7B.3 DEVELOPMENT OF AN “EVENTS-ORIENTED” APPROACH TO FORECAST VERIFICATION

Michael E. Baldwin*[1,2], S. Lakshmivarahan[3], John S. Kain[1]

[1] Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma, Norman, OK. Also affiliated with NOAA/NSSL
[2] Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma, Norman, OK. Also affiliated with NOAA/NWS/SPC
[3] School of Computer Science, University of Oklahoma, Norman, OK

1. INTRODUCTION

The goal for this work is to develop new forecast verification techniques that provide useful information on the accuracy of spatial fields containing high-amplitude, small-scale features or “objects”, such as a thunderstorm cluster which might be observed by radar or predicted by a state-of-the-art numerical weather prediction (NWP) model. Traditionally, “objective” forecast verification is performed by an automated system which compares values of forecast and observed variables valid at the same set of points in both time and space. The forecast values are compared to the observed values, and various statistics are computed to measure the accuracy of the forecast field. For meteorological fields that contain high-amplitude, small-scale features, small forecast errors in phase, displacement, or time lag can produce very large differences between forecast and observed scalar variables at specific locations. Traditional approaches to verification may actually represent a forecast system that never predicts small-scale, high-amplitude features as more accurate than one that predicts realistic features (Baldwin et al. 2001). For example, it is possible for a thunderstorm to produce 50mm of rain at one location, while 10km to the south no rain is observed. A forecast that correctly called for this high-amplitude spatial variation in rainfall except with a southward displacement error of 10km would find that the mean absolute error (MAE) from these two locations would be 50mm. Another forecast that called for 10mm at both locations would have a smaller MAE of 25mm, but this forecast does not accurately represent the spatial variation found in the observed rainfall data.

Despite the potential for large errors at specific points in time and space, predictions that contain spatial structures, scales, and amplitudes that are similar to those observed, albeit with phase/displacement errors, may be of considerable value to certain users. For example, forecasters at the Storm Prediction Center (SPC) are often faced with the following forecast problem: What will the dominant mode of convection be tomorrow? In other words, will the dominant events be isolated cellular convection or organized linear convection, or will there be a transition from one type to another? Due to the uncertainty involved in forecasting these kinds of events, the exact location and timing of their occurrence is not as critical as determining the type of event that is likely to occur over a general area in a given time period. Currently, NWP models do not contain enough resolution to explicitly predict spatial patterns of precipitation at the scales which are of critical importance to the SPC forecasters. Forecast output from future NWP models that can explicitly predict different types of small-scale spatial rainfall patterns to some degree of accuracy, even with errors in timing and phase, will be of considerable value to SPC forecasters. However, given the problems with “point-to-point” verification methods mentioned previously, the value of such forecasts will not be expressed when using traditional objective methods of measuring forecast accuracy.

On the other hand, when a person performs “subjective” verification, by visually comparing the forecast and observed fields, the comparison is much less tightly focussed. A human analyst will naturally take errors in phase or displacement into account. Other attributes of the fields will also be considered, for instance, a forecast field with spatial variation similar to the observed field might subjectively be considered of greater value than a forecast field with quite different spatial variability. In order to obtain more useful information on the accuracy of forecasts that contain high-amplitude small-scale features, there is a need to develop “objective” or automated techniques that mimic, as closely as possible, how a human subjectively assesses the skill of a forecast field or spatial maps of meteorological variables.

2. “EVENTS-ORIENTED” VERIFICATION PARADIGM

As a basis for the development of these new automated techniques, the paradigm of “point-to-point” objective verification is expanded to the verification of features, events, or “objects” (Neilley 1993), which are defined as meteorological phenomena. An event or object on a weather map can be defined as a region containing similar meteorological or statistical

**Corresponding author address:* Michael E. Baldwin, CIMMS, 1313 Halley Cir, Norman, OK, 73069
Email: Mike.Baldwin@noaa.gov

characteristics, properties, or *attributes*. Examples of attributes might be the range of values of temperature across a region, or the minimum pressure value of a surface cyclone ("the model is predicting a 982 mb low"). In order to reduce the dimensionality of the problem and to allow for easier interpretation of the results, one should choose a set of attributes that can describe the most important and discriminating aspects of an event in a concise fashion. For example, the i^{th} forecast event could be described by an *attribute vector* of *m* dimension $\mathbf{f_i} = (\alpha_i, \beta_i, \ldots, x_i, y_i)^T$ where x_i, y_i are the attributes associated with the spatial location of this event (perhaps latitude and longitude), and α_i, β_i,..., are attributes that could be associated with the shape, scale, amplitude, orientation, continuity, intermittancy, etc., of the event. Of course, observed events must be described with the same set of attributes, for example, the vector describing the j^{th} observed event would contain $\mathbf{o_j} = (\alpha_j, \beta_j, \ldots, x_j, y_j)^T$.

Depending upon how the events are identified within observed and forecast fields, one could end up with different numbers of observed (n_o) and forecast (n_f) events. For sake of explanation, assume that $n=n_o=n_f$ is the number of forecast and observed events. In order to measure the accuracy of the forecast and quantify the agreement between forecast and observed events, the similarity between these events can be measured. There are numerous possible choices of similarity/ dissimilarity measures, for example, the correlation coefficient between $\mathbf{f_i}$ and $\mathbf{o_j}$ is an example of a *similarity* measure, since the higher the correlation coefficient is, the more similar $\mathbf{f_i}$ and $\mathbf{o_j}$ are. Another possible candidate would be the generalized Euclidean distance, defined as $d_{ij} = (\mathbf{f_i} - \mathbf{o_j})^T \mathbf{A} (\mathbf{f_i} - \mathbf{o_j})$, a measure of dissimilarity. Here **A** is a weight matrix that could allow certain attributes to have greater weight than others, due to differences in units, relative importance, etc. Once the similarity measure has been chosen, overall summary verification scores or accuracy measures could then be obtained. This approach to verifying events would be analogous to the "measures-oriented" approach to verification (Brooks and Doswell 1996). A more comprehensive analysis of the verification information could also be obtained by examination of the joint distribution of forecast and observed events, dubbed the "distributions-oriented" approach by Brooks and Doswell (1996). For example, one could determine the *reliability* of the forecast by determining the average observed event given specific forecast events. This could be considered an extension to the verfication framework outlined by Murphy and Winkler (1987). However, since we assume that the dimension of the attribute vector is *m*, the joint distribution of forecast and observed events will be a *2m*-dimensional multivariate distribution and will require significant factorization to allow for interpretation of the results.

This section outlines the general framework that will be followed to perform an "events-oriented" verification. In order to establish the identity of events within forecast and observed fields and determine the attributes associated with them, there is a need to develop an automated, objective method to recognize events by organizing those regions within the spatial field that posses similar attributes. In order to determine regions within a dataset that exhibit similar characteristics, we naturally turn to the discipline of data mining.

3. CLASSIFICATION USING HISTOGRAM ANALYSIS

The initial goal of this work is to develop a robust automated technique to classify significant and interesting features within a two-dimensional spatial field of meteorological data, such as observed or predicted rainfall. Analysis of such a complex data set can be made at several levels; similiarity of the raw values of the variables at every point in space, 2-D image processing, spectral analysis, etc. As a first step in this multi-faceted analysis process, we choose to classify events by analyzing the similarity of bulk statistical measures representing the distribution of rainfall values across a region of fixed size, using hierarchical cluster analysis (Alhamed et al. 2002) as the classification tool. To validate this system, results from a small target data set are compared to a subjective classification of the rainfall patterns. If the results from this system agree with a subjective classification, we can assume that the choices of attributes and classification schemes are appropriate. In this case, we find that the system successfully classifies the cases into convective and non-convective events with over 90% accuracy. However, further refinement of the classification was less successful and leaves room for future improvement.

An initial target data set has been collected to test various data mining techniques. This data set consists of 1h accumulated rainfall analyses obtained from the NCEP "Stage IV" analysis system (Baldwin and Mitchell 1998) for the period covering late summer/early fall of 2000. The domain size was chosen to be fixed at 128 x 128 4km grid boxes, which is approximately 500km by 500km. A set of 48 separate precipitation events occuring at different times and locations across the United States was selected for inclusion in the target data set.

The selection criteria was based upon the occurence of "typical" rainfall patterns that occur often across the U.S. during the year. Each of these 48 events are considered "objects" for classification. Each case was subjectively classified (by a SPC meteorologist) into the following set of event classes and sub-classes:

CONVECTIVE:

Linear - The precipitation field is more or less consistent along a line, with a large variation in the direction normal to the line.

Cellular - The precipitation field consists of nearly circular-shaped features.

NON-CONVECTIVE:

Stratiform - The precipitation field shows little variation in any direction over a large area.

Orographic - The precipitation field is strongly tied to the shape of the terrain field.

The target data set is relatively small and manageable but well-populated with interesting rainfall events that are desirable for classification. Table 1 shows the distribution of the subjectively classified events across the four sub-classes, the events are not uniformly distributed since the majority of the events are linear or cellular.

Table 1: Subjective event classification for the target data set.

Event type (# of cases)	Case numbers
Convective: Linear (16)	1-16
Convective: Cellular (18)	17-34
Non-convective: Orographic (6)	35-40
Non-convective: Stratiform (8)	41-48

There is a large number of possible choices for attributes that could describe the rainfall pattern over a region. An obvious choice is the amount of rainfall at every point in space obtained from a gridded analysis. With this choice, one would expect the clustering algorithm to produce groups of objects that are similar in a "point-to-point" sense, which is exactly what we want to avoid in this new paradigm of "events-oriented" verification. Therefore, a logical choice for the attributes might be some sort of bulk statistical description of the overall distribution of rainfall across a region. To being this work, parameters of a theoretical statistical distribution fitted to the histogram representing the observed distribution of rainfall amounts across the region are selected as attributes. For the theoretical distribution, the gamma distribution was selected since it is well suited for rainfall data and has been widely used for rainfall histogram analysis (e.g., Wilks 1990). Due to the spatially correlated nature of rainfall, a robust method of parameter estimation of the gamma distribution is required, therefore we selected the generalized method of moments (GMM) estimation technique (Hamilton 1994).

GMM can be considered an extension to the more familiar method of moments technique for parameter estimation. In the method of moments technique, a set of equations are developed to cover the number of unknown parameters found in the model. In the case of the gamma distribution, there are two unknown parameters, α and β, therefore two equations relating these to known quantities are needed. Here, the two equations are found by equating the first two computed sample moments to the population moments. For example, the population mean of the gamma distribution is $\alpha\beta$ and the sample mean is $\bar{x}$ (which is known, computed from the observed data). The population variance (related to the second moment) is $\alpha\beta^2$ and the sample variance is σ^2. Equating these sample and population values provides a set of two equations and two unknowns. This system can easily be solved to find that $\alpha=\bar{x}^2/\sigma^2$ and $\beta=\sigma^2/\bar{x}$. These parameters fit the observed mean and variance exactly, but higher-order moments are not taken into account. In some cases, it may be desirable for the parameters to provide a better fit to the observed skewness (related to the 3rd moment) or kurtosis (related to the 4th moment). The GMM technique allows for this by adding higher-order moments to the equation set, resulting in an non-linear system of equations which can then be solved by least-squares methods. The spatial correlation in the observed data can affect the parameter estimation by modifying the weighting matrix used to determine the weighted sum of squared errors that are minimized by the least-squares optimization (see Baldwin and Lakshmivarahan 2002 for more details). In this work, we tried several different combinations of moments and values of the lag-correlation in the data in estimating the gamma parameters. These estimates of α and β are then used in a classification algorithm in order to find clusters of similar rainfall events.

Since classification is the desired data mining task in this work, hierarchical cluster analysis has been selected as the primary classification tool for this work. Here, objects will be clustered where objects are defined as rainfall events over regions of fixed size, and attributes are the parameters of the gamma distribution fitted to the observed rainfall distribution. The goal of this heirarchical classification scheme is to first group the cases into convective/non-convective classes, then further refine these classes into linear/cellular for the convective class and stratiform/orographic for the non-convective class. The hierarchical cluster analysis method that is chosen for this work is Ward's method

(Alhamed et al. 2002), which is based upon the fact that the total variance of all of the objects is constant and can be partitioned into the sum of between-cluster and within-cluster components. The criteria for adding an object to a cluster is minimizing the squared error, which is the same as minimizing the within-cluster variance, and therefore maximizing the between-cluster variance. This forces the objects found within a cluster to be similar while keeping the clusters as separate as possible. Ward's method has been found to produce good results for meteorological data in previous research (Alhamed et al. 2002).

The results show that this system produces four main clusters (Table 2). This shows that the cluster analysis successfully classified the cases into the subjectively determined convective/non-convective classes. For example, clusters 1 and 2 are unanimously populated by convective-type events (both linear and cellular). Cluster 3 is dominated by convective events, with 3 (out of 18) exceptions. Cluster 4 is dominated by non-convective events, with 1 (of 12) exception. Overall, there are only 4 out of 48 "mis-classified" events, resulting in a 92% classification accuracy. In addition, for this example there is a threshold value of β (=1.5) that cleanly separates the three convective clusters from the non-convective cluster. These results were similar to those found with three and four moments, and by increasing the lag-correlation (see Baldwin and Lakshmivarahan 2002 for more details).

Table 2: Cluster membership for the 2-moment, uncorrelated experiment.

Cluster # (# of members)	cases
1 (8)	1, 3, 14,15,31, 32, 33, 34
2 (10)	2, 4, 11, 13, 16, 21, 22, 24, 26, 30
3 (18)	6,7,8,9,10,12,17,18,19, 20, 23, 25, 27, 28, 29, 39, 42,47
4 (12)	5, 35, 36, 37, 38, 40, 41, 42,44, 45, 46, 48

Now we examine how well the cluster analysis classifies the cases into the four sub-classes (linear, cellular, stratiform, orographic). Returning to the 2-moment, uncorrelated experiment (Table 2), cluster 1 contains four cases that were subjectively classified as linear and four that were subjectively classified as cellular precipitation events. Cluster 2 is also evenly split among the linear and cellular precipitation events with five cases from each. Cluster 3 contains six linear events, nine cellular events, one orographic, and two stratiform events. Cluster 4 contains mainly stratiform (6) and orographic (5) events, with one linear event included. These results show that the CA did not produce clusters with a clear preference for a particular sub-class in this experiment. These results were similar to those found with three and four moments, and by increasing the lag-correlation value, with some variation.

These results should not come as a surprise, since two parameters (α,β) should be able to discriminate between two classes (convective, non-convective) quite well, but have some difficulty in further refining the classification. It is reasonable to expect that additional discriminants will be needed in order to increase the degrees of freedom and allow the classification system to identify finer and more specific classes of events. This sets the stage for future work where we will use; cluster analysis to classify events based upon similarity of the raw values at each point in space, principal component analysis to transform the data, image processing techniques to refine the selection of attributes, etc. The choice of attributes is obviously critical, attributes based upon some measure of the spatial variability and intermittence (e.g., Harris et al. 2001) of the fields could help in refining the classification.

Acknowledgements: This work was partially supported by COMET Cooperative Project S01-32796. Professor Daniel Wilks, Cornell University, kindly provided the software for maximum likelihood estimation of gamma distribution parameters. Ying Lin of NCEP/EMC provided archived Stage IV rainfall data. Ahmed Alhamed provided cluster analysis software. Other software for solving non-linear least squares and matrix inversion was obtained from the Netlib.org repository.

References

Alhamed, A., S. Lakshmivarahan, and D. J. Stensrud, 2002: Cluster analysis of multimodel ensemble data from SAMEX. *Mon. Wea. Rev.*, **130**, 226-256.

Baldwin, M. E., S. Lakshmivarahan, and J. S. Kain, 2001: Verification of mesoscale features in NWP models. Preprints, *9th Conf. on Mesoscale Processes*, Ft. Lauderdale, FL, Amer. Meteor. Soc., 255-258.

________, and ________, 2002: Rainfall classification using histogram analysis: An example of data mining in meteorology. Tech. Rep., School of Computer Science, University of Oklahoma, 21 pp. [Available from School of Computer Science, University of Oklahoma, 200 Felgar St, Norman, OK, 73019]

________, and K. E. Mitchell, 1998: Progress on the NCEP hourly multi-sensor U. S. precipitation analysis for operations and GCIP research. Preprints, *2nd Symposium on Integrated Observing Systems*, 78th AMS Annual Meeting, January 11-16, 1998, Phoenix, Arizona, 10-11.

Brooks, H. E., and C. A. Doswell III, 1996: A comparison of measures-oriented and distributions-oriented approaches to forecast verification. *Wea. Forecasting*, **11**, 288-303.

Hamilton, J. D., 1994: *Time Series Analysis*. Princeton University Press, 799pp.

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

Murphy, A. H., and R. L. Winkler, 1987: A general framework for forecast verification. *Mon. Wea. Rev.*, **115**, 1330–1338.

Neilley, P. P., 1993: Evaluating forecasts using an object-oriented approach. Preprints, *13th Conference on Weather Analysis and Forecasting*, Vienna, VA, 298-300.

Wilks, D. S., 1990: Maximum likelihood estimation for the gamma distribution using data containing zeros. *J. Clim.*, **3**, 1495-1501.