Neighborhood Verification: A Strategy for Rewarding Close Forecasts

ELIZABETH E. EBERT

Centre for Australian Weather and Climate Research, Bureau of Meteorology, Melbourne, Victoria, Australia

(Manuscript received 17 December 2008, in final form 16 April 2009)

ABSTRACT

High-resolution forecasts may be quite useful even when they do not match the observations exactly. Neighborhood verification is a strategy for evaluating the “closeness” of the forecast to the observations within space–time neighborhoods rather than at the grid scale. Various properties of the forecast within a neighborhood can be assessed for similarity to the observations, including the mean value, fractional coverage, occurrence of a forecast event sufficiently near an observed event, and so on. By varying the sizes of the neighborhoods, it is possible to determine the scales for which the forecast has sufficient skill for a particular application. Several neighborhood verification methods have been proposed in the literature in the last decade. This paper examines four such methods in detail for idealized and real high-resolution precipitation forecasts, highlighting what can be learned from each of the methods. When applied to idealized and real precipitation forecasts from the Spatial Verification Methods Intercomparison Project, all four methods showed improved forecast performance for neighborhood sizes larger than grid scale, with the optimal scale for each method varying as a function of rainfall intensity.

1. Introduction

High space and time resolution quantitative precipitation forecasts (QPFs) are becoming increasingly available for use in such weather-related applications as heavy rain, flooding, landslides, and other high-impact weather prediction and hydrological applications such as streamflow prediction and water management. High-resolution modeling allows for more realistic structure and variability in the rainfall patterns, including better representation of topographically influenced rainfall associated with mountains and coastlines. Simulated rain intensity distributions from high-resolution models have the potential to better match those measured from rain gauges and radar.

Quantifying the accuracy of high-resolution QPFs can be tricky, as traditional verification statistics based on point matches between the forecasts and observations from gauges and/or radar data severely penalize finescale differences that are not present in coarser-resolution forecasts. Moreover, the uncertainties in the observational data due to sampling, measurement, and representativeness errors have a greater impact on the verification statistics at high resolution than when observations are spatially averaged on a coarse grid (e.g., Tustison et al. 2001). Poorer verification scores measured by traditional verification approaches do not reflect the perceived added benefit of the high-resolution modeling, in terms of forecast realism.

In recent years many new spatial verification approaches have been proposed to try to more adequately reflect various aspects of the quality of high-resolution forecasts. Casati et al. (2008) and Gilleland et al. (2009) review these new approaches, which fall into one of four types: they may measure scale-dependent accuracy by filtering finer scales (neighborhood methods), isolate scale-dependent errors (scale separation methods), verify the macroscale properties of rain systems (features-based methods), or distort the forecast to better match the observations (field deformation). This paper focuses on the first approach.

Neighborhood, also known as fuzzy, methods measure the accuracy of the forecasts within space–time neighborhoods. All grid-scale values within a spatial and/or temporal neighborhood of the observation are considered to be equally likely estimates of the true value, thus giving a probabilistic flavor to the verification
process. Ebert (2008) called these techniques “fuzzy” because they allow a forecast to be partially correct and partially incorrect (as in fuzzy logic), and also because the space–time filtering can be thought of as a “fuzzifying” or blurring process. However, “neighborhood verification” is now the preferred term as it more clearly reflects the verification strategy, and avoids confusion with fuzzy set theory. The scales for which the forecasts have useful skill are determined by varying the sizes of the neighborhoods and performing the verifications at multiple scales and for multiple intensity thresholds.

The oldest and best-known neighborhood verification method is upscaling (averaging) the forecasts and observations to coarser space and time resolutions and verifying using the usual metrics (e.g., Zepeda-Arce et al. 2000; Yates et al. 2006). This enables an assessment of how similar the mean forecast value is to the mean observed value as the scale is increased. This may not be ideal for verifying QPFs for all applications since potentially useful information (such as potential for flash flooding) is lost in the averaging process. Other aspects of the rainfall distribution such as the maximum value or fractional rain coverage may be more relevant for some applications.

Many neighborhood verification methods have been independently proposed in the last decade to assess the quality of high-resolution forecasts (Brooks et al. 1998; Zepeda-Arce et al. 2000; Atger 2001; Damrath 2004; Germann and Zawadzki 2004; Weygandt et al. 2004; Theis et al. 2005; Marsigli et al. 2005; Rezacova et al. 2007; Segawa and Honda 2007; Roberts and Lean 2008). These methods address various aspects of the rainfall distribution, each one explicitly or implicitly making an assumption about what constitutes a useful forecast. Ebert (2008) described 12 neighborhood verification methods and showed that they could be implemented together within a framework to simultaneously investigate multiple aspects of the forecast spatiotemporal distribution. Because the resulting large array of statistics can be daunting, it was recommended that the user first identify the most important aspect(s) of the forecast to predict correctly, then select one or two neighborhood verification methods that specifically address those important aspects. The verification results for that method show the scales for which useful or optimal skill is achieved.

In this paper four neighborhood verification methods are described and demonstrated for idealized and real high-resolution precipitation forecasts included in the Spatial Forecast Verification Methods Intercomparison Project (Abijevych et al. 2009, hereafter AGBE; Gilleland et al. 2009). These methods are particularly suited for evaluating precipitation mean values, precipitation frequency, occurrence of extreme values, and similarity of the rain pattern to the observed pattern. The information that can be gained from each of the methods is highlighted.

2. Description of neighborhood verification methods

Neighborhood verification computes error metrics for the set of all spatial neighborhoods in the domain, where a neighborhood is a window of grid boxes centered on each individual grid box. The window can be round or square; experience shows little impact on the results, so square windows are used for simplicity. The observation against which each forecast neighborhood is verified can either be the observed value in the central grid box, or it can be a statistic calculated from the set of observations within the neighborhood, depending on the neighborhood method being used. The neighborhood size is systematically increased from $1 \times 1$, to $3 \times 3$, to $5 \times 5$, and so on, up to the largest square array that can fit within the full domain. (In practice not all neighborhood sizes may be useful or sensible, so to speed computation time a smaller number of neighborhood sizes, $n$, is usually chosen.) Since many forecast errors are related to the mistiming of a frontal passage or the onset of convection, it may be desirable to also include a number, $t$, of time windows (e.g., Theis et al. 2005).

For verifying precipitation forecasts, categorical verification scores are frequently used because they are less sensitive to outliers than are quadratic scores like the root-mean-square error. Categorical scores are based on the joint distribution of forecast and observed events, commonly known as a contingency table, and containing four elements: hits, misses, false alarms, and correct negatives. [More information on commonly used verification scores can be found in the textbooks of Wilks (2006) and Jolliffe and Stephenson (2003), or on the Internet (JWGV 2009).] In QPF verification an event is usually defined as the occurrence of rain greater than or equal to a given intensity threshold. When applied to neighborhoods, an event may refer to one or more rain...
occurrences meeting or exceeding the threshold, again depending on which neighborhood method is applied (e.g., Marsigli et al. 2005). By verifying using a variety of thresholds, \(R_1, \ldots, R_m\), one can assess how the forecast accuracy depends on the rain intensity.

So, while traditional grid-scale verification provides a single score on which to judge the forecast, or at most \(m\) values if categorical scores are computed for a range of thresholds, neighborhood verification provides \(n \times t \times m\) values that give the accuracy at multiple space and time scales and intensity values. If the forecasts are not sufficiently accurate at grid scale for a particular application, the scale at which the accuracy becomes “good enough” according to some criterion can easily be determined.

The first method demonstrated here is upsampling, where the forecasts and observations are averaged to increasingly larger scales before being compared using the usual continuous and categorical verification metrics (Zepeda-Arce et al. 2000; Yates et al. 2006). Upscaling is a good verification approach to use when it is important to know how well the forecast mean value agrees with the mean value of the observations, for example, in evaluating precipitation forecasts for a hydrological watershed. Since the upsampling method is familiar and easy to understand, it also makes a useful reference against which the results of other methods can be compared. The upsampling results in this paper show the Gilbert skill score (GSS; Gilbert 1884), which is the original name of the well-known equitable threat score (ETS) that penalizes both false alarms and misses.

The fractions skill score (FSS) method compares the forecast and observed rain fractional coverage, rather than rain amounts, within spatial neighborhoods (Roberts and Lean 2008). The observed fractional coverage is normally computed from gridded observations, which are most easily obtained from radar rainfall analyses. By focusing on the rain distribution, this method gives information on whether the forecast rain looks realistic compared to the observations, and is therefore a useful tool for modelers. The use of rain fractions is also more robust to the presence of random error in the observations. The fractions skill score is based on a variation of the Brier score used to verify probability forecasts:

\[
\text{fractions Brier score (FBS)} = \frac{1}{N} \sum_{n} (P_{\text{fcst}} - P_{\text{obs}})^2, \tag{1}
\]

where \(P_{\text{fcst}}\) and \(P_{\text{obs}}\) are the fractional forecast and observed rain areas in each neighborhood (analogous to the probability that a pixel in the neighborhood contains rain), and their squared difference is averaged over the \(N\) neighborhoods in the domain. The FBS is transformed into the (positively oriented) FSS by referencing it to the corresponding fractions Brier score for the mismatched case:

\[
\text{FSS} = 1 - \frac{1}{N} \left( \sum_{n} P_{\text{fcst}}^2 + \sum_{n} P_{\text{obs}}^2 \right). \tag{2}
\]

The FSS ranges from 0 for a complete mismatch to 1 for a perfect match.

Roberts and Lean (2008) show that the value of FSS above which the forecasts are considered to have useful skill (i.e., better than a uniform probability forecast of \(f_{\text{obs}}\), the observed rain fraction over the domain) is given by

\[
\text{FSS}_{\text{useful}} = 0.5 + \frac{f_{\text{obs}}}{2}. \tag{3}
\]

The smallest scale at which the FSS exceeds FSS_{useful} can be thought of as a “skillful scale.” Users may find this physical quantity easier to relate to than a skill score (Mittermaier and Roberts 2010). The skillful scale determined using the FSS is being computed at the Met Office to assess the quality of their high-resolution QPFs (M. Mittermaier 2008, personal communication).

Weather forecasters often consult high-resolution QPFs when preparing warnings of heavy rainfall. However, since it is unrealistic to expect the model to be able to pinpoint the precise location and timing of heavy rain, the model output is almost always interpreted rather than being taken at face value. Atger (2001) developed the multi-event contingency table (MECT) method to measure whether the forecasts succeed in predicting at least one occurrence of an event close to the observations, where “closeness” can be specified in terms of space, time, intensity, and any other important aspect. The observation is the value in the center of the neighborhood, which might represent the rain at a location of interest, perhaps a population center. This “user focused” view emphasizes the importance of forecasting and verifying for a particular location. In contrast, the previous two methods (upsampling and FSS) represent a “model focused” view in which observations are transformed to the scale of the model or neighborhood. Although the user-focused view is more demanding than the model-focused one, it is still not as tough as the traditional point-to-point verification, since skill can still be demonstrated when the forecast predicts rain close to the observation.

Using thresholds on spatial distance (forecast event within 10 km of the location of interest, within 20 km,
30 km, etc.) and rainfall intensity to define “multi events,” Atger (2001) generated contingency tables for each combination. From these he computed and plotted the probability of detection versus the false alarm rate to produce a relative operating characteristic (ROC) from the cloud of points. A single value that summarizes each point in the ROC diagram is the Hanssen and Kuipers discriminant (HK), also known as the true skill score, which is simply the difference between the probability of detection and the false alarm rate. The HK score is used within the neighborhood verification framework, with a value of 0 for no skill and 1 for perfect skill. The scale at which the HK peaks can be thought of as an optimal search diameter for forecast events, where the hit rate is high but without too many false alarms.

The fourth neighborhood verification method demonstrated here is the practically perfect hindcast (PP) method proposed by Brooks et al. (1998) for evaluating forecasts of rare events. Recognizing the difficulty in scoring well using traditional verification metrics such as the threat score (simply due to the high number of false alarms), this approach puts the verification score into context by comparing it against the score that would be achieved using a practically perfect hindcast. The practically perfect hindcast is obtained by objectively analyzing the observations onto a spatial map of occurrence probability, then defining the optimal threat (warning) area by the probability contour $P_{opt}$ that gives the best verification score over the whole domain. This approach represents the threat area that would have been drawn given perfect (prior) knowledge of the observations. The ratio of the actual score to the PP score indicates how close the forecast was to “perfect.” This approach is of particular interest to forecasters, who can also visually compare the actual and PP threat areas to assess their similarity.

Within the context of neighborhood verification, an observed event is defined as the occurrence of the grid-scale value in the center of the neighborhood meeting or exceeding the intensity threshold of interest, while a forecast event is defined as the neighborhood fractional occurrence meeting or exceeding $P_{opt}$. Brooks et al. (1998) chose the threat score as the value to optimize. Here, the Gilbert skill score is chosen since it is less sensitive to forecast bias (Baldwin and Kain 2006). The metric to be output by the neighborhood verification is the ratio of the actual and practically perfect scores: $\text{GSSratio} = \frac{\text{GSS}}{\text{GSS}_{pp}}$.

3. Neighborhood verification of idealized geometric and perturbed forecasts

The intercomparison project included five idealized geometric forecast cases and seven perturbed cases with prescribed errors, to test whether the verification methods gave appropriate and useful error information. The details of the idealized cases are given by AGBE. Although the idealized cases are best suited for testing features-based verification methods that diagnose location errors and other attributes of rain systems, they are also useful for investigating the behavior of other types of verification methods.

a. Geometric cases

The idealized geometric cases are shown in Fig. 1 of AGBE and are not reproduced here. The observed field in each of the five geometric cases was a large vertically oriented ellipse with a major (minor) axis of 200 (50) grid points and was filled with rainfall of intensity $12.7 \text{ mm h}^{-1} (0.5 \text{ in. h}^{-1})$, in which was embedded a smaller ellipse of $25 \text{ mm h}^{-1}$ offset east of center. The forecast fields were ellipses that were identical to the observations but displaced by 50 and 200 grid points (cases geom001 and geom002), or displaced horizontally and distorted to various degrees (cases geom003, geom004, and geom005). In particular, the geom005 forecast was displaced 125 points to the east and was elongated by a factor of 8 in the E–W direction. To these cases we add case “geom000,” which is a forecast that is identical to the observations.

Figure 1 shows the neighborhood verification results for the upscaling, FSS, MECT, and PP methods for the perfect forecast case geom000. A score is computed for each combination of spatial scale (neighborhood size) and intensity threshold. The lower-left score in each plot in Fig. 1 is the traditional grid-scale value most often computed for a low rain threshold.

Although one might expect all of the neighborhood verification methods to reflect perfect performance for all scales and intensities, in fact the MECT method does not. This surprising result was first noticed by Ament et al. (2008), who performed a similar set of idealized known-error experiments using neighborhood verification. They called this behavior “leaking scores,” and it occurs for methods that compare a neighborhood of forecasts to a single value in the center. In the case of MECT, forecasts for grid boxes located near the raining ellipse were counted as false alarms because there was at

---

3 Here, $P_{opt}$ is a function of the observation density and event frequency, and varies for each set of observations. An alternative implementation of the PP method would be to specify the probability contour in advance as, say, 0.5; however, this has not been done here.
least one predicted event but no observed event. Although truly perfect scores were achieved only at grid scale, the impact of the leaking scores was small except at the largest scales examined (>100 grid points), which was greater than the minor axis of the ellipse.

The neighborhood verification results for cases geom001 (small separation), geom002 (large separation), and geom005 (moderate separation, biased high and overlapping) are shown in Fig. 2. For geom001 there was no skill at grid scale because the forecast and observed ellipses did not overlap. As the scale was increased, some skill started to appear at the lower rain rates. The performance was best at the large scale for all methods; this is because the forecast and observed ellipses could both be enclosed in some of the neighborhoods. In fact, “useful” skill was achieved at the 129-point scale according to the FSS. The geom002 case had a much larger separation, with the result that the forecast showed no useful skill at any spatial scale.

Case geom005 had a significant area of overlap so that even at grid scale some small level of skill was indicated by each verification method for all but the heaviest rainfall. The skill generally improved with increasing scale, but not markedly; this is because the forecast and observed ellipses could both be enclosed in some of the neighborhoods. In fact, “useful” skill was achieved at the 129-point scale according to the FSS. The geom002 case had a much larger separation, with the result that the forecast showed no useful skill at any spatial scale.

According to the traditional grid-scale metrics, geom005 is the slightly better forecast, with a GSS of 0.08 compared to −0.01 for geom001. Since for each of these neighborhood methods the scores are positively oriented (higher values correspond to better performance), their differences between one forecast and another will indicate which one has the better performance. The difference between the neighborhood scores for geom001 and geom005 is shown in Fig. 3. At the smaller scales, geom005 indeed performed better than geom001 for all but the heaviest rain. At the larger scales (>129 points), geom001 performed much better than geom005, since the observed and forecast features may be nearly enclosed in the same neighborhood.

b. Perturbed cases

The seven perturbed cases in the intercomparison project were created from a 24-h model forecast (see section 4) of hourly rainfall valid at 0000 UTC on 1 June 2005 (Fig. 4). The idealized forecasts were created by translating and amplifying the observed rain to simulate known forecast errors. The ability of the neighborhood verification methods to describe the known errors is examined here for two of the cases: pert004 (field shifted 96 km to the east, 160 km to the south) and pert006 (field shifted 48 km to the east, 80 km to the south, rainfall multiplied by 1.5).

The neighborhood verifications for pert004 and pert006 and their differences are shown in Fig. 5. Traditional grid-based verification shows a higher value of GSS for the pert006 case (0.17) than the pert004 case (0.08), presumably because the smaller location error is less detrimental to the forecast than the imposed bias error. According to the upscaling and fractions skill score methods, the level of forecast performance for both forecasts improved with increasing spatial scale and peaked at the lowest rainfall threshold. This behavior is

<table>
<thead>
<tr>
<th>Spatial scale (pts)</th>
<th>Threshold (mm)</th>
<th>Upscaling – GSS</th>
<th>Fractions skill score – FSS</th>
<th>Multi-event cont. table – HK</th>
<th>Proc. perf. – GSS ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>257</td>
<td>1</td>
<td>1.00, 1.00</td>
<td>1.00, 1.00</td>
<td>1.00, 1.00</td>
<td>1.00, 1.00</td>
</tr>
<tr>
<td>129</td>
<td>1</td>
<td>1.00, 1.00</td>
<td>1.00, 1.00</td>
<td>1.00, 1.00</td>
<td>1.00, 1.00</td>
</tr>
<tr>
<td>65</td>
<td>1</td>
<td>1.00, 1.00</td>
<td>1.00, 1.00</td>
<td>1.00, 1.00</td>
<td>1.00, 1.00</td>
</tr>
<tr>
<td>33</td>
<td>1</td>
<td>1.00, 1.00</td>
<td>1.00, 1.00</td>
<td>1.00, 1.00</td>
<td>1.00, 1.00</td>
</tr>
<tr>
<td>17</td>
<td>1</td>
<td>1.00, 1.00</td>
<td>1.00, 1.00</td>
<td>1.00, 1.00</td>
<td>1.00, 1.00</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>1.00, 1.00</td>
<td>1.00, 1.00</td>
<td>1.00, 1.00</td>
<td>1.00, 1.00</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1.00, 1.00</td>
<td>1.00, 1.00</td>
<td>1.00, 1.00</td>
<td>1.00, 1.00</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1.00, 1.00</td>
<td>1.00, 1.00</td>
<td>1.00, 1.00</td>
<td>1.00, 1.00</td>
</tr>
</tbody>
</table>

FIG. 1. Neighborhood verification of a perfect idealized geometric forecast (refer to text). The intensity threshold varies along the x axis and the spatial scale varies along the y axis. The score value is shaded with darker shades indicating better performance, and also printed for clarity. Blank entries represent intensity-scale combinations for which there were no observations or the score was undefined.
typical for real precipitation forecasts (Zepeda-Arce et al. 2000; Roberts and Lean 2008), and reflects the difficulty in accurately predicting small-scale high-intensity precipitation features. The MECT results suggest that the optimal search diameter increases with rain threshold. For rain exceeding 10 mm h$^{-1}$, it has a value of about 260 km for pert004 and 130 km for pert006, similar to the imposed displacements of about 200 and 100 km, respectively. The skillful scales diagnosed by the FSS method resembled the MECT optimal search diameters. GSSratio for the practically perfect hindcast method peaked at large scales, indicating that greater "fuzzification" increased the resemblance between the real forecast and the practically perfect hindcast.

The difference plots (Fig. 5c) show that for most spatial scales and rainfall intensities, the biased pert006 forecast was more skillful than the unbiased pert004 forecast with the larger displacement. Only for the largest scales, when the displacement error was of lesser significance, did the pert004 forecast outperform pert006 and only according to the FSS and MECT methods.

Pert004 and pert006 differed in both their bias and their displacement errors. To isolate the effects of forecast bias on the verification results, pert006 was compared to pert003, which had exactly the same displacement as pert006 but was unbiased. It is reasonable to expect that pert003 will show better verification scores than pert006 since only one kind of error was applied. However,
Fig. 6 shows that this was not the case. The biased pert006 outperformed the unbiased pert003 for scales smaller than about 100–200 km and rain thresholds of 10–20 mm h$^{-1}$ or less, depending on the neighborhood method. The effect of the positive forecast bias was to increase the size of the rain area exceeding the various thresholds, allowing for greater overlap of the forecasts with the observations, thus leading to improved scores. This is consistent with the theoretical results of Baldwin and Kain (2006), who showed that for spatially displaced forecasts of infrequent events, the optimum score for commonly used verification metrics was achieved for a positively biased forecast. The comparison of the geometric cases geom001 and geom005 showed this behavior also (Fig. 3).

While the neighborhood verification results can be explained for the idealized cases, the scores shown so far do not always match with our initial expectations, nor have they proven very useful for identifying the nature of the errors. For some of the methods it is possible to compute other categorical scores such as frequency bias, probability of detection, and false alarm ratio, to better understand whether or not the errors are related to bias. The slightly unsatisfying results shown so far might discourage verifiers from considering neighborhood verification as a useful approach. However, the intent of
neighborhood verification is to give credit to close forecasts, not diagnose the source of the forecast errors. This point will be discussed further in section 5. As will be seen in the next section, the real strength of neighborhood verification is in showing at which scales the forecast has useful skill for each intensity threshold.

4. Neighborhood verification of high-resolution model QPFs from the 2005 Spring Program

Nine 24-h forecasts of 60-min rainfall from each of three different configurations of the Weather Research and Forecasting (WRF) model, produced as part of the Storm Prediction Center’s 2005 Spring Program (Kain et al. 2008), were included in the intercomparison project. The goal was to test whether the objective verification from the various spatial methods matched well with the participants’ subjective opinions. The models included a 2-km grid (Advanced Research WRF (WRF-ARW) simulation run by the Center for the Analysis and Prediction of Storms (CAPS), a 4-km grid WRF-ARW simulation run by the National Center for Atmospheric Research (NCAR), and another 4-km version of the Nonhydrostatic Mesoscale Model (WRF-NMM) run by the National Centers for Environmental Prediction (NCEP). These models will be abbreviated here as CAPS2, NCAR4, and NCEP4, respectively. All model runs were initiated at 0000 UTC and verified for the period 2300–2400 UTC. The output was remapped to the g240 Lambert conformal grid used by the stage II hourly radar-based rainfall analysis (Lin and Mitchell 2005), which provided the observational data.
The overall performance for each of the three models was computed by aggregating its neighborhood verification results over the nine 2005 Spring Program cases. The aggregation was done by summing the contingency table elements (upscaling, MECT, and PP methods) or squared errors (FSS) for all of the cases and computing the scores from the summed components. As in any systematic verification, aggregation over many cases characterizes the overall performance of the forecast system, giving users a better idea of what they can expect for independent forecasts.

The aggregated neighborhood verification results for the 24-h QPFs are shown in Fig. 7. The three models performed similarly in most respects. In general, the performance was better for larger scales and lighter intensity thresholds. In particular, rainfall at scales of about 250 km or more was predicted with useful skill according to the FSS. Even rainfall in excess of 20 mm h\(^{-1}\) was well predicted by the CAPS2 model when a scale of 500 km was considered (Fig. 7a), although this scale may be too large to be considered useful for many applications. The MECT method indicates that an optimal search diameter for forecast events located near observed events was on the order of 100–250 km, depending on the intensity of the event.

The grid-scale values of GSS were quite low, with a maximum of 0.10 for the NCAR4 model and a light rain threshold (1 mm h\(^{-1}\)). If the grid-scale verification results were the only ones considered, these scores would suggest that the models did a very poor job of predicting the rainfall 24 h in advance. When the rain fields were upscaled to 100 km or more, the GSS values increased to twice the grid-scale value, but were still rather small. Putting the skill into context using the PP method, at 100-km scale the rain forecasts had less than a quarter of the possible skill using practically perfect hindcasts, while they achieved more than half of the “practically perfect” skill at 500-km scale.

To help visualize the sort of forecast that would perform at the “typical” level of skill shown in Fig. 7, the forecast whose neighborhood verification results best matched the aggregate scores was identified. The NCAR4 forecast valid at 0000 UTC on 1 June 2005 is shown in Fig. 8; the verifying observations were shown in Fig. 4. The forecast rain systems looked similar in shape and intensity to the observations, but were displaced to the west, possibly as a result of a mistimed frontal passage. The forecast rain area in the southeastern United States was smaller than was observed, although more intense. The threat area for 1 mm h\(^{-1}\) rainfall diagnosed by the PP method at a scale of 260 km shows that the forecast area bore a reasonable resemblance to the practically perfect threat area (Fig. 9).

5. Discussion

Four neighborhood verification methods have been described and demonstrated here using idealized and real precipitation forecasts from the Spatial Forecast Verification Methods Intercomparison Project (Gilleland et al. 2009; AGBE). Neighborhood verification gives credit to “close” forecasts by relaxing the requirement for exact matches with the observations at grid scale. It thereby addresses many of the problems identified by Barnes et al. (2007) of close forecasts not being recognized as useful using traditional verification approaches. The upscaling method evaluates the average forecast values at increasing scales, while the FSS method compares the forecast and observed fractional coverage. The MECT method verifies the occurrence of a forecast event nearby an observed event, and the PP method examines the scale-dependent verification score within the context of that which would be achieved by a practically perfect hindcast. These four methods address four different aspects of forecast goodness. Other neighborhood methods are also available but were not demonstrated there (e.g., Ebert 2008).
One goal of the intercomparison project was to determine how well various spatial verification methods could diagnose known errors in high-resolution forecasts. The neighborhood verification methods did not turn out to be very good at this because they did not differentiate between different sources of error. A few surprising results emerged. Methods that compare a neighborhood of forecasts to a single observation in the center did not indicate perfect performance when applied to a perfect forecast, due to the influence of nearby grid boxes. Tests with perturbed forecasts showed that displaced forecasts that were biased high actually gave better scores than forecasts with identical displacements but no bias for most scales and intensities. This was predicted by Baldwin and Kain (2006), but it goes against our instinctive feeling that unbiased forecasts should be better. The relevant question is, what is meant by better? Model developers strive to predict realistic rain with as few errors in location and amount as possible; in their view an unbiased forecast is certainly better than a biased one. But from the point of view of the user who wants to get the best predictions for a particular location, forecasts with (possibly compensating) errors may be more useful if they provide more accurate or valuable rain forecasts for the location of interest. As long as the chosen verification metric reflects what is important to the user, then forecasts that are worse for the modeler may actually be better for the user!

Neighborhood verification methods address a variety of decision models concerning what makes a useful
forecast. This can also be said about the variety of verification scores commonly used for grid-scale verification. By performing the verification at a variety of spatial scales and intensity thresholds, the neighborhood approach makes it possible to determine for which scales a forecast has useful skill for events of a given intensity. In fact, two of the methods, FSS and MECT, lend themselves easily to a “skillful scale” interpretation. Trends in forecast performance over the years can be assessed by plotting the skillful scale for one or more key intensity thresholds, or by examining score differences as done in Fig. 6.

Finally, it is useful to compare neighborhood verification with other spatial verification methods. If the aim of the verification is to help diagnose the source of errors in the forecast, then neighborhood verification is not an appropriate approach. Scale-separation methods can isolate the scale-dependent error, whereas neighborhood methods merely identify the scales at which the errors of the filtered forecast are sufficiently small for a particular application. For forecasts with well-defined features, features-based methods such as the Method for Object-based Diagnostic Evaluation (MODE; Davis et al. 2006, 2009) and the contiguous rain area (CRA; Ebert and McBride 2000) are effective at diagnosing errors in location, size, intensity, and other attributes of the forecast. Features-based verification methods normally require specification of various parameters such as the threshold for defining an object and the search distance used in matching forecast and observed objects. As shown by Ebert and Gallus (2009), some verification results may be somewhat sensitive to the choice of parameters. Neighborhood verification methods do not have selectable parameters; rather, all combinations of threshold intensities and spatial scales (and temporal scales if time neighborhoods are used) are evaluated. Furthermore, unlike feature-based methods, neighborhood methods work well for verifying forecasts that are “messy” and do not contain well-defined or well-matched features.

Neighborhood verification is starting to be used routinely at the Met Office, Météo-France (F. Rabier 2008, personal communication), and the Australian Bureau of Meteorology. As high-resolution modeling becomes increasingly important as a source of numerical guidance for weather forecasters and other users, it is likely that neighborhood verification (along with other spatial verification approaches) will become more commonly practiced.

Acknowledgments. I would like to thank many colleagues for interesting and thought-provoking discussions of neighborhood verification, especially Dave Ahijevych, Felix Ament, Barb Brown, Barbara Casati, Ulrich Darrath, Eric Gilleland, Marion Mittermaier, Pertti Nurmi, and Francis Schubiger. The Forecast Evaluation and
Applied Statistics group at NCAR/RAL, along with Mike Baldwin, were instrumental in setting up and coordinating the Spatial Verification Methods Intercomparison Project. I also thank the three anonymous reviewers for providing many useful suggestions that improved the paper.

REFERENCES


