9A.4 SPATIAL VERIFICATION OF HIGH-RESOLUTION WEATHER FORECASTS: BEYOND TRADITIONAL METRICS

D. A. Ahijevych*, B. Brown, E. Ebert, E. Gilleland, C. Davis, and L. Holland
National Center for Atmospheric Research, Boulder, Colorado

1 INTRODUCTION

As models are refined and developed, it is imperative to have objective ways to evaluate forecast quality. This is important for comparing different model configurations or tracking performance over time. Greater computing power has allowed finer grid spacing and more explicit handling of previously unresolved circulations (which is crucial for distinguishing high impact events with intense peaks in wind or precipitation). The problem is, as grid spacing decreases, the traditional verification methods become swamped by small-scale errors and they often cannot discriminate between a somewhat-useful forecast and a totally useless forecast. For example, a high-resolution forecasted precipitation field may look very good and be quite useful, but if it is slightly offset from the observations, the traditional verification scores (such as critical success index and equitable threat score) will be dominated by false-alarms and misses due to slight displacement errors. Forecasted and observed events are unlikely to be matched up exactly on a point-by-point basis and the forecast is "doubly-penalized" for false alarms and misses associated with what is essentially the same entity. We would like our verification metric to be sensitive to displacement errors, but at the same time not give an inordinate amount of weight to trivial deviations from the observations (truth). It is with this in mind that we look at several innovative approaches to spatial forecast verification. These methods, which were discussed at a verification workshop in Feb. 2007, can be divided into three broad categories: *feature-based*, *neighborhood approach*, and *scale decomposition*.

2 FEATURE-BASED

Numerous methods have been proposed to look specifically at how well coherent features are forecasted. These methods are also referred to as features-based, object-oriented, and cell-identification techniques. The primary difference among these approaches is how they determine: (a) what constitutes an feature, (b) whether two spatially discontinuous features within a field should be treated as one feature or two separate features, and (c) how they match features from one field (e.g., the forecast field) to the other (e.g., the observation field), and (d) what sorts of diagnostics and/or summary measures they produce. Most of the methods determine (a) by applying a threshold to the raw field.

The contiguous rain area (CRA) approach of Ebert and McBride (2000) determines (b) based on enlarging the feature area and checking whether the features overlap, and (c) is attained by translating the forecast until a pattern matching criterion (e.g., maximum overlap) is met. Displacement, volume and pattern error are found as a consequence of this procedure. Various modifications to this procedure have been proposed (e.g., Grams et al., 2006).

The method developed by Davis et al. (2006), now called the Method for Object-based Diagnostic Evaluation (MODE), addresses (a) not solely by applying a threshold, but also by smoothing. Once features are identified, they are merged and matched by using a dual-threshold method and/or fuzzy logic. An initial threshold defines simple objects which can, in turn, be grouped according to whether they are enclosed by the same lower-threshold contour. The fuzzy logic utilizes information about centroid distance between two features, boundary distance, orientation, area ratio, and intersection area ratio and assigns weights and confidence to each component based on user preferences. The attributes that enter the fuzzy logic algorithm and the final *interest values* provide various diagnostic and summary measures about forecast quality.

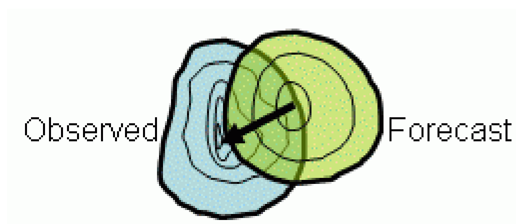

Fig. 1. The object-based approach defines contiguous entities (such as green and blue blobs above) and computes their distance (which may be measured in a number of ways).

The shear number of ways to determine (b) and (c) can be bewildering and make the ultimate choice of fuzzy logic parameters somewhat subjective. An intriguingly simple alternative that deserves further attention is defining the distance in terms of the Baddeley metric, which analytically summarizes differences in object placement and shape with a single value (Gilleland, 2007).

Marzban and Sandgathe (2006) apply statistical cluster analyses in order to define features, so that (c) is not an issue. The results for (d) are traditional verification scores displayed for varying numbers of features, referred to as clusters, in each of the forecast and observed fields.

Nachamkin (2004) uses a composite approach whereby (d) is addressed by looking at the conditional distributions of the forecast events given an observed event occurred and of the observed events given a forecast event occurred.

Micheas et al. (2006) address (b) by using a user-defined minimum object size criterion to tag all individual objects above the intensity threshold selected. The method determines (c) by matching based on proximity and intensity structure of the observed and forecasted objects. Consequently, some observed cells may be matched to multiple forecast objects yielding a higher penalty for over- or under-forecasting of cells. Procrustes shape analysis and a user-defined penalty function are subsequently employed to glean information about forecast performance in terms of rotation, dilation, translation, as well as intensity-based errors over the entire forecast domain.

Finally, Wernli et al. (2007) take a different approach to this general idea. They define features within an area of interest, but no attribution between precipitation objects in the forecast and observations is necessary. Their method, referred to as SAL for Structure, Amplitude and Location, considers these three independent components defined so that a perfect forecast would yield values of zero for all three.

3 NEIGHBORHOOD APPROACHES

Instead of just matching the forecast to the observation gridpoint-by-gridpoint, the neighborhood approach looks at the immediate neighborhood surrounding each point of interest. Statistics such as mean, max, or median, are computed for the neighborhood. The earliest and perhaps simplest of these methods is referred to as upscaling, whereby the forecasts and observations are merely averaged to consecutively coarser scales and compared with traditional scores (e.g., Yates et al., 2006; Zepeda-Arce et al., 2000; Weygandt et al., 2004). Atger (2001) uses a multiple-event contingency table approach that allows for several intensity thresholds to be evaluated as well as other dimensions such as spatial or temporal proximity. The Fractions Skill Score (FSS) of Roberts (2005) and Roberts and Lean (2007) compare the fractional coverage of events in windows surrounding the observations and forecasts. Damrath (2004) utilizes two approaches: one that uses a functional on the neighborhood that employs a proportion threshold within the neighborhood to determine whether an event has occurred or not, and one that employs a fuzzy logic technique that defines events as the probabilities themselves. Brooks et al. (1998) address the issue of rare event verification in this context using a practically perfect hindcast. Other scores under this general paradigm are investigated by Germann and Zawadzki (2004) and Rezacova et al. (2007). Marsigli et al. (2006) introduce a more general approach by comparing the distribution of observations in neighborhoods compared with the distribution of forecasts in neighborhoods.

Two important points should be considered when applying these neighborhood schemes, and the radius defining neighbors is increased (or decreased) in one field and not the other; or neighborhoods are only being considered for one field. First, because the functionals aggregate over the neighborhood, one must be cautious about interpreting comparisons between them because the representativeness of the values for the forecast and observation fields may, subsequently, be quite different. Second, the events may be very different when a threshold defines them. For example, if the event is, "precipitation exceeds 20 mm," then an average over 100 km^2 versus an average over 1000 km^2 results in comparisons of wildly different events.

Among some of the qualitative advantages of these approaches are: (i) the parsimony of the techniques, (ii) the familiarity of traditional scores, (iii) the ability to determine at which scales the forecast performs best, (iv) avoiding the double-penalty problem, and (v) the possibility to apply them to non spatially homogeneous sets of forecasts and observations.

The particular verification questions addressed by these procedures depend largely on: the traditional score utilized, the functions used to summarize the neighborhood values, and how the neighborhoods are determined. For complete information on the specific questions addressed and detailed summaries of each technique, please consult Ebert (2006); only a very brief summary is given here.

4 SCALE DECOMPOSITION

Scale-dependent error is addressed by the scale decomposition approaches. Here, "scale" refers to a single-band spatial filter (e.g., Fourier transforms, wavelets, etc.), whereby one investigates forecast performance by isolating the features at each scale (or wave number). These scales are, therefore, representative of physical features such as separate large-scale frontal systems to smaller scale convective showers. These approaches aim to: (i) assess the scale dependency error, (ii) determine the skill/no-skill transition scale (i.e., assess the scale dependency of the model predictability), and (iii) assess the capability of the forecast to reproduce the observed scale structure.

The intensity-scale (IS) technique of Casati et al. (2004) measures skill as a function of the scales and of the intensity (e.g., rainfall rates). Forecast and observation fields are transformed into binary images by thresholding for different intensities. These images are subsequently separated into the sum of different scale components using a two-dimensional Haar wavelet decomposition, and a skill score based on the mean square error (MSE) of these images is evaluated for each scale component and intensity threshold. The result is a Heidke skill score evaluated at different scales, thereby linking categorical scores with the scale verification approaches.

Mittermaier (2006) expanded the idea by presenting a method for aggregating results for individual (operational) forecasts produced from the intensity scale analysis, and compared the performance of the 12- and 4-km Unified Models against radar rainfall and gridded gauge analyses. The wealth of detailed information these methods provide is useful in a diagnostic context, but for operational verification, there is a need for a method for condensing this detail into manageable and easy to understand quantities.

Harris et al. (2001) look at multiscale statistical properties related to the spatio-temporal scale structure of the fields. In particular, they study the forecast performance by looking at the: Fourier spectrum, structure function, and moment-scale analyses. The method differs from the above methods in that they do not perform the verification on different scales separately. They also apply the technique to the forecast and observation fields separately so that they address the issue of assessing the capability of the forecast to reproduce the observed scale structure. Because the technique does not involve matching forecast phenomena to those of the observations, information about the marginal distributions are gleaned rather than their joint distributions.

5 FUTURE WORK

Some have noted similarities between the object-based methods and the theory of image morphing. Essentially, both rely on anchors or common landmarks in the forecast and observed fields. The correspondence between the forecast and the observation is proportional to the degree of warping that must be applied to the forecast field in order to match the observations. The body of literature on image warping is relatively untapped when it comes to meteorology, and it may offer novel analytic ways to compare geophysical fields (Keil and Craig (2007); Nehrkorn et al., 2003).

6 SUMMARY

A suite of new verification methods has recently emerged to deal with high resolution gridded forecasts. As grid spacing has decreased, forecasts have improved due to less reliance on sub-gridscale parameterization. Even though the forecasts look more realistic, this improvement is not captured well by traditional methods of spatial verification. Traditional methods that rely on a gridpoint to gridpoint comparison between the forecast and observation field will typically show lower skill for smaller grid spacing. This is a fundamental limitation of methods such as critical success index, false alarm ratio, and equitable threat score. Model developers can also artificially increase their skill scores by simply adjusting the bias (Mesinger, 2007). These concerns have led to alternate verification methods based on feature identification, bias-adjusted CSI, neighborhood averaging and/or spatial error decomposition. At a meeting in Boulder in Feb 2007, developers of these new methods convened to show how their unique methods could be applied to common set of gridded observations and forecasts. We present results from this meeting and touch upon some inherent strengths and weaknesses of the new verification methods.

7 REFERENCES

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

Baldwin ME and JS Kain, 2005. Sensitivity of several performance measures to displacement error, bias, and event frequency. Wea. Forecasting, 21:636-648

Brooks HE, M Kay and JA Hart, 1998. Objective limits on forecasting skill of rare events. 19th Conf. Severe Local Storms, Amer. Met. Soc., 552-555 16

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

Damrath U, 2004. Verification against precipitation observations of a high density network--what did we learn? Intl. Verification Methods Workshop, 15-17 September 2004, Montreal, Canada.

Davis CA, BG Brown, and RG Bullock, 2006: Object-based verification of precipitation forecasts, Part I: Methodology and application to mesoscale rain areas. *Mon. Wea. Rev.*, **134**, 1772-1784.

Ebert E. E., 2006: Fuzzy verification of high resolution gridded forecasts: A review and proposed framework. Submitted to Meteorol. Appl.

Ebert EE and JL McBride, 2000. Verification of precipitation in weather systems: determination of systematic errors. *J. Hydrology*, **239**, 179-202.

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

Gilleland E, TCM Lee, J Halley-Gotway, RG Bullock, and BG Brown, 2007. Quantifying differences between defined binary image objects of forecast and analysis of quantitative precipitation for forecast verification. Manuscript in preparation

Grams J. S., W. A. Gallus Jr., S. E. Koch, L. S. Wharton, A. Loughe, and E. E. Ebert, 2006: The Use of a Modified Ebert-McBride Technique to Evaluate Mesoscale Model QPF as a Function of Convective System Morphology during IHOP 2002. *Wea. Forecasting,* **21**, 288-306.

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

Kain JS, SJ Weiss, JJ Levit, ME Baldwin, and DR Bright, 2006. Examination of convection-allowing configurations of the WRF model for the prediction of severe convective weather: The SPC/NSSL Spring Program 2004. Wea. Forecasting, 21:167-181.

Keil, C., and G. C. Craig, 2007: A novel forecast quality measure to rank members in a regional ensemble forecasting system. Submitted to Mon. Wea. Rev.

Marsigli C, A Montani, and T Paccagnella, 2006. Verification of the COSMOLEPS new suite in terms of precipitation distribution. COSMO Newsletter No. 6. Available at http://www.cosmo-model.org/public/newsLetters.htm

Marzban, C., S. Sandgathe, 2006: Cluster analysis for object-oriented verification of fields. Wea. and Forcasting. **21**, 824-838

Mesinger F, 2007. Bias adjusted precipitation threat scores. Submitted to Wea. Forecasting.

Micheas A. C., N. I. Fox, S. A. Lack, and C. K. Wikle, 2006: Cell identification and verification of QPF ensembles using shape analysis techniques. Submitted to J. of Hydrology.

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

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

Rezacova, D., Z. Sokol, and P. Pesice, 2007: A radar-based verification of precipitation forecast for local convective storms. *Atmos. Res.*, **83**, 211-224.

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

Weygandt SS, AF Loughe, SG Benjamin, and JL Mahoney, 2004. Scale sensitivities in model precipitation skill scores during IHOP. 22nd Conf. Severe Local Storms, Amer. Met. Soc., 4-8 October 2004, Hyannis, M.A.

Venugopal, V., S. Basu, 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.

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

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