An Object-Oriented Multiscale Verification Scheme

STEVEN A. LACK

Cooperative Institute for Research in Environmental Sciences, University of Colorado, and NOAA/Earth System Research Laboratory, Boulder, Colorado

GEORGE L. LIMPERT

Department of Geosciences, University of Nebraska—Lincoln, Lincoln, Nebraska

NEIL I. FOX

Department of Soil, Environmental, and Atmospheric Sciences, University of Missouri—Columbia, Columbia, Missouri

(Manuscript received 10 December 2008, in final form 23 July 2009)

ABSTRACT

Object-oriented verification methodology is becoming more and more common in the evaluation of model performance on high-resolution grids. The research herein describes an advanced version of an object-oriented approach that involves a combination of object identification on multiple scales with Procrustes shape analysis techniques. The multiscale object identification technique relies heavily on a novel Fourier transform approach to associate the signals within convection to different spatial scales. Other features of this new verification scheme include using a weighted cost function that can be user defined for object matching using different criteria, delineating objects that are more linear in character from those that are more cellular, and tagging object matches as hits, misses, or false alarms. Although the scheme contains a multiscale approach for identifying convective objects, standard minimum intensity and minimum size thresholds can be set when desirable. The method was tested as part of a spatial verification intercomparison experiment utilizing a combination of synthetic data and real cases from the Storm Prediction Center (SPC)/NSSL Weather Research and Forecasting (WRF) model Spring Program 2005. The resulting metrics, including error measures from differences in matched objects due to displacement, dilation, rotation, and intensity, from these cases run through this new, robust verification scheme are shown.

1. Introduction

Verification in meteorology has three major goals. One is to simply assess the accuracy of forecasts of different types. The second is to make a comparison between different methodologies of observing the same phenomenon. The third goal is to provide feedback to a model or an end user for modifications of a forecasting model or observing platform. As models and observations have become increasingly complex, covering finer and finer resolutions, the need for advances in verification techniques have become equally important. Older methodologies (i.e., standard skill scores) may give false representations of a fine-resolution forecast's value. It is therefore necessary to develop tools that will match an end user's view of forecast value. For example, a hydrologist may be more focused on whether or not intensity was handled properly, whereas an aviation forecaster may be primarily concerned about the location of hazardous weather and the timing of the event.

One way to tackle this issue would be to assess the different components that constitute the error. Beyond simply generating statistics that measure the relative success or failure of a particular forecast, it is desirable to identify the contributions to the error. This deeper level of information can then be used to better interpret forecast products and the uncertainty in the forecast, while providing details on aspects of the forecast that need improvement. Such an approach would also allow users to adjust the weighting of the components to reflect their priorities.

Corresponding author address: Steven A. Lack, 325 Broadway, R/GSD5, Boulder, CO 80305.
E-mail: steven.a.lack@noaa.gov

DOI: 10.1175/2009WAF2222245.1

The particular verification of interest here is finescale model solutions of precipitation fields, quantitative precipitation forecasts (QPFs), as these remain the toughest obstacle to short-term forecasting, and are essential to the accurate prediction of severe weather threats. In essence, most cases of heavy rain and severe weather are characterized by discrete precipitation objects (convective storms), and the verification of the value of the forecasts depends on methods that can deal with such objects at a range of spatial scales to provide accurate and representative measures of forecast success.

Standard verification metrics such as probability of detection, critical success index, and the Heidke skill score (Wilks 2006) are good baseline statistics that can give a first-order indicator of skill. These scores are even more effective when combined with neighborhood methodologies or performed at different forecast scales. However, these scores alone can be misleading, especially in high-resolution models. A finescale convective product may show skill as part of a decision process that is not captured by these standard statistics; these common metrics may even show zero skill when calculated. Additional metrics are then needed to provide insights into the evaluation process. Object-oriented methods, also referred to as feature-based approaches, can be used in a supplementary nature to common metrics in an evaluation. Object-oriented methods attempt to capture how a forecaster perceives the matching of objects (such as precipitation fields) in a forecast field to the observed field. The major differences in the current methodologies that deal with objects include the following: what defines an object, how they match objects, and what information the method provides the end user (Gilleland et al. 2009).

Overall, object-oriented approaches have the capability to provide more intuitive information than other procedures when dealing with QPFs in finescale models. They provide numerous metrics, which may give the end user a variety of useful parameters to judge the specific attributes of a given forecast. There are some shortcomings to object-oriented approaches, namely, the possibility for counterintuitive matching to occur with rigid object identification schemes. The method presented herein describes a modified and robust Procrustes object-oriented verification scheme, originally described by Micheas et al. (2007). Some of the notable modifications include multiple-scale object identification, cell-by-cell verification metrics, and summary statistics of objects in the forecast and observed domains.

Section 2 contains the details of the verification scheme it self. This starts with an overview of the previously published Procrustes method upon which this scheme is based, followed by descriptions of the changes that have been made to that method to produce the system currently in operation. The novel Fourier transform–based object identification scheme that results in the delineation of objects at a range of spatial scales is included here, as well as means of dealing with unequal numbers of objects in the forecast and observed fields, and weighting of error components to provide comparable quantitative values. Section 3 contains the results of a controlled set of "fake" forecasts used to test the system, which include a series of idealized geometric objects and perturbed forecasts. Section 4 summarizes some interesting cases from the Storm Prediction Center/National Severe Storms Laboratory (SPC/NSSL) Spring Program 2005, and this is followed by a discussion of the verification approach as a whole and future directions and applications of the scheme.

2. Procrustes verification scheme

a. Original Procrustes verification scheme

The original Procrustes verification scheme described by Micheas et al. (2007) was devised to fill the need for a near-real-time object-based verification method for multiple realizations of a radar-reflectivity-based nowcaster. Convective objects in the domain are defined as contiguous pixels of reflectivity greater than a predefined intensity threshold. In addition, the defined object must reach or exceed a predefined size. Once convective objects are identified, each object is assigned an equal number of landmarks along the boundary of the defined objects based on a fixed angle from the centroid. An identification array is then created for each convective object and contains the centroid, minimum, maximum, and mean intensity, along with the bounding landmarks. This allows for a great reduction in data density, and object-based verification for multiple ensemble members is possible in near–real time.

Following Micheas et al. (2007), the full Procrustes fit in (1) is then applied as a matching baseline between two objects, the jth truth and the kth forecast realization (z^j and z^{kj}). The study herein matches one forecast field to one truth field instead of multiple realizations and, thus, $k = 1$:

$$\hat{z}_k^j = \hat{b}_{jk} + \hat{r}_{jk} e^{i\hat{\phi}_{jk}} z^{kj}. \tag{1}$$

From (1), sum of square differences can be derived from the Procrustes fit estimators: the translation component ($\hat{b}_{jk}$), the dilation component ($\hat{r}_{jk}$), and the rotation component ($\hat{\varphi}_{jk}$). Once these components are calculated from a truth object to a forecast object, a penalty function (D) is calculated, which contains the residual sum of squares

(RSS), referred to as the shape error based on the components of (1), and the sum of squares of the average, minimum, and maximum intensities (SS_{avg}, SS_{min}, and SS_{max}) given by

$$D = \mathrm{RSS} + \mathrm{SS}_{avg} + \mathrm{SS}_{min} + \mathrm{SS}_{max}. \quad (2)$$

The penalty function is minimized for all forecast objects given a truth object for all truth objects in the domain to yield the proper object match. A truth object may match the same forecast object as another truth object in the domain. Then, D is summed for all matches and becomes the total domain penalty and can be compared to another member of the ensemble. The smaller D yields the best forecast.

From the mechanics outlined above, it is evident there are several possibilities for improvement to this algorithm.

- The penalty function defined by (2) does not contain distinct information on the rotation, translation, or dilation individually that might be of use to an end user who is more inclined to match objects based on proximity.
- Minimum intensity error might not be of interest as a predefined threshold to identify objects in a continuous field (radar reflectivity) as used as a preprocessing step.
- A forecast object may match more than one truth object in a domain, but a forecast object may not end up being matched and subsequently not penalized in the domain, giving a misleading overall penalty for the domain.
- The ability to examine error components of individually matched objects may be of use to some end users.
- The object detection scheme may be too simplistic for some applications or in situations where one may want to stratify results on different spatial scales.

It is therefore necessary to come up with an alternative approach to identifying objects, especially when dealing with differences in convective mode. The ideas outlined above are explored in the following subsections.

b. *Overview of adjustments from the original Procrustes scheme*

The Procrustes verification technique, described originally in Micheas et al. (2007), was slightly modified for use with meteorological precipitation fields for this study. The adjustments made from the original version of the Procrustes scheme are minor and reflect changes in the interpretation of errors only and not with the original mathematical methodology. Some of these modifications have been addressed in Lack et al. (2007).

Within the current Procrustes framework, a flexible matching scheme was implemented. In the original scheme, matching objects was accomplished by minimizing the shape and intensity differences. The current version of the scheme allows for matching based on a user-weighted cost function. The original cost (penalty) function, D, in Micheas et al. (2007) utilized only the shape and intensity sum of squares errors. The new cost function in (3) is a cell-by-cell-based function that includes shape (RSS), intensity (SE_{avg}, SE_{max}, and SE_{min}), dilation (SE_D), rotation (SE_R), and translation (SE_T) error components. Each error component can be user weighted (w_i, where $\sum w_i = 1$) to put more emphasis on one or more of the variables above. Common weighting schemes include equal weighting on all variables, weighting just on translation, or eliminating the intensity terms by setting the weight to zero:

$$D = w_1 \mathrm{RSE}^{0.5} + w_2 \mathrm{SE}_{avg} + w_3 \mathrm{SE}_{max} + w_4 \mathrm{SE}_{min} + w_5 100(1 - \mathrm{SE}_D) + w_6 100(\mathrm{SE}_R) + w_7 \mathrm{SE}_T^{0.5}. \quad (3)$$

Once the cost for each pair of matched objects is compiled, an overall mean squared error is assessed for all matches in a forecast domain. These errors can be decomposed back into the original variables making up the cost function for each cell-by-cell match in the domain. It is also possible to have a set of cost functions and to minimize the result of this set of cost functions instead of applying a single cost function.

In the newest version of the Procrustes scheme, all observed (truth) objects are matched to one forecast object; additionally, all forecast objects are matched to one truth object. This allows for the identification of each cell as a hit, miss, or false alarm. The terminology of hit, miss, and false alarm is not what is meant when compiling standard dichotomous skill scores such as the probability of detection (POD) and the critical success index (CSI). A hit is simply an observed object being matched to the lowest cost forecast object. A miss is flagged only when two observed objects match to the same forecast object: the lower cost of the match is considered a hit, while the higher cost is considered a miss. A false alarm is when a forecast object is not used to match to an observed object. The importance of flagging hits, misses, and false alarms allows for important stratifications that can be made during the analysis of results by the end user.

A flexible object identification approach is an additional strength of the Procrustes verification scheme. The original scheme utilizes a minimum size and minimum intensity threshold to identify objects, alone. This scheme has power when dealing with noncontinuous

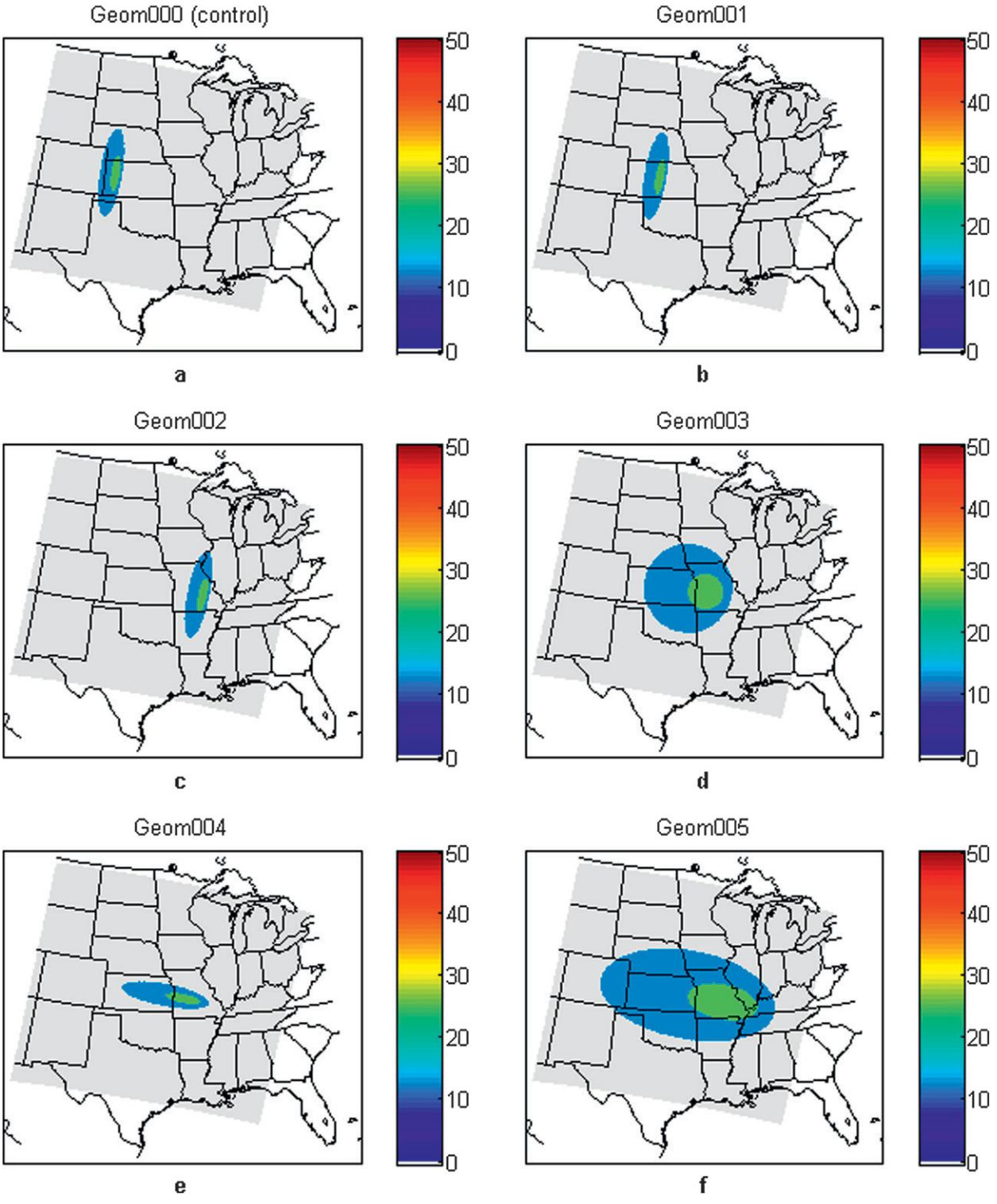

FIG. 1. The geometric cases used in this study to test the schemes: (a) GEOM000 [the control case (truth)], (b) GEOM001, (c) GEOM002, (d) GEOM003, (e) GEOM004, and (f) GEOM005. Intensity is in mm.

TABLE 1. Summary of verification statistics for the geometric cases. Note that perfect dilation is scored as a squared error (SE) of 1, and the magnitudes of the displacement are in pixels in this case.

	GEOM001	GEOM002	GEOM003	GEOM004	GEOM005
No. cells truth	1	1	1	1	1
No. cells forecast	1	1	1	1	1
Min intensity SE	0	0	0	0	0
Max intensity SE	0	0	0	0	0
Avg intensity SE	0	0	5×10^{-3}	0	5×10^{-3}
Dilation SE	1	1	0.19	1	0.11
Rotation SE	0	0	0	2.47	2.47
Translation SE	2500	40 000	6857	15 625	5277
Magnitude displacement (pixels)	50	200	125	125	125
Direction displacement	90	90	90	90	90
Residual SE	0	0	14 893	0	4244
Cell penalty	50	200	286	372	473
Hit, miss, or false alarm	Hit	Hit	Hit	Hit	Hit
Total penalty	50	200	286	372	473

fields such as discrete radar-derived video integrator and processor (VIP) levels. However, information can be lost when filtering out the lower-frequency information. The new adaptation of the Procrustes scheme allows for the minimum threshold object identification, as well as a multiscale approach, utilizing Fourier decomposition techniques. This method is described in section 2c.

c. *Multiscale object identification*

The crux of the new object identification scheme involves decomposing a radar image to identify structures of different scales within the image. This is particularly useful in radar-based nowcasters or models that can yield a postprocessed simulated reflectivity field. The idea is to divide objects by spatial scale in a manner that replicates a meteorological view of the hierarchy of precipitation structures. However, there are other potential benefits and this work only provides a demonstration of the possible means of applying this approach. One other application of this that has been explored is the division of individual radar reflectivity images into features corresponding to different storm types to which different reflectivity–rainfall (Z–R) relationships can be applied (Limpert et al. 2008). This application leads to descriptive terms for each scale that are not essential in the verification application, but allow association of the scales with commonly perceived meteorological distinctions.

Primarily, this method separates features of different scales. We have chosen a particular set of parameters, but one could choose to vary these, or add more strata. For convenience, the terminology employed refers to the identification of ''clusters,'' ''segments,'' and ''cells'' as representative structures of three different spatial scales that correspond to an intuitive meteorological classification. The cluster, segment, and cell identification schemes work by identifying structures of different spatial scales within an image. Structures of different scales may be contained within one another to represent a hierarchy of structures within a reflectivity image. That is to say that a cluster may contain multiple segments, each of which may contain multiple cells. However, there is no requirement that smaller-scale features be contained within larger ones.

The decomposition into different spatial scales is accomplished through a discrete Fourier transform (DFT). Applying Gaussian bandpass filters at multiple frequencies and recomposing the filtered images into the spatial domain, the three spatial scales are realized. The Fourier transform (FT) has been applied in many fields that use image processing and is a standard method for decomposing an image into multiscale constituents (Gonzalez and Woods 2002). The FT is performed on an image in which radar reflectivity (power) is measured in units of dBZ and the bandpass filters are dependent on the scale and resolution of the domain. The bands used for filtering the image were determined empirically by analyzing several cases. Once the bandpass filters have been applied, the image is recombined using the inverse DFT. The result is a series of several bandpass-filtered images in the spatial domain that show how much power is in an image at a given point within the frequency band that was passed by the filter. Examining the filtered images yields information about how much power is within an image at each point within the selected frequency band. The particular combination of parameters used in the cases examined here can be found in Limpert (2008). However, these selections would be changeable by a user based on the application.

As with the broader identification scheme, each cluster, segment, and cell must meet a series of criteria for that scale. Criteria must be met so that the ratios of the power in one frequency band to the power levels in

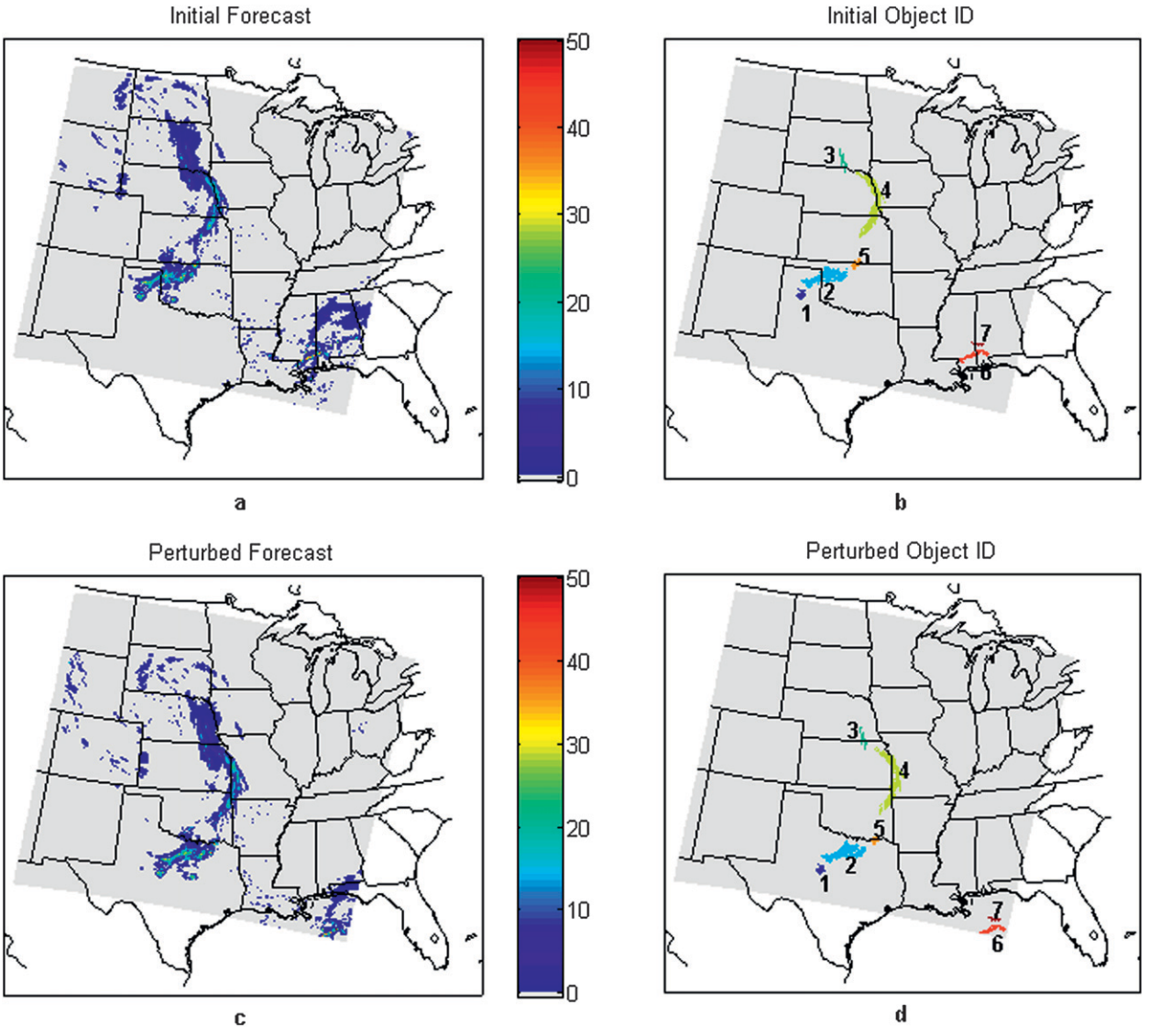

FIG. 2. A large perturbation example showing (a) the original (truth) field with (c) the corresponding identification of cells, and (b) the perturbed forecast field with (d) the corresponding identification of cells. Intensity is in mm.

other bands are below a defined threshold. The purpose of this is to require that, although the region is classified in more than a single band, the signals are of similar strengths. Without these additional criteria, it is likely that many larger convective systems would meet both requirements in a manner that is inconsistent with subjective expert analysis. Additionally, other power thresholds within frequency bands must be satisfied to meet each classification. Cluster identification only examines power within the lowest-frequency band associated with convection whereas segment and cell identification examine progressively higher-frequency bands. For each scale it is possible that very small regions may exceed all the thresholds and unrealistically small regions may be marked as objects that should only be identified as larger-scale structures. To prevent this from occurring, additional image processing is performed by specifying a filtering mask that ensures objects exceed particular size thresholds. This masking process also has the effect of removing detail around the edges of structures; however, this is only used for identifying objects and the original image is restored, having been tagged, prior to the verification process.

3. Results

The Procrustes scheme was utilized as part of an exercise from the National Center for Atmospheric

TABLE 2. Perturbed case 5 cell matches for (left) the standard cost function and (right) the cost function that puts more weight on shape-based matching. The average magnitude displacement and average direction displacement appear with the total domain penalty and the average cell penalty within the domain. Lower penalties are indicative of higher skill.

Standard cost function			Shape-weighted cost function		
Truth ID	Forecast ID	Hit, miss, or false alarm	Truth ID	Forecast ID	Hit, miss, or false alarm
1	2	Hit	1	1	Hit
2	4	Hit	2	2	Hit
3	4	Miss	3	3	Hit
4	3	Hit	4	4	Hit
5	4	Miss	5	5	Hit
6	6	Hit	6	6	Hit
7	6	Miss	7	7	Hit
2	1	False alarm			
2	5	False alarm			
6	7	False alarm			
Avg magnitude displacement	97.4		Avg. magnitude displacement	87.5	
Avg direction displacement	186		Avg direction displacement	149	
Total penalty	2066		Total penalty	108	
Avg cell penalty	270		Avg cell penalty	108	

Research (NCAR) Intercomparison Verification Workshop (Gilleland et al. 2009). The workshop involved examining numerous verification methodologies on similar datasets to illustrate the strengths of the verification procedures and not to be conclusive results on the abbreviated datasets used. During the meetings, it was determined that a set of synthetic cases were to be used in testing the schemes that included geometric objects and different perturbations from a single precipitation analysis field. In addition, several real cases from the SPC/NSSL Spring Program 2005 (Kain et al. 2008) were also examined and compared to the subjective results found from the NCAR ICP. The data for comparison include three versions of the Weather Research and Forecasting (WRF) model, including a 2-km Center for Analysis and Predictions of Storms (CAPS) run (WRFCAPS), a 4-km NCAR run, and a 4-km National Centers for Environmental Prediction (NCEP) run.

a. Geometric cases

The geometric cases show the power of employing object-oriented verification approaches. The geometric cases include rotations, dilations, and translations of a single elliptical object overlaid on a conterminous U.S. (CONUS) grid (Fig. 1). The elliptical object has two intensity levels embedded within it. The Procrustes scheme in this case utilized the simple object identification method of setting a minimum intensity and minimum size threshold. The Procrustes verification scheme results of the matching and associated error decompositions are shown in Table 1. The error decomposition shown is from the controlled geometric case (GEOM000) to the rest of the geometric cases (001–005). The Procrustes scheme accurately portrays the displacement (in terms of pixels) for each geometric case, although this could easily be converted to kilometers. In addition, the Procrustes scheme provides information on rotation and dilation error attributes. Overall, the worst forecast is GEOM005, which indicates that the forecast object is considerably larger than the observed object, there is large translation from forecast object to observed object, and the objects are rotated 90° out of phase. Although the single forecast object in each case matches the single observed object in this idealized situation (note in Table 1 that the cell penalty is equal to the total penalty), a fixed radius of interest can be applied within the scheme to ignore matches that exceed this radius. For example, if the radius were picked to be 150 pixels, the GEOM002 would be classified as a false alarm instead of a hit, although the penalty values would not change.

b. Perturbed cases

The perturbed forecasts start to show the flexibility of the Procrustes scheme to utilize different user-defined weighting schemes for matches. For the perturbed cases, the resulting translations found by the Procrustes scheme match with how the forecasts were actually displaced away from the original field in order to create the pseudo–forecast field when using the cost function in (3), where all terms are weighted equally. An interesting case is found in the perturbed case that has the largest displacement away from the original field. In this example, some identified objects are translated to the extreme that they actually become closer to differently shaped objects from the original field (Fig. 2). In this

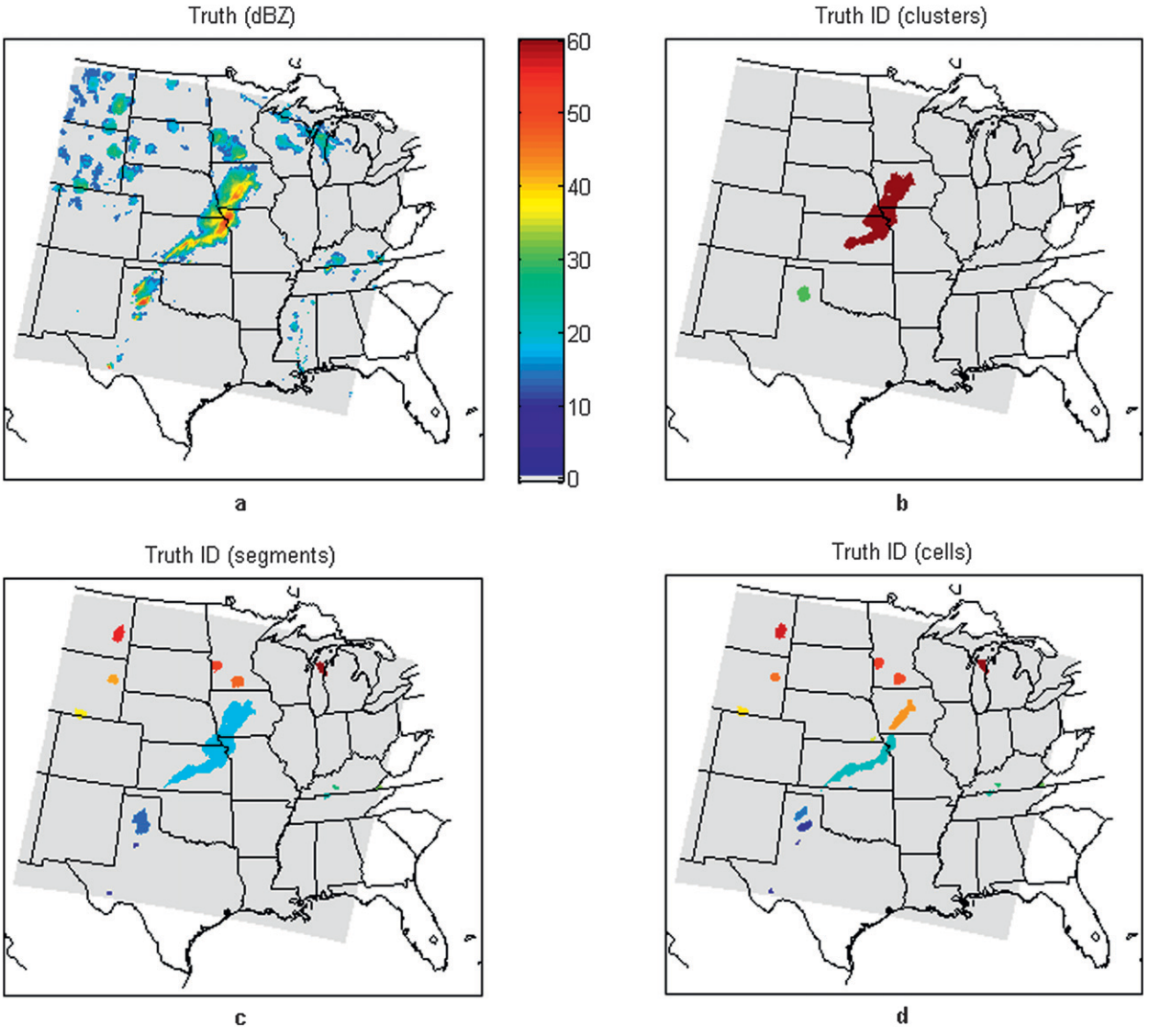

FIG. 3. (a) Stage II data converted to radar reflectivity (dBZ), with object identification at (b) the cluster scale, (c) the segment scale, and (d) the cell scale on 13 May 2005.

case, when utilizing the Procrustes cost function with equal weight on all variables, translation becomes a dominant component over shape. Therefore, the overall perturbation of the entire field is lost as some objects get matched due to their close proximity instead of by their shape. Changing the weighting scheme in favor of primarily shape (residual) errors allows the objects from the original field to match perfectly with the perturbed fields in terms of shape, and the perturbation vector is preserved. Additionally, it can be shown that using the cost function that weights shape more than translation actually minimizes the total penalty of the entire forecast over the domain. Table 2 shows the resultant matching of the objects based on the standard penalty function and the penalty function in which the residual error is weighted heavily. From Table 2, it is shown that the shape-based matching captures the true displacement of approximately 48 pixels to the east and 80 pixels to the south. Overall, the power of the Procrustes scheme to handle objects individually and have different weighted cost functions is advantageous to the end user that may have slightly different concerns in terms of forecast quality.

c. *Spring 2005 real cases*

The real forecast cases from the three WRF variants used during the SPC/NSSL Spring Program 2005 illustrates the usefulness of the modified Procrustes scheme.

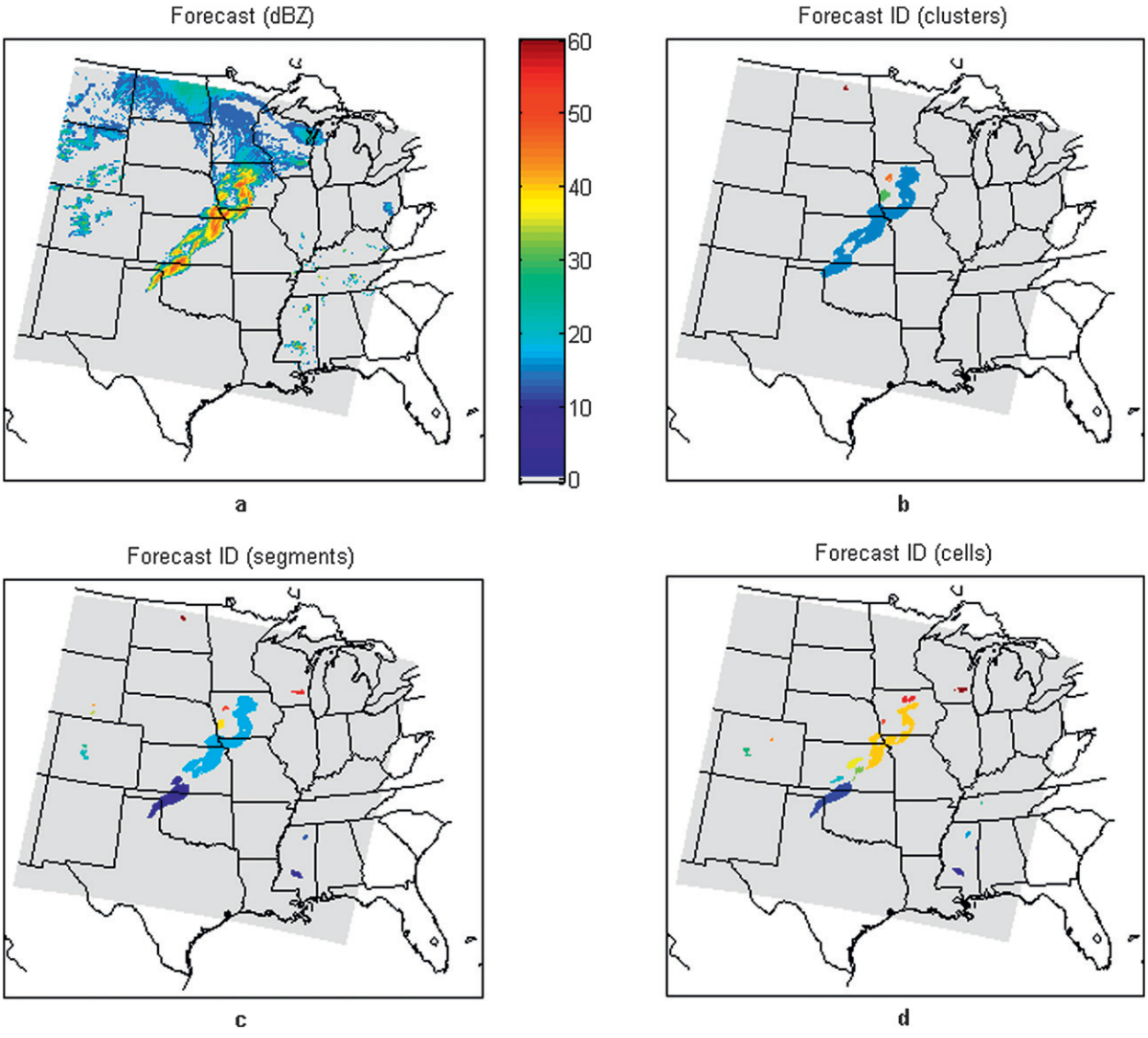

FIG. 4. (a) A 24-h forecast of 1-h precipitation accumulation converted to radar reflectivity (dBZ) for the WRF4NCAR run with object identification at (b) the cluster scale, (c) the segment scale, and (d) the cell scale valid at 0000 UTC 13 May 2005.

For this brief study the forecast of interest in these WRF runs is the 24-h lead time, 1-h precipitation accumulation field. The focus of this study is on sampling the verification information from real cases and on assessing the potential information that can be gleaned from Procrustes methodology metrics. To illustrate the full potential of the Procrustes scheme in comparing precipitation fields from these three models, the multiscale object approach is used in combination with the threshold approach for object detection. To produce the multiscale object identification, each forecast was modified from a discrete 1-h precipitation total field to a continuous radar reflectivity field using the standard Marshall–Palmer Z–R relationship for stratiform precipitation, where R is the rainfall rate (mm h^{-1}) and Z is the radar reflectivity factor (mm^6 m^{-3}):

$$Z = 200R^{1.6}. \quad (4)$$

Since all forecasts were transformed using the same Z–R relationship, the interpretations of the verification scores should not suffer. An example of the multiscale decomposition is shown for an observed and a forecast field on 13 May 2005 in Figs. 3 and 4, respectively. Overall, five object identification methods were used during the verification process, including cluster, segment, and cell identification within the multiscale scheme and a minimum size threshold of 100 pixels (1600 km^2)

TABLE 3. Relative ranking from worst to best of the three WRF runs with notes for each of the nine study days (courtesy of D. Ahijevych).

Date	Worst	Middle	Best	Notes
26 Apr 2005	CAPS	NCAR	NCEP	All above 3/5; no apparent difference
13 May 2005	NCEP	NCAR	CAPS	CAPS–NCAR above 3.5/5; NCEP less than 2.5/5
14 May 2005	NCEP	CAPS	NCAR	All performing poorly; NCEP is the worst
18 May 2005	CAPS	NCAR	NCEP	All above 3/5; no apparent difference
19 May 2005	CAPS	NCAR	NCEP	NCEP near 3/5; CAPS–NCAR near 2/5
25 May 2005	NCAR	NCEP	CAPS	All above 2.5/5; no apparent difference
1 Jun 2005	NCEP	CAPS	NCAR	CAPS–NCAR above 3.5/5; NCEP around 3/5
3 Jun 2005	CAPS	NCAR	NCEP	All near 3/5; no apparent difference
4 Jun 2005	NCEP	CAPS	NCAR	NCAR 3/5; CAPS 2.5/5; NCEP 2/5

and 50 pixels (800 km^2), both with minimum intensities of 20 dB*Z*. The multiscale approach inherently examines the echoes that are significant in terms of convective impact, while the threshold approach will retain some of the lighter-precipitation areas that are important for hydrological and model-feedback impacts. Summary statistics are shown for each of the 9 days of interest in the spatial verification intercomparison study. The summary statistics include the average cell penalty in the domain and the total penalty for the domain. The total penalty accounts for an additional penalty when there are misses and false alarms in the domain. The division of the summary statistics is necessary because one forecast may have a group of false alarms that are small and may have near proximity to a large truth cell, keeping the average cell penalty small, but the total domain penalty would be increased for each false alarm present. This allows the end user to glean more information from the total domain penalty alone without having to look at the cell-by-cell breakdowns for each of the nine individual forecasts.

The results of the five object identification variations are compared to the subjective evaluations for a baseline comparison. The subjective results were compiled by a survey of a group of experts during the NCAR ICP; a summary of these results is found in Table 3. In the subjective analysis the rankings go from 1 to 5, with 5 being the ''most accurate'' forecast against stage II precipitation information. Using a combination of Procrustes verification results with the subjective results, one may extract additional information as to what the subjective evaluators deemed to be important forecast aspects to capture for each of the nine study days. For example, for the 13 May 2005 valid time, the subjective analysis shows that the CAPS and NCAR versions of the WRF scored much higher than the NCEP version. The Procrustes scheme mimics the subjective results (Tables 4 and 5) as there is a higher total domain penalty for the NCEP version of the WRF and a relatively large difference between the NCEP version and the others. A glance to the average cell penalty in the domain may give further insight into the weaknesses of the NCEP run in this case. On the smaller of the convective impact scales (segments and cells in Table 4), the NCEP version of the WRF actually has a smaller average cell penalty. When contrasting to the total domain penalty, it can be noted that the NCEP version must have more false alarms; however, the false alarms combined with hits and misses must be in close proximity to the observed segment and cell convective impact objects within the domain. This is illustrated in Fig. 5, where there seems to be some overforecasting in the southeast United States in the NCEP model; however, it still seems to be relatively close to the observed objects. In this example, the NCAR model does not pick up as many smaller-scale convective impact objects in this location and thus results in a higher average cell penalty in the domain.

Overall, when examining the case using size and intensity-weighted object identification mechanics, the NCAR and CAPS versions of the WRF show signs of outperforming the NCEP version of the WRF on average for this abbreviated 9-day study (Table 5). Upon changing to the multiscale object detection mechanics within the Procrustes scheme, some intuitive results are found. First, the largest convective impact object scale (clusters) shows scores of all three versions of the WRF roughly on par for the 9 days of interest. Breaking down these larger objects into the smaller scales starts to shift the results in favor of the NCAR and CAPS versions of the WRF. As a result, for large-scale convective impact objects, each model has similar skill on average; thus, all forecasts have some utility on the largest convective scale. On the smaller convective scales, the NCAR and CAPS versions resolve the finer-scale structures within the larger-scale objects and the scattered convection that may exist in other regions within the domain. It is intuitive that models running over the same domain with approximately the same physics schemes produce similar results on the largest scales, while the smaller scales have larger and larger differences.

TABLE 4. Total domain penalty and associated average cell penalty within the domain for the nine study days of interest for the multiscale object identification scheme, including the cluster scale (largest), segment scale, and cell scale (smallest). Lower penalties are indicative of higher skill.

Cluster scale				Segment scale				Cell scale			
Total domain penalty				Total domain penalty				Total domain penalty			
	WRF2CAPS	WRF4NCAR	WRF4NCEP		WRF2CAPS	WRF4NCAR	WRF4NCEP		WRF2CAPS	WRF4NCAR	WRF4NCEP
26 Apr 2005	1849	2722	1068	26 Apr 2005	4526	4292	6635	26 Apr 2005	4823	3183	4388
13 May 2005	566	917	1903	13 May 2005	3350	3934	5616	13 May 2005	4433	4047	5958
14 May 2005	2256	2018	1100	14 May 2005	4706	4505	5188	14 May 2005	3897	4006	5097
18 May 2005	524	1168	576	18 May 2005	2916	2269	3534	18 May 2005	2699	2490	3764
19 May 2005	1126	1649	1179	19 May 2005	2264	1961	2556	19 May 2005	2648	2306	4582
25 May 2005	1126	801	836	25 May 2005	2449	2768	5823	25 May 2005	3266	4422	4171
1 Jun 2005	1343	1599	2419	1 Jun 2005	4559	5587	6041	1 Jun 2005	5266	6743	7133
3 Jun 2005	1722	870	1733	3 Jun 2005	4605	3998	3351	3 Jun 2005	5090	4733	4960
4 Jun 2005	1383	1143	2261	4 Jun 2005	5489	4139	7810	4 Jun 2005	5738	3598	9701
Avg	1322	1432	1453	Avg	3874	3717	5173	Avg	4207	3947	5528
Avg cell penalty				Avg cell penalty				Avg cell penalty			
	WRF2CAPS	WRF4NCAR	WRF4NCEP		WRF2CAPS	WRF4NCAR	WRF4NCEP		WRF2CAPS	WRF4NCAR	WRF4NCEP
26 Apr 2005	156	214	135	26 Apr 2005	163	167	165	26 Apr 2005	160	161	161
13 May 2005	189	183	272	13 May 2005	208	213	199	13 May 2005	219	202	190
14 May 2005	269	244	215	14 May 2005	193	186	178	14 May 2005	163	157	152
18 May 2005	262	234	192	18 May 2005	217	194	214	18 May 2005	202	178	221
19 May 2005	225	275	236	19 May 2005	170	180	225	19 May 2005	160	168	202
25 May 2005	263	190	234	25 May 2005	219	240	258	25 May 2005	215	264	215
1 Jun 2005	184	190	204	1 Jun 2005	167	172	171	1 Jun 2005	186	179	166
3 Jun 2005	250	165	184	3 Jun 2005	153	166	165	3 Jun 2005	158	163	155
4 Jun 2005	221	376	221	4 Jun 2005	217	200	200	4 Jun 2005	223	226	190
Avg	224	230	210	Avg	190	191	197	Avg	187	189	184

TABLE 5. Total domain penalty and associated average cell penalty within the domain for the nine study days of interest for the threshold object identification scheme using 100 and 50 pixels as the minimum size criteria with 20 mm as the intensity threshold. Lower penalties are indicative of higher skill.

100-pixel/20-mm threshold				50-pixel/20 mm-threshold			
Total domain penalty				Total domain penalty			
	WRF2CAPS	WRF4NCAR	WRF4NCEP		WRF2CAPS	WRF4NCAR	WRF4NCEP
26 Apr 2005	1332	1372	1458	26 Apr 2005	2939	2555	3333
13 May 2005	1640	1675	6569	13 May 2005	2144	2223	9846
14 May 2005	2359	1161	5687	14 May 2005	4229	2664	7793
18 May 2005	1239	1613	1451	18 May 2005	2006	2454	2241
19 May 2005	NA	NA	NA	19 May 2005	1691	1534	2602
25 May 2005	1182	1612	1953	25 May 2005	1522	1757	2092
1 Jun 2005	4822	4444	1769	1 Jun 2005	9246	8328	8598
3 Jun 2005	1798	1419	2656	3 Jun 2005	3377	2733	4319
4 Jun 2005	2100	1715	6410	4 Jun 2005	2769	2566	14 223
Avg	2059	1876	3494	Avg	3325	2979	6116
	Avg cell penalty				Avg cell penalty		
	WRF2CAPS	WRF4NCAR	WRF4NCEP		WRF2CAPS	WRF4NCAR	WRF4NCEP
26 Apr 2005	333	274	365	26 Apr 2005	319	319	362
13 May 2005	388	398	468	13 May 2005	407	417	489
14 May 2005	266	242	315	14 May 2005	306	270	328
18 May 2005	413	403	363	18 May 2005	391	346	345
19 May 2005	NA	NA	NA	19 May 2005	243	249	323
25 May 2005	286	297	334	25 May 2005	288	304	350
1 Jun 2005	582	513	375	1 Jun 2005	634	548	509
3 Jun 2005	335	235	310	3 Jun 2005	279	246	291
4 Jun 2005	258	211	493	4 Jun 2005	246	219	474
Avg	358	322	378	Avg	346	324	386

4. Conclusions and future work

Utilizing the newest version of the Procrustes verification scheme not only has power in examining many attributes of matched objects, including dilation, rotation, intensity, and translation, but also has power when examining convective impact objects on multiple size scales. The Procrustes scheme is useful when comparing forecasts; however, when used on individual forecast products, a single number cannot be produced to characterize the success or failure of the forecast. The object identification scheme has utility without the matching and Procrustes penalty functions by giving the end user an opportunity to characterize convection in a domain of interest. From the object identification scheme alone, a user can get information of size and intensity distributions of objects within the domain, which can be very powerful when evaluating meteorological models. An example would be when using a high-resolution convective model (WRF-simulated reflectivity) as a supplement to lower-resolution operational products such as the collaborative convective forecast product (CCFP) created by the Aviation Weather Center (AWC). This may give specific end-user information on ways to add structural information to the lower-resolution model by overlaying high-resolution data when comparing to the truth field.

We are currently working on a new version of the Procrustes scheme that has a built-in linear versus cellular discriminator. Using empirically derived classification mechanics based on aspect ratio and eccentricity along with cell size, identified objects are classified as linear or cellular. This can be a useful stratification in verification results on a case-by-case basis. This method may be further used in the future as a matching criterion. A radius of influence can be used to disallow matches if the distance between objects is significantly large as defined by the user. Overall, the Procrustes verification scheme is robust and can serve a variety of end users.

A major near-term future direction of this research is the assimilation of the Procrustes verification tool in the Network Enabled Verification Service (NEVS) currently being developed by the National Oceanic and Atmospheric Administration (NOAA) to serve the Federal Aviation Administration's (FAA's) Next Generation Air Transportation System (NexGen) (Madine et al. 2009). NEVS allows large, disparate verification datasets to be merged and queried with a high degree of flexibility. The addition of the object-oriented multiscale verification

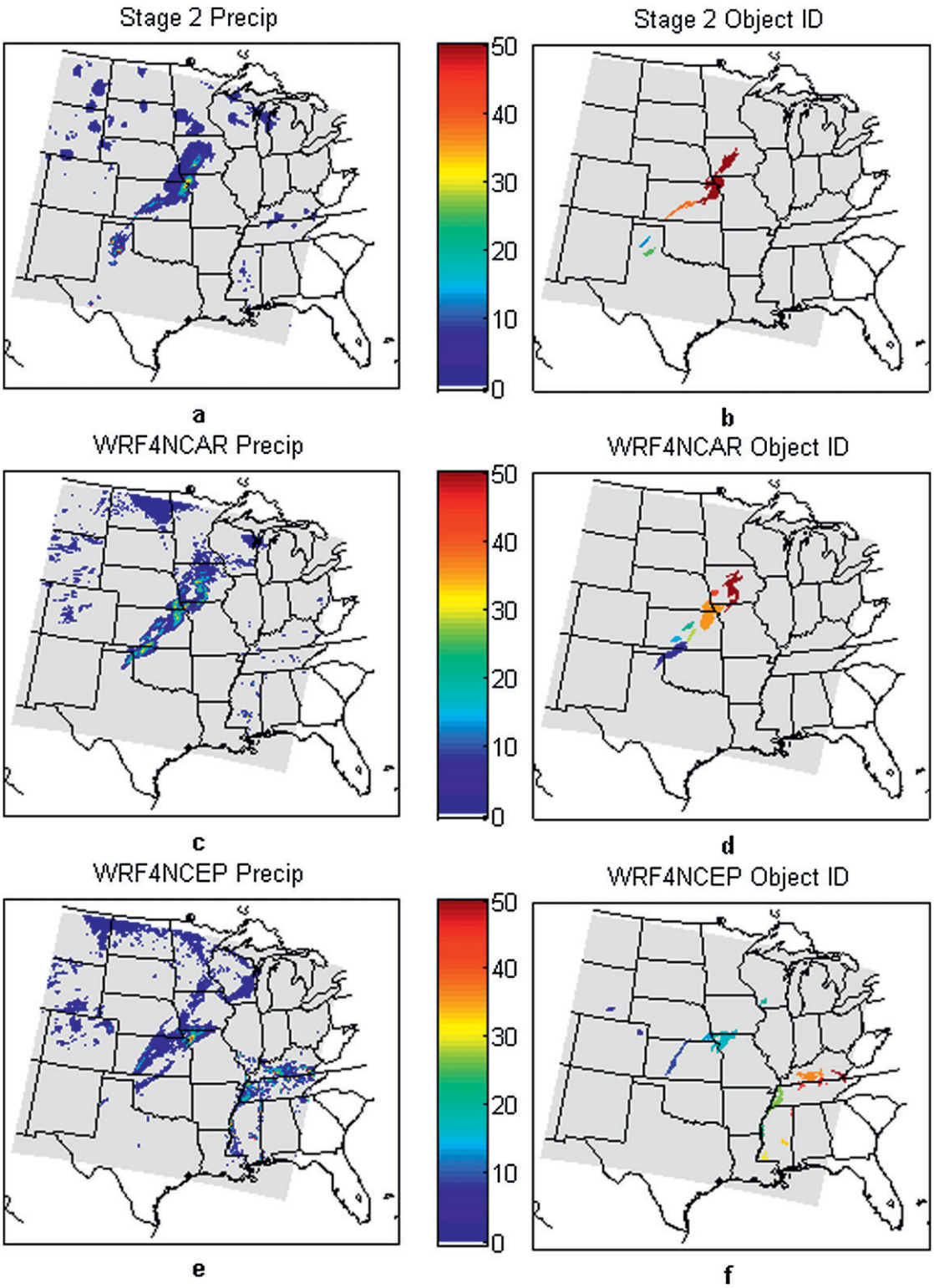

FIG. 5. Comparison of (a) the stage 2 precipitation (mm) observed field and (b) the objects identified using a minimum size of 100 pixels and minimum intensity of 20 mm with the 24-h forecast of 1-h precipitation accumulation from (c) the NCAR version and (e) the NCEP version of the WRF model and (d),(f) their associated object identifications valid at 0000 UTC 13 May 2005.

scheme to NEVS will enhance the utility of the service, most significantly by providing new metrics for convective forecasts used in air traffic management.

Acknowledgments. This research was made possible by National Science Foundation Grant ATM-0434213. George Limpert was funded by the USDA–ARS. The authors thank Sean Madine for helpful comments.

REFERENCES

Gilleland, E., D. Ahijevych, B. G. Brown, B. Casati, and E. E. Ebert, 2009: Intercomparison of spatial forecast verification methods. *Wea. Forecasting,* **24,** 1416–1439.

Gonzalez, R. C., and R. E. Woods, 2002: *Digital Image Processing*. 2nd ed. Prentice Hall, 793 pp.

Kain, J. S., and Coauthors, 2008: Some practical considerations regarding horizontal resolution in the first generation of operational convection-allowing NWP. *Wea. Forecasting,* **23,** 931–952.

Lack, S. A., N. I. Fox, and A. Micheas, 2007: An evaluation of a Procrustes shape analysis verification tool using idealized cases. Preprints, *22nd Conf. on Weather Analysis and Forecasting/ 18th Conf. on Numerical Weather Prediction,* Park City, UT, Amer. Meteor. Soc., 9A.5. [Available online at http://ams.confex.com/ams/pdfpapers/124714.pdf.]

Limpert, G. L., 2008: Evaluating and improving the performance of radar to estimate rainfall. M.S. thesis, Dept. of Atmospheric Sciences, University of Missouri—Columbia, 169 pp. [Available online at http://edt.missouri.edu/Summer2008/Thesis/LimpertG-073108-T11750/research.pdf.]

——, S. A. Lack, N. I. Fox, and E. J. Sadler, 2008: An automated method for detecting precipitation and cell type from radar products. Preprints, *Sixth Conf. on Artificial Intelligence Applications to Environmental Science/24th Conf. on Int. Interactive Information and Processing Systems for Meteorology, Oceanography, and Hydrology,* New Orleans, LA, Amer. Meteor. Soc., J2.4. [Available online at http://ams.confex.com/ams/pdfpapers/134609.pdf.]

Madine, S., and Coauthors, 2009: The Network-Enabled Verification Service (NEVS): Providing verification of weather forecast products in NextGen. Preprints, *25th Conf. on Int. Interactive Information and Processing Systems (IIPS) for Meteorology, Oceanography, and Hydrology,* Phoenix, AZ, Amer. Meteor. Soc., P1.16. [Available online at http://ams.confex.com/ams/pdfpapers/150274.pdf.]

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

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences*. 2nd ed. Academic Press, 627 pp.