Beyond Correlation and RMSE: Cutting-edge Tools for Evaluating Predictability

Earth System Predictability Across Timescales Workshop

Boulder, CO, U.S.A.

April 11, 2024

Eric Gilleland

Research Applications Laboratory

NSF National Center for Atmospheric Research

Introduction

There are many challenges to evaluating predictability when faced with large spatial data sets. First, one must obtain a verification data set that is, ideally, independent of the prediction set, and at the same time and scale. For the latter, it is seldom the case that they are at the same time and scale, and this representativeness uncertainty should be considered, though seldom are (cf. Tustison et al. 2001; Mittermaier 2018). A related topic concerns uncertainty in the observations, one of whose sources is representativeness error (Ferro 2017), which again is rarely considered.

Because observations are often not found on the same grid (e.g., they may be from point sources, radar at a different resolution, etc.), the usual approach is to create an "analysis" that is often found via the same prediction model that made the forecast, just with a much shorter lead time. Subsequently, the independence criterion is often not met. Neither this topic nor the issue of representativeness and/or observational error are discussed further in this poster, but they should be considered in any verification study.

The aim of this treatment is to address many of the issues that arise under the condition that a verification field is available at the same time and on the same grid as the prediction. As weather forecasts went to higher resolution grids, it was found that they often had worse verification results than did their coarser model counterparts. The reasons, however, had less to do with any superiority of the coarser model, and more to do with the verification tools, such as root-mean square error (RMSE) or correlation, as well as the myriad contingency-table based measures (Mass et al. 2002). The primary reasons for this paradox are twofold. One is the so-called "double-penalty" and the other is an over accumulation of errors in the higher resolution grid resulting from

many more chances to have errors. The double penalty results from a single error that is tallied twice rather than once. For example, a storm feature may be predicted perfectly in terms of its size, spatial shape, and intensity values, but is displaced spatially so that it does not overlap with the observation (cf. **Figure 1** left). Most traditional measures will tally a "miss" everywhere the green oval (marked with an "O") is in the two top rows of the figure, and then tally a false alarm everywhere there is a red oval (marked with an "F"), though the only error is a spatial displacement.

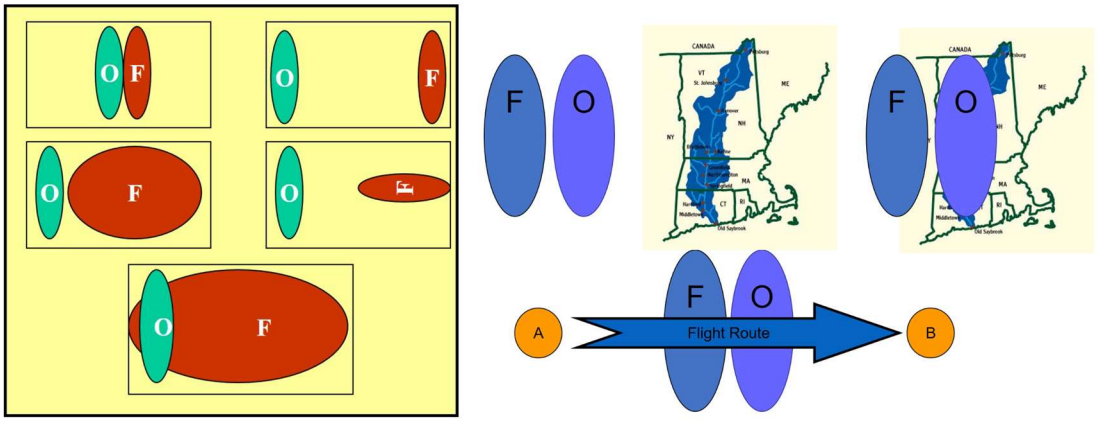

Figure 1: (Left) Graphical depiction of the double-penalty problem and the need for diagnostic information. Here, the green oval is the observed "storm" and the red represents various forecasts. Several traditional verification measures, such as RMSE and correlation, are identical for the two top comparisons and the one in the second row and second column. Some favor the bottom panel because it is the only one where the prediction overlaps with the observation. (Right) Depiction showing the need for user-specific diagnostic evaluations. For one user, the prediction is very bad (missed the watershed) and for the other it is very good (flight-path change needed). Figures used by permission of Barbara G. Brown.

Of course, a single summary measure like RMSE can only provide a small amount of information. It cannot diagnose particular problems with a prediction. The right side of **Figure 1** illustrates the need not only for diagnostic information but also that the information may differ depending on the specific user needs.

Because of relevant methodologies developed in other areas, such as computer vision, image analysis, and spatial statistics, a large number of new methods were introduced over a fairly short amount of time; in part fostered by the spatial forecast verification Inter-Comparison Project (ICP; https://projects.ral.ucar.edu/icp). Early reviews of the various methods can be found in Gilleland et al. (2009), where those methods were found to generally be categorized as being in one of two main categories, filter v. displacement. Each of which can itself be categorized into two main categories. The filter methods fall into either smoothing or band-pass methods, and the displacement methods were categorized as either feature-based or deformation. In this poster, some highlights of the latter categories of displacement methods are given, in part to illustrate additional challenges with spatial prediction evaluation.

Notation

In what follows, let $Z(\boldsymbol{s})$ be the verification at grid point $\boldsymbol{s} = (x, y) \in \mathcal{D}$ where $\mathcal{D}$ represents the entire grid domain. Similarly, let $\hat{Z}(\boldsymbol{s})$ be the prediction at grid point $\boldsymbol{s}$. In this treatment, $Z(\boldsymbol{s})$ and $\hat{Z}(\boldsymbol{s})$ will be quantitative precipitation (mmh^{-1}) but most of the methods apply to a much wider range

of variables. The shortest distance between two grid points $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$ is denoted by $d(\boldsymbol{s}_1, \boldsymbol{s}_2)$ and is generally assumed to be Euclidean distance. The shortest distance between a grid point $\boldsymbol{s}$ and a set of grid points, say $\mathcal{A}$, is denoted by $d(\boldsymbol{s}, \mathcal{A})$, and the grid giving $d(\boldsymbol{s}, \mathcal{A})$ for every point $\boldsymbol{s} \in \mathcal{D}$ is called a distance map. There exist fast algorithms for calculating the distance map efficiently.

Deformation Methods

In principle, the issues about double penalties and over accumulation of errors can be directly modeled in the context of image warping (cf. Sampson and Guttorp 1992; Dryden and Mardia 1998; Hoffman et al. 1995; Nehrkorn et al. 2003; Gilleland et al. 2010b,c as labeled in G22; Gilleland 2013). The model says that the verification at grid point $\boldsymbol{s}$ is the prediction at possibly some other grid point $\boldsymbol{s}' = W(\boldsymbol{s})$ plus additional (intensity) error. Namely,

$$Z(\boldsymbol{s}) = \hat{Z}\big(W(\boldsymbol{s})\big) + \varepsilon(\boldsymbol{s}) \,,$$

where $\varepsilon(\boldsymbol{s})$ is the intensity error and $W(\boldsymbol{s}) = \big(W_x(\boldsymbol{s}), W_y(\boldsymbol{s})\big)$ so that the x- and y-coordinates are each modeled independently according to the warping function. There are many warping functions that can be chosen. For example, the pair-of-thin-plate-splines (POTPS) warping function for the x-coordinate (and similarly for the y-coordinate) is given by

$$W_x(\boldsymbol{s}) = a_{x,0} + a_{x,1} \cdot x + a_{x,2} \cdot y + \sum_{i=1}^{n_c} b_{x,i} \cdot U\left(d(\boldsymbol{p}_{x,i}, \boldsymbol{s})\right),$$

where the constants $a_{x,0}$, $a_{x,1}$, and $a_{x,2}$ are linear coefficients representing affine transformations, and the coefficients $b_{x,i}$ represent non-linear transformations governed by the thin-plate spline $U(\cdot)$ that is a function of the shortest distance between a set of control points, $\boldsymbol{p}_{x,i}$, and $\boldsymbol{s}$, where $U(r) = r^2 \log r$. The entire deformation is controlled by the subset of points, $\boldsymbol{p}_{x,i}$. The fewer the control points, the less complicated is the deformation, and the more efficient the algorithm. To obtain a better match, one would need more control points.

Figure 2: Example of image warping applied to photographic images. Left is Johan Lindström and right is Finn Lindgren. In the middle is Johan's image warped so that his facial features are more aligned with those of Finn's. The black dots show the control points used and note that the entire deformation is based solely on these points. Here, Finn is the 0-energy field and Johan is the 1-energy field. Image provided by Johan Lindström.

Figure 2 shows an example of image warping applied to photographic images and **Figure 3** shows an example of image warping applied to a contrived verification/prediction pair based on one of the cases from **Figure 1**. In this case, the prediction is warped to better match the verification so that

the verification is called the 0-energy field because it is not warped, and the prediction is the 1-enregy field because it is warped. The 1-energy field maintains the intensities but they are displaced spatially to be better aligned with the 0-energy intensities. Note that the 1-energy field in this example is a re-scaling of the verification field, but a warping that involves a large rotation also provides a good match; demonstrating one issue with the image-warping approach in that it can be difficult to find the "best" warp as there are numerous different ways to warp the points to provide a good match.

Once a good warped image is found, traditional verification measures, such as RMSE, can be applied without concern about double penalties. The percent reduction in RMSE (or other measure) before and after warping can also be useful information. One can also look at the coefficients $a_{x,0}$, $a_{x,1}$, and $a_{x,2}$, as well as, $a_{y,0}$, $a_{y,1}$, and $a_{y,2}$, in order to glean the amount of affine displacement in order to make statements like, on average the prediction is too far east (as in the example of **Figure 3**). A summary of the non-linear part in the case of POTPS warping naturally falls out as something called the bending energy.

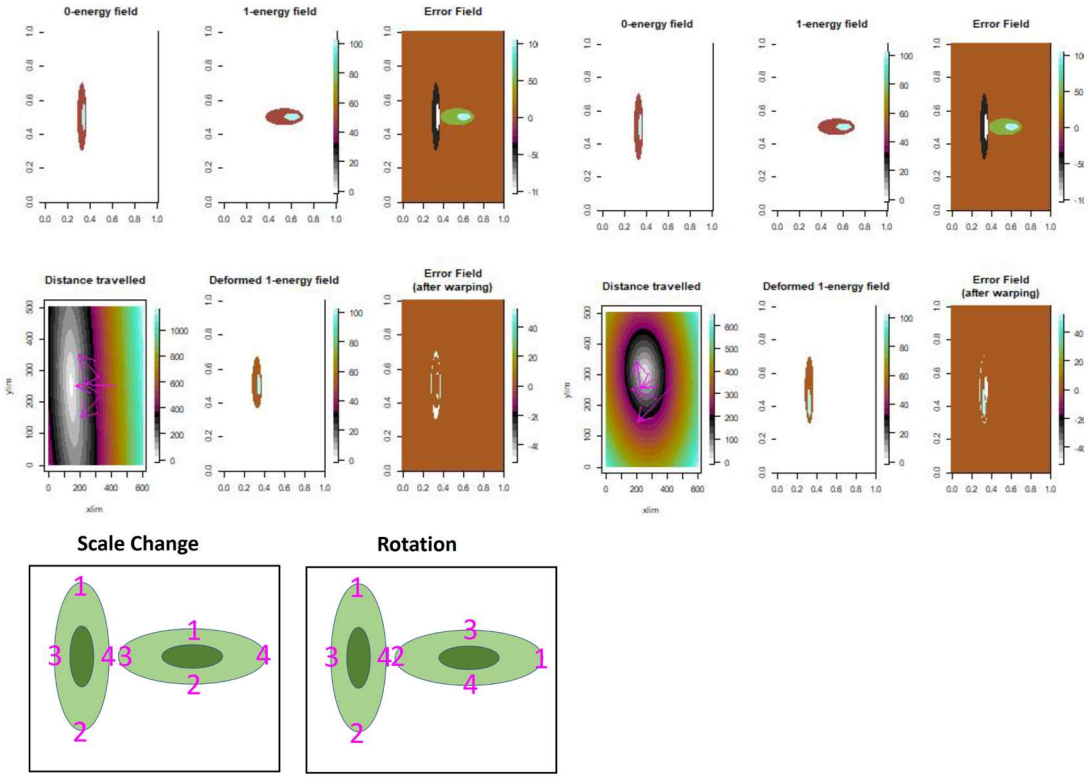

Figure 3: An example of image warping for a contrived prediction, labeled the 1-energy field, of the verification, labeled the 0-energy field. In this case, the prediction is a re-scaling of the verification, but it is approximately a rotation. There is not a unique image warp that provides a reasonable result. The left three columns depict the re-scaling warp and those on the right depict a rotation warp. The latter has a higher overall error, but still represents a good match to the verification. In the bottom, the two types of warp mappings are displayed where the control points are numbered so that 1 maps to 1, 2 to 2, etc. This example uses four control points that deform the entire image.

Assimilating all of this information, however, can be challenging unless a specific user has a preference for one type of error over another. For example, suppose the reduction in the RMSE is

very high for one prediction, say $\hat{Z}_1(\boldsymbol{s})$ and very low for another, say $\hat{Z}_2(\boldsymbol{s})$. However, the amount of warping for $\hat{Z}_2(\boldsymbol{s})$ is also considerably less than that for $\hat{Z}_1(\boldsymbol{s})$. Which prediction is better? Consider further the lack of uniqueness in finding a good fit, or the difficulty in finding a good fit for some cases, and the question becomes even murkier.

For example, **Figure 4** shows a contrived comparison from Gilleland et al. (2022) that demonstrates the difficulty in summarizing the results of image warping and other methods. Taking C6 to be the verification and C1 the prediction, C1 missed the two circles from C6 completely (no overlap) and has one circle that gives a false alarm. First, perhaps it is best in this case to count C6 circles as misses and the one from C1 as a false alarm (perhaps that is the true error). There is no way to tell, but certainly the image warping can be constrained to not warp too much.

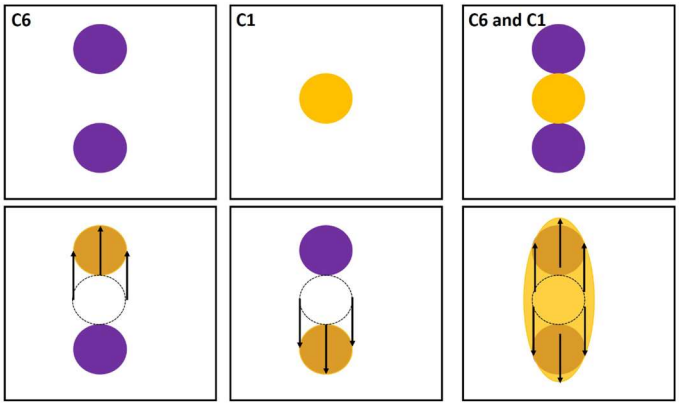

Figure 4: The contrived cases C6 v. C1 from Gilleland et al. (2020). Warping could be applied to move C1 up to match the top circle in C6 or down to match the lower circle (but maybe it is a true set of misses and false alarms!). Only one control point would be necessary for either of these deformations (though three are shown), but with more control points, the C1 circle could be stretch to overlap with both circles in C6. With even more (not shown) it could be stretched to overlap both circles, and squeezed to reduce the false alarms (taking C1 as the prediction).

If it is decided to warp C1, how should it be done? The figure illustrates three possibilities. Translate C1 north to match one circle, or south to match the other. Or, perhaps using more control points, stretch C1 to overlap with both. With more control points, it could also be squeezed until an almost perfect match is attained.

Feature-based applications

The main difference between feature-based and field-deformation approaches, such as image warping, is that feature-based methods are aimed at individual "features" within a field, whereas the latter morphs the entire field. In some cases, such as the contiguous rain area (CRA; Ebert and McBride 2000) method, apply field deformation to the individual features and then assess performance. In the case of CRA, the RMSE is decomposed into different sources, such as shift and volume, but is still applied to the entire field. Other approaches, such as the composite (Nachamkin et al. 2005) and structure-amplitude-location (Wernli et al. 2008) methods, use the individual feature elements to make distributional summaries about the prediction performance. Davis et al. (2009) merge (within each of the verification and prediction fields) and match the feature elements (across fields) in order to make various comparisons, including using these comparisons to employ traditional contingency-table measures.

The first challenge in a feature-based approach is to identify features in each field. Generally, this procedure needs to be performed automatically, which may result in features that are not meteorologically meaningful. While some utilize other meteorological information to identify features (such as in Wernli et al. 2008), the features are generally obtained by first identifying what are known in the computer-vision literature as connected components, or sometimes called isolated clusters (e.g., AghaKouchak et al. 2011). In most cases, the field is first smoothed, though this step is not necessary, it generally results in far fewer features that are very small in size; even when the field is smoothed, subsequent analyses may (or may not) be applied to the original underlying field using the smooth feature as a kind of mask. A threshold is then applied to the field in order to create a binary field, where typically values below the threshold are set to zero and everywhere else to one. Computer algorithms then determine the connected components, which include sets of one-valued grid points that are all connected to each other (cf. **Figure 5**). In some cases, as mentioned above, these individual blobs of points that are close together (but not touching) may subsequently be merged by some type of criterion analogous to a cluster analysis (e.g., Davis et al. 2006a,b as labeled in G22; Gilleland et al. 2008; Marzban and Sandgathe 2008).

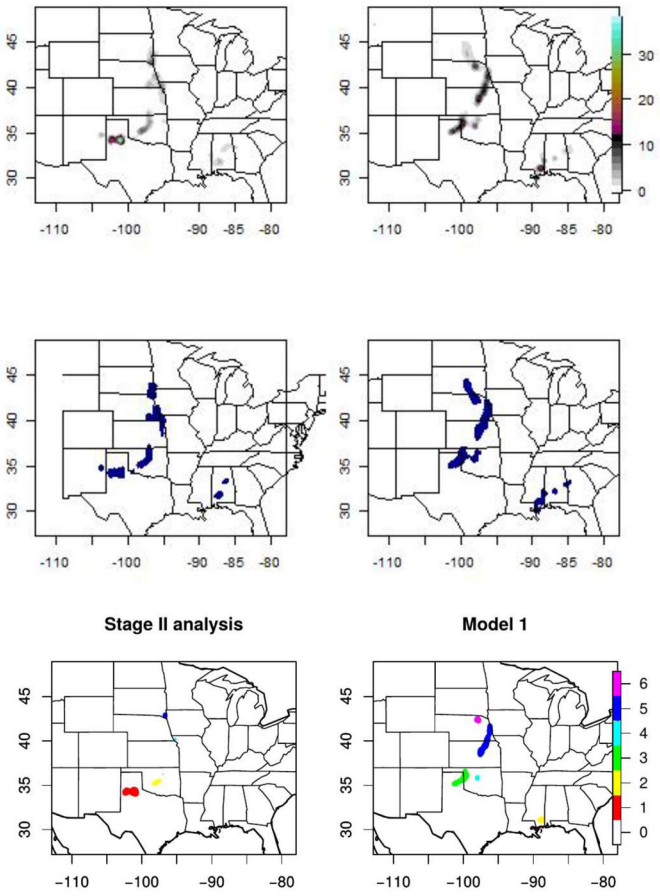

Figure 5: Two examples of features identified by first smoothing the respective fields, thresholding, and then identifying connected components. The top shows the original fields, followed by the binary fields. The colors in the bottom figure represent identified features. Here, the features have not been merged or matched. That is the next challenge!

Once the features are identified, one might compare summaries of the features across fields without directly comparing individual features, as in the distributional approaches mentioned above. In other cases, the next challenge is to determine the best matching (or not) of features across fields. Again, this process can be rather challenging, and often different methods result in different comparisons. In fact, many methods involve tunable parameters where results can be highly sensitive to their choices. Davis et al. 2006a,b employ a fuzzy logic procedure based on various feature properties, such as centroid, axis angle (if the feature is long and skinny), ratio of the convex hull to the area, etc. Gilleland et al. (2008) employ a dissimilarity measure, called Baddeley's Δ, described in further detail below in a type of cluster analysis, and the CRA method uses a simple proximity approach.

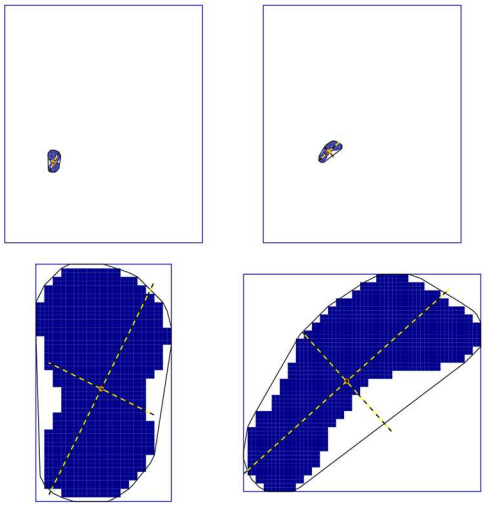

Figure 6: Example of two individual features and some properties that might be of interest. The top two panels show the features on the original field, and the bottom two are zoomed in to the features. The outline around the features is the convex hull. The ratio of this value to the area tells how concave or convex a shape is. For example, if the feature is a circle, then the ratio would be one because there would be no white space. The more C-shaped the feature is, the lower the ratio. The centroid of the featue is marked by the intersection of the features axis lines (the axis lines are not relevant if the feature is, for example, a circle). The angle of the major axis is the orientation angle.

The challenge is to summarize the overall performance based on the merged and matched features. In some cases, the summary measure used in the cluster-analysis process may double as the overall performance measure, such as the maximum interest value used by Davis et al. (2006a,b). The next section deals with dissimilarity measures that may be used in this approach, and discusses the many challenges associated with them.

Dissimilarity Measures

The task of comparing two features can be summarized as finding a measure $d(\mathcal{A}, \mathcal{B})$ for two sets of one-valued grid points, $\mathcal{A}$ and $\mathcal{B}$. Determining $d(\boldsymbol{s}_1, \boldsymbol{s}_2)$ and $d(\boldsymbol{s}, \mathcal{A})$ are relatively straightforward. While some options exist for $d(\boldsymbol{s}, \mathcal{A})$, it makes sense to use the shortest distance to the nearest point in the set $\mathcal{A}$ of one-value grid points, and is what is meant, herein, by $d(\boldsymbol{s}, \mathcal{A})$. For this section, it is assumed that the individual features are on an otherwise empty field, as some measures rely on the entire domain, $\mathcal{D}$. Note that these methods are also useful if applied to a binarized version of the original field, without identifying individual features.

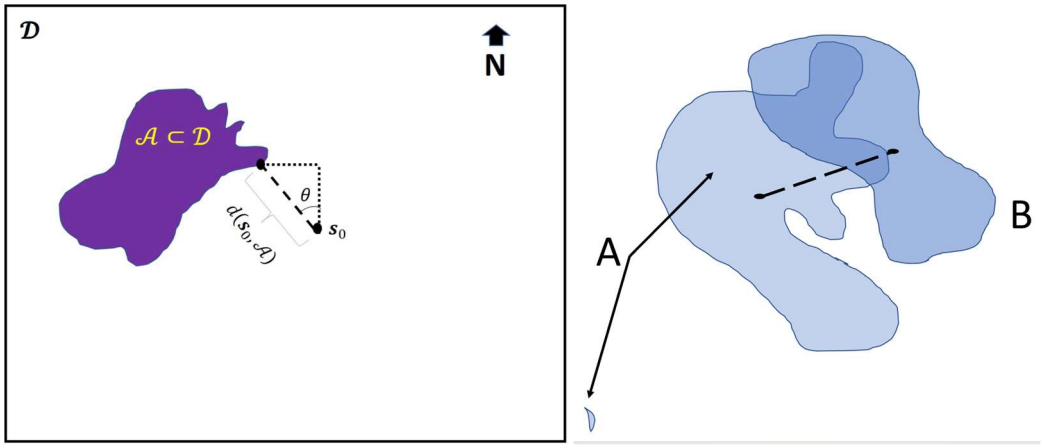

Figure 7: Left: Finding the distance between a point s_0 and a set of points, $\mathcal{A}$, is relatively straightforward. Usually, the shortest distance to the nearest point in the set is used, and is illustrated here. The angle, θ, is the bearing of the point s_0 to the set $\mathcal{A}$ relative to north. Right: comparing two sets of points, $d(\mathcal{A}, \mathcal{B})$, is a much more difficult task. The dashed line shows the centroid distance between the two sets, but many other possibilities exist.

One of the first dissimilarity measures to be used in spatial forecast verification, as it is often called, is the centroid distance (cf. Davis et al. 2006a,b). This measure gives the distance between two features' centroids. **Figure 8** illustrates the drawbacks of this measure, but it should be noted that there are many situations where this measure is appropriate, as well. All of the cases in the left column give a perfect match according to this measure despite some very large differences. On the other hand, the cases on the left column are less than perfect matches despite that human observers might suggest that they are better.

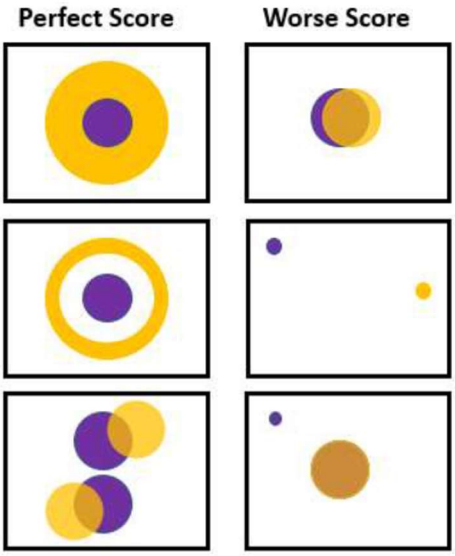

Figure 8: Examples illustrating the drawbacks of the centroid distance as a dissimilarity measure. The purple set of grid points is compared against the gold set.

It is relatively easy to find cases where the centroid distance might not be very useful, but every dissimilarity measure has its flaws, and is the subject of Gilleland et al. (2020). Nevertheless, some are very useful. When applied to an entire field without identifying features, they can make for very

simple, straightforward summaries of performance that while not comprehensive, are easy to digest and understand; and in most cases can be used operationally. They do rely on binary fields, so they are (in most cases) about the spatial alignment of intensities rather than about the intensities themselves. Therefore, they should be used in conjunction with other measures such as frequency bias, comparisons of mean intensities and their standard errors, etc.

The distance-map based dissimilarity measures (cf. Dubuisson and Jain 1994) are of particular importance because they can be calculated efficiently thanks to fast computational algorithms for finding distance maps. One of the earliest of these methods is the Hausdorff method, which is given by:

$$H(\mathcal{A},\mathcal{B}) = \max\left\{\max_{\boldsymbol{s}\in\mathcal{B}} d(\boldsymbol{s},\mathcal{A}), \max_{\boldsymbol{s}\in\mathcal{A}} d(\boldsymbol{s},\mathcal{B})\right\}.$$

Note that H gives the maximum of shortest distances of all of the points in each set to the nearest point in the other (i.e., look closely at the subscripts of the maxima inside the outer maximum). The usual criticism of this measure is that it can be highly sensitive to even a small change in one or both sets.

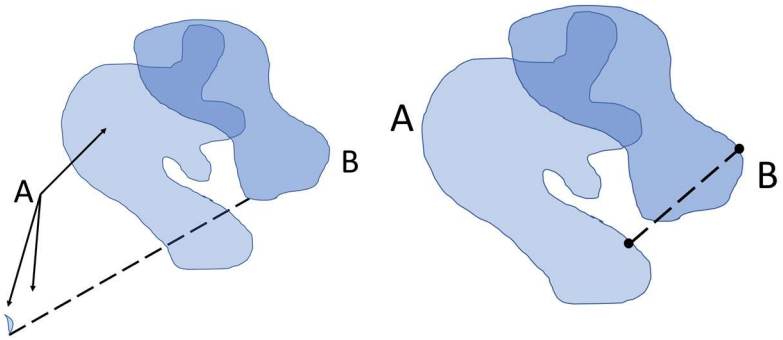

Figure 9: Illustration of the sensitivity of the Hausdorff metric to a small change in one or both of B. When the small disconnected part of set A is removed, the value of the Hausdorff metric changes drastically. The length of the dashed line in each figure is the Hausdorff metric.

One modification aimed at mitigating this effect is Baddeley's Δ (1992a,b as labeled in G22). This measure is given by

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

where $|\mathcal{D}|$ is the size of the domain, the summation is over every point in $\mathcal{D}$, $\omega(\cdot)$ is a special type of function but is usually chosen to be $\omega(t) = \min(t, c)$ for a chosen constant c called the cutoff, and p is a user-chosen parameter. Usually, $p = 2$ is chosen, which gives the Euclidean norm of the differences in distance maps (possibly after filtering out distances larger than a certain amount with the cutoff value). If $p = 1$, the result is simply the mean of the difference in distance maps, and in the limit as $p \to \infty$, the Hausdorff distance falls out. **Figure 10** illustrates how Δ is calculated from the distance maps. First, the individual distance maps for each field are found (top row), their difference is then found (bottom row left without a cutoff and right with a cutoff). Then, the L_p norm of the resulting difference is obtained to give Δ.

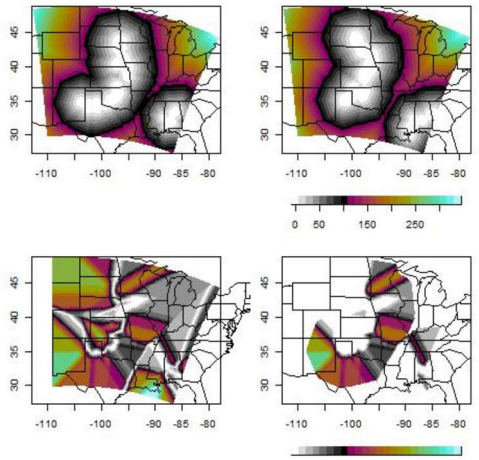

Figure 10: Illustration of the use of distance maps to calculate Baddeley's Δ. Top row shows the distance maps for two binary fields. Bottom left shows the difference between the two images in the top row. Bottom right is the same except after having applied a cutoff value.

A reason for the cutoff value is that because the norm is taken over the entire domain, the measure is susceptible to boundary effects. In particular, it may give a different result for two identical features separated by the same amount, depending on whether they are close to a boundary or not. The difference is usually small, but nevertheless, discouraging.

A particularly useful measure is the mean-error distance (MED). It is given by:

$$M(\mathcal{A},\mathcal{B}) = \frac{1}{|\mathcal{B}|}\sum_{\boldsymbol{s}\in\mathcal{B}} d(\boldsymbol{s},\mathcal{A}).$$

It can be thought of as a conditional measure of one set given the other. The main arguments against its use in other topic areas is its lack of symmetry in that $M(\mathcal{A},\mathcal{B}) \neq M(\mathcal{B},\mathcal{A})$. However, as demonstrated in Gilleland (2017), this asymmetry can be exploited to glean information about misses v. false alarms, thereby making the measure very valuable.

Gilleland et al. (2020) used various contrived cases, such as those in **Figure 4** to test the behavior of these and other spatial verification methods; those primarily aimed at spatial alignment errors, such as the dissimilarity measures. A summary of those findings is provided in **Table 1**. The testing can be grouped into several categories of behavior. The first considers pathological cases that arise rather often in meteorology. Namely, when nothing is predicted (e.g., no rain), as well as what happens when only one or a few grid points have values of one and everywhere else is zero. None of the existing methods at the time handled these cases, meaning that special handling is necessary for these situations.

Table 1: Summary of the findings in Gilleland et al. (2020). MED is mean-error distance and FoM is Pratt's Figure of Merit (not discussed here).

	Handles Pathological Cases well?	No positional effects?	Sensitive to frequency bias?	Useful for rare events?	Reward partial perfect match?	Correctly penalize despite partial perfect match?
Centroid distance	No	Yes	No	No	No	No
Baddeley's Δ	No	No	Yes	No	Yes	No
Hausdorff	No	Yes	No	Yes	No	No
MED	No	Yes	No**	Yes**	Yes**	Yes**
FoM	No	Yes	Yes	Unclear	No	Yes

The next category involves positional effects of features within the field. It was already mentioned earlier that Δ is sensitive to this issue. Frequency bias has to do with whether a prediction can be hedged by over predicting a phenomenon in order to achieve a better score. Only the centroid distance, Hausdorff distance and MED were immune (in the case of MED, it is immune only through the asymmetrical property when looking at both directions, $M(\mathcal{A},\mathcal{B})$ and $M(\mathcal{B},\mathcal{A})$).

The next test category is similar to the pathological cases. It is specifically intended for the case when a user is interested in high-intensity storms that are small in spatial extent, as many thunderstorms are. In these cases, a high threshold for creating the binary fields will yield very small features that are very important to predict well. The Hausdorff distance excels at this test, and the MED is generally useful in this setting, but other measures fail to produce here. Lastly is the idea of a partial perfect match, and whether the measure rewards or penalizes for such a scenario. Δ and MED are the only measures, here, that reward a prediction for this scenario.

In order to find a measure that performs better under some of the scenarios, such as the pathological cases, that were not handled well by existing methods, Gilleland (2021) proposed two new measures. The first is given by:

$$\sqrt{G}(\mathcal{A},\mathcal{B}) = \left(\sqrt{y_1 \cdot y_2}\right)^{1/3},$$

where $y_1 = n_{\mathcal{A}} + n_{\mathcal{B}} - 2n_{\mathcal{AB}}$ with $n_{\mathcal{A}}$ representing the number of grid points in the set $\mathcal{A}$, etc. (i.e., the symmetric difference) and $y_2 = M(\mathcal{A},\mathcal{B}) \cdot n_{\mathcal{B}} + M(\mathcal{B},\mathcal{A}) \cdot n_{\mathcal{A}}$, a weighted sum of the MED in both directions. The symmetric difference in the first term penalizes for lack of overlap between the two sets of grid points while the second term utilizes the utility of the MED and its information about both misses and false alarms. The square root in the above equation is added here in order to give it the same units (grid points) as the other measures discussed here, such as the Hausdorff distance.

The second measure proposed in Gilleland (2021) was created to provide an index from zero to one for ease of interpretability. It is similar to $\sqrt{G}$ in that they both make use of the same product $y_1 \cdot y_2$. This product is generally a very large number, so it needs to be scaled down. The former is scaled down by taking the cubed root (and here also the square root). The latter uses a user-chosen parameter β and is given by:

$$G_\beta(\mathcal{A},\mathcal{B}) = \max\left\{1 - \frac{y_1 \cdot y_2}{\beta}\right\}.$$

Table 2 shows the results of these new measures applied to the tests of Gilleland et al. (2020). Clearly, they perform well, noting that they penalize for a partially perfect match rather than reward for them. G is not useful for rare event predictions, whle G_β is useful provided β is chosen in such a way as to inform about them.

Table 2: The Gilleland et al. (2020) tests applied to the measures proposed in Gilleland (2021).

	Handles Pathological Cases well?	No positional effects?	Sensitive to frequency bias?	Useful for rare events?	Reward partial perfect match?	Correctly penalize despite partial perfect match?
G	Yes	Yes	Yes	No	No	Yes
G_β	Yes*	Yes	Yes	Yes*	No	Yes

References

For brevity on this poster, most references are not given explicitly here. The main reference for this poster is:

Gilleland, E., 2022. Comparing spatial fields with SpatialVx: Spatial forecast verification in R. doi: 10.5065/4px3-5a05, (abbreviated as G22 herein)

and references therein (most figures are borrowed directly from this unpublished manuscript). For a very long list of spatial evaluation methods, see: https://projects.ral.ucar.edu/icp/references.html.

Specific references cited here but are not in the above paper or webpage:

Dubuisson M.-P. and A. K. Jain, 1994. A modified Hausdorff distance for object matching. *Proceedings of 12th International Conference on Pattern Recognition*, Jerusalem, Israel, **1**, 566-568, doi: 10.1109/ICPR.1994.576361.

Ferro, C. A. T., 2017. Measuring forecast performance in the presence of observation error. *Q. J. R. Meteorol. Soc.*, **143**, 2665 – 2676, doi: 10.1002/qj.3115.

Mittermaier, M. P. 2018. How interpolation and resolution can affect verification scores: A study based on the Fractions Skill Score. *Meterologische Zeitschrift*, **28** (3), 181 – 192, doi: 10.1127/metz/2018/0890.