Spatial Verification Using A True Metric

Meijun Zhu[1],Valliappa Lakshmanan [2,31],
Pengfei Zhang [2,3], Yang Hong [4],
Kesong Cheng [1] and Sheng Chen [4]

Abstract

Verifying high-resolution forecasts is challenging because forecasts can be considered good by their end-users even when there is no pixel-to-pixel correspondence between the forecast and the verification field. Many of the verification methods that have been proposed to address the verification of high-resolution forecasts are based on filtering, warping or searching within a neighborhood of pixels in the forecast and/or the verification fields in order to retain the capability to use a simple metric. This is because it is necessary for a verification score to be a metric to allow comparisons of forecasts. In this paper, we devise a computationally simple scalar spatial verification metric that is capable of ordering forecasts without preprocessing the fields. The metric is based on the insight that in the verification problem, the observation field can be considered a reference field that forecast fields are ordered against. This new metric is demonstrated on synthetic and real model forecasts of precipitation.

1. Introduction

The verification of high-resolution forecasts differs from the verification of high-resolution nowcasts because forecasts tend to have significant position errors. At the time scale of most nowcasts, position errors are only on the order of a few pixels, leading to considerable pixel overlap with the verification field. On the other hand, position errors in forecasts can be so high that there

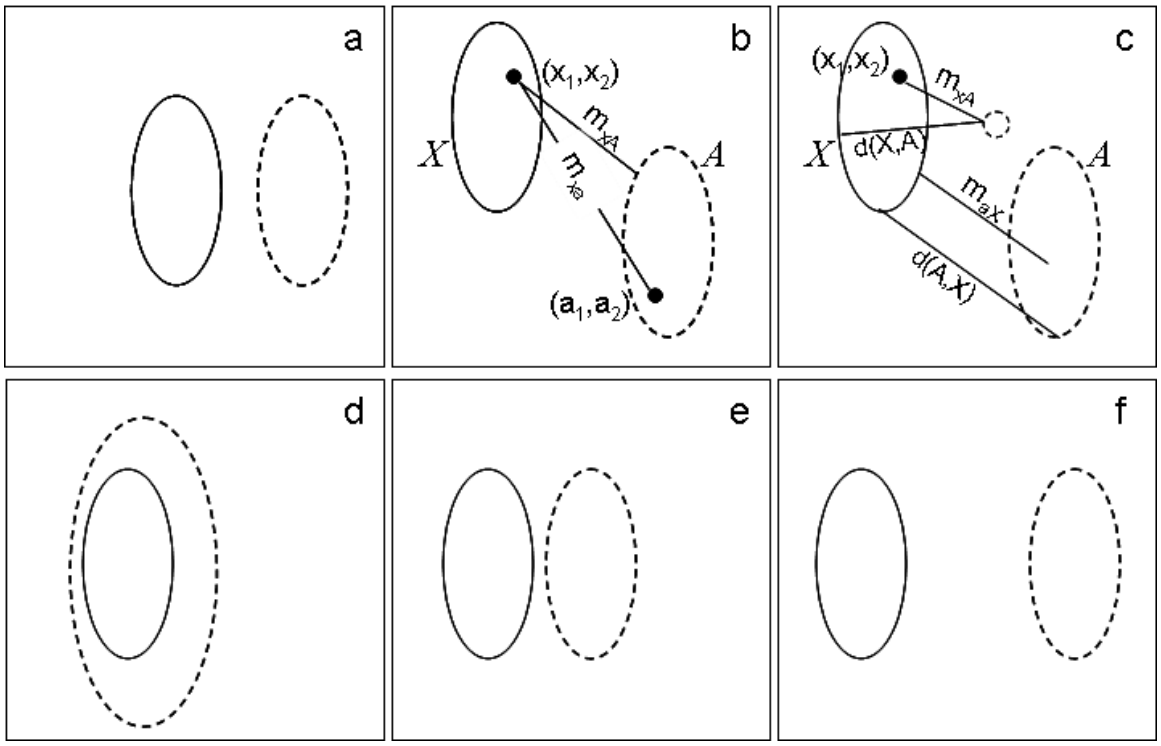

Figure 1: Schematic plots of observation and forecast fields of different scenarios referenced in the text. In all cases, solid lines represent the observation field while the dashed line represents the forecast valid for the time of the observation. (a) A position error leads to double penalty if using RMSE. (b) Euclidean metric function, between points and between a point and a set. (c) The Euclidean distance (d) between two sets is not symmetric. (d) $dist_{DV} = 0$ i.e. it does not penalize overforecasts. (e,f) $dist_{OV}$ is the same but the $dist_{DV}$ is larger in the case of (e).

is no overlap at all between the forecast and observation fields. Therefore, it is necessary to reward a forecast for an "almost-correct" position i.e. to carry out the evaluation beyond simple pixel-to-pixel correspondence. In the absence of such a reward, the verification method will suffer from a "double-penalty" problem (Gilleland et al. 2009; Ahijevych et al. 2009). For example, if the Root Mean Square Error (RMSE) is computed by differencing the corresponding pixels of the model forecast and the verifying field and then computing the average of the squared differences, the RMSE will show the impact of a double penalty because there are two areas of large differences even though the underlying problem is simply a position shift (See Figure 1a).

Several verification techniques have been introduced to address this problem of comparing forecasts with observations when there are significant displacement errors in the forecast.[2] Neighborhood approaches change the way that the error is computed: instead of computing errors by directly differencing the two fields, a neighborhood around each grid-point is searched in both the fields and the statistical properties of the set of pixels in the neighborhoods such as fractional coverage and mean value are compared (Ebert 2009). Pixel-to-pixel correspondence requirements can be avoided by comparing properties of the pixels in the entire domain of interest, as was done by Wernli et al. (2009). Filtering-based methods (e.g. Casati and Wilson (2007)) decompose the fields into images of different resolutions and compute pixel-to-pixel scores on them. The resolution at which double penalty errors start to show up is an indirect measure of the position error although the errors may well not be due to location. Non-parametric optical flow and warp approaches have been employed (Alexander et al. 1999; Keil and Craig 2009) to modify one of the fields before computing a pixel-to-pixel difference. The amount of warping that is required is an indirect measure of the position error. Davis et al. (2006); Marzban and Sandgathe (2006) describe cluster-based approaches where each field is segmented into objects and objects in the fields compared for position and size errors. Lakshmanan and Kain (2010) introduced Gaussian Mixture Models as a way to fit images into parametric models and then compared the parameters of these models on the two images to extract position, amplitude and rotation errors.

One way to categorize these methods, different from the four-category classification proposed by Gilleland et al. (2009), is to consider their intent. The neighborhood, filtering and warp methods all intend to modify the image or the range of pixels so that pixel-to-pixel (or super-pixels to super-pixels) error measurements work. The object-oriented and parametric fit approaches take the approach of avoiding the problematic pixel-to-pixel comparisons by instead comparing groups

[2]The methods handle displacement errors, but the grids themselves have to be mapped to the same projection and resolution before any verification is carried out.

of pixels or parametric fits to those pixels.

Such methods of windowing, warping or filtering images before comparing them yield rich multi-aspect measures of the goodness of a forecast. For example, the Gaussian Mixture Model approach decomposes the difference between the observation field and a model forecast into position, rotation and amplitude errors. The wavelet and neighborhood approaches provide an indication of what scale the forecast performs best. The object-based approaches provide a variety of evaluation measurements for each object in the observation and forecast fields.

In spite of the richness of the verification measures introduced in the literature, it has been our experience that end-users gravitate towards simple and intuitive scalar measures of performance. Our goal, then, is to devise an intuitive scalar measure of model performance that can be computed without extensive preprocessing of model forecast fields. A distance metric is certainly intuitive – the "farther" away a forecast is from the observation field, the worse it is. It is also a scalar and has the benefit of naturally encompassing position errors in model forecasts. Our goal in this paper is to devise a distance metric that can be used to gauge how close a forecast is to the observation. It should be noted that the metric introduced in this paper is a distance, not a skill score. Unlike a skill score, the distance is not bounded[3] – the larger the distance, the worse the forecast.

a. *Need for a metric*

It is worth stepping back a little and asking why so many verification techniques warp, filter or window (search in the neighborhood of pixels) images before computing error measurements. We suggest that one key reason is that it is very important for the final location error measurement to be a metric if a suitable one can be found.

A function $m(A, B)$ is a metric if it is symmetric ($m(A, B) = m(B, A)$), positive ($m(A, B) \geq$

[3]Our distance measure is bounded by the size of the grid, but not bounded in the sense that a Probability of Detection is bounded between 0 and 1.

0; $m(A, b) = 0$ if and only if $A = B$) and satisfies the triangle inequality ($m(A, B) + m(B, C) \geq m(A, C)$).

It is important that verification measurements are metrics because in the absence of it being a metric, we may obtain unreasonable results when comparing two forecasts. The positivity property notes specifically that the distance between two objects is zero is equivalent to the fact that these two objects are identical. This is important because in a perfect forecast, the sets of pixels corresponding to the observation and forecast fields will be identical and it is necessary to recognize a perfect forecast.

The triangle inequality property is essential to carry out a fair measurement. Think about this scenario: Let O be the observation, F_1 and F_2 be two forecasts. If we measured that the distance between O and F_1 is 100 units, and the distance between O and F_2 is 10 units, we would say that F_2 is a *better* forecast. However, if the verification measurement does not satisfy the triangle inequality property, we may find that the distance between F_1 and F_2 is, say, 0.5 units or even less. Considering the expected variance in computed distances, we may not be convinced that F_2 is really better since it is almost the same as F_1 (the distance between them is almost zero).

The symmetric property guarantees that every set has equal right to be fairly measured: the distance from set A to set B is always the same as the distance from set B to set A.

Given that it is important that verification measurements be metrics, how hard is that to realize in practice? Are not all intuitive measurements metrics?

b. *A metric between two sets*

Although the definition of a metric seems intuitive, many reasonable measurements turn out to not be metrics especially when considering the model verification problem. This is because it is non-trivial to define a distance between two sets of points that is a metric. It is easy to define a metric between a point and a set of points, but not so easy to define a metric between two sets.

Take, for example, the most intuitive measurement of all. Suppose we were to use the Euclidean metric as our distance function. For two points $x = (x_1, x_2),\ a = (a_1, a_2) \in \mathbb{R}^2$, the Euclidean metric function is:

$$m_{xa} = \sqrt{(x_1 - a_1)^2 + (x_2 - a_2)^2}. \tag{1}$$

This is shown in Figure 1b. Given the Euclidean metric betwen two points, we can define another metric, this time between any point $x \in \mathbb{R}^2$ and a set $A \subset \mathbb{R}^2$ as:

$$m_{xA} = \min_{a \in A} m_{xa} \tag{2}$$

i.e. as the distance between the point x and the closest point to it in the set A (See Figure 1b).

It can easily be noted that metric m_{xA} is overly sensitive. It is defined exclusively by the closest point and can therefore be unduly affected by noise in the data. Consider the scenario in Figure 1c where the forecast field has a single non-zero pixel close to the observations. Because this pixel is closest to the observations, all the m_{xA} will be evaluated on the basis of this one point, leading to m_{xA} being much smaller for all x in X than for the scenario in Figure 1b. Yet, the forecast is not that much better. Thus, even intuitively appealing metrics such as the Euclidean metric turn out to be problematic for verification.

The difficulty in devising a metric does not stop there. Our definition of m_{xA} is a metric function between a point and a set. We have yet to define a metric function between two sets X and A – this is what would be needed to find the distance between the pixels of the forecast and observed images. One possibility is to define it as the intuitively appealing maximum of all possible m_{xA}:

$$d(X, A) := \max_{x \in X} m_{xA} \tag{3}$$

But this Euclidean distance between two sets turns out to not even be a metric function as it is not symmetric. In other words, $d(X, A)$ can be different from $d(A, X)$ (See Figure 1c for an illustration). Using the minimum of all possible m_{xA} does not work either because it does not

satistfy the positivity property. The sets A and X need to only overlap, not be identical, for the distance to be zero.

The traditional solution to this problem is to enforce symmetry using the Hausdorff metric defined as (Rucklide 1996):

$$m_H(A, B) = \max\{d(A, B),\ d(B, A)\}. \tag{4}$$

It is interesting to observe that when B includes only one point x, the distance from x to set A is different from the Hausdorff distance between x and set A: the former measures the point x to the closest point to it in set A, the latter measures x to the farthest point to it in set A. The maximum operation in the Hausdorff metric makes it very susceptible to noise. One possible way to address this, called the Partial Hausdorff Distance (PHD), is to use, say the 75th percentile, rather than the maximum. However, this is not a metric (Rucklide 1996), so methods such as those of Venugopal et al. (2005) that are based on the PHD are also unlikely to be true metrics. Baddeley (1992) replaced the maximum in the definition of the Hausdorff metric with an L^p norm and this was employed for model forecast verification by Gilleland et al. (2008). Similar to the Hausdorff metric, such a metric may suffice when the objective is to compare objects that consist of contiguous sets of pixels i.e. if there will not be noisy pixels elsewhere in the image that have to be considered part of the distance computation. If these Hausdorff type metrics are not preceded by a step of object identification or noise removal, they are always sensitive to noise. This is because the Euclidean metric function $m_{x,A}$ that Hausdorff type metrics are built on is itself problematic for spatial *field* verification, as opposed to verifying objects extracted from those fields.

2. Verification metric

In this paper, we introduce an easily computable metric that has been devised specifically for the model verification problem. Rather than consider a generic pair of binary images, we recognize that, in model verification, there is an observation field which is quite special and a set of forecast

fields each of which has to be evaluated. Our metric will use the observation field as a reference field so as to come up with a measure that is (a) a true metric, and so can be used to rank forecasts, (b) evaluatable between two sets and does not require pixel-to-pixel correspondence.

The metric can be computed directly from the images. It is not necessary to filter, warp, window, identify objects or fit parameters to the images. It should be noted, though, that our metric is defined on sets of pixels and, so, it requires a threshold to be specified by the end-user. Pixels with a data value greater than the threshold will be considered part of the set and those with a data value less than the threshold will be considered outside the set, thus images are converted into binary images first. When we show the results of our technique, we will demonstrate the results on a variety of thresholds.

Let O be the set consisting of pixels in the observation field that are above a user-specified threshold. Let A, B, C be sets consisting of pixels in forecast fields. The verification metric that we propose is as follows.

Definition(Verification metric) *The verification metric between two sets A and B is given by*

$$metr_V(A, B) := \lambda_1 dist_{OV}(A, B) + \lambda_2 dist_{DV}(A, B). \tag{5}$$

i.e. a weighted sum of two distances that are defined below.

The overlap-based distance $dist_{OV}$ is given by:

$$dist_{OV}(A, B) = \sqrt{\sum_{\forall i} \sum_{\forall j} (a_{ij} - b_{ij})^2}, \tag{6}$$

where a_{ij} b_{ij} are characteristic functions of sets A, B, respectively. i.e. a_{ij} is 1 if the pixel i, j is in the set A, a_{ij} is 0 if the pixel i, j is not in the set A. b_{ij} is defined similarly. Recall that the pixel i, j is in the set A if its value is above a user-specified threshold. Thus, this is simply the root mean square error computed on the binary field.

Next, we introduce the observation distance. The observation distance $dist_{ob}$ is the average

distance of every observation point to a forecast field:

$$dist_{ob}(O,A) = \begin{cases} \frac{1}{N(O)} \sum_{i=1}^{N(O)} m_{oA}(o_i, A) \quad \text{if} \quad N(O) \cdot N(A) \neq 0 \\ 0 \ \text{ if } \ N(O) = 0 \ \ and \ \ N(A) = 0, \\ \text{D otherwise}, \end{cases} \tag{7}$$

where m_{oA} is the Euclidean metric function of Equation 2, o_i are the pixels in the observation field, $N(O), N(A)$ are the number of pixels in the sets O and A, respectively, i.e. the number of pixels in the corresponding images that are above the threshold. The number D in the definition of $dist_{ob}$ is a number larger than the maximum possible distance. One possible choice is the length of the diagonal of the grids being compared. This upper limit value of $dist_{ob}$ will be reached if the observation field or the forecast field is an empty set.

The above distance is used to compute the observation-based displacement $dist_{DV}$ between the sets A and B as:

$$dist_{DV}(A,B) = |dist_{ob}(O,A) - dist_{ob}(O,B)|. \tag{8}$$

The relative weights of the two component distances (of Equations 6 and Equation 8) in the verification metric is quite subjective. We use $\lambda_1 = \lambda_2 = \frac{1}{2}$ throughout this paper for simplicity. Users could choose different weights depending on whether overlap error is more or less important than displacement. This depends on the purpose of the forecast and is very much a subjective decision to be made by the user.

As defined, the units of the measurement are in pixels. It can be converted into a true distance (in, for example, km) by multiplying by the appropriate pixel dimensions.

Simplified form (Verification metric): Since this metric will mainly be used for verification, what is of interest is $metr_V(O,A)$ in which case, one of the terms in $dist_{DV}$ drops away (since $dist_{ob}(O,O) = 0$), leaving:

$$metr_V(O,A) := \lambda_1 dist_{OV}(O,A) + \lambda_2 dist_{ob}(O,A). \tag{9}$$

Other than to prove the triangle inequality, when we will need the more general form, this simplified definition with $\lambda_1 = \lambda_2 = \frac{1}{2}$ is what we will term the verification metric.

a. *Explanation of the verification metric*

Note the special role that the observation field, O, plays in $dist_{DV}$ (See Equation 8). The distance between any two fields A and B is computed as the sum of the distances between each of those fields and O. In other words, the observation field is the reference field against which forecasts are compared as far as their displacement is compared. The overlap between forecasts, on the other hand, is compared directly from the two fields (See Equation 6). Of course, if we are comparing a forecast field to an observation field, one of the terms in $dist_{DV}$ is zero and both comparisons take place on an image-to-image level.

Further $dist_{DV}$ does not penalize overforecasts. For example, consider the scenario in Figure 1d. For every point in the observation, there is a point in the forecast field that exactly matches. Therefore, $dist_{ob}$ is zero, leading to a zero $dist_{DV}$. Thus, one way of thinking about the overlap term $dist_{OV}$ is as the penalty for overforecasts. On the other hand, $dist_{OV}$ is based on strict pixel-to-pixel correspondence and is, therefore, insenstive to position errors – the scenarios in Figure 1e and Figure 1f have the same $dist_{OV}$ but the $dist_{DV}$ of Figure 1f is larger, leading to a larger value in the verification metric. In this view, $dist_{DV}$ provides the position-error sensitivity to the verification metric. It is also apparent that $dist_{OV}$ is subject to double-penalty issues but this is not a serious problem because the verification metric as a whole is sensitive to position errors.

It should be noted that the verification metric is a distance and not a bounded skill score. The larger the images being compared, the larger the maximum distance can be. The images being compared should, however, be of the same size and resolution. In practice, this can be achieved by cropping or subsampling the larger or more detailed image to meet the dimensions and resolution of the smaller, coarser image.

It should also be noted that the metric is extremely sensitive to the observation field, because distances are defined by using the observation as the reference field. This is because the verification metric is designed to compare two forecasts given the same observation. The verification metric should not be used to compare two forecasts at two different times – a forecast 100 km displaced from the observation might be acceptable when there are only a few observations, but may not be acceptable when the entire domain is full of observations.

3. Experiments

We computed the verification metric on a fake geometric and on a perturbed dataset from a verification methods intercomparison project (Gilleland et al. 2009; Ahijevych et al. 2009) that was established to improve the understanding of the characteristics of various spatial forecast verification methods. To enable reasonable inter-comparision, the verification methods were carried out on synthetic and real fields with known errors. The methods were also applied to real model forecasts from an experiment conducted by Kain et al. (2008). The results of the verification metric on the different datasets that were created by the intercomparison project are presented below.

a. *Geometric cases*

The "geometric" cases (also from the Intercomparison Project (Ahijevych et al. 2009)) were defined on a 601×501 grid and were mapped to a 601×501 subsection of the NCEP storage grid 240. The geometric cases illustrate three types of error: 1) displacement, 2) frequency, and 3) aspect ratio. The results of the verification are shown in Figure 2. The description of the results are given in Table 1.

Because the verification metric is defined on binary images, the fields were thresholded at zero i.e. pixels with a value above zero were assumed to be part of the object and pixels in the "white" background were assumed to be outside it. In particular, this means that even though the objects

have two intensity levels, they are treated as a single intensity level.

It might be helpful to delineate the steps to compute the verification metric on geom001 (see Figure 2). First, the observation field (geom000) and forecast field (geom001) are both thresholded at zero. Thus, there are two binary images. The first image consists of pixels whose value is 1 within the ellipse of geom000 and 0 outside. The second image is similar, except that the ellipse corresponds to the points in geom001. From these two binary images, the verification metric needs to be computed using Equation 9. The second step, then, is to compute $dist_{OV}$, defined in Equation 6. a_{ij} is 1 within the first ellipse while b_{ij} is 1 within the second ellipse. If the ellipses had overlapped, the difference $a_{ij} - b_{ij}$ would have been zero at points of overlap. Here, however, the ellipses do not overlap. Thus, the difference has a magnitude of 1 where either a_{ij} or b_{ij} is 1. Therefore, $dist_{OV}$ is equal to the square root of twice the size of the ellipse measured in pixels. The third step is to compute $dist_{ob}$ using Equation 7. Both the observation and the forecast have some valid points, so the answer is not simply the length of the diagonal of the grids being compared. Instead, the Euclidean distance from every observation point to the closest point in the forecast field needs to be computed. For every point within the ellipse in geom000, we need to find the closest point in geom001. It should be noted that we will find the closest point, not the corresponding point. Because geom001 consists of the ellipse displaced right, the closest points will all consist of points on the left-most boundary of the ellipse in geom001. For the points on the left boundary of the ellipse in geom000, m_{oA} will be 50, the known displacement. For points inside the ellipse and on the right boundary of the ellipse in geom000, this distance will be less, as it is always the distance to the left The average of these distances over all the points in the ellipse of geom000 is $dist_{ob}$. The final step is to average $dist_{OV}$ and $dist_{ob}$. This is the verification score for geom001.

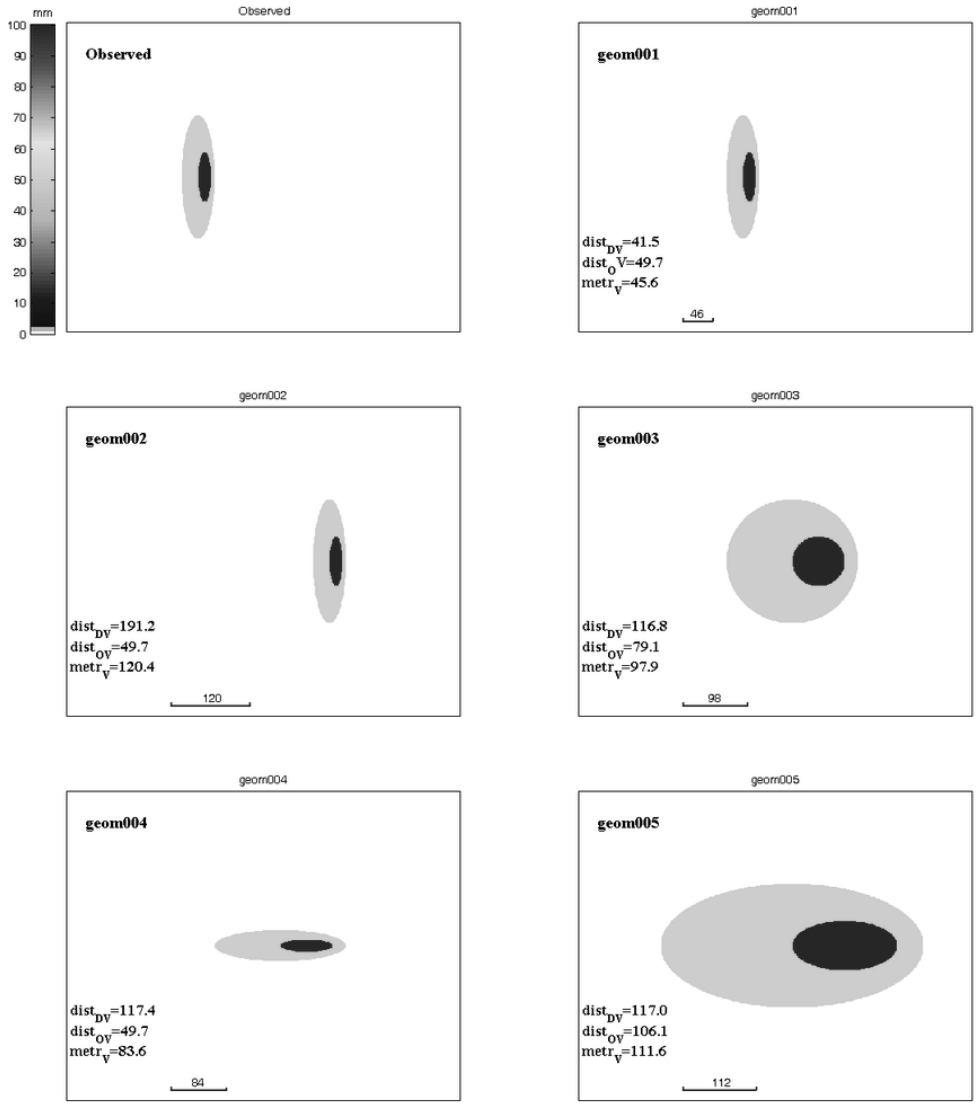

Figure 2: Verification metric (in pixels) for the fake geometric cases from Ahijevych et al. (2009). See Table 1 for details on the transformations.

Table 1: Verification metric on fake geometric images from Ahijevych et al. (2009) The position of the fake image in Figure 2 is indicated in this table.

Data set	Description	$metr_V$
geom000 (Top left)	Original	0
geom001 (Top right)	50 pts. right	46
geom002 (2nd row left)	200 pts. right	120
geom003 (2nd row right)	125 pts. right, too big	98
geom004 (3rd row left)	125 pts. right, rotated	84
geom005 (3rd row right)	125 pts. right, huge	112

Table 2: Verification metric on perturbed images from Ahijevych et al. (2009) when considered at thresholds of 0mm and 20mm. The position of the perturbed image in Figure 3 is indicated in this table.

Data set	Description	$metr_V, 0mm$	$metr_V, 20mm$
fake000 (Top left)	Original	0	0
fake001 (Top right)	3 pts. right	80.0	18.5
fake002 (2nd row left)	6 pts. right	91.8	20.9
fake003 (2nd row right)	12 pts. right	103.6	24.7
fake004 (3rd row left)	24 pts. right	116.6	33.7
fake005 (3rd row right)	48 pts. right	131.7	53.8
fake006 (Bottom left)	12 pts. right, 10 pts down	103.6	29.6
fake007 (Bottom right)	12pts. right, 20 pts down, minus 2mm	103.3	24.3

b. *Perturbed cases*

The "perturbed" set of cases from the Intercomparison Project (Ahijevych et al. 2009) consist of observed data from the 2005 NSSL/SPC Spring Experiment described in Kain et al. (2008). The observed data were subjected to various transformations such as shifting the entire image by a known number of pixels or multiplying the pixel value by a known amount. The results of the verification are shown in Figure 3. The results of verification at two different thresholds are shown in Table 2.

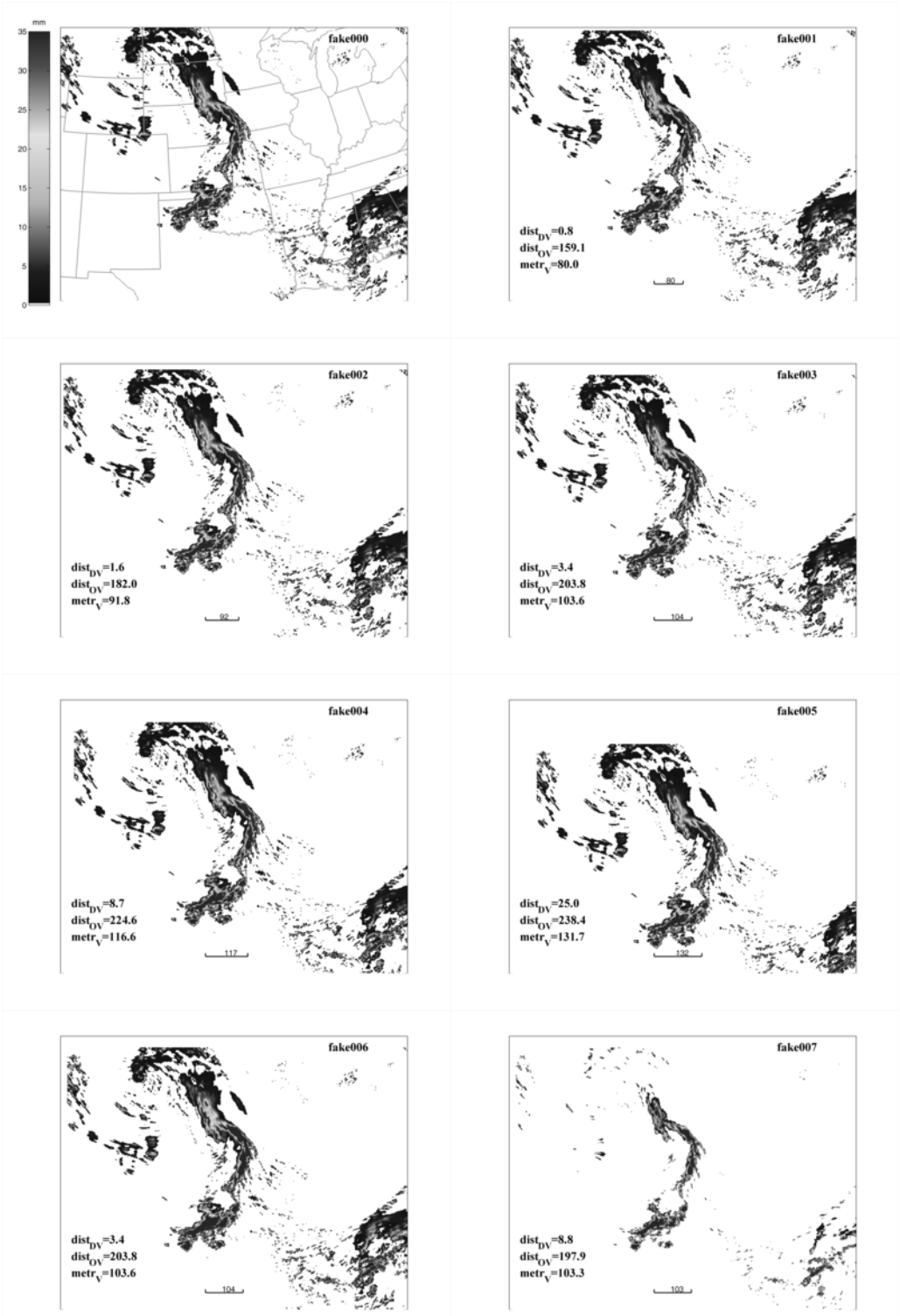

Figure 3: Verification metric for perturbed images from Ahijevych et al. (2009) thresholded at 0 mm. Details on the perturbations are in Table 2.

c. *May 14 and 19, 2005*

We also analyzed observed data and model runs from the 2005 NSSL/SPC Spring Experiment described in Kain et al. (2008) and used for intercomparisons by Gilleland et al. (2009). The observed data from May 14, 2005 were compared with 24 hour forecasts of one hour rainfall accumulation carried out on May 13, 2005. We repeated the experiment with observations from May 19, 2005 and model forecasts from May 18, 2005.

The observations and model forecasts (from the CAPS, NCAR and NCEP models) are shown in Figures 4 and 5. The images cover the lower 48 states of the United States. The NCEP model forecast was produced at the National Centers for Environmental Prediction (NCEP) using a Weather Research and Forecasting (WRF) model whose core was a Nonhydrostatic Mesoscale Model (Janjic et al. 2005) with a 4.5km grid spacing and 35 vertical levels. The NCAR model forecast was produced at the National Center for Atmospheric Research using the Advanced Research WRF (ARW; Skamarock et al. (2005)) core with a 4km grid spacing and 35 vertical levels. The CAPS was produced at the Center for Analysis and Prediction of Storms at the University of Oklahoma (also using the ARW core) with a 2km grid spacing and 51 vertical levels. All three forecast systems used initial and lateral boundary conditions from the North American Mesocale Model (Rogers et al. 2009). The observations are from the Stage II rainfall accumulation dataset produced by NCEP (Baldwin and Mitchell 1998).

Since the verification metric depends on the threshold used to evaluate the image, we show the impact of thresholding the image by illustrating the images at two thresholds. How the verification metric varies as the threshold is varied is shown in Figure 6.

4. Discussion

It should be pointed out that the verification metric introduced in this paper emphasizes the location error at the expense of fine structures in the forecast since the initial step, of converting the

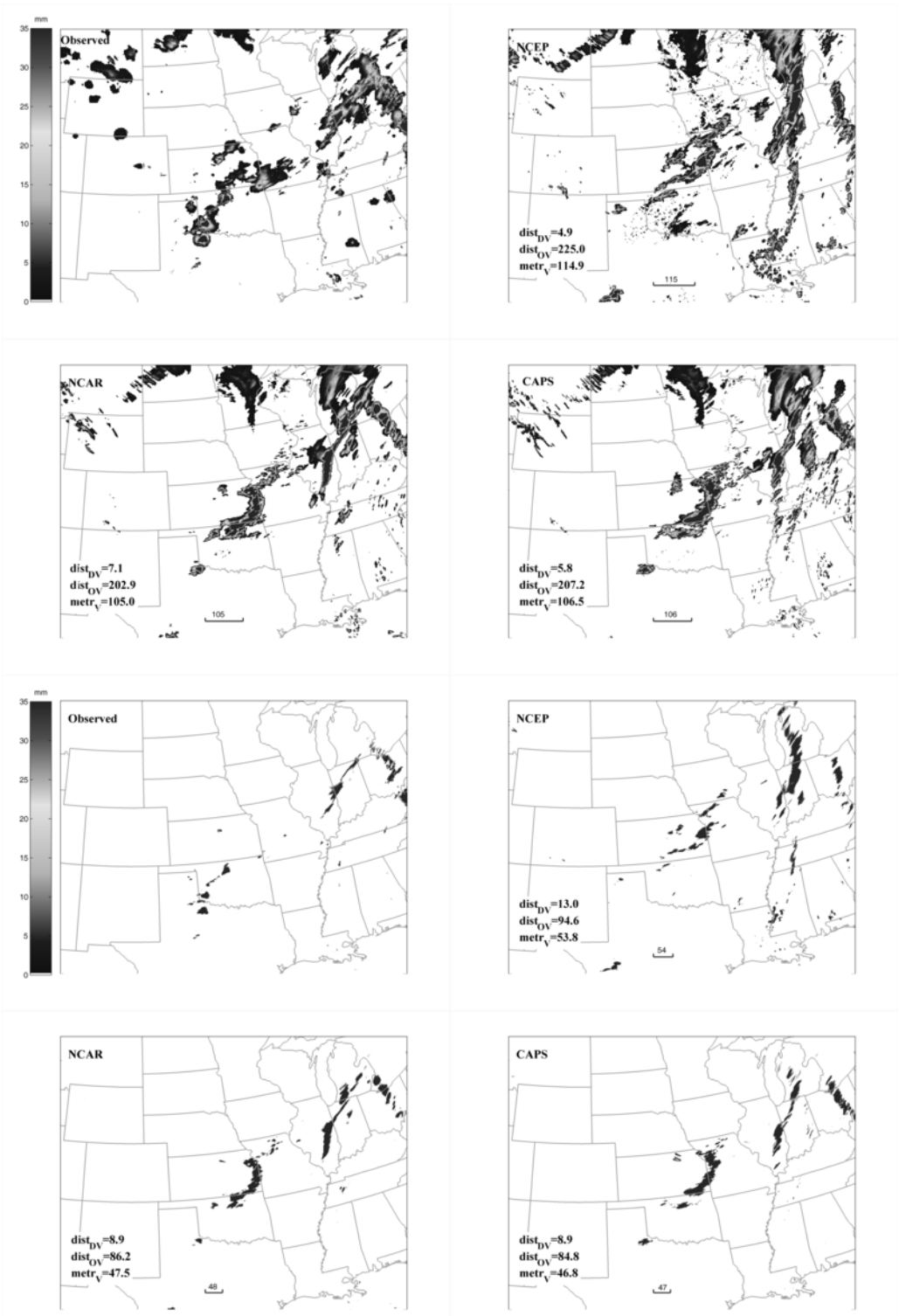

Figure 4: Verification metric for 24 hour precipitation forecasts valid for May 14, 2005 against the observations on that day. Fields are shown thresholded at 0 mm (top two rows) and 20 mm (bottom two rows).

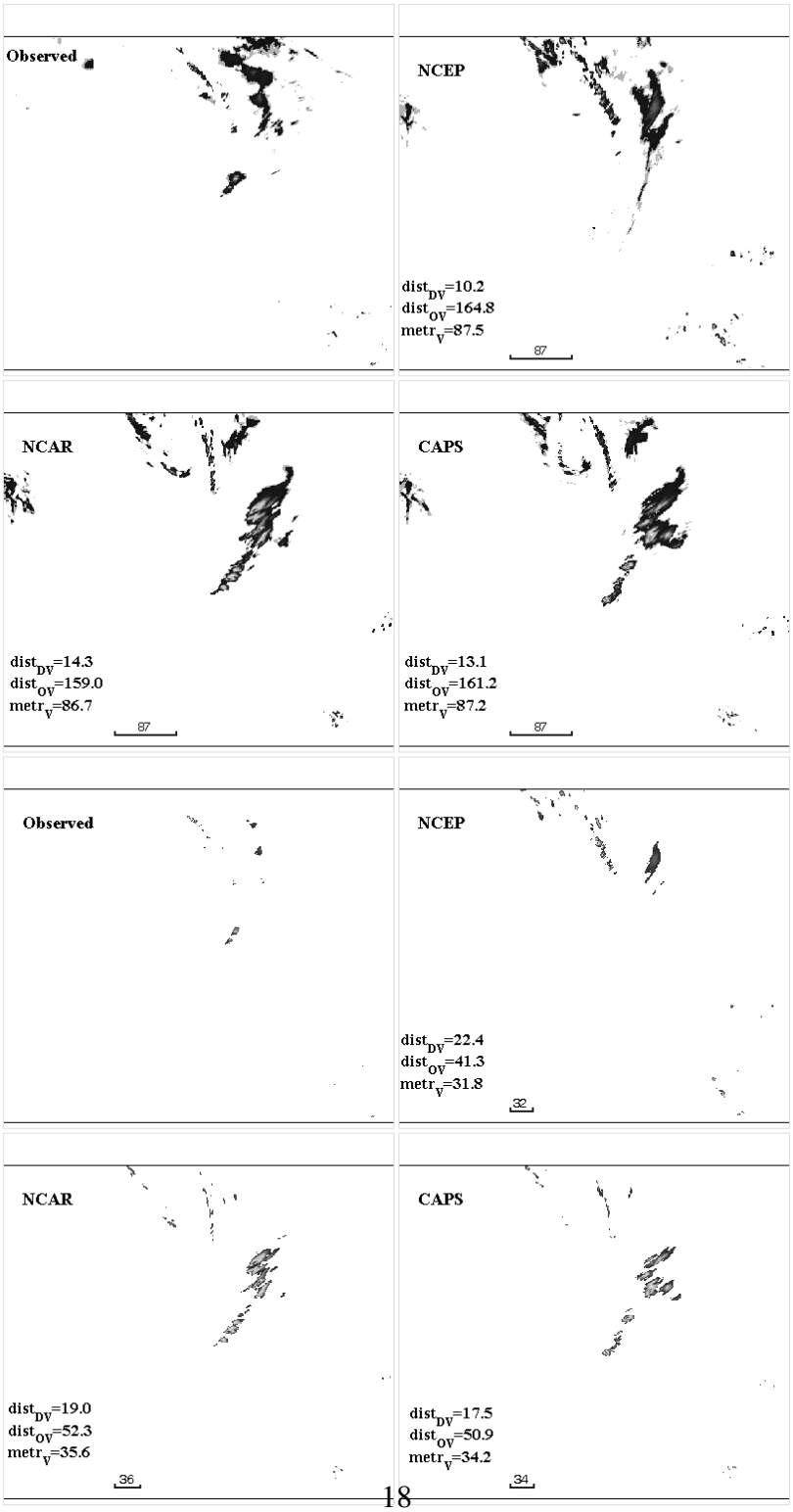

Figure 5: Verification metric for 24 hour precipitation forecasts valid for May 19, 2005 against the observations on that day. Fields are shown thresholded at 0 mm (top two rows) and 20 mm (bottom two rows).

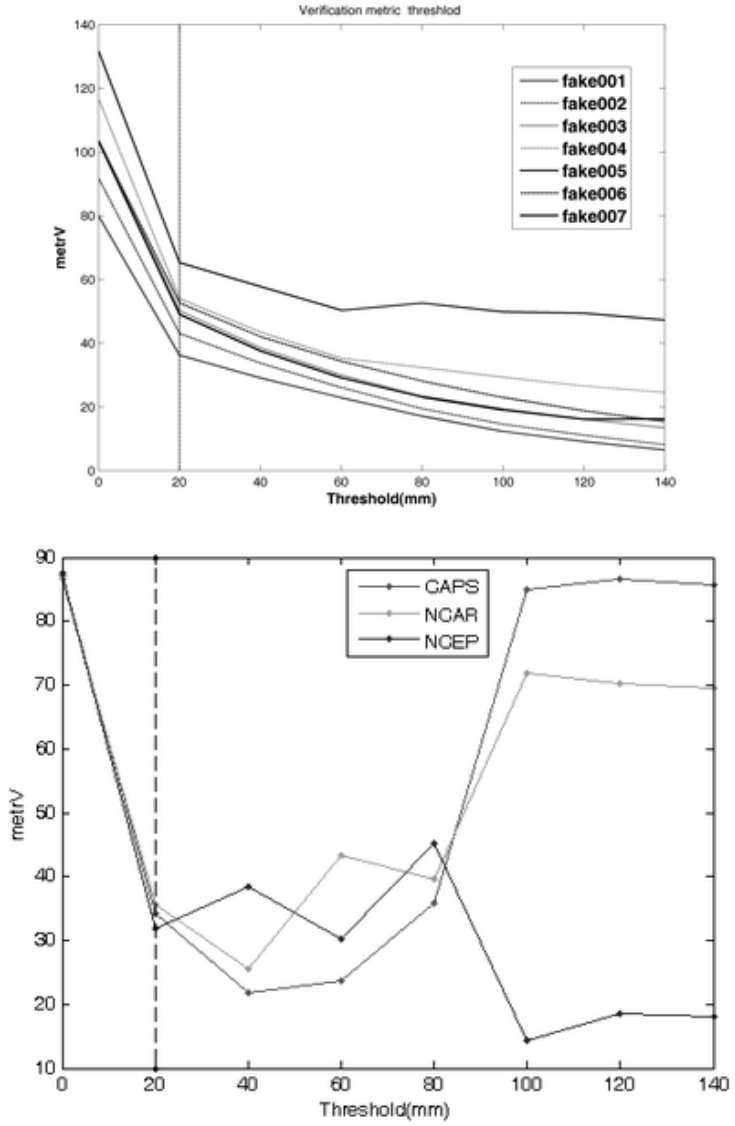

Figure 6: Top: verification metric (in pixels) by threshold for the perturbed cases from Ahijevych et al. (2009). See Table 2 for details on the perturbations. Bottom: verification metric (in pixels) by threshold for various model forecasts valid for May 14, 2005.

fields to binary by applying a threshold, treats all pixels above the threshold identically regardless of how much above the threshold the pixel's value is. It is possible to use a graph such as that in Figure 6 to derive the variation of the metric by threshold and compute a scalar metric such as the area under the curve to obtain a simple scalar metric that takes into account all the pixel values. For simplicity of analysis, however, we concentrate on a single threshold in the cases that follow.

a. *Geometric cases*

In the case of the fake geometric images, the verification metric penalizes the highly displaced forecast geom002 the most and the overforecast of geom005 nearly as much, demonstrating the impact of $dist_{DV}$ and $dist_{OV}$ respectively. As would be expected, the forecast exhibiting a small displacement (geom001) is declared the best and the value of $metr_V$ (46 pixels) is close to the known displacement of 50 pixels. The overforecast in geom003 and rotation in geom004 receive intermediate scores. It should be noted that $metr_V$ lies in the range $0, D$ and so there is no way to specify the threshold beyond which a forecast is bad. A forecast that exhibits a 50-pixel displacement may be considered bad for some applications, tolerable for others and very useful in some cases.

Comparing these results with that of Keil and Craig (2009) and Davis et al. (2006), we suggest that our ranking (01, 04, 03, 05, 02) is more pleasing than that of either Keil and Craig (2009) (01, 02, 05, 04, 03) or of Davis et al. (2006) (04, 03, 05, 02, 01). Note that our ranking places the slightly displaced figure in geom001 highest whereas the method of Davis et al. (2006) favors the much larger overforecast in geom004 because it happens to overlap slightly with the observation. On the other hand, both our method and the method of Keil and Craig (2009) pick geom001 as the best, but differ on the second-place entry. Our method picks rotated forecast of geom004 because it is shifted the least (125 points) whereas Keil and Craig (2009) pick the highly displaced forecast in geom002 because it happens to have a similar size to the observed phenomenon.

b. *Perturbed cases*

In the case of the perturbed images, the relative ranking of forecasts 2 to 6 is mostly independent of the threshold (See Figure 6 and Table 2). A lower value of the verification metric indicates a better forecast since it indicates that the forecast is less distant from the observation. The verificiation metric shows that the order of the forecasts is, from best to worst: fake001 (3 points right), fake002 (6 points right), fake003 (12 points right), fake006 (12 points right and 10 points down), fake004 (24 points right) and fake005 (48 points right). Comparing with the displacement of the images, this is quite reasonable. The only forecast that varies in relative ranking is fake007 which has both a position error and a reduction in pixel magnitude. The better ranking of fake007 than fake006 indicates the reduction in pixel magnitude happens to match up with the structures of forecast, even though the displacement of fake007 before the reduction of intensity is larger than that of fake006.

c. *May 14 and 19, 2005*

In the case of the real forecasts valid for May 14, 2005, the relative ranking of the three forecasts falls into three broad intervals (See bottom panel of Figure 6). Below a threshold of 25mm, all three forecasts have very similar performance but the NCEP forecast is worse than the CAPS and NCAR ones. Between 25mm and 100mm, the CAPS forecast is better than the other two. Beyond 125mm, the NCAR forecast is best. Based on this, we can determine that all three forecasts have similar performance in predicting the gross structure of precipitation, that the CAPS model is better at predicting moderate rainfall on that day and that the NCAR forecast is better at predicting extreme rainfall.

The number of forecast points becomes zero beyond some threshold and results in the displacement measure becoming the arbitrary value D. This explains the saturation in the value of the verification metric beyond a certain threshold. The fact that there is a jump to this saturation value indicates that using the length of the diagonal of the grids might not be the best choice of a value

for D.

d. *Fast computation*

The verification metric requires the computation of Euclidean distances between every pair of points in the observation and model forecast fields. It is necessary to employ distance transforms in order to compute distances in a computationally efficient manner. Readers wishing to implement the verification metric and requiring computational efficiency are directed to a survey of exact Euclidean distance transforms by Fabbri et al. (2008). In particular, we suggest the use of ordered propogation (Cuisenaire and Macq 1999) or the two-step decomposition proposed by Saito and Toriwaki (1994) depending on the number of points above threshold. If the number of points is small relative to the number of pixels in the image, ordered propogation will be more efficient whereas if the two are comparable, Saito's method should be prefered. The survey paper of Fabbri et al. (2008) was accompanied by source code which is freely available on the internet – readers wishing to implement the verification of this paper are strongly encouraged to use these suggested resources rather than employ brute force to calculate distances.

e. *Summary*

In this paper we suggested that by using true metrics for spatial verification, it is possible to use a simple scalar number to capture the goodness of a forecast even if there is no pixel-to-pixel correspondence. Further, we devised a verification metric (Equation 9) and showed that it was suitable for verifying model forecasts.

Acknowledgements

Funding for Lakshmanan and Zhang was provided under NOAA-OU Cooperative Agreement NA17RJ1227. Zhu would like to thank Biqun Chen for valuable discussion which leads to current

form of Equation 8.

Appendix A: Proof that $metr_V$ is a metric

Definition: For a given set Ω and its elements $A_1, A_2, A_3, ...$ in Ω: a function $m(x, y)$ is called a metric function if it satisfies:

1. Symmetric property: $m(A_i, A_j) = m(A_j, A_i)$;
2. Triangle inequality: $m(A_i, A_j) + m(A_j, A_k) \geq m(A_i, A_k)$;
3. Positivity: $m(A_i, A_j) \geq 0$; And $m(A_i, A_j) = 0$ if and only if $A_i = A_j$.

We wish to prove that the $metr_V$ (Equation 9) is a true metric and satisfies all three of the above conditions.

It is easy to see that $dist_{DV}(\cdot, \cdot)$ satisfies symmetric properties. We show that $dist_{DV}(\cdot, \cdot)$ also satisfies the triangle inequality. In fact, for given A, B, C (without loss of generality, we may assume that these sets are all distinct), We have that

$$\begin{aligned} dist_{DV}(A, B) + dist_{DV}(B, C) & \\ &= |dist_{ob}(O, A) - dist_{ob}(O, B)| + |dist_{ob}(O, B) - dist_{ob}(O, C)| \\ &\geq |dist_{ob}(O, A) - dist_{ob}(O, C)| \\ &= dist_{DV}(A, C). \end{aligned}$$

However, $dist_{DV}$ does not satisfy the positivity property, since from $dist_{DV}(A, B) = 0$ we can only conclude that $dist_{ob}(O, A) = dist_{ob}(O, B)$, which may not yield $A = B$. So, we wish to caution that it is the sum of $dist_{DV}$ and $dist_{OV}$ that is a metric. One should not pull out any of these terms by themselves and use them to evaluate forecasts.

Since we have shown that $dist_{DV}$ satisfies symmetric and triangle inequality properties, we need to show that $dist_{OV}$ satisfies symmetric and triangle inequality properties, and to show that $metr_V$ satisfies the positivity property.

Let A, B, C be three sets and a_{ij}, b_{ij}, c_{ij} be corresponding characteristic functions. We first observe that

$$\begin{aligned}
&(a_{ij} - b_{ij})^2 + (b_{ij} - c_{ij})^2 - (a_{ij} - c_{ij})^2 \\
&= 2b_{ij}^2 - 2a_{ij}b_{ij} - 2b_{ij}c_{ij} + 2a_{ij}c_{ij} \\
&= 2b_{ij}(b_{ij} - a_{ij}) + 2c_{ij}(a_{ij} - b_{ij}) \\
&= 2(a_{ij} - b_{ij})(c_{ij} - b_{ij}) \\
&\geq 0.
\end{aligned}$$

The last inequality follows from the fact that a_{ij}, b_{ij},c_{ij} are either 1 or 0. Thus

$$\begin{aligned}
dist_{OV}(A, B) + dist_{OV}(B, C) &= \sqrt{\sum_{i}^{m}\sum_{j=1}^{n}(a_{ij} - b_{ij})^2} + \sqrt{\sum_{i}^{m}\sum_{j=1}^{n}(b_{ij} - c_{ij})^2} \\
&\geq \sqrt{\sum_{i}^{m}\sum_{j=1}^{n}(a_{ij} - b_{ij})^2 + \sum_{i}^{m}\sum_{j=1}^{n}(b_{ij} - c_{ij})^2} \\
&\geq \sqrt{\sum_{i}^{m}\sum_{j=1}^{n}(a_{ij} - c_{ij})^2} \\
&= dist_{OV}(A, C).
\end{aligned}$$

This yields that the triangle inequality holds for $dist_{OV}(\cdot, \cdot)$. It is obvious from the definition that $dist_{OV}(A, B) = dist_{OV}(B, A)$.

To show that $metr_V$ satisfies the positivity property, we first observe that

$$metr_V(A, B) = \lambda_1 dist_{OV}(A, B) + \lambda_2 dist_{DV}(A, B) \geq 0.$$

Also, we observe that $metr_V(A, B) = 0$ if and only if $dist_{ob}(O, A) = dist_{ob}(O, B)$, and $dist_{OV}(A, B) = 0$, which is equivalent to say $A = B$. Thus $metr_V$ satisfies the positivity property.

There is a geometric way to view $dist_{OV}$. If we introduce a characteristic function for set A and B:

$$\chi_A(x) = \begin{cases} 1 \ \ if \ \ x \in \ A \\ 0 \ \ if \ \ x \notin \ A, \end{cases} \qquad \chi_B(x) = \begin{cases} 1 \ \ if \ \ x \in \ B \\ 0 \ \ if \ \ x \notin \ B. \end{cases}$$

Then

$$dist_{OV}(A, B) = \sqrt{\int |\chi_A - \chi_B|^2 dx} = ||\chi_A - \chi_B||_{L^2},$$

where $|| \cdot ||_{L^2}$ is usually called L^2 norm. From this, one can view the metric as the square root of the area of the difference between sets A and B:

$$dist_{OV}(A, B) = ||\chi_A - \chi_B||_{L^2} = \sqrt{area\{(A \cup B) \setminus (A \cap B)\}}.$$

References

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

Alexander, G., J. Weinman, V. Karyampudi, W. Olson, and A. Lee, 1999: The effect of assimilating rain rates derived from satellites and lightning on forecasts of the 1993 superstorm. *Mon. Wea. Rev.*, **127**, 1433–1457.

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

Baldwin, M. and K. Mitchell, 1998: Progress on the NCEP hourly multi-sensor u.s. precipitation analysis for oeprations and GCIP research. *2nd Symp. on Integrated Observing Systems*, Amer. Meteor. Soc., Phoenix, AZ, 10–11.

Casati, B. and L. Wilson, 2007: A new spatial-scale decomposition of the Brier score: Application to the verification to lightning probability forecasts. *Monthly Weather Review*, **135**, 3052–3069.

Cuisenaire, O. and B. Macq, 1999: Fast euclidean distance transformation by propagation using multiple neighborhoods. *Computer Vision and Image Understanding*, **76**, 163 – 172.

Davis, C., B. Brown, and R. Bullock, 2006: Object-based verification of precipitation forecasts. part i: Methodology and application to mesoscale rain areas. *Monthly Weather Review*, **134**, 1772–1784.

Ebert, E., 2009: Neighborhood verification: A strategy for rewarding close forecasts. *Weather and Forecasting*, **24**, 1498–1510.

Fabbri, R., L. Costa, J. Torelli, and O. Bruno, 2008: 2D Euclidean distance transforms: a comparative survey. *ACM Computing Surveys*, **40**, 44, doi: 10.1145/1322432.1322434 http://doi.acm.org/10.1145/1322432.1322434.

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

Gilleland, E., T. Lee, J. Gotway, R. Bullock, and B. Brown, 2008: Computationally efficient spatial forecast verification using baddeley's delta image metric. *Monthly Weather Review*, **136**, 1747–1757.

Janjic, Z., T. Black, M. Pyle, H. CHuang, E. Rogers, and G. DiMego, 2005: High resolution applications of the WRF NMM. *21st Conf. on Weather Analysis and Forecasting/17th Conf. on Numerical Weather Predicition*, Amer. Meteor. Soc., Washington, DC, 16A.4.

Kain, J. S., S. J. Weiss, D. R. Bright, M. E. Baldwin, J. J. Levit, G. W. Carbin, C. S. Schwartz, M. L. Weisman, K. K. Droegemeier, D. B. Weber, and K. W. Thomas, 2008: Some practical considerations regarding horizontal resolution in the first generation of operational convection-allowing NWP. *Weather and Forecasting*, **23**, 931–952.

Keil, C. and G. Craig, 2009: A displacement and amplitude score employing an optical flow technique. *Wea. and Forecasting*, **24**, 1297–1308.

Lakshmanan, V. and J. Kain, 2010: A Gaussian mixture model approach to forecast verification. *Weather and Forecasting*, **25**, 908–920, doi:10.1175/2010WAF2222355.1.

Marzban, C. and S. Sandgathe, 2006: Cluster analysis for verification of precipitation fields. *Weather and Forecasting*, **21**, 824–838.

Rogers, E., G. DiMego, T. Black, M. Ek, B. Ferrier, G. Gayno, Z. Janjic, Y. Lin, M. Pyle, V. Wong, W. Wu, and J. Carley, 2009: The NCEP north american mesoscale modeling system: Recent changes and future plans. *23rd Conf. on Weather Analysis and Forecasting/19th Conf. on Numerical Weather Predicition*, Amer. Meteor. Soc., Omaha, NE, 2A.4.

Rucklide, W., 1996: *Effective Visual Recognition Using the Hausdorff Distance*. Springer, New York, 178 pp.

Saito, T. and J.-I. Toriwaki, 1994: New algorithms for euclidean distance transformation of an n-dimensional digitized picture with applications. *Pattern Recognition*, **27**, 1551 – 1565.

Skamarock, W., J. Klemp, J. Dudhia, D. Gill, D. Barker, W. Wang, and J. Powers, 2005: A description of the Advanced Research WRF version 2. Technical Report NCAR/TN-468*STR, National Center for Atmospheric Research, Boulder, CO, available from UCAR Communications, P.O. Box 3000, Boulder CO 80307.

Venugopal, V., S. Basu, and E. Foufoula-Georgiou, 2005: A new metric for comparing precipitation patterns with an application to ensemble forecasts. *J. Geophysical Research*, **110**.

Wernli, H., C. Hofmann, and M. Zimmer, 2009: Spatial forecast verification methods intercomparison project – application of the SAL technique. *Weather and Forecasting*, **24**, DOI: 10.1175/2009WAF2222271.1.