16B.2 THE SPATIAL VERIFICATION METHOD INTERCOMPARISON PROJECT

David Ahijevych*, Eric Gilleland, and Barbara G. Brown
National Center for Atmospheric Research, Boulder, Colorado

Elizabeth E. Ebert
Centre for Australian Weather and Climate Research, Melbourne, Australia

1. INTRODUCTION

The numerical simulation of weather events has progressed to the point that mesoscale phenomena such as squall-lines and hurricanes are routinely forecasted. The simulated reflectivity field and precipitation distribution have realistic features and spatial structure that can provide valuable guidance to forecasters on the mode of convective evolution (Weisman et al., 2008). However, the traditional verification scores often do not reflect this improvement. Small errors in the position or timing of convective features result in both false alarms and missed events that dominate the 2x2 contingency table that serves as the foundation of traditional categorical verification scores (Wilks, 2006). This problem is only exacerbated by tighter grid spacing. Several traditional scores such as critical success index (CSI or threat score) and Gilbert skill score (GSS or equitable threat score) have been used for decades, but their utility is limited when it comes to diagnosing displacement error or an incorrect mode of convective organization.

The spatial verification method Intercomparison Project, or ICP, was organized to explore better ways of evaluating high-resolution numerical model forecasts. The ICP stemmed from a verification workshop originally held in Boulder, CO. The new methods often do not require one-to-one matches between forecast and observed events at the grid scale in order to give credit to a good forecast. A literature review of the new methods is given by Gilleland et al. (2009). In Gilleland et al. (2009), four main categories of methods are identified and described, a convention that is continued here: neighborhood, scale separation, feature-based and field deformation. These categories are illustrated in Fig. 1.

As part of the ICP, a set of idealized gridded precipitation forecasts with prescribed displacement, intensity, and frequency bias errors was created to demonstrate the capabilities of different forecast verification methods. ICP participants were asked to apply their methods to the common set of fake and real forecasts. This paper describes the test cases and summarizes results from the method inter-comparison.

All methods could detect bias error, and the features-based and field deformation methods were also able to diagnose displacement error. The best approach for capturing errors in aspect ratio was the field deformation approach. For real cases, some new spatial verification methods agreed better with the subjective assessment of the forecasts than did the traditional verification statistics, confirming their ability to account for realistic spatial structure and close forecasts.

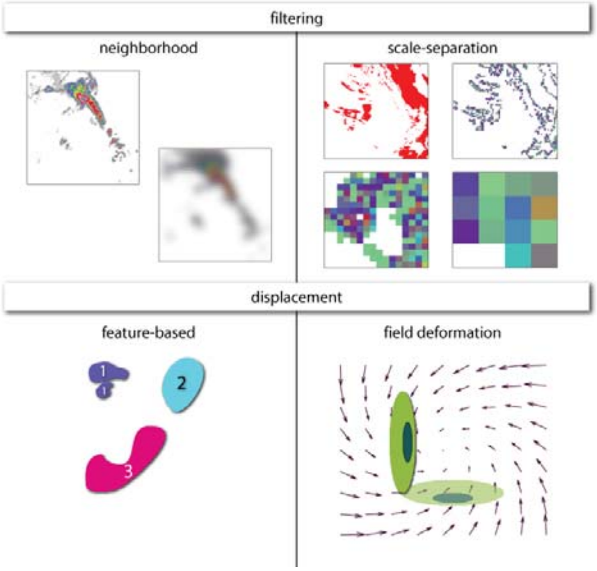

Fig. 1. This is a pictorial representation of the four new verification categories explored in the ICP. The panels are meant to show off unique aspects of each technique, not to illustrate any particular case.

The reader may use this document alongside Gilleland et al. (2009) to determine which methods are appropriate for their needs, and then refer to more detailed papers on the individual methods, many of which form part of a *Weather and Forecasting* special collection on the Spatial Verification Methods Intercomparison Project (Casati, 2009; Davis et al., 2009; Ebert 2009; Ebert and Gallus, 2009; Gilleland et al., 2009b; Keil and Craig, 2009; Lack et al., 2009; Lindstrom et al., 2009; Marzban and Sandgathe, 2009; Marzban et al., 2009; Mittermaier and Roberts, 2009; Nachamkin, 2009; Wernli et al., 2009).

2. TEST CASES

a) *Geometric cases*

To explore the variety of new methods, we present simple geometric forecast and observation patterns that embody general forecast errors. The observation is named geom000 and the forecasts are geom001-

* *Corresponding author address:* David Ahijevych, NCAR, Boulder, CO; e-mail: ahijeyvc@ucar.edu

geom005 (Figure 2). As with our more realistic cases, we assume these patterns represent 1-h accumulated rainfall at 24-hour lead time. The precipitation zones are simple ellipses as defined in Ahijevych et al. (2009). They could also be thought of as idealized storm cells or mesoscale convective systems with a high intensity core embedded within a region of low intensity.

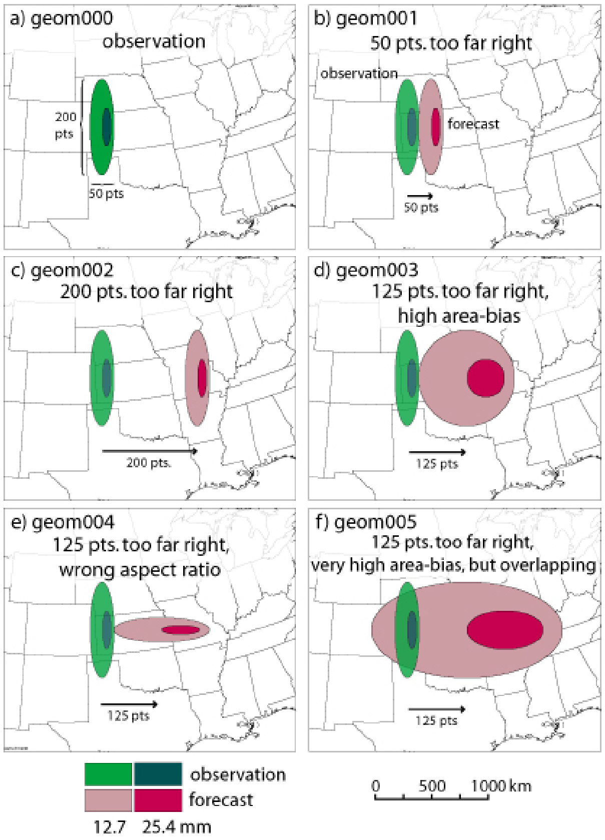

Figure 2. Five simple geometric cases derived to illustrate specific forecast errors. In all panels, the forecasted feature (red) is positioned to the right of the observed feature (green). Note, the forecast and observation features only overlap in geom005.

There are no structural differences among the geometric features except area and aspect ratio. All features are centered on the same y-coordinate; the ratio of the area enclosed by the high intensity core to the area enclosed by the low-intensity ellipse is always the same; the high-intensity core is always in the same right-of-center position relative to the low-intensity ellipse.

The geometric forecasts illustrate three types of error: 1) displacement, 2) aspect ratio, and 3) frequency bias. Knowledge of these errors could be useful for model development and improvement and could be informative for users of the forecasts. The first two types of error are especially difficult to discern with traditional verification methods. When the forecast and observed areas of precipitation do not overlap, as in geom001-geom004, then traditional scores such as CSI, GSS, and Hanssen-Kuipers (H-K) indicate no skill (Table 1). The geom003 case has even worse probability of false detection, H-K, HSS, and GSS than geom001, geom002, and geom004 because the larger forecast object results in more false alarms and fewer correct forecasted null events.

Table 1. Traditional verification scores applied to geometric cases. These statistics were calculated with the grid_stat tool, part of the Model Evaluation Tools (MET) verification package (NCAR, 2009).

Traditional score	geom1/ 2/4	geom00 3	geom005
accuracy	0.948	0.870	0.811
multiplicative freq. bias	1.000	4.018	8.034
mult. intensity bias	1.000	4.022	8.044
RMSE	3.514mm	5.572mm	6.875mm
bias-corrected RMSE	3.514mm	5.451mm	6.327mm
correlation coefficient	-0.024	-0.050	0.203
prob. of detection	0.000	0.000	0.876
prob. of false detection	0.027	0.107	0.191
false alarm ratio	1.000	1.000	0.891
Hanssen-Kuipers (H-K)	-0.027	-0.107	0.685
Threat score (CSI)	0.000	0.000	0.107
Gilbert skill score	-0.013	-0.021	0.084
Heidke skill score	-0.027	-0.043	0.155

The first two geometric cases, geom001 and geom002, represent pure displacement error. The geom001 forecast feature is touching the observation, and the geom002 case is displaced further to the right. The forecast and observed features do not overlap in either case. The spatial scale suggested by the map background suggests that both of these forecasts are very poor because 24-h forecasts of synoptic-sized precipitation zones are typically more skillful than this. However, the forecast value ultimately depends on the forecast application. Even forecast geom002 might be valuable to some users as the shape, orientation, and distribution of intensity are perfect.

Despite the obvious differences in the quality of these two forecasts, traditional verification metrics (Table 1) do not indicate any differences in performance. In contrast, some of the new spatial verification methods are able to distinguish differences in performance for these two cases and quantify the displacement error.

The geom004 case illustrates an error in aspect ratio. Although this case is similar to a simple rotation and displacement of the observed feature, the error is not quite that straightforward. In particular, width of the forecast area is four times larger than the width of the observation area, and the forecast height is one-fourth the observed height.

The final type of simple forecast error isolated in the geometric figures is frequency bias. The geom003 and geom005 cases are both stretched in the *x*-dimension, making the forecast too large. The displacement error is the same as in geom004 (125 points to the right). Traditional bias measures do pick up the difference in forecast and observed areas (multiplicative bias in Table

1), and the RMSE is largest for geom005, but the behavior of some other traditional scores is troubling when comparing these three cases (geom003-geom005). In particular, geom005 has a much higher forecast frequency bias than geom003 or geom004, but because it overlaps the observation, its false alarm ratio, H-K, GSS, and CSI scores are superior to the scores for all of the other geometric cases (Table 1). To be fair, a hydrologist might actually prefer geom005, even if it is considered to be extremely poor by modelers and other users. Nevertheless, a larger CSI value does not indicate that the forecast is better overall, which is why these types of scores can be misleading when used in isolation. For example, it is best practice to show bias alongside GSS and CSI whenever possible because GSS and CSI can easily be improved at the expense of increasing the bias (Baldwin and Kain, 2006).

b) Perturbed cases

In addition to the geometric shapes, some ICP participants evaluated a set of perturbed precipitation forecasts from a high resolution (2-km) numerical weather prediction model. Starting with an artificial observed precipitation field based on the 24-h forecast of 1-hour accumulated precipitation provided by the Center for Analysis and Prediction of Storms (CAPS) valid on 1 June 2005 at 0000 UTC, perturbed forecasts were made by shifting the entire field to the right and downward by different amounts. Additional details about the model configuration are in Kain et al. (2008). In the final two perturbed cases, the displacement error was held constant, but the intensity was multiplied by 1.5 in pert006 and had 1.27 mm subtracted from it in perl007. This paper does not describe the verification results for the seven perturbed cases, but interested readers can consult individual papers that evaluate them (Casati, 2009; Ebert 2009; Ebert and Gallus, 2009; Gilleland et al., 2009b; Keil and Craig, 2009; Lack et al., 2009; Lindstrom et al., 2009; Marzban and Sandgathe, 2009; Marzban et al., 2009; Mittermaier and Roberts, 2009; Nachamkin, 2009; Wernli et al., 2009).

c) Real cases

For the realistic precipitation examples, we use nine cases that were originally presented to a panel of twenty-six scientists attending a workshop on spatial verification methods to obtain their subjective assessments of forecast performance. The panel's subjective scores are presented here as an alternative viewpoint, not a definitive assessment of forecast performance. The panel looked at three different model forecasts of one-hour accumulated precipitation between forecast hours 23 and 24 and were asked to compare them to the observed precipitation from the stage II analysis (Lin and Mitchell, 2005). They rated the models' performance on a scale from 1 to 5, ranging from poor to excellent. For fairness, the models were ordered randomly and unlabeled. Figure 3 shows the observed precipitation fields in the same order as presented to the panel. To conserve space we do not show the 27 accompanying forecasts (3 for each case), but they are available online at http://www.ral.ucar.edu/projects/icp/datacases.html.

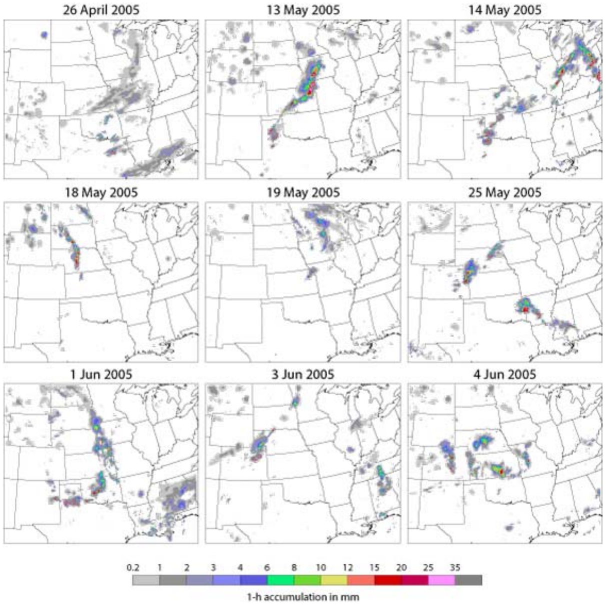

Figure 3. Nine cases were chosen for the subjective evaluation. These are the observed 1-h accumulated precipitation fields at 00 UTC for the dates indicated at the top of each panel. For space considerations, the corresponding 27 model forecasts (9 days x 3 models) are not shown.

The evaluators were not asked to consider the usefulness of the forecasts from the standpoint of any particular user (e.g., water-manager, farmer, SPC forecaster) or to focus on a particular region, but to subjectively evaluate the forecast as a whole. Afterwards, several of the panel members indicated that more guidance was needed in these areas, because the usefulness of a forecast depends greatly on the perceived needs of the user and the geographical area of concern; sometimes a model did well in one region and poorly in another. But in order to keep it simple, participants were asked to simply give an overall impression of the models' skill and were left to themselves to decide what mattered most.

Although we cycled through the nine cases twice to increase the stability of the overall responses and to assess the natural variability from one trial to the next, several aspects of our survey added uncertainty to the results. First, the panel members had varying professional backgrounds, including meteorologists, statisticians, and software engineers. Meteorologists were more likely to consider realistic depictions of mesoscale structure (such as in the stratiform precipitation area of a mesoscale convective system) as an important criterion defining a "good" forecast, and may have focused on different features than scientists with a pure mathematical background. Participants were not asked to focus on a particular region, nor were examples of a "good" forecast provided.

The three forecast models that provided precipitation forecasts were run for the 2005 Spring Program sponsored by the Storm Prediction Center and National Severe Storms Laboratory (SPC/NSSL) (http://www.nssl.noaa.gov/projects/hwt/sp2005.html). Two of the three numerical models (provided by NCAR and NCEP Environmental Modeling Center [EMC]) were run on a 4-km grid, while one was run on a 2-km grid (CAPS), denoted wrf4ncar, wrf4ncep, and wrf2caps, respectively. Additional information on the model configurations can be found in Kain et al. (2008). All forecasts and observations were interpolated onto the same ~4 km grid as the geometric cases. For space considerations, we do not present them here, but all the forecasts are available online at http://www.ral.ucar.edu/projects/icp/. Later, we focus on two cases from the perspective of each new verification method category and compare them to the subjective scores and traditional metrics.

3. APPLICATION OF NEW METHODS FOR GEOMETRIC CASES

We now apply the new verification methods to the set of geometric cases to ascertain whether they can correctly diagnose the applied errors. Specifically, we ask:

- Does geom001 score better than geom002 and is the error correctly attributed to displacement?
- Can the method detect the aspect ratio error in geom004?
- Can the method discriminate between the high frequency bias in geom003 and the very-high bias in geom005? How about the equal displacement of the feature centroids?

Traditional methods do not diagnose displacement or structure errors and with a single intensity threshold, they also do not provide information about performance on different spatial scales. Table 2 summarizes the answers to these questions for the geometric cases, which are discussed in greater detail below. The paper of Gilleland et al (2009) answers a different set of questions that address the nature of the information provided by the various spatial verification methods.

Table 2. This table indicates which categories of verification method clearly diagnose the types of error illustrated in the geometric cases. A method may be sensitive to a type of error, but not clearly diagnose it.

	Method Category			
Error type *geometric case*	Neighborhood	Scale-separation	Feature based	Field deformation
displacement *geom001* *geom002*	Indirectly	No	Yes	Yes
frequency bias *geom003* *geom005*	Yes	Indirectly	Yes	Yes
aspect ratio – “quasi-rotation” *geom004*	No	No	possible	Yes

a) Neighborhood methods applied to geometric cases

In general, the neighborhood methods look in progressively larger space-time neighborhoods about each grid point and compare the set of probabilistic, continuous or categorical values from the forecast to the observation (Gilleland et al., 2009; Ebert 2009). For our cases, the time neighborhood is ignored.

While they do not usually provide direct information on feature displacement and aspect ratio, neighborhood methods are sensitive to these errors. For example, with the simple upscaling method, Ebert (2009) found that the forecast with small displacement error (geom001) showed skill greater than zero for any neighborhood larger than the native grid spacing, but for geom002, there was no skill for any reasonably-sized neighborhood. The same was found by Mittermaier and Roberts (2009) for Fractional Skill Score (FSS; Roberts and Lean, 2008), a method that compares forecasted and observed fractional coverage of events within neighborhoods. Mittermaier notes, however, that if an unreasonably-large neighborhood is used, the skill does reach perfection for unbiased forecasts such as geom001 or geom002. Using FSS, Mittermaier and Roberts (2009) show that the geom001 forecast has skill at scales above 200 km. Aptly, this neighborhood size corresponds exactly to the prescribed separation between the two objects.

The incorrect aspect ratio and displacement found in geom004 produces a FSS profile that reaches skillful values at 550 km, but one cannot easily tell whether a change in score is due to displacement, rotation or both. The forecasts with high frequency biases (geom003 and geom005) do not exhibit useful skill for any scale examined, but the FSS is lower for geom005 at larger scales relative to geom003. The geom005 case has slightly greater FSS at the smaller scales because of the forecast-observed area overlap. Ebert (2009) applies additional neighborhood methods to the geometric cases including Multi-event contingency table (Atger, 2001) and practically perfect hindcast (Brooks et al., 1998), with similar results.

While the composite method of Nachamkin (2009) is categorized as a features-based approach by Gilleland et al. (2009), it adopts a neighborhood-like strategy to measure performance at different scales. Nachamkin (2009) uses a metric called the conditional bias difference to assess the difference between forecast and observation at different neighborhood sizes centered on contiguous regions of precipitation. As expected, the geom001 forecast performed better than geom002 at all neighborhood sizes, though it performed poorly for the smallest neighborhoods. The conditional bias difference also indicated that geom003 was the worst forecast at intermediate distances, but for large distances geom005 was worst. Because of its overlap with the observation, geom005 was best in the smallest neighborhoods. As the neighborhood gets smaller, the results converge to traditional metrics,

In summary, the neighborhood methods indicate that forecast geom001 is more skillful than geom002 as desired. The method does not explicitly characterize error as displacement, amplitude, or structure, so the aspect ratio error in geom004 is difficult to diagnose. The bias errors in geom003 and geom005 are also difficult to diagnose using the particular neighborhood scores shown in Ebert (2009); however, computing and displaying the neighborhood frequency bias in addition to the GSS addresses this problem.

b) Scale-separation applied to geometric cases

Casati (2009) revisits the Intensity-scale separation (IS) method introduced in Casati et al. (2004) and applies it to the geometric cases of the ICP. She uses wavelets to decompose the error between the observed binary field and the forecast binary field at different intensity thresholds. Following her method[*], we show the average IS skill scores for geom001 and geom002 using a precipitation threshold of greater than zero (Figure 4). For geom001, the error is concentrated at a spatial scale of 128 km because the width of the features is close to 128 km and so is the displacement . But for geom002, the error shifts toward the 512-km and 1024-km scales, which reflects the larger displacement (~800km). For the geom003 and geom005 cases, in which the forecast area is too large, the skill is reduced for the largest scale (2048 km), with a larger drop for the severe over-forecast in geom005. While the IS skill score is certainly sensitive to the displacement and frequency bias errors illustrated in the geometric cases, it does not tell how much of the total error is caused by each error type. However, the frequency bias errors do have prominent signatures in the wavelet energy spectra (Casati, 2009). The incorrect aspect ratio of geom004 is not picked up clearly by either the IS method nor the energy spectra.

[*] These average scores were calculated from a set of 64 overlapping, but randomly positioned tiles encompassing the features of interest. The "tiling" method reduces the variability associated with the starting position of the wavelet. A more rigorous approach was applied by Casati et al. (2009) in which all possible tile positions were used. That is why her Fig. 7 has slightly different IS skill scores using the low intensity threshold.

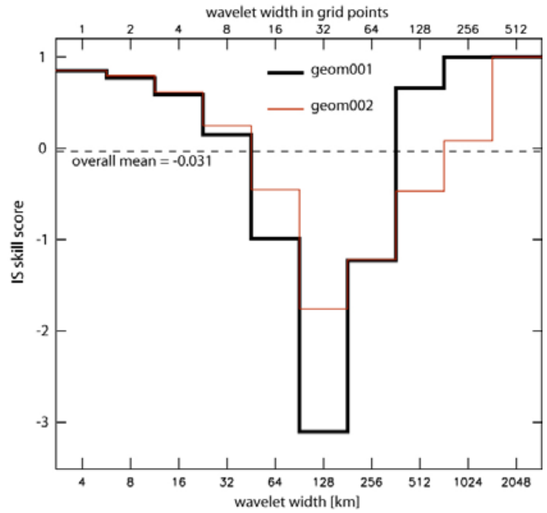

Figure 4. This figure shows the IS skill scores for geom001 (black) and geom002 (red). As described in Casati (2009), the difference field between the binary forecast and observation for non-zero precipitation was decomposed into its constituent spatial scales. For the case with the small displacement error (geom001), the low IS skill scores are concentrated at a wavelet width of 128 km, while in geom002, low IS skill scores spread to larger spatial scales, consistent with the larger displacement error.

Similar to the spectral decomposition method (Harris et al., 2001), the variogram method (Marzban and Sandgathe, 2009; Marzban *et al*., 2009) describes global aspects of the precipitation field. Essentially, variograms measure texture as a function of scale. For sparse fields such as precipitation, the variogram results depend on whether or not zero-pixels are included in the analysis. For example, the displacement, rotation, and frequency errors are only evident if zero pixels are included (Marzban *et al*., 2009). Moreover, if the zero pixels are included, the frequency biases of geom003 and geom005 are very prominent.

Scale-separation methods exhibit similar behavior to the neighborhood methods with the geometric cases. Both types of methods are sensitive to displacement error, but they do not directly measure it. The scale separation methods are also not sensitive to the aspect-ratio error. Using this strategy, the stretching in the longitude-dimension for geom003-geom005 would show up as different spatial fluctuation in the longitudinal direction. Like the neighborhood methods, the scale separation methods can be sensitive to the position of the precipitation within the domain; Casati (2009) suggests a tiling method in which the statistics are calculated within a square subdomain that is shifted grid-point by grid-point around the perimeter of the domain. The statistics are calculated at each possible position and then averaged. This methodology is analogous to the neighborhood approach, where the best practice is to calculate statistics at each possible grid point instead of using every n-th grid point (i.e. overlapping neighborhoods, not discrete).

c) Features-based methods applied to geometric cases

In general, the features-based methods (Gilleland et al., 2009) divide a gridded field into objects by identifying clusters of points above an intensity threshold. Some methods apply a smoothing operator or band-pass filter before identifying the features. The choice of smoothing operator and threshold will determine what type of objects are defined, whether they be small, isolated high intensity rain cores, or larger mesoscale precipitation shields. In the case of cluster analysis (Marzban et al., 2009), the gridpoints are grouped into a prescribed number of clusters. This number determines the relevant spatial scale.

In the geometric cases, depending on the threshold, either the high-intensity or the low-intensity ellipses represent the objects. As long as the forecast ellipse is matched to the observed ellipse, attributes such as position, size and shape can be compared. Beyond a critical distance, no match occurs and hence no diagnostic information is derived.

The SAL quality measure (Wernli et al. 2008), which separately considers aspects of structure (S), amplitude (A), and location (L), captures the location error and frequency bias in the geometric and perturbed cases (Wernli et al., 2009). It does not provide an actual distance, but instead it gives a normalized location error (L), with higher values associated with greater displacement error. For geom001, L = 0.11, and for geom002, L = 0.39. The frequency bias of forecasts geom003 and geom005 is reflected in the S and A terms, as the forecast objects are both too large (S), and the domain-average precipitation is too high (A) (Wernli et al., 2009). SAL does not capture the aspect ratio error in geom004.

Similar to SAL, the contiguous rain area (CRA) method (Ebert and Gallus, 2009) treats displacement, volume (amplitude), and structure (pattern) as separate components of error, but for matched objects in the forecast and observed fields. The CRA method gave very intuitive results for the combinations of displacement error and frequency bias in geom001-geom003. The proportion of error attributed to pattern was largest (63%) in geom004, the case with incorrect aspect ratio; but geom003 and geom005 also had pattern error (34% and 41%, respectively) because of their stretching in the x-dimension.

The Method for Object-based Diagnostic Evaluation (MODE; Davis et al., 2009) looks at characteristics of the objects such as location, area, intensity, curvature, and orientation, and attempts to match objects in the forecast and observed field based on these characteristics. MODE diagnoses the displacement error and frequency bias of each forecast object in the geometric cases perfectly (Davis et al., 2009). MODE can evaluate the distribution of intensity values (based on percentiles) within an object, but not the relative position of those intensity peaks. Therefore, the aspect error in geom004 is diagnosed as a simple rotation.

The position of the high-intensity ellipse within the low-intensity ellipse is not one of the matching criteria.

The Procrustes object-oriented verification scheme originally described by Micheas et al. (2007) is similar to MODE in that forecast and observed objects are matched and merged based on their attributes. Moreover, prior to matching, the objects may be identified using a filter in Fourier space as opposed to physical space (Lack et al., 2009). The quality of the forecast is determined by the fraction of matched objects, and the amount of displacement, dilation, and rotation that is necessary to better match the forecast object to the observed object. Lack et al. (2009) show that the Procrustes method measures the right amount of displacement and frequency bias in the geometric cases. As in MODE, the location of the higher intensity ellipse within the low-intensity ellipse is not a factor because a single intensity threshold is used to identify the features. Therefore, the dilation error in geom004 is indistinguishable from rotation error.

d) Field deformation methods applied to geometric cases

Field deformation methods attempt to morph the forecast and/or observation fields to look like each other, minimizing a score such as RMSE. They can quantify the overall dissimilarity between the forecast field and the observation field. As long as the search radius is large compared to the displacement error, the error in the forecast objects is consistent with a subjective evaluation of the error. Interestingly, the field deformation method is the only one to truly capture the aspect ratio error in geom004. Other methods rely on a single intensity threshold and treat the case as simple rotation. As seen in Figure 5, the field deformation vectors clearly attribute the error to dilation and contraction, not rotation. This figure comes from Gilleland et al. (2009b), a paper in which the image warping technique is applied to the geometric cases.

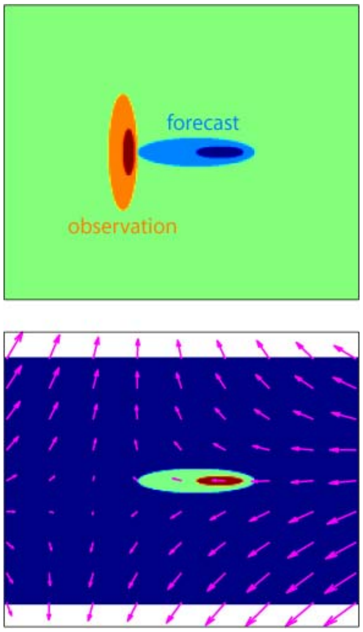

Figure 5. This figure, adapted from Gilleland et al. (2009b), shows the forecast and observation field for the geom004 case (top panel). The image warping technique (Gilleland et al., 2009b) attempts to morph the forecast to the observation, and the resultant displacement vectors are shown in the bottom panel with the original forecast field. The aspect ratio error is clearly illustrated by the deformation in the displacement vector field.

Keil and Craig (2009) use a pyramidal matching algorithm to derive the displacement vector field. Contributions from observation and forecast space are averaged to give a scalar amplitude and displacement score. The two components are combined into a single displacement and amplitude score (DAS). For geom001 the contribution of the displacement component dominates the DAS, while for geom002, when the feature is outside the search distance, only the amplitude error contributes. The geom002 forecast is not close enough to the observed precipitation area to be considered the same object, so it is treated as a false alarm, and the displacement vectors attempt to remove it by shrinking its area.

Marzban and Sandgathe (2009b) apply optical flow to a set of forecast objects very similar to geom001 and geom002. They show that small displacement errors (smaller than the scale of interest) have very simple and easy to understand optical flow fields. As in the DAS method of Keil and Craig (2009), as the forecast object gets further from the observation and beyond the scale of interest, the optical flow vectors tend to converge about the forecast object in an attempt to shrink the false alarm. About the observed object, the optical flow diverges in an attempt to expand forecasted precipitation where the feature was missed.

The Forecast Quality Index (FQI; Venogopal et al., 2005) straddles the features-based and field-deformation categories. It does not generate a displacement vector field, but it quantifies the location error in a global sense. The numerator of the FQI is the partial Hausdorff distance (PHD), a scalar that quantifies the distance between binary images. The PHDs for geom001 and geom002 (41 and 191 grid pts) are slightly less than the actual displacements, but a slight modification of the PHD formulation (using the 100th percentile instead of the 75th) results in perfect diagnosis of the displacement. The frequency bias in geom003 and geom005 is not detected because without zero pixels, the mean and variance are perfect for the geometric forecasts. FQI can be altered to use all pixels, but this alternate formulation may not be desired because a) it is sensitive to the amount of empty space in the domain, and b) the PHD is already sensitive to aspect ratio and frequency bias. For example, in cases geom003-geom005, where the forecast objects are all displaced by 125 points, the PHDs (based on the 75th percentile) are 145, 141, and 186 points, each distance reflecting the 125-point rightward shift, plus stretching in the x-dimension.

4. APPLYING NEW METHODS TO REAL CASES

a) Traditional and subjective scores for real cases

Real precipitation forecasts from nine days were shown to a panel of experts and were subjectively evaluated. First some classic verification scores are presented. We will focus on the wrf4ncep model in this paper, but additional cases and model comparisons are found in accompanying papers in the ICP special collection (Casati, 2009; Davis et al., 2009; Ebert 2009; Ebert and Gallus, 2009; Gilleland et al., 2009b; Keil and Craig, 2009; Lack et al., 2009; Lindstrom et al., 2009; Marzban and Sandgathe, 2009; Marzban et al., 2009; Mittermaier and Roberts, 2009; Nachamkin, 2009; Wernli et al., 2009). The traditional scores for the wrf4ncep model for all nine days are shown in Figure 6. All three models tended to overdo precipitation, with wrf4ncep being most prone to over-prediction. Except for 25 May, the frequency bias was usually over 2 for a threshold of 6 mm.

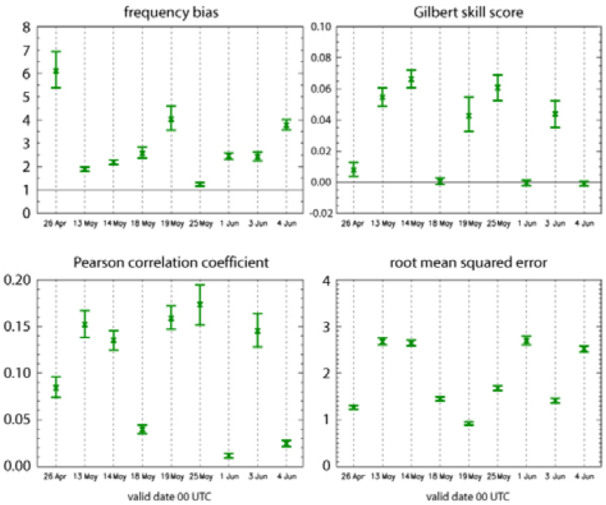

Figure 6. Traditional verification scores for nine 24-h forecasts of 1-h accumulated precipitation from the wrf4ncep model. The metrics include frequency bias and GSS for a precipitation threshold of 6 mm (top row) and the Pearson correlation coefficient and RMSE (mm) (bottom row). 95% bootstrap confidence intervals were calculated using the percentile interval method in the MET package (NCAR, 2009). The bootstrap replicate sample size was 0.8 times the number of matched data pairs (0.8 x 601 x 501), and the data were sampled 1000 times with replacement.

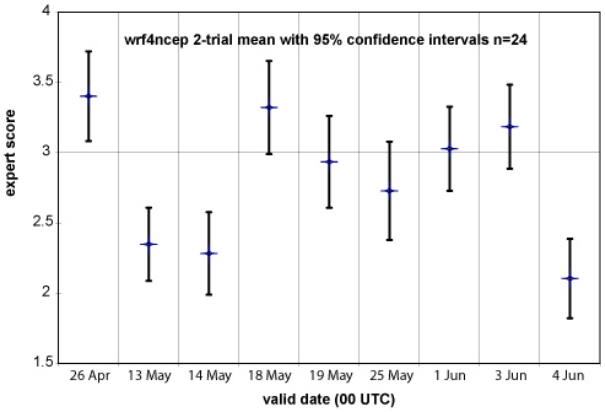

Figure 7. Mean subjective scores for the wrf4ncep model. Twenty-four participants rated nine forecasts on a scale from 1 to 5 with 1 being poor and 5 being excellent. The group evaluated the cases twice and these scores are based on the 2-trial mean. The capped vertical bars are the +/-2 standard error, or 95% confidence interval, for the true mean assuming the sample mean is normally distributed.

As expected with small grid spacing, the traditional scores are quite low (Figure 6) with GSS< 0.1 for a threshold of 6 mm. Precise prediction of convective precipitation is nearly impossible 24 hours in advance and these low scores reflect that difficulty. At thresholds of 6 mm and above, the wrf4ncep forecasts valid on 18

May, 1 June and 4 June all have negative GSS, suggesting no skill, but their subjective scores are quite variable (Figure 7). Looking exclusively at 1 Jun and 4 Jun, the 2-trial mean subjective score for 4 Jun is significantly lower than the score for 1 Jun (with a significance level of less than 0.0001) based on a paired two-tail student-t test with 23 degrees of freedom. The panel members were not asked to explain why they rated the 1 June wrf4ncep forecast better than the 4 June forecast, but these positive aspects of the 1 June forecast could play a role: the 1 June forecast captured the overall shape of the long band of convective precipitation curling from North Dakota to Texas (Figure 8) and the heavy precipitation cores in the Texas panhandle were forecast close to observed precipitation cores. On the other hand, for 4 June, one's attention is drawn to the strong north-south band of forecasted precipitation in Missouri and Arkansas; it is obviously a false alarm. Based on the subjective reviews, one might consider the 1 Jun forecast superior to the 4 Jun forecast, but the GSS says both forecasts are equally poor (Figure 6). Do the new verification methods make a distinction?

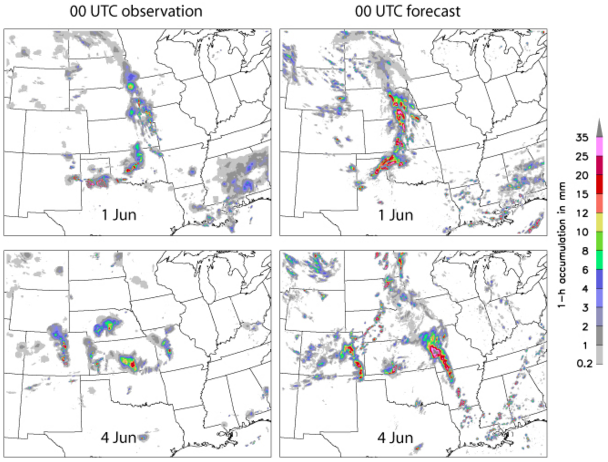

Figure 1. Stage II 1-h precipitation observation (left) and 24-h wrf4ncep forecast (right) valid at 00 UTC on 1 Jun and 4 Jun 2005.

b) Neighborhood methods applied to real cases

According to the neighborhood methods applied by Ebert (2009), the 1 Jun case scored better than the 4 Jun case. Using the FSS, Mittermaier and Roberts (2009) showed that for 1 Jun, the useful distance scale extended down to 60 km and 370 km for thresholds of 1 mm and 32 mm, respectively. For 4 Jun, the forecast was not useful at any threshold or neighborhood size tested.

Using the composite method Nachamkin (2009) found wrf4ncep had the best conditional bias difference (CBD) for high precipitation thresholds (12.7 mm) and neighborhoods less than 124 km. At larger neighborhoods, wrf4ncep's tendency to predict too much precipitation overwhelmed the CBD and its performance dropped below the other two models. Indeed, the wrf4ncep model had very good spatial forecasts for a few high impact events, such as on 25 May and 3 Jun (not shown), but generally predicted too much precipitation.

c) Scale-separation applied to real cases

The IS method was applied in aggregate to the nine real cases (Casati, 2009). No significant difference was found among the three models in terms of IS score. Differences were more apparent in the energy at each spatial scale (Casati, 2009). All three models over-forecasted moderate to high intensity precipitation (2-16 mm) across all spatial scales, but the wrf4ncep bias was worse for small spatial scales (Casati, 2009).

Marzban and Sandgathe (2009) used variograms to assess the texture of the 1 Jun and 4 Jun wrf4ncep forecasts. A forecast that has the same texture as the observations will have a similar variogram. It can be plotted as a function of scale. On 1 Jun, the wrf4ncep variogram is fairly close to the observations except at the high end of the scale spectrum. At the largest distances (above 2700 km), the 4 Jun variogram is closer to the observations. The subjective scores were better for 1 Jun. If the experts were concerned about texture, this suggests that variogram distances above 2700 km were not important to the experts. It could be argued that performance above this scale is irrelevant since it is too large to be of much use.

d) Features methods applied to real cases

In terms of the features-based methods, the precipitation overforecasting was reflected in too many forecast objects, too large matched forecast objects, and too intense matched forecast objects (MODE, Procrustes, CRA). Some investigators focused on the nine cases from the subjective evaluation, while others expanded their analysis to a superset of 32 cases from the time period (MODE [Davis et al., 2009] and Cluster Analysis [Marzban et al., 2008]). Overall model performance ranking depended on the method used and the spatial scale of interest. Cluster analysis on a similar dataset by Marzban et al. (2008) showed that wrf2caps was best, followed by wrf4ncep, and finally wrf4ncar. The Procrustes feature-based method (Lack et al., 2009) suggested the largest scales were predicted uniformly well by all three models, but wrf4ncar and wrf2caps were better than wrf4ncep at the small scales. Davis et al. (2009) looked at 32 cases with MODE and found that the wrf4ncar model performed better than the wrf4ncep model based on the MMIF[†] metric, whereas models performed nearly

[†] The primary variable used for Davis et al.'s (2009) MODE evaluation was "total interest" derived from a fuzzy-logic algorithm that compared several attributes of forecast and observed rain features. The median of the set of maximum interest values for forecast or observed

identically based on the MMIO metric. The primary reason for the poorer MODE performance of the wrf4ncep model was a few forecasts with copious false alarms.

Do the features methods prefer the 1 Jun wrf4ncep forecast over 4 Jun? In most respects, the answer is yes. For example, the MODE-based MMI was clearly higher for 1 Jun (Davis et al., 2009). The CSI curve in Marzban et al.'s (2009) cluster analysis for 1 Jun was slightly higher than 4 Jun for all cluster sizes, suggesting 1 Jun was better—although Marzban et al. (2009) note this difference may not be statistically significant. Using the Procrustes method, Lack et al. (2009) found that the total penalty and the average penalty for matching convective segments and cells was consistently less in the 1 Jun forecast than in the 4 Jun forecast. But for the largest (cluster) scale, 4 Jun had smaller penalties on average.

e) Field deformation methods applied to real cases

For field deformation methods applied to individual cases, results depended on the precipitation threshold. Using a threshold of greater than zero, the FQI (Venogopal et al., 2005) was clearly better for the 1 Jun wrf4ncep forecast in terms of both its distance and intensity components. The optical flow technique (DAS; Keil and Craig, 2009) gives a better score on 4 June (1.37) than on 1 June (1.40) for a threshold of 1.52 mm, evidently a result of the large over-prediction of heavy precipitation along the front (Figure 8). Focusing on more intense convection (10 mm), the DAS diverges even further from the subjective ranking. On the other hand, a lower threshold of 1 mm results in slightly better DAS for 1 June (1.22) than for 4 June (1.26), in line with the subjective ranking. This suggests the panel of experts may have been more concerned with low-intensity rain areas on these two dates.

Marzban et al. (2009) compared all cases together and summarized the results of the optical flow technique with joint histograms of displacement vector magnitude and direction. No systematic shift was evident in the joint histograms but analysis of individual cases indicated that the wrf2caps and wrf4ncar models had similar performance and the wrf4ncep forecast was sometimes much better or worse (i.e., its relative performance was more variable).

5. SUMMARY

We constructed simple precipitation forecasts to which we applied some of the latest spatial verification methods. These simple geometric cases illustrated potential problems with traditional scoring metrics. Displacement error was easily diagnosed by the feature-based and field deformation methods, but the signal was not as clear-cut in the neighborhood and scale separation methods, sometimes getting mixed with frequency bias error. Errors in aspect ratio were reflected in some of the scores for neighborhood and scale separation approaches, but were correctly diagnosed by only a couple of specialized configurations of the features methods. Typically, the features-based methods treat aspect ratio error as rotation and/or displacement. The field deformation methods seemed to have the best ability to directly measure errors in aspect ratio.

For the more realistic cases that we tested, each method provided different aspects of forecast quality. Compared to the subjective scores, the traditional approaches were particularly insensitive to changes in perceived forecast quality at moderate hourly precipitation thresholds (≥ 6 mm). In these cases, the newer feature-based, neighborhood, and field deformation methods appeared to give credit for close forecasts of precipitation features or resemblance of overall texture (scale separation methods and variograms), even though the forecasts did not line up exactly with the observations. This was particularly evident in the comparison of the 1 Jun and 4 Jun forecasts from the wrf4ncep model. The overall 1 Jun forecast appeared to be fairly good relative to the 4 Jun forecast according to the subjective evaluation. This assessment was consistent with the results from most of the new methods that accounted for spatial structure or close forecasts.

While this paper is not focused on model resolution or comparing the dynamical cores of different models, the results of the subjective evaluation did suggest, at first glance, that the wrf2caps and wrf4ncar models were more skillful than the wrf4ncep model. This general finding was corroborated by the Procrustes approach (Lack et al., 2009), MODE (Davis et al., 2009), and others. Some methods, like cluster analysis, variogram, and optical flow, found wrf4ncep to be less consistent, but sometimes superior to the other two models (Marzban et al., 2009).

It should be pointed out that the four general categories into which we have classified the various methods are only used to give a general idea of how a method describes forecast performance. Some methods fall only loosely into a specific category (e.g., cluster analysis, variograms, FQI). Further, it is conceivable to combine the categories to provide even more robust measures of forecast quality. This has been done, for example, in Lack *et al*. (2009) who apply a scale separation method as part of a features-based approach. Results shown here should not only assist a user in choosing which methods to use, but might also point out potentially useful combinations of approaches to method developers and users.

Upon examining the results from the subjective evaluation, it became clear that a more rigorous experiment with more controlled parameters would be preferred. A more robust evaluation with a panel of experts would undoubtedly require pinning down the

objects (denoted MMIF or MMIO, respectively) defined a single metric of forecast quality.

region of interest, isolating the potential users' needs, and providing a concrete definition of a good forecast. This type of exercise, which would be best done in collaboration with social scientists and survey experts, is left for future work.

ACKNOWLEDGEMENTS

Thanks to Mike Baldwin for providing the Spring 2005 NSSL/SPC cases for the subjective evaluation. Thanks to Christian Keil, Jason Nachamkin, Bill Gallus, and Caren Marzban for helpful comments. Randy Bullock and John Halley-Gotway wrote the MET statistical software package and graciously answered numerous questions. This work was supported by NCAR. NCAR is sponsored by the National Science Foundation.

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,* submitted.

Atger, F., 2001: Verification of intense precipitation forecasts from single models and ensemble prediction systems. *Nonlin. Proc. Geophys*., **8**, 401-417.

Baldwin, M. E., and J. S. Kain, 2006: Sensitivity of Several Performance Measures to Displacement Error, Bias, and Event Frequency. *Wea. Forecasting*, **21**, 636–648.

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

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

Davis, C. A., B. G. Brown, R. Bullock, and J. Halley-Gotway, 2009. The method for object-based diagnostic evaluation (MODE) applied to WRF forecasts from the 2005 SPC Spring Program, submitted to *Wea. Forecasting*.

Dey C. H., 1998: Office Note 388, GRIB (edition 1). U.S. Dept. of Commerce, NOAA/NWS [Available online at http://www.nco.ncep.noaa.gov/pmb/docs/on388/tableb.html#GRID240].

Ebert, E., 2009: Neighborhood verification – a strategy for rewarding close forecasts. *Wea. Forecasting*, accepted.

Ebert, E. E., and W. A. Gallus, 2009: Toward better understanding of the contiguous rain area (CRA) method for spatial forecast verification. *Wea. Forecasting*, in press.

Gilleland, E., D. Ahijevych, B. G. Brown, B. Casati, and E. E. Ebert, 2009a: Intercomparison of spatial forecast verification methods. *Wea. Forecasting*, accepted.

Gilleland, E., J. Lindstrom, and F. Lindgren, 2009b: Analyzing the image warp forecast verification method on precipitation. *Wea. Forecasting*, in preparation.

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

Kain J. S., S. J. Weiss, D. R. Bright, M. E. Baldwin, J. J. Levit, et al., 2008: Some practical considerations regarding horizontal resolution in the first generation of operational convection-allowing NWP. *Wea. Forecasting*, **23**, 931-952.

Keil, C., and G. C. Craig, 2009: A displacement and amplitude error based score employing an optical flow technique, to be submitted

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

Lin, Y., and K. E. Mitchell, 2005: The NCEP Stage II/IV hourly precipitation analyses: development and applications. Preprints, 19th Conf. on Hydrology, American Meteorological Society, San Diego, CA, 9-13 January 2005, Paper 1.2.

Marzban, C., S. Sandgathe, H. Lyons, and N. Lederer, 2009: Three Spatial Verification Techniques: Cluster Analysis, Variogram, and Optical Flow. *Wea. Forecasting*, submitted.

Marzban, C. and S. Sandgathe 2009: Verification with variograms. *Wea. Forecasting*. In press.

Marzban C., S. Sandgathe, and H. Lyons, 2008: An Object-oriented verification of three NWP model formulations via cluster analysis: An objective and a subjective analysis. *Mon. Wea. Rev.,* **136**, 3392-3407.

Micheas, A., N.I. Fox, S.A. Lack, and C.K. Wikle, 2007: Cell identification and verification of QPF ensembles using shape analysis techniques. *Journal of Hydrology*, **344**, 105-116.

Mittermaier, M. P. and N. Roberts, 2009: Intercomparison of spatial forecast verification methods: Identifying skillful spatial scales using the Fractions Skill Score. Submitted to *Wea. Forecasting*.

Nachamkin, J. E., 2009: Application of the composite method to the spatial forecast verification methods inter-comparison dataset. Submitted to *Wea. Forecasting*.

NCAR, cited 2009: Model Evalution Tools (MET) Users Page. [available online at http://www.dtcenter.org/met/users.]

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

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.

Weisman, M. L., C. Davis, W. Wang, K. W. Manning, and J. B. Klemp, 2008: Experiences with 0-36h explicit convective forecasts with the WRF-ARW model. *Wea. Forecasting*. **23**, 407-437.

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.

Wernli, H., C. Hofmann and M. Zimmer, 2009: Spatial forecast verification methods inter-comparison project – application of the SAL technique. Sumitted to *Wea. Forecasting*.

Wilks, D. S., 2006: *Statistical methods in the atmospheric sciences*. 2nd edition. Elsevier, 627 pp.