Object-oriented Verification of Reflectivity Fields based on Cluster Analysis: A Report

Caren Marzban*

Department of Statistics, University of Washington, Seattle, WA 98195,
and
Center for Analysis and Prediction of Storms
University of Oklahoma, Norman, OK 73019

August 5, 2006

1 Project Description, Method, and Goals

It is now clear that the performance of numerical weather prediction (NWP) models must be assessed within some type of object-oriented scheme. Considerable progress has been made in this direction (Baldwin, Lakshmivarahan, and Kain 2001, 2002; Brown et al. 2002, 2004; Bullock et al. 2004; Chapman et al. 2004; Du and Mullen 2000; Ebert and McBride 2000). The main thrust of these works is to identify and delineate objects in the two fields (observation and forecast) in a meteorologically meaningful fashion in order to quantify model forecast quality. This injection of meteorology or mental models into the analysis is one of the strengths of these approaches.

An alternative methodology proposed by Marzban and Sandgathe (2006a) does not require quantification of the objects. Specifically, a statistical procedure is employed to perform Cluster Analysis (CA) on the observed and forecast field, separately, thereby automatically identifying "objects" in the two fields. By contrast to the aforementioned papers, the identified objects are not parameterized at all. Although this allows for the possibility that a given cluster may not be physically meaningful, it does have the advantage of allowing for automatic, objective verification.

*marzban@stat.washington.edu

A desirable feature of the CA-based approach is that one can assess the performance of an NWP model as a function of spatial scale. The specific CA algorithm employed is an iterative one, wherein the number of clusters is varied from N, the total number of grid points in the data, down to 1. As such, the number of clusters constitutes a measure of scale; a large number of clusters corresponds to verification on a small scale, and a small number of clusters is associated with large-scale verification. The two quantities, N_o and N_f - number of clusters in the observed field, and that in the forecast field - span a 2-d space, over which one can compute some scalar measure of performance. The resulting "error surface" captures the quality of the forecasts on different scales.

An example, of an "error surface" for precipitation forecasts from an NCAR WRF model forecast is shown in Figure 1. It can be seen that CSI is generally higher when an equal number of clusters exists in both fields. This is not surprising; however, what is surprising is that the height of the surface along the diagonal undulates. The highest CSI values appear on the scale of 5-10 clusters, followed by another peak (albeit lower) at 40-50 clusters; even on the scale of 80-100 clusters there is another low and broad peak. One explanation of this pattern is that the observation and the forecast field have inherent scales, and they better match one another on those scales. Away from the inherent scales, the agreement between the two fields is less.

Instead of performing CA on the two fields separately, and then matching the clusters between the two fields, one can perform one CA on the *combined set* of observations and forecasts. Each cluster will have some number of points belonging to the observation field, n_o, and some number of points originating from the forecast field, n_f. Then, a comparison of these numbers can indicate whether the cluster should be counted as a hit, a false alarm, or a miss. For example, a cluster with $n_o/(n_o + n_f)$ less than some threshold is considered a false alarm, and one with a proportion 1-$n_o/(n_o + n_f)$ exceeding some threshold is considered a miss; a cluster with any other ratio is defined to be a hit. As such, a measure of performance such as the Critical Success Index (CSI) can be computed upon a single application of CA. This variation on the original CA-based method is called Combinative Cluster Analysis or CCA; its advantage being a more efficient and direct approach to the computation of categorical performance measures, such as CSI. Further details of CCA are presented in an article based on this work, by Marzban and Sandgathe (2006b).

The main goal of this study has been to perfect CCA, and then employ it to assess the quality of reflectivity forecasts for three mesoscale NWP model formulations - the NWS/NCEP NMM at four km resolution (nmm4), the NCAR WRF at two km resolution (arw2) and at four km resolution (arw4). CA is performed not only in the 2-dimensional space of (x, y) values, labeling grid coordinates, but also in the 3d space of (x, y, z) values, with z representing reflectivity. The former captures performance only in terms of spatial placement of the clusters, and their size, whereas the latter also includes reflectivity in the assessment of forecast quality.

2 Results and Conclusion

An extensive data set dealing with reflectivity forecasts and observations is currently being compiled by Mike Baldwin of Purdue University. Here, however, only three specific days - May 12 , May 13, and June 4, 2005 - from that data set are employed. Figure 2 displays the observations and the 24hr forecasts according to arw2, arw4, and nmm4. The analysis is performed on only data whose reflectivity exceeds 40dbz for the May 12 and 13 data, and 35dbz for the June 4 data. Although this is done mostly to assure that the sample size of the three days is comparable, it does imply that the model comparisons performed here refer only to high-reflectivity regions. In an upcoming analysis, CCA will be performed on the “un-thresholded” data.

The manner in which CA performs the clustering can be viewed at any iteration of CA. An example with 13 clusters is shown in Figure 3, for June 4, for arw2 forecasts. The top two panels display the hits, with the observation points plotted on the left, and the forecast points plotted in the right panel. Clearly, it is desirable that forecasts should populate these two panels much more so than the remaining panels, which, in these examples, they do. Moreover, the matched colors between the two panels indicate a reasonable matching of the clusters between the two fields. Note that the southern extension of the line of reflectivity located in the left (western) third of the display is forecast too far to the east and the northern extension of the reflectivity in the right or eastern third of the display is forecast to extend too far to the north, yet CCA identifies these areas as “hits”. In other words, these features are forecast with only an error in location (or timing). This is a good example of the strength of object-oriented methods such as CCA.

The false alarms are shown in the middle two panels, and the misses are displayed in the bottom two panels. The arw2 forecast for this particular date misses a significant portion of the eastern extension of the reflectivity band. This is easily determined by CCA along with a few other misses.

As for the performance of the various model formulations, Figure 4 displays the results for arw2, arw4, and nmm4. The manner in which the error-bars for CSI are computed is described in Marzban and Sandgathe (2006b). It should be noted that the error-bars are only crude estimates of standard errors, and not true confidence intervals. As such, they should be interpreted with caution. Referring to Figure 4, the following observations can be made:

1) On all three dates arw2 outperforms arw4 and nmm4 to some extent. This should be expected considering its ability to better resolve smaller-scale features and convective precipitation. (The difference between arw2 and arw4 is least on May 13);

2) This outperformance is generally true across most scales examined, except on the extremes (see next items).

3) On larger scales (small cluster numbers), there are very large variations in

CSI. This reflects the merging of generally unrelated areas of reflectivity together into single clusters causing erratic verification results. Visually scanning the forecasts for the three dates (Figure 2) reveals that no fewer than 10 to 15 clusters should be considered for such a large region.

4) On smaller scales (larger number of clusters), the differences between the models appear to diminish.

5) The performance of all three models falls off with increasing cluster number (smaller scale). Appendix B provides some theoretical explanation of this and other behaviors. It is not surprising that all 3 models have lower performance when forecasting at smaller scales than on larger scales.

In summary, the main conclusions of this study are 1) that the proposed methodology for automatic object-oriented verification appears to be sound, producing reasonably meaningful results; and 2) in our small demonstration sample (involving only three days), the methodology suggests that arw2 forecasts are marginally superior to those of arw4 and nmm4, in terms of the spatial placement of clusters and their reflectivity. And this appears to be true across a wide range of spatial scales. Currently, the analysis is being extended to a larger data set (i.e. more days), and the results will be reported as soon as they are available.

3 Relevance to DTC and WRF Users

The work has demonstrated that CCA can be employed to evaluate mesoscale model performance while avoiding the subjectivity and labor intensiveness of human evaluation or the pitfalls of non-object-oriented automated verification. Providing CCA results to WRF users can aid them in assessing quality in a fashion that is taylored to their specific needs. For instance, a user can pick the model formulation that performs better on the specific spatial scale of interest. In short, CCA offers an object-oriented, automatic, and scale-sensitive framework for verification. Although, the current version of CCA could in principle be made available to WRF users, it would be more beneficial to incorporate into CCA more of the approaches followed by other groups (e.g., at NCAR/RAL). R. Bullock (at NCAR/RAL) and C. Marzban have arranged for a collaboration, taking place in the latter part of August, 2006, that is apt to progress in that direction.

4 Impact on Operational Forecasting

The methodology developed in this project has three facets: The first involves an assessment of forecast quality in an object-oriented setting, and the second is that the assessment can be made objectively and automatically. These two facets have only an indirect impact on operational forecasting, in that the finds can only suggest the use of one model over another, or one scale over another. The third facet, however, has a more direct impact on operational forecasting, because it can convey to the forecaster the conditions leading to good or bad forecasts. This type of information is contained in, for example, Figure 3, where one can identify the type of clusters or weather phenomena that the NWP model does, or does not, forecast correctly.

5 Related publications

Two papers have been produced, with different levels of support from the DTC visitor program. The first paper (Marzban and Sandgathe, 2006a) was mostly complete before the support from DTC, although some of the later improvements to the paper and the underlying methodology did fall within the period of support from DTC. That paper has been accepted for publication in *Weather and Forecasting* and is in press. By contrast, the second paper (Marzban and Sandgathe, 2006b) was based almost entirely on work supported by DTC's visitor program. A portion of the work was also supported by the National Science Foundation.

Acknowledgements

The author would like to acknowledge Scott Sandgathe, Mike Baldwin, Barbara Brown, Chris Davis, and Randy Bullock for contributing to all levels of this project. Partial support for this project was provided by the Weather Research and Forecasting Model Developmental Testbed Center (WRF/DTC), and the National Science Foundation.

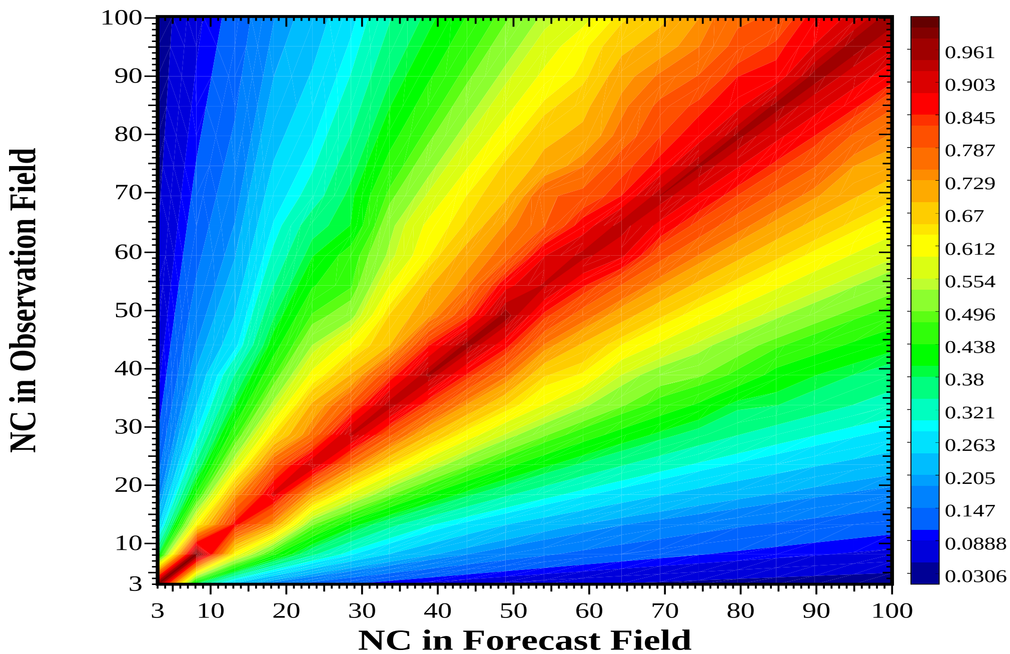

Figure 1. An example of a CSI performance surface for arw2 24hr forecasts, according to the methodology of Marzban and Sandgathe (2006a), with the CA performed in (x, y, z).

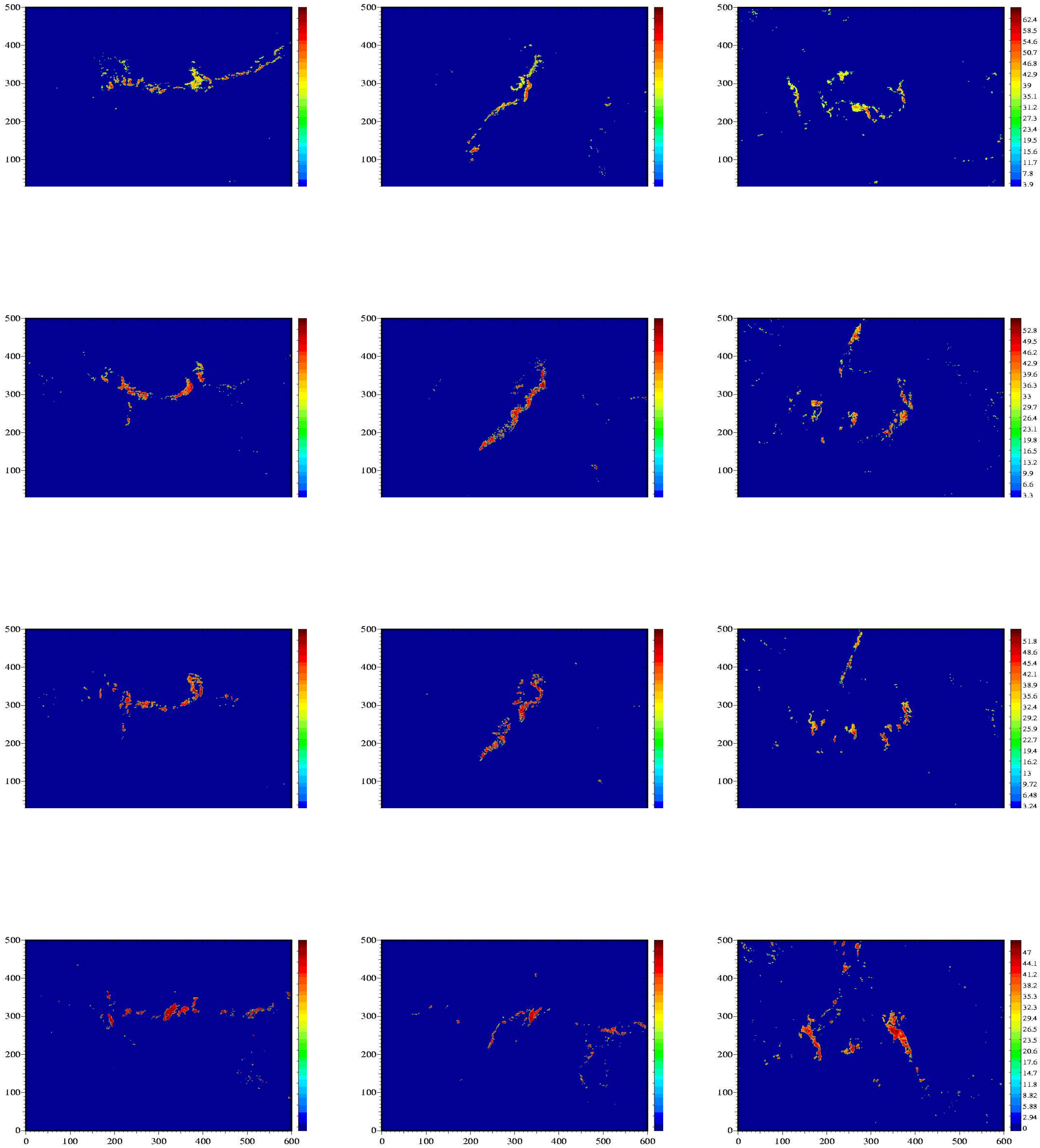

Figure 2. Reflectivity observations (top row), and the corresponding 24hr forecasts according to arw2, arw4, and nmm4 (from top to bottom). The columns refer to the three dates examined - May 12, 13, and June 4, 2005. The coordinates of the region are 70W/30N, 93W/27N, 67W/48N, 101W/44N, which covers the US East of the Mississippi.

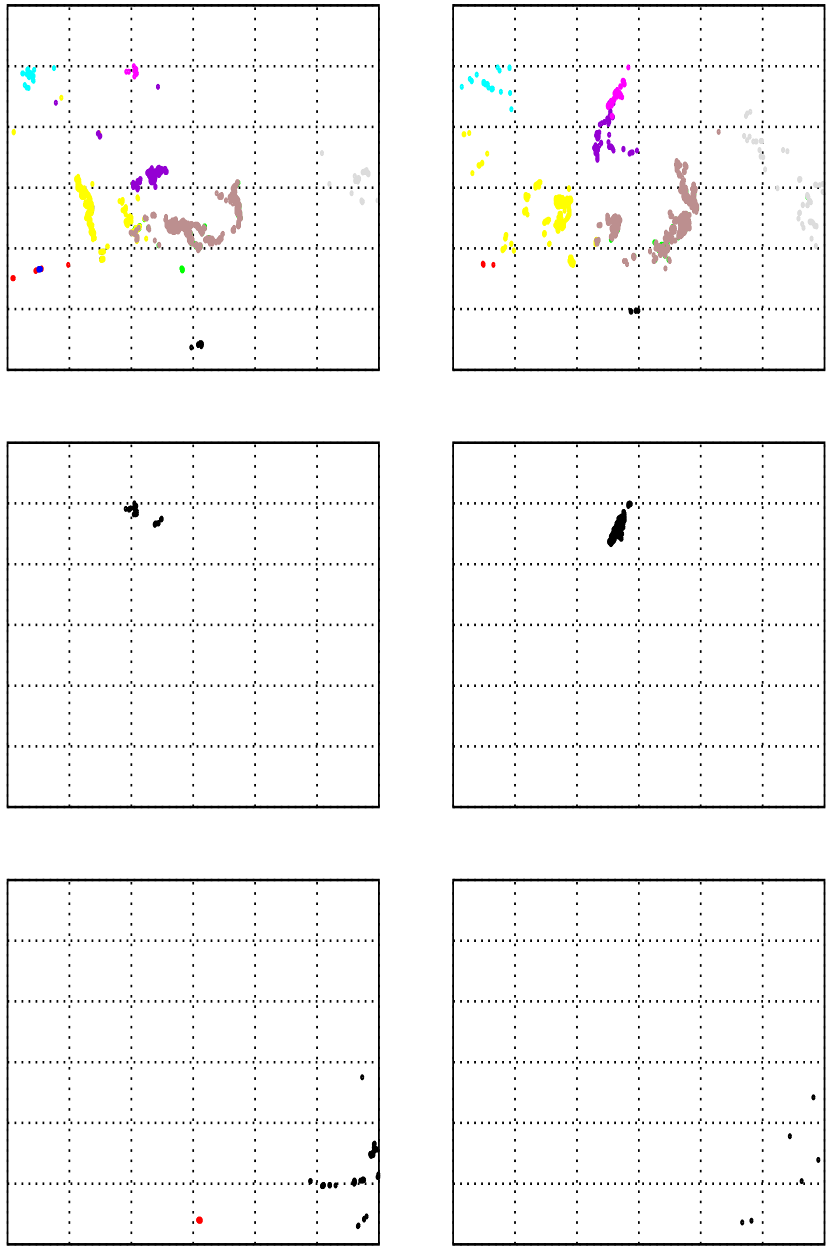

Figure 3. The observed and arw2 forecast fields on June 4, partitioned according the scheme described in text. The colors represent different clusters according to CA, and similar-color clusters in adjacent panels indicate matched pairs. These images are extracted from CA at the iteration corresponding to 13 clusters.

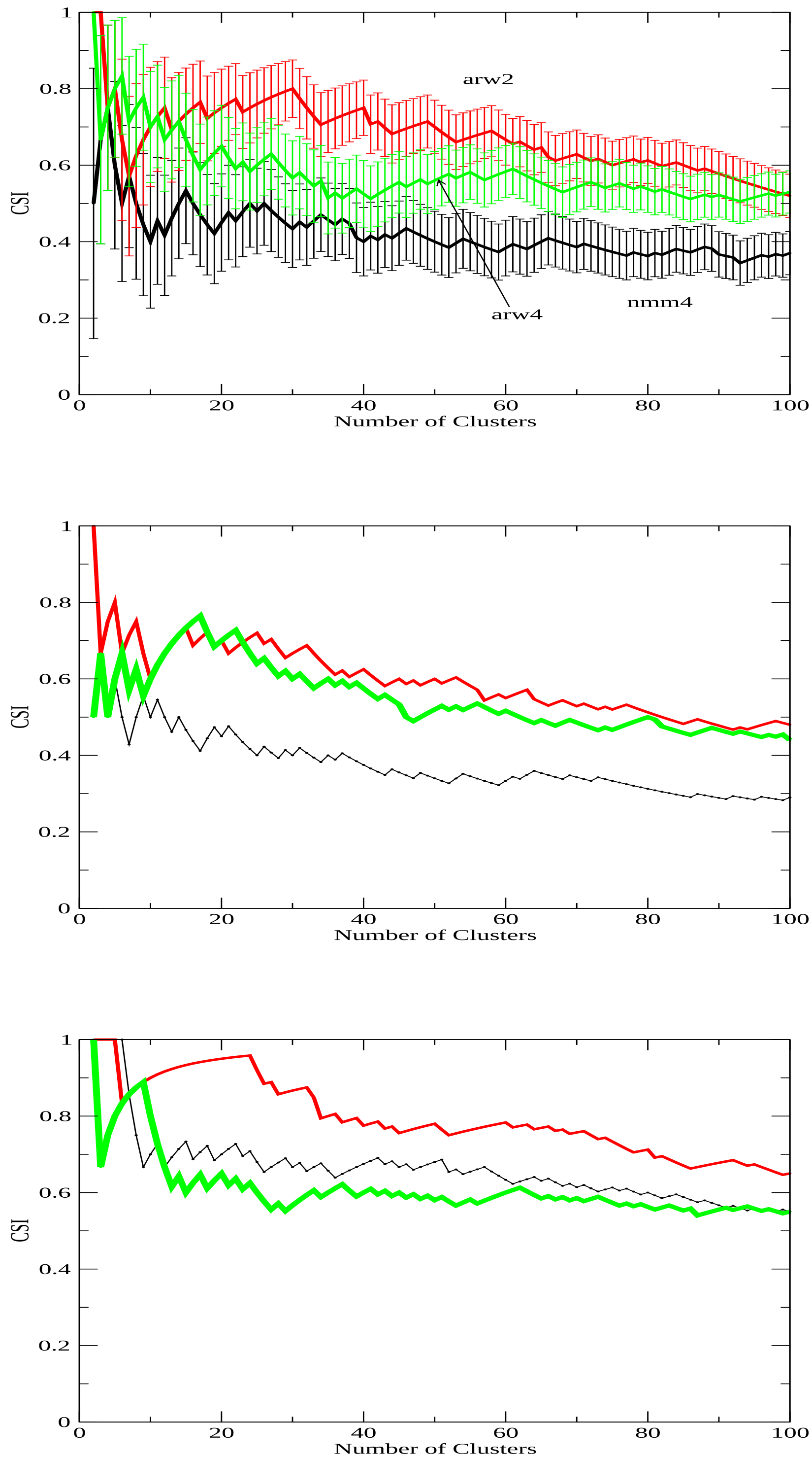

Figure 4. The CSI values for the 3 models on May 12 (top), May 13 (middle) and June 4 (bottom).

References

Baldwin, M. E., S. Lakshmivarahan, and J. S. Kain, 2002: Development of an "events-oriented" approach to forecast verification. 15th Conf. Numerical Weather Prediction, San Antonio, TC. 12-16 August 2002.
Available at http://www.nssl.noaa.gov/mag/pubs/nwp15verf.pdf .

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.

Brown, B.G., J.L. Mahoney, C.A. Davis, R. Bullock, and C.K. Mueller, 2002: Improved approaches for measuring the quality of convective weather forecasts, Preprints, 16th Conference on Probability and Statistics in the Atmospheric Sciences. Orlando, 13-17 May, American Meteorological Society (Boston), 20-25.

Brown, B.G., R. Bullock, C.A. Davis, J.H. Gotway, M. Chapman, A. Takacs, E. Gilleland, J. L. Mahoney, and K. Manning, 2004. New verification approaches for convective weather forecasts. Preprints, 11th Conference on Aviation, Range, and Aerospace, Hyannis, MA, 3-8 October.

Bullock, R., B.G. Brown, C.A. Davis, K.W. Manning, and M. Chapman, 2004: An Object-oriented Approach to Quantitative Precipitation Forecasts. Preprints, 17th Conference on Probability and Statistics in the Atmospheric Sciences. Seattle, 11-15 January, American Meteorological Society (Boston).

Chapman, M., R. Bullock, B. G. Brown, C. A. Davis, K. W. Manning, R. Morss, and A. Takacs, 2004: An Object Oriented Approach to the Verification of Quantitative Precipitation Forecasts: Part II - Examples. Preprint, AMS 2004.

Du, J., and S. L. Mullen, 2000: Removal of Distortion Error from an Ensemble Forecast. *Mon. Wea. Rev.*, **128**, 3347-3351.

Ebert, E. E., J. L. McBride 2000: Verification of precipitation in weather systems: determination of systematic errors. *Jour. Hydrology*, **239**, 179-202.

Marzban, C., S. Sandgathe, 2006a: Cluster analysis for verification of precipitation fields. Accepted at *Wea. Forecasting.*

Marzban, C., S. Sandgathe, 2006b: Cluster Analysis for Object-oriented Verification of Fields: A Variation. Submitted to *Mon. Wea. Rev.*.