Verification with Variograms

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

(Manuscript received 29 January 2008, in final form 7 January 2009)

ABSTRACT

The verification of a gridded forecast field, for example, one produced by numerical weather prediction (NWP) models, cannot be performed on a gridpoint-by-gridpoint basis; that type of approach would ignore the spatial structures present in both forecast and observation fields, leading to misinformative or noninformative verification results. A variety of methods have been proposed to acknowledge the spatial structure of the fields. Here, a method is examined that compares the two fields in terms of their variograms. Two types of variograms are examined: one examines correlation on different spatial scales and is a measure of texture; the other type of variogram is additionally sensitive to the size and location of objects in a field and can assess size and location errors. Using these variograms, the forecasts of three NWP model formulations are compared with observations/analysis, on a dataset consisting of 30 days in spring 2005. It is found that within statistical uncertainty the three formulations are comparable with one another in terms of forecasting the spatial structure of observed reflectivity fields. None, however, produce the observed structure across all scales, and all tend to overforecast the spatial extent and also forecast a smoother precipitation (reflectivity) field. A finer comparison suggests that the University of Oklahoma 2-km resolution Advanced Research Weather Research and Forecasting (WRF-ARW) model and the National Center for Atmospheric Research (NCAR) 4-km resolution WRF-ARW slightly outperform the 4.5-km WRF-Nonhydrostatic Mesoscale Model (NMM), developed by the National Oceanic and Atmospheric Administration/National Centers for Environmental Prediction (NOAA/NCEP), in terms of producing forecasts whose spatial structures are closer to that of the observed field.

1. Introduction

Numerical prediction models typically produce a gridded forecast of some quantity. Observations are generally gathered at specific points, not on a grid, and then interpolated to the grid using various numerical or dynamic methods. The problem of assessing the quality of the forecasts, therefore, is equivalent to the problem of comparing two gridded fields, or two digital images. The comparison of the two fields can be done in numerous ways. The simplest methods compare (e.g., subtract) the two images from one another pixel by pixel. This method, however, is inadequate because, for example, it does not reward a forecast for producing the correct structure of the field and penalizes it harshly for not placing the structure in the correct place. Such issues have given rise to a variety of spatial verification methods. A taxonomy of the methods has been attempted by Casati et al. (2008), but an admittedly imperfect classification can be based on the emphasis placed on the spatial covariance structure of the fields. The following is a sample of references where many attributes of forecast performance are provided, but the underlying methods place less emphasis on the covariance structure: Baldwin et al. (2002); Briggs and Levine (1997); Brown et al. (2002), (2004); Bullock et al. (2004); Casati et al. (2004); Chapman et al. (2004); Du et al. (2000); Ebert (2008); Ebert and McBride (2000); Hoffman et al. (1995); Marzban and Sandgathe (2006), (2007), (2008); Marzban et al. (2008); Nachamkin (2004); Roberts (2005); Skamarock (2004); Skamarock et al. (2004); and Venugopal et al. (2005). At the other

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

DOI: 10.1175/2009WAF2222122.1

extreme, some of the main works that place more emphasis on the covariance structure of the fields are those of Gebremichael et al. (2004), Germann and Joss (2001), Germann and Zawadzki (2002), Harris et al. (2001), and Zepeda-Arce et al. (2000). This classification is by no means exclusive; for example, Skamarock (2004) and Skamarock et al. (2004) could belong to the latter group, because they rely on spectral analysis, which is in turn related to the covariance structure of the field. The current work belongs to the second group, because the role played by the spatial covariance structure of the fields is central and explicit. However, as shown below, the variogram can be computed in two distinct ways: one places it more in the spectral domain, while the other conveys information that one may easily consider object oriented.

Although from a meteorological point of view it is perfectly reasonable to view a forecast field in terms of its constituent objects, there are numerous other facets of a forecast that are also important in assessing the quality of the forecasts. At the most basic level, one may compare two fields in terms of the mean of the variable of interest (e.g., precipitation or reflectivity) across the entire field. Another interesting measure is the variance of the variable. Both of these measures are directed at a comparison of the distribution (or histogram) of the two fields. In this line of thinking, the underlying question is whether or not the distribution of the variable across the forecast field is comparable to that observed.

Such a verification method implicitly accounts for a number of summary measures that are relevant to meteorologists, including the mean and the variance of the forecast variable across the field. However, a distribution does not assess how quickly the variable changes from grid point to grid point. In other words, the distribution of the variable conveys no information on the spatial structure of the field. This spatial structure is an important facet of the quality of the forecast field. Also, it is evident that the spatial structure is a concept that is contingent on the spatial scale. For example, a given field may have a complex spatial structure on very small scales and an almost trivial spatial structure on the large scale. Said differently, the comparison of the spatial structure of a forecast field to that of an observed field must be performed within a framework that allows for an exploration of different spatial scales.[1]

In image processing circles, one notion of spatial structure is called texture. It is a measure of the graininess of an image (i.e., field) and assesses how quickly changes occur as a function of distance. In spatial statistics (Cressie 1993; Ripley 1991), a quantity called a variogram effectively gauges texture. Intuitively, and loosely speaking, it quantifies the spatial extent of correlations. If the value of the variable changes between two pixels (i.e., two grid points) in some incoherent way, then one can conclude that the underlying variable has no correlation on the scale separating those pixels. By contrast, if distant pixels vary in some coherent fashion, then one may suspect an underlying spatial correlation that extends to long distances. These ideas are made more formal in the next section, where the defining equation for the variogram is given.

Variograms have already been employed in many meteorological applications, quite apart from verification problems. Given that they appear naturally within the context of interpolation, most applications utilize variograms within that context. For example, Şen (1997) uses variograms as a basis of an interpolation scheme for performing analysis in NWP models. Greene et al. (2002) utilize kriging (Cressie 1993)—wherein one fits variograms—to interpolate wind fields. Germann and Joss (2001) use variograms to find the spatial variation of the precipitation rate in the European Alps, the dependence of this rate on temporal and spatial averaging, and how precipitation measurements from two or more instruments can be compared. Germann and Zawadzki (2002) utilize a correlation function motivated in Germann and Joss (2001) to address the temporal extent to which precipitation is predictable as a nonlinear response in a dynamic model; they, too, rely on variograms for summarizing the spatial structure of the field. Berrocal et al. (2007) use concepts from spatial statistics at large to improve the quality of probabilistic forecasts.

As for verification, the comparison of two fields in terms of their spatial structures has been pioneered by Gebremichael et al. (2004), Germann and Joss (2001), Germann and Zawadzki (2002), Harris et al. (2001), and Zepeda-Arce et al. (2000). The current work is more closely connected with Harris et al. (2001), wherein three methods are described: one based on spectral analysis, one based on a generalized structure function, and another called moment-scale analysis. Of particular relevance to the current work is the second method wherein a generalized structure function for a spatial variable Z is defined as $E[|Z(x + y) - Z(x)|^q]$. The special case with $q = 2$ is equal to the variogram. They focus on $q = 1$ because of three reasons. First, the structure function with $q = 1$ is more robust with respect to outliers in the increment $|Z(x + y) - Z(x)|$. Second, the $q = 1$ structure function allows for scaling relations that are naturally connected to the Hurst exponent in the theory of turbulence (Davis et al. 1996). Finally,

[1] The term scale has different meanings in different fields. Throughout this article, it refers to a physical distance.

the $q = 1$ results are used for performing their third method, that is, the moment-scale analysis. In spite of these arguments in favor of $q = 1$, in this paper the ordinary variogram (i.e., $q = 2$) is employed, and for the following reasons: First, we have confirmed that the distribution of the increments does not suffer appreciably from outliers, and when unambiguous outliers do exist, their effects are tamed because of a resampling procedure described below. Second, within the context of verification, there is no need to relate variograms to scaling relations arising in the theory of turbulence. Finally, this article does not perform the moment-scale analysis.

There is one other difference between the current work and the other aforementioned spatial methods, and that relates to spatial intermittency, or the mixed discrete-continuous nature of reflectivity and precipitation fields (Kundu and Bell 2003; Sapozhnikov and Foufoula-Georgiou 2007). This issue is further addressed in the next section. Here, suffice it to say that the above-mentioned studies perform their analyses on only the portion of the field where the forecast variable is nonzero. For fields like reflectivity or precipitation this is a problem, because the spatial structure of the field is highly sensitive to the inclusion of the zero regions. Yoo and Ha (2007) specifically examine the effects of zero measurements on the correlation structure of rainfall. In the current study, both analyses are performed—with and without the inclusion of the zero regions—because they capture different facets of the quality of the forecasts. As will be shown below, these variograms are capable of revealing that the NWP models generally overforecast the spatial extent of features and that they oversmooth the reflectivity field.

It is worth mentioning that the variogram-based approach does not provide a complete assessment of forecast quality. Although, two different types of variograms are computed here, capturing different facets of performance, other performance measures should also be consulted in order to provide a more complete picture. It does, however, have a few virtues worthy of emphasis. First, as shown below, the variogram can readily identify the overuse of smoothing in NWP models. Also, it can be computed from irregularly placed points (e.g., from the true observations). As such, the variogram of the observations can be computed prior to analysis or interpolation.

In this article, after describing the data, and reviewing the concept of a variogram, the methodology is illustrated on a number of synthetic examples, followed by a verification of 24-h reflectivity forecasts from three NWP formulations: the University of Oklahoma 51-level, 2-km resolution Weather Research and Forecasting (WRF) model (arw2); the National Center for Atmospheric Research (NCAR) 35-level, 4-km resolution WRF (arw4); and the National Oceanic and Atmospheric Administration/National Centers for Environmental Prediction (NOAA/NCEP) 35-level, 4-km NMM (nmm4). A resampling approach is employed to provide a measure of the sampling variation of the variogram. The work ends with a number of conclusions, and a discussion of the underlying assumptions of the proposed methodology.

2. The data

In addition to a few synthetic datasets, also examined is a realistic dataset from the Storm Prediction Center/National Severe Storms Laboratory (SPC/NSSL) Spring Program 2005 (Weiss et al. 2005) involving pairs of observations and 24-h forecasts of reflectivity. The dates span from 19 April to 4 June 2005, and have been interpolated onto a polar stereographic grid with a grid interval of approximately 4 km. Thirty dates are selected, for which a matching pair of observation and forecast exists, and all reflectivity values below 20 dBZ are set to zero, in order to focus on only intense reflectivity (corresponding to significant precipitation events). The dimension of the field is 501×601 grid lengths. An example of the data from one of the days is displayed in Fig. 1, showing the spatial distribution of the observations and the 24-h forecasts according to arw2, arw4, and nmm4, for 13 May 2005. The coordinates of the four corners of the region are 30°N, 70°W; 27°N, 93°W; 48°N, 67°W; and 44°N, 101°W—covering most of the United States from the Rocky Mountains to the Eastern Seaboard. All of the 30 observed fields are shown in Fig. 2, but for brevity the corresponding forecast fields from arw2 only are presented.

Since variograms measure the spatial structure of an image, it is important to discuss how these fields are produced. The observations are a mosaic of the coverage from the National Weather Service radar network. The lowest radar tilt elevation is chosen, providing near-contiguous coverage over the eastern United States. The 24-h forecasts are for model reflectivity at 1-km altitude. The aforementioned 20-dBZ threshold eliminates much of the radar ground clutter from the observations, which is a result of choosing the lowest radar tilt elevation. The radar observations increase in elevation as a function of distance from the radar due to both the angle of the tilt above ground and the curvature of the earth. Meanwhile, the model data are restricted to a single elevation. Restricting the data to significant precipitation via the 20-dBZ threshold should reduce or eliminate any impact this difference could have on the spatial structure of the fields. Weiss et al. (2005), and Baldwin and Elmore (2005) describe the development of the data in more detail.

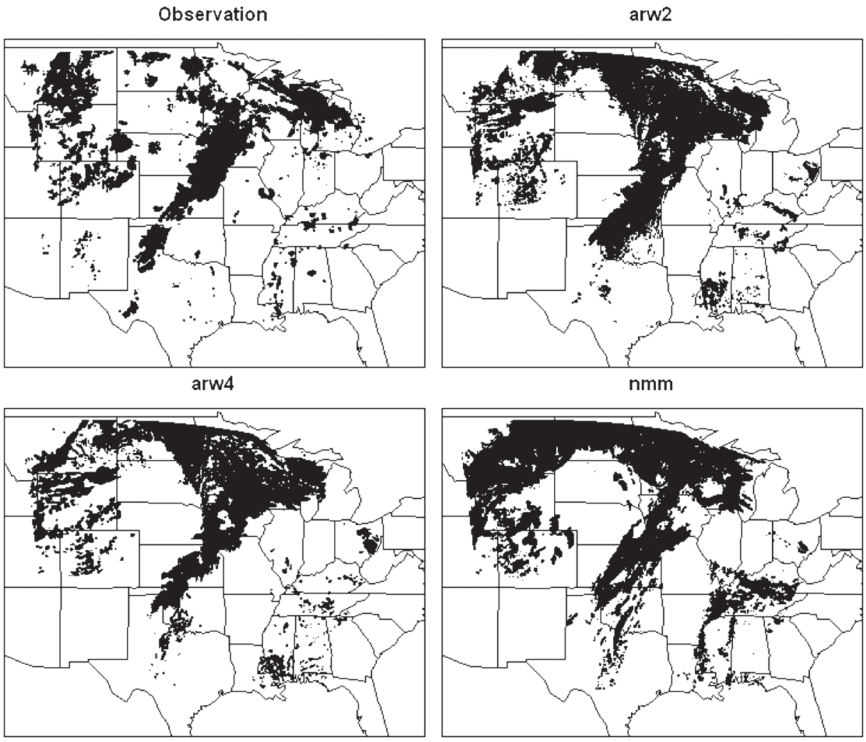

FIG. 1. The spatial distribution of the observation field, and the three forecast fields according to arw2, arw4, and nmm4 for 13 May 2005.

The variogram is sensitive to the resolution of the model or analysis. In the examples considered here, the radar-observed reflectivity smoothed to a 4-km grid is compared to 4-km horizontal resolution models and a subsampled 2-km resolution model. This consistency among the sampling resolutions is critical to variogram interpretation. Also, and more critical, the observed fields are not truly observed; they are all an interpolation of the observations using various algorithms as applied to a grid at a specified resolution. Modern data assimilation systems employ a variety of quality control metrics, as well as dynamical and other constraints, to massage observations into an "analysis." This can have a significant effect on the texture or smoothness of the field and should be carefully considered. This "computational smoothing" (Harris et al. 2001) of the observation field when comparing a model of much lower resolution may be advisable; even smoothing (or subsampling) a very high-resolution model to the scale of the observations may be necessary as one goes to subkilometer model resolutions.[2]

[2] An interesting test would be to compare the variograms of several different operational analyses or forecasts. Variograms would reveal the overuse of smoothing within a model or, more likely, the use of specific algorithms that have a secondary effect of smoothing the forecast field.

3. Methodology

The spatial structure of a field can be summarized in numerous ways, but three common measures are the autocovariance, the autocorrelation, and the variogram (Cressie 1993; Ripley 1991). For a field $Z(x)$, where x denotes spatial coordinates, they are defined as

$$\begin{aligned} &\mathrm{cov}[Z(x), Z(x+y)], \\ &\mathrm{cov}[Z(x), Z(x+y)]/\sqrt{V[Z(x)]V[Z(x+y)]}, \\ &V[Z(x) - Z(x+y)], \end{aligned} \tag{1}$$

where $Z(x + y)$ is the field value a distance y away from x, and V and cov are the variance and covariance operators with respect to some probability density function (pdf). In general, all three quantities depend on x and y; however, if the pdf of the field Z is constant in space—a condition referred to as *stationarity*—then all three quantities depend only on the distance y between points. In that case, they are usually called the autocovariance function, the autocorrelation function (or correlogram), and the variogram, and are plotted as a function of the distance y. There are two other conditions, both weaker than stationarity, that are assumed frequently. The first, called *covariance stationary* (or *second-order stationary*), refers to the condition wherein

FIG. 2. (top) The 30 observation fields, with reflectivity exceeding 20 dBZ, examined here and (bottom) the corresponding arw2 forecasts.

only the first two moments of the pdf are constant in space. Under this condition, all three quantities depend only on the distance y. An even weaker condition, referred to as the *intrinsic hypothesis* (Matheron 1963), refers to when only the variogram depends on y. This is a weaker condition than either form of stationarity because it refers to the variance of the *difference* between variables. It is this more general condition that has led to the variogram being used more frequently than the other two measures. As in the verification of nonspatial fields, the choice of these measures is not unique. Although the other two quantities can also be considered to be verification measures, and may even lead to different conclusions, because of the popularity of the variogram it is selected here as a summary measure of the spatial structure of the field.

Throughout this paper a variogram refers to what is technically an empirical (semi-) variogram, namely an estimate of that mentioned above. Specifically, the estimate used here is the method-of-moments estimator computed as

$$\gamma(y) = \frac{1}{2N}\sum_{i,j}^{N}(z_i - z_j)^2, \tag{2}$$

where z_i is the value of the field measured at some number of locations labeled by an index i. The sum is over all pairs of points a distance y apart, and N is the number of such pairs. A variogram refers to the plot of $\gamma(y)$ as a function of distance y. Given that it is a function of distance, and not a distance from any specific point (like an origin), it summarizes how much the value of the variable Z varies between points, as a function of scale. The points may be regularly or irregularly spaced, but in all of the examples considered in this work, the field is defined on a square grid. Consequently, not all values of y occur in a sample. Generally, some binning is called for in order to produce a reasonably continuous-looking variogram. Two useful properties of the variogram follow from Eq. (2): 1) the variogram of a sum of two fields is equal to the sum of the variograms; as a result, adding a constant to a field, does not affect the variogram, but 2) multiplying the field by a constant leads to the variogram being multiplied by the square of that constant.

In spatial statistics, the general shape of the variogram is often relatively simple: beginning low, rising, and eventually leveling off. For this reason, one often summarizes the variogram with three quantities: the y intercept, the limiting value of the variogram over large scales, and the distance at which that value is obtained; these quantities are called the nugget, the sill, and the range, respectively (Cressie 1993). The nugget reflects the variability at distances smaller than the sample spacing. The sill refers to the maximum variance reached by the variogram, and the range is the distance at which the sill is reached. In principle, measurements separated by distances larger than the range are uncorrelated.

The autocovariance function and the autocorrelation function are closely related to the power spectrum (or variance spectrum, or periodogram) in Fourier spectral analysis. Specifically, the Fourier transform of one is equal to the other. Spectral methods have already been employed for verification purposes (Briggs and Levine 1997; Casati et al. 2004; Skamarock 2004; Harris et al. 2001). For example, Skamarock (2004) and Skamarock, et al. (2004) show that the kinetic energy spectrum of WRF agrees with the observed spectrum. As such, one may expect that verification based on variograms may not add information beyond spectral methods. However, there are at least four reasons that warrant a variogram analysis. 1) The information contained in a variogram is equivalent to that in the autocovariance (and autocorrelation) function only under the condition of stationarity. As mentioned above, the assumption underlying the utility of the variogram is weaker than stationarity or even covariance stationarity. Given that most realistic fields are unlikely to strictly satisfy any of these conditions, the variogram is apt to provide information different from that in a spectral analysis. 2) Another technical reason to examine the variogram is related to its estimation. Cressie (1993, section 2.4) notes that the estimate of the autocorrelation function is sensitive to bias and trends in the field. Furthermore, that estimate is sensitive to deviations from the normality of Z. The estimator for the variogram $\gamma(y)$ depends only on the differences in the field values and, so, is less affected by such deviations. 3) In a verification setting, variograms are more natural because they make the spatial scale explicit, without the need to perform any Fourier transformation into frequency space. 4) Even in applications unrelated to verification, variograms generally provide different information than power spectra. For example, Maillard (2001) compares and contrasts the two methods for texture classification and finds that the variogram approach slightly outperforms the Fourier spectral method. Mela and Louie (2001) use both methods for estimating correlation length and fractal dimensions; they argue in favor of using both methods because of the added interpretability of the data.

As mentioned in the introduction, an important issue that arises in modeling reflectivity or precipitation fields is the spatially intermittent nature of the field, reflected in a distribution that has a delta-function peak at zero. Nonzero precipitation can be well modeled with a

lognormal distribution, but a realistic forecast field consists of many grid cells at which the precipitation is zero. Many methods have been proposed for addressing the mixed discrete-continuous nature of such fields (Kedem et al. 1990; Barancourt et al. 1992). The latter work proposes an idea that is more conducive to the approach adopted here, because it relies directly on variograms. Specifically, first, one approximates the field with a binary field, where any nonzero precipitation is replaced by a constant. In spatial statistics, this type of categorical variable is referred to as an indicator variable (Journel 1983). The interpolated field, then, models only the intermittent component of the field; through interpolation (kriging), it highlights regions where the precipitation is nonzero. Finally, the variability of the nonzero precipitation values is then assessed by examining the spatial structure of the fields within these regions. In this way, the authors refer to the variability captured in the latter as the ''inner variability.''

This two-step procedure is computationally intensive and is worthwhile if modeling the field is the focus of the work. Even for verification purposes, this decomposition of the variability into ''inner'' and ''outer'' is important, but only in diagnosing–decomposing the forecast errors. However, in a situation where several forecast models are to be compared, objectively and automatically (without human intervention), it is more desirable for a performance measure to provide a more all-embracing assessment of the forecast quality, beyond size and displacement errors. To that end, a variogram is computed across the entire field (i.e., zero and nonzero grid points). Such a variogram would be sensitive to size, shape, displacement, and intensity errors. By contrast, one can compute the variogram across only the nonzero grid points. In this work, both sets of variograms are computed: one across only nonzero grid points and another across the entire field. The former assesses the texture of the field and is useful for comparing the spatial properties of the two fields, but it does not include size and displacement errors. The second set of variograms does include that information.

To gain some sense of the sampling variation of the variograms, two resampling approaches are adopted. One is a standard bootstrap approach where samples of the same size as the original sample are taken, with replacement, from each field, and the variogram is computed. The number of bootstrap trials is 40, but it has been confirmed that a number of trials as small as 10 yields similar results. This bootstrap approach is used when the variogram is computed across only nonzero portions of the field. To apply the same procedure to the entire field is possible but computationally infeasible; instead, samples of size 50 000 are taken from each field, without replacement, again 40 times. Technically, neither approach yields proper confidence intervals, precisely because of the spatial structure of the fields. There exist many variations of the standard bootstrap that do allow for spatial (and temporal) correlations in the data; subsampling (Politis et al. 1999), and the block bootstrap (Politis and Romano 1994) are but two. However, the extra effort is not worthwhile in the current application, because confidence intervals in the strict sense of the concept are not necessary; the purpose of the intervals produced here is simply to convey a general sense of the variability. Suffice it to say that proper confidence intervals would be wider than any computed here. In fact, in this work, the sampling variation of the variograms is displayed through box plots, rather than individual intervals. This way, one obtains a more complete picture of the sampling distribution of the variogram.

Finally, given that the primary aim of this study is model comparison, for the purpose of selecting the ''better'' model, it is important to decompose the sampling variation reflected in these box plots into ''within day'' and ''between day'' components, because it is the latter that is important in selecting the better model.[3] This is further explained in section 5.

4. Synthetic examples

It is instructive to consider a few synthetic examples in order to develop an intuition for the type of variogram associated with a given field. The top panel in Fig. 3 shows a background consisting of a random field with an ''object'' introduced at a specific location. The background is an uncorrelated field generated by a sequence of floating-point random numbers uniformly distributed between 0 and 1. The ''object'' is a bivariate Gaussian with parameters $\mu_x = \mu_y = 30$, $\sigma_x = \sigma_y = 30/4$, $\rho = 0$, contaminated by the same fluctuations as the background. The variogram associated with the field alone (excluding the object) coincides with a horizontal line with a y intercept given by the variance of the variable across the entire field; this variogram is shown in the bottom panel of Fig. 3 as a dashed horizontal line. The nontrivial variogram also shown in the bottom panel (as a sequence of circles) is that associated with

[3] This terminology is based on the standard terminology in the analysis of variance, where the variance in a nonhomogeneous sample is decomposed into its within-group and between-group variances.

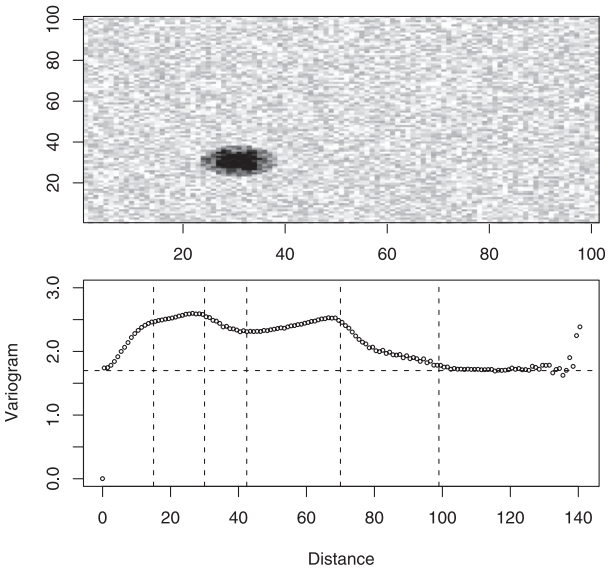

FIG. 3. (top) A synthetic field consisting of a uniform background and a well-delineated object. (bottom) The variogram of the uniform field (dashed horizontal line) and that of the whole field including the object. The vertical lines are described in the text.

the field including the object. As mentioned above, the bin size for the quantity denoted "distance" is chosen so as to produce a reasonably continuous curve.[4]

Several features are noteworthy. First, note that the object is approximately 15 grid lengths (or pixels) across, and it is centered at (30,30). From left to right, the first vertical line is at 15, that is, the size of the object, where one can see a slight bend in the variogram. The next bend is at 30, that is, the location of the object. The next line is at $30\sqrt{2}$, which is the distance between the center of the object and the origin. The line at 70 also corresponds to a bend in the variogram; note that $70 = 100 - 30$ is the distance between the center of the object and the right-most side of the field. Finally, the distance between the object and the upper-right corner of the field is $70\sqrt{2}$, and this line is the right-most vertical line drawn in the variogram plot. At each of these values, corresponding to either the size of the object or its distance from the various sides of the field, the variogram displays some sort of a bend. In other words, the shape of the variogram is determined by the size and location of the objects in the field. Needless to say, the size of the grid can also affect the variogram.

[4] The role of the bin size in variograms is exactly the same as the role of the bin size in histograms. Its determination is part science, part empirical.

This variogram can be explained intuitively by imagining a stick of some length sliding across the field. For a small stick, that is, on small scales, the variogram tends to increase with distance because as the stick gets larger the typical size of $(z_i - z_j)^2$ increases. However, the rate of change (i.e., the slope of the variogram) is affected when the stick is large enough so as to not fit entirely inside the object. On such scales, at least one end of the stick is outside the object, and so, the spatial coherency of the object causes the variogram to increase at a lower rate, hence, the bend at 15. Other changes occur when the stick is sufficiently large to extend from the object to the edges of the field—at 30 and 70. On these scales the variogram decreases with increasing distance because relatively more sticks have both ends outside of the object. As such, the variogram is driven toward the variogram of the background (i.e., the horizontal line). That convergence is complete when the stick is larger than the largest possible stick that allows for no end of the stick to be within an object, namely at $70\sqrt{2}$. The slightly increasing trend between $30\sqrt{2}$ and 70 arises because the size of the object allows for one end of the stick to be inside the object, thereby partially compensating for the convergence to the background value. In short, whereas a smooth continuous field might generate a smooth variogram (usually increasing with distance), an object placed within the boundaries of a square box will generate variograms resembling that in Fig. 3. Although, it is not shown here, a larger object tends to produce a variogram that extends farther in the vertical direction.

As mentioned in the previous section, the nugget, the sill, and the range are often sufficient to capture the general shape of the variogram. For the variogram in Fig. 3, however, all of these quantities are ambiguous. The nugget may be considered to be at 0 or at 1.7 (the horizontal line). The sill may be considered to be at 2.5 where the variogram reaches its peak, or 1.7 where it levels off. Consequently, the range is also completely ambiguous. These ambiguities arise because of the mixed discrete-continuous nature of the field.

Although many realistic fields have objects within them, the actual nature of the fields is more continuous than a disc placed against a uniformly distributed background. More realistic examples are offered by considering a random Gaussian field. Such a field is more realistic, because it is still a random field, but the field itself displays some nonrandom structure, mimicking the object of the previous example. Figure 4 shows one such field (top-left panel; this field is generated by GaussRF, a function in an R package called RandomFields, which is freely available online at http://www.r-project.org). The corresponding variogram is shown in

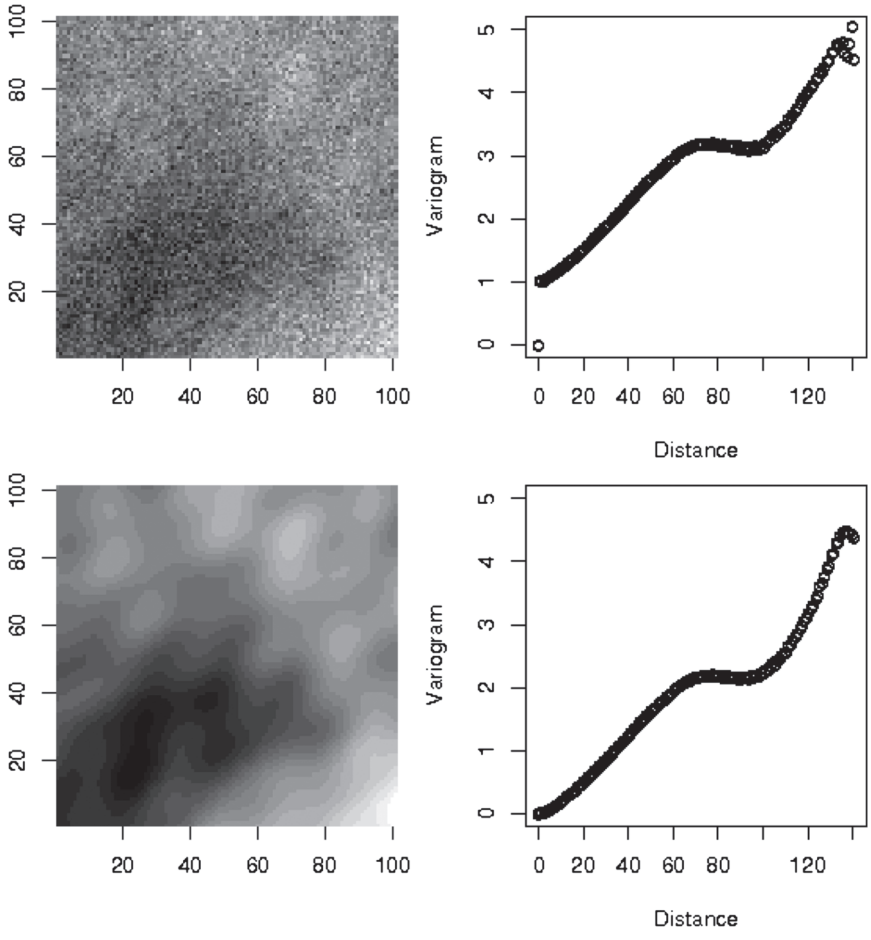

FIG. 4. Examples of (left) random Gaussian fields and (right) the corresponding variograms. The top field has a nugget of 1, while the nugget is zero for the bottom field. Lower nugget values are generally associated with smoother fields, and vice versa.

the top-right panel of Fig. 4. As expected, the variogram generally increases with distance, with some humps and bumps caused by the structure of the field.

A random Gaussian field is defined in terms of numerous parameters, all of which visually affect the field and the corresponding variogram. For example, the top field in Fig. 4 is generated with mean = 0, variance = 4, and nugget = 1. The bottom field in Fig. 4 shows the same field but with the nugget parameter set to 0. The visual effect is an overall smoothing of the field. This change in the nugget is also reflected in the y intercept of the variograms. In other words, increasing the nugget generally produces more texture in the field, resulting in an upward shift of the variogram, without affecting its overall shape. It turns out that this type of behavior is, in fact, seen in the real data examined in this paper; for some days, the variogram of the observed field is simply a shifted version of that of the forecast field. In other words, the forecast and the observed variograms differ only in their nuggets. The other parameters defining a random Gaussian field affect the visual appearance of the field but in ways that are not easily describable.

As a final synthetic example, consider the situation depicted in Fig. 5. The top-left panel in Fig. 5 shows two circular objects with a random Gaussian field interior, placed against a background of zero values. The top-right panel in Fig. 5 shows the same image shifted by 15 grid lengths in x and y. One may treat these two fields as an observed and a forecast field, respectively. The middle row shows the corresponding variograms computed across the entire field (including the zero background); collectively their shape resembles that in Fig. 3 and can be explained in the same manner in terms of the size and location of the objects. The panels in the bottom row in Fig. 5 display the variograms computed over only the nonzero portion of the fields, that is, over the objects. These two variograms are identical, as they

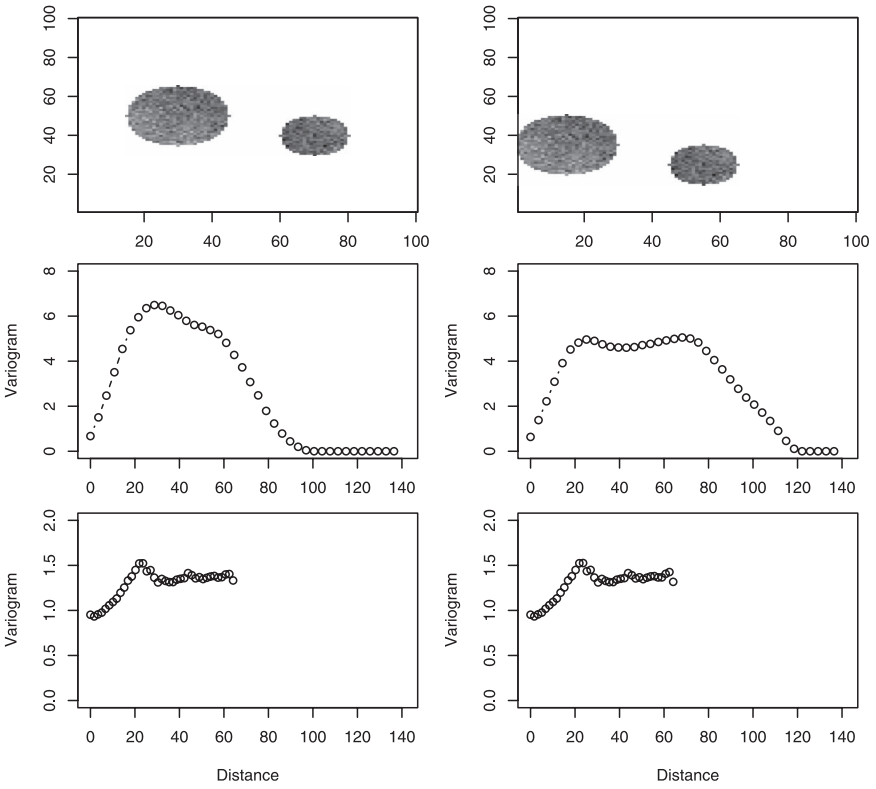

FIG. 5. Examples illustrating the effects of a displacement error on the two types of variograms. (top left) Two objects with a random Gaussian field interior, placed against a zero background. (top right) The same image but with the objects shifted by 15 grid lengths in the x and y directions. The corresponding variograms are shown in the second row (when it is computed across entire field) and in the bottom row (when it is computed across only nonzero grid points).

should be because they assess the texture of the field—a quantity invariant under translation. As such, although these variograms are measuring some facet of performance, they do not reflect displacement errors. By contrast, the middle variograms do. This example illustrates the utility of computing both types of variograms. One is more useful for assessing texture only, while the other measures some combination of texture as well as size, shape, and displacement errors. Clearly, for an objective and automatic verification of a large set of forecasts, the latter is more appropriate, for it is sensitive to more types of errors.

5. Results

We consider 13 May 2005 first. On this day, a large, occluded frontal system crosses the Midwest approaching the Great Lakes. Figure 6 shows the variograms for the observed field and for the three forecast fields (in three colors), computed across the entire field (top), and only the nonzero grid points (bottom). Each variogram is actually a sequence of box plots, reflecting the sampling variation of the variogram. They are produced by taking 40 random samples from the fields (as described in section 3) and producing the variograms for each sample. Instead of plotting 40 different variograms, the box plot of the variogram is plotted for every value on the x axis (i.e., scale).[5]

In the top panel in Fig. 6, the general behavior of the variogram for the observed field (black) is consistent with the synthetic examples given above. The variogram initially rises but only up to a scale of about 1000 km, at which point it begins a decreasing trend toward the background value. This behavior is consistent with the existence of an ''object'' against a background field, as

[5] The line within a box gives the median of the 40 variograms, the top and bottom sides of the box denote the third and first quartiles, respectively, and the whiskers display the range of the values. In this work, the width of the box plots is fixed, in the sense that it does not convey any information.

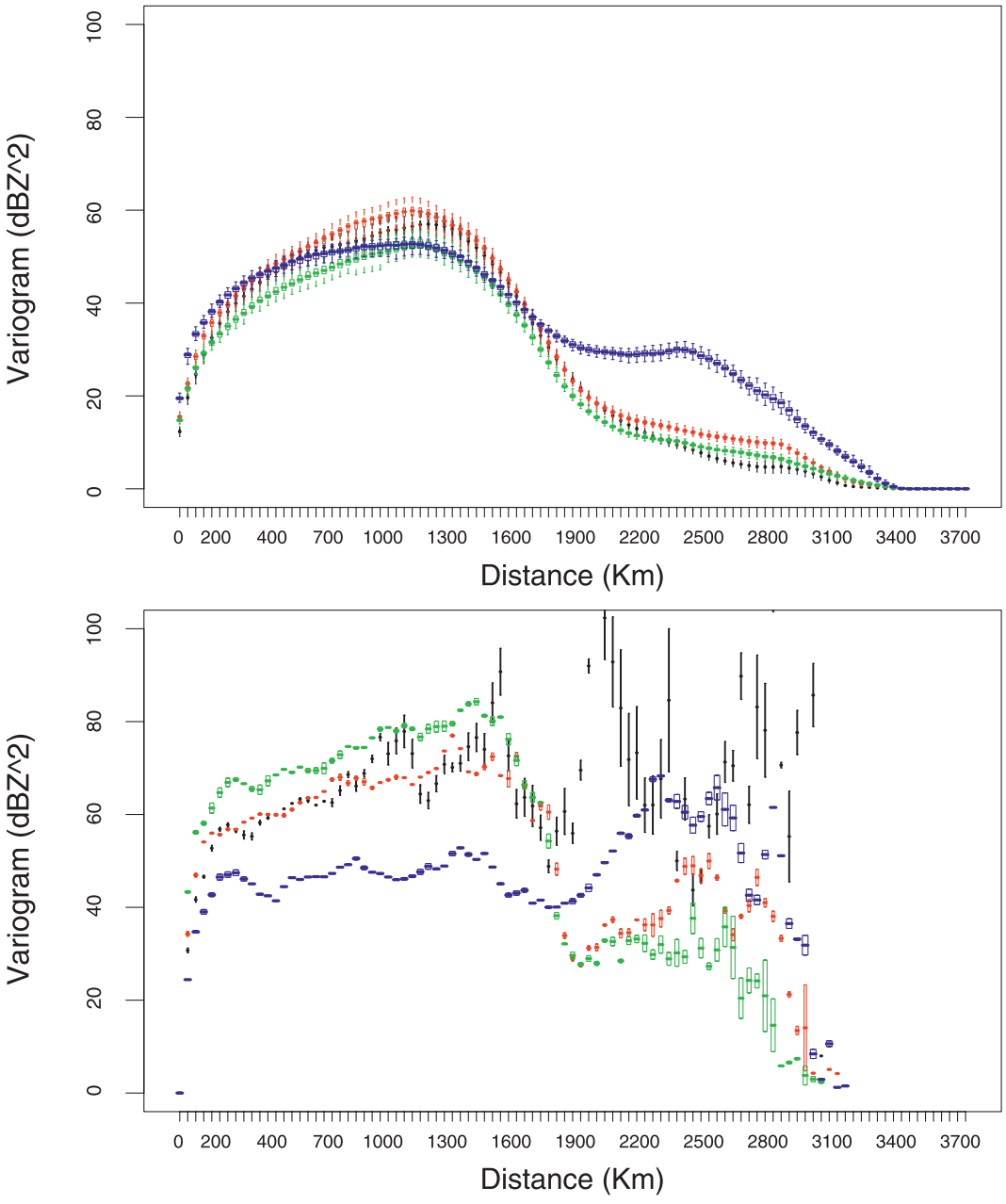

FIG. 6. (top) The variogram for the observed field (black), and the arw2 (red), arw4 (green), and nmm4 (blue) simulations. The box plots displaying the sampling variation are based on 40 resamples. (bottom) As in the top panel but here the variograms are computed only for grid points with nonzero value.

confirmed by examining the actual field (Fig. 1). The background is clearly noiseless, with all grid values set to zero (a consequence of the reflectivity threshold at 20 dBZ), and the object is the large frontal occlusion over the Midwest. The peak of the variogram at distances of the order of 1000 km suggests that the system is approximately of that size, or that it is about 1000 km away from the boundary of the field. This ambiguity may be troublesome if one desires to diagnose the variogram unambiguously; however, loss of information is an unavoidable consequence of any summary measure, and the variogram is no exception.

The arw2 simulation (red) produces a variogram similar to that observed. However, there are regions in the variogram where there is no overlap between the box plots of the two variograms. Specifically, on short scales (500–1000 km), and on long scale (>2000 km), the arw2 variogram is higher than observed. By contrast,

arw4 (green) produces a lower variogram on short scales. nmm4 (blue) yields a similar variogram to arw4, but only on smaller scales (<1400 km); on larger scales, the variogram for nmm4 is clearly much higher than that of the observed field, or of arw2 and arw4. One may be tempted to attribute the higher variogram of the nmm4 simulation on larger scales to the lower resolution of the model (4.5 versus 2 km for arw2); however, that explanation would not be justified because another low-resolution model, namely arw4, yields a variogram comparable to both that of the observed field and of arw2. In this case, the correct explanation of the difference can be seen in the actual forecast field; the nmm4 forecast in Fig. 1 clearly has a more solid (i.e., more area exceeding the 20-dBZ threshold) feature in the northwest region of the occlusion and is much more scattered in the observed field as well as in the arw2 and arw4 forecasts. It is this feature that gives rise to the discrepancy on large scales. One may also wonder if the higher value of the variogram for nmm4 on larger scales is due to a larger extent of the reflectivity exceeding 20 dBZ because, as mentioned in section 4, a larger object can produce a higher variogram. However, as explained in the next section, that fact does not entirely explain the higher variogram values, because all three models produce larger areas of high reflectivity than that observed.

The bottom panel in Fig. 6 shows the variograms when they are computed only over nonzero grid points. Again, the overall shape of these variograms is consistent with that seen in the synthetic example (Fig. 5, bottom row). Note that the three conventional summary measures for a variogram—nugget, sill, and range—are less ambiguous in these variograms, especially for nmm4; although the three forecasts appear to have different sills, their range is of the order of 200 km. In other words, reflectivity values beyond 200 km are mostly uncorrelated. More importantly, whereas nmm4's variogram computed across the whole field (Fig. 6, top) is comparable to that observed on smaller scales, according to the bottom panel in Fig. 6, it tends to oversmooth the field, much more so than in the arw2 and arw4 simulations.

It is worth pointing out that the box plots at longer distances are larger mostly due to smaller samples. Given a finite field, there exist more grid point pairs that are close to one another than those farther away. In fact, on the largest scale (e.g., ~3700 km) the variogram is based on only the values of reflectivity on the boundary of the field. This places a restriction on the largest scales, which can be reasonably assessed. The size of the box plots is a visual reminder of that scale. Moreover, the variogram at larger scales is apt to be most affected by factors beyond sampling. For example, forecasts near boundaries often have different physical characteristics as compared to those within a field. Also, the three NWP formulations considered here have different physical properties in the way the forecasts are generated near the boundaries. For these reasons, the variogram for large scales should be interpreted with caution.

Although other qualitative differences exist between the variograms in Fig. 6—and their diagnosis can be useful—from a model verification–selection point of view, the more relevant question is whether the above-noted features persist across multiple days. To answer that question, variograms are produced for all 30 days. The resulting individual variograms are not shown here, because our main interest is in a comparison of two variograms—based on a forecast and an observation. Then, it is natural to examine the *differences* between variograms. Figure 7 shows the differences between variograms computed across the entire field: in black for arw2 observation, red for arw4 observation, and green for nmm4 observation.

Evidently, there is a great deal of variation in the shape of these variograms. A perfect forecast field would produce a curve overlapping the x axis. Given that the underlying variograms are computed across the entire field (including zero grid points), deviations from the x axis reflect errors in texture as well as in size and location. According to Fig. 7, however, the ideal situation does not arise for any of the curves for all 30 days. In other words, none of the models produce variograms that are consistent with observations across all 30 days. On some days (e.g., 18 May 2005), all three curves are relatively close to the horizontal line at 0, suggesting that all three models produce high quality forecasts. On other days (e.g., 10 May 2005), none of the models produce a variogram comparable with the observed variogram, on any scale. There exist days (e.g., 11 May 2005) when the agreement with observed variograms is near perfect on small scales, but abysmal on larger scales. Finally, although the models generally appear to produce higher variograms than observed, there are a few exceptions—most notably on 26 April on larger scales, when all three models produce large negative differences. For 28 May, on larger scales, arw2 and arw4 produce negative differences, but nmm4 produces positive differences. In general, for most of the 30 days, arw2 and arw4 produce comparable variograms, and nmm4 produces variograms larger than that observed. This suggests that nmm4 may be producing either larger objects, or generally higher reflectivity values, than arw2 or arw4.

Again, such figures can be employed to examine the forecasts for a given day. However, the box plots in

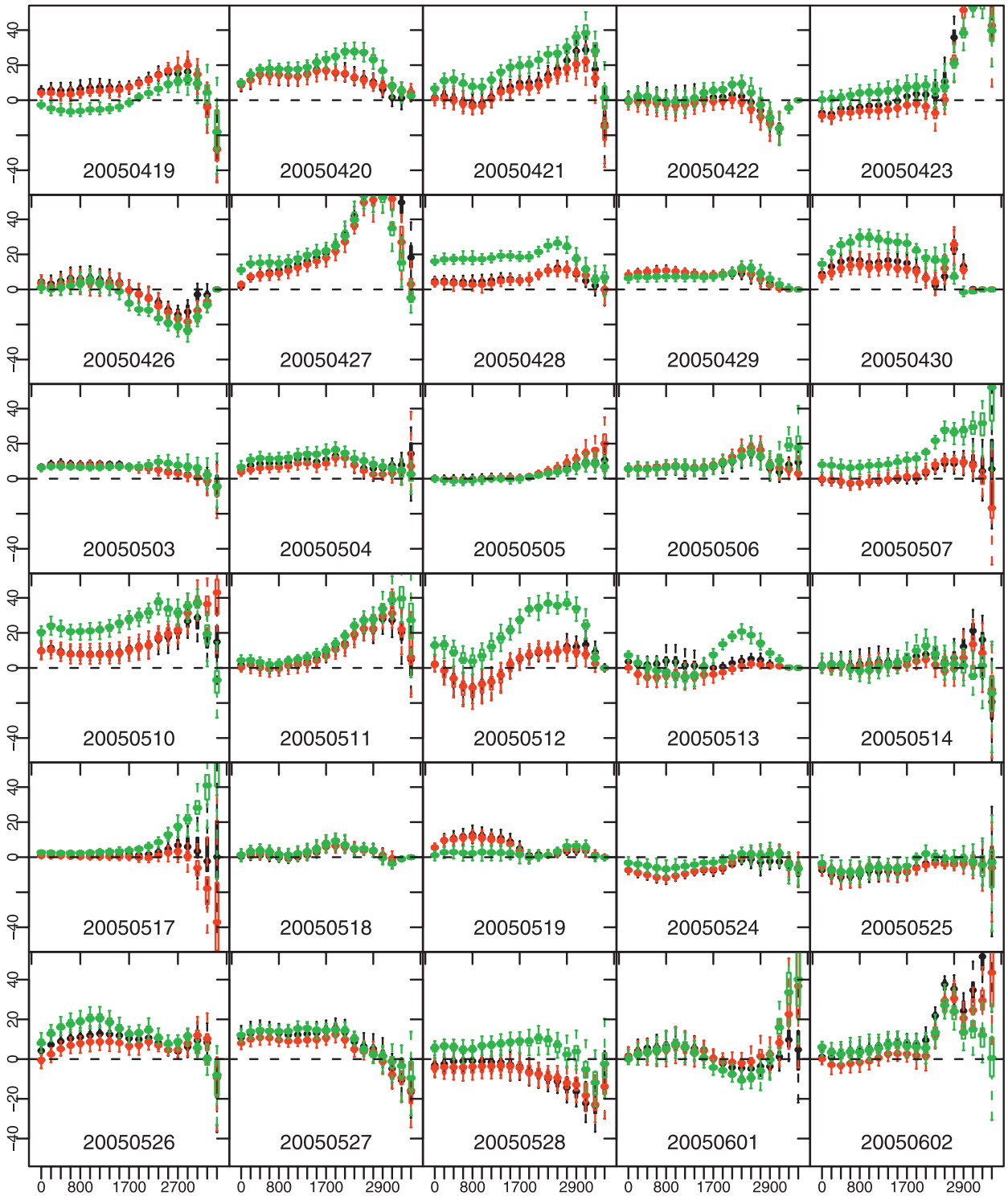

FIG. 7. The differences between variograms: arw2-observation (black), arw4 observation (red), and nmm4 observation (green).

Fig. 7 display the within-day variation of the differenced variogram. For the purpose of comparing the three models, it is more important to examine the between-day variation of the differenced variograms. To that end, we examine the distribution (across 30 days) of the median of the differenced variograms, for a given scale. Figure 8 shows a series of box plots summarizing that distribution, each computed from 30 medians

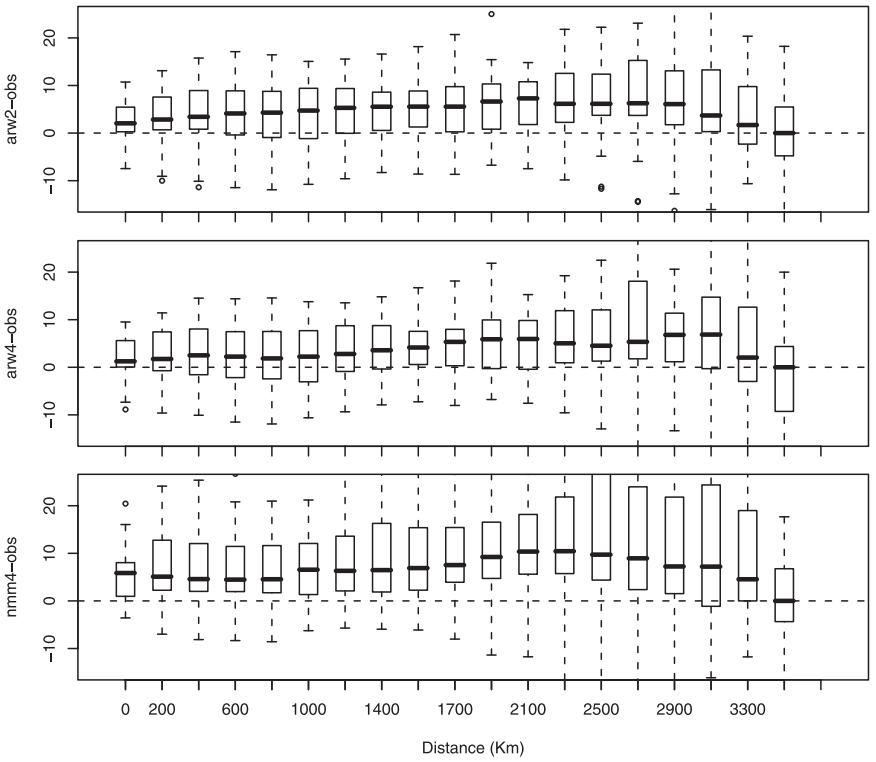

FIG. 8. The variation in the differences between variograms between days: (top to bottom) arw2 observation, arw4 observation, and nmm4 observation, respectively. The variograms are computed across the entire field.

(1 day^{-1}), versus distance.[6] An ideal model would produce a series of box plots with a significant overlap with the dashed line at zero. In other words, the quantity on the y axis measures forecast error. From top to bottom, the three panels in Fig. 8 refer to arw2 observation, arw4-observation, and nmm4-observation. According to Fig. 8, all three models generally produce higher variograms than observed. One explanation is that the forecast extent of reflectivity exceeding 20 dBZ is larger than that observed, for all three models; that is, they produce more large spatially coherent features. On smaller scales (<1700 km), the differences between the model and observed variograms are not statistically significant; however, at larger distances (>1400 km), the differences are more significant, especially for nmm4. The generally wider box plots for nmm4 also suggest that it produces a wider range of forecasts with more diverse errors. On the largest scales possible in this study, namely the size of the field itself (~3500 km), all three models agree with the observations, but that only means that the three models produce forecasts whose variances across the entire field are consistent with that observed. This is not too surprising, given the 20-dBZ threshold adopted in this study.

Figures 7 and 8 are based on variograms computed across the entire field. The analog of Fig. 7 for variograms computed across only nonzero grid points is not shown. Figure 9 shows the error measure (analog of Fig. 8) when variograms are computed across only nonzero grid points, that is, measuring texture errors. Given that all the box plots cover the x axis, all three models generally capture the texture of the observed field. On other hand, based on the fact that the median errors (middle line of the box plots) are generally below the x axis, it follows that all three models oversmooth the field. The exceptions are at the largest scales, where arw4 and nmm4 appear to produce forecasts that are coarser than observed. However, as mentioned previously, such conclusions regarding large scales should be treated with caution.

6. Supplementary analysis

According to Fig. 8, the variogram of the forecast field (computed across the entire field) is generally higher

[6] To be more specific, the between-day variation is computed as follows: For a given value of the x axis (i.e., for a given scale), the medians of the box plots across the panels in Fig. 7 are aggregated, and their distribution is summarized as a box plot. It is these box plots that are shown in Fig. 8.

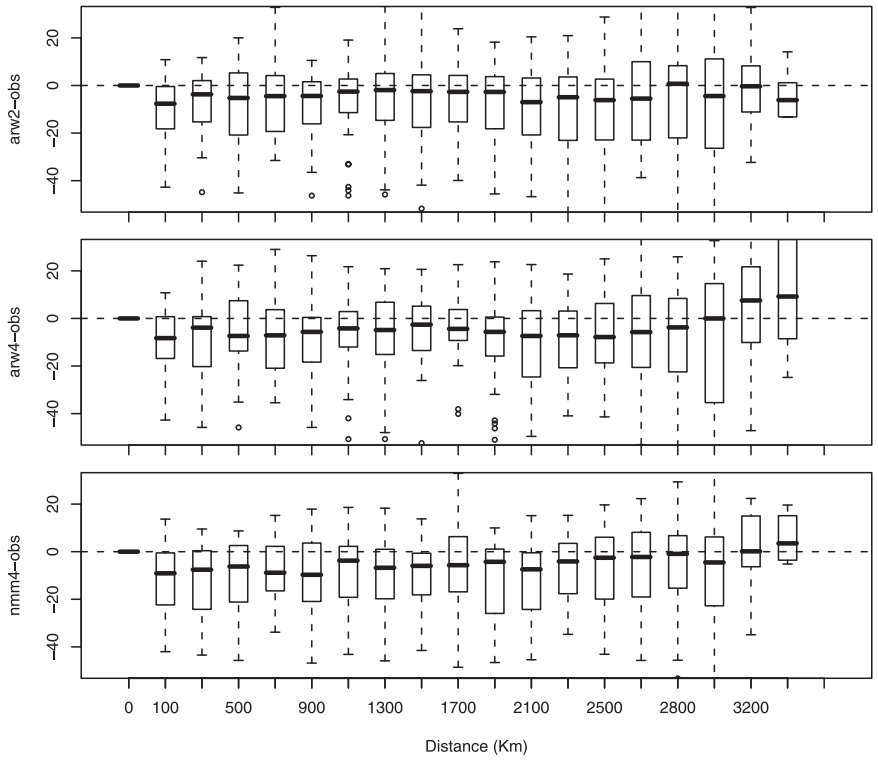

FIG. 9. As in Fig. 8 except that the variograms are computed only across nonzero grid points.

than that of the observed field. It is possible that this is an artifact of the observed field, and not of the forecast. What has been called the observed field is not truly observed; it is a result of a number of analyses, all of which affect the structure of the field. For example, reflectivity data are generally noisy and are frequently smoothed during analysis, thereby explaining the discrepancy with the forecasts. However, the relative performance of the three models (among themselves) remains unaffected by any changes affecting the observed field. In other words, the proposed methodology can still assess the quality of a set of forecasts in a relative way, as well as indicating differences between the forecast and analyzed data.

Another question concerns the higher variograms associated with nmm4 as compared to arw2 and arw4. For the case of 13 May 2005, that discrepancy was explained above by examining the actual forecast and observed fields and noting that the frontal occlusion on that day is much larger, and more continuous, in the forecast of nmm4, one which is much less pronounced in the forecasts of arw2 and arw4, at least in terms of area exceeding 20 dBZ. However, for a majority of the 30 days, nmm4 produces variograms (across the whole field) that are higher than those of arw2 and arw4. As noted in section 4, higher variograms can be produced by larger objects. It is very likely that nmm4 produces forecasts with larger (and more dense) precipitation areas than arw2 or arw4.

To verify this, statistical tests are performed on three simple summary measures: the percent of grid points with reflectivity exceeding 20 dBZ (called coverage), the mean of the reflectivity across all grid points in the field (called mean reflectivity), and the mean of the reflectivity over only the grid points whose reflectivity exceeds 20 dBZ (called alternative mean reflectivity). Figure 10 shows forecast versus observed scatterplots of these three quantities for the three models. From the top row in Fig. 10, the preponderance of points above the diagonal line suggests that all three models produce reflectivity exceeding 20 dBZ over a larger number of grid points than is observed. However, it is also evident that nmm4 produces more coverage than either arw2 or arw4.

The second row in Fig. 10 shows that the three models also produce higher levels of mean reflectivity than is observed. Again, nmm4 produces higher levels of reflectivity across an entire forecast field than either arw2 or arw4, which may simply be due to its overforecasting of the area of reflectivity.

The coverage and mean reflectivity are overestimated by the three models, and so it is beneficial to examine the forecasts in more detail. Examining the actual observations and model forecasts for the 30-day experiment

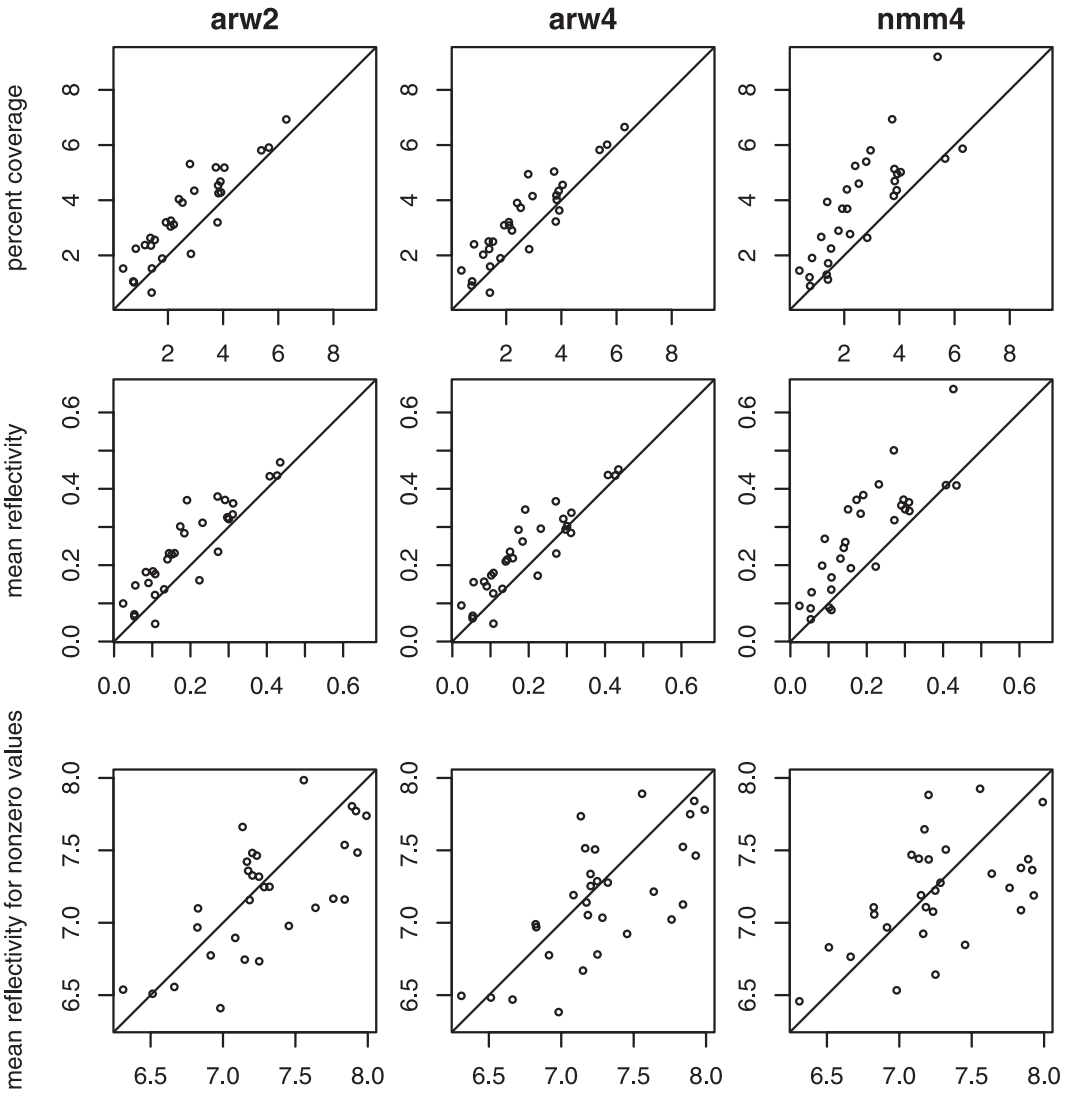

FIG. 10. Scatterplots of (top) forecast vs observed percent coverage, (middle) mean reflectivity across the entire field, and (bottom) mean reflectivity across only nonzero grid points. The three columns refer to (left) arw2, (middle) arw4, and (right) nmm4.

(not shown), it is evident that all three models overforecast the extent of the nonzero reflectivity. When examining the alternative mean reflectivity (bottom panel in Fig. 10) where only points with nonzero reflectivity are included, one notes that, even though nmm4 has higher coverage than arw2 and arw4 (top-right panel in Fig. 10) and higher mean reflectivity, the reflectivity based only on these points is generally less biased than those from arw2 and arw4. In short, all three models are overforecasting the extent of the reflectivity.

To quantify these findings, a number of statistical tests are performed. A paired two-sided *t* test performed on the 30 forecasts' coverage values and the corresponding observed values yields p values <0.001 for all three models. A similar test performed on the mean reflectivity results in similarly small p values. As such, the evidence provided from the data suggests that the models do not produce the observed coverage or mean reflectivity across 30 days. In fact, the models generally produce higher values of coverage and mean reflectivity than is observed. A similar test performed on the alternative mean reflectivity yields nonsignificant p values, suggesting that the data are consistent with the models in terms of their alternative mean reflectivity. All of these conclusions are consistent with the panels shown in Fig. 10. In addition, they suggest that the higher variogram values produced by nmm4, as compared with arw2 and arw4, are due to the higher coverage of the

reflectivity produced by nmm4. This conclusion is strengthened by the errors shown in Fig. 9, because they preclude errors in the texture of the forecasts from being the culprit.

7. Conclusions and discussion

Forecast and observation fields are compared with respect to their spatial structures, as summarized by their variograms. Two types of variograms are computed: one sensitive to texture, as well as size, shape, and displacement errors, and the other sensitive only to texture. The fact that the first type of variogram is sensitive to the size and displacement error qualifies it as an object-oriented verification measure. Similarly, given the affinity between the second type of variogram and the texture, it can be considered to be a spectral method. Although these two variograms provide a useful assessment of the forecast quality in terms of smoothness, they complement other, more standard measures such as accuracy.

A resampling framework is set up to assess the sampling variation of the variogram, thereby allowing comparisons of different forecast models. The framework is then employed to compare forecasts from three different models with radar-based observations of reflectivity. It is found that arw2 and arw4 produce highly similar structures, both different from that observed; nmm4 creates a spatial structure that is least similar to that of the observed field. More specifically, it is found that 1) the three models (especially nmm4) overforecast the spatial extent of features, as shown by the variograms computed on zero and nonzero values (Fig. 8); 2) all three models oversmooth the reflectivity field, as shown by the variograms computed on the nonzero values only (Fig. 9); and 3) additional analysis of more traditional measures confirms that the models generally produce larger coverage areas and mean reflectivity than is observed.

In addition to the questions addressed in the previous section, there is another question that is almost hypothetical, but worth asking. Many interpolation schemes (e.g., kriging) involve fitting the variogram with some theoretical model, first. The variogram model is then employed to develop the interpolating model. As such, a forecast field that has been smoothed in this way will necessarily yield higher quality forecasts, in terms of variograms, when compared with forecasts that have been smoothed in some variogram-independent fashion. Since the spatial structure is an important facet of forecast quality, this suggests that one should incorporate variograms into the analysis phase of NWP modeling, as is proposed by Şen (1997) and Greene et al. (2002).

The above analysis makes a few assumptions whose removal points to future research. For example, it has been assumed that the fields are isotropic. This is clearly not true in the fields examined here, because the atmosphere is not isotropic. One way to remove this assumption is to use directional variograms. It has also been assumed that the variogram is constant across the entire field. Although it is more difficult to assess the validity of this assumption, it will be interesting to allow for the variogram itself to vary as a function of region. Of course, this complicates the verification task, but it is entirely arguable that verification should be performed in a region-dependent fashion anyway. After all, it is possible, if not likely, that one model outperforms another model in one specific region, but not in another, such as more stratiform midlatitude systems or more convective tropical systems. One other assumption is that knowledge of performance assessed in terms of variograms may aid a model developer in altering the model for the purpose of improving its forecasts in terms of variograms. To address this assumption, it will be interesting to discover which model parameters affect variograms. Finally, given that a variogram can be computed for a sample of irregularly placed points, it is possible to compute it for ''raw'' observations (prior to analysis or interpolation), and compare it with the variogram computed from gridded forecasts; in this way, any concerns regarding undesirable smoothing of the observations can be precluded. All of these issues are currently under investigation.

Finally, one may wonder if for verification problems, the variogram has utility only for high-resolution models. The variogram assesses the spatial structure of a field on all spatial scales, from the grid spacing to the size of the field itself. The resolution of a model simply manifests itself as a (lower) cutoff on the x axis of the variogram. In other words, the variogram can be computed for all length scales (i.e., at any resolution), but its values for length scales nearing model resolution should be interpreted with caution. In general, this methodology can be used on forecasts from any model and allows for assessing yet another facet of forecast quality.

Acknowledgments. Michael Baldwin is acknowledged for providing the data employed for the analysis in this paper. Don Percival and Tilmann Gneiting are thanked for numerous discussions related to spatial statistics and for pointing us to the R package GaussRF, which was used in a number of the spatial statistical procedures performed in this article. Partial support for this project was provided by National Science Foundation Grant 0513871 and Office of Naval Research Grant N00014-01-G-0460/0049.

REFERENCES

Baldwin, M. E., and K. L. Elmore, 2005: Objective verification of high-resolution WRF forecasts during 2005 NSSL/SPC Spring Program. Preprints, *21st Conf. on Weather Analysis and Forecasting/17th Conf. on Numerical Weather Prediction,* Washington, DC, Amer. Meteor. Soc., 11B.4. [Available online at http://ams.confex.com/ams/pdfpapers/95172.pdf.]

——, S. Lakshmivarahan, and J. S. Kain, 2002: Development of an "events-oriented" approach to forecast verification. Preprints, *19th Conf. on Weather Analysis and Forecasting/15th Conf. Numerical Weather Prediction,* San Antonio, TX, Amer. Meteor. Soc., 7B.3. [Available online at http://ams.confex.com/ams/pdfpapers/47738.pdf.]

Barancourt, C., J. D. Creutin, and J. Rivoirard, 1992: A method for delineating and estimating rainfall fields. *Water Resour. Res.,* **28,** 1133–1144.

Berrocal, V. J., A. E. Raftery, and T. Gneiting, 2007: Combining spatial statistical and ensemble information in probabilistic weather forecasts. *Mon. Wea. Rev.,* **135,** 1386–1402.

Briggs, W. M., and R. A. Levine, 1997: Wavelets and field forecast verification. *Mon. Wea. Rev.,* **125,** 1329–1341.

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 Conf. on Probability and Statistics in the Atmospheric Sciences,* Orlando, FL, Amer. Meteor. Soc., 1.6. [Available online at http://ams.confex.com/ams/pdfpapers/29359.pdf.]

——, 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, M. Chapman, K. W. Manning, and R. Morss, 2004: An object-oriented approach to quantitative precipitation forecasts. Preprints, *17th Conf. on Probability and Statistics in the Atmospheric Sciences/20th Conf. on Weather Analysis and Forecasting/16th Conf. on Numerical Weather Prediction,* Seattle, WA, Amer. Meteor. Soc., J12.4. [Available online at http://ams.confex.com/ams/pdfpapers/71819.pdf.]

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

——, and Coauthors, 2008: Forecast verification: Current status and future directions. *Meteor. Appl.,* **15,** 3–18.

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/20th Conf. on Weather Analysis and Forecasting/16th Conf. on Numerical Weather Prediction,* 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, A., A. Marshak, W. Wiscombe, and R. Cahalan, 1996: Multifractal characterization of intermittency in nonstationary geophysical signals and fields: A model-based perspective on ergocity issues illustrated with cloud data. *Current Topics in Nonstationary Analysis,* G. Treviño et al., Eds., World Scientific, 97–158.

Du, J., S. L. Mullen, and F. Sanders, 2000: Removal of distortion error from an ensemble forecast. *Mon. Wea. Rev.,* **128,** 3347–3351.

Ebert, E. E., 2008: Fuzzy verification of high-resolution gridded forecasts: A review and proposed framework. *Meteor. Appl.,* **15,** 51–64.

——, and J. L. McBride, 2000: Verification of precipitation in weather systems: Determination of systematic errors. *J. Hydrol.,* **239,** 179–202.

Gebremichael, M., and W. F. Krajewski, 2004: Assessment of the statistical characterization of small-scale rainfall variability from radar: Analysis of TRMM ground validation datasets. *J. Appl. Meteor.,* **43,** 1180–1199.

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.

Greene, J. S., W. E. Cook, D. Knapp, and P. Haines, 2002: An examination of the uncertainty in interpolated winds and its effect on the validation and intercomparison of forecast models. *J. Atmos. Oceanic Technol.,* **19,** 397–401.

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

Journel, A. G., 1983: Non parametric estimation of spatial distributions. *Math. Geol.,* **15,** 445–467.

Kedem, B., L. S. Chiu, and G. R. North, 1990: Estimation of mean rain rate: Application to satellite observations. *J. Geophys. Res.,* **95,** 1965–1972.

Kundu, P., and T. L. Bell, 2003: A stochastic model of space–time variability of mesoscale rainfall: Statistics of spatial averages. *Water Resour. Res.,* **39,** 1328–1343.

Maillard, P., 2001: Developing methods of texture analysis in high resolution images of the Earth. *Anais X SBSR,* Foz do Iguacu, Brazil, INPE, 1309–1319.

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

——, and ——, 2007: Verification via optical flow. *Proc. Third Int. Verification Methods Workshop,* Reading, United Kingdom, ECMWF.

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

——, ——, 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.

Matheron, G., 1963: Principles of geostatistics. *Econ. Geol.,* **58,** 1246–1266.

Mela, K., and J. N. Louie, 2001: Correlation length and fractal dimension interpretation from seismic data using variograms and power spectra. *Geophysics,* **66,** 1372–1378.

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

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.

Ripley, B. D., 1991: *Statistical Inference for Spatial Processes*. Cambridge University Press, 148 pp.

Roberts, N. M., 2005: An investigation of the ability of a storm-scale configuration of the Met Office NWP model to predict flood-producing rainfall. Forecasting Research Tech. Rep. 455, Met Office, Exeter, United Kingdom, 80 pp.

Skamarock, W. C., 2004: Evaluating mesoscale NWP models using kinetic energy spectra. *Mon. Wea. Rev.,* **132,** 3019–3032.

——, M. E. Baldwin, and W. Wang, 2004: Evaluating high-resolution NWP models using kinetic energy spectra. *Proc. Fifth WRF/14th MM5 Users' Workshop,* Boulder, CO, NCAR, 53–56.

Sapozhnikov, V. B., and E. Foufoula-Georgiou, 2007: An exponential Langevin-type model for rainfall exhibiting spatial and temporal scaling. *Nonlinear Dynamics in Geosciences,* A. A. Tsonis and J. B. Elsner, Eds., Springer, 87–100.

Şen, Z., 1997: Objective analysis by cumulative semivariogram technique and its application in Turkey. *J. Appl. Meteor.,* **36,** 1712–1724.

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,** D08111, doi:10.1029/2004JD005395.

Weiss, S. J., J. Kain, J. Levit, M. Baldwin, D. Bright, G. Carbin, and J. Hart, 2005: NOAA Hazardous Weather Testbed: SPC/NSSL Spring Program 2005—Program overview and operations plan. Storm Prediction Center, Norman, OK, 61 pp. [Available online at http://www.spc.noaa.gov/exper/Spring_2005/2005_ops_plan.pdf.]

Yoo, C., and E. Ha, 2007: Effect of zero measurements on the spatial correlation structure of rainfall. *Stoch. Environ. Res. Risk Assess.,* **21,** 287–297.

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.