



Novel measures for summarizing high-resolution forecast performance

Eric Gilleland

Research Applications Laboratory, National Center for Atmospheric Research, Boulder, Colorado, USA

Correspondence: Eric Gilleland (ericg@ucar.edu)

Received: 19 July 2020 – Revised: 31 December 2020 – Accepted: 13 January 2021 – Published: 16 February 2021

Abstract. In ascertaining the performance of a high-resolution gridded forecast against an analysis, called the verification set, on the same grid, care must be taken to account for the over-accumulation of small-scale errors and double penalties. It is also useful to consider both location errors and intensity errors. In the last 2 decades, many new methods have been proposed for analyzing these kinds of verification sets. Many of these new methods involve fairly complicated strategies that do not naturally summarize forecast performance succinctly. This paper presents two new spatial-alignment performance measures, G and G_β . The former is applied without any requirement for user decisions, while the latter has a single user-chosen parameter, β , that takes on a value from zero to one, where one corresponds to a perfect match and zero corresponds to the user's notion of a worst case. Unlike any previously proposed distance-based measure, both handle the often-encountered case in which all values in one or both of the verification set are zero. Moreover, its value is consistent if only a few grid points are nonzero.

1 Introduction

Gaging the performance of a forecast from a high-resolution verification set (i.e., the pair of forecast and observation fields) is challenging because of the potential for small-scale errors to be over-accumulated in summary measures and the well-known double penalty issue whereby a forecast is penalized for both a miss and a false alarm in the face of a single displacement error (Mass et al., 2002). Many advances in forecast verification techniques, known as spatial verification methods, have occurred in the last few decades (e.g., Ebert, 2008; Rossa et al., 2008; Gilleland et al., 2009; Brown et al., 2011; Weniger et al., 2016; Wikle et al., 2019, Sect. 6.3.5); and independently in the hydrology literature (e.g. Wealands et al., 2005; Koch et al., 2015, 2016, 2018). Many of the methods provide a wealth of diagnostic information about specific ways in which a forecast performs well or poorly; a notable example is the Method for Object-based Diagnostic Evaluation (MODE, Davis et al., 2006a, b, 2009). Nevertheless, a single summary measure is often desired and is often required within more complicated techniques such as MODE.

This paper introduces new summary measures that rectify several drawbacks of other summary measures.

1.1 Background on distance-based summary measures

Several summary measures are available for measuring errors in the location, size, and sometimes shape of some type of weather event area, usually defined as being wherever a meteorological parameter exceeds (or meets and exceeds) a specified threshold (e.g., Brunet et al., 2018; Gilleland, 2011, 2017; Gilleland et al., 2020). Most of these measures were introduced long before spatial verification began and are sometimes referred to as binary image measures as they were generally applied to photographic images (e.g., Peli and Malah, 1982; Baddeley, 1992a, b; Pratt, 2007). In practice, a binary field is usually created from a meteorological field, $Z(\mathbf{s})$, where $\mathbf{s} = (x, y) \in \mathcal{D}$ is a grid point location within the domain, \mathcal{D} , of the field by setting values above a threshold, u , to one and the rest to zero. In this case, the binary field $I_{Z>u}(\mathbf{s}) = 1$, where $Z(\mathbf{s}) > u$ and $I_{Z>u}(\mathbf{s}) = 0$ otherwise. Of course, other rules could be employed, and often summary measures of this kind are applied for several different threshold values.

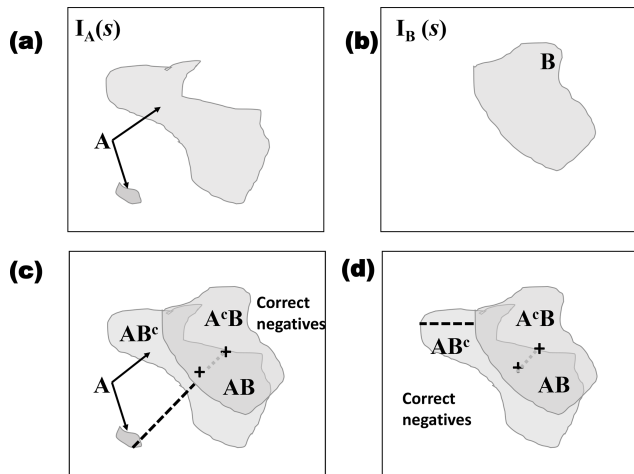


Figure 1. Illustration of two possible binary fields, namely (a) I_A and (b) I_B . Labels A and B indicate the areas (sets) of one-valued grid points, s , in each field, respectively. Set A contains two disconnected areas (i.e., connected components), where B has one connected component. Panel (c) shows the same fields as in panels (a) and (b) overlaid on top of each other, along with a black dashed line indicating the Hausdorff distance and a gray dotted line indicating the centroid distance (+ signs show the centroid for each field). Here, AB indicates the area where A and B overlap (i.e., $I_A(s) = I_B(s) = 1$), AB^c where $I_A(s) = 1$ and $I_B(s) = 0$ (the superscript c stands for the set complement), and A^cB where $I_A(s) = 0$ and $I_B(s) = 1$. White areas are correct negatives (i.e., where I_A and I_B are both zero valued). Panel (d) is the same as panel (c) but where the smaller connected component from A is removed.

When considering a single location, say s_1 , it is easy to derive a sensible summary metric between s_1 and another location, say s_2 , or even a set of several locations, i.e., $s \in A$. For example, the length of the shortest possible distance between s_1 and the set A , denoted by $d(s_1, A)$, could be used. However, summarizing the similarities or differences between two sets of locations is a very challenging task.

To simplify the discussion, it is useful to introduce some notation, which is depicted in Fig. 1. The figure displays two binary fields, $I_A(s)$ and $I_B(s)$. The set of locations s , where $I_A(s) = 1$, is labeled as A and similarly for B . Define AB to be the set of $s \in \mathcal{D}$, where $I_A(s) = I_B(s) = 1$, AB^c is the intersection of the sets A and the complement of set B , or where $I_A(s) = 1$ and $I_B(s) = 0$, A^cB where $I_A(s) = 0$ and $I_B(s) = 1$, and, finally, A^cB^c where $I_A(s) = I_B(s) = 0$ (correct negatives). Let n_A (n_B) denote the number of grid points, where $I_A(s) = 1$ ($I_B(s) = 1$), and let n_{AB} be the number of grid points where both $I_A(s) = 1$ and $I_B(s) = 1$ (i.e., the area of overlap between A and B). Finally, let N be the total number of grid points in \mathcal{D} , i.e., the size of the domain.

Table 1 gives the equations for the distance-based measures described subsequently.

One of the most well-known binary image measures is the Hausdorff distance metric (Baddeley, 1992a, b). This met-

ric is very useful, generally, but has some drawbacks. For example, it has a high sensitivity to small changes in one or both fields. Figure 1 illustrates this metric for a comparison between $I_A(s)$ and $I_B(s)$, when A has two connected components (see Fig. 1 caption) and when the smaller connected component is removed. Although the two versions of A are very similar, $H(A, B)$ is very different. On the other hand, this metric can be particularly useful in the forecast verification domain when the interest is in verifying small-scale high-intensity events that can be isolated after applying a high threshold specifically because of its high sensitivity to noise.

Several modifications to the Hausdorff method have been proposed, and perhaps the most well-known is the partial Hausdorff distance (PHD_k), which replaces the innermost maxima with the k th largest order statistic, mean error distance (MED), and Baddeley's Δ metric (Peli and Malah, 1982; Baddeley, 1992a, b). One summary measure, proposed by Venugopal et al. (2005), which was specifically designed for spatial forecast verification, utilizes the PHD_k but also incorporates an intensity component, and is known as the forecast quality index (FQI). Because it involves intensity and location in its evaluation of forecast performance, it is omitted from Table 1 and is instead given here by the following:

$$\text{FQI}(A, B) = \frac{\text{PHD}_k(A, B)}{\bar{\psi}} \left/ \frac{2\mu_A\mu_B}{\mu_A^2 + \mu_B^2} \cdot \frac{2\sigma_A\sigma_B}{\sigma_A^2 + \sigma_B^2} \right., \quad (1)$$

where $\bar{\psi}$ is the average of $\text{PHD}_k(A, C_i)$ over $i = 1, \dots, m$ surrogate fields, C_i (randomly generated fields with the same probability density and spatial correlation structure as A) and μ_A (μ_B) are the mean intensities over nonzero grid points from the original fields from which A and B are derived, and σ_A (σ_B) are the analogous standard deviations. As applied to a verification set, the surrogate fields are made with reference to the observation field so, using the notation from Eq. (1), the observed/analysis field is always A . The range of values FQI takes is $[0, \infty)$, with zero representing a perfect match and increasing values implying decreasing forecast quality.

One alternative summary measure to compare two sets A and B is known as Pratt's figure of merit (FoM; Peli and Malah, 1982; Baddeley, 1992a, b; Pratt, 2007, Eq. 15.5-1). FoM has one user-chosen parameter, α , which is a scaling constant that is typically set to one-ninth, corresponding to when $d(\cdot, \cdot)$ is normalized so that the smallest nonzero distance between neighboring s is one, and $0 < \text{FoM}(A, B) \leq 1$ with $\text{FoM}(A, B) = 1$ only if $A = B$. Like MED, FoM is not symmetric so that, generally, $\text{FoM}(A, B) \neq \text{FoM}(B, A)$. Figure 2 shows an example similar to one found by Peli and Malah (1982, and also shown in Baddeley, 1992b) that demonstrates how FoM is not sensitive to the pattern of displacement errors. In the example, the set A overlaps with B_2 but not B_1 , although it is more similar in shape to B_1 . Nevertheless, $\text{FoM}(A, B_1) \approx \text{FoM}(A, B_2)$ and $\text{FoM}(B_1, A) = \text{FoM}(B_2, A)$.

Table 1. Equations for distance-based measures discussed here. Let $\mathbf{s} = (x, y) \in \mathcal{D}$ represent a grid point (coordinate) in the domain \mathcal{D} , and let N be the size of the domain with $A, B \subset \mathcal{D}$ representing sets of grid points for which the corresponding value is one (in the binary field). Then let $d(\mathbf{s}, A)$ be the shortest distance from \mathbf{s} to A and, similarly, for $d(\mathbf{s}, B)$. If a field is empty of one-valued grid points, define $d(\mathbf{s}, A) = D$ for some large value D , such as N . Let n_A and n_B represent the number of grid points in the sets A and B , respectively. Furthermore, let $I_A(\mathbf{s}) = 1$, if $\mathbf{s} \in A$, or zero otherwise, which is similar for $I_B(\mathbf{s})$. ω is any continuous function on $[0, \infty]$ that is concave, and p is a user-chosen parameter. Finally, α in figure of merit (FoM) is a user-selectable parameter (see text).

Measure name	Measure's equation
Hausdorff distance	$H(A, B) = \max \left\{ \max_{\mathbf{s} \in B} [d(\mathbf{s}, A)], \max_{\mathbf{s} \in A} [d(\mathbf{s}, B)] \right\}$
Partial Hausdorff distance	$\text{PHD}_k(A, B) = \max \left\{ \begin{matrix} k\text{th largest}[d(\mathbf{s}, A)], \\ k\text{th largest}[d(\mathbf{s}, B)] \end{matrix} \right\}_{\mathbf{s} \in A}$
Baddeley's Δ	$\Delta(A, B) = \left[\frac{1}{N} \sum_{\mathbf{s} \in \mathcal{D}} \{ \omega(d(\mathbf{s}, A)) - \omega(d(\mathbf{s}, B)) \}^p \right]^{1/p}$
Mean error distance	$\text{MED}(A, B) = \frac{1}{n_B} \sum_{\mathbf{s} \in B} d(\mathbf{s}, A)$
Pratt's FoM	$\text{FoM}(A, B) = \frac{1}{\max\{n_A, n_B\}} \sum_{\mathbf{s} \in B} \frac{1}{1 + \alpha d^2(\mathbf{s}, A)}$

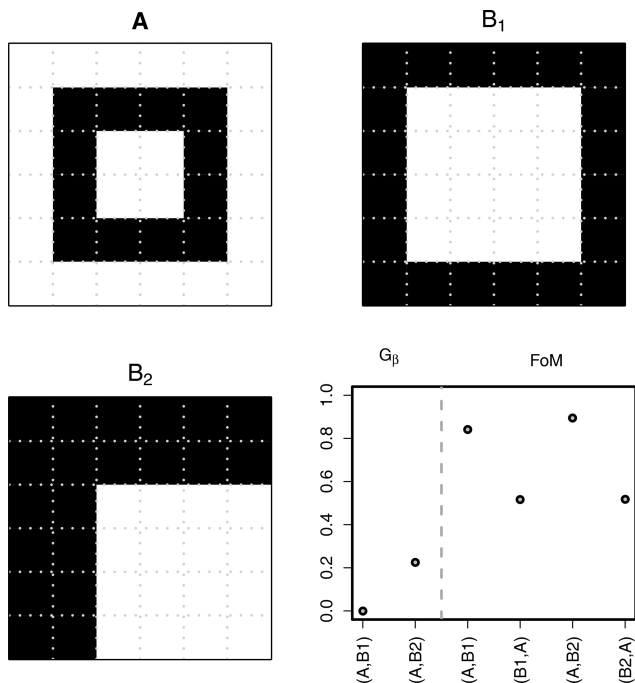


Figure 2. Similar to an example from Peli and Malah (1982). Suppose A represents the truth, and B_1 and B_2 are competing predictions for A . $G_\beta(A, B) = G_\beta(B, A)$ and $H_\eta(A, B) = H_\eta(B, A)$, but FoM does not share this symmetry property. Note that, when using $\alpha = 1/9$, $\text{FoM}(B_1, A) = \text{FoM}(B_2, A) \approx 0.5$ and $\text{FoM}(A, B_1) \approx \text{FoM}(A, B_2) \approx 0.9$, indicating an insensitivity to shape. G_β , using $\beta = 648 = 6^4/2 = N^2/2$, yields different values when comparing A with B_1 as opposed to A with B_2 . Therefore, B_2 is favored because of the amount of overlap with A , whereas B_1 does not overlap with A at all. Nevertheless, G_β indicates a poor match for both B_1 and B_2 when compared with A at this level of β .

The same properties pointed out above also apply to MED (Baddeley, 1992b), except that MED is not bounded above by one. Additionally, MED is relatively sensitive to background noise, which is usually not desired unless the background noise is important (see an example in Gilleland, 2017, where the noise could be a severe storm).

Distance maps are useful for both describing and computing most of the distance-based methods. A distance map, $d(\cdot, A)$, gives $d(\mathbf{s}, A)$ for every $\mathbf{s} \in \mathcal{D}$ and can be calculated efficiently because of various distance transform algorithms (see Meijster et al., 2000; Brunet and Sills, 2017). Most, if not all, distance-based methods fail when one or both of the fields is empty (Gilleland et al., 2020). Some invoke a special rule so that they will be defined, but then their values tend to differ considerably if $I_A(\mathbf{s}_0) = 1$ at a single grid point \mathbf{s}_0 . For example, $d(\mathbf{s}, A) = \infty$ when A is empty, or in practice, $d(\mathbf{s}, A) = D$, where D is some large number such as N . If, instead, $I_A(\mathbf{s}_0) = 1$ at a single point \mathbf{s}_0 and $I(\mathbf{s}) = 0$ at all other points $\mathbf{s} \neq \mathbf{s}_0$, then $d(\cdot, A)$ will be much smaller than D at many locations, typically resulting in drastically different summary measure values even though the two fields are practically identical. Moreover, these measures may differ greatly, depending on where \mathbf{s}_0 is located within \mathcal{D} .

Some of the measures, such as MED, are susceptible to hedging because they are not very sensitive to frequency bias (Gilleland, 2017). Finally, most of these summary measures have zero as the value for a perfect match (e.g., if a field is compared against itself) but can be less straightforward to interpret in terms of what constitutes a poor forecast.

This paper presents the following two new measures: one that is free of any user-chosen parameters and another with a single user-chosen parameter that falls on a scale from zero to one, where one corresponds to a perfect match and zero to the user's idea of poor. The metrics incorporate not just the spatial alignment errors but also the frequency bias. Two

additional summary measures are proposed that are based on the second location-only one but also incorporate a measure of intensity error that allows for the forecast to not be a perfect match without resorting to any computationally intensive or difficult-to-implement algorithms.

The new measures are computationally efficient, are not affected if the domain size is increased, cannot be hedged by increasing or decreasing the forecast frequency, and provide a single, easy-to-interpret summary of forecast performance that accounts for the amount of overlap between the forecast and observation and a graduated measure of the severity of any lack of overlap. Moreover, unlike all other distance-based summary measures, both of the new proposed metrics handle pathological cases naturally (see Sect. 4.1).

The equitable threat score (ETS) is also applied for some of the cases as a baseline comparison. The ETS is a hit-heavy traditional verification measure, so it does not account for the issues associated with high-resolution forecasts. Using the nomenclature of Fig. 1, ETS is defined by $(n_{AB} - r)/(n_{AB} + n_{AB^c} + n_{A^cB} - r)$, where $r = (n_{AB} + n_{AB^c})(n_{AB} + n_{A^cB})/N$. ETS can be thought of as a strict, or even automaton, measure because it requires the overlap n_{AB} to be very high in order to pass muster.

1.2 The intensity conundrum

Because the main issue with traditional grid-point-to-grid-point verification is largely centered on the over-accumulation of errors in the face of spatial displacement errors, the most natural approach is to deform one or both of the forecast and observation fields so that they align better in space. Examples of this approach in the literature include a pyramidal matching algorithm (Keil and Craig, 2007, 2009), based on an approach that had been developed for the detection and tracking of cloud features in satellite imagery (Zinner et al., 2008), optical flow (Marzban and Sandgathe, 2010), and image warping (Gilleland et al., 2010; Gilleland, 2013). A feature-based method such as MODE can also be considered as a deformation approach in that individual features may be matched together and assumed to be aligned. In the case of MODE, the distributional intensity information is performed on these individual matched feature comparisons, so if it accumulated over features, then the result would be analogous to a deformation type of method. Beyond these approaches, only distributional information about intensities can be obtained.

Of the deformation methods proposed, the image-warping procedure is the only one that has an elegant statistical model that allows for a natural method to ascribe uncertainty information. Deformations that yield optimal, or near-optimal, spatial alignments are not unique, however. Moreover, while non-sensible and non-physical deformations can be penalized in the optimization procedure, there is no guarantee that they will not result; for example, it is possible for the entire field to be compressed into a ball or pushed out of the

domain. The biggest challenge, however, may be simply in determining how best to weight the amount of deformation required versus the resulting error intensities.

The vast majority of spatial verification methods proposed are applied on binary fields usually obtained by setting all values below the threshold to zero and those above to one. Examples include the intensity-scale approach (Casati et al., 2004) and MODE, which also includes a distributional component for the intensities and all of the distance-based approaches discussed in Sect. 1.1. In this work, a distributional measure of intensity performance is proffered that is combined with the new spatial alignment measure, G_β , described in Sect. 2. It is not a fully satisfactory measure, as will be seen, but it can be useful in some situations. Moreover, it represents a best-case deformation summary – if there were no considerations about how much deformation should be allowed.

2 The new measures

Having a summary measure that penalizes misses and false alarms in a consistent and sensible manner is desired. It should be sensitive to small changes in one or both fields but not overly so. If the (binary) forecast contains a set, A , of one-valued grid points and the observation set, B , then the value of the measure should not change if A and B are found in a different part of the domain, or are rotated, as long as they are the same relative to each other (see the first few circle cases from Gilleland et al., 2020 that are also analyzed subsequently here).

Therefore, the new measures involve a term, denoted by y_1 , that measures the lack of overlap by way of the size of the symmetric difference. Another term, y_2 , provides a measure of the average distance of A to B and B to A , weighting each by the size of the two sets in order to mitigate the impact of this term in the event that a set is small.

Let $y = y_1 y_2$ where $y_1 = n_{A\Delta B} = n_{AB^c \cup A^cB} = n_A + n_B - 2n_{AB}$ and $y_2 = \sum_{s \in B} d(s, A) + \sum_{s \in A} d(s, B) = \text{MED}(A, B) \cdot n_B + \text{MED}(B, A) \cdot n_A$, with MED as in Table 1. Both new measures are functions of the product $y_1 y_2$. This product multiplies an area and a distance with units of grid points squared and grid points, respectively. The first measure, therefore, is a cubed root of this product so that the resulting value has the same units as most of the other distance-based measures (i.e., grid points) described in Sect. 1.1. The second is a unitless quantity that is rescaled so that it falls between zero and one, with zero being a bad score and one being a perfect score. The new performance measures applied to compare two binary fields, $I_A(s)$ and $I_B(s)$ (as described in Fig. 1), are given by the following:

$$G(A, B) = y^{1/3}, \quad (2)$$

and

$$G_{\beta}(A, B) = \max \left\{ 1 - \frac{y}{\beta}, 0 \right\}, \quad (3)$$

where $\beta > 0$ is a user-chosen parameter. Equation (3) represents a fuzzy-logic scaling of y that falls between zero (exceptionally poor forecast) and one (perfect forecast) in the same vein as an interest value in the MODE approach (Davis et al., 2006a, b, 2009). Indeed, G_{β} applied to individual features would be a good alternative to the centroid distance that is often used as part of MODE.

The term y_1 , which, again, is a penalty for lack of overlap between A and B , has a maximum possible value of N , which happens when $n_A + n_B = N$ and $n_{AB} = 0$. In the case of a perfect match (i.e., $A = B$), $y_1 = 0$. It is the size of the symmetric difference, $A \triangle B$, between the two sets, and its units, therefore, are numbers of grid points squared.

$MED(A, B)$ is a measure of average distance from B to A in terms of the distance $d(\cdot, \cdot)$ described previously. In general, $MED(A, B) \neq MED(B, A)$. Because MED is not sensitive to the occurrence rate, each MED in y_2 is multiplied by its respective size, n_A or n_B . This compensation ensures that a large MED will not overly dominate the resulting measure when sets A or B are small. The units, here, for y_2 are grid points so that the units for the product $y_1 \cdot y_2$ are numbers of grid points cubed. Therefore, G from Eq. (2) has numbers of grid points as its units.

For a perfect match, $MED(A, B) = MED(B, A) = 0$ and, subsequently, $y_2 = 0$. The maximum value for MED is D (where D is a large value as described in Sect. 1.1), which would occur when $n_A = N$ and $n_B = 0$ for $MED(B, A)$ or, similarly, when $n_A = 0$ and $n_B = N$ for $MED(A, B)$. Because $MED(A, B)$ is undefined if $n_B = 0$, the term $MED(A, B) \cdot n_B$ is defined to be zero in this situation and, similarly, for $MED(B, A) \cdot n_A$ when $n_A = 0$. Tempering the MED in this way ensures that small amounts of noise will not overly affect the results, unlike untempered MED. Gilleland (2017) argued that MED's sensitivity to noise could be useful when smaller-scale spatial events are important (e.g., storm activity), and it should be noted that G_{β} can be useful in this capacity through careful choice of the β argument as described below. The Hausdorff distance, as mentioned previously, is particularly well suited to such a verification.

The product $y = y_1 y_2$ provides a measure of the size of the area given by AB^c and $A^c B$ magnified by a measure of the distance between A and B . The function in Eq. (3) decreases linearly in y from one (perfect match) to zero (very poor forecast). The maximum possible value of $y_1 y_2 = N^2 \cdot D$, which provides some motivation for choosing β and is to be chosen with N and the base rate in mind. Experimentation has found that $\beta = N^2/2$ is a sensible choice when the base rate is large but small relative to N (see Fig. 2 and the cases in Gilleland et al., 2020), but if penalizing location errors more greatly is desired, then smaller values for β can be used. For example, Fig. 3a, c, and e demonstrate both G_{β} , with

three different choices of β , and FoM, with three different choices of α , applied to a comparison between two identical circles, where one circle is translated horizontally by increasing amounts. If β is small ($\beta = N\sqrt{N}$), then G_{β} decreases rapidly with increasing translation errors and equals zero once it is separated by 20 grid squares, or the point at which there is no longer an overlap between the two circles. When $\beta = 200^4/2$ or 200^4 , G_{β} decreases more slowly, giving more credit to forecasts with larger displacement errors. When $\beta = 200^6$, G_{β} decreases but remains very close to one for all circles. The FoM applied with three different values of α demonstrates that this scaling constant does not affect the result very much, and the measure decreases rapidly with increasing translation errors.

Figure 3b, d, and f show a similar analysis for G_{β} and FoM but for different types of bias. For the highest values of β applied, the smaller radii are not small enough for G_{β} to drop much below the perfect value, though they do drop. In this case, the lowest choice of β perhaps supplies the optimal behavior. FoM behaves well, and in this case, the asymmetry of the measure is apparent as the gray and black lines reflect a similar pattern but do not give the same result for each comparison.

Because G and G_{β} are designed, in part, to seamlessly handle the pathological cases, they will typically yield good scores at high thresholds because the areas that exceed those thresholds are typically very small in both fields. If such high intensities over such small areas are important (e.g., severe thunderstorms), then a lower choice of β can be used. Figure 4, for example, shows a precipitation forecast only for values exceeding 40 mm h^{-1} . The observation is not shown but, to the naked eye, appears identical (i.e., empty). However, the observation has no precipitation exceeding 40 mm h^{-1} anywhere near the enlarged region of southern Louisiana and Alabama displayed, where, clearly, the forecast has rather intense false alarms. Over the entire domain, the observation field has three connected components with areas of 18, 132, and 11 grid points, whereas the forecast has only two such areas of five and 37 grid points (both shown in Fig. 4). That is, there are also areas with intense precipitation that were missed by the forecast. Taking, as an exceptionally large area for this activity, about 200 grid squares and supposing that an average distance of 20 grid squares ($\approx 80 \text{ km}$) is too large of a translation error, then $\beta = 200 \cdot 20 = 4000$ is an appropriate choice for β . Indeed, $G_{\beta} = 0$ for this choice with this verification set and a threshold of 40 mm h^{-1} .

The flip side of choosing β is that a particular user might only be interested in egregious errors so that a higher choice of β will be appropriate as the score will be more forgiving. In general, when reporting results for G_{β} , it is important to specify the choice of β and the reasoning behind it. When possible, physical reasoning should be employed.

A modification is presented for incorporating intensity information into G_{β} , denoted $G_{\beta, \text{IL}}$, where IL stands for intensity and location. G_{β} can be thought of as a type of sum-

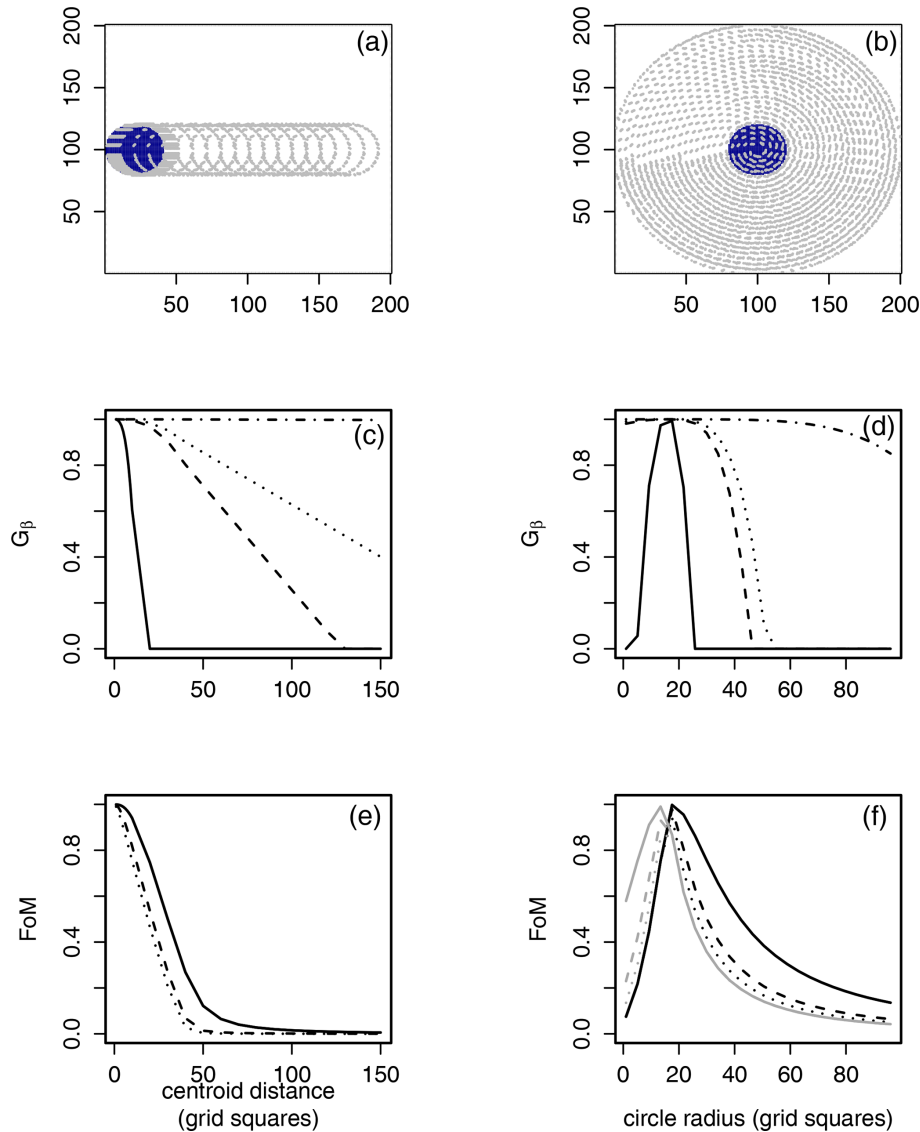


Figure 3. G_β (c, d) and FoM (e, f) applied to pairs (a) consisting of the leftmost blue solid circle and identical, horizontally translated circles (outlined for clarity) and (b) the blue solid circle with circles centered on the same location but with varying radii. Each set of comparisons is on a 200×200 domain. The blue solid circles have radii of 20 grid points. G_β lines are $\beta = 200^3 = N\sqrt{N}$ (solid), $\beta = 200^4/2 = N^2/2$ (dashed), $\beta = 200^4 = N^2$ (dotted), and $\beta = 200^6 = N^3$ (dot dashed). FoM lines are $\alpha = 1/100$ (solid), $\alpha = 1/9$ (dashed), and $\alpha = 1/2$ (dotted). $\text{FoM}(A, B) \neq \text{FoM}(B, A)$, generally, but is symmetric for the left column comparisons. Black lines for FoM in panel (f) indicate $\text{FoM}(A, B)$, and gray lines indicate $\text{FoM}(B, A)$.

mary of the amount of deformation that would be required if one could completely transmogrify the field so that all intensity values were matched with the nearest such value in the other field. Therefore, any type of distributional measure of the similarities in intensities completes the picture for this paradigm. Of course, a calibrated forecast should already match the distribution of the observed field in this way so that the additional measure may be less important than the alignment errors. It is given by the following:

$$G_{\beta, \text{IL}}(A, B) = \omega \cdot G_\beta(A, B) + (1 - \omega) \cdot \theta(A, B), \quad (4)$$

where $\theta(A, B)$ is $\max\{\rho(A, B), 0\}$, with ρ the linear correlation coefficient between sorted values of the intensities within the set A and those in the set B , and ω is a user-chosen weight dictating the importance of spatial alignment errors versus intensity errors, which are defined distributionally.

If the distribution of values in A is the same as that in B , then the relationship between their sorted values should be a straight line, and correlation measures the strength of this linear relationship. Because the sorted values are both monotonically increasing, $\theta(A, B)$ will generally be high. Nevertheless, when the distributions are different, their value will de-

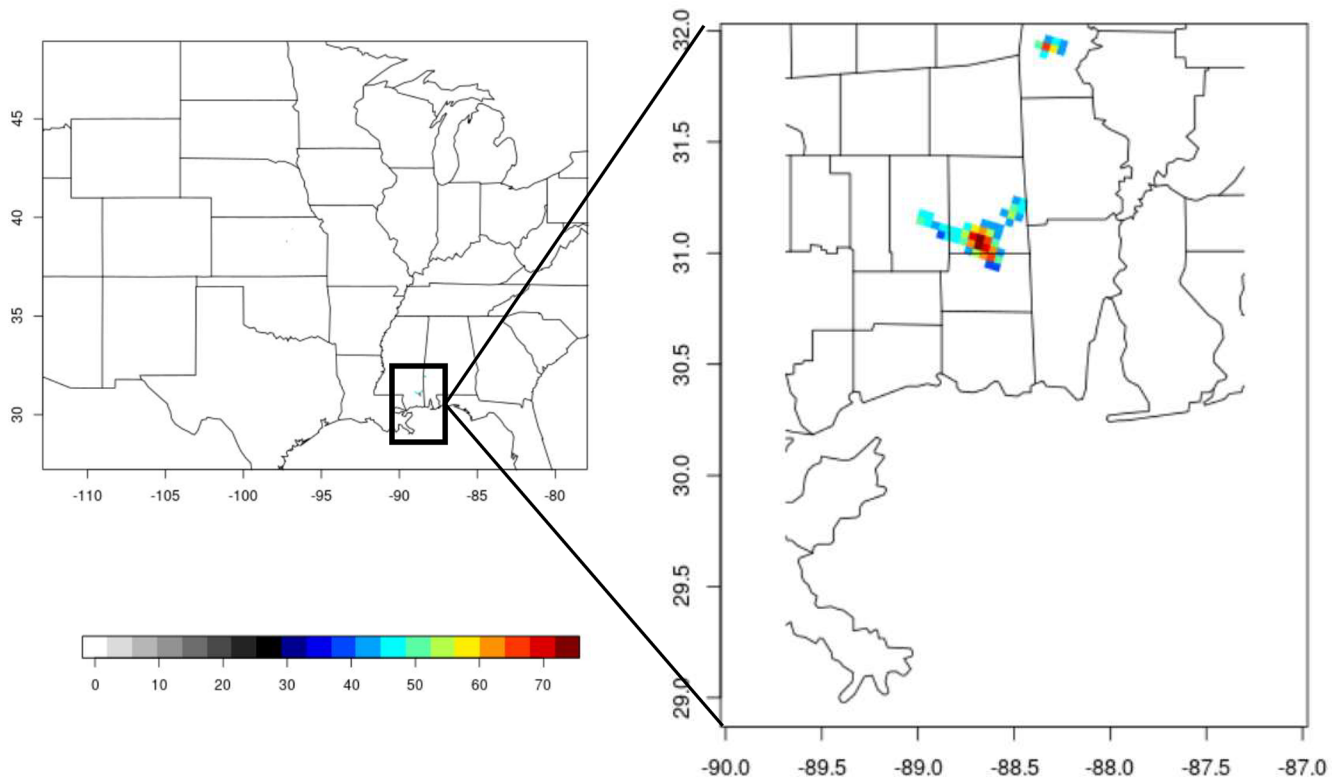


Figure 4. An example of 24 h accumulated precipitation forecast displaying only those values that exceed 40 mm h^{-1} . The forecast is the Inter-Comparison Project (ICP) wrf4ncar case for 1 June 2005 (Ahijevych et al., 2009).

crease. In general, ω should be higher than one-half in order to weigh the more important spatial alignment errors more heavily.

Equation (4) assesses intensity performance distributionally through the sorted intensity values within the sets A and B . If $n_A < n_B$ or $n_B < n_A$, then the larger set of sorted values is linearly interpolated to have the same size as the smaller set. A difficulty arises when $n_A = 0$ or $n_B = 0$. For $\theta(A, B)$, it is unclear what to do for such cases. If $n_A = n_B = 0$, then $\theta(A, B)$ could be taken to be one, and if the non-empty set is small, then it could be taken to be something near to one, but the solution to this problem will be application dependent and is not discussed further here.

The motivation for using the sorted values in θ is that y_1 measures lack of overlap, and y_2 provides a sense of the closeness of the lack of overlap, so θ can be thought of as the intensity error remaining if they could be realigned spatially so that they match up in an optimal way. In this way, it can be thought of as a quick and easy field deformation type of summary. Another interpretation, already mentioned, is that it is a distributional comparison of the nonzero-valued intensities analogous to those from a quantile–quantile (q–q) plot. In fact, a q–q plot of the values in the set A compared with those in B is a scatter plot of exactly the same sorted, and possibly interpolated to the same size, intensities.

3 Test cases

Several test cases are employed in order to demonstrate the behavior of G_β and $G_{\beta, \text{IL}}$. Comparisons of G_β are made with FoM and $G_{\beta, \text{IL}}$ with FQI from Eq. (1) because these existing measures appear to be the most similar. The first set of test cases is applied to understand G_β and comes from the recently proposed set of comparisons by Gilleland et al. (2020),¹ namely the pathological, circle, ellipse, and noisy comparisons.

For $G_{\beta, \text{IL}}$, 3 h accumulated precipitation (in millimeters) from the core case of the Mesoscale Verification Inter-Comparison over complex Terrain project is used (MesoVICT; <https://ral.ucar.edu/projects/icp/>, last access: 11 February 2021; Dorninger et al., 2013, 2018). The precipitation case utilizes the Vienna Enhanced Resolution Analysis (VERA, Steinacker et al., 2000) as observation and the Canadian high-resolution model (CMH; Dorninger et al., 2013, 2018). The case is a subset of the minimum core case from the MesoVICT project and involves several valid times between 20 and 22 June 2007. Here, only 3 h of accumulated precipitation for the 12 h lead times is analyzed. Figures 5

¹Gilleland et al. (2020) is open access to facilitate easy viewing of the figures for these test cases. It is recommended that the reader refer to these figures when reading the results section.

to 7 show plots of these cases, along with diagnostic plots described in Sect. 5.

Finally, in order to provide a more thorough test of $G_{\beta,IL}$, the 32 spatial forecast verification ICP real test cases are also analyzed. These contain two models, namely the Advanced Research Weather Research and Forecasting Model (ARW-WRF; Skamarock et al., 2005) and the nonhydrostatic mesoscale model (NMM). Stage II reanalysis is used as the observations. These cases were previously analyzed by Davis et al. (2009) for MODE, Gilleland et al. (2010) and Gilleland (2013) with image-warping and hypothesis-testing procedures, and Gilleland (2017) using MED. The forecasts were initialized at 00:00 UTC (early evening), with a 24 h lead time, and were part of the 2005 National Severe Storms Laboratory and Storm Prediction Center Spring Program (Baldwin and Elmore, 2005). See Kain et al. (2008) and Ahijevych et al. (2009) for more information on these cases.

The development of spatial verification measures largely grew out of the specific application of verifying quantitative precipitation forecasts, which is the reason why the ICP and MesoVICT cases are focused on these types of fields. Nevertheless, G and G_{β} , and most of the spatial verification methods, can be applied to other types of variables such as wind speed. In the case of the distance-based measures, including G and G_{β} , as long as the variable field can be sensibly converted into a binary field, then the measures can be applied.

4 Application to the new geometric cases from MesoVICT

Gilleland et al. (2020) applied the Hausdorff distance, Baddeley's Δ , and MED to the test cases developed therein, so those results can be compared with the results for G and G_{β} in this section. Because FoM has not been considered on these cases previously, and because it is the most similar to G_{β} , it will also be evaluated and compared here. For brevity, G is only applied to the circle cases as results are generally consistent with G_{β} , which means essentially comparing the functions $f_1(x) = x^{1/3}$ with $f_2(x) = 1 - x/\beta$ for $x/\beta \geq 0$.

4.1 Pathological cases

The pathological comparisons represent challenging, but often encountered, situations where fields are empty, full, or contain only some small amounts of noise. Each individual field is on a 200×200 domain, as described in Table 2, and is labeled P1 to P7. A total of two comparisons also involve a circle of radius 20 centered at (100, 100) and labeled C1.

All of the distance-based measures tested by Gilleland et al. (2020) were either undefined or gave erratic results for the pathological comparisons. Table 2 demonstrates that G_{β} , using $\beta = N^2/2 = (200 \cdot 200)^2/2$, behaves naturally even for the most difficult of these tests. The comparisons of P1 against P1 (henceforth denoted by P1P1 and similarly for other comparison labeling) and P2P2 yield G_{β} values that are

identically one, which is desired as they are perfect matches. When one or a few one-valued grid points are added to the empty field (i.e., cases P3, P4, P5, P6, and P7), the comparisons P1P3, P1P4, P1P5, and P6P7 in the figure yield G_{β} values that are very close to, but slightly less than, one. Because these fields are largely identical outside of this small amount of noise, a G_{β} near one is the result that is wanted.

On the other hand, the comparison between P1 and P2, which represents a worst-case scenario of an empty field ($I_{P1}(s) = 0$ at every $s \in \mathcal{D}$) compared to the full field, P2 ($I_{P2}(s) = 1$ at every $s \in \mathcal{D}$) yields $G_{\beta}(P1, P2) = 0$, as desired. The comparisons P2P5 and P2P6 are comparisons that are nearly identical to the comparison P1C1, except that, again, P5 and P6 have four nonzero-valued grid points. $G_{\beta} = 0$ for these comparisons, agreeing with the subjective judgment. It is also zero for the comparison P2C1, which compares the full field against a field with a relatively small circle of one-valued grid points.

Finally, $G_{\beta}(P1, C1)$ has a relatively higher value than for $G_{\beta}(P2, C1)$, if $\beta = (200 \cdot 200)^2$ (right), and for $\beta = (200 \cdot 200)^2/2$ (left), its value is zero (the same as for P2C1). For the higher β , the large amount of correct negatives allows for a better score for P1C1 than P2C1. The zero-value of G_{β} for P2C1 is, quite correctly, suggestive of a poor match as desired, regardless of whether $\beta = (200 \cdot 200)^2$ or $(200 \cdot 200)^2/2$.

Pratt's FoM is not defined whenever it is calculated from the empty field P1 because there is no non-empty set of one-valued grid points from which to calculate the distance map (including FoM(P1, P1), FoM(P1, P2), etc.). If using the rule that $d(\cdot, A) = D$ for large D when A is empty, then FoM will be near zero for any comparison from P1. FoM for each of the remaining comparisons is very close to zero, indicating that each pair of fields is a poor match, which is not consistent with a subjective evaluation. That is, the empty field creates a situation where FoM is either not defined, or it suggests a bad match regardless of whether the match is very good or very bad.

4.2 Circle comparisons

Figure 8 shows results for G from Eq. (2) applied to the circle comparisons, and Fig. 9 shows the results for G_{β} and FoM applied to the circle and ellipse comparisons proposed by Gilleland et al. (2020, Figs. 5 and 6). Small dotted vertical lines separate similar comparisons as a visual aid. The first five circle comparisons demonstrate that neither G nor G_{β} is sensitive to the positioning of the one-valued grid points within the domain. That is, they give the same value for the first three comparisons, which are identical comparisons, apart from their positions within \mathcal{D} , which is similar for the next two comparisons (C1C4 and C3C4). This property is desirable but is not achieved by all summary measures, including Δ . The first three comparisons represent identical circles that are translated so that they touch at one point but do not overlap. G_{β} penalizes them for the lack of overlap but

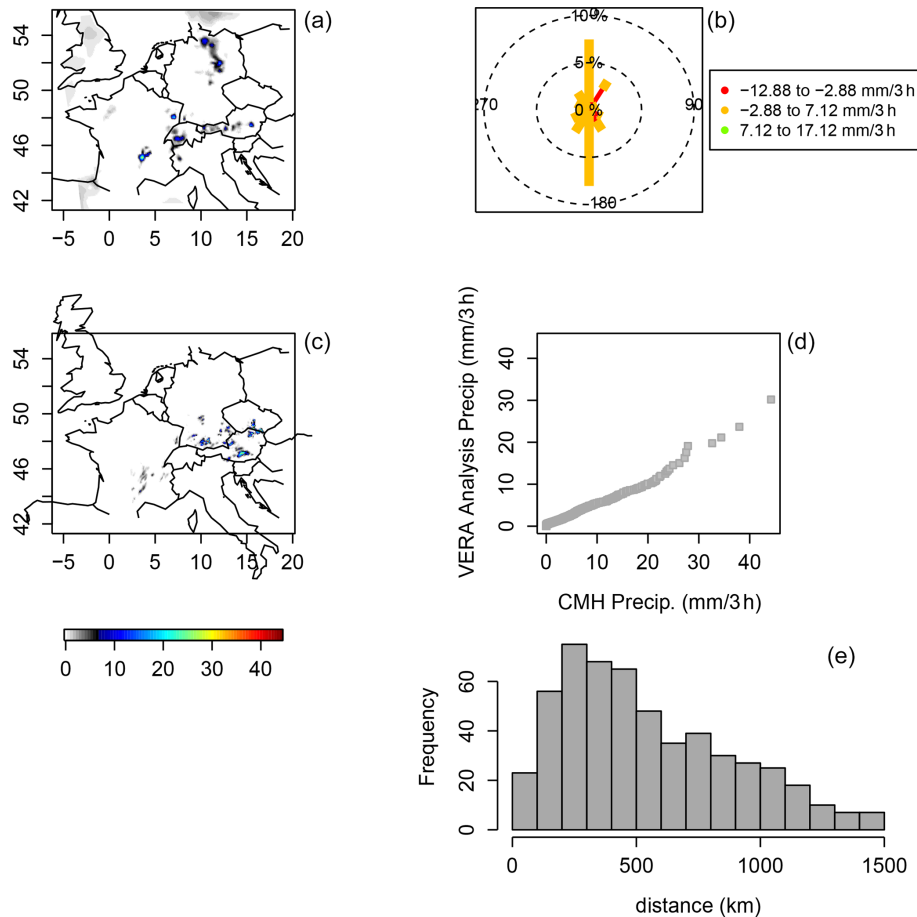


Figure 5. MesoVICT case 1 verification set, ending on 20 June 2007, at 18:00 coordinated universal time (UTC), for 3 h accumulated precipitation, given in millimeters (a, c). VERA analysis (a) and CMH with a 12 h lead time (b). Circle histogram (analogous to a wind rose diagram) showing the sorted model values, minus the sorted analysis values, and binned by the bearing from the analysis to the model (referenced from the north) of these sorted values (b). A q–q plot of VERA against CMH over all values (d). A histogram for the distances between the sorted values of the analysis and the model (e). See Sect. 5 for a detailed explanation of this figure.

otherwise gives these cases a relatively high mark. G is less readily interpreted, but its relative value for the comparisons is consistent with the other distance-based measures applied in Gilleland et al. (2020). C1C4 and C3C4 are also identical circles but have been translated further apart, and this additional translation error is penalized more than with the first three cases. FoM, while generally not symmetric, is symmetric for the first 11 comparisons and for C1C10, C6C12, and C13C14. FoM penalizes the first five comparisons considerably, yielding values very near to zero.

C2C5 and C3C5 are similar comparisons as for C1C2, for example, except that these circles overlap as they are not translated as far. Their G_β values are subsequently very near to one, but their FoM values are relatively low at about 0.6. $G \approx 288.86$ for both of these comparisons, which puts them in a tie for second in terms of ranking the quality of the matches across the comparisons (C13C14 is better, according to G , with a value of about 263.99).

C2C11 is a comparison in which C11 is the union of the three circles C1, C3, and C4 against C2 and was originally introduced in Gilleland (2017) in testing the MED. It was found that $MED(C2, C11)$ gave the same value as it gave for the C1C2, C2C3, and C2C4 comparisons, but $MED(C11, C2)$ was much smaller. C1C6 is a similar comparison, but with only two additional circles, and results are similar for this comparison. G and G_β are symmetric, so there is no difference between, for example, $G_\beta(C2, C11)$ and $G_\beta(C11, C2)$, and both penalize this comparison because of the lack of overlap and the additional circles. FoM does not distinguish this comparison at all from the C1C2, C2C3, and C2C4 comparisons, again indicating its inability to distinguish shape errors.

C7 and C8 have the same upper circle as C6, but the lower circle is offset, with C7 more so than C6. $G(C6, C7) = G(C1, C2)$ and $G_\beta(C6, C7) = G_\beta(C1, C2)$ because, despite having a perfect match with the upper circle, the error with the lower circle in C6C7 is identical to the error between C1

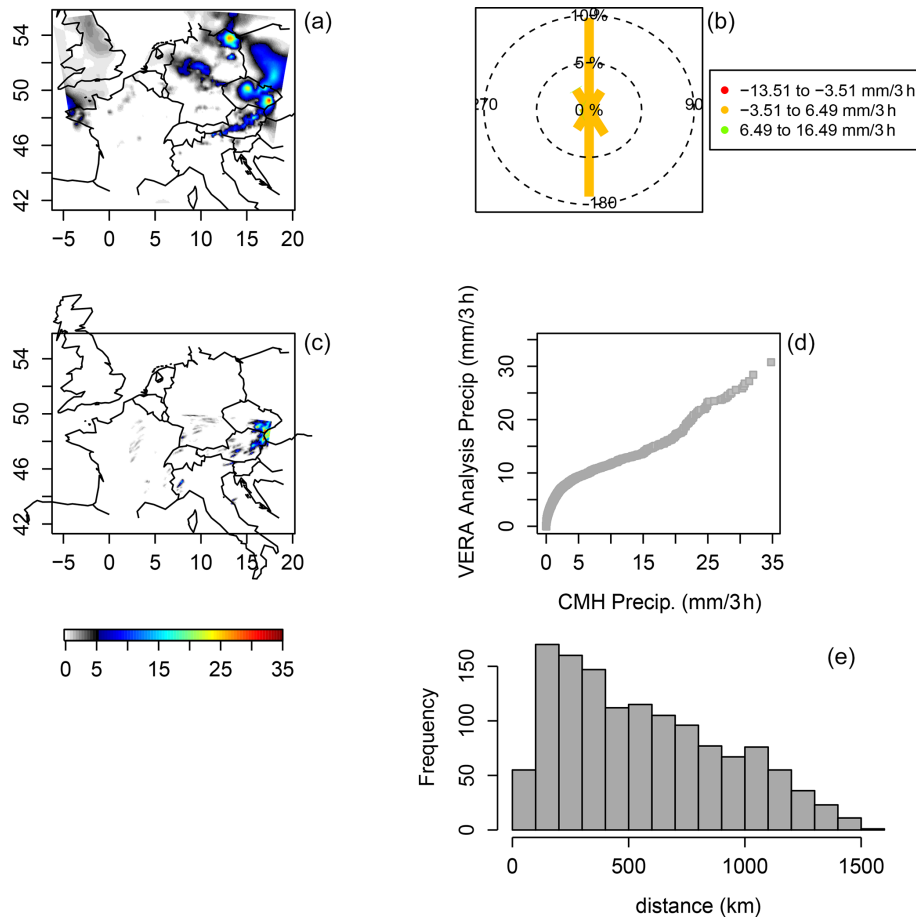


Figure 6. Same as Fig. 5 but ending on 21 June 2007.

and C2. FoM, on the other hand, is much higher for C6C7 than C1C2. Thus, FoM rewards more for having some more parts correct than G_β .

C9 is a circle with the same center as C1 (centroid distance is zero) but is much larger. $G_\beta(C1, C9) = 0$, using $\beta = 200^3/2$, whereas FoM is very low but not identically zero. $G(C1, C9) \approx 1367.67$ is a high value and shows this case to be the worst of the circle comparisons. C1C10 is a similar comparison, but the larger circle is replaced with a ring so that C10 does not intersect with C1, and there is a gap between them. Again, the centroid distance is zero (perfect score), but both G_β and FoM are zero (worst score possible). $G(C1, C10) \approx 1231.91$ is also a high value, indicating a poor match, but is lower than for C1C9, which contrasts with Δ , which was found to have approximately the same value for these comparisons.

C6C12, similarly, has a centroid distance of zero this time because the two circles in C12 are translated equally but in opposite directions from the two circles in C6, and each circle in C12 overlaps with the two circles in C6. G_β is relatively high for this case, but lower than for the first three comparisons, penalizing for the additional area that is not

overlapped. $G(C6, C12) \approx 637.59$ is again difficult to interpret in terms of good versus bad but suggests that, relative to the other comparisons, it is much better than the most egregious error comparisons.

Finally, G_β and FoM disagree completely about how good of a match there is between C13 and C14. Both sets of circles include much smaller circles than before, so G_β (with $\beta = N^2/2$) does not penalize the relatively small translation errors as much. On the other hand, FoM greatly penalizes them for this case, ignoring the circle sizes.

As discussed in Sect. 2, the FoM result for C13C14 (see Fig. 10a) might be the more desired result for this case if these relatively small circles represent an event such as severe weather rather than unimportant noise. In such a case, G_β can still be effective if the value of β is chosen with the physical interpretation in mind. For example, the base rate for C1 (i.e., n_{C1}/N) is a little larger than 0.03, whereas it is less than 0.01 for C13 and C14. The total size $n_{C13} = n_{C14} = 362$, and assuming each grid square represents, say, 4 km, then one might consider an error of 10 grid squares (≈ 40 km) to be too large. In fact, the largest circles in C13 and C14 have a radius of eight grid squares. Therefore, an

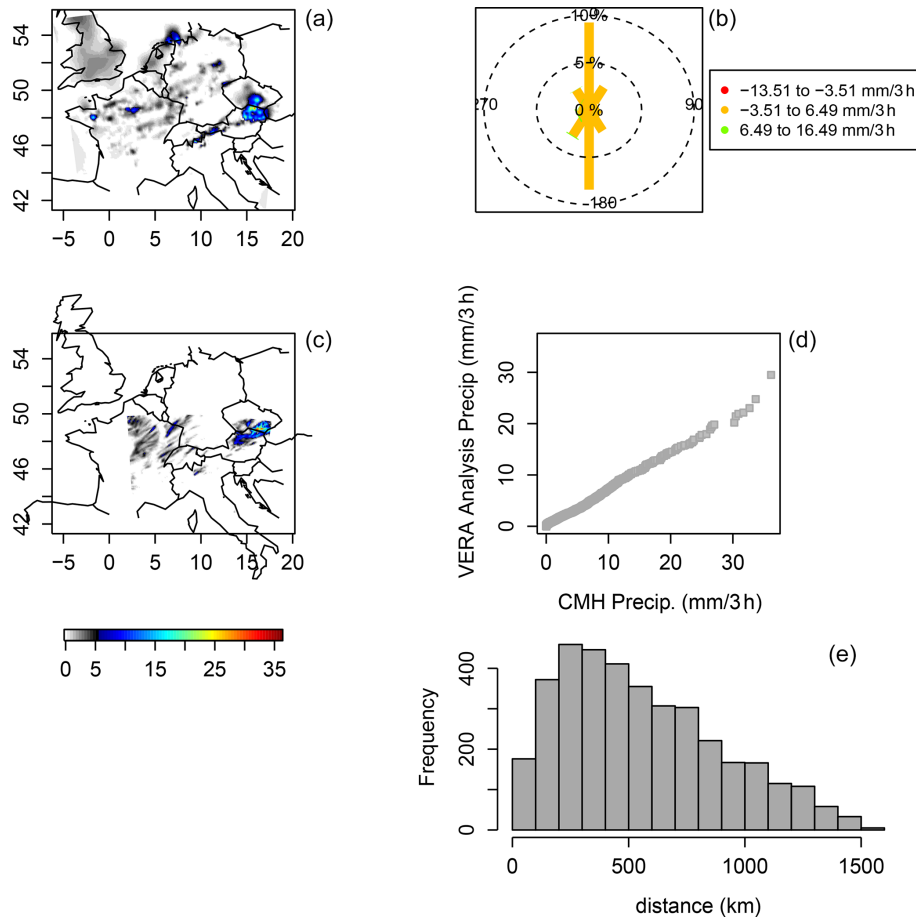


Figure 7. Same as Fig. 5 but ending on 22 June 2007.

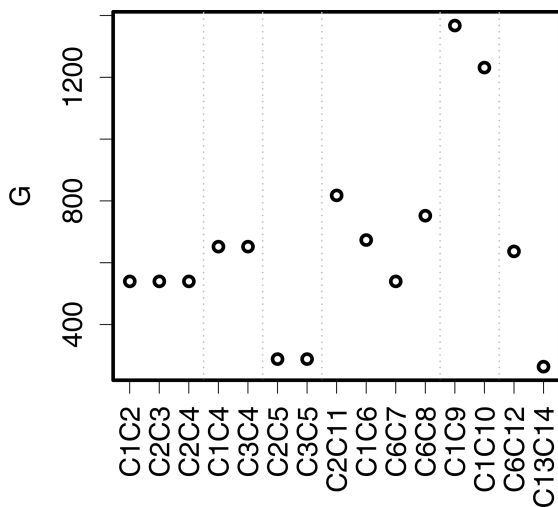


Figure 8. Results for G from Eq. (2) applied to the circle comparisons from Gilleland et al. (2020). Smaller values are better, and zero signifies a perfect match. Dotted vertical lines separate similar comparisons as a visual aid.

upper bound for a very poor match for y from Eq. (3) might be considered to be about $350 \cdot 10 = 3500$. Figure 10b shows $G_{\beta}(C13, C14)$ for varying values of β . The left vertical line is through $\beta = 3500$, and the right vertical line is through $\beta = N^2/2 = 20\,000$ that is used for the above results. In other words, if the small circles in the figure are important, then a smaller value of β should be used, and if they represent noise, then a higher value is in order. $G_{\beta=3500}$ correctly informs that these two cases are not close enough, whereas $G_{\beta=N^2/2}$ indicates a fairly good match.² G gives the lowest value for this comparison out of any of the other comparisons, indicating it is the best match. If the application needs

²An alternative strategy to employ when applying G_{β} for severe storm types of situations is to use it within a feature-based approach. In this setting, a bounding box can be placed around the union of the spatial sets to be compared (i.e., individual features rather than all sets over the entire domain) so that the size of the domain is the size of the bounding box, say B , and $\beta = B^2/2$. Such an approach introduces an additional complexity in that the individual features need to be identified within each field and matched across fields. Both procedures require the user to make choices about parameters and methods, and the results may be quite sensitive to these choices.

Table 2. Description (top) of the pathological fields from Gilleland et al. (2020). Each field is on a 200×200 domain, \mathcal{D} , with $\mathbf{s} = (x, y)$ giving the x and y coordinates such that $(1, 1)$ is the lower left corner. Results (bottom) for G_β for comparisons of the pathological cases. The value ≈ 1 indicates that $G_\beta < 1$ but $G_\beta = 1.00$ after rounding to two (or even more) decimal places.

Case identifier	Description
P1	Empty field ($I_{P1}(\mathbf{s}) = 0$ at every $\mathbf{s} \in \mathcal{D}$)
P2	No zero-valued grid points ($I_{P2}(\mathbf{s}) = 1$ at every $\mathbf{s} \in \mathcal{D}$)
P3	One nonzero-valued grid point in the lower left corner ($I_{P3}(\mathbf{s}) = 1$ at $\mathbf{s} = (1, 1)$ and $I_{P3}(\mathbf{s}) = 0$ everywhere else)
P4	One nonzero-valued grid point in the upper right corner ($I_{P4}(\mathbf{s}) = 1$ at $\mathbf{s} = (200, 200)$ and $I_{P4}(\mathbf{s}) = 0$ everywhere else)
P5	One nonzero-valued grid point in approximately the center of the field ($I_{P5}(\mathbf{s}) = 1$ at $\mathbf{s} = (100, 100)$ and $I_{P5}(\mathbf{s}) = 0$ everywhere else)
P6	Four nonzero-valued grid points in the corners ($I_{P6}(\mathbf{s}) = 1$ at $\mathbf{s} \in \{(1, 1), (200, 1), (1, 200), (200, 200)\}$ and $I_{P6}(\mathbf{s}) = 0$ everywhere else)
P7	Four nonzero-valued grid points in the middle of each boundary ($I_{P7}(\mathbf{s}) = 1$ at $\mathbf{s} = \{(1, 100), (100, 1), (100, 200), (200, 100)\}$ and zero everywhere else)
G_β	(P1, P1) (P1, P2) (P2, P2) (P1, P3) (P1, P4) (P1, P5) (P2, P5) (P2, P6) (P6, P7)
$\beta = N^2/2$	1 0 1 ≈ 1 ≈ 1 ≈ 1 0 0 ≈ 1
$\beta = N^2$	1 0 1 ≈ 1 ≈ 1 ≈ 1 0 0 ≈ 1
	$G_\beta(\text{P1}, \text{C1})^a$ $G_\beta(\text{P2}, \text{C1})^a$
$\beta = N^2/2$	0 0
$\beta = N^2$	0.36 0

^a C1 is a circle of radius 20 centered at the point $(100, 100)$.

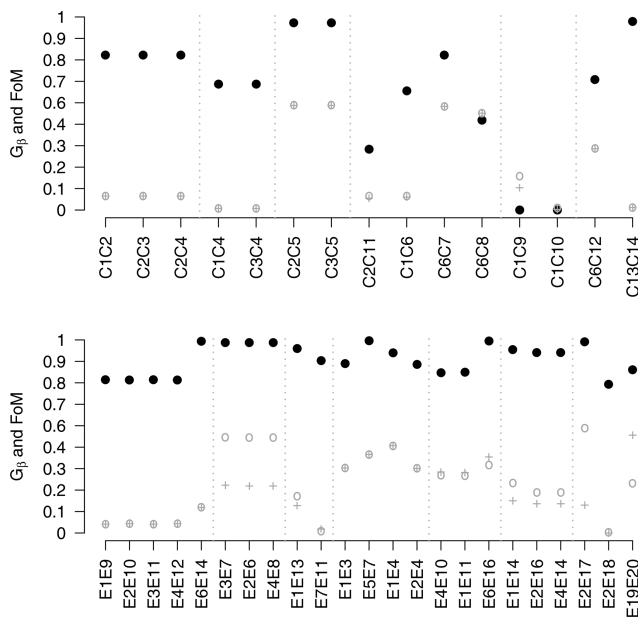


Figure 9. G_β and FoM results for the geometric comparisons from Gilleland et al. (2020), which are on a 200×200 gridded domain. At the top are the circle comparisons, and at the bottom are the ellipse comparisons. Black solid circles indicate G_β values (using $\beta = 200^4/2 = N^2/2$), gray circles are $\text{FoM}(A, B)$, and gray + signs are $\text{FoM}(B, A)$, where A represents the first case in the label, e.g., C1 from C1C2 and B the second; $\alpha = 1/9$. Dotted vertical lines separate similar comparisons as a visual aid.

these cases to be more heavily penalized, then G is not the ideal choice of measure, and the Hausdorff distance would be recommended.

4.3 Ellipse comparisons

The ellipse comparisons (Gilleland et al., 2020, Figs. 7 and 8) are shown in the lower panel of Fig. 9, and abstract the types of situations common over complex terrain. The vertical dotted lines again provide a visual aid to separate different types of errors represented in the comparisons, namely, (i) translation-only errors, followed by (ii) size-only errors, (iii) size and translation errors, (iv) rotation only errors, (v) rotation and translation errors, and (vi) rotation, translation, and size errors, and (vii) the final three comparisons involve a single ellipse compared against several small ellipses within a larger elliptical envelope and three ellipses compared against a larger blob that roughly follows the shape of these ellipses.

G_β and FoM disagree considerably on the quality of the match for the translation-only errors, where, in this case, G_β perhaps gives a value (around 0.8) that is more consistent with a subjective evaluation because the ellipses are identical – apart from a relatively small translation (no overlap). FoM, on the other hand, gives a value very close to zero. They also disagree on E7E11, which is a similar comparison, except that one of the ellipses is much smaller than the other. G_β again gives a high value (higher than for the first several comparisons), and FoM is again very low and even closer to

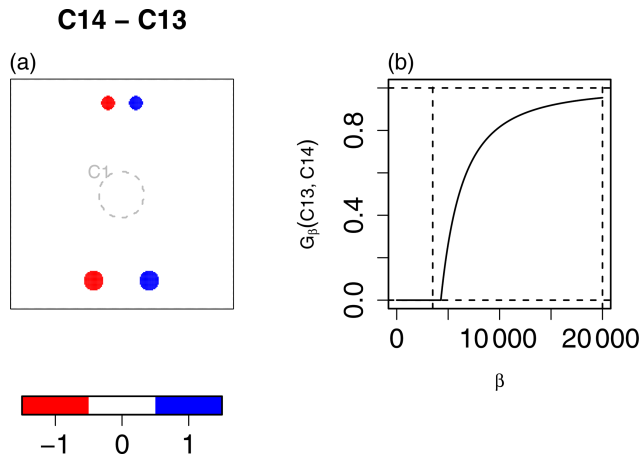


Figure 10. (a) The C13C14 comparison shown as the difference $C14 - C13$. The size of the one-valued set is small relative to the size of the domain. The base rate is less than 0.01 for both fields. The total size of these sets is 362. For comparison in terms of the spatial areas of the circles, the gray dashed circle in the center shows C1. (b) $G_\beta(C13, C14)$ applied with a range of different β values. The left vertical line is through $\beta = 3500$, and the right vertical line is through $\beta = N^2/2 = 20000$.

zero in this particular comparison. Finally, they also disagree on E2E18 in the same way. This case again involves a translation error and essentially a size error. One might agree more with FoM that this match quality is poor, but if the small-scale detail is less important than the overall size and shape of the extent of, say, the type of convective activity in complex terrain for which this comparison was designed, then G_β provides a sensible result.

Otherwise, the two measures generally agree with each other in relative terms; that is, they rank the remaining comparisons similarly. G_β generally results in fairly high values for all of these comparisons, whereas FoM is below 0.6 for every one, giving its highest values slightly above 0.5 for the size-only errors when calculated as FoM(larger ellipse, smaller ellipse) as shown by the gray circles. However, when calculated as FoM(smaller ellipse, larger ellipse), depicted with the gray + signs, FoM is below 0.3, indicating a much poorer match.

4.4 Noisy comparisons

The noisy comparisons from Gilleland et al. (2020) are analyzed with G_β , using $\beta = (200 \cdot 200)^3/2$ and FoM (Fig. 11). The first comparison is the same as that for C1C4 but with additional noise, as described in the Fig. 11 caption. $G(C1, C4) \approx 0.65$, which is a bit lower than it is for $G(N1, N2) \approx 0.87$. In this case, the amount of overlap is about the same, but because of the noise that is scattered about the entire domain in both fields, there is less average distance between N1 and N2 than there is between C1 and C4. Indeed, Gilleland et al. (2020) show MED values

that are smaller by about 10 grid squares for the N1N2 case, which explains the lower G_β value here. The difference is not great, however, demonstrating that G_β is not overly sensitive to noise.

The latter two noisy comparisons again involve the C1C4 comparison but with a single point of noise added to the C4 field. In the first, the single point is added at the center of C1, and in the last it is added to the lower left corner of the domain. $G(C1, N3) \approx 0.76$ is closer to the value for C1C4 and is nearly identical for C1N4. All of the measures applied in Gilleland et al. (2020) were fairly sensitive to this noise, which is positioned far away from any other nonzero grid point. Because of the differences in scale, the measured effect cannot be compared, but both 0.76 and 0.65 are reasonably close together on the zero to one scale. Moreover, differentiating between the two fields is desired, but it is important to note the size of the difference relative to the error.

FoM is also applied to these comparisons, and while it ranks them the same as G_β (i.e., N1N2 has the highest value followed by C1N3 and C1N4, resp.), it suggests a very poor match between each set of comparisons. It also gave a low grade to the C1C4 comparison, and the value changes in the same direction as for G_β but only very slightly, as it was already very close to zero.

5 G_β and $G_{\beta,IL}$ applied to real test cases

Figure 5 shows a snapshot of the MesoVICT core case verification set valid on 20 June 2007 at 18:00 UTC. Comparing Fig. 5a and c, it is clear that CMH (Fig. 5d) missed 3 h precipitation in northern Germany and over forecast precipitation in southern Germany, Austria, and a small part of the Czech Republic at this valid time. The circle histogram (Fig. 5b) shows the difference in sorted CMH-modeled 3 h precipitation minus the sorted VERA analysis values binned by the bearing of the grid points associated with these sorted values from VERA to those of CMH using north as reference. The northern misses and southern false alarms stand out with the long petals in the north and south directions. The q-q plot (Fig. 5d) shows a generally good agreement in terms of the distribution of 3 h accumulated precipitation values but with a slight tendency for the CMH to forecast larger values. Finally, the histogram (Fig. 5e) shows the distances between the sorted values, which are generally very far apart, with a high density of distances beyond 500 km.

Figure 6 is the same as Fig. 5 but ends the 3 h period 21:00–24:00 UTC on 21 June. At this valid time, there is considerably more missed 3 h accumulated precipitation covering much of northern Germany, the Czech Republic, and western Poland (the rest of Poland is outside of the domain). It also misses the precipitation found off the coast of, as well as over, Brittany, France. Figure 6b again shows a north–south issue, while the q-q plot (Fig. 6d) reveals that, for this valid time, the forecast is now under-predicting lower values

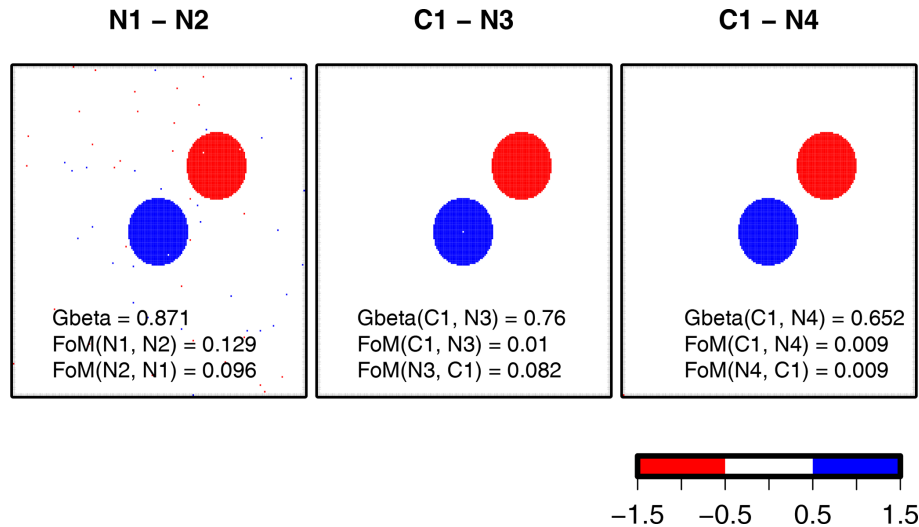


Figure 11. The noisy cases from Gilleland et al. (2020), with results for $G_{200^4/2}$ and FoM rounded to three decimal places. Images show the difference between each case. Cases N1 and N2 are the same as C1 and C4 but with additional random noise ($I_{N1}(s) = 1$ and $I_{N2}(s) = 1$ at randomly chosen s). N3 and N4 are again the same as C4 but with $I_{N3}(s) = 1$ at $s = (100, 100)$ and $I_{N4}(s) = 1$ at $s = (1, 1)$.

of 3 h accumulated precipitation but otherwise captures the distribution well for values above about $5 \text{ mm}(3 \text{ h})^{-1}$. The histogram of distances (Fig. 6e) between sorted values again shows the long distances between them.

Finally, Fig. 7 shows the same graphic displays but ends the 3 h period 21:00–24:00 UTC on 22 June. CMH again misses precipitation in the north, this time near the easternmost Frisian islands through to the northwestern coast of Germany, and some scattered showers in northern Germany. The circle histogram (Fig. 7b) again shows the north–south bias, but the q–q plot (Fig. 7d) suggests that the model matched the distributional properties of the analysis very well at this valid time. The histogram of distances (Fig. 7e) between sorted values is similar to the previous two valid times.

Much information about forecast performance is provided in Figs. 5 to 7. Subjective analysis of these figures would suggest that the forecast on 20 June is the best, followed by 22 June, with 21 June clearly being the worst. In practice, it is not always possible to diagnose forecast performance with this much detail, and a good summary measure should rank the forecasts valid on these 3 d as a subjective inspection of the graphs in these figures would. Of course, subjective evaluation is also limited, and these summary measures can be useful in identifying problems that are not so easily seen (e.g., Ahijevych et al., 2009; Gilleland, 2017).

Figure 12 shows the results after applying the FQI to the core case 3 h accumulated precipitation fields from MesoVICT for values above $2 \text{ mm}(3 \text{ h})^{-1}$. Because the FQI involves randomly generated surrogate fields in its calculation, its value will vary each time it is computed. Therefore, 100 values are obtained, using 10 surrogate fields each, in order to obtain a distribution of values. Day 2 is clearly the worst

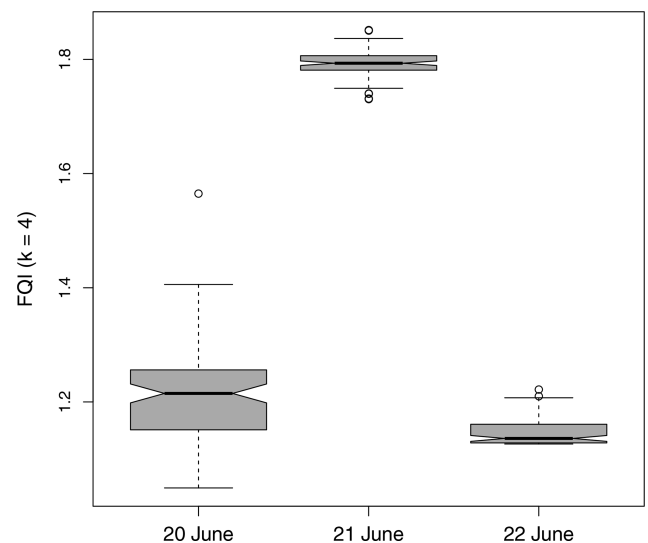


Figure 12. FQI applied to the MesoVICT case 1 12 h lead time forecasts for 3 h accumulated precipitation (in millimeters) ending at 18:00 UTC in 2007 for values above $2 \text{ mm}(3 \text{ h})^{-1}$. In each case, 100 FQI (using $k = 4$) values are found using 10 surrogate fields each, and the box plots represent the distribution of FQI values for each set of surrogates.

forecast of the 3 d, according to FQI. Day 1 has the most variability but is found to be the second best for most of the 100 iterations, with its 95 % confidence interval (CI) for the median (shown via the notches in each box) clearly not overlapping with the FQI median notches of day 3. The FQI suggests that day 3 is the best forecast day of the three. Subjectively, one might argue that the day 1 is the best, though the forecast did not fare very well in any case. When applied with

$\beta = N^2/2$, $G_{\beta,IL}$ is near one, suggesting that the choice of β is too large, which is expected as it is the relatively small spatial areas of intense precipitation that are of interest, and G_{β} with $\beta = N^2/2$ is not bothered by such small-scale phenomena. Instead, a much smaller value of β is necessary.

To ascertain a better choice of β , Fig. 13 shows results for G_{β} and $G_{\beta,IL}$ applied to these same three cases for three different thresholds and choices of β , as well as two choices of weight ($\omega = 1/2, 3/4$) for $G_{\beta,IL}$. In fact, using the logic of choosing β to be on the scale of the phenomena of interest, $G_{\beta} = 0$ and $G_{\beta,IL}$ simply give the resulting correlation of the sorted values within the threshold excess areas multiplied by $1 - \omega$. Such low values for these fields should be desired. Nevertheless, the results shown in Fig. 13 are for much larger choices of β in order to give a sense of the sensitivity in choosing this parameter. The value of $N^2/2$ for these fields is of the order of 800 million, so the choices in Fig. 13 are very large but substantially smaller than $N^2/2$, and they represent values where, for these types of fields, G_{β} moves from zero to near one. Both G_{β} and $G_{\beta,IL}$ at these levels of β clearly indicate the poorest performance of the model on 21 June, with a slight preference for the day 1 over day 3, which is arguably consistent with subjective evaluation.

Figure 14 displays the results from applying $G_{\beta,IL}$, $G_{2,\beta,IL}$ (a modification of G_{β} discussed briefly in Sect. 6.2), and ETS to the 32 ICP real cases. Both of the new measures generally agree with previous results using MODE, image warping, and MED that ARW-WRF has a similar performance to NMM but slightly better. While there are some occasions for the lowest threshold where $G_{\beta,IL}$ suggests poor performance for NMM, most show fairly good performance, especially for the two highest threshold choices, which agrees with subjective evaluation and suggests that NMM occasionally has too much scatter that is omitted as the threshold increases. The ETS, on the other hand, suggests close to no skill for both models across the range of values, which is to be expected from a grid-point-by-grid-point type of measure.

Comparing Fig. 14 with Fig. 5 from Gilleland et al. (2010) shows that there is reasonably good agreement between the two methods in terms of which valid times have the best performance and which ones do not and which of the two models is better. The image-warping procedure from Gilleland et al. (2010) is a relatively cumbersome procedure that is difficult to implement in practice, while $G_{\beta,IL}$ and $G_{2,\beta,IL}$ are straightforward and efficient to compute. On the other hand, no thresholding procedure is required with the image-warping method.

Figure 15 compares rankings for the nine ICP real cases that are a subset of the 32. These cases were introduced and evaluated subjectively in Ahijevych et al. (2009), and they were also evaluated using a summary, called IWS, based on image-warping results by Gilleland et al. (2010). These results are included for comparison with G_{β} and $G_{\beta,IL}$. The subjective scores cannot be taken as truth, and indeed, there was a large amount of variability among the subjective eval-

uators' opinions. Rankings for all of these methods are more similar for the NMM model than for ARW. With some notable exceptions, such as 13 May and 4 June for the ARW model, the IWS rankings are fairly similar to those of G_{β} and $G_{\beta,IL}$.

6 Discussion and conclusions

A total of three new measures are presented, dubbed G , G_{β} and $G_{\beta,IL}$, which provide meaningful summaries of forecast performance. The first two are spatial alignment and distance based, and the third additionally includes intensity error information in a distributional sense. The latter two measures each range from zero to one, where one represents a perfect match and zero a very poor forecast with the notion of very poor, determined by a user-chosen parameter, β . Desirable properties that have proved challenging for competing measures include that they provide (i) sensible information, even when one or both fields are empty of values, (ii) sensible information, when one or both fields are mostly empty of values, and (iii) information that is not overly sensitive to noise.

Although considerable effort has been made to rigorously test G and G_{β} , tests for $G_{\beta,IL}$, here, are less rigorous and should be seen as evidence that the measure provides reasonable information. Further testing will be necessary to know how well it can inform about forecast performance under all possible situations. Indeed, the intensity portion of the measure sets a low bar for a forecast to pass muster. However, if a single summary is needed, then it seems to perform well. However, G or G_{β} , combined with the traditional frequency bias, along with a distributional summary of intensity errors (e.g., differences in mean intensity or, better, the Kullback–Leibler divergence; Kullback and Leibler, 1951) provide a relatively comprehensive summary of forecast performance with little complexity and computational efficiency.

Table 3 grossly summarizes the findings in Gilleland et al. (2020) for some of the measures tested there and for G and G_{β} . Each column specifies an attribute, where an affirmative answer is a positive and a negation is a negative for the measure. The first column refers to whether or not the measures behave well for all of the pathological cases. The next column, labeled “No positional effects,” refers to the first five circle cases in which the first three and second two sets A and B are completely identical, apart from their positions within the field. An affirmative answer means that the measure gives the same value for each set of identical cases. The third column refers to the comparison C2C11 as contrasted against any of the first three circle comparisons. An affirmative answer means that the measure provides sensible information for such biases. The “Useful for rare events” column pertains to whether the measure gives a poor grade to the case C13C14 or not. The last two columns refer to the pair of comparisons C6C7, as contrasted with any of the first three com-

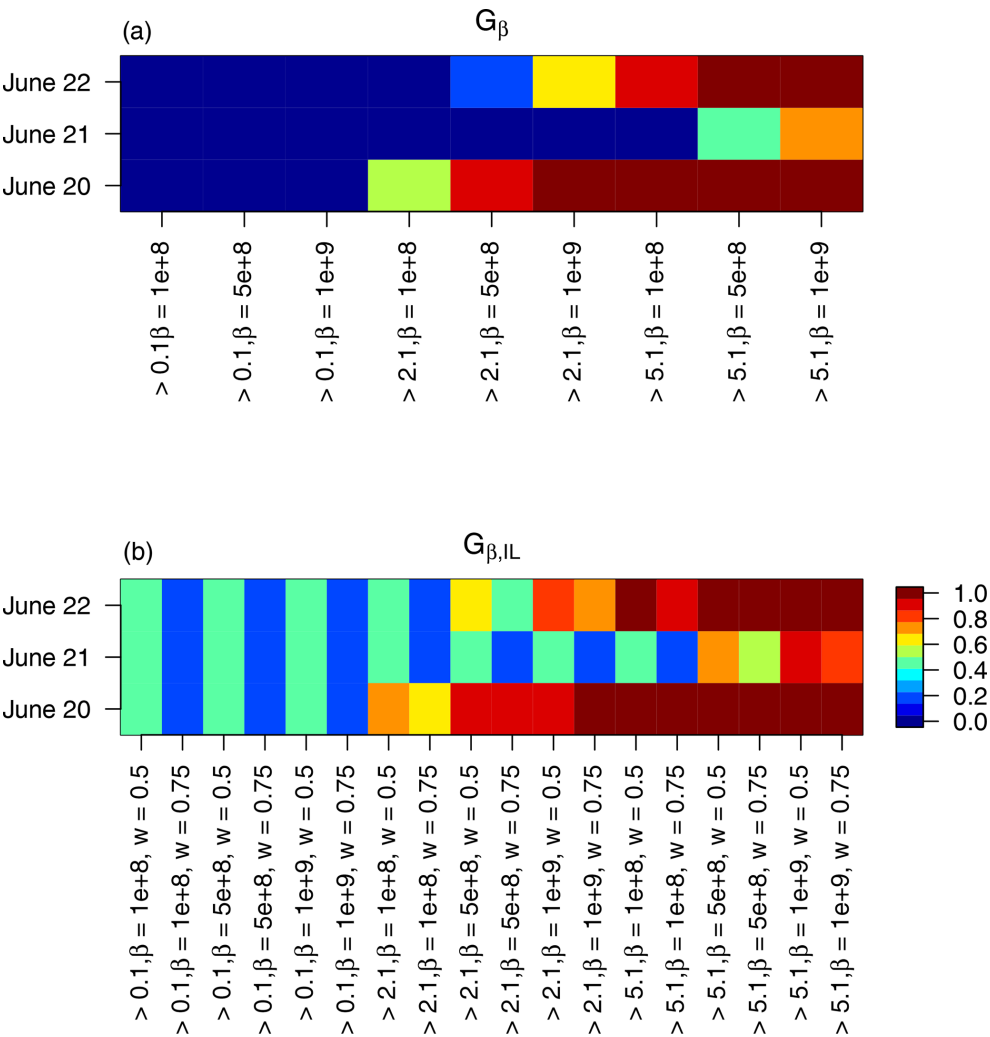


Figure 13. G_β (a), $G_{\beta,IL}$ (b) with $\beta = 1 \times 10^8$, 5×10^8 , and 1×10^9 and with thresholds of 0.1, 2.1, and 5.1 mm (3 h)^{−1} applied to the MesoVICT core case.

Table 3. Summary of the performance of the distance-based measures as informed through testing on the geometric cases from Gilleland et al. (2020). For each column, affirmative answers represent a positive attribute of the measure and negations represent a negative attribute. See the text for a description of each column heading.

	Good pathological handling	No positional effects	Sensitive to frequency bias	Useful for rare events	Reward partial perfect match	Correctly penalize despite partial match
Centroid distance	No	Yes	No	Yes	No	No
Baddeley's Δ	No	No	Yes	No	Yes	No
Hausdorff distance	No	Yes	No	Yes	No	No
MED	No	Yes	No ^a	Yes	Yes ^a	Yes ^a
FoM	No	Yes	Yes	Unclear	No	Yes
G	Yes	Yes	Yes	No	No	Yes
G_β	Yes	Yes	Yes	Yes ^b	No	Yes ^b

^a MED is an asymmetric measure that provides sensible information for these situations in one direction.
^b With an appropriate choice of β parameter.

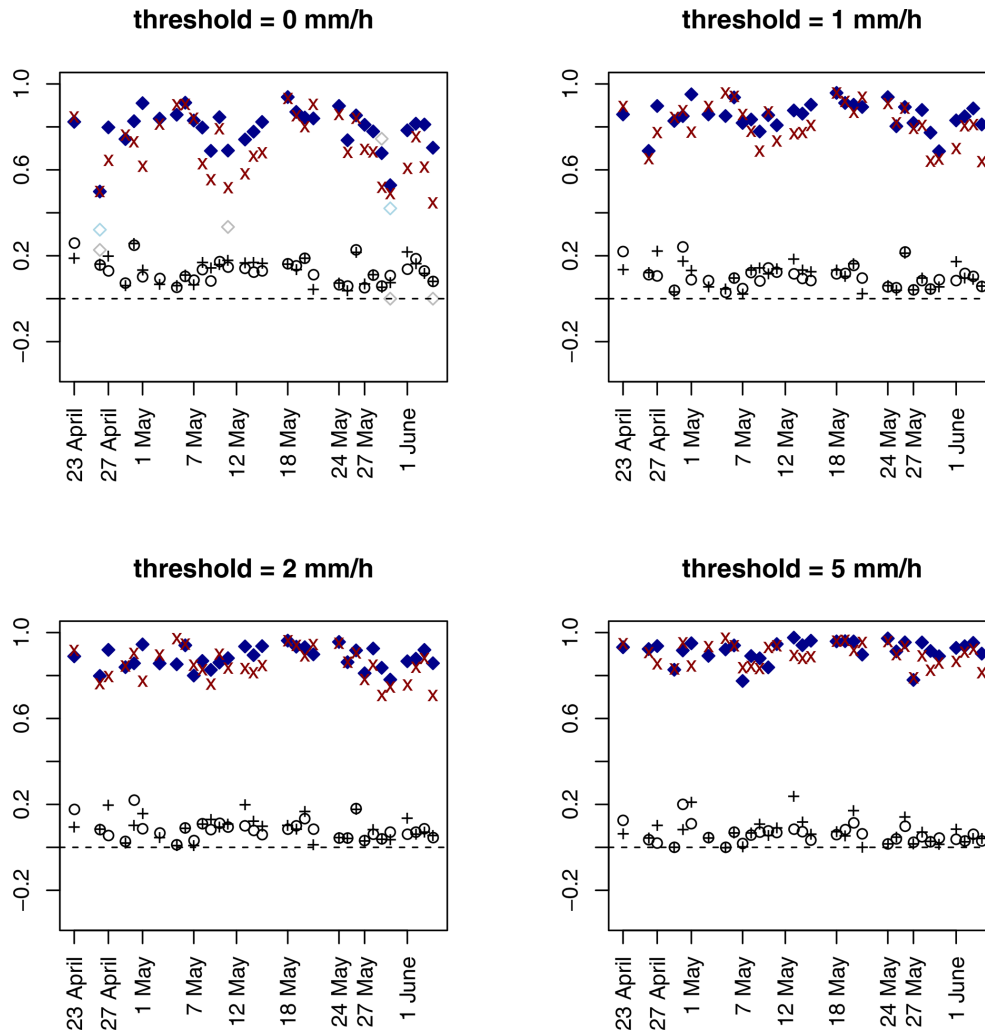


Figure 14. Results from applying $G_{\beta,IL}$, $G_{2,\beta,IL}$ (see Sect. 6.2), and ETS to the 32 ICP real cases. Blue solid diamonds are $G_{\beta,IL}$ applied to ARW-WRF, red crosses are for $G_{\beta,IL}$ applied to NMM, and open diamonds are $G_{2,\beta,IL}$ applied to ARW (light blue) and NMM (gray). Black + signs are ETS for ARW-WRF, and black circles are ETS applied to NMM. The ordinate axis limits are from $-1/3$ to 1, corresponding to the possible range for ETS, and the dashed line provides a visual aid for zero.

parisons and C1C9 and C1C10, respectively. An affirmative answer in the penultimate column means that the measure gives a higher score to C6C7 than, for example, C1C2. An affirmative answer in the last column means that the measure differentiates between C1C9 and C1C10 rather than yielding the same value. A negation for a measure in both of these last columns is particularly undesirable although one might prefer a negation in the penultimate column for certain applications. That is, the error part of C6C7 is identical to the error in C1C9, and the partially correct part of C6C7 could be contrasted with the correct negatives in C1C9 (they are both correct in this same region). Therefore, a negation in the last column seems to be the more egregious error.

As can be quickly surmised from Table 3, the only negatives for G are in the “Useful for rare events” and “Reward partial perfect match” columns. The latter is the only neg-

ative among these attributes for G_{β} , where, again, a careful choice of β is necessary in order to be useful in this setting. In terms of the handling of comparison of C6C7 versus C1C2, the measure gives the same result. Thus, it does not give an extra reward for having obtained a perfect match in another part of the field. Thus, the measure will penalize for any defect in the field.

The fact that G is not useful for rare events is by design. The construction of both G and G_{β} is made with the aim of handling the pathological cases, while still providing sensible information for most situations, which has clearly been accomplished. The rare-event situation is not the focus, although G_{β} is useful in this setting, provided an appropriate choice of β is used as described in the text. Moreover, the one situation in which the Hausdorff distance is particularly

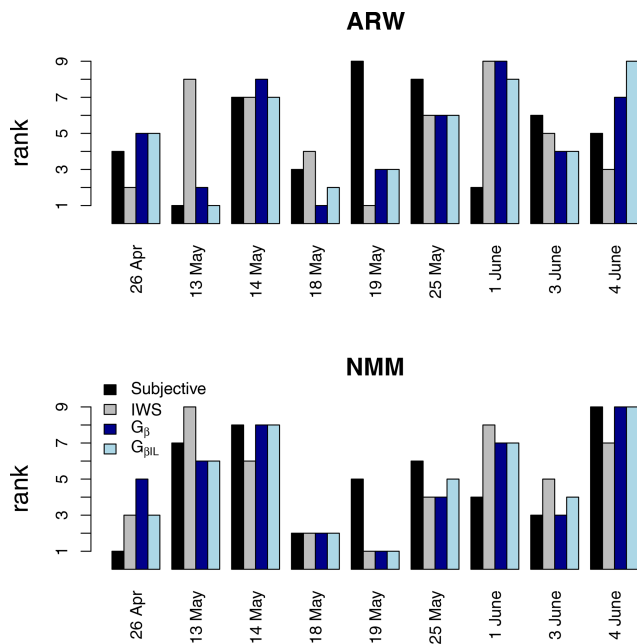


Figure 15. Comparison of the rankings (1 is best and 9 is worst) of the nine ICP real test cases from Ahijevych et al. (2009), including the subjective rankings described therein, and IWS, which is the image-warping score proposed by Gilleland et al. (2010). These cases are a subset of the 32 ICP real cases. A threshold of 2.1 mm h^{-1} was applied for G_β and $G_{\beta,IL}$. Because the spatial area of precipitation for these cases is relatively small compared to the domain size, $\beta = N/2$ is used here.

useful is the “rare event” setting so that it can provide complementary information to the measures introduced here.

6.1 Some thoughts about intensity information in spatial verification

The intensity summary suggested, here, for $G_{\beta,IL}$ is based on the distribution of errors so that, in particular, it is possible to obtain a perfect score when the two fields are not identical. The MesoVICT core case studied here shows an example in which any distributional measure for intensities is meaningless because the areas of intensity are separated by far too much distance. Therefore, the summary is not entirely satisfactory. It is meaningful when the spatial alignment of errors is much closer, and G or G_β can be used to help inform about the existence of such an issue.

Despite the fact that many methods have been proposed for solving the issue of double penalties and the over-accumulation of small-scale intensity errors, most of these methods do not directly assess intensity information. Most of them target intensity only indirectly through a thresholding procedure, like the one described in Sect. 1.1 and applied here. For example, the fractions skill score (Roberts and Lean, 2008) applies a smoothing filter to binary fields obtained through the threshold process in order to obtain infor-

mation about frequency of occurrence within neighborhoods of points. Similarly, the intensity-scale skill score (Casati et al., 2004) applies a wavelet decomposition to the verification sets after a fair amount of processing, including reducing the fields to binary fields via thresholding before applying the wavelet decomposition. Other filter-based approaches smooth the intensity values before summarizing the performance, which is also unsatisfactory. Similarly, measures like FQI incorporate only distributional information about the intensities.

The feature-based approach proposed in Davis et al. (2006a, b, 2009) also applies a thresholding process but maintains the intensity information within features. However, the intensity performance is also only incorporated in the final analysis through distributional values, usually in the upper quartile of intensities within a spatial feature. Because these comparative summaries are applied to individual features within the fields, MODE can be thought of as a crude deformation approach, so that its distributional summaries are more satisfactory (e.g., the gross errors of the MesoVICT core case will be correctly handled).

The issue with assessing forecast performance in terms of the intensities stems from the inability to apply a grid-point-by-grid-point accounting of their errors. Once the requirement that their precise locations be used is lifted, then it becomes less clear which values should be compared. Subsequently, thresholding, smoothing, and distributional summaries of intensity performance are the most relevant solutions, and much work has focused on accounting for their spatial alignment errors instead.

One of the earliest methods proposed for verifying forecasts spatially is based on deformations (see Gilleland, 2013, and references therein). G_β , and more so $G_{\beta,IL}$, can be thought of as potentially optimistic image warp results. In particular, the method presented in Gilleland et al. (2010) tries to optimize an objective function that penalizes for translations that are non-physical or too long. G_β and $G_{\beta,IL}$ summarize how much lack of overlap there is, and, on average, the distance of any lack of overlap. Thus, if everything could be perfectly aligned (the potential with or without the need for non-physical deformations), then it provides a summary of how much deformation might be required to do so. $G_{\beta,IL}$ summarizes the intensity errors in a similar way to the warping procedure of Gilleland et al. (2010) through a distributional measure over the sorted values from each field only (i.e., a perfect match). They are optimistic because they do not consider what it might take to actually realign the fields, and it is possible that no physically meaningful realignment is possible.

Letting RMSL_0 be the root mean square loss (RMSL)³ calculated grid point by grid point in the traditional manner, and RMSL_1 the RMSL calculated over the sorted values only (this time it includes all grid points in each field), then, in the following:

$$\frac{\text{RMSL}_0 - \text{RMSL}_1}{\text{RMSL}_0} \times 100 \quad (5)$$

would be the potentially perfect percent reduction in RMSL analogous to that defined in Gilleland et al. (2010) and would represent an upper bound on the error reduction percentage capable of being achieved by the image-warping procedure. This percent reduction in error is analogous to Eq. (2) from Mittermaier (2014) but where the reference RMSL is RMSL_0 . It is possible for a forecast to match the observation in intensity values perfectly, even if they are not located at precisely the same grid point locations, so that $y_3 = 0$. Indeed, a report of each of RMSL_0 , RMSL_1 , and the percent reduction in RMSL is a reasonably satisfactory summary of the intensity performance because it informs about the importance of displacement errors. Ebert and McBride (2000) apply translation-only deformations to individual features within the verification sets and summarize the RMSL similarly, as in Eq. (5), but where they are able to break it down into RMSL attributed to total, displacement, volume, and pattern types of errors. Such a breakdown is attractive but the processing involved to obtain them is considerable.

Another issue about summarizing intensity information is that many forecasts are first calibrated before applying any verification so that the intensity values should be fairly close in a distributional sense. Therefore, measures that include intensity components like those suggested here, i.e., FQI and MODE, among many others, will generally not be affected much by the intensity terms in their equations. Subsequently, the summaries proposed herein represent highly efficient, interpretable, and sensible summaries that provide a useful alternative.

6.2 Potential modifications to the proposed summary measures

A possible generalization of G_β ($G_{\beta, \text{IL}}$) that would allow a user to weight each component of y from Eq. (3) differently would be to replace y_1 and y_2 with $y_1^{\gamma_1}$ and $y_2^{\gamma_2}$. However, an initial investigation of implementing this additional complexity in G_β found the measure to be highly erratic for small changes in γ_1 and γ_2 , making it difficult to obtain an interpretable measure.

Another possible modification to G_β is to introduce an additional user-chosen parameter $0 \leq \alpha < \beta$ in the following

manner:

$$G_{\alpha, \beta}(A, B) = \max \left\{ 1 - \frac{y - \alpha}{\beta - \alpha}, 0 \right\}. \quad (6)$$

The effect is to allow for a perfect score for a model that is very close to the observation but not a perfect match. It could be thought of as an adjustment for errors that are not considered important. In fact, G_β given in Eq. (3) is the same as Eq. (6) but with $\alpha = 0$.

Another modification that is similar to $G_{\beta, \text{IL}}$, denoted here by $G_{2, \beta, \text{IL}}$, modifies Eq. (3) through the y term as follows:

$$G_{2, \beta, \text{IL}}(A, B) = \max \left\{ 1 - \frac{y_1 y_2 (1 + y_3)}{\beta}, 0 \right\}, \quad (7)$$

where y_1 and y_2 are as in Eq. (3), and y_3 is the mean absolute loss (MAL) between the sorted original values at locations within the sets A and B . The RMSL (as in Rezacova et al., 2007) could be used in the place of MAL if applying a greater penalty for large intensity differences is desired. If $y_3 = 0$, then $G_{2, \beta, \text{IL}}$ reduces to G_β . Otherwise, the measure provides an additional penalty for a lack of distributional agreement in the intensities.

Figure 14 displays results for this modification along with those for $G_{\beta, \text{IL}}$, and it can be seen that the results are similar for both measures, apart from a few cases where the multiplication of the intensity term causes a more drastic penalization of the forecasts than $G_{\beta, \text{IL}}$ (top left panel). A difficulty with $G_{2, \beta, \text{IL}}$ is that the multiplication of the intensity term in Eq. (7) makes choosing β more difficult because the intensity loss information may depend on the specific variable of interest, and the units for y_3 will generally differ considerably from y_1 and y_2 .

The summary measure, y , in both G_β and $G_{\beta, \text{IL}}$, passes the requirements to be a true mathematical metric. $\max\{1 - y/\beta, 0\}$ simply rescales the metric to be between zero and one and reorients it so that one, instead of zero, corresponds to a perfect match. Gilleland (2017) argued in favor of the violation of the symmetry property for the mean error distance (MED) because information about misses and false alarms can be inferred. G_β can be modified to be asymmetric in this way by only including one of the MED terms in the y_2 term of Eq. (3) and by removing one of the n_A or n_B terms from y_1 ; that is, an asymmetric $G_\beta^a(A, B)$ would have $y_1^a = n_B - n_{AB}$ and $y_2^a = \text{MED}(A, B) \cdot n_B$.

The building blocks for G_β and $G_{\beta, \text{IL}}$ themselves make for a good summary vector for forecast performance. For example, knowledge of n_A , n_B , n_{AB} , $\text{MED}(A, B)$, $\text{MED}(B, A)$, and, in the case of non-binary fields, y_3 with MAL or RMSL calculated over the areas A and B , without first sorting them, provides a wealth of diagnostic information about forecast performance. Examples include which component(s) contributed most to a particular G_β or $G_{\beta, \text{IL}}$ value, the frequency bias, the average distance from observed areas of activity to those that were forecast (and vice versa), and the potential

³Here, loss is taken to mean forecast values minus observed values. In the weather forecast verification literature, loss is referred to as error. The loss nomenclature is used here so as not to be confused with the statistical meaning.

and actual intensity error in addition to the percent reduction in error possible if the forecast field could be perfectly realigned with the observation field.

The proposed measures are designed for 2D spatial fields here, but because of the simplicity of the measures, they can be easily extended to any number of dimensions. For example, a third dimension of time could be added in order to verify an entire spatiotemporal verification set with a single summary number. Such a scheme would provide limited information, but it would nevertheless provide a simple summary that would allow for a ranking between competing gridded forecast models over numerous time points. Many variables also have an important vertical component so that the measures could be extended for the horizontal and vertical dimensions. Additionally, although the analysis here is focused on gridded verification sets, the measure can be extended to a verification set where both spatial fields are irregular point locations. Further study would be needed to investigate these applications, both in terms of their behavior in these settings and their feasibility in terms of computational efficiency.

Spatial forecast verification is only one avenue in which the measures introduced here can be applied. For example, G could be especially useful as the loss function when predicting a spatial-exceedance region as in Zhang et al. (2008); Cressie and Suesse (2020).

Code and data availability. The R (R Core Team, 2018) programming language was used for all analyses carried out in this paper, especially for functions from the packages spatstat (Baddeley and Turner, 2005), fields (Nychka et al., 2017), maps (Becker et al., 2018), and SpatialVx (Gilleland, 2019). The ICP/MesoVICT data are all freely available at <https://ral.ucar.edu/projects/icp/> (last access: 11 February 2021). The summary measures proposed here are now available as part of the SpatialVx package beginning with version 0.7-1.

Competing interests. The author declares that there is no conflict of interest.

Acknowledgements. This work was sponsored in part by the National Center for Atmospheric Research (NCAR).

Review statement. This paper was edited by Dan Cooley and reviewed by Manfred Dorninger and one anonymous referee.

References

Ahijevych, D., Gilleland, E., Brown, B. G., and Ebert, E. E.: Application of spatial verification methods to idealized and NWP-gridded precipitation forecasts, *Weather Forecast.*, 24, 1485–1497, 2009.

- Baddeley, A. J.: Errors in binary images and an L_p version of the Hausdorff metric, *Nieuw Arch. Wiskunde*, 10, 157–183, 1992a.
- Baddeley, A. J.: An error metric for binary images, in: *Robust Computer Vision Algorithms*, edited by: Forstner, W. and Ruwiedel, S., Wichmann, 402 pp., 59–78, available at: <https://pdfs.semanticscholar.org/aa50/669b71429f2ca54d64f93839a9da95ceba6b.pdf> (last access: 14 May 2020), 1992b.
- Baddeley, A. and Turner, R.: spatstat: An R Package for Analyzing Spatial Point Patterns, *J. Stat. Softw.*, 12, 1–42, available at: <http://www.jstatsoft.org/v12/i06/> (last access: 11 February 2021), 2005.
- Baldwin, M. E. and Elmore, K. L.: Objective verification of high-resolution WRF forecasts during 2005 NSSL/SPC Spring Program, in: *21st Conf. on Weather Analysis and Forecasting/17th Conf. on Numerical Weather Prediction*, Amer. Meteor. Soc., Washington, D.C., 11B4, 2005.
- Becker, R. A., Wilks, A. R., Brownrigg, R., Minka, T. P., and Deckmyn, A.: maps: Draw Geographical Maps, R package version 3.3.0, available at: <https://CRAN.R-project.org/package=maps> (last access: 11 February 2021), 2018.
- Brown, B. G., Gilleland, E., and Ebert, E. E.: *Forecasts of Spatial Fields*, John Wiley and Sons, Ltd., 95–117, 2011.
- Brunet, D. and Sills, D.: A generalized distance transform: Theory and applications to weather analysis and forecasting, *IEEE T. Geosci. Remote. Sens.*, 55, 1752–1764, <https://doi.org/10.1109/TGRS.2016.2632042>, 2017.
- Brunet, D., Sills, D., and Casati, B.: A spatio-temporal user-centric distance for forecast verification, *Meteorol. Z.*, 27, 441–453, <https://doi.org/10.1127/metz/2018/0883>, 2018.
- Casati, B., Ross, G., and Stephenson, D. B.: A new intensity-scale approach for the verification of spatial precipitation forecasts, *Meteorol. Appl.*, 11, 141–154, <https://doi.org/10.1017/S1350482704001239>, 2004.
- Cressie, N. and Suesse, T.: Great expectations and even greater exceedances from spatially referenced data, *Spat. Stat.-Neth.*, 37, 100420, <https://doi.org/10.1016/j.spasta.2020.100420>, 2020.
- Davis, C., Brown, B., and Bullock, R.: Object-Based Verification of Precipitation Forecasts. Part I: Methodology and Application to Mesoscale Rain Areas, *Mon. Weather Rev.*, 134, 1772–1784, 2006a.
- Davis, C., Brown, B., and Bullock, R.: Object-Based Verification of Precipitation Forecasts. Part II: Application to Convective Rain Systems, *Mon. Weather Rev.*, 134, 1785–1795, 2006b.
- Davis, C. A., Brown, B. G., Bullock, R., and Halley-Gotway, J.: The Method for Object-Based Diagnostic Evaluation (MODE) Applied to Numerical Forecasts from the 2005 NSSL/SPC Spring Program, *Weather Forecast.*, 24, 1252–1267, 2009.
- Dorninger, M., Mittermaier, M. P., Gilleland, E., Ebert, E. E., Brown, B. G., and Wilson, L. J.: MesoVICT: Mesoscale Verification Inter-Comparison over Complex Terrain, Tech. rep., No. NCAR/TN-505+STR, <https://doi.org/10.5065/D6416V21>, 2013.
- Dorninger, M., Gilleland, E., Casati, B., Mittermaier, M. P., Ebert, E. E., Brown, B. G., and Wilson, L.: The set-up of the Mesoscale Verification Inter-Comparison over Complex Terrain project, *B. Am. Meteorol. Soc.*, 99, 1887–1906, <https://doi.org/10.1175/BAMS-D-17-0164.1>, 2018.

- Ebert, E. E.: Fuzzy verification of high-resolution gridded forecasts: a review and proposed framework, *Meteorol. Appl.*, 15, 51–64, 2008.
- Ebert, E. E. and McBride, J. L.: Verification of precipitation in weather systems: determination of systematic errors, *J. Hydrol.*, 239, 179–202, [https://doi.org/10.1016/S0022-1694\(00\)00343-7](https://doi.org/10.1016/S0022-1694(00)00343-7), 2000.
- Gilleland, E.: Spatial Forecast Verification: Baddeley's Delta Metric Applied to the ICP Test Cases, *Weather Forecast.*, 26, 409–415, 2011.
- Gilleland, E.: Testing Competing Precipitation Forecasts Accurately and Efficiently: The Spatial Prediction Comparison Test, *Mon. Weather Rev.*, 141, 340–355, 2013.
- Gilleland, E.: A New Characterization within the Spatial Verification Framework for False Alarms, Misses, and Overall Patterns, *Weather Forecast.*, 32, 187–198, 2017.
- Gilleland, E.: SpatialVx: Spatial Forecast Verification, R package version 0.6-6, available at: <http://www.ral.ucar.edu/projects/icp/SpatialVx> (last access: 11 February 2021), 2019.
- Gilleland, E., Ahijevych, D., Brown, B. G., Casati, B., and Ebert, E. E.: Intercomparison of Spatial Forecast Verification Methods, *Weather Forecast.*, 24, 1416–1430, 2009.
- Gilleland, E., Lindström, J., and Lindgren, F.: Analyzing the image warp forecast verification method on precipitation fields from the ICP, *Weather Forecast.*, 25, 1249–1262, 2010.
- Gilleland, E., Skok, G., Brown, B. G., Casati, B., Dorninger, M., Mittermaier, M. P., Roberts, N., and Wilson, L. J.: A novel set of verification test fields with application to distance measures, *Mon. Weather Rev.*, 148, 1653–1673, <https://doi.org/10.1175/MWR-D-19-0256.1>, 2020.
- Kain, J. S., Weiss, S. J., Bright, D. R., Baldwin, M. E., Levit, J. J., Carbin, G. W., Schwartz, C. S., Weisman, M. L., Droege-meier, K. K., Weber, D. B., and Thomas, K. W.: Some practical considerations regarding horizontal resolution in the first generation of operational convection-allowing NWP, *Weather Forecast.*, 23, 931–952, <https://doi.org/10.1175/WAF2007106.1>, 2008.
- Keil, C. and Craig, G.: A displacement-based error measure applied in a regional ensemble forecasting system, *Mon. Weather Rev.*, 135, 3248–3259, 2007.
- Keil, C. and Craig, G.: A displacement and amplitude score employing an optical flow technique, *Weather Forecast.*, 24, 1297–1308, 2009.
- Koch, J., Jensen, K. H., and Stisen, S.: Toward a true spatial model evaluation in distributed hydrological modeling: Kappa statistics, Fuzzy theory, and EOF-analysis benchmarked by the human perception and evaluated against a modeling case study, *Water Resour. Res.*, 51, 1225–1246, <https://doi.org/10.1002/2014WR016607>, 2015.
- Koch, J., Siemann, A., Stisen, S., and Sheffield, J.: Spatial validation of large-scale land surface models against monthly land surface temperature patterns using innovative performance metrics, *J. Geophys. Res.-Atmos.*, 121, 5430–5452, <https://doi.org/10.1002/2015JD024482>, 2016.
- Koch, J., Demirel, M. C., and Stisen, S.: The SPAtial EFficiency metric (SPAEF): multiple-component evaluation of spatial patterns for optimization of hydrological models, *Geosci. Model Dev.*, 11, 1873–1886, <https://doi.org/10.5194/gmd-11-1873-2018>, 2018.
- Kullback, S. and Leibler, R. A.: On information and sufficiency, *Ann. Math. Stat.*, 22, 79–86, <https://doi.org/10.1214/aoms/117729694>, 1951.
- Marzban, C. and Sandgathe, S.: Optical flow for verification, *Weather Forecast.*, 25, 1479–1494, <https://doi.org/10.1175/2010WAF2222351.1>, 2010.
- Mass, C. F., Ovens, D., Westrick, K., and Colle, B. A.: Does Increasing Horizontal Resolution Produce More Skillful Forecasts?, *B. Am. Meteorol. Soc.*, 83, 407–430, [https://doi.org/10.1175/1520-0477\(2002\)083<0407:DIHRPM>2.3.CO;2](https://doi.org/10.1175/1520-0477(2002)083<0407:DIHRPM>2.3.CO;2), 2002.
- Meijster, A., Roerdink, J. B. T. M., and Hesselink, W. H.: A General Algorithm for Computing Distance Transforms in Linear Time, in: *Mathematical Morphology and its Applications to Image and Signal Processing*, edited by: Goutsias, J., Vincent, L., Bloomberg, D. S., Computational Imaging and Vision, Springer, Boston, MA, 18, https://doi.org/10.1007/0-306-47025-X_36, 2000.
- Mittermaier, M. P.: A strategy for verifying near-convection resolving model forecasts at observing sites, *Weather Forecast.*, 29, 185–204, 2014.
- Nychka, D., Furrer, R., Paige, J., and Sain, S.: fields: Tools for spatial data, R package version 10.0, <https://doi.org/10.5065/D6W957CT>, available at: <https://github.com/NCAR/Fields> (last access: 11 February 2021), 2017.
- Peli, T. and Malah, D.: A study on edge detection algorithms, *Comput. Vision Graph.*, 20, 1–21, 1982.
- Pratt, W. K.: *Digital Image Processing*, 4th edn., John Wiley and Sons, Inc., Hoboken, New Jersey, 782 pp., 2007.
- R Core Team: R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, available at: <https://www.R-project.org/> (last access: 11 February 2021), 2018.
- Rezacova, D., Sokol, Z., and Pesice, P.: A radar-based verification of precipitation forecast for local convective storms, *Atmos. Res.*, 83, 221–224, <https://doi.org/10.1016/j.atmosres.2005.08.011>, 2007.
- Roberts, N. M. and Lean, H. W.: Scale-Selective Verification of Rainfall Accumulations from High-Resolution Forecasts of Convective Events, *Mon. Weather Rev.*, 136, 78–97, 2008.
- Rossa, A. M., Nurmi, P., and Ebert, E. E.: Overview of methods for the verification of quantitative precipitation forecasts, edited by: Michaelides, S. C., Springer, New York, NY, USA, 418–450, 2008.
- Skamarock, W. C., Klemp, J. B., Dudhia, J., Gill, D. O., Barker, M., Wang, W., and Powers, J. G.: A description of the Advanced Research WRF version 2, Ncar tech. rep. ncar/tn-468+str, Boulder, CO, USA, 2005.
- Steinacker, R., Häberli, C., and Pötschacher, W.: A transparent method for the analysis and quality evaluation of irregularly distributed and noisy observational data, *Mon. Weather Rev.*, 128, 2303–2316, 2000.
- Venugopal, V., Basu, S., and Foufoula-Georgiou, E.: A new metric for comparing precipitation patterns with an application to ensemble forecasts, *J. Geophys. Res.*, 110, D08111, <https://doi.org/10.1029/2004JD005395>, 2005.
- Wealands, S. R., Grayson, R. B., and Walker, J. P.: Quantitative comparison of spatial fields for hydrological model assessment

- some promising approaches, *Adv. Water Resour.*, 28, 15–32, 2005.
- Weniger, M., Kapp, F., and Friederichs, P.: Spatial Verification Using Wavelet Transforms: A Review, *Q. J. Royal Meteor. Soc.*, 143, 120–136, <https://doi.org/10.1002/qj.2881>, 2016.
- Wikle, C. K., Zammit-Mangion, A., and Cressie, N.: *Spatio-Temporal Statistics with R*, Chapman and Hall/CRC, Boca Raton, Florida, USA, 396 pp., available at: <https://spacetimewithr.org> (last access: 11 February 2021), 2019.
- Zhang, J., Craigmole, P. F., and Cressie, N.: Loss function approaches to predict a spatial quantile and its exceedance region, *Technometrics*, 50, 216–227, <https://doi.org/10.1198/004017008000000226>, 2008.
- Zinner, T., Mannstein, H., and Taferner, A.: Cb-TRAM: Tracking and monitoring severe convection from onset over rapid development to mature phase using multi-channel Meteosat-8 SEVIRI data, *Meteorol. Atmos. Phys.*, 101, 191–210, <https://doi.org/10.1007/s00703-008-0290-y>, 2008.