An Object-oriented Verification of Three NWP Model Formulations via Cluster Analysis: An objective and a subjective analysis

Caren Marzban[1,2,3]*, Scott Sandgathe[1], Hilary Lyons[2]

[1] Applied Physics Laboratory, University of Washington, Seattle, WA 98195

[2] Department of Statistics, University of Washington, Seattle, WA 98195,

and

[3] Center for Analysis and Prediction of Storms

University of Oklahoma, Norman, OK 73019

December 2, 2007

*Corresponding Author: marzban@stat.washington.edu

Abstract

Recently, an object-oriented verification scheme was developed for assessing errors in forecasts of spatial fields. The main goal of the scheme was to allow the automatic and objective evaluation of a large number of forecasts. However, processing speed was an obstacle. Here, it is shown that the methodology can be revised to increase efficiency, allowing for the evaluation of 32 days of reflectivity forecasts from three different mesoscale numerical weather prediction model formulations. It is demonstrated that the methodology can address not only spatial errors, but also intensity, and timing errors. The results of the verification are compared with those performed by a human/expert.

For the case when the analysis involves only spatial information (and not intensity), although there exist variations from day to day, it is found that the three model formulations perform comparably, over the 32 days examined and across a wide range of spatial scales. However, the higher resolution model formulation appears to have a slight edge over the other two; the statistical significance of that conclusion is weak but nontrivial. When intensity is included in the analysis, it is found that these conclusions are generally unaffected. As for timing errors, although for specific dates a model may have different timing errors on different spatial scales, over the 32 day period the three models are mostly "on time." Moreover, although the method is non-subjective, its results are shown to be consistent with an expert's analysis of the 32 forecasts. This conclusion is tentative because of the focused nature of the data, spanning only one season in one year. But the proposed methodology now allows for the verification of many more forecasts.

1 Introduction

It is now evident that forecasts with a spatial structure should be verified in a manner which accounts for that structure. Such methods are generally referred to as "object-oriented" because they acknowledge the existence of objects in both the forecast and observed fields, and attempt to quantify the quality of the former in terms of various error components, including displacement, size, and intensity. The landmark papers are as follows: Baldwin et al. 2002, 2001; Brown et al. 2004, 2002; Bullock et al. 2004; Casati et al. 2004; Chapman et al. 2004; Davis et al. 2006a,b; Du and Mullen 2000; Ebert and McBride 2000; Hoffman et al. 1995; Marzban and Sandgathe 2006, 2007; Nachamkin 2004; and Venugopal, et al. 2005.

In a sequence of two papers, Marzban and Sandgathe (2006, 2007) have demonstrated the utility of cluster analysis in identifying/defining the objects in one or both fields (observed and forecast). Cluster analysis (Everitt 1980) refers to a set of statistical techniques designed to identify structures in data. The generality of the methodology allows for the objects to be not only 2-dimensional (as in a gridded field), but multi-dimensional entities that include spatial information as well as other dimensions, including the intensity of the field, or the time at which it is recorded. As such, the verification procedure based on cluster analysis has three desirable features: 1) It is object-oriented, 2) allows for a multitude of quantities to be included in the definition and identification of an object, and 3) is fully automated, allowing for non-subjective verification of many forecasts.

Marzban and Sandgathe (2006) proposed to perform cluster analysis on a forecast field and an observation field, separately. The two fields are then compared in terms of the best pairing of clusters within them. This approach allows one to compute any measure of error between the two fields in an object-oriented fashion. An alternative was proposed in (Marzban and Sandgathe 2007), where the clustering is performed on the combined set of forecasts and observations. This approach was named Combinative Cluster Analysis (CCA). In CCA one identifies clusters in the two fields, simultaneously. A single cluster in the combined set can be considered a "hit", if it consists of comparable proportions of forecast and observed grid points Otherwise, it amounts to a "miss" or a "false alarm." In this way, it is possible to produce a contingency table reflecting the quality of a single forecast field. One may then summarize the contingency table by a scalar measure, e.g., the Critical Success Index (CSI), to assess the overall quality of a single forecast field.

Both approaches rely on an iterative/hierarchical variant of cluster analysis, wherein the number of clusters in a field (any field) is varied systematically. On the one hand, one may begin with a single cluster containing the entire data, and then proceed to break it up into ever smaller clusters, ending with as many clusters as cases.[1] Alternatively, the procedure may begin with as many clusters as cases, and proceed to combine them into ever larger clusters, ending with a single cluster containing all the data. Either way, the number of clusters is varied iteratively. The

[1] A "case" refers to a grid point whose reflectivity meets some requirement, e.g., reflectivity $> 20dBz$. This usage of the term is consistent with the statistical usage. For example, one speaks of performing cluster analysis on some number cases.

latter version is called hierarchical agglomerative cluster analysis (HAC); it is the one underlying both approaches examined by Marzban and Sandgathe (2006, 2007), and is also used here.

As mentioned previously, it is important to emphasize that cluster analysis is not constrained to two dimensions, or even to spatial variables. The basic variables of the method can be any set of quantities defined to characterize a cluster/object. In this article, p denotes the dimension of the space in which cluster analysis is performed. And in a verification scheme, it is desirable for the variables to include information about spatial location. Here, if the analysis is done on $p = 2$ spatial coordinates only, it is referred to as an (x, y) analysis. In addition to (x, y) analysis, Marzban and Sandgathe also examine some $p = 3$ dimensional examples - an $(x, y, log(\text{precipitation}))$ analysis (Marzban and Sandgathe 2006), and a $(x, y, \text{reflectivity})$ analysis (Marzban and Sandgathe 2007).

When cluster analysis is employed for verification, the virtue of the hierarchical approach is that it allows one to examine the quality of a forecast field on different spatial scales, i.e., number of clusters. It is worth noting that in an object-oriented framework, the number of clusters in a field is a more appropriate notion of scale than one based on a physical notion of distance. This is so because different clusters/objects in a given field (forecast or observed) may be wildly different in terms of their size. Furthermore, as mentioned above, cluster analysis may be performed in a space that includes non-spatial variables (e.g., reflectivity). In such cases, it would simply make no sense to employ a quantity based on length alone to explore different spatial

scales. As such, the number of clusters, NC, is a generalized notion of scale, beyond spatial scale. In short, in hierarchical clustering (agglomerative or not) one can assess performance as a function of the number of clusters in a field, with the latter (inversely) related to scale. Therefore, henceforth, "scale" refers to the number of clusters, NC.

In contrast to hierarchical clustering, there exists a class of clustering methods generally called k-means (Everitt 1980). The main advantage of k-means is speed. It is generally much faster than a hierarchical method, because it assumes that the data consist of precisely k clusters, and coerces the cases to fall into those clusters. Although its general disadvantage is that it can be somewhat sensitive to initial choice of the k clusters, this is generally not a concern in large data sets. In the verification scheme, k-means does have another disadvantage: it does not allow an exploration of different scales, because the number of clusters k is fixed. The idea of running multiple k-means clusterings with k itself varying from N, the sample size, to 1, turns out to be unfeasible. However, k-means can serve as an initial clustering method, to reduce the number of clusters from N to some reasonable number (say 100), before hierarchical clustering is employed to perform the remainder of the clustering from 100 clusters down to 1. Indeed, this is one of the steps taken in this work to expedite CCA. Other steps for expediting CCA are described below. The relatively slow performance of CCA allowed for the verification of only a handful of forecast days in Marzban and Sandgathe (2006, 2007). A fast CCA opens the possibility of applying the methodology to a large number of forecasts, and thereby comparing different

forecast models in an objective and statistically reliable way.

The main goal of the current work is to apply CCA to the verification of reflectivity forecasts for 32 days, from two high resolution versions of the Weather Research and Forecast model (WRF), and the NOAA Mesoscale Model (NMM). To that end, several methodological revisions to CCA are introduced and described. The three models are compared in terms of their CSI values as computed by the revised CCA in (x, y). The results are compared with those based on an expert's assessment of the forecasts. An attempt is also made to assess errors in the timing of events. Additionally, a $p = 3$ dimensional analysis is performed, and some technical issues in that work are addressed.

2 The Data and Method

The data set consists of pairs of 32 days of observations and 24hr-forecasts of reflectivity exceeding 20 dbz. This corresponds to the forecast of light to heavy precipitation. The 32 days span the dates April 19 to June 4, 2005, and the grid spacing is 4.7625 km. Figure 1 displays the observations and the 24hr forecasts according to the University of Oklahoma $2km$ resolution WRF (arw2), the NCAR $4km$ resolution WRF (arw4), and the National Weather Service $4km$ NMM (nmm4), for May 13, 2005. This forecast is one of the 32 forecasts which will be examined in this paper. The coordinates of the four corners of the region are 70W/30N, 93W/27N, 67W/48N,

101W/44N, covering the US, East of the Mississippi.[2] The data come from the 2005 Spring Experiment, described in (Baldwin and Elmore 2005; Kain et al. 2005; 2006).

Although CCA is thoroughly discussed in Marzban and Sandgathe (2007), it is reviewed here briefly. CCA amounts to performing HAC on the combined set of a forecast and observation field. CCA has several parameters, one of which is referred to as the "hit threshold." It determines whether a given cluster should be considered a hit, a miss, or a false alarm. The hit threshold pertains to the proportion of grid points in the cluster that belong to the observed field. For example, a hit threshold of 0.1 means that if less than 10% of a cluster is composed of observed points, then it will be classified as a false alarm. Also, if less than 10% of a cluster consists of forecast points, then it is classified as a miss. Otherwise, the cluster is identified as a hit. Although different hit thresholds affect the overall evaluation of the forecasts, in the context of model comparison (i.e., the goal of this work), the choice of the hit threshold is unimportant. Several different thresholds were examined and it was determined that, within a reasonable range of values, the choice of the threshold did not vary the assessed relative performance of the three models. For that reason, the bulk of the analysis here is performed at a hit threshold of 0.1.

As mentioned previously, the typical HAC approach has two draw-backs: it is prohibitively slow for efficient processing of multiple, large fields, as well as needlessly exhaustive in the number of clusters produced by the procedure. The revisions to CCA discussed here involve combining multiple clustering techniques, and sampling

[2]We are grateful to Michael Baldwin for providing the data for this analysis.

approaches to produce large improvements in computational efficiency. An algorithmic sketch of this methodology is as follows:

1) Only grid points whose reflectivity exceeds 20dbz are kept for analysis. In terms of the implied precipitation, this means that the verification is performed on light to heavy precipitation.

2) The clustering is performed on $p = 2$ spatial coordinates, (x, y). As such, the comparison of the three models is done in terms of the quality of their spatial content. A $p = 3$ example is also performed for comparison, The relative weight of the three coordinates, i.e., the metric, is discussed.

3) The two coordinates (x, y) are standardized by subtracting the respective mean, and dividing by the pooled (across observation and forecast) standard deviation.[3] In the $p = 3$ dimensional analysis, a transformation is applied in order to assure all three coordinates are on the same footing. This issue is further addressed in section 6.

4) k-means clustering is performed on the combined data set, at a specified number of clusters, $k = 100$. This step clusters the data into 100 clusters much faster than hierarchical clustering can. Although the 100 resulting clusters are somewhat sensitive to the choice of the initial clusters, the differences are thought to be unimportant, because the 100 clusters will be furthered clustered by the hierarchical approach (step

[3]Note: This is different from the standardization adopted by Marzban and Sandgathe (2006, 2007), where unpooled standard deviations are employed. In fact, in the $p = 2$ dimensional case, standardization with the pooled standard deviation is unnecessary, since the x and y coordinates are already on the same scale.

7, below).

5) The procedure is further expedited by performing the analysis on a sample of size n taken (with replacement) from each of the k clusters.

6) A final step is taken in order to improve computational efficiency. Some details of this step are presented in the appendix. Briefly, instead of performing CCA on N p-dimensional vectors, it is performed on k $(n * p)$-dimensional vectors.

7) CCA is performed, and CSI curves (Marzban and Sandgathe 2007) are produced. These are "curves" that display CSI values as a function of NC, the number of clusters in a field.

8) Steps 5 through 7 are repeated many times (101) to assess the influence of sampling on CSI curves. This type of resampling is often called bootstrapping in statistics. Only the average (over the 101 samples) of the CSI curves is computed for comparing the forecasts of the three models.

9) To assess timing errors, the entire procedure is applied to observations and forecasts with a time lag between them. The time lag values examined here are -3hrs to +3 hrs. The introduction of a time-lag calls for a generalization of CSI curves to CSI surfaces (i.e., CSI as a function of NC and time-lag).[4]

10) All of the ensuing results are compared with a human expert's assessment of the quality of the forecasts. This is not an exact science as only one trained forecaster

[4]These CSI surfaces are different from those of Marzban and Sandgathe (2006); there, CSI is plotted as a function of two NC's, one for each of the fields.

is considered. Furthermore, the forecasts are very complex fields, and so, the expert's assessments are only qualitative.

Elaboration of some of the above steps follows:

Observation and forecast fields consist of a 501×601 grid of reflectivity values. This is the common grid for the forecast and observation fields. For the data at hand, there are approximately 300,000 points (i.e., grid points) with non-zero (in dBz) reflectivity in each field. That number is reduced to about 10,000 for reflectivity exceeding 20dbz.

In practice, there is little interest in every possible cluster number, especially when the number of clusters is comparable to the number of points in the fields. Therefore, k-means clustering with $k \sim 100$ serves as a reasonable and efficient starting point for clustering to fewer clusters. k-means does not produce the same clusters as a hierarchical technique, but it can be thought of as a technique with which to dynamically reduce the resolution of the field based on objects (as opposed to nearby points).

Step 6, above, alludes to a transformation of data for the purpose of improving computational efficiency. Although computational efficiency is one of the reasons for the transformation, the primary reason is more technical. The k-means clustering produces cluster assignments for each point for the k clusters. However, HAC does not allow clustering to begin at some arbitrary initial clustering, because it requires a dissimilarity metric for every pair of points (i.e., a matrix of distances between every

pair). As such, to take advantage of the cluster initialization from k-means, n points are selected from each cluster, and are "transposed" to a $(n * p)$-dimensional vector. Thus, instead of performing HAC on N p-dimensional vectors, this transposition allows one to perform HAC on k $(n * p)$-dimensional vectors. The transposition is described further in the appendix. Here $n = 25$ points are sampled randomly (with replacement) from each of the $k = 100$ clusters.

For large N, performing HAC on p-dimensional vectors would be computationally infeasible; but the transposition yields $k \ll N$ cases, with each case being a vector of dimension $n * p$. It turns out that the computational efficiency of HAC is not particularly sensitive to the dimension of the points over which clustering is performed, as the dissimilarity between pairs of points is computed just once and the bulk of the processing typically occurs in the repeated search to select the next optimal combination of clusters.

3 Results

Armed with CSI values for the three models over 32 days, and over a range of scales, a number of comparisons can be made. The most relevant ones can be divided into two classes - one class where the models are compared in terms of their actual CSI values, and another where the models are ranked in terms of their CSI, and then compared based on their rankings. The conclusions from the two classes are somewhat different, because they address different facets of model performance.

3.1 Comparisons of CSI Values

Figure 2 shows the mean (over 101 samples) CSI curves for all 32 days, for the three models; arw2, arw4, and nmm4, colored black, red, and blue, respectively. Let us begin with a discussion of CSI curves for a specific date: May 13, whose fields are shown in Figure 1..[5] Not surprisingly, there is a significant amount of overlap between some of the models and for some values of NC. In other words, on this particular day the models appear to perform comparably. However, one may note that nmm4 does appear to produce systematically low CSI values across the full range of NC values, suggesting that it is the worse of the three. On larger scales ($10 < NC < 40$), arw2 (black) appears to be the best of the three, while on smaller scales ($60 < NC < 100$) that status is occupied by arw4 (red). However, these conclusions should be qualified: First, they pertain to the forecasts on a single day, and second, they ignore sampling variations. The first limitation is overcome here, by examining such CSI curves for forecasts made on 32 days. As for sampling variations: it turns out that a larger portion of it can be attributed to between-days variations than within-days variations. Therefore, to decide which of the models (if any) is best, one must examine all 32 days.

In order to get a sense of the numerical scale of these CSI values, it is beneficial to compute them for a "random" forecast. However, it is important to assure that the random field has a spatial structure similar to actual forecasts. In other words, a random field consisting of white noise would be inappropriate. The number of points

[5]Hereafter, "CSI curves" refers to the curve resulting from averaging over $n = 101$ samples.

in the random forecast field must also be comparable to that of actual forecast fields. Such a random field is shown in Figure 1.[6]

Note that it does visually resemble real forecasts. For each of the 32 forecasts, a similar random field is produced. The corresponding CSI curves are shown as the green, dashed lines in Figure 2. Clearly, the CSI curves for the real forecasts are generally higher than those of a random forecast, certainly for cluster numbers exceeding 10 or so. For smaller cluster numbers (i.e., on larger scales), real forecasts and the random forecast are indistinguishable in terms of the resulting CSI curves.[7]

One can now address the question of how the three models perform over all 32 days. The sampling variations (over the 101 samples) are not shown for clarity, but are implicitly taken into account in the following observations. There is only

[6]What is a random Gaussian field? First, consider a sequence of random numbers, and note that not all random numbers are the same. For example, the numbers may be uniformly distributed over some interval, or be a sample drawn from a Gaussian with a specified mean and variance, etc. Now, moving to the 2-dimensional case, it is relatively easy to show that if one organizes a sequence of $n \times m$ uniformly distributed random numbers, then the resulting field (or "image") will appear also random in a uniform fashion. Generating a random field which has some nontrivial spatial structure calls for specifying something about that structure. One way is to assume that the numbers are distributed according to a bivariate Gaussian with specified means and covariance matrix. In the field of spatial statistics, there are families of popular covariance structures, and methods for simulating them. The one employed in this work is the so-called "stable family" whose simulation is described in Gneiting, et al. 2005.

[7]Recall that for $NC = 1$ one would include all valid points in both the observed and forecast fields. Since our "random" field is distributed across the entire domain, it scores well for the May 13 case where a significant portion of the domain has observed precipitation.

one day for which the CSI curves for the three models have little overlap across the full range of scales, namely April, 27. In that case, the ranking of the models, in order of decreasing performance is arw2, arw4, and nmm4. This ranking, although physically arguable, is clearly not generally true. On most days (22) arw2 and arw4 are comparable across the full range of scales, as expected for nearly identical models executing at only slightly different resolution; meanwhile, nmm4 is equally likely to be better or worse. On a few days the models actually switch rank at some scale. For example, on 5/3, nmm4 outperforms the other two for cluster numbers below 30; but on smaller scales immediately above $NC = 30$, arw2 appears to be the best of the three. Finally, on some days (e.g., 5/27) the models are not only comparable to each other but also comparable to a random forecast.

Although the CSI curves in Figure 2 speak more to the complex relationship between performance and spatial scale, it is possible to summarize the results in a way that is conducive to a comparison of the three models "averaged" over the 32 days. One way is to compare the three models two at a time. Figure 3 shows the boxplot of the difference between the CSI curves of the models in a pair-wise fashion. Each panel, therefore, compares two of the models.[8]

[8]For technical reasons, it is difficult to assess the statistical significance of these differences. For example, a paired t-test may seem appropriate, but then the issue of multiple testing (i.e., for 100 NC values) becomes a thorny issue. One may address that concern via a Bonferroni adjustment, but such a correction ignores the clear dependence of CSI as a function of NC. For such reasons, a rigorous statistical test of the hypothesis of equal-means is not performed here. The boxplots, however, do provide a useful and visual tool for qualitatively assessing the difference between the

Evidently, all of the boxplots cover the horizontal axis at 0. In other words, there does not appear to be a strong statistically significant difference between the three models. A more rigorous comparison could be performed using a multivariate t-test, but some of the assumptions of that test are violated here. For that reason, only graphic means, such as the boxplots in Figure 3 are examined in this study.

If one were to relax the requirement of statistical significance, the boxplots in Figure 3 suggest some tentative conclusions. The conclusions are better organized if they pertain to three distinct scales, broadly defined as $NC < 20$, $20 < NC < 60$, and $NC > 60$. Consider the middle range, first: The left panel in Figure 3 implies that CSI values for arw2 are generally higher than those of arw4. In fact, these differences appear to be statistically significant, in the sense that zero (i.e., the horizontal line) is just outside of the interquartile range of the boxplots. The middle panel suggests that arw4 is generally worse than nmm4, and the right panel implies that arw2 is generally better than nmm4. In short, across the 32 days, CSI values of the three models, for NC values between 20 and 60, can be ordered as CSI(arw2) > CSI(nmm4) > CSI(arw4).

For $NC < 20$, a similar analysis of Figure 3 implies that arw2 and arw4 are comparable, and both are superior to nmm4; i.e., CSI(arw2) $\sim$ CSI(arw4) > CSI(nmm4). And for $NC > 60$, arw2 and nmm4 are comparable, with both superior to arw4; i.e., CSI(arw2) $\sim$ CSI(nmm4) > CSI(arw4).

In short, on larger scales ($NC < 20$), arw2 and arw4 are comparable, with both

means.

being superior to nmm4. On mid-range scales ($20 < NC < 60$), arw2 outperforms nmm4, which in turn is superior to arw4. Finally, on smaller scales ($NC > 60$), arw2 and nmm4 are comparable, with both being better than arw4.

One can make an educated guess at the meaning of these results; although the model developers themselves will likely have a better interpretation. Starting with arw2 and arw4, essentially the same model at different resolutions, the higher resolution of arw2 is less important in comparison to arw4 at larger scales; therefore, they perform similarly. At smaller scales, however, higher resolution has a greater impact on model skill, allowing arw2 to score higher; i.e., resolve and predict smaller features more reliably than arw4. When considering arw (2 or 4) versus nmm4, the results reveal the affect of both resolution and different model numerics and physics. The CSI data indicates that the model formulation in nmm4 is better at smaller scales than in arw; arw2 (higher resolution than nmm4) performs comparably to nmm4, and arw4 (same resolution) performs worse than nmm4. However, at larger scales, the arw model seems to have the edge over nmm4 (i.e., at $NC \sim 20$, arw4 is better than nmm4 even though both have the same $4km$ resolution.)

Of course, many caveats apply to these conclusions: The entire data is restricted to only 32 days, in one Spring, of a single year over a region representing predominately the midwest, from the Gulf to the Canadian border. Moreover, both the arw and nmm models have undergone many upgrades since the Spring 2005 experiment. What is being demonstrated here is the ability to derive meaningful conclusions from an automated analysis of highly complex, very high resolution, weather predictions.

3.2 Comparisons of CSI-based Ranks

The above tests compare the three models in terms of their CSI values. It is also possible to compare them in terms of their rank - 1, 2, or 3. The following lead-in questions set the stage for the analysis. In how many of the 32 days, does CSI (at, say $NC = 20$) suggest that arw2 is the best (i.e., rank=1) of the three models? In how many, does CSI suggest that arw2 has rank=2; and rank=3? Similarly, what are the answers to these questions if they pertain to arw4, and nmm4? The results can be tabulated as a contingency table, with the rows representing the three models, and the columns denoting the rank. For NC=20 and 60, the results are shown in Table 1. The choice of $NC = 20$ and 60 is based partly on meteorological considerations, and so, is explained in section 5. Note that in rankings where a "tie" occurs, the models are given the same score, i.e., if there is a tie for first place, there will be two ranks of 1 and one rank of two. If there is a tie for second place, there will be one rank of 1 and two ranks of 2; hence, there is a preponderance of rank=2 and significantly fewer rank=3.

Given that these contingency tables address the association between the choice of the model and a CSI-based rank, one can perform a chi-squared test of that association. For the $NC = 20$ and $NC = 60$ contingency tables, the p-values of the test are 0.03, and 0.01, respectively. Therefore, at a significance level of 0.05, both of these p-values are significant, implying that there is statistically significant association between the choice of the model and a CSI-based rank. In short, a knowledge of CSI can generally predict the rank of a model, and vice versa. Said differently, CSI curves

can distinguish between the ranking of the models, at a statistically significant level.[9]

Since the associations are statistically significant, one may further diagnose these tables.[10] For example, it is clear that arw2 rarely obtains rank=3, for either $NC = 20$ or 60; arw2 is also approximately equally likely to obtain a rank of 1 or 2. arw4 obtains a rank of 2 on the majority of days, again for both NC values. nmm4 follows a slightly different pattern: at larger scales ($NC = 20$), it ranks second most frequently; but on smaller scales ($NC = 60$), its rank is a tie between 1st and 2nd.

To summarize these observations: For $NC = 20$, both arw4 and nmm4 rank 2nd, but arw2 ranks equally between 1 and 2. In this sense, one may conclude that arw2 is the better of the three at larger scales, as noted previously. On smaller scales ($NC = 60$), arw2 can still be considered the best of the three models, of course aided by its higher resolution; however, nmm4 follows closely, implying that a $2km$ resolution version of nmm may score better than arw2.

[9]Two other but equivalent conclusions are as follows: 1) The three models are not homogeneous with respect to their rank; 2) for at least one of the three ranks, the proportion of the three ranks (for each model) are not identical for the three models.

[10]Technically, one should convert all of the cell counts into row proportions. One can then compare the proportion of rank=1 cases in arw2, with that in arw4, etc. However, given that the row-marginals of these tables are equal (i.e., 32), one can compare the actual counts.

4 Timing Error

By virtue of clustering in (x, y) space, the above analysis implicitly takes into account the spatial error of the forecasts. To account for errors in the magnitude of reflectivity as well, one may perform cluster analysis in (x, y, z) space, where z =reflectivity; see section 6. But what about timing errors? Although one can set-up a frame for clustering in (x, y, t), or even in (x, y, z, t), it is more instructive to perform a series of (x, y) analyses for observations and forecasts that differ in their valid time. Here, for each observation day, in addition to a comparison with the "default" forecast for that hour, 12 additional forecasts are also examined at hourly lags from -6hr to +6hr. Again, specializing to a given day, the observations and the forecasts for the three models are shown in Figure 4, but only for hourly lags from -3hr to +3hr (from bottom to top) to economize on figures.

Visually detecting a time lag in these very complex weather patterns is difficult; however, a visual inspection of the forecasts in Figure 4 suggests that the nmm4 forecast is noticeably different, at least for values above 20dBz, from arw2 and arw4. Looking specifically at the development in nmm4, some of the hourly predictions appear to match better with a previous observation ($\sim$ -3hr) than the verifying observation.

This type of timing-error can be quantified in the current verification scheme. For each of the 32 days, it is possible to produce a plot of CSI as a function of the number of clusters as well as the time lag. Figure 5 shows a sample of two such dates: May 13

and April 23, 2005. These two dates are selected for display, because they show two different patterns of time lags. Consider May 13 first: On large scales (small NC), arw2 and arw4 have their highest CSI values at lag$\sim$ 0; in other words, on large scales these models are neither fast nor slow. On smaller scales (large NC), however, they both have their highest CSI values at positive lag ($\sim 2hr$), implying that they are slow on small scales. nmm4 shows a slight peak-CSI at a negative lag ($\sim -2hr$) across all scales. This means that on this day nmm4 is too fast on all scales.[11] By contrast, on April 23, all three models are too fast because their highest CSI values occur at negative lags, across the full range of scales. It is important to point out that these two dates are atypical cases, and are discussed here only to illustrate the diagnostic nature of the verification scheme. A visual inspection of all 32 plots (not shown) suggests that the three models are "on time" on the average.

5 Expert Opinion

As mentioned previously, the main aim of the verification methodology discussed here is to allow for the automatic verification of large number of forecasts. The question remains as to whether or not the results of the verification agree with human/expert assessment. To that end, three tests are performed.

[11] The lags reflect difference in the observation time rather than the forecast, i.e. the forecast time is fixed. A peak at a negative lag is, therefore, a peak at an earlier observation, i.e., the forecast is fast. Similarly, a positive lag implies a slow forecast.

The first test involves an expert's (Sandgathe) visual assessment of the overall quality of the three forecasts combined. This is possible because the three forecast models' predictions are quite similar to each other on each day; in fact, they are closer to one another than to the corresponding observations. (At this point, there is a desire to recommend more research in model physics, i.e., all models appear to be similarly lacking; however, that point will be reserved for a different forum.) Based on a visual inspection of the three forecasts (e.g., Figure 1), a subjective score (VIS) ranging from 0 to 10, representing poor to excellent forecasts, respectively, is assigned to each day. Given that CSI is the objective measure of model/forecast quality adopted in this study, the question is if VIS is correlated with CSI across the 32 days. It is important to point out that the VIS scores are assigned *prior* to the expert's viewing of the CSI scores. The results are shown in Figure 6, for three values of hit threshold (0.01, 0.1, 0.2), and two values of NC: 20 and 60. $NC = 20$ is selected because the fields appear visually to have 2 to 7 major "systems" and 15 to 30 major clusters on each day. As such, 20 is a physically meaningful cluster number. $NC = 60$ is selected because it corresponds to where the CSI curves (e.g., Figure 2) become constant. Interestingly, $NC = 20$ and 60 are also suggested by the boxplots shown in Figure 3. The correlation coefficient, r, between VIS and CSI is also shown.

It can be seen, that CSI and VIS are generally well-correlated, with r values in the 0.7-0.8 range. These scatterplots confirm that the relation between CSI and VIS is generally linear, regardless of hit threshold and NC. As such, it follows that the verification method described here yields CSI values consistent with a human/expert's

assessment of forecast quality. This conclusion is intended to be only qualitative in nature. The forecasts cover a large area involving multiple weather systems and tradeoffs between "good" forecast for one system and a "poor" forecast for another system on a given day are very subjective. A more rigorous study would require multiple experts, careful attention to inter-operator variability, reducing the forecast area to smaller regions, and a more thorough definition of forecast quality.

The above test compares an expert's assessment of the forecasts of the general skill of the three models on any given day. In other words, it answers the question of how well the models perform in general. The second test is more difficult visually and attempts to verify whether or not CSI is sufficiently sensitive to model skill to accurately rank the model forecasts on a given day.[12] To that end, the three model forecasts are ranked visually for each of the 32 days. Again, ties are awarded the same ranking, resulting in scores skewed toward lower numbers (higher rank). The data are summarized in Table 2, and a chi-squared test of the association between model and visual rank yields a p-value of 0.04. So, at a significance level of 0.05 one can conclude that visually ranking the models distinguishes between the models at a statistically significant level. Inspection of Table 2 indicates again that the models should be ranked, from best to worst, as arw2, nmm4, arw4, in general agreement with Table 1 as discussed in section 3.2, with the caveat that in this visual test, the

[12]Ranking the actual forecast fields visually is much more difficult than ranking the models based on CSI, as a model may do well in one geographic region of the field, or on one weather complex, while doing poorly in another. For this reason, the following analysis should be considered qualitative, in spite of the appearance of statistical significance tests.

number of clusters is ignored. Note that this particular ranking of the three models is the same as that based on the values of CSI itself, in mid-range NC values, as discussed in section 3.1.

The last test or comparison is to ensure "clear winners" and "clear losers" are identified by the CCA methodology. In this instance, only those cases where either the visual inspection indicates a clear best or worst model, or the CSI gives a clear indication of a best or worst model, are compared. Of the twenty two cases so indicated, there is agreement on ten, disagreement on two, and either the visual or CSI do not indicate a clear (for CSI, statistically significant, for visual, too difficult to assess) winner on the remaining ten. Again, in order to perform this test more accurately, the region should be subdivided into smaller regions and additional experts should be employed to visually verify the fields.

6 A 3d Example

The above CCA is performed in a $p = 2$ dimensional space, but it is instructive to examine at least one $p = 3$ dimensional space because it illuminates some important issues. On the one hand, by performing the analysis in (x, y, z), one expects the resulting clusters (at each iteration of HAC) to be more physically sensible than in an analysis performed in (x, y). That is the main reason for even pursuing this methodology in a higher-dimensional space. On the other hand, there are (at least) two complexities. First, the resulting clusters *when viewed in* (x, y) may appear to be

unphysical. For instance, two spatially adjacent clusters may be labeled as distinct, because they differ in terms of their non-spatial coordinate. Although, this may create some problems in a visual assessment of the clusters, in truth the analysis is more faithful than an analysis in (x, y), because it relies on more information (even if that information may not be easily displayed in a 2d plot of the clusters).

The second complexity with performing the analysis in a higher-dimensional space is that it raises a question which is difficult to answer objectively. Specifically, what metric should be used to compute the distance between two points in a p-dimensional space? In other words, how should the non-spatial components be weighted against the spatial components? Said differently, what metric should be used for computing distances? One way to address this issue is to introduce a "knob" controlling the strength of each non-spatial component relative to the spatial ones. This solution may be feasible if there is only one non-spatial variable, e.g., reflectivity; however, it becomes impractical for large values of p.

Alternatively, one may appeal to some other criteria for choosing the metric; for example, one based on tolerance. One expert or a consensus of experts might advise "tolerances" for forecast errors for spatial and non-spatial coordinates. These tolerances may be used to standardize the different coordinates to a common tolerance scale. For example, one might tolerate a forecast error of $20km$. As for errors in reflectivity itself, one might tolerate a forecast error of one quarter of the difference in a reflectivity of 50 dbz (very heavy rain) and 20 dbz (very light rain). Then distance errors and reflectivity errors may be converted to "tolerance units" by dividing the

spatial component of the distance by $20km$, and the reflectivity component by $7.5 = (50dBz - 20dBz)/4$. This is one of many possible criteria.

Figure 7 shows the CSI surfaces for three analyses of the May 13 data. The first one (top) is a $p = 2$-dimensional analysis in (x, y); it is Figure 5 reproduced here for easier comparison). The other two are $p = 3$ dimensional analyses in $(x, y, \log(z))$, where z denotes reflectivity; in one case (middle plot) the three components are standardized individually (i.e., they are weighted equally), and in the other case (bottom plot) the metric is based on the aforementioned tolerance considerations. In going from the $p = 2$ (top panel) to the $p = 3$ analysis with equal weight given to all the components (middle panel), it is evident that arw2 becomes more sensitive to the lag. This conclusion is based on the "crest" appearing in the middle panel of the arw2 column in Figure 7. Another notable difference is that the above-mentioned "slowness" of arw4 is much less pronounced. Also, nmm4 is not affected by the inclusion of reflectivity in the analysis; it appears to be fast on that date.

Most of these features disappear when a tolerance-based metric is employed for computing distances. Indeed, the resulting figures (bottom panel in Figure 7) resemble those of the (x, y) analysis (top panel). It follows that the inclusion of reflectivity in the analysis does not drastically affect the assessment of performance; the three models are comparable in terms of the amount of reflectivity assigned to the various clusters, at least within the range of tolerances injected into the distance metric.

7 Summary and Discussion

This is arguably a small sample for any definitive conclusions regarding the comparison of the three model formulations. Moreover, the three model formulations have changed significantly since these cases were forecast in Spring of 2005. But the data do appear to provide a valid test of the CCA methodology, and the methodology does appear to provide insight into the model formulations. The CSI curves, not surprisingly, indicate little difference between the various model formulations. However, there appears to be a tendency for arw2 to outperform nmm4, which in turn outperforms arw4, in the 20 to 60 cluster range, the range that is likely the most significant for the cases evaluated here. The evaluated arw4 formulation performs statistically worse in that range as born out also by both visual and rank-based evaluations. Although the large, complex region under evaluation makes visual ranking difficult, comparison of visual rankings and CCA CSI scores indicates that CCA is faithfully capturing model performance.

A methodological contribution of the current study is the advent of the transposed HAC (section 2, and appendix). Without it, it would be simply impossible to analyze a large number of forecasts in any reasonable length of time. In order to assess the efficiency of the transposed HAC method, a few benchmark tests have been performed to compare transposed HAC with traditional HAC. The comparison was performed on a field containing about 20,000 points (even after implementing the 20dBz threshold). The key parameters of the method were set at $k = 100$ (used for k-means), $n = 25$

(the number of sample points), and 101 CSI re-samples (i.e., yielding 101 different CSI curves to be averaged). To assess the speed of the two methods the number of points to be clustered was varied from 200 to 4,000 for both methods, continuing up to 20,000 only for transposed HAC. All tests were performed on a computer using a relatively modest Intel(R) Pentium(R) 3.00 GHz CPU.

The results are striking though not surprising. For transposed HAC, the procedure ran in the range of 30 seconds to 45 seconds depending on the number of points, from 200 to 20,000; this suggests that the timing in this range is perhaps dominated by the CSI re-sampling. For traditional HAC, runtime grows exponentially: for less than 1500 points it ran in less than 30 seconds, growing to about 10 minutes for 4,000 points. For 20,000 points, traditional HAC simply did not execute due to memory limitations on the computer; the matrix of pairwise distances repeatedly scanned in each HAC step has a very large number of elements (i.e., $20,000\ (20,000-1)/2$), which traditional HAC cannot accommodate. The transposed HAC technique introduces vast improvement in computational efficiency making the analysis of large datasets practical.

Acknowledgements

The authors would like to acknowledge Michael 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 by the National Science Foundation

grant number 0513871.

8 Appendix: Transposed Hierarchical Clustering

In this appendix, the method of "transposing" the data will be outlined. Further details will be presented elsewhere[13]

A hierarchical clustering method (HAC) is inherently iterative. The agglomerative version (adopted here) begins by assigning every data case to a cluster, and ends when all the cases fall into a single cluster. In the current application, a "case" refers to the coordinates of a grid point, but as described by Marzban and Sandgathe (2007), the point may reside in a larger-dimensional space. For example, the clustering may be performed in (x, y, z), where x and y denote the cartesian coordinates of a grid point, and z refers to the reflectivity at that point.

The iterative nature of HAC is a desirable feature in the current application, because it allows exploration of the clusters on different scales. However, for data sets involving a large number of cases the procedure is prohibitively slow. To expedite the procedure, a revision is considered: For simplicity, and without loss of generality, consider a 1-dimensional data set whose cases are denoted x_i, with $i = 1, 2, 3, ..., n$. HAC begins with n clusters, and proceeds to identify the nearest pair, which in turn are merged into a new cluster.

[13]Marzban C., H. Lyons, S. Sandgathe, C. Fraley, 2007: A method for expediting hierarchical clustering methods. In preparation.

Now, suppose one were to identify all the nearest pairs, but not combine them into clusters. Instead, suppose the two members of each pair are viewed as coordinates of a new "case" viewed in a 2-dimensional cartesian space. Figure 8 illustrates the idea on a hypothetical data set consisting of six cases, $x = 11, 12, 14, 15, 18, 19$. Let $C_{i,j,k,...}$ denote the cluster with the elements $i, j, k, ...$ HAC begins with the six clusters $C_1 = 11, C_2 = 12, C_3 = 14, ..., C_6 = 19$. Traditional HAC would yield the following sequence of iterations: $C_{1,2}, C_{3,4}, C_{5,6}, C_{1,2,3,4}, C_{1,2,3,4,5,6}$. However, consider a transposition that maps the 1-dimensional data to 2-dimensions according to the scheme displayed in Figure 8. For example, the two cases, $x = 11$ and $x = 12$, get mapped to the single case at $(11, 12)$. In this space, there are only three cases present, and so, the size of the data has been reduced by 50%. HAC on the new data can then proceed in a traditional fashion.

The method is being called "transposed HAC", because with each case being a p-dimensional vector, the original data, which can be viewed as an $n \times p$ matrix is "transposed" into an $\frac{n}{2} \times 2p$ matrix. As another example, consider $n = 6$ and $p = 2$. The data $(x_i, y_i), i = 1, 2, ..., n$, are mapped into 4 dimensions, according to the scheme shown in Figure 8. The only ambiguity in this map has to do with the order of the points. For instance, one might map two nearest neighbors (x_1, y_1) and (x_2, y_2) to the point (x_1, x_2, y_1, y_2), or (x_1, y_1, x_2, y_2), etc. It has been confirmed that the final results of the analysis (e.g., CSI curves) are insensitive to the ordering of the points. This is so, because in the re-sampling phase of the procedure (step 5, section 2), the order of the points changes from sample to sample.

To make contact with the analysis performed in the body of the paper, note that the above-mentioned pairing of the cases is equivalent to performing some type of clustering (e.g., k-means) on the original data.

9 References

Baldwin, M. E., K. L. Elmore, 2005: Objective verification of high-resolution WRF forecasts during 2005 NSSL/SPC Spring Program. Preprints, 21st Conference on Weather Analysis and Forecasting/17th Conference on Numerical Weather Prediction., Washington, DC, USA, American Meteorologocal Society, 11B.4.

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.nsslnoaa.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).

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

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.

Davis, C. A., B. Brown and R. Bullock, 2006a: Object-based verification of precipitation forecasts. Part I: Methodology and application to mesoscale rain areas. *Mon. Wea. Rev.*, **134**, 1772-1784.

Davis, C. A., B. Brown and R. Bullock, 2006b: Object-based verification of precipitation forecasts. Part II: Application to convective rain systems. *Mon. Wea. Rev.*, **134**, 1785-1795.

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

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

Everitt, B. S., 1980: *Cluster Analysis*. Second Edition, Heinemann Educational Books. London.

Gneiting, T., H. Sevcikova, D.B. Percival, M. Schlather, and Y/ Jiang, 2005: Fast and Exact Simulation of Large Gaussian Lattice Systems in R^2; Exploring the limits. University of Washington, Department of Statistics, Technical Report no. 477; available at http://www.stat.washington.edu/www/research/reports/2005/tr477.pdf .

Hoffman, R. N., Z. Liu, J.-F. Louis, and C. Grassotti, 1995: Distortion representation of forecast errors. *Mon. Wea. Rev.*, **123**, 2758-2770.

Kain, J. S., S. J. Weiss, M. E. Baldwin, G. W. Carbin, D. A. Bright, J. J. Levit, and J. A. Hart, 2005: Evaluating high-resolution configurations of the WRF model that are used to forecast severe convective weather: The 2005 SPC/NSSL Spring Program. Preprints, 21th Conference on Weather Analysis and Forecasting/17th Conference on Numerical Weather Prediction, Washington, D. C., Amer. Meteor. Soc., CD-ROM, 2A.5.

Kain, J. S., S. J. Weiss, J. J. Levit, M. E. Baldwin, and D. R. 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.

Marzban C., H. Lyons, S. Sandgathe, and C. Fraley, 2007: A method for expediting hierarchical clustering methods. In preparation.

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

Marzban, C., and S. Sandgathe, 2007: Cluster Analysis for Object-Oriented Verification of Fields: A Variation. Accepted by Monthly Weather Review.

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

Venugopal, V., S. Basu and E. Foufoula-Georgiou, 2005: A new metric for comparing precipitation patterns with an application to ensemble forecasts. *J. Geophys. Res.*, **110:** D8, D08111 10.1029/2004JD005395.

Figure Captions

Figure 1. Top four panels: The Observation field, and the 3 forecast fields according to arw2, arw4, and nmm4 for May 13, 2005. Middle Panels: Same, but for reflectivity exceeding 20dbz. All the analysis in this study is performed on grid points whose reflectivity exceeds 20dbz. Bottom figure: A random forecast field employed for exploring the range of CSI values.

Figure 2. Mean (over 101 samples) CSI curves for all 32 days for arw2 (black), arw4 (red), nmm4 (blue) forecasts. The dashed (green) line corresponds to the CSI curve for random forecasts (e.g., shown in Figure 1). The x-axis denotes the number of clusters, NC.

Figure 3. Boxplots of the difference between CSI curves, computed over 32 days, versus the number of clusters.

Figure 4. From left to right, the columns correspond to the observations, arw2, arw4, and nmm4 forecasts, respectively, on May 13, 2005. From bootom to top the valid times are (approximately) 21Z-27Z.

Figure 5. Levelplots of CSI as a function of the number of clusters and the time lag, for the three models; May 13 (top), April 23 (bottom).

Figure 6. Average (over three models) CSI versus VIS, for threshold = 0.01 (top), 0.1 (middle), and 0.2 (bottom), for NC=20 (left) and NC=60 (right). The corresponding correlation coefficient, r, is also shown.

Figure 7. The csi-surfaces when the analysis is performed in (x, y) space (top), in (x, y, z) space, with each variable contributing equally to distance calculations (middle), and with the three variables weighted differently (bottom); see text.

Figure 8. An illustration of the transposition which maps a clustering of 6 cases in 2 dimensions into a clustering of 3 cases in 4 dimensions.

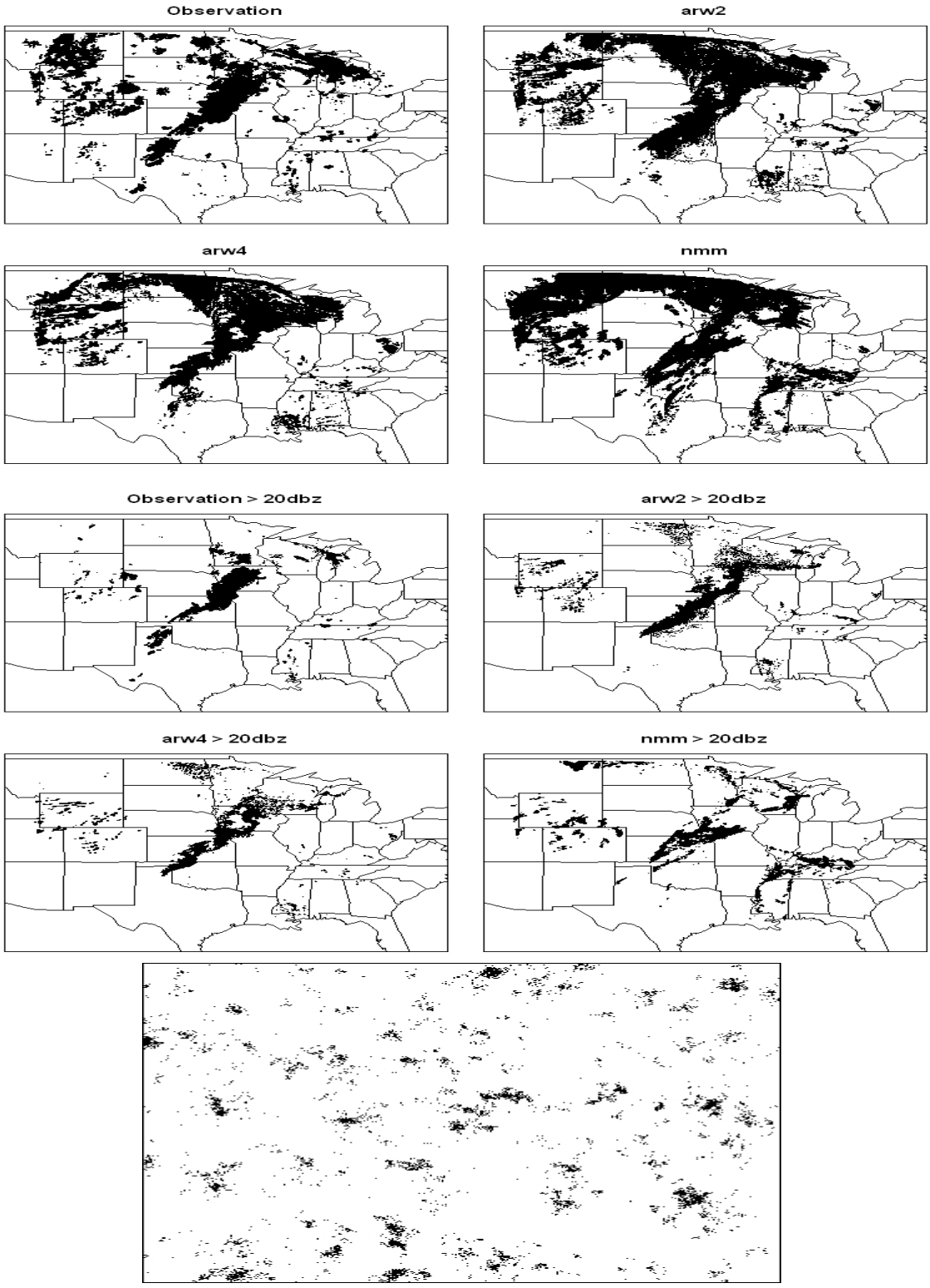

Figure 1. Top four panels: The Observation field, and the 3 forecast fields according to

arw2, arw4, and nmm4 for May 13, 2005. Middle Panels: Same, but for reflectivity exceeding 20dbz. All the analysis in this study is performed on grid points whose reflectivity exceeds 20dbz. Bottom figure: A random forecast field employed for exploring the range of CSI values.

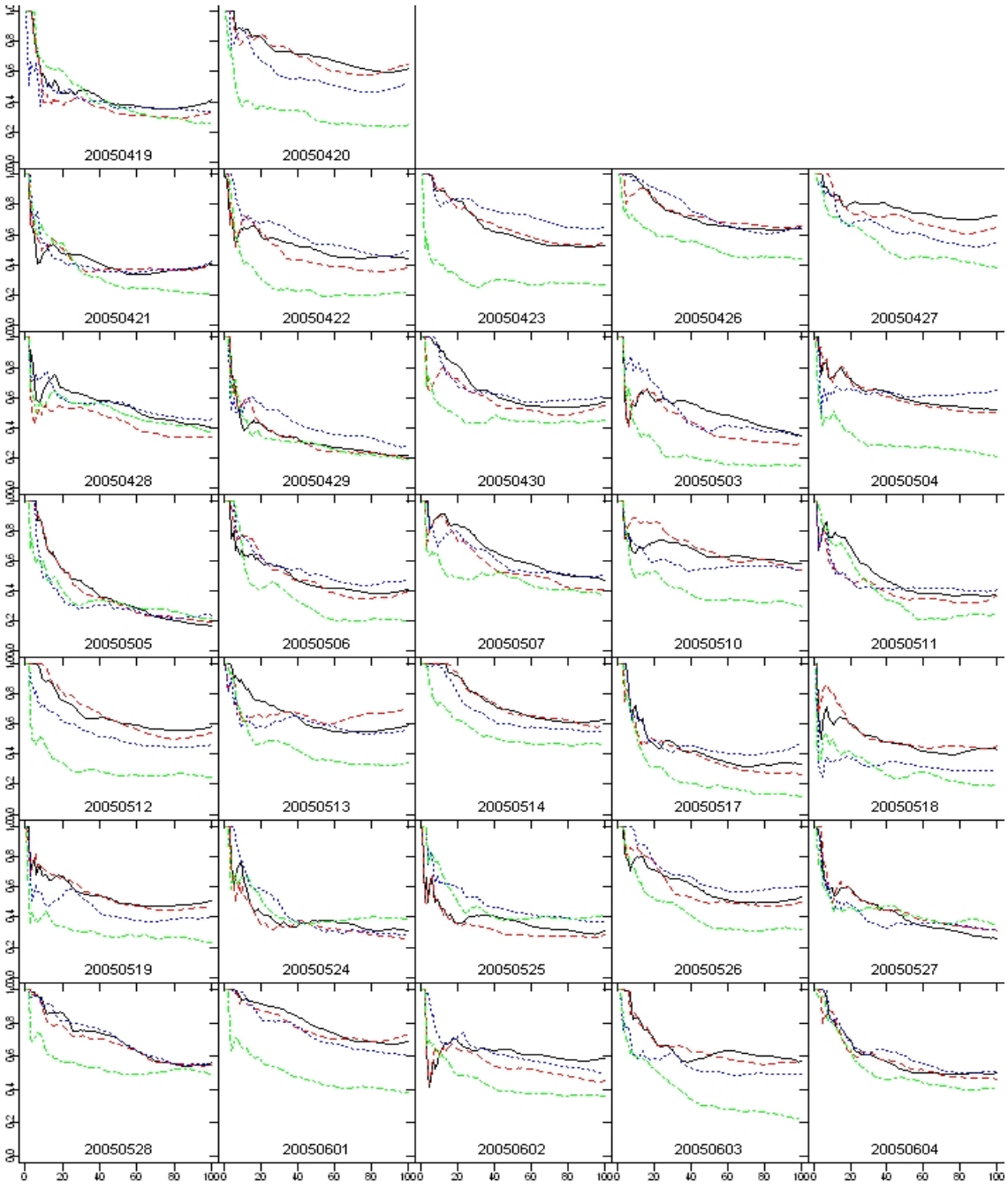

Figure 2. Mean (over 101 samples) CSI curves for all 32 days for arw2 (black), arw4 (red), nmm4 (blue) forecasts. The dashed (green) line corresponds to the CSI curve for random forecasts (e.g., shown in Figure 1). The x-axis denotes the number of clusters, NC.

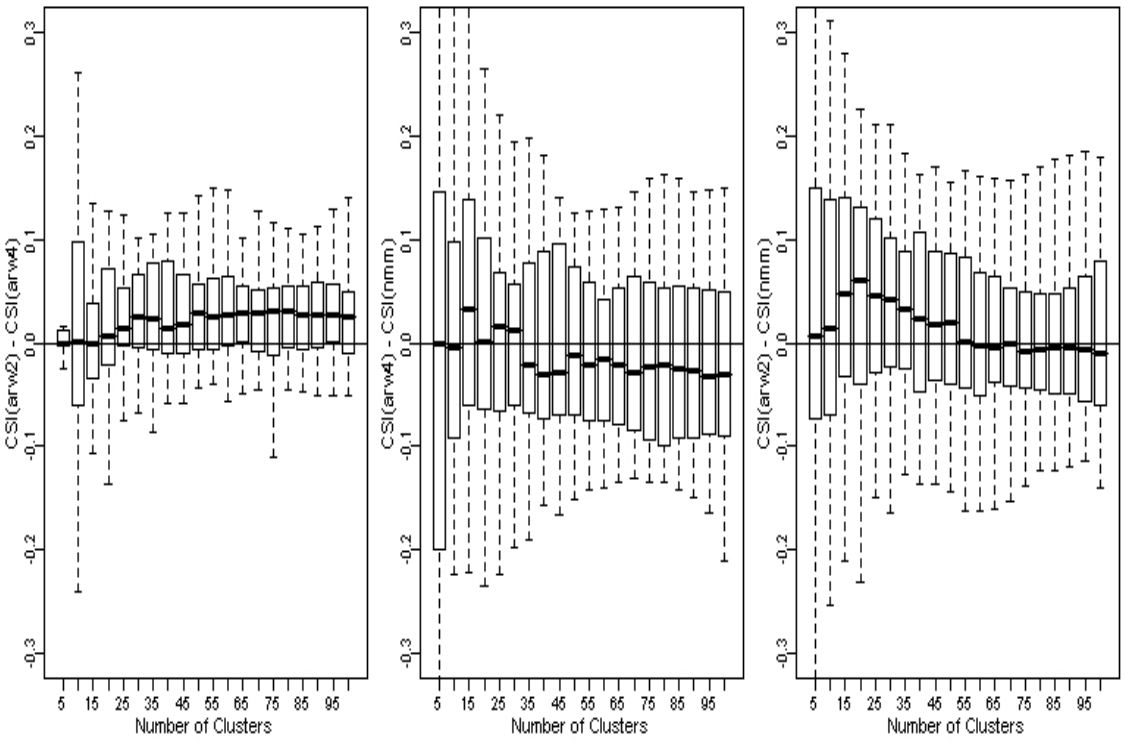

Figure 3. Boxplots of the difference between CSI curves, computed over 32 days, versus the number of clusters.

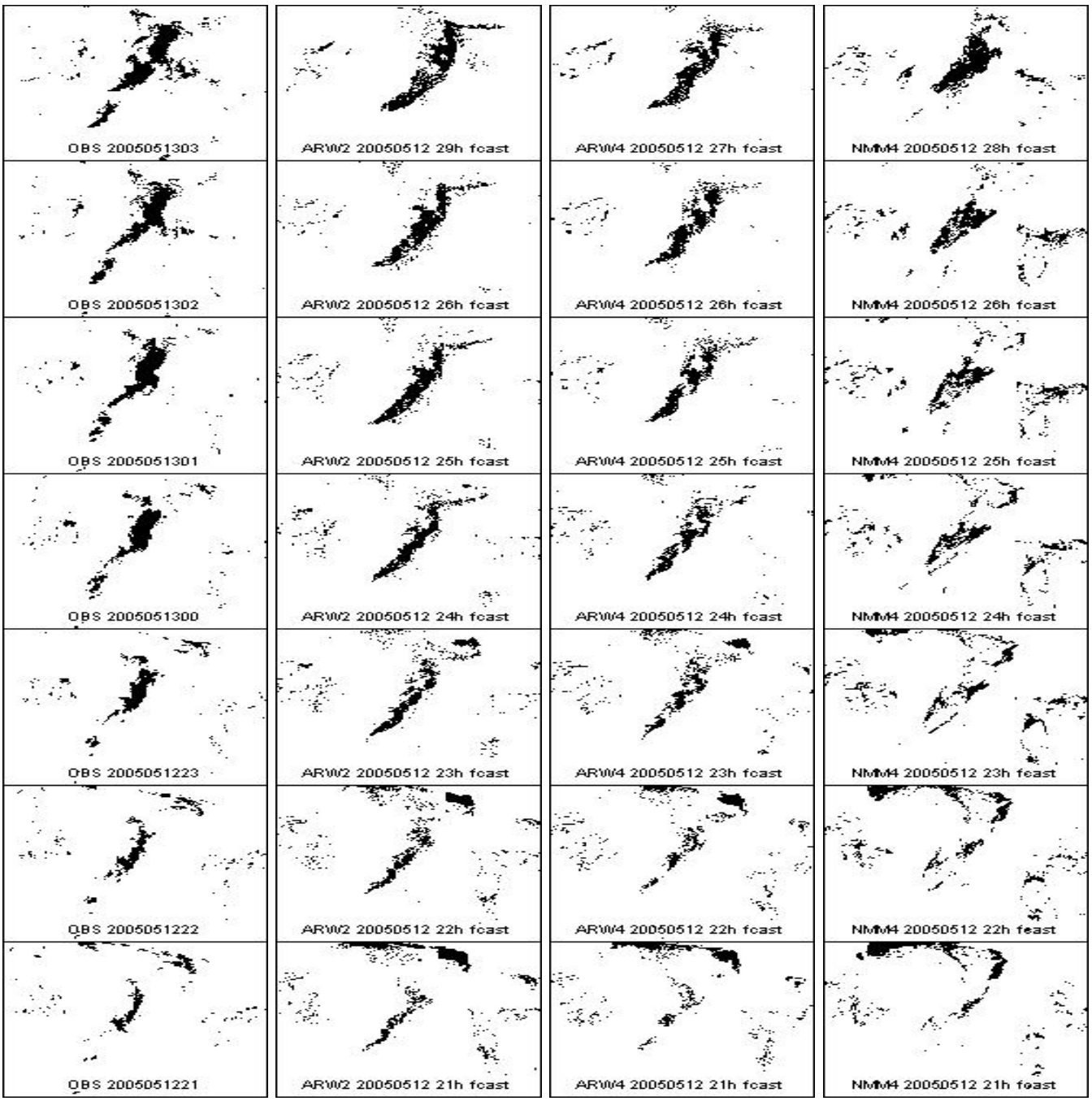

Figure 4. From left to right, the columns correspond to the observations, arw2, arw4, and nmm4 forecasts, respectively, on May 13, 2005. From bottom to top, the valid times are (approximately) 21Z-27Z.

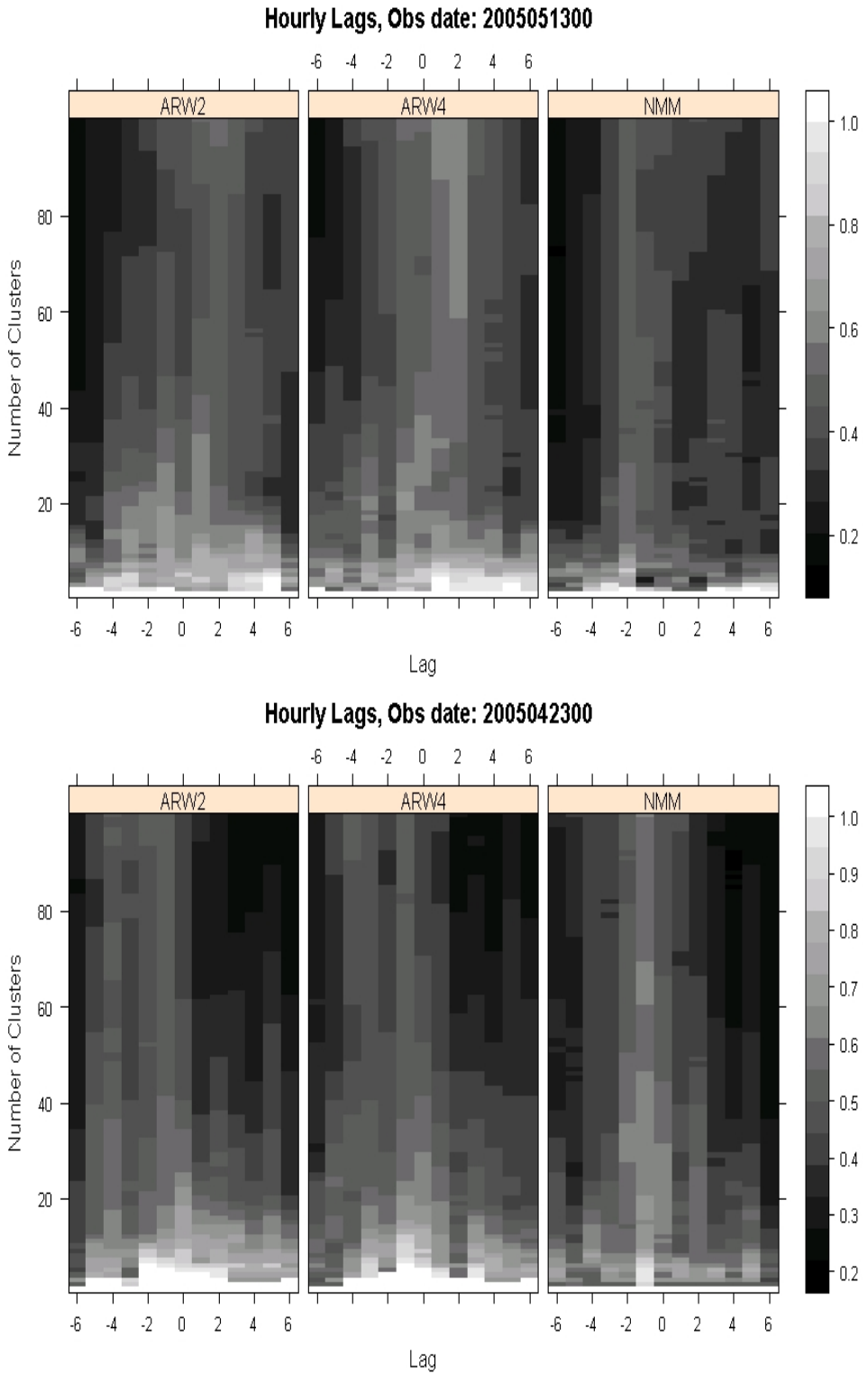

Figure 5. Levelplots of CSI as a function of the number of clusters and the time lag, for the three models; May 13 (top), April 23 (bottom).

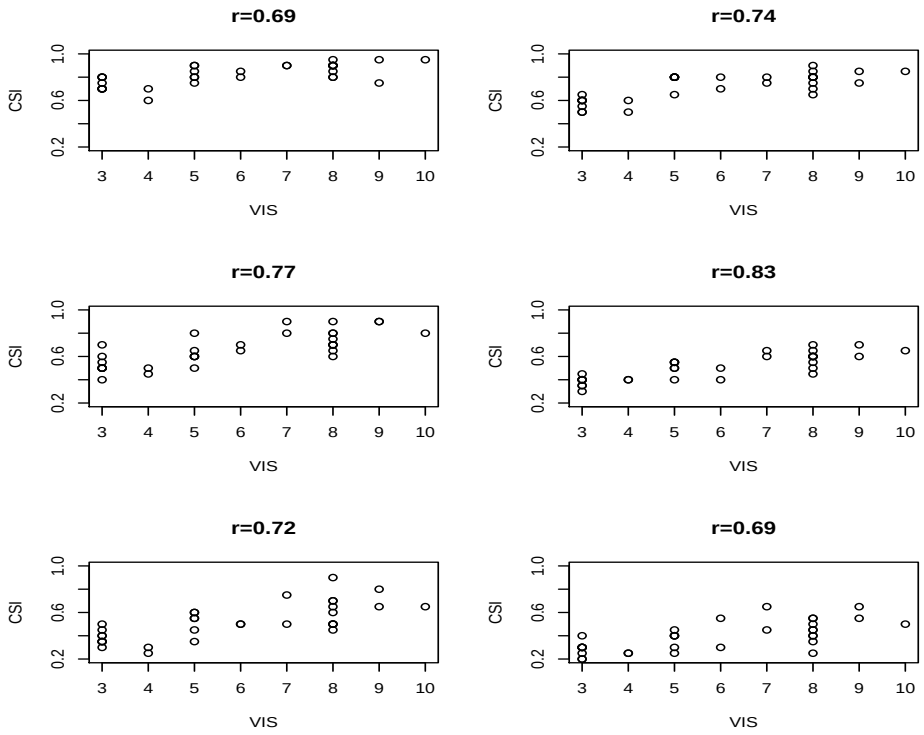

Figure 6. Average (over three models) CSI versus VIS, for threshold = 0.01 (top), 0.1 (middle), and 0.2 (bottom), for NC=20 (left) and NC=60 (right). The corresponding correlation coefficient, r, is also shown.

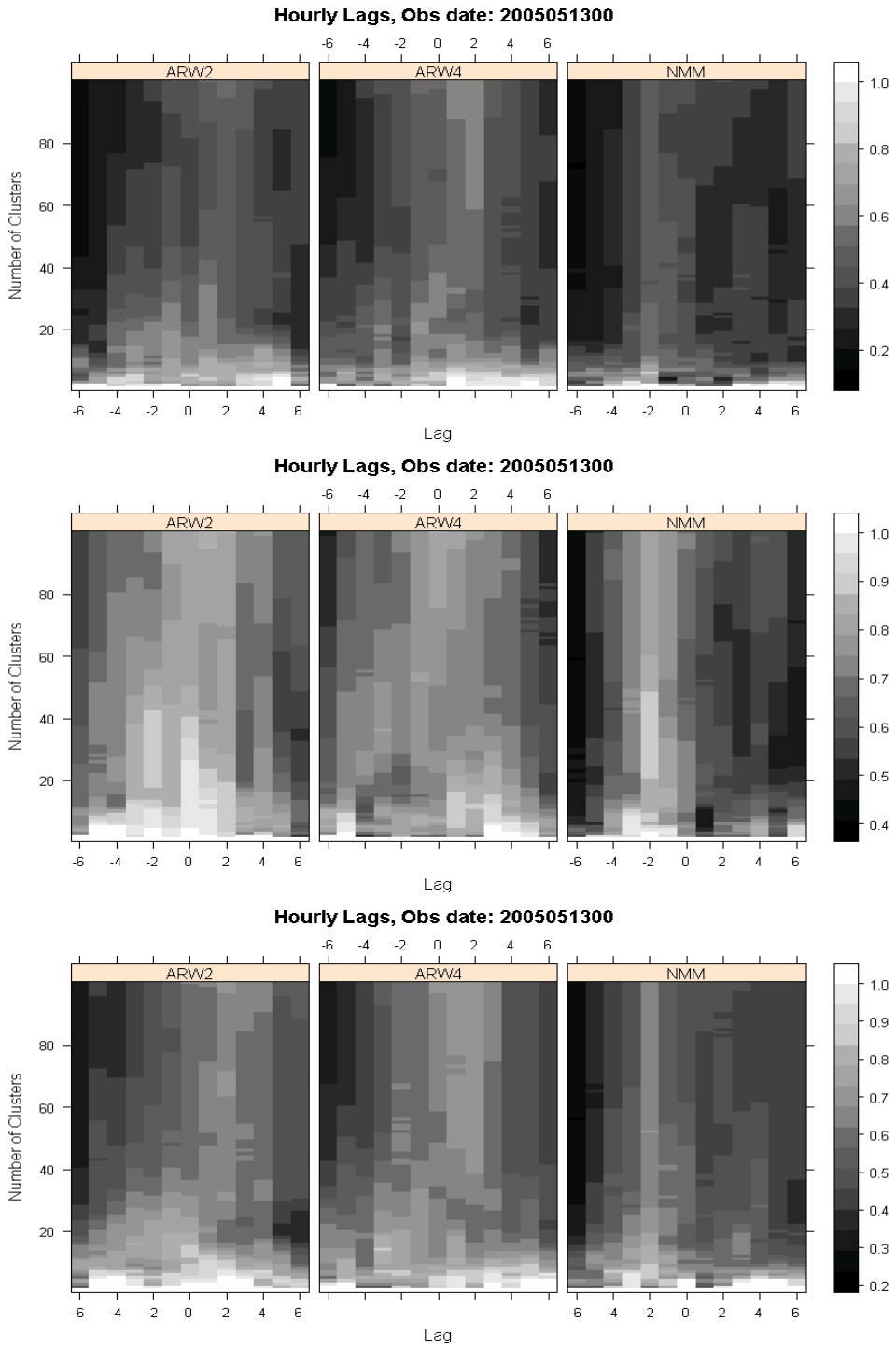

Figure 7. The csi-surfaces when the analysis is performed in (x, y) space (top), in (x, y, z) space, with each variable contributing equally to distance calculations (middle), and with the three variables weighted differently (bottom); see text.

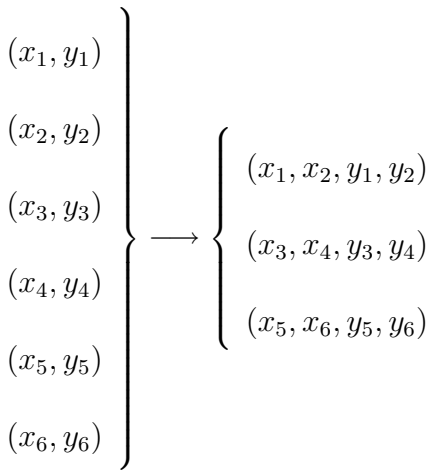

Figure 8. An illustration of the transposition which maps a clustering of 6 cases in 2 dimensions into a clustering of 3 cases in 4 dimensions.

Table 1. The contingency table reflecting the association between the three models (rows), and a ranking of the three models according to CSI at $NC = 20$, and $NC = 60$.

	Rank based on CSI					
	NC=20			NC=60		
Model	1	2	3	1	2	3
arw2	12	19	1	14	18	0
arw4	7	23	2	7	16	9
nmm4	7	17	8	12	11	9

Table 2. The contingency table reflecting the association between the three models (rows), and a ranking of the three models according to a visual inspection of the forecast field and the corresponding observation field.

	Rank		
Model	1	2	3
arw2	12	18	2
arw4	3	22	7
nmm4	12	16	4