Intercomparison of Spatial Forecast Verification Methods

ERIC GILLELAND, DAVID AHIJEVYCH, AND BARBARA G. BROWN

National Center for Atmospheric Research, Boulder, Colorado*

BARBARA CASATI

Ouranos, Montreal, Quebec, Canada

ELIZABETH E. EBERT

Center for Australian Weather and Climate Research, Melbourne, Victoria, Australia

(Manuscript received 3 February 2009, in final form 24 May 2009)

ABSTRACT

Advancements in weather forecast models and their enhanced resolution have led to substantially improved and more realistic-appearing forecasts for some variables. However, traditional verification scores often indicate poor performance because of the increased small-scale variability so that the true quality of the forecasts is not always characterized well. As a result, numerous new methods for verifying these forecasts have been proposed. These new methods can mostly be classified into two overall categories: filtering methods and displacement methods. The filtering methods can be further delineated into neighborhood and scale separation, and the displacement methods can be divided into features based and field deformation. Each method gives considerably more information than the traditional scores, but it is not clear which method(s) should be used for which purpose.

A verification methods intercomparison project has been established in order to glean a better understanding of the proposed methods in terms of their various characteristics and to determine what verification questions each method addresses. The study is ongoing, and preliminary qualitative results for the different approaches applied to different situations are described here. In particular, the various methods and their basic characteristics, similarities, and differences are described. In addition, several questions are addressed regarding the application of the methods and the information that they provide. These questions include (i) how the method(s) inform performance at different scales; (ii) how the methods provide information on location errors; (iii) whether the methods provide information on intensity errors and distributions; (iv) whether the methods provide information on structure errors; (v) whether the approaches have the ability to provide information about hits, misses, and false alarms; (vi) whether the methods do anything that is counterintuitive; (vii) whether the methods have selectable parameters and how sensitive the results are to parameter selection; (viii) whether the results can be easily aggregated across multiple cases; (ix) whether the methods can identify timing errors; and (x) whether confidence intervals and hypothesis tests can be readily computed.

1. Introduction

Small-scale variability in high-resolution weather forecasts presents a challenging problem for verifying forecast performance. Traditional verification scores provide incomplete information about the quality of a forecast because they only make comparisons on a point-to-point basis with no regard to spatial information [Baldwin and Kain (2006); Casati et al. (2008); see Wilks (2005) and Jolliffe and Stephenson (2003) for more on traditional verification scores]. For example, a forecast feature with the correct size and structure might yield very poor verification scores if the feature is displaced slightly in space because it will be penalized once for missing the observations and again for giving a false alarm; this is known as the ''double penalty.'' Higher variability (e.g.,

* The National Center for Atmospheric Research is sponsored by the National Science Foundation.

Corresponding author address: Eric Gilleland, Research Applications Laboratory, National Center for Atmospheric Research, P.O. Box 3000, Boulder, CO 80307-3000.
E-mail: ericg@ucar.edu

DOI: 10.1175/2009WAF2222269.1

as often occurs with higher-resolution forecasts) leads to a greater likelihood of having a larger amount of small-scale intensity error. Under such circumstances, this double penalty can become problematic in judging the true quality of a forecast. Similarly, because of the spatial coherence of the features, it is likely that a displaced forecast feature will score better by simply inflating the forecasted values at each grid point, thereby inflating the spatial extent of the feature. Baldwin and Kain (2006) investigated the sensitivity of several commonly used scores [Gilbert skill score (GSS; also known as the equitable threat score), true skill score (TSS), odds ratio, etc.] to changes in event frequency, spatial displacement, and bias. Among their findings, they showed that the behavior of several scores, including GSS, TSS, and odds ratio, were highly sensitive to event frequency. For more frequently occurring events, the scores were found to be more sensitive to displacement errors.

In response to these undesirable properties of traditional verification methods when applied to high-resolution forecasts, researchers have proposed numerous new verification methods. Here, attention is focused on the verification of gridded forecasts with an observation field that is on the same grid (though this is not necessary for some methods), but note that methods that address the verification of forecasts on one scale against observations on a different scale do exist (e.g., Tustison et al. 2003).

The majority of the new techniques can be broadly grouped into four categories, illustrated schematically in Fig. 1: (i) neighborhood (or fuzzy), (ii) scale separation (or scale decomposition), (iii) features based (or object based), and (iv) field deformation. The first two categories can be more generally described as filtering methods as both apply a spatial filter to one or both fields (or sometimes to the difference field), and then calculate verification statistics on the filtered fields. The filter is usually applied at progressively coarser scales to provide information about the scale(s) at which the forecast has skill. The neighborhood methods apply a smoothing filter, whereas the scale-separation techniques apply several single-bandpass spatial filters (Fourier, wavelets, etc.) so that performance at separate scales can be evaluated independently.

The features-based and field deformation categories are similar in that they both try to ''fit'' the forecast to the observations as well as possible. They then give information about how much the forecast field needs to be manipulated spatially (displacement, rotations, scaling, etc.) and quantify the residual errors to obtain a more meaningful notion of skill. The primary difference between the two is that the features-based methods first identify features of interest (e.g., storm cells), and analyze each feature separately, whereas the field deformation approaches analyze the entire field or a subset thereof. Therefore, the two categories can be broadly thought of as displacement methods because in contrast to finding the scale at which the desired skill is achieved, these methods describe how much spatial movement is required in order to match the forecast field to the observed field. Of course, not all of the methods fall nicely into one of these four categories, and it is possible for some methods to fit into more than one category. Discrepancies such as these are pointed out where appropriate.

Because new spatial verification methods have only fairly recently been introduced, many of them are not yet in wide usage. Potential users may be confused as to which one(s) may be most appropriate for their particular application. There is a need to analyze their characteristics and determine how they compare to one another. Specifically, it is important to know the kinds of information each method provides [e.g., the verification question(s) addressed], the type of data required (regular grid, normally distributed, etc.), how well suited each method is for operational or diagnostic use, and whether different approaches provide similar information but from different perspectives. It is also important to know whether a statistical model can be formulated to characterize uncertainty concerning forecast performance, or failing that, whether a nonparametric method such as bootstrap resampling can be utilized instead. These are the impetuses for the spatial verification method intercomparison project (ICP; information available online at http://www.ral.ucar.edu/projects/icp/), which is a metaverification project including many of the researchers who have proposed these methods.[1] Specifically, the goals of the ICP are to critically compare the various approaches both qualitatively and quantitatively, assessing to what extent they provide useful information beyond that given by traditional verification methods. The project also seeks to identify differences and commonalities among the methods, as well as to characterize the information they provide for certain well-defined cases. Although the project does not compare every single strategy introduced, it includes a good fraction of them at the time of writing.

The ICP has made available a number of test cases. The cases currently being studied include real examples of quantitative precipitation forecasts and verifying radar–gauge precipitation analyses, as well as perturbations from one of these real cases so that results can be compared

[1] Participants in the ICP are volunteers who predominantly are applying their own methods. Anyone is welcome to participate, however, and may begin by visiting the ICP Web site (given above) and signing up on the e-mail list.

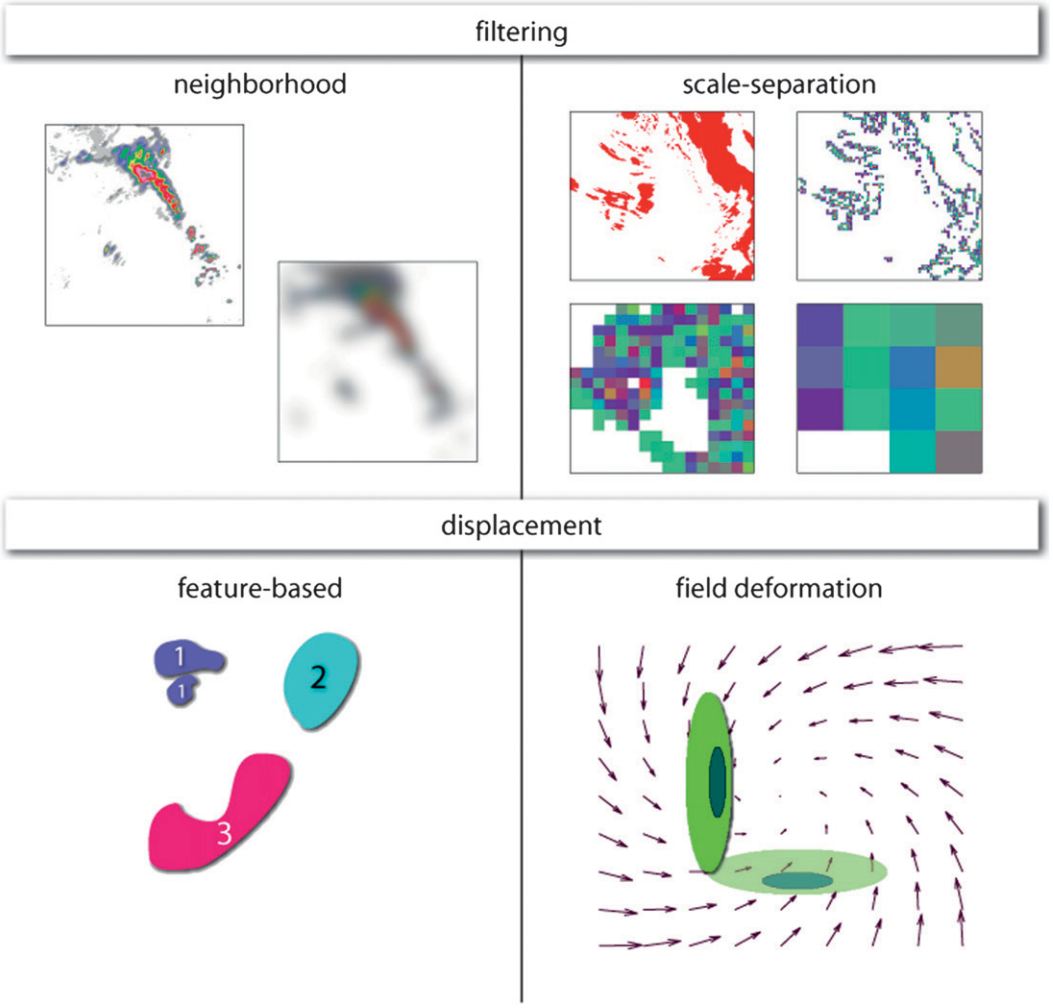

FIG. 1. Schematic representations of the four categories of verification methods reviewed in this paper. (top) The neighborhood and scale-separation methods can both be considered "filtering" approaches while (bottom) the feature-based and field deformation methods fall under the "displacement" category.

with "known" errors. Additionally, some very simple geometric cases are being used and have, thus far, proven to be very useful for gaining a better understanding of the verification methods because of their simple structures. The geometric, perturbed, and real cases are described in detail in Ahijevych et al. (2009, hereafter AGBE) along with select quantitative results from most of the methods described herein. These two summary papers, along with the unabridged papers on each method, compose this special collection of *Weather and Forecasting*.

The next section provides a literature review along with summary information for the various methods considered by the ICP. Section 3 compares each method qualitatively, based on a series of verification questions aimed at determining which methods are useful for which types of users. Finally, section 4 summarizes the methods and provides some discussion on the goals of the ICP, difficulties associated with such an intercomparison, and some intuition about which types of methods are best suited for particular applications.

2. Description and comparison of methods by type

The following subsections give some details about each of the four general categories of spatial verification methods. Table 1 provides a list of abbreviations used here for the individual methods represented in the ICP along with the general category in which they belong and some references.

Before discussing the new methods, a review of certain aspects of traditional verification statistics is presented (see also Wilks 2005; Jolliffe and Stephenson 2003). When comparing continuous values (e.g., temperature,

TABLE 1. List of individual methods considered in this paper, and the ICP, along with their abbreviations used here. References listed are not comprehensive; see the text and the references for further representative works.

Abbreviation	Description	Method type	Reference(s)
BCETS	Bias-corrected ETS	Traditional	Mesinger (2008)
CA	Cluster analysis	Features based*	Marzban and Sandgathe (2006, 2008)
Composite	Composite method	Features based*	Nachamkin (2005, 2009)
CRA	Contiguous rain area	Features based	Ebert and McBride (2000); Ebert and Gallus (2009)
DIST	Distributional method	Neighborhood	Marsigli et al. (2006)
FQI	Forecast quality index	Field deformation*	Venugopal et al. (2005)
FQM–DAS	Forecast quality measure–displacement amplitude score	Field deformation	Keil and Craig (2007, 2009)
FSS	Fractions skill score	Neighborhood	Roberts (2005); Roberts and Lean (2008); Mittermaier and Roberts (2009)
IS	Intensity scale	Scale separation	Casati et al. (2004); Casati (2009)
IW	Image warping	Field deformation	E. Gilleland, J. Lindström, and F. Lindgren (2009, unpublished manuscript); Lindström et al. (2009)
MODE	Method for Object-based Diagnostic Evaluation	Features based	Davis et al. (2006, 2009)
MSV	Multiscale variability	Scale separation	Zapeda-Arce et al. (2000); Harris et al. (2001); Mittermaier (2006)
Neighborhood	Neighborhood based methods	Neighborhood	Ebert (2008, 2009)
Procrustes	Cell identification and Procrustes shape analysis	Features based	Micheas et al. (2007)
Procrustes2	Multiscale cell identification and Procrustes shape analysis	Scale separation–Features based	Lack et al. (2009)
SAL	Structure, amplitude, and location	Features based	Wernli et al. (2008, 2009)
Traditional	Point-based comparison	Point	Jolliffe and Stephenson (2003)
VGM	Variogram	Scale separation*	Marzban and Sandgathe (2009)

* A method that only loosely belongs to the given method type.

relative humidity, precipitation amount) statistics such as RMSE, mean error, mean absolute error, etc. can be calculated using forecasts and observations at each grid point. Alternatively, the values can be categorized into levels, for example by thresholding, and statistics based on the resulting contingency tables can be calculated. The simplest case is a 2×2 contingency table (e.g., 24-h accumulated precipitation less than or equal to 0.1 mm versus greater than 0.1 mm). In such a case, a hit occurs at a grid point when an event is correctly predicted (i.e., in this case, 24-h accumulated precipitation exceeding 0.1 mm was forecast and it occurred). Similarly, correct negatives result from correctly forecasting nonevents (24-h accumulated precipitation less than or equal to 0.1 mm). A false alarm occurs when the forecast predicts the event, but it does not occur, and a miss is an observed event that was not forecast. Finally, a correct negative occurs when no precipitation greater than 0.1 mm is forecast and none occurs. In terms of spatial verification, one can imagine partially overlapping areas of forecast and observed rainfall, where the overlap area represents the hits, the area where rain was observed but not forecast represents the misses, the area where rain was forecast but not observed represents the false alarms, and the area with no rain forecast or observed represents the correct negatives.

Numerous verification statistics utilize the information from such a contingency table. The GSS statistic is frequently used because it corrects for the number of hits expected to occur by chance. When the forecast is unbiased, then higher values of GSS indicate better forecasts without sensitivity to large numbers of correct negatives. However, the GSS results are misleading if the forecast is not unbiased (i.e., the frequency of forecast events is different from the frequency of observed events). Certainly, if a forecast over- or underpredicts the spatial extent of the observations, then the forecast is biased. The bias-corrected GSS introduced by Mesinger (2008) corrects for this problem by assuming a nonlinear relationship between false alarms and misses.

a. Neighborhood approaches

The neighborhood (also known as fuzzy) approaches compare values of forecasts and observations in space–time neighborhoods relative to a point in the observation field. Properties of the fields within neighborhoods

(e.g., mean, maximum, existence of one or more points exceeding a certain threshold) are then compared using various statistical summaries, which are often simply the traditional verification statistics. Such comparisons are typically done for incrementally larger neighborhoods so that it is possible to determine the scale at which a desired level of skill is attained by the forecast. The top-left panel in Fig. 1 depicts this general idea by showing a field that has been upscaled by averaging the values of neighbors of grid points within a certain radius of each other. The result is a smoothed version of the original field. Summary statistics, such as traditional verification statistics, can be applied to the smoothed field. The process is typically repeated using increasingly larger neighborhoods.

Several independently developed techniques were summarized and interpreted into a common framework by Ebert (2008). Among some of the qualitative advantages of these approaches are (i) the parsimony of the techniques, (ii) the use of many of the familiar traditional scores, (iii) the ability to determine at which resolutions the forecast performs best, and (iv) reduction of the double-penalty problem. The particular verification questions addressed by these procedures depend largely on the traditional score utilized and how the neighborhoods are aggregated, as well as how the neighborhoods are defined. A brief summary is given here.

The earliest and perhaps simplest of these methods is referred to as upscaling, whereby the forecasts and observations are averaged to consecutively coarser scales and compared using traditional scores (e.g., Yates et al. 2006; Zepeda-Arce et al. 2000; Weygandt et al. 2004). One of the disadvantages of upscaling is the loss of small-scale variability that is crucial for depicting high-impact events such as extreme wind or precipitation. This variability may be properly captured by a high-resolution model, but the high-intensity wind or precipitation event locations may be displaced slightly from the observed high-intensity events. Traditional methods would not give credit for these near misses, but various neighborhood methods can by preserving the original grid values and looking in the neighborhood around each grid point for events.

The fractions skill score (FSS) of Roberts (2005) and Roberts and Lean (2008) compares the fractional coverage of events (occurrences of values exceeding a certain threshold) in windows surrounding the observations and forecasts (see also Mittermaier and Roberts 2009). Theis et al. (2005) compared the forecast fractional coverage in the neighborhood to the occurrence of an observed event at a point. Marsigli et al. (2006) take a more general approach by comparing moments of the distribution of observations in the neighborhoods with the moments of the distribution of forecasts in the neighborhoods. Damrath (2004) utilizes two approaches: one that employs a proportion threshold within the neighborhood to determine whether an event has occurred or not, and one that employs a fuzzy logic technique that defines events as the probabilities themselves. Scores that evaluate the forecast intensities were proposed by Germann and Zawadzki (2004) and Rezacova et al. (2007). Brooks et al. (1998) address the issue of rare-event verification by comparing the performance of the forecast to that achieved using a ''practically perfect hindcast'' obtained by objectively analyzing the observations.

Event frequency can be very sensitive to both thresholds and neighborhood size. For this reason, the neighborhood methods generally consider a range of spatial resolutions and intensity thresholds. Atger (2001) uses a multievent contingency table approach that allows for several intensity thresholds to be evaluated as well as other dimensions such as spatial or temporal proximity.

b. Scale separation/decomposition

Scale-separation approaches provide information on forecast performance on distinct scales. The different scales are obtained with a single-band spatial filter (Fourier transforms, wavelets, etc.), whereby one investigates forecast performance by isolating the features at each scale. These scales are often representative of physical features, such as large-scale frontal systems or small-scale convective showers. This is depicted in the top-right panel in Fig. 1, where the large red blob represents a large-scale storm system and the graph with the multicolored blob shows the individual small-scale storms within the system. The scale-separation approaches aim to (i) assess the scale dependency of the error, (ii) determine the skill–no-skill transition scale (i.e., assess the scale dependency of the model predictability), and (iii) assess the capability of the forecast to reproduce the observed scale structure in the observations.

One important difference between the neighborhood and scale-separation approaches is that the neighborhood approach essentially smoothes the forecasts over a range of increasing scales and the filtered fields always broadly resemble the original field because the large scale is retained. At the same time, the filtered fields cannot be combined to reproduce the original field because the smallest scales are lost in the smoothing process. In contrast, the scale-separation approaches treat each scale independently; the filtered fields may not resemble the original field, but they may be combined to reproduce the original field.

Briggs and Levine (1997) introduced the first scale-separation verification approach using wavelets. They assessed geopotential height fields by decomposing them

with a 2D wavelet filter, and then evaluating traditional continuous scores (correlation, ratio of the variances, and RMSE) for each scale component.

The intensity-scale (IS) technique of Casati et al. (2004) measures skill as a function of the scales and of the intensity. Recalibrated forecast and observation fields are transformed into binary images by thresholding for different intensities, and the point-to-point difference between the binary fields is taken to obtain binary error fields. These error fields are subsequently separated into the sum of different scale components using a two-dimensional Haar wavelet decomposition, and a skill score based on the mean squared error of these images is evaluated for each scale component and intensity threshold. The result is a Heidke skill score evaluated at different scales, thereby linking categorical scores with the scale-separation approach. Casati (2009) modifies this approach by not recalibrating the forecast field, and subsequently computing the energies and the relative difference in energy at each scale. This allows one to assess the bias between the two fields at different scales.

The wealth of detailed information this method provides is useful within a diagnostic context, but for operational verification, a method is needed to condense this detail into manageable and easy to understand quantities. To address this need, Mittermaier (2006) expanded the IS idea by presenting a method for aggregating results for individual (operational) forecasts produced from the intensity-scale analysis. The approach was applied to compare the performance of the 12- and 4-km versions of the Unified Model against radar rainfall and gridded gauge analyses. Casati (2009) presents a different approach for aggregating the IS statistics, and provides confidence intervals on these statistics using a bootstrapping approach.

Harris et al. (2001) look at multiscale statistical properties related to the spatiotemporal scale structure of the two fields. In particular, forecast performance is investigated by evaluating the Fourier spectrum, structure function, and moment-scale analyses. The method differs from other spectral decomposition methods in that verification is not performed separately on different scales. Because the technique investigates the two fields individually, information about the marginal distributions is gleaned rather than the joint distribution, and an assessment can be made on whether the forecast looks realistic in terms of its scale-dependent spatial properties.

Marzban and Sandgathe (2009) and Marzban et al. (2009) investigate variogram plots of forecast and observed fields to compare the *textures* of the fields, which gives an indication of the similarities–discrepancies of a marginal distributional property of the two fields. It is easy to show that the covariance of the coefficients from a wavelet decomposition implies that the field is a random surface also described by a covariance function. The variogram and covariance are equivalent characterizations for a random surface that is stationary and Gaussian distributed so that the variogram of the field is also tied to the covariance of the coefficients of the wavelet decomposition. In this way the variogram technique is a type of scale-separation method (cf. Nychka 1998; Ogden 1996). It should be emphasized, however, that the variogram approach (VGM), like the Harriss et al. (2001) approach, investigates distributional properties of the two fields and, therefore, differs in its aims from those outlined at the beginning of this section. Finally, VGM computes the variogram in two different ways by calculating it (i) on only the nonzero elements of the field and (ii) on the entire field. The first approach yields information on the texture of rain areas, and the second yields information about size and displacement errors, as well as the texture of the entire field. When applied to the entire field, large discrepancies in the spatial extent or locations of storms can affect the resulting variogram shape insofar as such differences affect the texture of the entire field. Because of the sensitivity to size and displacement errors, Marzban and Sandgathe (2009) refer to VGM as a kind of object-based technique. Here, we would not classify it as such because it does not identify individual features. In fact, it is a method that does not fit well into any of the four categories. It would perhaps be best categorized into a fifth category of field distribution methods because it most directly compares a distributional property of the two fields.

Lack et al. (2009) introduce a new multiscale approach that overlaps with the features-based approach. A Fourier transform is used to associate signals within convection to different spatial scales. A user-defined weighted cost function is then employed to match objects, characterize them as being more linear versus more cellular, and identify matches as hits, misses, or false alarms. The technique is an advancement of the method proposed in Micheas et al. (2007), which is discussed in the next section.

c. Features-based approaches

Numerous methods have been proposed to look specifically at how well the forecast captures the overall structure of meteorological features. These methods are referred to as features-based, object-based, and cell-identification techniques. The primary differences among these approaches are how they determine (i) what constitutes a feature, (ii) whether spatially discontinuous features within a field should be treated as one feature or separate features, (iii) how they match features from one

field to the other, and (iv) what sorts of diagnostics and/or summary measures they produce. Most of the methods identify features by applying a threshold to the fields. The general schematic is shown in the bottom-left panel of Fig. 1, where three features are identified, the first of which is a merging of two spatially discontinuous objects. Attributes such as size, shape, and average intensity over each object can be calculated for each individual object within a forecast or observation field. Additionally, they can be compared against identified features in the corresponding field (e.g., based on spatial proximity), and statistics pertaining to how well these features compare can be calculated.

The contiguous rain area (CRA) approach of Ebert and McBride (2000) associates forecast and observed features that overlap. [To associate features that are close but not touching, the features can be smoothed before checking for overlap; see Ebert and Gallus (2009)]. Optimal matching is attained by translating the forecast until a pattern-matching criterion is met (e.g., minimum squared error or maximum correlation between the gridpoint values in the observed and forecast features). Displacement, volume, and pattern errors are found as a natural consequence of this procedure, and their contributions to the total error can be quantified. Various modifications to this approach have been proposed to fine-tune it for short-period forecasts (e.g., Ebert et al. 2004; Grams et al. 2006).

Baldwin and Lakshmivarahan (2003) present a features-based technique aimed at discriminating between three phenomena in a rainfall field: linear, cellular, and stratiform precipitation. The technique is multifaceted with an initial step involving a hierarchical statistical clustering analysis to isolate stratiform events. To distinguish between cellular and linear events, they employ techniques from geostatistics, in particular the correlogram. Principal component analysis is then employed to determine attributes that provide unique information on the resulting correlation matrix.

The method developed by Davis et al. (2006), now called the Method for Object-based Diagnostic Evaluation (MODE), addresses feature identification not solely by applying an intensity threshold to the field, but also by a convolution procedure whereby the fields are first smoothed over space and then thresholded. Once contiguous nonzero pixels (i.e., features) are identified, they are merged and matched by an algorithm utilizing information about various attributes (e.g., centroid position, total area, area overlap, intensity distribution, orientation angle, and boundary separation). Gilleland et al. (2008) propose an alternative method for merging and matching features for MODE based solely on a binary image distance measure, known as Baddeley's Δ metric. Another simple option is the partial Hausdorff distance (Venugopal et al. 2005). MODE assigns user-defined weights and confidence to these attributes and combines them in a fuzzy logic algorithm to produce a total interest for each feature pair. To get a representative total interest value over all pairs, Davis et al. (2009) devised a metric called the median of the maximum interest. This measure summarizes all of the total interest values for each combination of forecast feature and observed feature into a single number.

Nachamkin (2004) uses composites of wind events to examine the distribution of forecasted events relative to those observed, and vice versa. Nachamkin et al. (2005) perform the technique for precipitation fields, and Nachamkin (2009) applies the procedure to the ICP test cases. Events are defined as spatially contiguous regions of precipitation intensities or amounts above a particular threshold. The method first identifies all of the observed rainfall events, centers them on a common grid, and composites them. Differences between the observed composite and the corresponding forecasted rainfall composite give clues about the model's tendency to overforecast, underforecast, and misplace precipitation events. The converse approach is also employed, whereby one obtains the conditional distribution of observed precipitation given that a forecast event occurred. In Nachamkin (2004), evidence from this verification technique pointed to shortcomings in the model's convective parameterization scheme. We classify this approach as features based because it identifies individual features, but it does not provide information about individual features. Similar to the VGM approach, it might be better classified into a fifth category of field distribution.

The technique proposed by Marzban and Sandgathe (2006, 2008) applies hierarchical statistical cluster analyses whereby features are identified by clusters at each iteration of the procedure. Verification measures such as the critical success index are calculated by defining hits, misses, and false alarms based on the proximity of clusters between the two fields using a distance metric. In Marzban and Sandgathe (2006), clustering is performed on the combined fields while keeping information about the separate fields (or underlying clusters) intact. The cluster approaches are roughly similar to the neighborhood methods in the sense that a cluster defines a neighborhood, and the number of clusters determines the scale.

Micheas et al. (2007) propose a technique for verification of cell forecasts (referred to here as Procrustes). Features (or cells) are identified by finding clusters of nonzero pixels meeting a user-defined size criterion (e.g., four or more pixels). Matching is carried out based on centroid distances (location) or a shape criterion of the observed and forecasted features. The procedure

requires there to be the same number of cells in each field. Consequently, some features in one field may be matched to multiple features in the other, yielding a higher penalty for the over- or underforecasting of cells. Procrustes shape analysis and a user-defined penalty function are subsequently employed to glean information about forecast performance in terms of rotation, dilation, translation, and intensity-based errors over the entire forecast domain. Lack et al. (2009) modify this approach in several ways. In particular, they employ a multiscale technique for identifying objects, which is a large departure from other features-based approaches. Additionally, they employ cell-by-cell verification metrics along with summary statistics of forecast and observed objects. We shall refer to this modified approach as Procrustes2 hereafter.

Wernli et al. (2008, 2009) take a different approach to features-based verification. They define features within a relatively small area of interest such as a hydrological watershed, but no merging or matching of features in the forecast and observation fields is necessary. Their method, referred to as SAL (for structure, amplitude, and location), considers these three distinct components defined so that a perfect forecast would yield values of zero for all three. The approach has the advantage of providing useful, but parsimonious information about forecast performance when compared with the other features-based techniques.

Other features-based methods include cyclone-tracking techniques (e.g., Templeton and Keenan 1982; Marchok 2002), which gave inspiration to some of the features-based techniques described here. These methods are not discussed further here, however, because the emphasis of this special collection is on the verification of quantitative precipitation forecasts.

d. Field deformation verification

Some of the earliest proposed techniques for verifying gridded forecast–observation fields fall under the category of field deformation verification (e.g., Hoffman et al. 1995; Alexander et al. 1999). The field deformation, or morphing, approaches essentially involve spatial manipulation of the forecast field to make it appear as much like the observation field as possible (e.g., to minimize a score such as RMSE). They produce a field of distortion vectors (see Fig. 1, bottom-right panel), which is then evaluated either diagnostically or analytically. As mentioned previously, most of the techniques described in this section could be applied to individual features and subsequently, used within the framework of the features-based methods. Conversely, as was done in earlier papers (e.g., Alexander et al. 1999), it is possible to attempt to identify features in a field in order to inform the procedures as to optimal movements. Therefore, the primary distinction we make between the features-based and field deformation methods is that the field deformation techniques primarily work on an entire field (or subfield) at once instead of identifying and describing individual feature-to-feature comparisons within a field. This is illustrated in Fig. 1 by the vector field applied to the entire region in the lower-right panel versus the individual features in the panel to the left.

Keil and Craig (2007, 2009) employ a kind of nonparametric image warp that is very similar to optical flow, which combines the magnitude of a displacement vector determined by a pyramidal matching algorithm and the local squared difference of observed and morphed forecast intensity fields. Forecast performance information is summarized by a metric incorporating the amount of movement and amplitude change. This is referred to as the forecast quality measure (FQM) in Keil and Craig (2007) and a modified version, intended to replace FQM, is called the displacement and amplitude score (DAS) in Keil and Craig (2009). The modification in DAS allows false alarms to be handled correctly, where the FQM does not. Marzban et al. (2009) investigate the use of optical flow techniques for verification purposes and propose a method for summarizing the resulting vector fields. The method of Nehrkorn et al. (2003) distorts the forecast field using a technique referred to as feature calibration and alignment (FCA), which is perhaps more closely related to the scale-separation techniques, but does involve a distortion of the entire forecast field to better match the observed field. Dickinson and Brown (1996) and Alexander et al. (1999) applied image warping techniques using polynomial warp functions. Image warping requires the selection of control points (also called landmarks or tie points). Dickinson and Brown (1996) chose their control points automatically using covariate information, whereas Alexander et al. (1999) manually selected them to match important features and ensure physically meaningful warps. J. Lindström, E. Gilleland, and F. Lindgren (2009, unpublished manuscript, hereafter LGL) use a similar technique, but do not attempt to identify features for the control points; they also use a thin-plate spline instead of polynomial functions for the warping function. They found that informative warps are determined using a relatively small number of regularly spaced grid points for control points, at least for the test cases analyzed so far in the ICP.

Venugopal et al. (2005) introduce an image comparison metric called the forecast quality index (FQI), which combines both distance between two binary images (created by thresholding the observation and forecast fields) and intensity errors. The numerator of

TABLE 2. Summary table of some of the verification questions from section 3. Columns should be interpreted as, ''Can the method(s) account for, or provide information about . . . .''

Abbreviation	Scales?	Location errors?	Intensity errors?	Structure errors?	Hits, misses, false alarms, and correct negatives?
BCETS	No*	Indirectly	Yes	No	Yes
CA	Yes (see text)	Yes (see text)	No*	No	Yes
Procrustes	No*	Yes	Yes	Yes	Yes (see text)
Procrustes2	Yes	Yes	Yes	Yes	Yes
Composite	No*	Yes (see text)	Average intensities	Yes	Yes
CRA	No*	Yes	Yes	Yes	Yes
DIST	Yes (see text)	Indirectly	Yes	No	Yes
FQI	No*	Yes	Yes	No	No
FQM–DAS	No*	Yes	Yes	Yes	Yes (see text)
FSS	Yes (see text)	Indirectly	Yes	No	Indirectly
IS	Yes (see text)	Indirectly	Yes	No	Indirectly
IW	Yes (see text)	Yes	Yes	No*	Yes (see text)
MSV	Yes (see text)	Indirectly	Yes	No	No
MODE	Yes (see text)	Yes	Yes	Yes	Yes (see text)
Neighborhood	Yes (see text)	Indirectly	Yes	No	Yes
SAL	No*	Yes	Yes	No	No
Traditional	No*	No	Yes	No	Yes
VGM	Yes	No (see text)	No (see text)	Yes*	No

* A method that does not directly provide information about the specific topic, but either is sensitive to the type of error, can be easily modified, or can be applied to different fields (e.g., different thresholds, or resolution fields) to allow for the question to be addressed.

the index is a measure of distance between two binary images based on the normalized partial Hausdorff distance (Huttenlocher et al. 1999) and the denominator is a measure of the intensity error based on the means and standard deviations of the nonzero pixels. In producing the binary images for the partial Hausdorff distance, one can eliminate pixels with values below a particular intensity and isolate the smaller high-intensity cores. To avoid dependence on the percentage of nonzero pixels, the partial Hausdorff distance is normalized by the mean distance for several stochastic realizations, or ''surrogates,'' of the truth field. The surrogates have the same probability density function and spatial correlation structure as the truth field. Because this technique does not displace the forecast field to better match the observed field, it is not precisely a field deformation method. However, it measures a spatial displacement across the entire field, and this is the reason we classify it in this category.

3. Verification questions

The primary goal of the ICP is to supply advice concerning the types of information provided by each spatial verification method on the forecast quality. Some questions, addressing some specific issues of interest, are presented in this section, and qualitative answers regarding how each technique addresses such issues are discussed. Some of the answers are also summarized in Table 2.

a. How do the methods inform about performance at different scales?

There are at least two potential interpretations of this question. First, the verification end user might want to find the spatial scale over which a forecast has a desired level of skill. The neighborhood methods fall under this description of scale. The second interpretation of this question refers to the scales associated with a single-band spatial filter, where features at each of these scales are isolated and analyzed. The scale-separation methods are the only methods detailed here that can account for this second type of scale performance.

Other ideas of ''scale'' are implicit in various methods. For example, Marzban and Sandgathe (2006, 2008) refer to each iteration (number of clusters) of the hierarchical cluster analysis method as a scale. As noted earlier, this is similar to the idea of scale, or resolution, employed by the neighborhood methods, except that the idea of neighbors is defined quite differently; the ''filter,'' for example, is less of a smoothing filter, and is more similar to the idea of the features-based methods at this stage, which is why it has been categorized as such in this paper. Indeed, most of the features-based methods can address scale in this sense by identifying features using varying thresholds or filters, recognizing that higher-intensity features generally have smaller scales.

Generally, the field deformation approaches do not address ''scale'' issues directly, but certainly could be applied to the fields at different scales. The FQM–DAS

method of Keil and Craig (2007, 2009) morphs the field through a series of changes at different resolutions, but the ultimate score and vector field do not directly inform about scale performance. For methods such as those in LGL, the number of control points chosen to define the warp determines the complexity of the warp so that fewer points will only give larger-scale information, and more points yield increasingly finer-scale feedback. Further, all field deformation methods give information about "scaling" errors in terms of localized divergence and convergence of the vector field, which can be interpreted as under- and overforecasting, respectively.[2]

Features-based methods are not generally designed to inform about scale errors but can be used in this manner. For example, they can identify features of different sizes. MODE identifies features through a spatial smoothing technique combined with a threshold. Quilt plots showing various metrics of forecast performance as a function of the threshold and convolution radius can show how much the field needs to be smoothed by convolution to provide skill for any given threshold (i.e., it answers the question concerning the resolutions at which the forecast has skill at identifying particular feature attributes). Davis et al. (2009) use a quilt plot of the median of maximum interest to examine performance as a function of scale and to find the combinations of the convolution radius and the threshold at which a forecast has skill in terms of this summary measure.

In summary, any proposed method (including the traditional scores) could potentially provide information at different scales by simply applying the method to the field(s) at different resolutions. The neighborhood methods explicitly evaluate the forecast's performance at different scales. Some of the scale decomposition methods investigate how well a forecast is able to reproduce the observed field's scale structure (namely, Zepeda-Arce et al. 2000; Harris et al. 2001), whereas the IS approach assesses the forecast performance at the individual wavelengths associated with features of different scales.

b. *How do the methods provide information on location errors?*

Location errors are quite common as forecasts are made on increasingly finer grid resolutions. Traditional scores do not provide any direct information on these types of errors. In fact, as mentioned previously, traditional methods doubly penalize location errors as both misses and false alarms. Therefore, a major impetus for developing new verification methods is to obtain diagnostic information about location errors. For example, is the forecast basically correct, but missing the spatial target by x kilometers? Are there systematic errors in the forecast locations of storms?

The bias-corrected ETS (BCETS) answers these questions indirectly by accounting for the effects of bias on the GSS (i.e., ETS) so that the only remaining influence is the placement of the forecast. The method does not provide specific information on displacement amplitude or direction. The same can be said for scale-separation methods. Forecasts with more displacement error have error at larger scales, but quantifying the displacement is not straightforward and direction is not considered. AGBE illustrate this with two simple geometric cases.

The features-based and field deformation methods all directly provide information about location errors, assuming that the forecast sufficiently resembles the observations so that corresponding features can be associated or matched. A possible exception is the CA method, which allows for location errors but does not return information about them apart from a lower score. The composite method is a good way to summarize the systematic displacement error for a large number of precipitation events. The neighborhood approaches, as well as IS, do not give location error information, except perhaps indirectly (most scores improve as location errors decrease), and they do not provide information on the direction or magnitude of the location errors. Because wavelet decomposition provides both scale and location information, it should be possible to utilize it to inform about location errors for different scale features, but to the best of our knowledge, this has not yet been proposed in the literature. The Procrustes2 technique informs directly about location errors for different scale features. When applying VGM to the entire field, the structure and location of an object will affect the shape of the variogram plot so that VGM can be sensitive to location errors.

c. *Do the methods provide information on intensity errors and distributions?*

Results for this question are similar to those for the previous question, except that neighborhood, IS, and traditional methods all provide information on intensity errors and distributions. The CA method can be modified slightly to give this information by thresholding fields to obtain CA scores as a function of intensity. The composite method gives information on the average difference between the intensities for the forecast and observation fields. Although the FQM–DAS and FQI metrics combine information on location and intensity errors, one can examine the location and intensity

[2] Generally, it is appropriate to morph only the forecast field so that comparisons of multiple forecasts are made to the same observed field. Therefore, diverging vectors indicate underforecasting and convergent vectors overforecasting.

components of the total scores separately. The volume component of the error from the CRA approach provides information on intensity errors. A multiplicative intensity error across the entire field will result in a variogram shifted in the ordinate axis, but an additive one will not show up at all. Therefore, intensity errors are not readily discriminated via the VGM approach.

d. Do the methods provide information on structure errors?

Similar to the question on scales, the answer to this question is sensitive to exactly what is meant by *structure*. One interpretation concerns individual structures within a field (e.g., a large-scale convective region), and the other concerns the overall spatial structure of the field. The answer to the second question is generally yes for some of the scale decomposition approaches [e.g., Harris et al. (2001) and Zepeda-Arce et al. (2000) assess scale-invariant parameters related to the spatiotemporal organization of precipitation fields], but no for most other approaches.

Techniques that define objects in forecast and observation fields can assign geometric attributes to these objects. These attributes (e.g., area, length, and orientation) summarize the structures of objects. Physical meaning can be assigned if the structure has a repeatable relationship to the behavior or character of a feature. For instance, highly elongated rainfall regions would be expected to occur more often with strong frontal zones. Circular rain areas might be individual thunderstorms if small, and complexes of thunderstorms if large.

The only methods that provide this type of information are the MODE, CRA, Procrustes, and Procrustes2 methods. Not surprisingly, these are all features-based methods. Field deformation methods supply vector fields showing the optimal movement of the forecast over the entire region. Inspection of such fields can be useful in determining how well a forecast was able to capture the observed storm structures, at least on a case-by-case basis. The FQM–DAS summarizes information relevant to this issue and can be easily aggregated over multiple cases. VGM will give information about the structure of the field as a whole, but does not generally inform about individual structures.

e. Do the approaches have the ability to provide information about hits, misses, false alarms, and correct negatives?

Certain methods such as traditional and neighborhood methods provide an obvious means of determining hits, misses, and false alarms, while some methods are sensitive to these values without measuring them directly.

For example, although the FSS method does not calculate them directly, it is sensitive to false alarms and misses, as well as hits. Generally, the features-based methods are able to provide this information, except for the SAL procedure. The Procrustes method described in Micheas et al. (2007) gives only limited information here; it does identify the number of observed cells versus the number of forecasted cells, which may be useful in identifying the over- and underforecasting of precipitation areas. The currently implemented version of this method (Procrustes2), however, incorporates hits, misses, and false alarms more directly. Otherwise, the strategy with features-based approaches is to interpret matched features as hits, unmatched observed features as misses, and unmatched forecast features as false alarms. There is no obvious means for determining correct negatives from this paradigm, but reasonable definitions can be applied. Because matched features may still be very different in some ways from each other (e.g., large differences in spatial extent), it is often useful to weight the various attributes of the match to quantify its goodness. The DIST method provides this information directly, while the IS method provides it indirectly (in particular, the IS skill score is equivalent to the Heidke skill score). The vector fields computed by the field deformation methods provide information about over- and underforecasting by indicating divergence and convergence (i.e., stretching and squeezing).

f. Do the methods do anything that is counterintuitive?

More research is required to answer this question, but there is agreement among ICP participants that the composite, DIST, neighborhood, scale-separation, and traditional scores do not do anything unintuitive. The features-based methods are all susceptible to counterintuitive merging and matching (e.g., schemes that rely more on intensity for matching may match objects that are far apart). This is hardly surprising as two different human observers are likely to merge and match objects differently from each other. A strategy to reduce counterintuitive matching is to impose restrictions or penalty functions on the differences in location, intensity, size, etc., between forecast and observed features. Field deformation methods may ''explode'' a forecast feature that is much smaller than the observed feature or ''implode'' a forecast feature that is not present in the observations.

g. Do the methods have selectable parameters and how sensitive are the results to parameter choice?

Although this question is seemingly innocuous, the interpretation of what constitutes a parameter varies

depending on how one looks at the problem, or what information is desired from the verification. All of the features-based methods except for SAL clearly have selectable parameters such as a threshold for feature definition, minimum size criterion, and search radius; some, such as MODE, have a great many such parameters. The CA method currently only has one selectable parameter: the ratio of observation to forecast overlap in defining a hit or a miss. There are numerous ways to carry out the CA method, but once a procedure is decided upon, there are no user-selectable parameters. Tunable parameters for the FQI method include the (i) choice of threshold, (ii) percentile distance in the partial Hausdorff distance calculation, and (iii) number of surrogate fields to evaluate.

In general, the neighborhood methods have relatively few selectable parameters, which can include (depending on the specific method) the number and size of the neighborhoods and the choice of score used to evaluate the forecast. Methods such as FSS and traditional scores generally do not have selectable parameters unless one considers thresholds as parameters. For the DIST method, one must choose the number of observations a forecast grid box must encompass in order to be considered in the computations. The FQM–DAS has only the search range as a parameter, but also uses the maximum value of the image and the image size to normalize the score.

Scale-separation techniques also generally have fewer selectable parameters, although one might need to choose which scales are meaningful. The VGM is merely a plot of the empirical variogram with error bars obtained via resampling methods, but it is necessary to bin distances. Deciding on a bin width, or method of binning, could be thought of as a selectable parameter.

Image warping allows some user-defined choices involving the penalty function (the penalty function protects against nonphysical warps), and if a distribution of the errors is assumed, then some additional parameters will be introduced. However, such parameters are chosen in part to optimize the procedure based on the properties of a particular variable field and partly depending on the particular user's needs. For example, for a precipitation field, it might be reasonable to assume that the forecast is not going to be spatially displaced more than, say, 100 km in any direction. Therefore, it would make sense to more heavily penalize large movements of the control points beyond this range. The warp could still make such a move, but it becomes increasingly less likely with higher penalties on such transformations. Regardless of how the penalties are chosen, however, the method gives consistent information about the overall forecast quality.

h. Can the results be easily aggregated across multiple cases?

The answer to this question is generally positive for all methods, except for the original implementation of the Procrustes approach (this is readily achieved via Procrustes2). For MODE, and other similar features-based methods, it is possible to aggregate over cases in a number of ways. For example, because it is usually possible to calculate the x–y displacement between matched objects, circle histograms can be used to give information (aggregated across several cases, or for a single case) about potential directional biases of forecast objects along with other information (centroid distances, intensity differences, etc.). Otherwise, whatever summary measures are calculated for a single snapshot for a features-based method can be aggregated over several cases. In fact, the original motivation for the CRA method was to investigate systematic errors over many cases.

Several aggregation options are available for field deformation approaches. For example, it is possible to summarize the vector fields for subregions of each case and subsequently aggregate these summaries over multiple cases. Care should be taken in doing so as it is certainly possible for some effects to be canceled out (e.g., a forecast that is generally displaced to the east in the daytime and to the west at night). Otherwise, low-dimensional metrics (e.g., FQM–DAS, FQI) can be gleaned for each field, and these can be easily aggregated over many cases.

i. Can the methods be used to identify timing errors?

Any of the methods could be applied to different times to find the time that has the best match and, therefore, give information about timing errors. The only methods that appear to not *potentially* have the ability to directly provide the information either from a single field or by incorporating temporal information (e.g., via multiple fields simultaneously) are the IS, the composite, and the traditional measures. For many of the methods, this information is obtained indirectly by spatial offset, false alarms, and missed forecasts. For the FSS and other neighborhood methods, timing errors can be inferred if different temporal windows are used. A new three-dimensional version of MODE is being developed specifically for identifying timing errors (R. G. Bullock 2008, personal communication). It should be possible to account for timing errors with image warping at the expense of greater complexity in the optimization routine.

However, practical considerations, including computational requirements, may limit the ability to extend some of the new methods to the time dimension. In theory, it is simple to add another dimension to the

software, but significant overhead is often associated with such a modification. Even if the software is already encoded to use the time dimension, the sheer volume of data can overwhelm the system unless the spatial domain is severely reduced from what could be handled when only a single time is considered. The time dimension also adds another pair of boundaries to the dataset (start time and end time), and almost all methods must make special exceptions for data at the boundaries or include a large enough buffer to avoid the issue. Given the importance of correctly predicting the onset and duration of high-impact events, efforts to adapt verification methods to quantify timing errors should be strongly encouraged.

j. Can confidence intervals or hypothesis tests be readily computed for the method?

For most methods proposed, confidence intervals (CIs) or hypothesis tests have not been considered. Bootstrapping can generally be employed for most situations, but sometimes the cost in efficiency may be high, at least for operational use. The term ''readily'' implies that if CI's or hypothesis tests have not been applied already, then the answer to this question can be yes only if it is clear that a method can be easily implemented by an average user.

The Procrustes method employs a Bayesian paradigm for finding confidence intervals, and the CA and composite methods both easily allow for confidence intervals because the statistics evaluated by these methods can be assumed to be normally distributed. The CRA method tests for significance of correlation when matching objects. Bootstrap methods are currently being implemented for efficient computation of confidence intervals for the IS scores (i.e., without requiring the wavelet decompositions to be redone). Because the image warping method is based on a statistical model, CI's can be efficiently computed provided a reasonable distribution for the movement errors is determined. For large amounts of data, it may be possible to implement a normal approximation to find confidence limits for some of the neighborhood approaches. When categorical scores are used in neighborhood approaches, then bootstrap confidence intervals for large datasets can be obtained by resampling the contingency table entries for each neighborhood size and threshold.

4. Discussion

This paper reviews some of the new approaches that have recently been developed to provide more diagnostic and informative evaluations of high-resolution forecasts of spatial fields, such as precipitation. Each of the approaches is designed to evaluate particular attributes of forecast performance. Each approach performs better than others for certain situations, and for obtaining certain verification information. Broadly speaking, the various methods can be categorized into neighborhood, scale-separation, features-based, and field deformation approaches.

When adopting new forecast evaluation approaches, it is important to consider their characteristics and the forecast performance attributes they assess. The goal of the ICP is to provide more specific information about the methods' capabilities and assist users by providing valuable information to help them decide which one(s) is best suited to their needs. To meet this goal, it is important to include cases that represent a wide range of forecast situations. For example, some methods might be ideally suited for forecasts of widespread precipitation, while others perform better for convective systems. Grid scales and domains may also factor into the ability of a verification method to accurately measure forecast performance. Therefore, it will be important to include cases on different scales and for different fields of interest (e.g., wind). For the initial evaluation, however, the focus is on a set of summer precipitation forecasts for the Midwest. Idealized cases with known forecast errors were also tested with each method and the results from this exercise are summarized in the companion paper by AGBE. These cases provide a baseline against which the performance of the methods can be compared.

When evaluating the capabilities of new verification approaches, the issue of hedging (forecasting other than one's true belief in order to improve the verification score) is generally also of concern. However, given that most of the new methods considered here are advanced diagnostic techniques for investigating high-resolution spatial forecasts, it is unlikely (but not impossible) that one would tune a model to obtain the best performance by hedging. For example, for a traditional verification metric such as a threat score one could hedge the results and increase the TS by simply increasing the forecast bias. Nevertheless, it will be worth investigating how each method could be hedged, if at all, to artificially improve the verification results.

In summary, methods classified here as neighborhood methods apply different scores (e.g., traditional scores) to filtered versions of one or both of the forecast and observed fields. The filter is a smoothing filter applied to increasingly larger neighborhoods of each grid point. The type of filter defines the specific method, and different verification questions are addressed depending on the chosen score (e.g., a different conclusion may be reached using HK versus GSS). The scale-separation

approaches apply a spatial bandpass filter and examine forecast performance (again, often based on traditional scores) at various wavenumbers. In contrast to the filter techniques, which attempt to find the scales at which a forecast is skillful, the features-based and field deformation approaches attempt to determine how much a forecast needs to be manipulated spatially in order to obtain a skillful forecast. The features-based approaches are concerned more with identifying physically meaningful structures in each field and comparing these structures across fields. Field deformation approaches, on the other hand, address the entire field at once without concern for individual structures in the fields. Some methods, such as CA and FQI, are more difficult to classify into these categories, but they can be loosely included in the features-based and field deformation categories, respectively.

The results presented in this paper are a preliminary qualitative summary of the methods contained in the other papers in this special collection, which include much more detail regarding the capabilities of the various approaches. Furthermore, new methods and changes to the methods described herein may have been invoked in the other papers for this collection; included here are the methods as described in published and in-progress papers (known to these authors) at the time of writing. We believe that the results of the ICP presented in these papers represent the first study of this type, in which such a large variety of new approaches have been compared on a common ground, with a focus on the information that can be provided to the users of the methods. Ideally, this effort will be the first in a series of such collaborative investigations.

Acknowledgments. The authors and participants thank Mike Baldwin for supplying the model and observation data for the ICP, as obtained from the Storm Prediction Center/National Severe Storms Laboratory (SPC/NSSL) Spring 2005 Program. We also thank all of the participants of the intercomparison project who took part in the planning meetings. We especially thank Efi Foufoula-Georgiou, Steven Lack, Chiara Marsigli, Caren Marzban, Fedor Mesinger, Marion Mittermaier, Jason Nachamkin, and Scott Sandgathe for valuable input and suggestions for this paper.

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. *Wea. Forecasting,* in press.

Alexander, G. D., J. A. Weinman, V. M. Karyampudi, W. S. Olson, and A. C. L. 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.

Atger, F., 2001: Verification of intense precipitation forecasts from single models and ensemble prediction systems. *Nonlinear Processes Geophys.,* **8,** 401–417.

Baldwin, M. E., and S. Lakshmivarahan, 2003: Development of an events-oriented verification system using data mining and image processing algorithms. Preprints, *Third Conf. on Artificial Intelligence,* Long Beach, CA, Amer. Meteor. Soc., 4.6. [Available online at http://ams.confex.com/ams/pdfpapers/57821.pdf.]

——, and J. S. Kain, 2006: Sensitivity of several performance measures to displacement error, bias, and event frequency. *Wea. Forecasting,* **21,** 636–648.

Briggs, W. M., and R. A. Levine, 1997: Wavelets and field forecast verification. *Mon. Wea. Rev.,* **125,** 1329–1341.

Brooks, H. E., M. Kay, and J. A. Hart, 1998: Objective limits on forecasting skill of rare events. Preprints, *19th Conf. on Severe Local Storms,* Minneapolis, MN, Amer. Meteor. Soc., 552–555.

Casati, B., 2009: New developments of the intensity-scale technique within the Spatial Verification Methods Inter-Comparison Project. *Wea. Forecasting,* in press.

——, G. Ross, and D. B. Stephenson, 2004: A new intensity-scale approach for the verification of spatial precipitation forecasts. *Meteor. Appl.,* **11,** 141–154.

——, and Coauthors, 2008: Forecast verification: Current status and future directions. *Meteor. Appl.,* **15,** 3–18.

Damrath, U., 2004: Verification against precipitation observations of a high density network—What did we learn? *Int. Verification Methods Workshop,* Montreal, QC, Canada, WMO. [Available online at http://www.bom.gov.au/bmrc/wefor/staff/eee/verif/Workshop2004/presentations/5.3_Damrath.pdf.]

Davis, C. A., B. G. Brown, and R. G. Bullock, 2006: Object-based verification of precipitation forecasts. Part I: Methodology and application to mesoscale rain areas. *Mon. Wea. Rev.,* **134,** 1772–1784.

——, ——, ——, and J. Halley Gotway, 2009: The Method for Object-based Diagnostic Evaluation (MODE) applied to WRF forecasts from the 2005 Spring Program. *Wea. Forecasting,* **24,** 1252–1267.

Dickinson, S., and R. Brown, 1996: A study of near-surface winds in marine cyclones using multiple satellite sensors. *J. Appl. Meteor.,* **35,** 769–781.

Ebert, E. E., 2008: Fuzzy verification of high resolution gridded forecasts: A review and proposed framework. *Meteor. Appl.,* **15,** 51–64.

——, 2009: Neighborhood verification: A strategy for rewarding close forecasts. *Wea. Forecasting,* in press.

——, and J. L. McBride, 2000: Verification of precipitation in weather systems: Determination of systematic errors. *J. Hydrol.,* **239,** 179–202.

——, and W. A. Gallus Jr., 2009: Toward better understanding of the contiguous rain area (CRA) method for spatial forecast verification. *Wea. Forecasting,* **24,** 1401–1415.

——, L. J. Wilson, B. G. Brown, P. Nurmi, H. E. Brooks, J. Bally, and M. Jaeneke, 2004: Verification of nowcasts from the WWRP Sydney 2000 Forecast Demonstration Project. *Wea. Forecasting,* **19,** 73–96.

Germann, U., and I. Zawadzki, 2004: Scale dependence of the predictability of precipitation from continental radar images. Part II: Probability forecasts. *J. Appl. Meteor.,* **43,** 74–89.

Gilleland, E., T. C. M. Lee, J. Halley Gotway, R. G. Bullock, and B. G. Brown, 2008: Computationally efficient spatial forecast verification using Baddeley's delta image metric. *Mon. Wea. Rev.,* **136,** 1747–1757.

Grams, J. S., W. A. Gallus Jr., S. E. Koch, L. S. Wharton, A. Loughe, and E. E. Ebert, 2006: The use of a modified Ebert–McBride technique to evaluate mesoscale model QPF as a function of convective system morphology during IHOP 2002. *Wea. Forecasting,* **21,** 288–306.

Harris, D., E. Foufoula-Georgiou, K. K. Droegemeier, and J. J. Levit, 2001: Multiscale statistical properties of a high-resolution precipitation forecast. *J. Hydrometeor.,* **2,** 406–418.

Hoffman, R. N., Z. Liu, J.-F. Louis, and C. Grassotti, 1995: Distortion representation of forecast errors. *Mon. Wea. Rev.,* **123,** 2758–2770.

Huttenlocher, D. P., R. H. Lilien, and C. F. Olson, 1999: Object recognition using subspace methods. *IEEE Trans. Pattern Anal. Mach. Intell.,* **21,** 951–956.

Jolliffe, I. T., and D. B. Stephenson, 2003: *Forecast Verification. A Practitioner's Guide in Atmospheric Science*. Wiley and Sons, 240 pp.

Keil, C., and G. C. Craig, 2007: A displacement-based error measure applied in a regional ensemble forecasting system. *Mon. Wea. Rev.,* **135,** 3248–3259.

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

Lack, S., G. L. Limpert, and N. I. Fox, 2009: An object-oriented multiscale verification scheme. *Wea. Forecasting,* in press.

Marchok, T., 2002: How the NCEP tropical cyclone tracker works. Preprints, *25th Conf. on Hurricanes and Tropical Meteorology,* San Diego, CA, Amer. Meteor. Soc., P1.13. [Available online at http://ams.confex.com/ams/pdfpapers/37628.pdf.]

Marsigli, C., A. Montani, and T. Paccagnella, 2006: Verification of the COSMOLEPS new suite in terms of precipitation distribution. *COSMO Newsletter,* No. 6, Consortium for Small-Scale Modeling, 134–141. [Available online at http://www.cosmo-model.org/content/model/documentation/newsLetters/newsLetter06/default.htm.]

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

——, and ——, 2008: Cluster analysis for object-oriented verification of fields: A variation. *Mon. Wea. Rev.,* **136,** 1013–1025.

——, and ——, 2009: Verification with variograms. *Wea. Forecasting,* **24,** 1102–1120.

——, ——, H. Lyons, and N. Lederer, 2009: Three spatial verification techniques: Cluster analysis, variogram, and optical flow. *Wea. Forecasting,* in press.

Mesinger, F., 2008: Bias-adjusted precipitation threat scores. *Adv. Geosci.,* **16,** 137–142.

Micheas, A. C., N. I. Fox, S. A. Lack, and C. K. Wikle, 2007: Cell identification and verification of QPF ensembles using shape analysis techniques. *J. Hydrol.,* **343,** 105–116.

Mittermaier, M. P., 2006: Using an intensity-scale technique to assess the added benefit of high-resolution model precipitation forecasts. *Atmos. Sci. Lett.,* **7,** 35–42.

——, and N. Roberts, 2009: Intercomparison of spatial forecast verification methods: Identifying skillful spatial scales using the fractions skill score. *Wea. Forecasting,* in press.

Nachamkin, J. E., 2004: Mesoscale verification using meteorological composites. *Mon. Wea. Rev.,* **132,** 941–955.

——, 2009: Application of the composite method to the Spatial Forecast Verification Methods Intercomparison Dataset. *Wea. Forecasting,* **24,** 1390–1400.

——, S. Chen, and J. Schmidt, 2005: Evaluation of heavy precipitation forecasts using composite-based methods: A distributions-oriented approach. *Mon. Wea. Rev.,* **133,** 2163–2177.

Nehrkorn, T., R. Hoffman, C. Grassotti, and J.-F. Louis, 2003: Feature calibration and alignment to represent model forecast errors: Empirical regularization. *Quart. J. Roy. Meteor. Soc.,* **129,** 195–218.

Nychka, D. W., 1998: Spatial process estimates as smoothers. *Smoothing and Regression,* M. G. Schimek, Ed., Wiley, 393–423.

Ogden, T., 1996: *Essential Wavelets for Statistical Applications and Data Analysis*. Birkhäuser, 224 pp.

Rezacova, D., Z. Sokol, and P. Pesice, 2007: A radar-based verification of precipitation forecast for local convective storms. *Atmos. Res.,* **83,** 221–224.

Roberts, N. M., 2005: An investigation of the ability of a storm-scale configuration of the Met Office NWP model to predict flood-producing rainfall. Forecasting Research Tech. Rep. 455, Met Office, 80 pp.

——, and H. W. Lean, 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Mon. Wea. Rev.,* **136,** 78–97.

Templeton, J. I., and T. D. Keenan, 1982: Tropical cyclone strike probability forecasting in the Australian region. Bureau of Meteorology Tech. Rep. 49, Melbourne, VIC, Australia, 18 pp. [Available from Bureau of Meteorology, GPO Box 1289K, Melbourne, VIC 3001, Australia.]

Theis, S. E., A. Hense, and U. Damrath, 2005: Probabilistic precipitation forecasts from a deterministic model: A pragmatic approach. *Meteor. Appl.,* **12,** 257–268.

Tustison, B., E. Foufoula-Georgiou, and D. Harris, 2003: Scale-recursive estimation for multisensory quantitative forecast verification: A preliminary assessment. *J. Geophys. Res.,* **108,** 8377–8390.

Venugopal, V., S. Basu, and E. Foufoula-Georgiou, 2005: A new metric for comparing precipitation patterns with an application to ensemble forecasts. *J. Geophys. Res.,* **110,** D08111, doi:10.1029/2004JD005395.

Wernli, H., M. Paulat, M. Hagen, and C. Frei, 2008: SAL—A novel quality measure for the verification of quantitative precipitation forecasts. *Mon. Wea. Rev.,* **136,** 4470–4487.

——, C. Hofmann, and M. Zimmer, 2009: Spatial Forecast Verification Methods Intercomparison project: Application of the SAL technique. *Wea. Forecasting,* in press.

Weygandt, S. S., A. F. Loughe, S. G. Benjamin, and J. L. Mahoney, 2004: Scale sensitivities in model precipitation skill scores during IHOP. Preprints, *22nd Conf. on Severe Local Storms,* Hyannis, MA, Amer. Meteor. Soc., 16A.8. [Available online at http://ams.confex.com/ams/pdfpapers/81986.pdf.]

Wilks, D. S., 2005: *Statistical Methods in the Atmospheric Sciences. An Introduction*. 2nd ed. Academic Press, 648 pp.

Yates, E. S., S. Anquetin, V. Ducrocq, J.-D. Creutin, D. Richard, and K. Chancibault, 2006: Point and areal validation of forecast precipitation fields. *Meteor. Appl.,* **13,** 1–20.

Zepeda-Arce, J., E. Foufoula-Georgiou, and K. K. Droegemeier, 2000: Space–time rainfall organization and its role in validating quantitative precipitation forecasts. *J. Geophys. Res.,* **105,** 10 129–10 146.