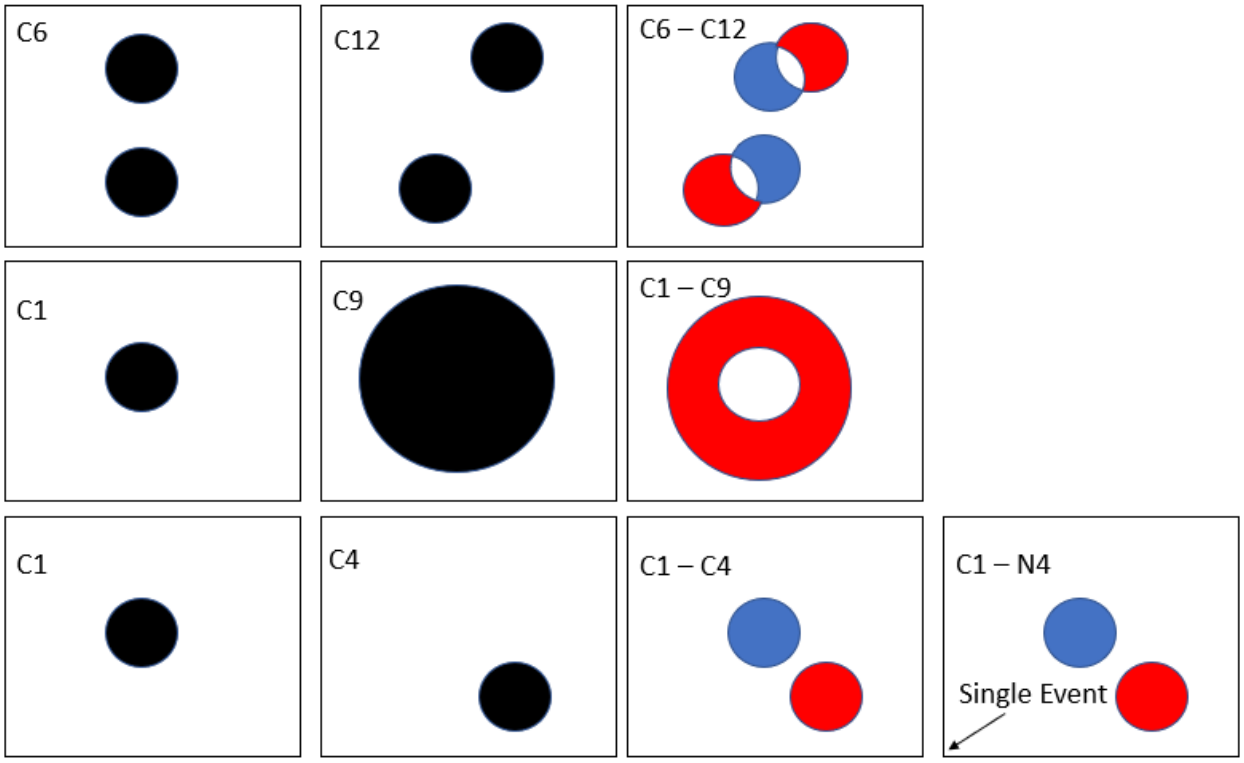

Figure 7: Graphic illustration of some of the comparisons proposed in Gilleland et al (2019). The left two columns show the individual fields containing a binary image. The third column shows the difference of the field in the first column less that of the second (red = −1, white = 0 and blue = 1). The last panel in the bottom row illustrating the difference between C1 and N4 is the same as the comparison of C1 with C4 except that N4 is contaminated with a single event in the lower left corner.

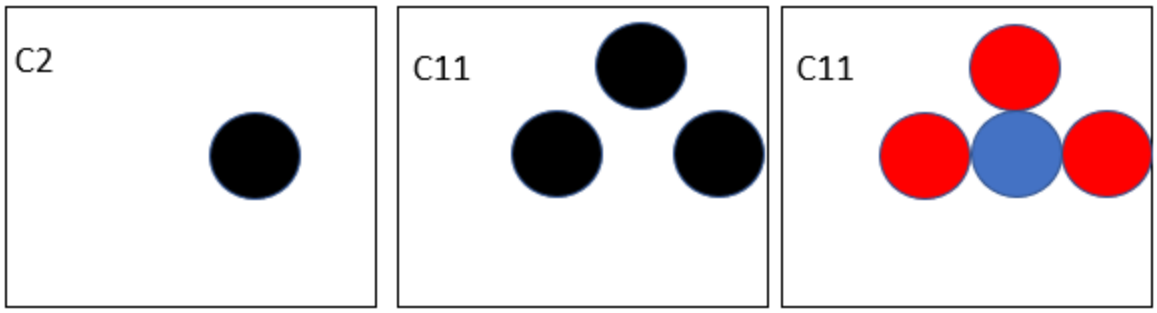

Figure 8: Another example from Gilleland et al (2019), but originally used in Gilleland (2017).

Sub-domains

One caveat about the measures in MET concerns the situation where a user wants to have summary information over sub-domains rather than an entire domain. For practical purposes, MET calculates the distance maps over the entire domain once rather than separately over each sub-domain. In the case of measures such as MED, if a subdomain has no objects, then NA is returned. If it has an object in one field in the subdomain and not the other, then NA will be returned for one direction, but a value is returned for the other direction. It is interpreted to mean that it is the average distance from the object in the sub-domain to the nearest object outside the sub-domain (from the other field). In the case of Δ, however, the interpretation is not as pragmatic, but one can still make sense of it. Namely, it is *essentially* the contribution of that sub-domain to the metric Δ over the sub-domain.

Other work

Apart from adding the distance-based measures to MET and updating the necessary documentation, some additional work was performed per the proposed plan.

The initial plan was to add certain hypothesis testing capabilities to MET, but it was decided that their implementation in an operational context is not recommended because diagnostic plots should be inspected to ensure that assumptions are met, and remediations made in the case they are not. Instead, a write-up was prepared to help users utilize output from MET in order to perform hypothesis testing in the competitive forecast verification setting. The document, in addition to previously prepared documents (Gilleland 2010a;b), should prove useful to many researchers, and possibly some operational users too. It is currently available at https://ral.ucar.edu/staff/ericg/HypoTestingWriteUp.pdf. A citable URL from UCAR's OpenSky platform will also be available soon.

Image warping was also determined to be beyond the scope of an operational usage within a framework such as MET because of the difficulty in automating the procedure without human input, as well as the obstacle of validating the results over different software packages, etc. The issue stems from the numerous parameters that must be estimated (generally at least 200 control points) and the fact that no unique warp can be readily interpreted as "best." The

implementation is still being investigated for use in SpatialVx, and good progress was made during the visit toward this end.

References

Abdou, I. E. and W. K. Pratt, 1979. Quantitative design and evaluation of enhancement/thresholding edge detectors. *Proceedings of the IEEE*, **67**, 753 – 763.

Baddeley, A. J., 1992. *Robust Computer Vision Algorithms*, chap. An error metric for binary images, 59–78. W. Forstner and S. Ruwiedel, Eds., Wichmann, 402 pp.

Gilleland, E., 2010a. Confidence intervals for forecast verification. NCAR Technical Note NCAR/TN-479+STR, 71pp, URL http://dx.doi.org/10.5065/D6WD3XJM.

Gilleland, E., 2010b. Confidence intervals for forecast verification: Practical considerations, (OpenSky citable URL, http://n2t.net/ark:/85065/d7445kgz).

Gilleland, E., 2011. Spatial Forecast Verification: Baddeley's Delta Metric Applied to the ICP Test Cases. *Weather Forecast.*, **26** (3), 409 - 415, doi: 10.1175/WAF-D-10-05061.1.

Gilleland, E., 2017. A new characterization in the spatial verification framework for false alarms, misses, and overall patterns. Weather Forecast., 32 (1), 187 - 198, doi: 10.1175/WAF-D-16-0134.1.

Gilleland, E., T.C.M. Lee, J. Halley Gotway, R.G. Bullock, and B.G. Brown, 2008. Computationally efficient spatial forecast verification using Baddeley's Δ image metric. *Mon. Wea. Rev.* **136** (5), 1747 - 1757, doi: 10.1175/2007MWR2274.1.

Gilleland, E., G. Skok, B. G. Brown, B. Casati, M. Dorninger, M. P. Mittermaier, N. Roberts, and L. J. Wilson, 2019. A novel set of verification test fields with application to distance measures. Submitted to Monthly Weather Review on 3 August 2019 (temporarily available at: https://ral.ucar.edu/staff/ericg/NewGeomCasesFinalVersion_modified.pdf).

Peli, T. and D. Malah, 1982. A study on edge detection algorithms. *Computer Graphics and Image Processing*, **20**, 1 – 21.

Pratt, W. K., 1977. *Digital Image Processing*. John Wiley and Sons, New York, NY, U.S.A.

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, doi:10.1016/j.atmosres.2011.09.004.