
# Three Spatial Verification Techniques: Cluster Analysis, Variogram, and Optical Flow

CAREN MARZBAN

*Applied Physics Laboratory, and Department of Statistics, University of Washington, Seattle, Washington*

SCOTT SANDGATHE

*Applied Physics Laboratory, University of Washington, Seattle, Washington, and College of Oceanic and Atmospheric Sciences, Oregon State University, Corvallis, Oregon*

HILARY LYONS

*Department of Statistics, University of Washington, Seattle, Washington*

NICHOLAS LEDERER

*Applied Physics Laboratory, University of Washington, Seattle, Washington*

(Manuscript received 8 January 2009, in final form 29 June 2009)

ABSTRACT

Three spatial verification techniques are applied to three datasets. The datasets consist of a mixture of real and artificial forecasts, and corresponding observations, designed to aid in better understanding the effects of global (i.e., across the entire field) displacement and intensity errors. The three verification techniques, each based on well-known statistical methods, have little in common and, so, present different facets of forecast quality. It is shown that a verification method based on cluster analysis can identify ''objects'' in a forecast and an observation field, thereby allowing for object-oriented verification in the sense that it considers displacement, missed forecasts, and false alarms. A second method compares the observed and forecast fields, not in terms of the objects within them, but in terms of the covariance structure of the fields, as summarized by their variogram. The last method addresses the agreement between the two fields by inferring the function that maps one to the other. The map—generally called optical flow—provides a (visual) summary of the ''difference'' between the two fields. A further summary measure of that map is found to yield useful information on the distortion error in the forecasts.

## 1. Introduction

Variograms and correlograms are both invariant under additive intensity errors. Under multiplicative intensity errors, however, only the correlogram is invariant; that is, a correlogram captures displacement (and shape–size) error only, not additive or multiplicative intensity errors.

It is now clear that the quality of forecasts of gridded parameters such as precipitation or temperature cannot be evaluated by a simple gridpoint by gridpoint comparison of the forecast field with the observed field. This issue has been thoroughly discussed in the literature, and a summary is provided in Ahijevych et al. (2009). Also discussed in that work are three datasets designed to diagnose the inner workings of a number of verification techniques for a proper assessment of spatial–gridded forecasts. Among those techniques, three have been examined previously by the authors of this article; they are referred to as the cluster analysis (CA) method (Marzban and Sandgathe 2006, 2008; Marzban et al. 2008), the variogram (VGM) method (Marzban and Sandgathe 2009a), and the optical flow (OF) method (Marzban and Sandgathe 2007, manuscript submitted to *Wea. Forecasting*, hereafter MSI; Marzban and Sandgathe 2009b,

*Corresponding author address:* Caren Marzban, Dept. of Statistics, University of Washington, Box 354322, Seattle, WA 98195-4322.
E-mail: marzban@stat.washington.edu

DOI: 10.1175/2009WAF2222261.1

manuscript submitted to *Wea. Forecasting*, hereafter MSII). The three methods have little in common and, so, examine completely different facets of forecast quality.

The CA method can be called object oriented in the sense described by Baldwin et al. (2002), Brown et al. (2004), Bullock et al. (2004), Chapman et al. (2004), Davis et al. (2006a,b, 2009), Ebert and Gallus (2009), and Ebert and McBride (2000). Given that it relies on the identification of spatially compact objects, it is suited for fields consisting of mixed (continuous and discrete) distributions, such as precipitation fields. The VGM method is closely related to ideas put forth by Gebremichael et al. (2004), Germann and Joss (2001), Germann and Zawadzki (2002), Harris et al. (2001), and Zepeda-Arce et al. (2000). It is designed to compare two fields in terms of their covariance structures. And the OF method is related to techniques examined by Bowler et al. (2004), Du and Mullen (2000), Gilleland et al. (2010, manuscript submitted to *Wea. Forecasting*), Hoffman et al. (1995), and Keil and Craig (2007, 2009). A classification of all of these techniques (and more) has been attempted in Gilleland et al. (2009).

Not all of the above works deal with the verification problem per se, but when they do, the primary task is to first assess some scalar measure of forecast error and, then, decompose it into components that may have some diagnostic value. Some of the more emphasized components have been the displacement error, intensity error, and size error. The three methods presented here assess the decomposition of error differently. The CA method is based on the notion of a distance between clusters in the forecast field and those in the observed field. The distance may be measured either in the Euclidean space spanned by $x$ and $y$ coordinates, or it may be measured in the three-dimensional space consisting of the Euclidean plane $(x, y)$ and the intensity of the field (denoted $z$). The CA method performed in $(x, y)$ assesses only displacement error, while the $(x, y, z)$ analysis gauges a combination of displacement and intensity error. The precise combination is determined by a metric introduced in the three-dimensional space. In this paper, the majority of the analysis is in $(x, y, z)$ and, so, the results convey a combination of displacement and intensity errors.

The VGM approach addresses the decomposition differently. As shown by Marzban and Sandgathe (2009a), a comparison of two fields in terms of their variograms can be performed in two different ways: one is sensitive to both displacement and intensity error, while the other is insensitive to displacement error. It is also possible to compute a similar quantity, called the correlogram, which is insensitive to intensity. The bulk of the analysis in this paper is based on a version of the variogram that is sensitive to both displacement and intensity errors, but an example of the correlogram is also given. Examples of the variogram that is insensitive to displacement are given in Marzban and Sandgathe (2009a).

TABLE 1. The five geometric forecasts and their definitions. The numbers denote the magnitude of shift.

| | |
|---|---|
| geom001 | 50 → |
| geom002 | 200 → |
| geom003 | 125 →, bigger |
| geom004 | 125 →, wrong orientation |
| geom005 | 125 →, very big (i.e., overlapping) |

The OF method is also capable of assessing displacement and intensity errors. As shown in MSII, the simplest OF model assesses a combination of the two errors, but a simple generalization allows for gauging the two components separately. Again, the OF analysis done here is based on the simple model, but examples of the decomposition can be found in MSII.

The next section describes the three datasets and is followed by a section reviewing the three verification methods. When appropriate, each method is applied to the three datasets, and the results are presented. The paper ends with a summary of the conclusions and a discussion.

## 2. The data

The three datasets examined here are described in Ahijevych et al. (2009). The first, referred to as the geometric set, consists of an observed field that involves a single elliptical object. The object has low and constant intensity on its periphery, but high and constant intensity in its interior regions; the object is asymmetric in the sense that the region of high intensity is not at its center. The geometric set consists of five forecast fields, with varying amounts of displacement and/or spatial stretching applied to the observed field. These five forecast fields are referred to as geom001–geom005, and for the sake of completeness are briefly defined in Table 1, where arrows indicate the directions of the displacement.

The second dataset, called the perturbed set involves a realistic observed precipitation field and seven forecast fields that are generated by applying varying amounts of displacement and intensity scaling to the observed field. The forecast fields are labeled as pert001–pert007, and the underlying transformations are succinctly displayed in Table 2.

The third dataset pertains to precipitation observed on nine dates in 2005: 26 April; 13, 14, 18, 19, and 25 May; and 1, 3, and 4 June. Each observed field is accompanied by three 24-h forecast fields from three formulations of the Weather Research and Forecasting (WRF) model, referred to as wrf2caps, wrf4ncar, and wrf4ncep. Details of these models can be found in Ahijevych et al. (2009).

TABLE 2. The seven perturbed forecasts and their definitions. The numbers denote the magnitude of shift.

| | |
|---|---|
| pert001 | (3 →, 5 ↓) |
| pert002 | (6 →, 10 ↓) |
| pert003 | (12 →, 20 ↓) |
| pert004 | (24 →, 40 ↓) |
| pert005 | (48 →, 80 ↓) |
| pert006 | pert003 × 1.5 |
| pert007 | pert003 − 1.27 mm |

## 3. The three methods

Each of the methods has multiple variations, only one of which is examined here. Moreover, each method has user-dependent parameters that are also fixed here. Detailed information about these choices can be found in the corresponding references. In this section, each method is described briefly, and an example of its "output" is presented. The example is one of the forecasts from the perturbed dataset involving only a shift.

The CA method identifies clusters or objects in the combined field of the forecast and the corresponding observation. The clusters are then assayed for their number of grid points that belong to the observed field, and the number of grid points belonging to the forecast field. If these two numbers are comparable, indicating significant overlap of the two fields, then the cluster is identified as a hit; otherwise, the cluster is a false alarm or a miss, depending on which of the two numbers is larger. The details of this matching criterion are discussed in Marzban and Sandgathe (2008) and Marzban et al. (2008). From the numbers of hits, misses, and false alarms, one computes the critical success index (CSI) as a measure of performance.[1] It is important to point out that the entire clustering procedure can be performed in a multidimensional space that includes, but is not limited to, the two spatial coordinates. In fact, the results reported here are based on three coordinates: the two spatial coordinates, plus intensity. In the current analysis the last coordinate has been weighted so as to contribute a third as much as the spatial coordinates; again, this choice is user dependent.

Cluster analysis techniques may be divided into two major classes: those wherein the number of clusters, NC, is specified a priori, and those for which the number is variable. An example of the former is k-means clustering, while an example of the latter is hierarchical agglomerative clustering (Everitt 1980). The latter begins by partitioning the field into $n$ clusters, where $n$ is the number of grid points. The technique then finds the closest clusters and joins them into a larger cluster. The procedure continues until there exists a single cluster consisting of all grid points. As such, NC varies from $n$ to 1. If the verification is performed on a large number of clusters, then one can argue that verification is done on a small scale. By contrast, large-scale verification is done when the number of clusters is small. By computing CSI for every value of NC, one obtains a "CSI curve," which effectively summarizes the forecast quality as a function of scale.

As an illustration of the technique, the top panel in Fig. 1 shows an example of partitioning a joint observed–forecast field into 100 clusters. The clustering algorithm used to generate the clusters in Fig. 1 is the aforementioned

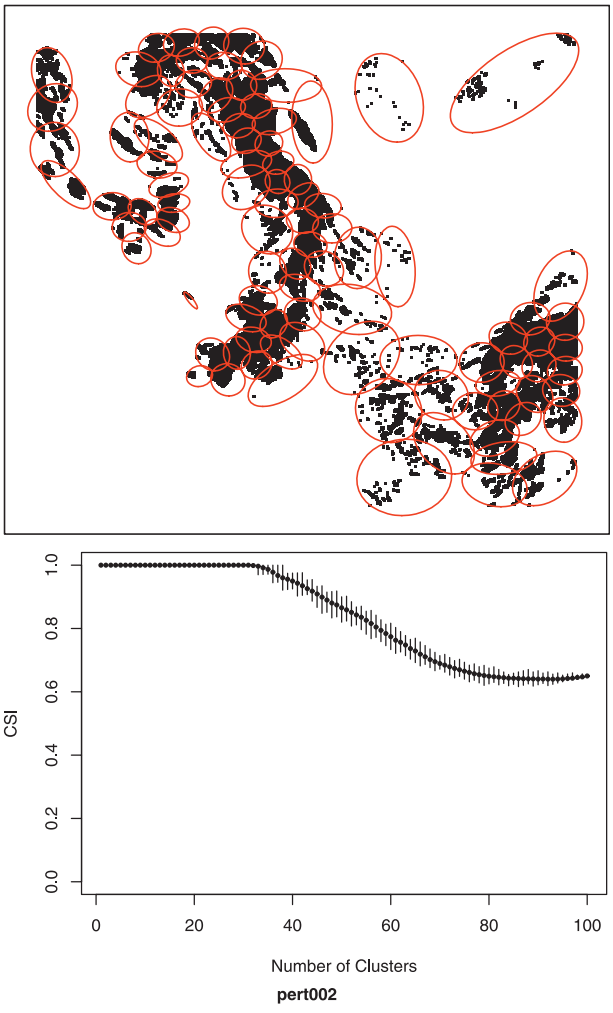


FIG. 1. (top) An example of partitioning a joint observed–forecast field into 100 clusters, and (bottom) the CSI curve from the CA method. The forecast is pert002. Each error bar is the central 90% interval.

[1] The reasons for selecting CSI as the measure of performance have been addressed in Marzban and Sandgathe (2006). CSI does have a number of "defects," such as not being a measure of skill or being misleading in rare-event situations (Marzban 1998), but these problems are not important in the current application.

k-means algorithm; in its simplest form, it assumes that the clusters are elliptical in shape. The algorithm used in the CA method begins with the result of the k-means algorithm, but further clustering is performed with the hierarchical agglomerative algorithm. The CSI curve corresponding to the forecast pert002 is shown in the bottom panel of Fig. 1. The error bars are the central 90% interval, generated via bootstrapping. In this particular instance, one may conclude that the forecasts are perfect (CSI = 1) on large scales corresponding to 30 or fewer clusters. On smaller scales, where the field is partitioned into more than 30 clusters, CSI falls off monotonically. For this example, the CA is performed only on the spatial coordinates (i.e., excluding intensity); other examples that also include intensity are examined below.

The central quantity in the VGM method is the variogram, an empirical plot of the mean-squared difference between the field values at two points, as a function of the distance separating the two points (Cressie 1993). The mean is computed over *all* points separated by a fixed distance. Consequently, the variogram assesses variations in the field as a function of scale. In spatial statistics, it is used to summarize the covariance structure of a field, while in image processing it gauges the texture of an image. For verification purposes, the variogram is useful because it allows a comparison of a forecast field and an observation field, as a function of scale. Marzban and Sandgathe (2009a) propose two versions of the VGM method; in one, the variogram is computed across all grid points in a field. However, if a field consists of a mixed continuous-discrete quantity, such as precipitation, then it also makes sense to compute the variogram across only nonzero grid points in the field. It can be shown that a performance measure based on the former variogram assesses forecasts in terms of all the components of error (displacement, intensity, and size). The latter variogram is invariant under global displacements (i.e., shifts), and so, it is insensitive to displacement errors. In this paper, only the former variogram is computed, because as mentioned in the introduction, these verification techniques are intended for (nondiagnostic) model comparison. The final ''product'' of the procedure is a plot of the difference between the variogram of the forecast field and that of the observed field, denoted ''delta variogram.'' If the delta variogram overlaps the horizontal line at $y = 0$, that would indicate that the forecasts are ''perfect'' (as far as texture is concerned) at all scales. The top-left panel in Fig. 2 shows the variograms for the observed field (black) and the forecast pert006 (red), and the right panel shows the delta variogram. More specifically, each variogram is actually a sequence of boxplots summarizing the sampling distribution of the difference between the observed and the forecast variograms, based on 50 bootstrap samples.[2] Recall that pert006 corresponds to a global-spatial shift and a multiplicative error in intensity; the delta variogram displays the difference between the two variograms. The fact that the boxplots do not cover the horizontal line at 0 implies that the two variograms are statistically distinct.

As mentioned above, the variogram is affected by displacement and intensity errors. By contrast, the correlogram is insensitive to intensity errors. Its definition is already given in Marzban and Sandgathe (2009a); suffice it to say that it is a two-dimensional generalization of the Pearson correlation coefficient and, so, is insensitive to the magnitude of the field. As such, the difference between two correlograms, called delta correlogram here, assesses only the displacement error. The bottom row in Fig. 2 shows the corresponding correlograms (left) and the delta correlogram (right). Although the difference between the correlograms in the left panel is not visually clear, the delta correlogram makes it abundantly evident that there is a difference; the difference is only on large scales (i.e., only the boxplots for large $x$ values do not overlap the $y = 0$ line). This is what one would expect, because a global shift (underlying pert006) is indeed a large-scale effect.

The final method is based on the idea that any field can be mapped to any other field, and that certain features of the map can be used as verification measures. The map is often referred to as the optical flow (OF). If the OF field is allowed to be completely general, then one can use nonparametric methods for estimating the map. Such methods are generally numerically intensive and do not allow analytic solutions. Keil and Craig (2007, 2009) have examined this approach. An alternative is to impose certain constraints on the map that allow for an analytic solution. Of course, because of the constraints, the formalism is more restrictive than the nonparametric approach, but it does provide for better illustration. One popular constraint is from Lucas and Kanade (1981), where it is assumed that the OF field is locally constant. This approach is examined for verification purposes by MSI and MSII. The result of the procedure is a 2D field of vectors, one per grid point, mapping the forecast to the observed field. At each grid point, the OF vector is computed from the field values at all neighboring grid points. The extent of the neighborhood is quantified by a window of size $W$. The magnitudes and directions of

[2] The sampling performed here assumes independence of the field values at different grid points. This assumption is clearly invalid for most meteorological fields. Marzban and Sandgathe (2009a) discuss this issue and propose two alternative sampling schemes that do allow for spatial (and temporal) correlations in the data: subsampling (Politis et al. 1999) and the block bootstrap (Politis and Romano 1994).

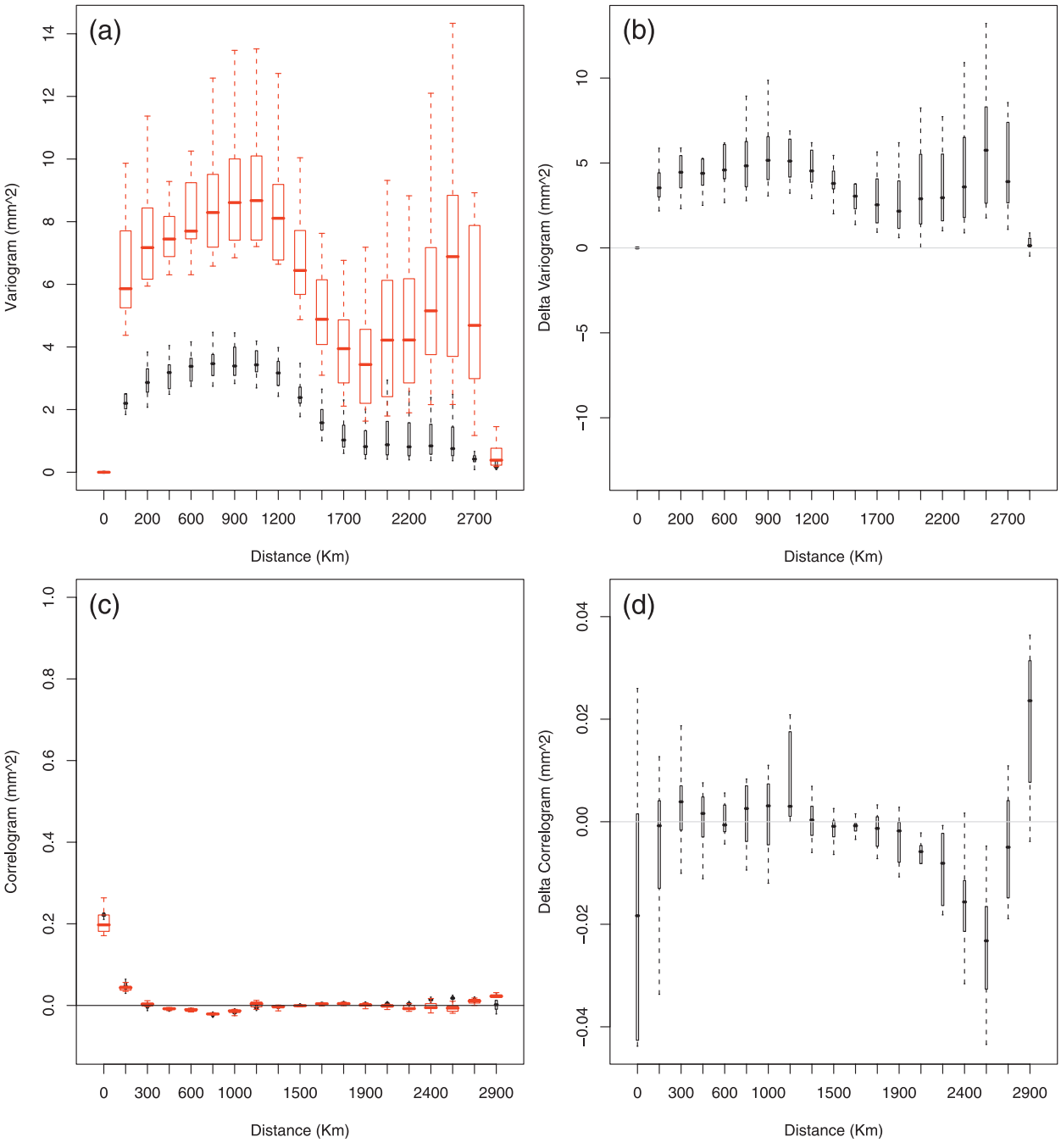


FIG. 2. (top left) Variograms for an observed field (bottom points), and a corresponding forecast field, pert006 (top points). (top right) The difference between the two variograms. (bottom panels) The corresponding correlogram.

these vectors—specifically, their joint histogram—can be used as a summary measure for distortion error (i.e., size, displacement, and intensity error combined). A "peak" displayed in the joint histogram indicates a coherent, large-scale transformation, that is, an overall shift of the forecast field relative to the observed field. Figure 3 shows the joint histogram for pert001, and will be discussed further, below. For now, recall that the magnitude of the shift generating pert001 is $\sqrt{3^2 + 5^2} = 5.8$, and the angle is $360/2\pi \ \text{atan}(-5/3) = -59°$. These numbers are entirely consistent with the joint histogram in Fig. 3. This example, along with other simulated cases examined in MSII, suggest that the joint histogram is a useful tool in summarizing the OF field.

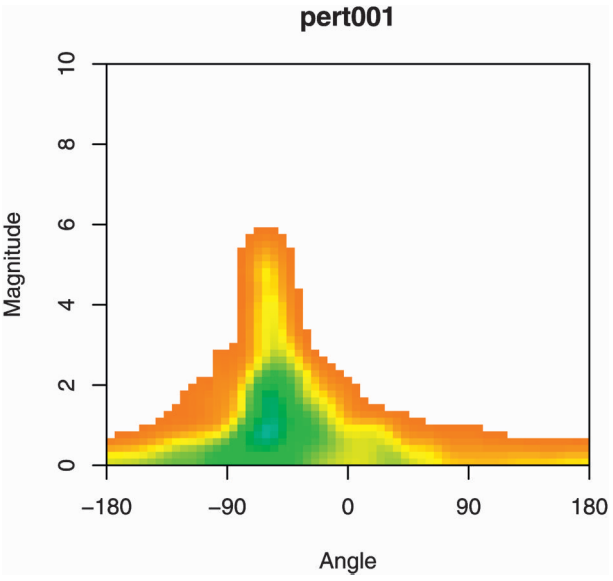


FIG. 3. The joint histogram summarizing the OF field mapping the observed field to the forecast field pert001.

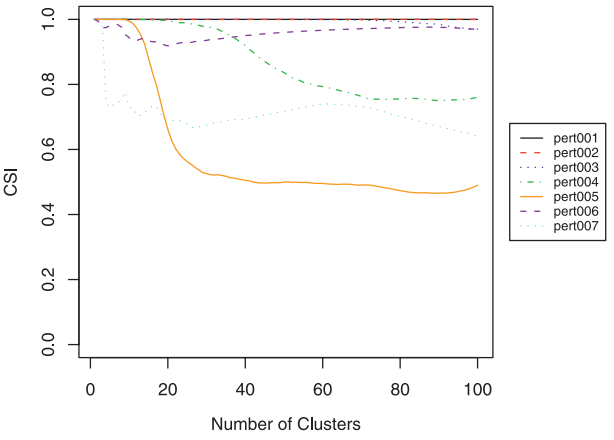


FIG. 4. The CSI curves for the seven perturbed forecasts.

One of the limitations of the assumption of a locally constant OF field is that the method will produce unphysical OF fields if the amount of shift between an observed and a forecast object is large relative to the scale of interest. In the Lucas–Kanade formulation of OF, the scale is specified by the size of the window, $W$, over which the OF is computed. In other words, for sufficiently small $W$, the OF field will appear to be unphysical. This is not a defect of the methodology, but it simply reflects the physical requirement that sufficiently distant objects cannot be, and should not be, mapped to one another, for they may in fact be distinct objects, that is, a miss or a false alarm. It is also true that increasing the window size will render the OF field more physical; after all, although two distant objects should not be necessarily matched on small scales, it is more reasonable to match them on larger scales. In short, an OF field is highly dependent on the scale over which it is computed.

To conclude this section, it is worth pointing out that an important difference between the three methods is in the way scale is quantified: in CA, the number of clusters in a field addresses the scale. In the VGM method, the scale is gauged in terms of the distance between two points, and in the OF method, it is quantified through the size of the window for which a single OF vector is computed.

## 4. CA results

Given that the geometric dataset contains only one unambiguous cluster, the CA method is not a natural method. The resulting CSI curves (not shown) all begin at CSI = 1 for NC = 1, and drop abruptly to zero for larger numbers of clusters. This is consistent with the fact that the fields do not have a wide range of scales (again, because each field consists of a single unambiguous object).

The perturbed cases provide for more insight into the method. Figure 4 shows the CSI curves (without the error bars, for visual ease) for the seven cases. Recall that the first five cases correspond to increasingly larger shifts in the forecasts relative to the observed field. Also recall that the only errors in these five forecasts are spatial, not involving changes in intensity. For the smallest shifts (i.e., pert001 and pert002), the CSI curves are constant at CSI = 1. This can be understood by noting that the current CA method is based on clustering in three dimensions: two spatial coordinates plus intensity. The equivalence of the CSI curves at CSI = 1 simply means that the spatial component of the error is sufficiently lower than the intensity component so as to not affect performance. In other words, the perfect forecast of intensity dominates the imperfect spatial structure of the forecast field. The notion of ‘‘perfect’’ forecasts requires more qualifications, which are presented in the discussion section.

For larger shifts (pert003) the CSI is 1, but only for NCs less than 70. Said differently, on smaller scales at which the field can be resolved as having 70 or more clusters, forecast quality is no longer perfect. This pattern continues for even larger shifts (e.g., pert004), for which the CSI drops below 1 for NCs larger than 20. For pert005, the drop in CSI occurs at NC = 10. In short, for larger shifts, the imperfection in the forecast can be detected at even larger scales. This is a desirable behavior of CSI curves.

Larger shifts also cause the CSI curve to fall off with NCs more drastically. In fact, the drop in CSI is larger for larger shifts. This simply confirms that the CSI is

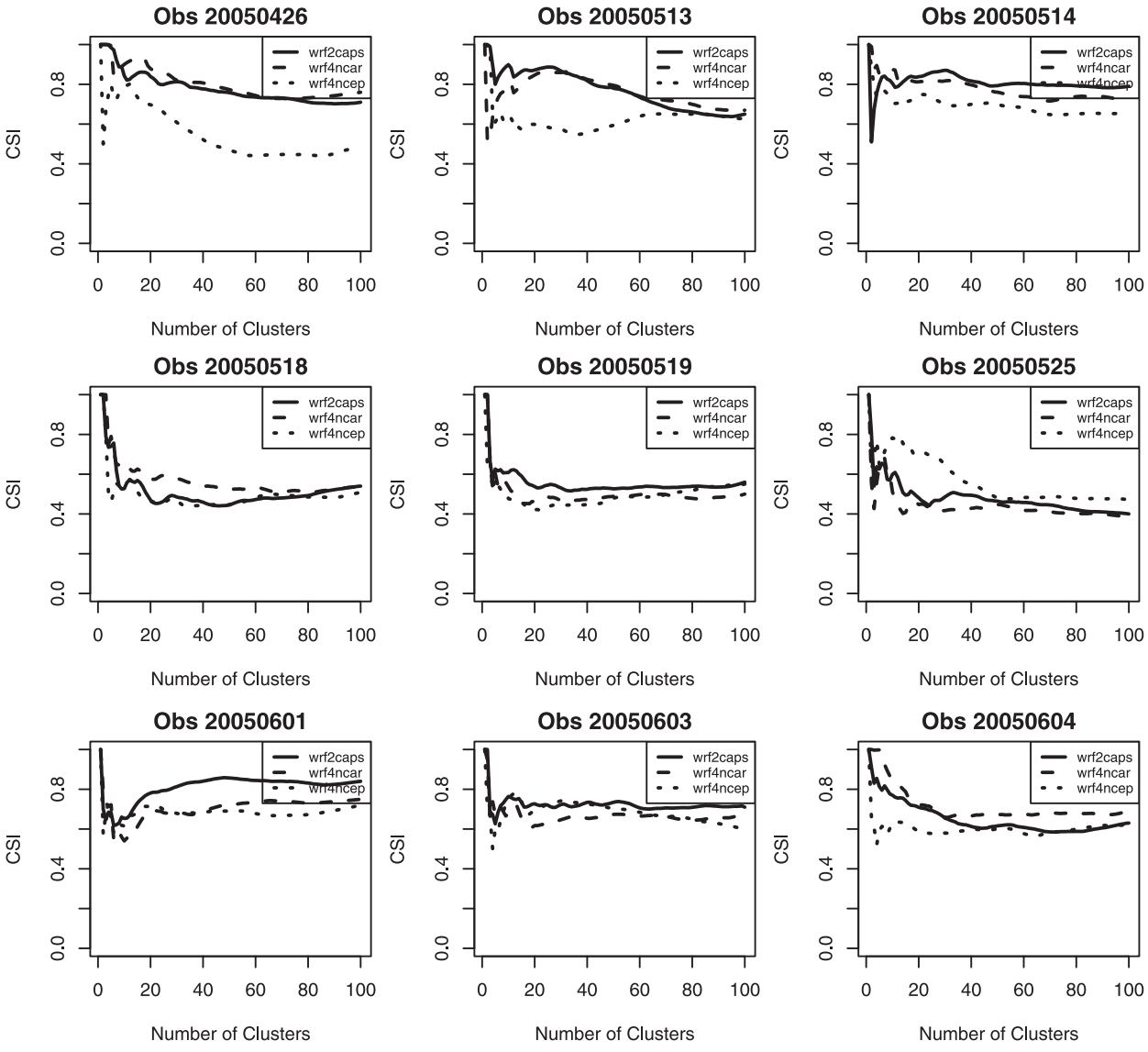


FIG. 5. The CSI curves for the nine dates during spring 2005. The three curves in each panel correspond to wrf2caps, wrf4ncar, and wrf4ncep.

negatively affected by displacement errors, as it should be. Similarly, CSI is negatively affected by intensity errors, as evidenced by the lower CSI curve for pert006 and pert007, relative to pert003. It is interesting to note that the CSI curve for pert006 (corresponding to a shift plus a multiplicative intensity error of 1.5) is higher than that of pert007 (corresponding to a shift plus an additive intensity error of $-1.27$ mm), but one cannot conclude that the CA method is more or less sensitive to additive or multiplicative errors, in general, because such a conclusion would also have to be contingent on the relative magnitude of the additive and multiplicative intensity errors.

Figure 5 shows the CSI curves for the nine dates during spring 2005, for the three NWP forecasts (wrf2caps, wrf4ncar, and wrf4ncep). It can be seen that wr2caps and wrf4ncar are comparable across all scales; although the sampling variations of these curves are not shown, paired $t$ tests indicate that the differences are not statistically significant, even when the point estimates (i.e., the curves) appear to be different. As for wrf4ncep, on small scales with NC $>$ 10, it can be better or worse than wrf2caps or wrf4ncar depending on the date. As such, its relative performance is difficult to assess. On larger scales with NC $<$ 10, it appears to be no better than wrf2caps or wrf4ncar. In short, no particularly simple pattern can be inferred from these nine dates; apart from the possibility that wrf4ncep is generally more variable than wrf2caps or wrf4ncar. A comparison of the reflectivity

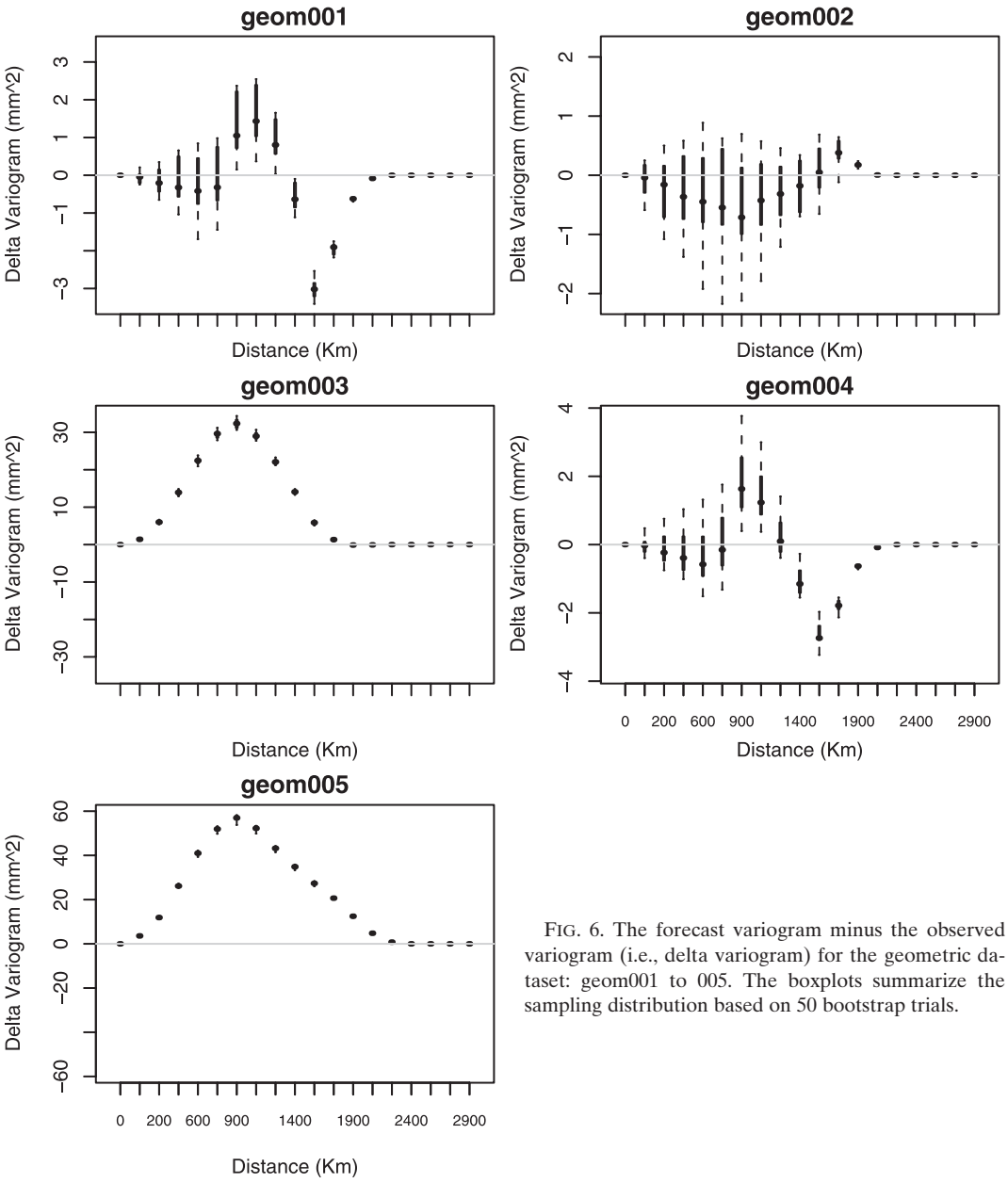


FIG. 6. The forecast variogram minus the observed variogram (i.e., delta variogram) for the geometric dataset: geom001 to 005. The boxplots summarize the sampling distribution based on 50 bootstrap trials.

forecasts from these three models, across 30 days, has been performed by Marzban et al. (2008).

## 5. VGM results

The effects of a global shift on a variogram can be seen in the top two panels of Fig. 6. Consider the case of a 50 gridpoint shift, first (geom001). Evidently, the only region where the boxplots do not overlap the horizontal line at zero is on larger scales (i.e., about 1800 km). On smaller scales, the variogram has significant overlap with the horizontal line at $y = 0$. This behavior is exactly what one would expect, because a global shift is clearly a large-scale change. The geom002 panel in Fig. 6 shows the effects of an even larger shift of 200 grid points. One might expect an even more pronounced large-scale effect, but according to Fig. 6, there is no significant difference between the observed and the forecast fields. The resolution of this ''paradox'' is in the realization that a shift of 200 happens to lead to a forecast that is a mirror image of the observed field (at least with regard to the spatial location of the object). It is easy to show that a variogram is invariant under such transformations. The reason the variograms of the two fields are not *exactly* the

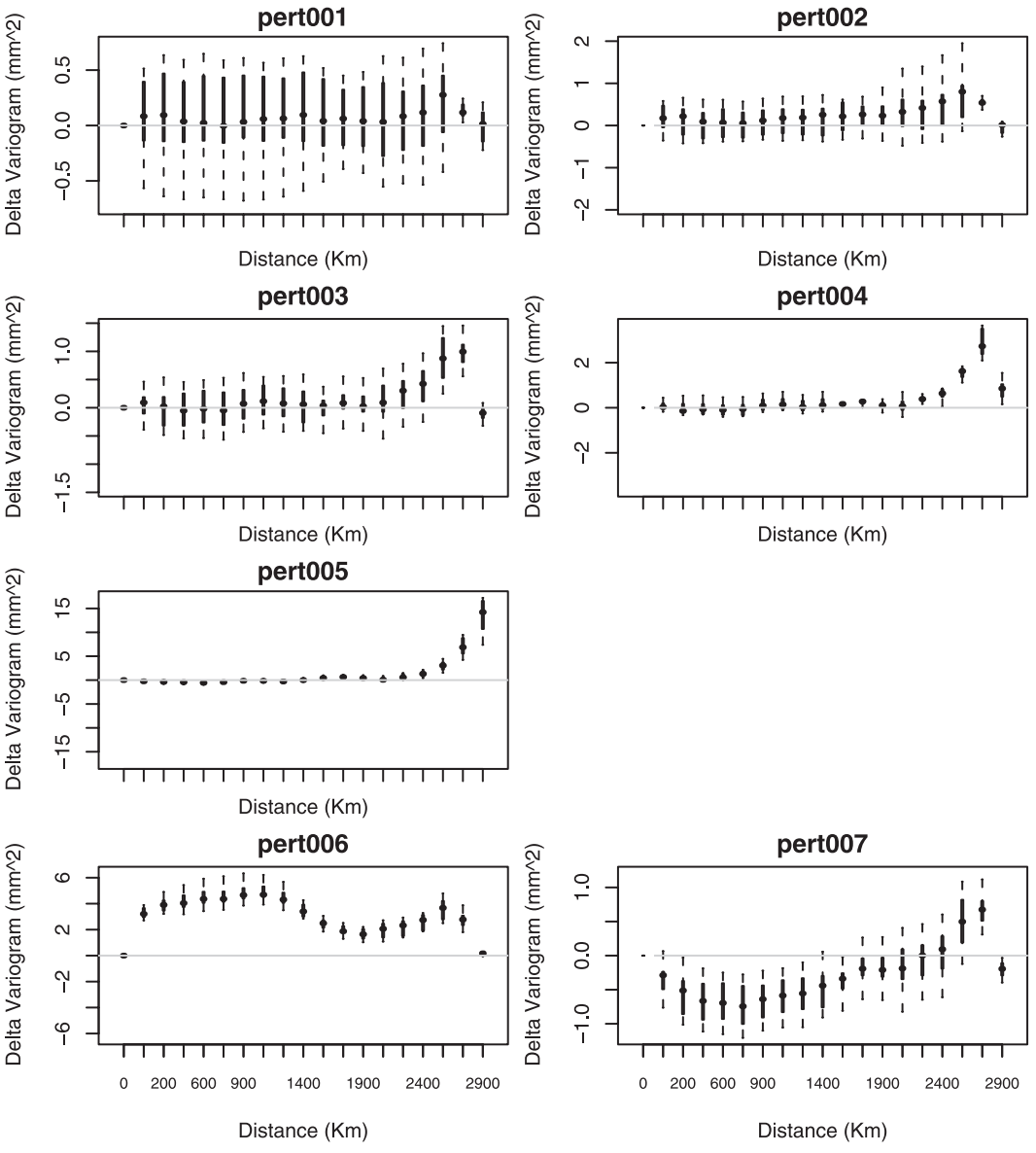


FIG. 7. As in Fig. 6, but for the perturbed dataset.

same is that the object itself is asymmetric, breaking the invariance of the variogram. However, this is not of practical concern, because it is unlikely to come across a realistic forecast field that is a mirror image of the observed field.

The geom003 and geom005 panels in Fig. 6 differ mostly in the magnitude of the variogram difference. This is consistent with the fact that the only difference between the underlying forecasts is the size of the forecast object. Finally, the geom004 panel resembles the geom001 panel. In other words, the transformation generating geom004 leaves the variogram invariant. This is, again, an artifact of the manner in which geome004 is generated.

The variograms for the perturbed dataset are equally easy to interpret. The panels pert001 through pert005 in Fig. 7 show delta variogram for shifted forecasts of varying degrees. It is clear that the difference between the forecast and observed variogram is only at the larger distances, and this is precisely what one would expect, since a global shift is in fact a large-scale transformation. The last two panels in Fig. 7 also convey what one would expect from changes in intensity: if the forecast intensity is multiplied by a positive number, then the variogram shifts to higher values, and if the intensity is reduced uniformly by some number, then the variogram is reduced by some amount. In particular, pert007, which includes

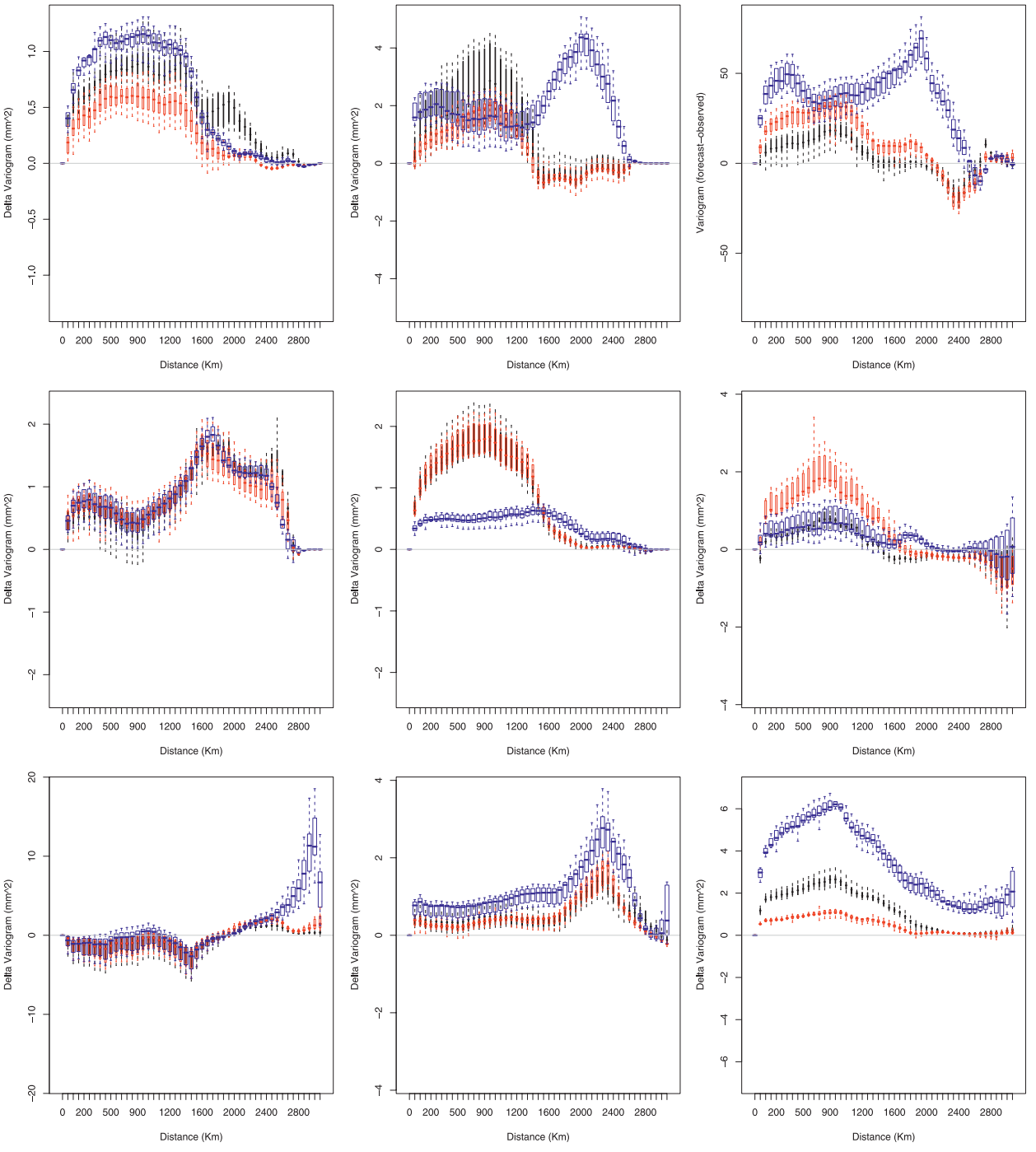


FIG. 8. Delta variogram for the spring 2005 cases. The forecasts are from wrf2caps (black), wrf4ncar (red), and wrf4ncep (blue). The panels are arranged as in Fig. 5.

both a displacement error and an additive error in intensity, has a delta variogram that reflects these errors accordingly: on smaller scales (<1500 km), the delta variogram is below zero, reflecting a reduction in intensity of 1.27 mm, but on larger scales, it is above zero, as a result of the spatial component of the error in the forecast.

As for the spring 2005 dataset, the delta variograms are shown in Fig. 8. The forecasts are from wrf2caps (black), wrf4ncar (red), and wrf4ncep (blue). On 26 April (Fig. 8, top left), wrf4ncar is clearly better than either wrf2caps or wrf4ncep, and this is true of all scales. However, on scales larger than about 1500 km, wrf2caps

is inferior to wrf4ncar and wrf4ncep. By contrast, on 13 May (Fig. 8, top middle), wrf4ncar is far worse than wrf4ncep on these large scales, and wrf2caps is statistically equivalent to wrf4ncar. On one day (18 May), all three forecasts are statistically equivalent. This is also the case on 1 June, except on larger scales (>2500 km) where wrf4ncep is worse. Another notable point is that on 3 June (Fig. 8, bottom right), wrfncep is worse than the other two models, over all scales. Other specific observations can be made for each date, but no simple pattern emerges across the nine dates. One may simply note that wrf2caps and wrf4ncar are generally comparable, but wrf4ncep is more variable—the same conclusion noted in the CA analysis, above. A similar analysis but across 30 days has been performed in Marzban and Sandgathe (2009a).

## 6. OF results

As mentioned previously, the Lucas–Kanade OF formulation can produce unphysical OF fields under certain conditions, for example, when a shift is large compared to the window size $W$. In this paper, only three values of $W$ are examined: $W = 20$, 40, and 60. This range is sufficiently wide to expose some discussion-worthy features in the perturbed dataset and in the spring 2005 data. However, the shifts underlying the geometric dataset are much larger than can be captured by this range of $W$ values. For this reason, the OF fields associated with the geometric dataset will not be presented here. A more complete analysis should examine a wider range for $W$, but there exist complications, some of which are addressed in the discussion section here and in MSII. More general versions of the OF method exist that are capable of handling such situations (Keil and Craig 2007, 2009).

Figure 9 shows the resulting joint histograms for the perturbed dataset, based on $W = 40$. First, recall that (Table 2) all of the transformations involve a shift in a direction $-59°$. All of the joint histograms in Fig. 9 do indeed have a peak at $-59°$. As such, the OF method correctly captures the direction of the shift. However, for the largest shift (pert005), there are two peaks in the joint histogram. As for the magnitude of the OF vectors (i.e., displacement error), it is clear that the peak of the joint histogram is consistent with the magnitude of the displacement error only for small shifts. For example, the magnitude of the largest shift (i.e., for pert005) is $\sqrt{48^2 + 80^2} = 93.3$, but neither of the peaks in the joint histogram is around that value. This is an example of the aforementioned unphysical result. Again, this is not a failure of the method, but rather a reflection of the requirement that two distant objects may, in fact, be a miss and a false alarm and, so, should not be mapped to one another. In fact, on the scale established by $W = 40$, the larger shifts underlying pert004 or pert005 are sufficiently large to justify that interpretation.

Finally, the last two cases in the perturbed dataset (pert006 and pert007) yield joint histograms that resemble that of pert003. It may appear that the OF method does not capture the intensity error, but that is not true. In general, the OF is affected by intensity errors of any kind, additive or multiplicative (MSII). The reason the joint histograms for pert006 and pert007 are similar to that of pert003 is that the displacement error (common to both pert006 and pert007) overwhelms the intensity error. Recall that the OF fields presented here reflect a combination of displacement and intensity errors.

All of the above conclusions are contingent on the scale at which the OF is computed; the window size for Fig. 9 is 40. An examination of the joint histograms for both smaller and larger window sizes (not shown) indicates that the conclusions are generally robust across different scales. The only difference is that the amount of shift at which the method begins to ''fail'' increases with window size. This makes sense, because the answer to the question of how far should two objects be before one declares them as different objects (not the same object, but shifted) depends on the scale of interest.

As for the spring 2005 dataset, all the joint histograms have been examined for wrf2caps, wrf4ncar, and wrf4ncep, for all nine dates, and at three window sizes (20, 40, and 60). There exist too many figures to reproduce here; however, they generally fall into three broad classes. In one class the joint histograms are bimodal; these are reminiscent of those found in the perturbed dataset with large shifts. They suggest that on these scales the forecasts are of extremely poor quality. The dates that fall into this class are 13, 14, 18, and 19 May. The top row in Fig. 10 shows the joint histograms for the three models for 19 May.

The joint histogram of one date in particular, 25 May, shows signatures of a shift (Fig. 10, middle row). The shift is most pronounced in wrf4ncar, and least noticeable in wrf4ncep. The direction of the shift is about 85° (i.e., due north), and the magnitude is about five grid lengths, (i.e., about 20 km). The remaining joint histograms (1–4 June) all appear to be uniformly distributed across all angles and magnitudes. As such, the forecasts do not appear to be shifts of the observed fields. The bottom row in Fig. 10 shows the joint histogram for 2 June.

Comparison of the joint histograms of the three models across different scales ($W = 20$, 40, and 60) indicates the same conclusions found for the CA and VGM methods, namely that wrf2caps and wrf4ncar are similar, and that wrf4ncep is more variable.

MSII also consider the OF field resulting from the average of all the OF fields across multiple days. This

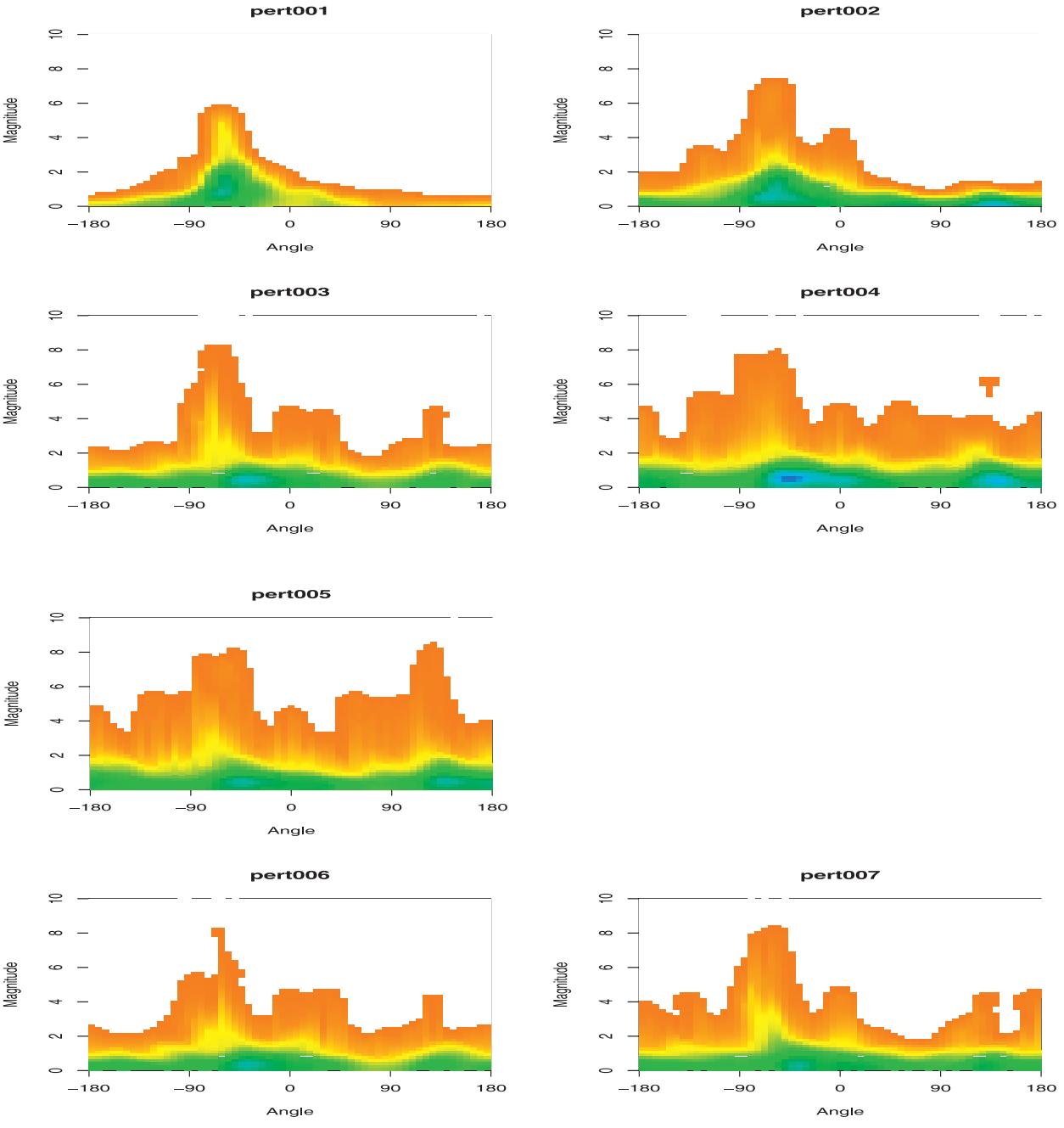


FIG. 9. The joint histogram summarizing the OF field for the seven perturbed cases.

type of an average OF field is a useful summary of forecast quality; however, it is not shown here because it is apt to be unreliable, given the small number of days (i.e., nine).

## 7. Summary and discussion

Three verification methods are applied to several datasets. The data are designed to assess how the methods measure global displacement and intensity errors. The three methods have little in common and, so, together provide a more complete picture of forecast quality. One method is based on cluster analysis (CA) and partitions a gridded field into "objects," while another method examines the spatial–covariance structure of the field itself in terms of the variogram (VGM). A third method infers a map relating the forecast and observed fields, where the map is based on ideas from optical flow (OF).

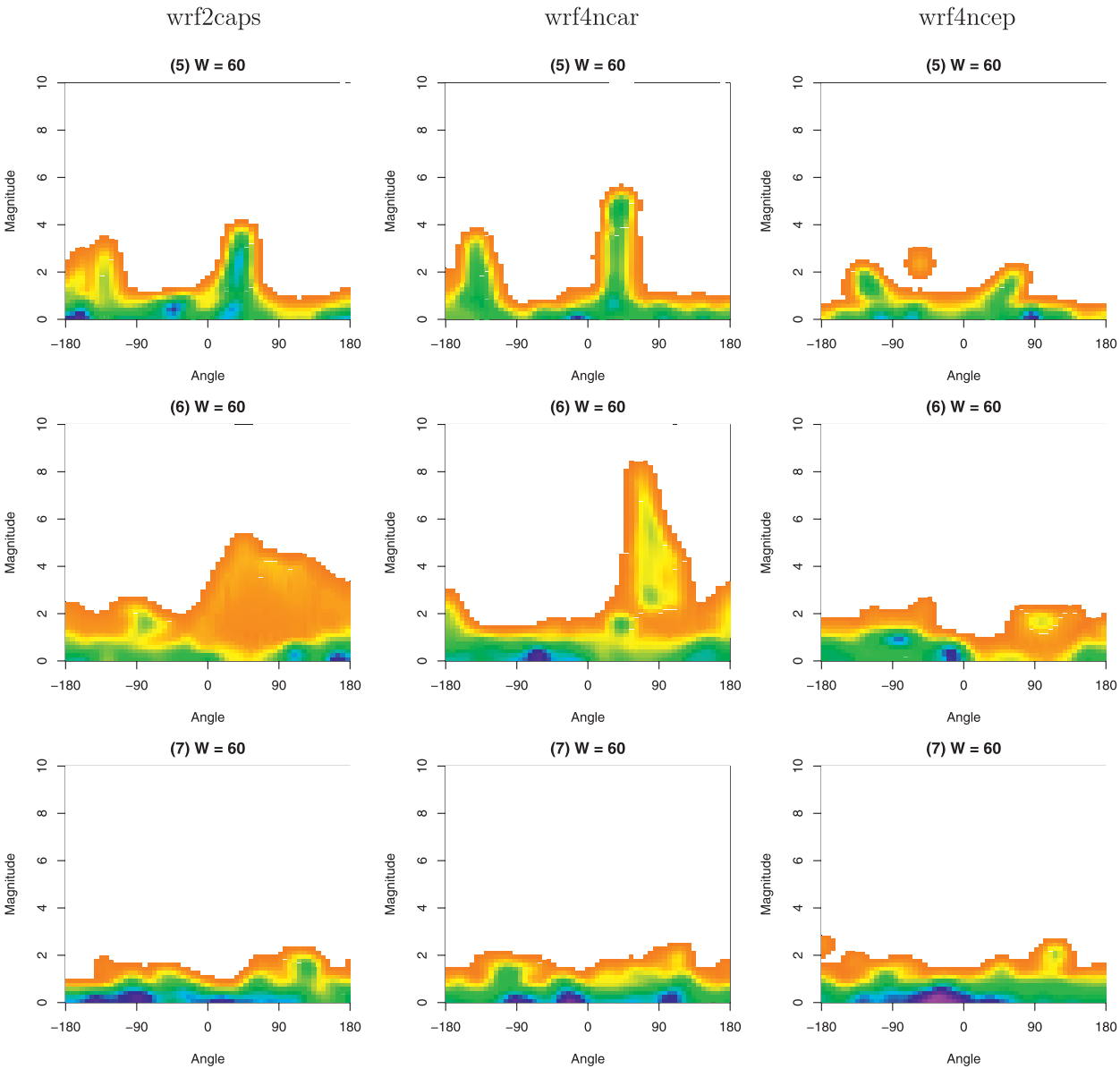


FIG. 10. The joint histograms for the three models for three of the nine dates: (top) 19 May, (middle) 25 May, and (bottom) 1 Jun: (left) wrf2caps, (middle) wrf4ncar, and (right) wrf4ncep.

The findings can be summarized as follow: When applied to forecasts that are a shifted version of an observed field, the CA method implies that larger spatial shifts lead to lower CSI curves across all scales, but the drop in CSI is more exaggerated on smaller scales. Moreover, both multiplicative and additive errors in intensity lead to lower CSI values across all scales. The VGM method is also sensitive to both spatial shifts and errors in intensity. Specifically, larger spatial shifts manifest themselves as large-scale changes in the variogram (i.e., for larger $x$ values), and both multiplicative and additive errors in intensity simply shift the variogram across all scales. By contrast, the OF method is mostly insensitive to intensity errors and is more suited to detecting spatial errors. In particular, the joint histogram of the magnitude and direction of the OF vectors reflects the magnitude and the direction of spatial shifts.

With regard to forecasts from wrf2caps, wrf4ncar, and wrf4ncep, a comparison across nine days during spring 2005, using the three verification methods, yields a complex set of conclusions that are difficult to summarize. However, it does appear to be the case that wrf2caps and wrf4ncar produce comparable forecasts across all scales; wrf4ncep's forecasts are different, and

also more variable across the nine dates. These findings can be understood by noting that wrf2caps is the same general model formulation as wrf4ncar, yet run at higher resolution. Therefore, while generally similar, wrf2caps does appear to resolve features better in some cases based on the CA and VGM methods. On the other hand, wrf4ncep is based on a different model formulation and performs differently, although not necessarily worse than the other two models.

There exist numerous issues in all three methods that legitimately place them in the realm of research. For example, in the CA method, a forecast that is generated by a small shift of the observed field will lead to CSI = 1; that is, it will deem the forecast to be perfect. One explanation, already offered above, is that the perfect forecast of intensity dominates the spatial component of error. But the issue is more complex and requires a qualification of the notion of perfect. One may wonder how the forecast can be deemed perfect when it is known to be a shifted version of the observed field. The resolution follows when one notes that CSI is a measure of performance *after* the clusters in the two fields have been matched with one another in some optimal sense. Those that are matched lead to hits, and the remaining clusters lead to false alarms or misses. In other words, CSI does not assess the quality of the match itself. Said differently, displacement errors and/or intensity errors are not embodied in CSI. These components *can* be computed. For example, the average distance between the clusters can be used to asses the quality of the match. Given that distance in the 3D space has a spatial and an intensity component, one can even view them as measures of displacement and intensity error, separately, and plot them as a function of NC. One difficulty is in deciding how the size of two clusters should effect a measure of distance between them. Even armed with an unambiguous and reasonable notion of distance, it is still unclear how the false alarms and misses should be incorporated, because for such objects the very notion of a displacement error is ill-defined.

The VGM method, too, has several issues that are the subject of research. For instance, it was mentioned above that one can compute a variogram for nonzero grid points only. One can even compute it for nonzero grid points that have been set to a constant intensity value. All of these variograms measure the covariance structure of the field, but it is unclear what they say within the verification context. After all, they can be interpreted as different summary measures of forecast quality. This ambiguity in summarizing forecast quality also manifests itself in the OF method. For example, as shown here, the joint histogram is suitable for highlighting global spatial shifts in a forecast relative to an observed field. But this may be less useful in realistic situations, because it is unlikely that a forecast field is simply a shifted version of an observed field. Moreover, the joint histogram, by virtue of being a 3D quantity, does not naturally lend itself to presentation on different scales; a scalar summary measure would be more amenable, because one would be able to simply plot it as a function of $W$. All of these issues are currently under investigation.

*Acknowledgments.* Michael Baldwin, Barbara Brown, Chris Davis, Randy Bullock, Tilmann Gneiting, Valliappa Lakshmanan, and Dustin Lennon are all acknowledged for contributing to different portions of this project. We are also grateful to Chris Wikle for reviewing the article and pointing out the benefit of using a correlogram for verification. Partial support for this project is provided by the National Science Foundation (Grant 0513871) and the Office of Naval Research (Grants N00014-01-G-0460/0049 and N00014-05-1-0843).

## REFERENCES

Ahijevych, D., E. Gilleland, B. G. Brown, and E. E. Ebert, 2009: Application of spatial verification methods to idealized and NWP-gridded precipitation forecasts. *Wea. Forecasting,* **24,** 1485–1497.

Baldwin, M. E., S. Lakshmivarahan, and J. S. Kain, 2002: Development of an "events-oriented" approach to forecast verification. Preprints, *15th Conf. on Numerical Weather Prediction,* San Antonio, TX, Amer. Meteor. Soc., 7B.3. [Available online at http://ams.confex.com/ams/pdfpapers/47738.pdf.]

Bowler, N. E. H., C. E. Pierce, and A. Seed, 2004: Development of a precipitation nowcasting algorithm based on optical flow techniques. *J. Hydrol.,* **288,** 74–91.

Brown, B. G., and Coauthors, 2004: New verification approaches for convective weather forecasts. Preprints, *11th Conf. on Aviation, Range, and Aerospace,* Hyannis, MA, Amer. Meteor. Soc., 9.4. [Available online at http://ams.confex.com/ams/pdfpapers/82068.pdf.]

Bullock, R., B. G. Brown, C. A. Davis, K. W. Manning, and M. Chapman, 2004: An object-oriented approach to quantitative precipitation forecasts: Part I—Methodology. Preprints, *17th Conf. on Probability and Statistics in the Atmospheric Sciences,* Seattle, WA, Amer. Meteor. Soc., J12.4. [Available online at http://ams.confex.com/ams/pdfpapers/71819.pdf.]

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. Preprints, *17th Conf. on Probability and Statistics in the Atmospheric Sciences,* Seattle, WA, Amer. Meteor. Soc., J12.5. [Available online at http://ams.confex.com/ams/pdfpapers/70881.pdf.]

Cressie, N. A. C., 1993: *Statistics for Spatial Data.* Wiley Series in Probability and Mathematical Statistics, John Wiley and Sons, 900 pp.

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.

——, ——, and ——, 2006b: Object-based verification of precipitation forecasts. Part II: Application to convective rain systems. *Mon. Wea. Rev.,* **134,** 1785–1795.

——, ——, ——, and J. Halley-Gotway, 2009: The method for object-based diagnostic evaluation (MODE) applied to WRF forecasts from the 2005 SPC Spring Program. *Wea. Forecasting,* **24,** 1252–1267.

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. *J. Hydrol.,* **239,** 179–202.

——, and W. A. Gallus Jr., 2009: Toward better understanding of the contiguous rain area (CRA) method for spatial forecast verification. *Wea. Forecasting,* **24,** 1401–1415.

Everitt, B. S., 1980: *Cluster Analysis*. 2nd ed. Heinemann Educational Books, 136 pp.

Gebremichael, M., W. F. Krajewski, and G. Ciach, 2004: Assessment of the statistical characterization of small-scale rainfall variability from radar: Analysis of TRMM ground validation datasets. *Sixth Int. Symp. on Hydrological Applications of Weather Radar,* Melbourne, VIC, Australia, Australian Bureau of Meteorology, 9 pp. [Available online at http://www.cawcr.gov.au/bmrc/basic/old_events/hawr6/hyrometeorology/GEBREMICHAEL_assess.pdf.]

Germann, U., and J. Joss, 2001: Variograms of radar reflectivity to describe the spatial continuity of Alpine precipitation. *J. Appl. Meteor.,* **40,** 1042–1059.

——, and I. Zawadzki, 2002: Scale dependence of the predictability of precipitation from continental radar images. Part I: Description of the methodology. *Mon. Wea. Rev.,* **130,** 2859–2873.

Gilleland, E., D. Ahijevych, B. G. Brown, B. Casati, and E. E. Ebert, 2009: Intercomparison of spatial forecast verification methods. *Wea. Forecasting,* **24,** 1416–1430.

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

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

Keil, C., and G. C. Craig, 2007: A displacement-based error measure applied in a regional ensemble forecasting system. *Mon. Wea. Rev.,* **135,** 3248–3259.

——, and ——, 2009: A displacement and amplitude score employing an optical flow technique. *Wea. Forecasting,* **24,** 1297–1308.

Lucas, B. D., and T. Kanade, 1981: An iterative image registration technique with an application to stereo vision. *Proc. Imaging Understanding Workshop,* Pittsburg, PA, Carnegie-Mellon, University, 121–130.

Marzban, C., 1998: Scalar measures of performance in rare-event situations. *Wea. Forecasting,* **13,** 753–763.

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

——, and ——, 2008: Cluster analysis for object-oriented verification of fields: A variation. *Mon. Wea. Rev.,* **136,** 1013–1025.

——, and ——, 2009a: Verification with variograms. *Wea. Forecasting,* **24,** 1102–1120.

——, ——, and H. Lyons, 2008: An object-oriented verification of three NWP model formulations via cluster analysis: An objective and a subjective analysis. *Mon. Wea. Rev.,* **136,** 3392–3407.

Politis, D. N., and J. P. Romano, 1994: The stationary bootstrap. *J. Amer. Stat. Assoc.,* **89,** 1303–1313.

——, ——, and M. Wolf, 1999: *Subsampling*. Springer, 347 pp.

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