Comparing the distance-based methods

Eric Gilleland

National Center for Atmospheric Research, Boulder, Colorado, U.S.A.

Introduction

The distance-based methods are primarily aimed at informing about location errors and are typically applied to a binary field that is usually attained through a thresholding process. Each measure has its idiosyncrasies and favor different types of errors over others. Some may provide complementary information, and the aim for this discussion is to get a sense of how they compare with each other. To that end, the geometric cases from the first spatial forecast verification Inter-Comparison Project (ICP, https://ral.ucar.edu/projects/icp/; Ahijevych et al. 2009; Gilleland et al. 2009;2010) are utilized.

A brief description of each method is provided in Table 1. Note that some methods have selectable parameters, such as Baddeley's Δ, the generalized performance metric (henceforth denoted G), Pratt's figure of Merit (henceforth, F), and Zhu's measure (henceforth, Z). In the table, however, these parameters have been replaced with the values used in this comparison, which are also the most frequently applied choices.

Pay particular attention to the grid points, $\boldsymbol{s}$, over which the summary measures may be summed or maximized. For example, the Hausdorff distance is a maximum between two maximum distances. The first distance is maximized over all of the (shortest) distances from only those $\boldsymbol{s}$ that fall in the set B, while the second is maximized over all the distances from only the $\boldsymbol{s}$ that fall inside the set A. Figure 1 demonstrates Hausdorff distance through an illustration. At each $\boldsymbol{s}$ that falls within one of the two sets, the distance from that grid point is taken to every grid point in the other set. For a grid point in the set B, the shortest of the values is denoted by $d(\boldsymbol{s}, A)$, and for an $\boldsymbol{s}$ in A, it is denoted $d(\boldsymbol{s}, B)$. Then, $d(\boldsymbol{s}, A)$ and $d(\boldsymbol{s}, B)$ are found for each grid point in the two sets. The maximum of these resulting distances is the Hausdorff distance. It is clearly extremely sensitive to small changes in the domain, depending on where those changes occur.

For $M(A, B)$, the average of the distances from $\boldsymbol{s}$ in the set B to those in the set A are taken. That is, $d(\boldsymbol{s}, A)$ is averaged over all of the $\boldsymbol{s}$ that fall in the set B; similarly, for $F(A, B)$. On the other hand, the summation in Baddeley's Δ is over all $\boldsymbol{s}$ in the entire domain. As a result, Δ is slightly sensitive to the location of the sets within the domain.

G is a measure that multiplies the lack of overlap between A and B by the average distances of this lack of overlap, but where each average distance $d(\boldsymbol{s}, A)$ and $d(\boldsymbol{s}, B)$ is first tamed by the relative size of

each set. It is therefore the only distance-based measure that gives reasonable values when faced with empty fields and nearly empty fields.

Table 1: Equations for each of the distance-based measures compared here. Let $\boldsymbol{s} = (x, y) \in \mathcal{D}$ represent a grid point (coordinate) in the domain $\mathcal{D}$, N be the size of the domain with $A, B \subset \mathcal{D}$ representing sets of grid points whose corresponding value is one (in the binary field). Then let $d(\boldsymbol{s}, A)$ be the shortest distance from $\boldsymbol{s}$ to A, and similarly for $d(\boldsymbol{s}, B)$. If a field is empty of one-valued grid points, define $d(\boldsymbol{s}, A) = D$ for some large value D, such as N. Further, let n_A and n_B represent the number of grid points in the sets A and B, respectively, and let n_{AB} represent the number of grid points in both sets. Further, let $p_A = \frac{n_A}{N}$ and $p_B = \frac{p_B}{N}$, $I_A(\boldsymbol{s}) = 1$ if $\boldsymbol{s} \in A$ and zero otherwise, similarly for $I_B(\boldsymbol{s})$.

Method Name	**Method Equation**
Hausdorff distance (grid points)	$H(A,B) = \max\left\{\max_{\boldsymbol{s}\in B}[d(\boldsymbol{s},A)], \max_{\boldsymbol{s}\in A}[d(\boldsymbol{s},B)]\right\}$
Baddeley's Δ	$\Delta(A,B) = \left[\frac{1}{N}\sum_{\boldsymbol{s}\in\mathcal{D}}\{d(\boldsymbol{s},A) + d(\boldsymbol{s},B)\}^2\right]^{1/2}$
Mean-error distance	$M(A,B) = \frac{1}{n_B}\sum_{\boldsymbol{s}\in B} d(\boldsymbol{s},A)$
G_β	$G_\beta(A,B) = \max\left\{1 - \frac{2}{N}(n_A + n_B - 2n_{AB})(M(A,B)p_B + M(B,A)p_A), 0\right\}$
Pratt's Figure of Merit	$F(A,B) = \frac{1}{\max\{n_A, n_B\}}\sum_{\boldsymbol{s}\in B}\frac{1}{1 + d^2(\boldsymbol{s},A)/9}$
Zhu's measure	$Z(A,B) = \frac{1}{2}\sqrt{\sum_{\boldsymbol{s}\in\mathcal{D}}\left(I_A(\boldsymbol{s}) - I_B(\boldsymbol{s})\right)^2} + \frac{1}{2}M(A,B)$

Zhu's measure, $Z(A,B)$, averages the root-mean-square error (RMSE) between the two binary fields with $M(A,B)$. Subsequently, none of $M(A,B)$, $F(A,B)$ and $Z(A,B)$ is symmetric in the sense that

$M(A,B) \neq M(B,A)$. Gilleland (2017) argued that this lack of symmetric can be exploited to inform about misses v. false alarm types of errors.

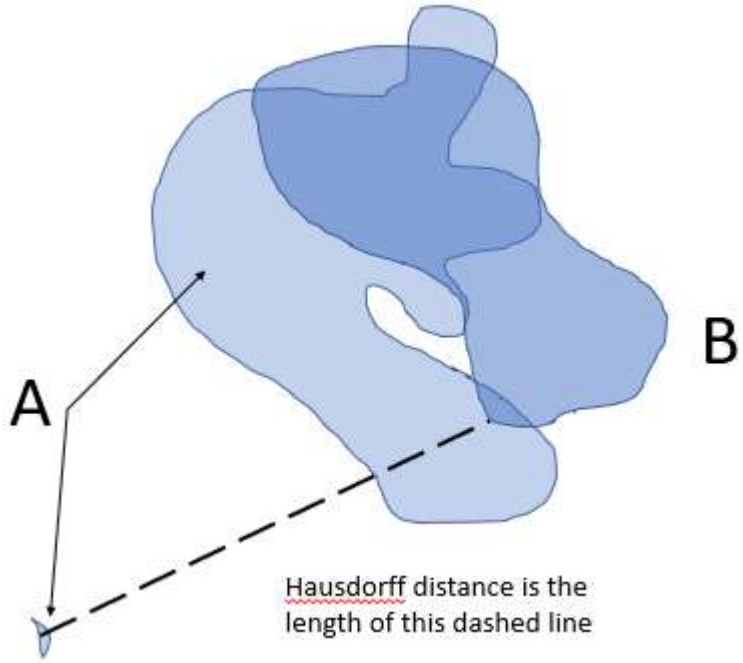

Figure 1: Two binary fields have been superimposed onto each other. The areas where the fields are one-valued, respectively, are labeled A and B. The set A consists of two isolated clusters (connected components) whereas B is just one such blob. If the smaller isolated component from A were removed, the Hausdorff distance would be much smaller.

Rankings

It is informative to test the various methods by ranking the set of geometric test cases from the ICP. A ranking from best to worst by each method should shed light on how they inform about the closeness between two spatial (binary) fields.

Table 2 summarizes the rankings over all measures and cases with some color-coding to assist in seeing the differences in rankings. Figure 2 to Figure 7 show the rankings by each measure. Where rankings are identical, in both the tables and the figures, the results are combined. Table 2 also summarizes what types of errors are represented by each case both pictorially and with words. How a subjective evaluator would rank the cases may vary, and may depend on the particular application. So, there is no right or wrong set of rankings, but some may agree/disagree more with most subjective evaluators.

Table 2: Rankings of each geometric case for each method summarized. Translation errors are given by numbers of grid points and denoted by pts.

	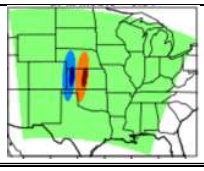	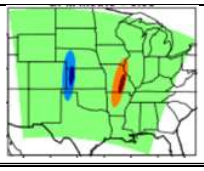	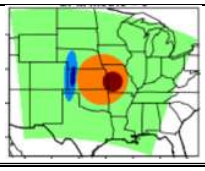	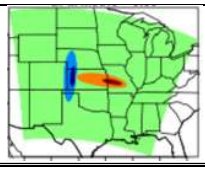	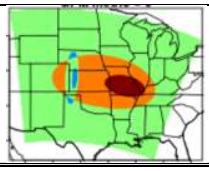
Method	geom001 translation-only error (50-pts)	geom002 translation-only (200-pts)	geom003 translation (125-pts) and large area bias	geom004 translation (125-pts) and aspect-ratio	geom005 translation (125-pts) and huge area bias (but overlapping)
$H(A,B)$	**Best**	**Tied for 2**	**Tied for 2**	**Tied for 2**	**Worst**
$G(A,B)$	**Best**	**3 (near tie for worst)**	**Tied for worst**	**2**	**Tied for worst**
$M(A,B)$ and $Z(A,B)$ Miss	**2 (near-tie with 3)**	**Worst**	**3 (near tie with 2)**	**4**	**Best**
$M(A,B)$ and $Z(A,B)$ False Alarm	**Best**	**Worst**	**3 (near tie with 2)**	**2 (near tie with 3)**	**4**
$F(A,B)$ Miss and False Alarm	**2**	**Worst**	**4**	**3**	**Best**
$\Delta(A,B)$	**Best**	**Worst**	**3**	**2**	**4**

Figure 2 shows the five geometric cases ranked according to G. This measure is a unitless measure that falls between zero (very bad forecast) to one (perfect forecast). Technically, the user can select a parameter to determine how quickly the measure decreases to zero, and how bad the forecast has to be before it reaches zero. However, Gilleland (2021) found that half the domain size for this parameter yields meaningful results across several different domains. Therefore, half the domain size is used here.

The first geometric case is usually considered the best by the subjective observer, but some users might prefer the fifth case because it is the only one that actually overlaps with the observations. G gives a fairly high mark (about 0.84) to the first case and zero to the fifth case. The scaling error, which was intended to be a rotational error, is the second-best case according to G, favoring the proximity and overall, less false alarm areas of the forecast than the two larger-area forecasts. Nevertheless, 0.59 is a relatively bad mark, indicating that the forecast is poor. The next best case yields a $G \approx 0$, so while it favors this thin but far away case over the two large forecast cases, it is nearly tied with them for worst. Unless any overlap is highly coveted, this measure gives the most reasonable results. In general, this measure does reward for overlap, but it also penalizes for lack of overlap, and there is too much lack of overlap in this geometric test case to warrant even a better subjective interpretation.

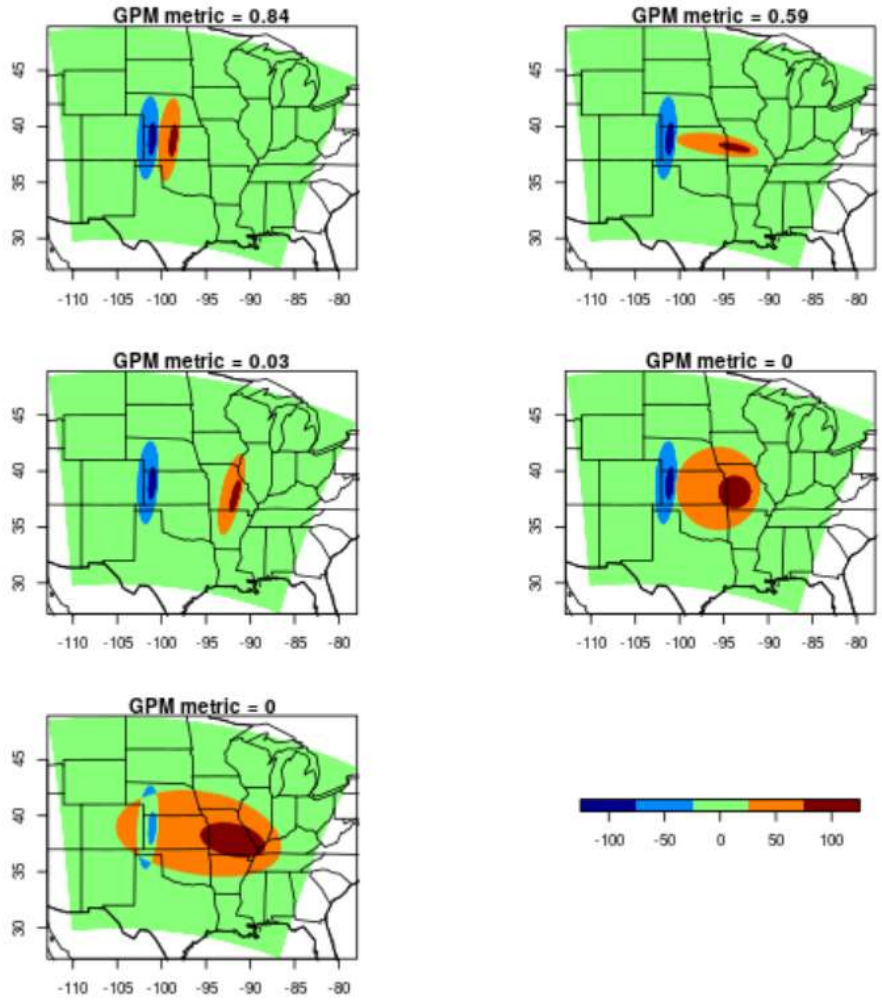

Figure 2: ICP geometric test cases ranked from best (top left) to worst (tie between middle right and bottom left) according to G_β from Gilleland (2021) using half the domain size for the β parameter (as shown in Table 1). Note that middle left is nearly as bad according to G_β (for this choice of β) as the worst two cases. Each geometric case consists of a small (higher intensity) ellipse surrounded by a larger (lower intensity) ellipse. The "observed" set is the blue ellipse in each field and the orange is the forecast (shown is F – O). The measure is calculated for the binary fields derived by thresholding only the values above zero; so the entire areas encompassing the larger ellipses.

The Hausdorff distance arguably gives the next best rankings of these cases (Figure 3). H falls between zero (best) and infinity (increasing implies worsening). It chooses the same best and worst cases as G but gives a tie to middle three cases. Some issues with this metric, however, include the aforementioned sensitivity to small changes in the field (e.g., Figure 1) and the related fact that it does not provide sensible information for the empty-field case or small frequency cases (Gilleland et al. 2020). Again, G does not suffer from any of these failings.

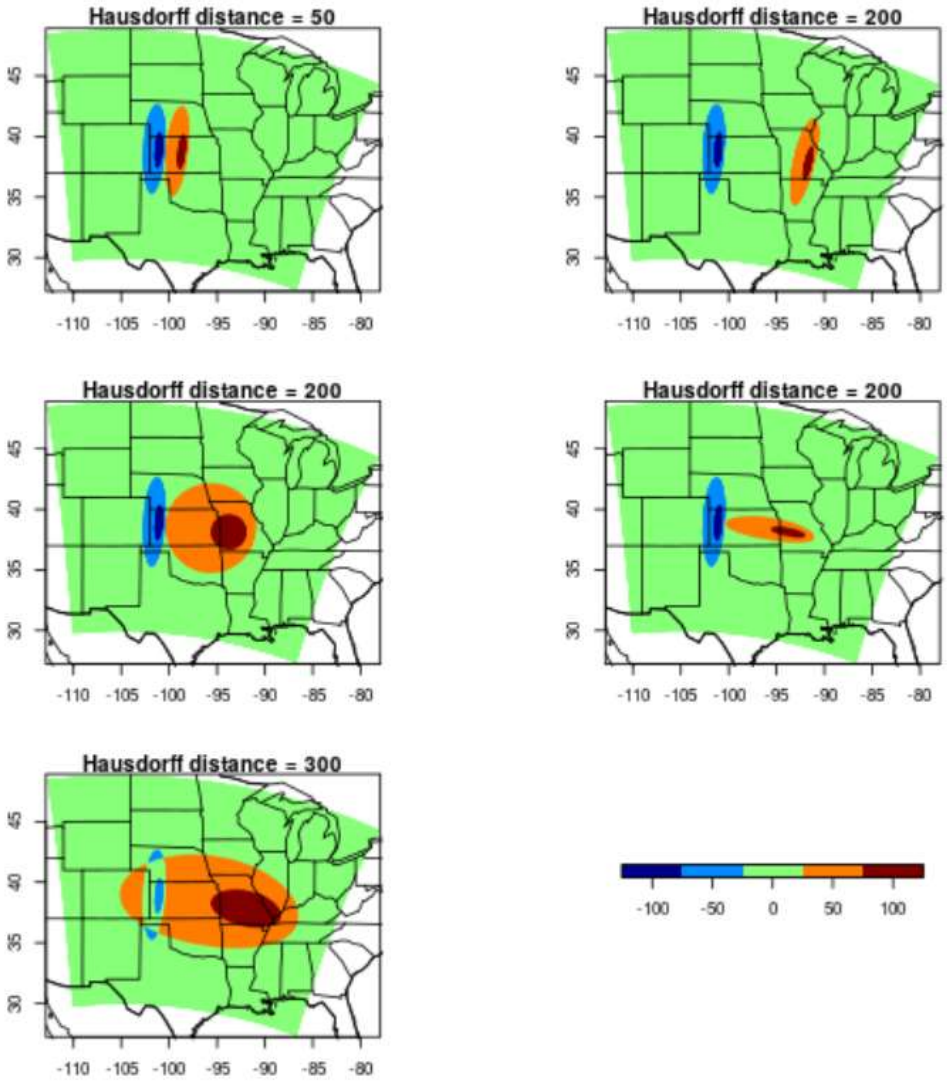

Figure 3: Same as Figure 2 but ranked according to the Hausdorff distance (units are number of grid points).

Figure 4 shows the rankings for the MED and Zhu's measure for the "miss" version; that is, it gives an average distance from the "observed" ellipse to the "forecast" ellipse. In this case, the rankings are identical because Zhu's measure is just an average of the RMSE and the MED, so it is not surprising that they rank the cases identically. In this case, they both greatly prefer the case that has a huge forecast bias because it is the only one that overlaps with the "observed" ellipse and from the perspective of the observation, that would mean it is the best. For both of these measures, it is important to also look at the false alarm versions (Figure 5). Now, the huge bias case, while not considered the worst by either measure, is fairly low in the rankings (ranked second worst to the huge translation-only error).

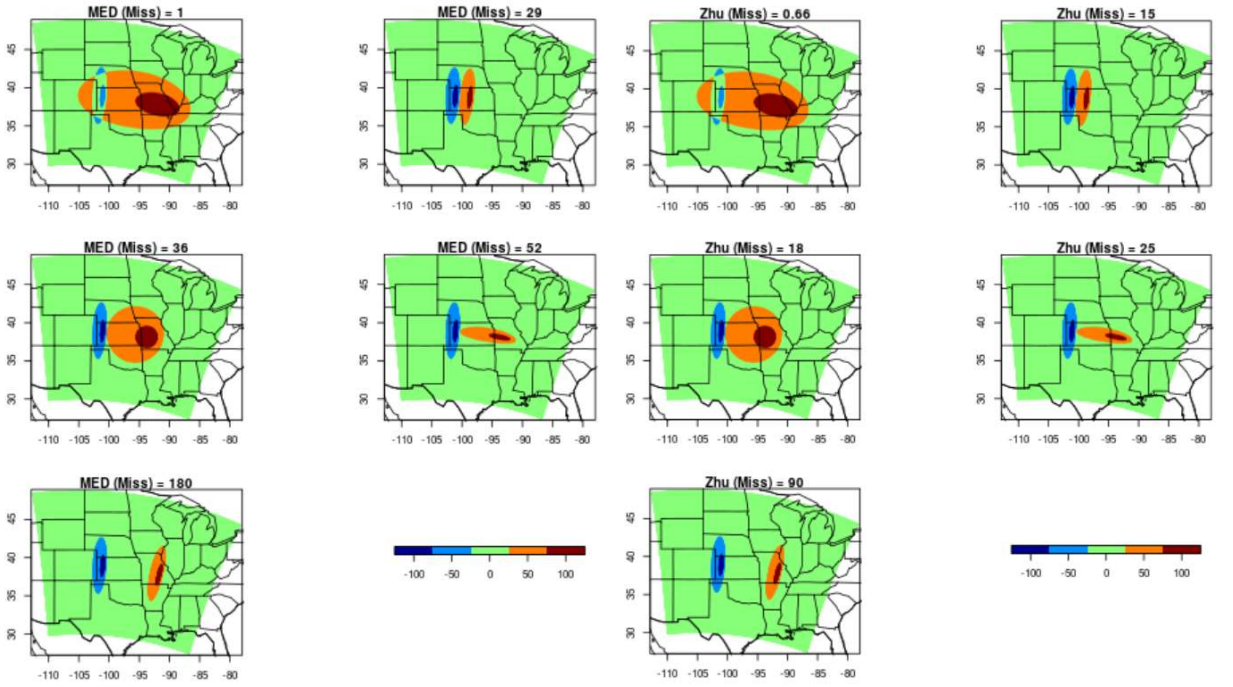

Figure 4: Same as Figure 2 but for MED (Miss, left) and Zhu (Miss, right). Lower values for both measures mean better matches between the two ellipses.

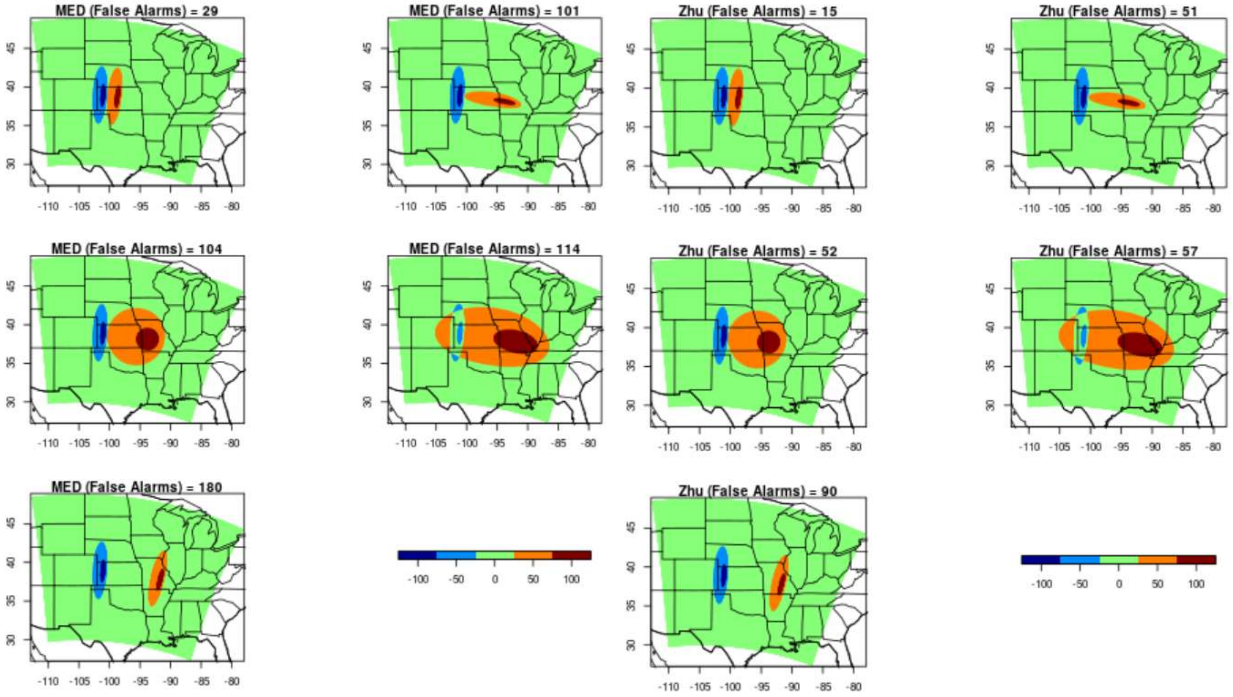

Figure 5: Same as Figure 4 but for the false alarm versions of the measures.

FoM also has a "miss" and "false alarm" version, but nevertheless ranks these geometric cases identically (Figure 6). Its rankings are almost identical to those of the "miss" version of MED and Zhu.

Only one pair of rankings are switched (the third and fourth). It is unfortunate that it ranks the cases this way for both versions because it is a questionable ranking that does not shed any additional light on the performance of each forecast.

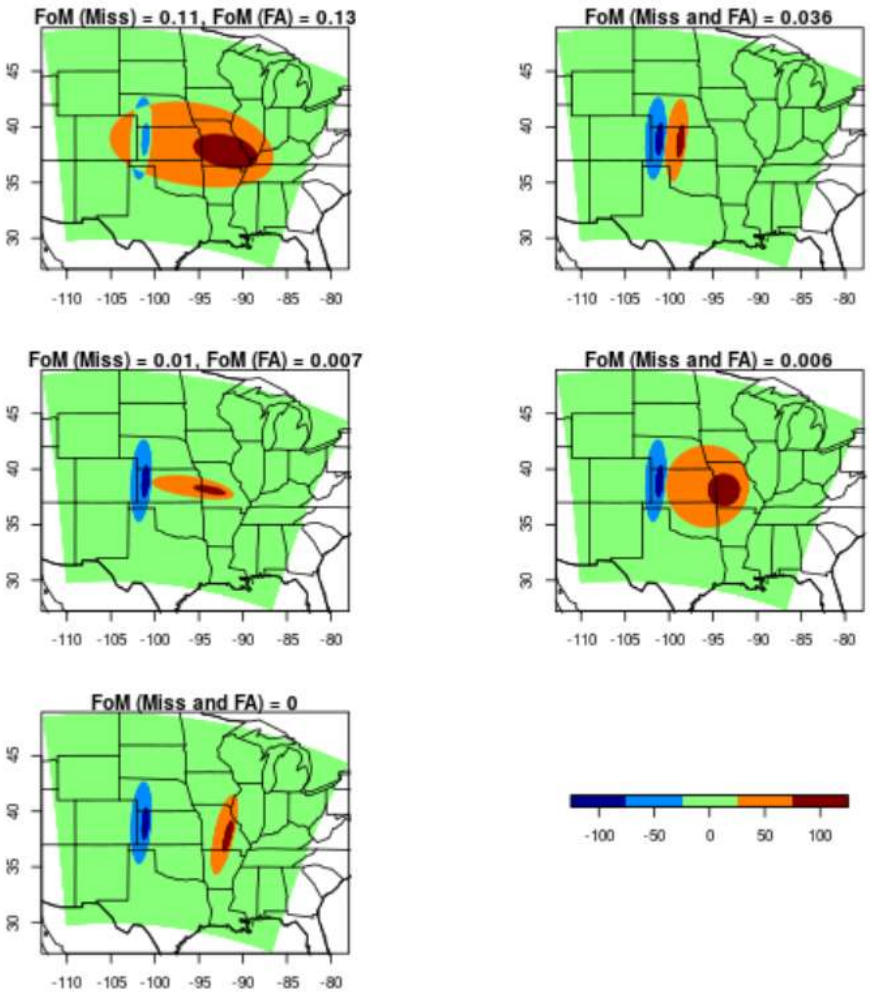

Figure 6: Same as Figure 2 but for FoM. Both the "miss" and "false alarm" versions of this measure rank the cases identically. It gives a low value for each case.

Finally, Figure 7 displays the results for Baddeley's Δ metric. The rankings are sensible, but it is strange that the values differ so much for three of the worst cases, whereas the Hausdorff and G_β consider them more-or-less equally poor. It is also disappointing that the huge translation-only error should be considered much worse than the large and huge bias cases.

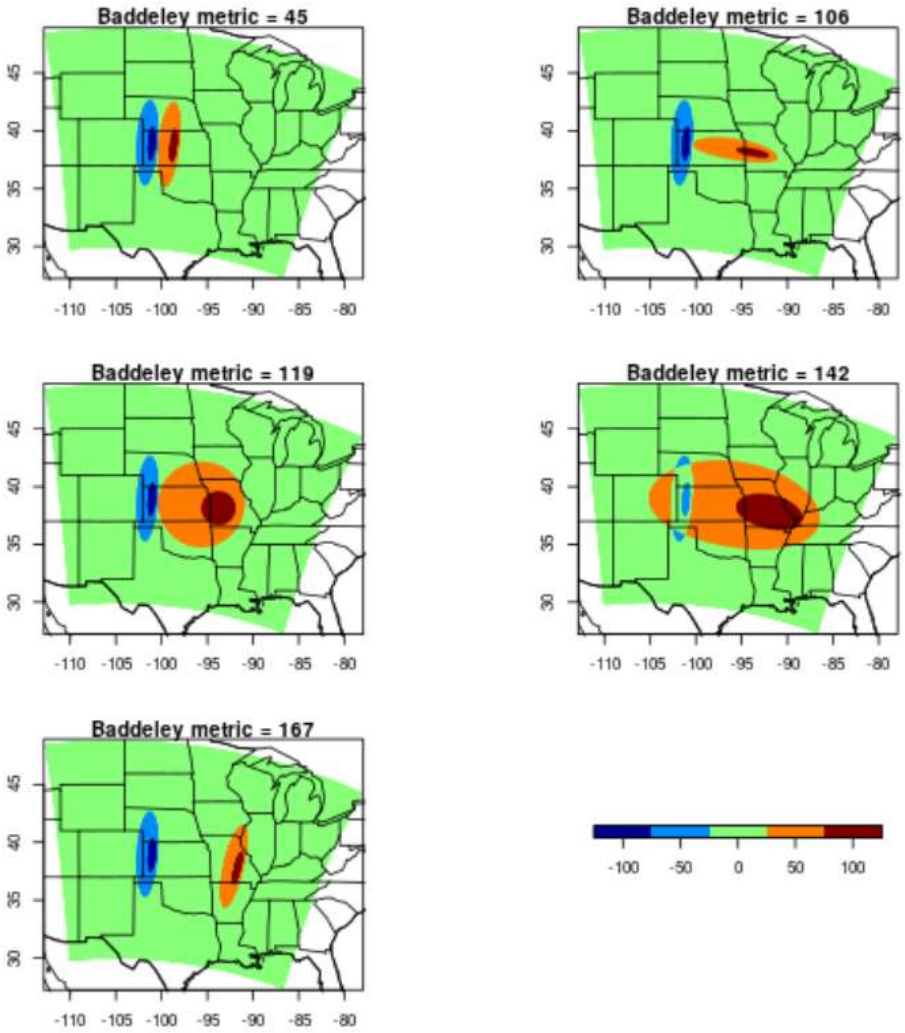

Figure 7: Same as Figure 2 but for Baddeley's Δ metric.

Fine v. Coarse Scale Forecast Performance

One of the major reasons behind the push for new verification methods was that subjectively better, high-resolution forecasts often had worse verification scores than subjectively poorer coarse-resolution counterparts. But do the new measures also give contradictory results in this sense? In this section, a single forecast is smoothed to a coarser resolution and each is compared with the corresponding original (finer scale) forecast by ranking the results of each distance-based measure across a range of thresholds.

Figure 8 shows one of the real cases from the ICP with one of the corresponding forecasts and a smoothed version of this forecast. The smoothing was carried out using a convolution smoothing technique with a disk-shaped kernel having radius of 25 grid points. This radius was chosen so as to not leave any doubt about which forecast model is "better" from a subjective standpoint. Clearly the coarse-resolution version is not as accurate as the original high-resolution model.

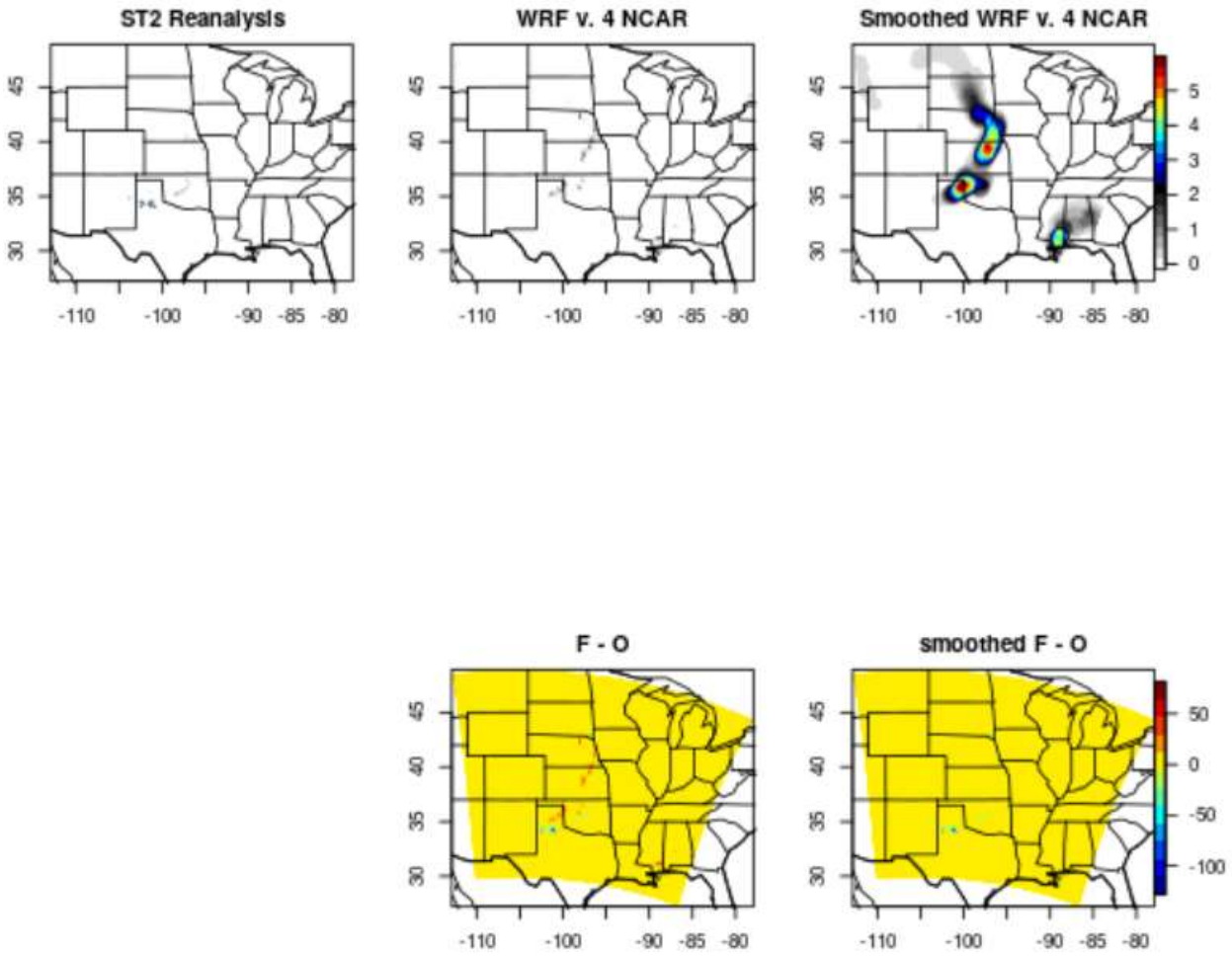

Figure 8: Stage II reanalysis ("observation", top left), WRF v. 4 NCAR (top middle) and a convolution-radius smoothing (with radius = 25 grid points) of WRF v. 4 NCAR (top right) for precipitation (mmh^{-1}) with 24-h lead time. The differences forecast minus "observation" for each competing forecast model (bottom row). Valid time is 1 June 2005 at 0000 UTC.

Table 3: Top table are results of the distance-based measures between the Stage II reanalysis and WRF 4 NCAR and bottom table are those for the smoothed version of WRF 4 NCAR. Coloring is to help visualize which are ranked better for the original v. smoothed forecasts. Yellow means better and red worse. Other colors indicate a near tie.

	$H(F_1,O)$	$\Delta(F_1,O)$	$G(F_1,O)$	$M(F_1,O)$	$M(O,F_1)$	$F(F_1,O)$	$F(O,F_1)$	$Z(F_1,O)$	$Z(O,F_1)$
>0.1	**89.73**	**26.58**	**0.61**	**5.24**	**5.03**	**0.63**	**0.54**	**2.69**	**2.59**
>1.1	**146.62**	**33.06**	**0.86**	**9.33**	**11.52**	**0.42**	**0.37**	**4.70**	**5.80**
>5.1	**344.42**	**51.00**	**0.98**	**14.59**	**21.85**	**0.11**	**0.16**	**7.30**	**10.94**

	$H(F_2,O)$	$\Delta(F_2,O)$	$G(F_2,O)$	$M(F_2,O)$	$M(O,F_2)$	$F(F_2,O)$	$F(O,F_2)$	$Z(F_2,O)$	$Z(O,F_2)$
>0.1	**282.46**	**52.92**	**0.15**	**5.06**	**6.05**	**0.43**	**0.57**	**2.62**	**3.12**
>1.1	**317.95**	**78.62**	**0.75**	**15.86**	**9.40**	**0.30**	**0.44**	**7.97**	**4.74**
>5.1	**360.17**	**131.60**	**0.99**	**62.63**	25.48	**0.00**	**0.01**	**31.32**	**12.75**

Table 3 shows the results of the distance-based measures for the forecast (top) and its smoothed version (bottom). Colors help to visualize which forecast is ranked better with yellow indicating better and red worse. Other colors mean that the values indicate a near tie. Figure 9 to Figure 11 show the corresponding distance maps for each threshold choice, which helps to visualize where precipitation amounts exceed the threshold. White (inside the domain, cf. Figure 8 bottom) indicates that the precipitation exceeded the threshold in that region and warmer colors indicate no precipitation in the vicinity. For the most part, the distance-based measures provide good evaluations in the sense that they prefer the more realistic and closer matched higher-resolution forecast. However, there are some discrepancies. Notably, asymmetric measures all indicate that from the forecast's perspective, the smoothed forecast is better, but by the same token, from the point of view of the observations, the higher-resolution forecast is better. The G_β is not a good measure for extreme events because it down weights fields that are emptier thereby giving a higher score. In this case, for the highest threshold, it slightly favors the smoother forecast.

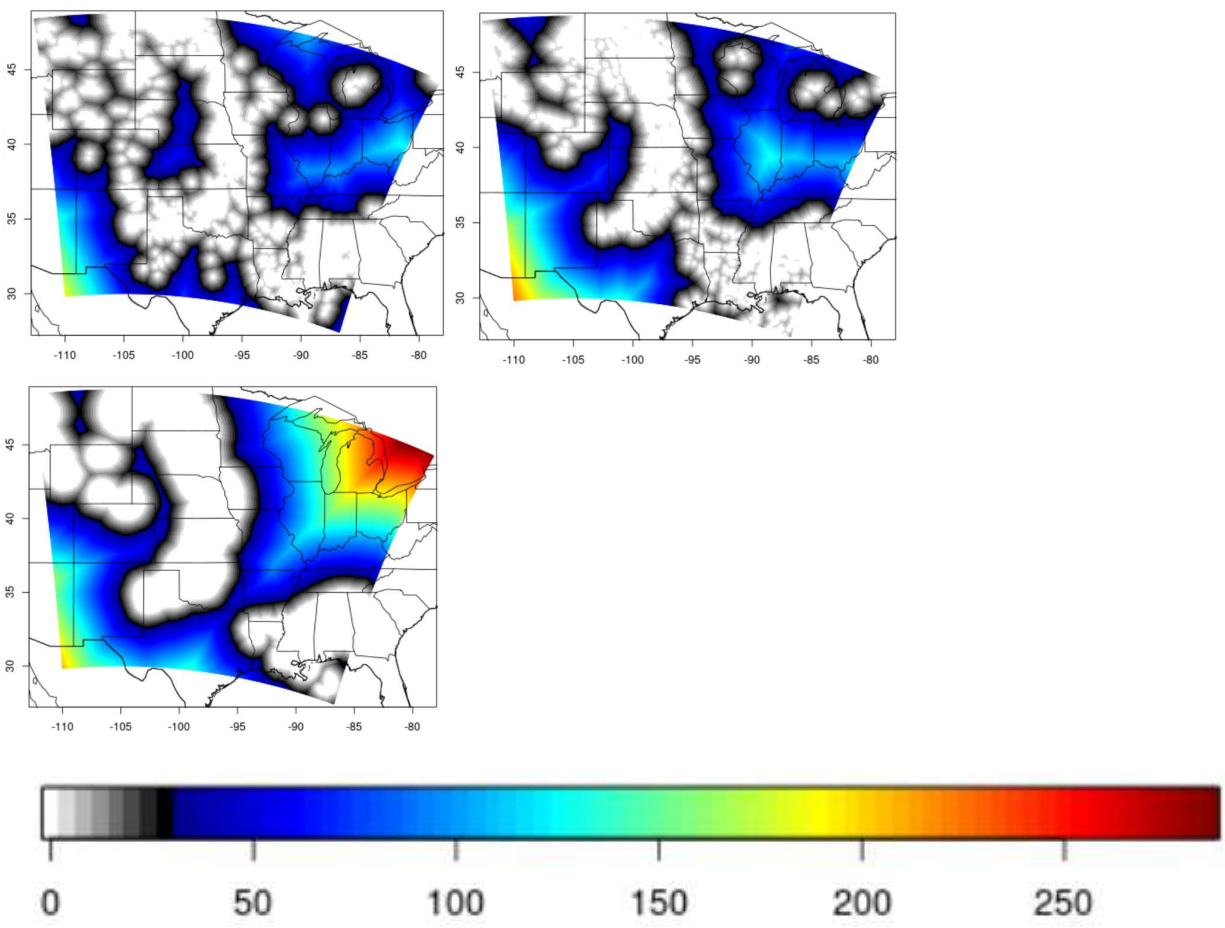

Figure 9: Distance maps for precipitation > 0.1 mmh^{-1} valid 1 June 2005. Top left is the Stage II reanalysis, top right is WRF 4 NCAR and bottom left is the smoothed version of WRF 4 NCAR.

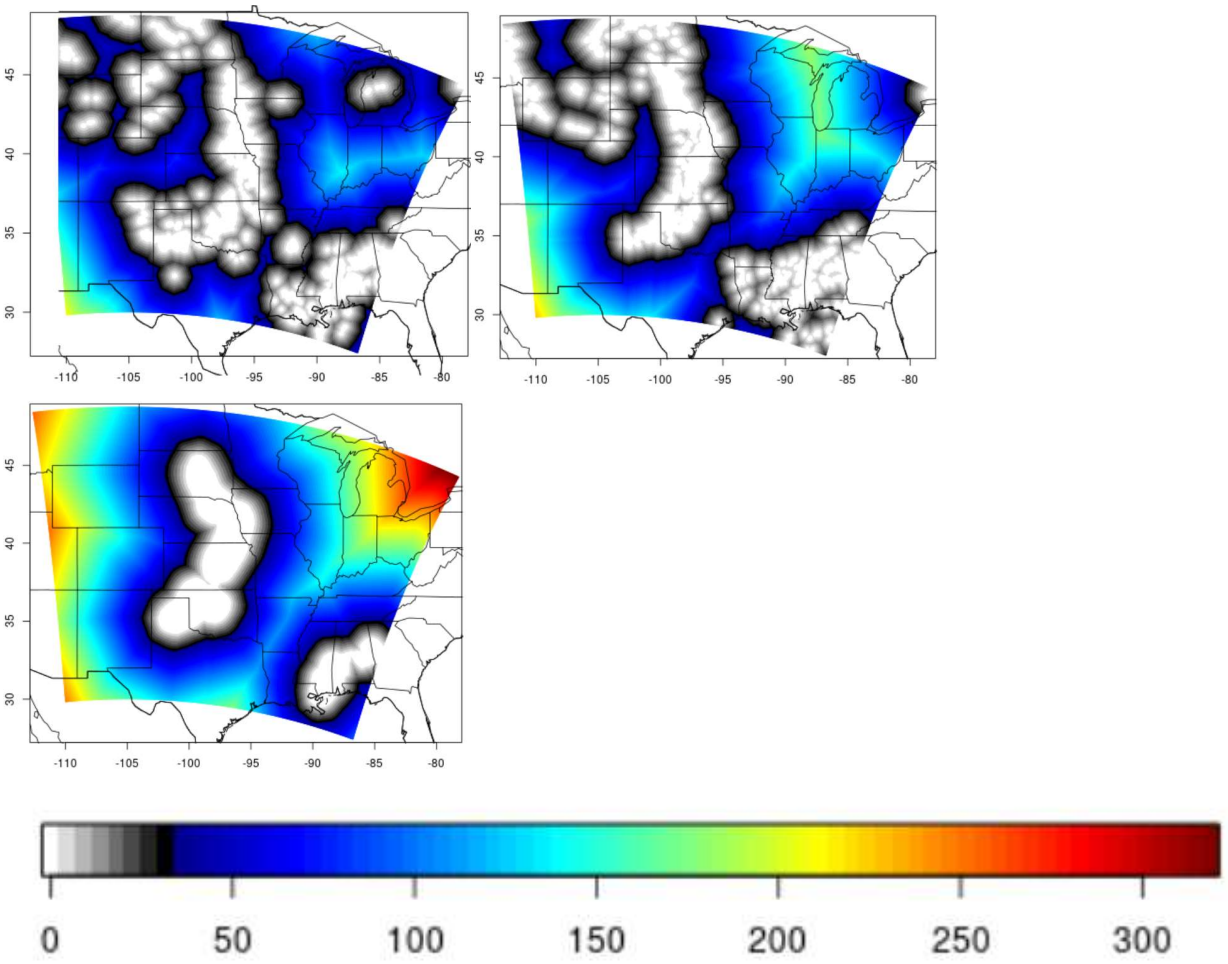

Figure 10: Same as Figure 9 but for precipitation > 1.1 mmh^{-1}.

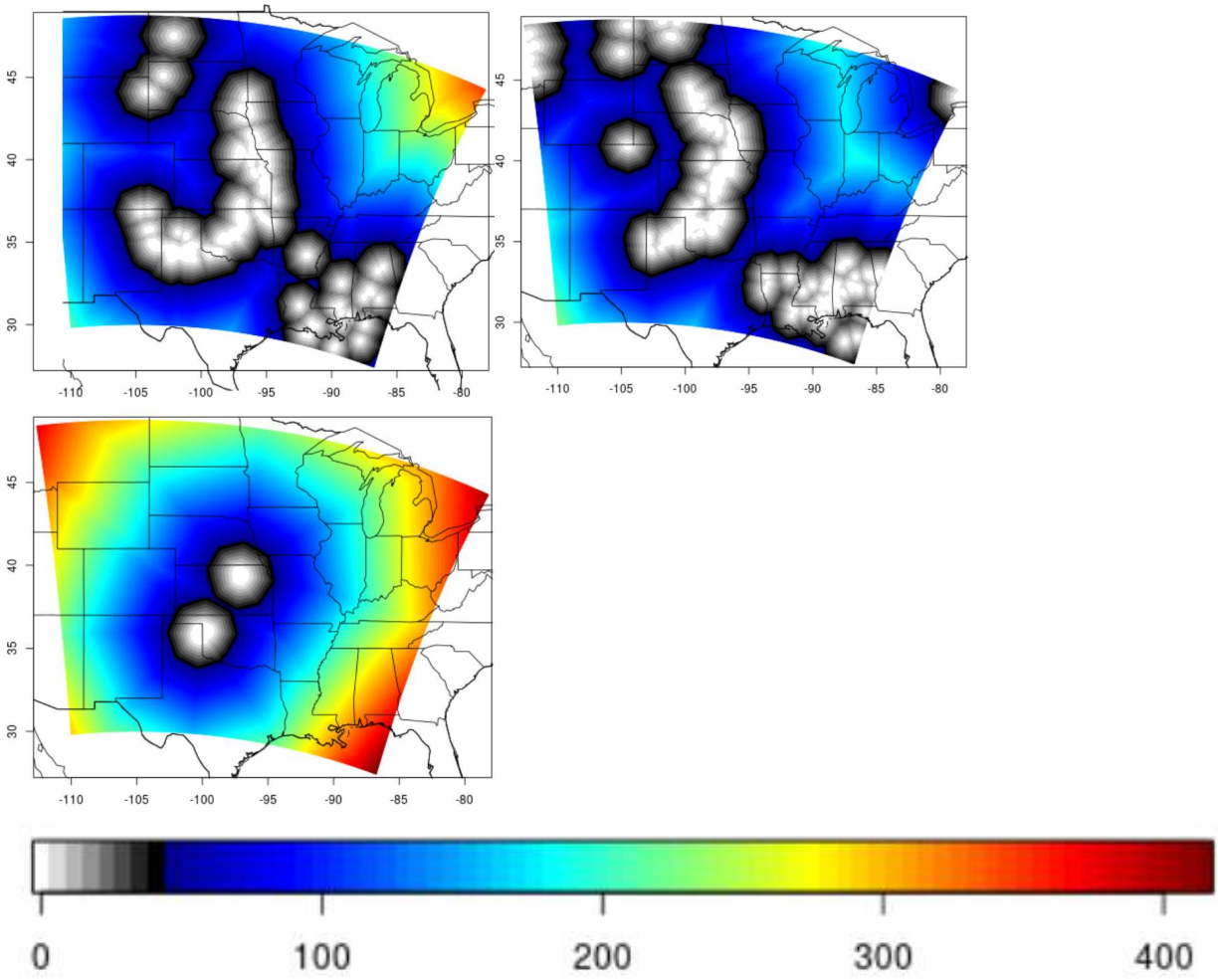

Figure 11: Same as Figure 9 but for precipitation > 5.1 mmh^{-1}.

Acknowledgments

This work was sponsored in part by the National Center for Atmospheric Research (NCAR), which is a major facility sponsored by the National Science Foundation under Cooperative Agreement No. 1852977.

References

Ahijevych, D., E. Gilleland, B.G. Brown, and E.E. Ebert, 2009. Application of spatial verification methods to idealized and NWP gridded precipitation forecasts. *Weather Forecast*., **24** (6), 1485 - 1497, doi: 10.1175/2009WAF2222298.1.

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

Gilleland, E., 2021. New novel generalized forecast performance metrics for high-resolution verification sets. *Advances in Statistical Climatology, Meteorology and Oceanography*, **7** (1), 13 - 34, doi: 10.5194/ascmo-7-13-2021.

Gilleland, E., D. Ahijevych, B.G. Brown, B. Casati, and E.E. Ebert, 2009. Intercomparison of Spatial Forecast Verification Methods. *Weather Forecast.*, **24**, 1416 – 1430, doi: 10.1175/2009WAF2222269.1.

Gilleland, E., D.A. Ahijevych, B.G. Brown and E.E. Ebert, 2010. Verifying forecasts spatially. *Bull. Amer. Meteor. Soc.*, **91** (10), 1365 – 1373, doi: 10.1175/2010BAMS2819.1.

Gilleland, E., G. Skok, B. G. Brown, B. Casati, M. Dorninger, M. P. Mittermaier, N. Roberts, and L. J. Wilson, 2020. A novel set of verification test fields with application to distance measures. *Monthly Weather Review*, **148** (4), 1653 - 1673, doi: 10.1175/MWR-D-19-0256.1.