A new metric for comparing precipitation patterns with an application to ensemble forecasts

V. Venugopal, S. Basu, and E. Foufoula-Georgiou

St. Anthony Falls Laboratory, University of Minnesota, Minneapolis, Minnesota, USA

Received 28 August 2004; revised 16 December 2004; accepted 10 February 2005; published 26 April 2005.

[1] Ensemble forecasting can be seen as serving two purposes: (1) by comparison of the control and ensemble members to the observed precipitation field, one can assess the forecast performance probabilistically; and (2) by comparison of ensemble members to the control forecast, one can assess the "diversity" of an ensemble and quantify the uncertainty of the forecast. Both problems are grounded to the basic requirement of being able to compare spatially nonhomogeneous, intermittent fields and come up with low-dimensional metrics that can summarize this comparison. Several standard metrics exist (e.g., root mean square error (RMSE), Brier score, and equitable threat score (EqTh)) and are adopted in many operational studies. We studied (1) a fine-scale ensemble precipitation forecast produced from the Advanced Regional Prediction System (ARPS) and (2) forecasts from multiple models (e.g., the 1998 Storm and Mesoscale Ensemble Experiment (SAMEX '98)) for the purpose of exploring how the selection of the performance metric can affect inferences about the quality and uncertainty of a forecast. We propose a new measure called forecast quality index, which combines image analysis and nonlinear shape comparison features, and we show that it is a more robust and informative metric compared to traditional metrics such as RMSE and EqTh.

Citation: 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.

1. Introduction

[2] Ensemble prediction at the mesoscale has been actively explored in the scientific community and operational ensemble prediction systems exist in most forecast centers (e.g., ECMWF [see *Buizza*, 1997; *Buizza and Palmer*, 1995] and NCEP [see *Toth and Kalnay*, 1993, 1997]). As the spatial resolution of the forecasts becomes finer (due to improved computing resources) and also as the need for hydrologically useful precipitation forecasts at a resolution (grid size) of a few square kilometers becomes more acute, producing a large number of ensemble members from a numerical weather prediction model becomes a problem. This problem has driven research on how to generate ensembles such that, with the fewest possible members, they are able to characterize the probability structure of the forecast and thus define its uncertainty. Toward that end, several methodologies based on the fastest modes of growth have been promoted [e.g., *Toth and Kalnay*, 1997]. Along similar lines, efforts to identify the "best member" of an ensemble and dynamically evolve the state of the system around it as the forecast proceeds, have been explored, although under intense controversy [e.g., see *Roulston and Smith*, 2003; *Bright and Nutter*, 2004]. Both of the above problems have been exacerbated by the fact that at high spatial resolutions, the variability of precipitation increases (higher spread of PDFs) and more realizations are needed to accurately define the higher moments (tails) of this distribution. Also the fine resolution forecasts make the problem of summarizing (ideally, with a single measure) the comparison of ensemble members with observations and among themselves more difficult.

[3] It makes intuitive sense that the selection of a particular measure/metric (we use measure and metric interchangeably in the rest of the paper) to compare precipitation patterns versus another measure might generally lead to different inferences about the forecast performance and ensemble spread. Standard measures are typically based on root mean square error (RMSE), equitable threat score (EqTh), correlation coefficient and Brier score [see, e.g., *Ebert et al.*, 2003; *Jolliffe and Stephenson*, 2003] (see also http://www.bom.gov.au/bmrc/wefor/staff/eee/verif/verif_web_page.shtml). Here, we introduce a new measure called forecast quality index (FQI) and show that, in comparison to RMSE and EqTh, it provides important additional information that is able to pick up the difference in patterns such that robust inferences can be made. Section 2 introduces the new metric and demonstrates its potential on some illustrative examples, while section 3 discusses two case studies: (1) a set of ensemble forecasts generated by the Advanced Regional Prediction System (ARPS) and (2) multimodel

0148-0227/05/2004JD005395

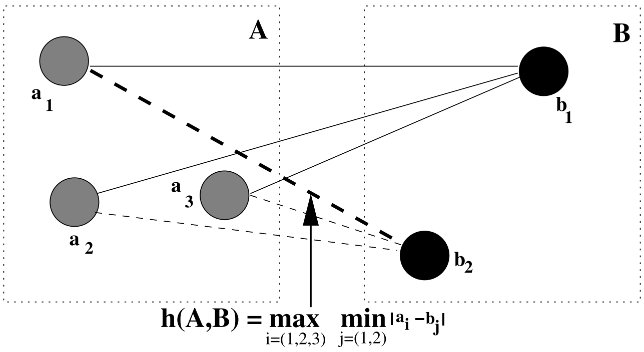

Figure 1. Schematic illustrating the idea of Hausdorff distance between two sets, A and B. For each point in A, the minimum distance to all points in B is measured (see broken lines). The maximum distance in this set of minimum distances is known as the forward distance, and is denoted by $h(A, B)$ (see thick broken line between a_1 and b_2). This process is repeated from B to A, and the resulting maximum is known as the backward distance, and is denoted by $h(B, A)$. The maximum of $h(A, B)$ and $h(B, A)$ is the Hausdorff distance between the sets A and B.

ensemble forecasts from the 1998 Storm and Mesoscale Ensemble Experiment (SAMEX '98). Finally, section 4 provides conclusions and discusses possible future work.

2. Forecast Quality Index: A Theoretical Background

2.1. A Brief Survey of Image Comparison Measures

[4] Objective image quality measures are important in image processing applications, and are a subject of continuing interest [*Gesù and Starovoitov*, 1999]. These measures can be broadly classified into two types: those that are mathematically defined (RMSE, signal-to-noise ratio (SNR) etc.), and those that are formulated based on human perceptions [see *Wang and Bovik*, 2002; *Pappas and Safranek*, 2000]. In this work, we focus our attention on mathematical measures, which can be further classified into two types: (1) amplitude-based and (2) distance-based. We briefly discuss below the pros and cons of each type of measures and how a combination of the two (a hybrid metric) could use the best qualities of each and perhaps serve as a bridge between them.

2.1.1. Amplitude-Based Measures

[5] Coefficient of correlation, RMSE and SNR are often used as a way to compare images. *Wang and Bovik* [2002] formulated a measure, which they termed a universal image quality index (UIQI), based on the first two moments (mean and standard deviation/covariance) of the given images. Mathematically, it is defined as

$$\mathrm{UIQI}(\mathrm{R}_1, \mathrm{R}_2) = \frac{\sigma_{R_1,R_2}}{\sigma_{R_1}\sigma_{R_2}} \cdot \frac{2\mu_{R_1}\mu_{R_2}}{{\mu_{R_1}}^2 + {\mu_{R_2}}^2} \cdot \frac{2\sigma_{R_1}\sigma_{R_2}}{{\sigma_{R_1}}^2 + {\sigma_{R_2}}^2} \qquad (1)$$

where R_1 and R_2 represent the fields being compared, μ_{R_1}, μ_{R_2} are the means, σ_{R_1}, σ_{R_2}, the standard deviations, respectively of the two fields, and σ_{R_1,R_2} represents the covariance between the two fields. It is evident from the equation above that the proposed measure is a combination of three properties: (1) correlation (the left most term on the right-hand side), (2) brightness (bias) and (3) distortion (variability). (The terms brightness/bias and distortion/variability are used in the image processing (geophysics) community, respectively.) Each of these components in themselves have been used extensively, but the combination of all three is a new approach. For instance, two images could be perfectly correlated (in the linear sense) but differ in magnitude, i.e., biased. This is taken into account by the second component. The combination of these three components results in a range of UIQI between -1 and 1. When UIQI is 1, it means that we have an exact match of the two images. The smaller the value of UIQI, the more the distortion. That said, however, it can immediately be seen that UIQI is an entirely amplitude-based measure, and thus would not be efficient in characterizing whether two patterns differ because of discrepancies in their amplitudes or simply because the two patterns are displaced. While the covariance could be made a function of lag (to account for displacements), the complication that arises from such a formulation is the need to define an objective function (penalization factor) that would enable one to estimate which lag yields the maximal similarity between the images being compared. It is quite well-known that numerical weather prediction models often produce forecasts that are displaced (both in time and space). To account for this kind of problem, we turn our attention to distance-based measures, which are useful primarily for binary images (each pixel is either 0 or 1).

2.1.2. Distance-Based Measures

[6] For binary image comparison, the Hausdorff distance is a natural choice. Intuitively, the Hausdorff distance metric measures the degree of mismatch between two finite sets A and B by measuring the distance of the point in the set A that is farthest from any point in the set B and vice versa. In other words, if $H(A, B) = d$, then every point of A must be within a distance d of some point of B and vice versa [*Huttenlocher et al.*, 1999]. Given two finite point sets $A = \{a_1,.., a_p\}$ and $B = \{b_1,.., b_q\}$, the Hausdorff distance metric (see Figure 1 for a schematic) is defined as:

$$H(A,B) = \max(h(A,B), h(B,A))$$

where

$$h(A,B) = \max_{a\in A}\left(\min_{b\in B} \parallel a - b \parallel\right)$$

and $\|.\|$ is some underlying Hölder norm on the points of A and B. A Hölder norm $\mathcal{L}_p$ is a generalization of the Euclidean distance between two points in $\mathcal{R}^n$, and is defined as $\mathcal{L}_p = \|\bar{\mathbf{r}}\|_p = (\sum_{i=1}^n |r_i|^p)^{1/p}$ for some n-dimensional vector $\bar{\mathbf{r}}$ [see, e.g., *Golub and van Loan*, 1996]. For two dimensions, i.e., $\mathcal{R}^2$, the (well-known) Euclidean distance is given as $\sqrt{r_x^2 + r_y^2}$. In this work, we have used the "taxi cab" distance, which is an $\mathcal{L}_1$ norm given by $\|\bar{\mathbf{r}}\|_1 = |r_x| + |r_y|$. In other words, taxi cab distance is measured as the sum of x and y movement in the Euclidean plane. For instance, the distance from (0,0) to (1,1) in the Euclidean plane is $\sqrt{2}$, but the distance in a taxi cab geometry would be 2. For more details on the taxi cab distance, see *Krause* [1986].

[7] In the case of binary fields or images, only the nonzero pixels form a valid set. For example, in the case of two binary images R_1 and R_2, the Hausdorff distance can be re-written as follows:

$$H(R_1, R_2) = \max(h(R_1, R_2), h(R_2, R_1))$$

where

$$\begin{aligned} h(R_1, R_2) &= \max_{R_1(i,j) \in NR_1} \left(\min_{R_2(l,m) \in NR_2} \| R_1(i,j) - R_2(l,m) \| \right) \\ &= \max_{R_1(i,j) \in NR_1} \left(\min_{R_2(l,m) \in NR_2} (|i-l| + |j-m|) \right) \end{aligned}$$

NR_1 and NR_2 are the sets of nonzero pixels of images R_1 and R_2 respectively and the spatial coordinates are represented by (i, j) and (l, m). The above classical definition of the Hausdorff distance is very sensitive to outliers. An outlier is often the point where the Hausdorff distance is achieved, giving too large a value for H [*Moeckel and Murray*, 1997]. Using a generalization of the Hausdorff distance by taking the kth percentile distance rather than the maximum, one can avoid this outlier problem. This generalized Hausdorff distance is known as the partial Hausdorff distance (PHD) and rarely obeys the metric properties (see *Rucklidge* [1996] for a discussion on this). The Partial Hausdorff distance can be written as:

$$PHD_k(R_1, R_2) = \max(h(R_1, R_2), h(R_2, R_1))$$

where

$$\begin{aligned} h(R_1, R_2) &= \underset{R_1(i,j) \in NR_1}{k^{th}} \left(\min_{R_2(l,m) \in NR_2} \| R_1(i,j) - R_2(l,m) \| \right) \\ &= \underset{R_1(i,j) \in NR_1}{k^{th}} \left(\min_{R_2(l,m) \in NR_2} (|i-l| + |j-m|) \right) \end{aligned} \quad (2)$$

In this work, we have used the 75th percentile.

2.1.3. Hybrid Measures

[8] Extending the distance-based measure for binary images to real valued images, is a logical step. This kind of hybrid measure would have the merits of both the amplitude and the distance-based measures. The simplest approach will be to generalize the classical Hausdorff distance, by adding a component related to magnitudes, as follows [*Gesù and Starovoitov*, 1999]:

$$GH(R_1, R_2) = \max(h(R_1, R_2), h(R_2, R_1))$$

where

$$\begin{aligned} h(R_1, R_2) &= \max_{R_1(i,j) \in NR_1} \left(\min_{R_2(l,m) \in NR_2} \| R_1(i,j) - R_2(l,m) \| \right) \\ &= \max_{R_1(i,j) \in NR_1} \left(\min_{R_2(l,m) \in NR_2} \{(|i-l| + |j-m|) + \lambda |R_1(i,j) - R_2(i,j)|\} \right) \end{aligned}$$

[9] The limitation of this measure is the free normalization parameter λ, which will drastically affect the magnitude of this generalized Hausdorff distance. It is noted however that in practical applications such as quantitative precipitation forecast verification, selecting the parameter λ is not trivial, as one has to combine distance (in km) and rainfall intensity (in mm/hour) into one single measure.

2.2. Proposed Measure for QPF Verification

[10] In this work, we made an effort to combine both measures (amplitude-based and distance-based) to create a new metric called the forecast quality index (FQI). The proposed index is defined as:

$$\mathrm{FQI}(\mathrm{R}_1, \mathrm{R}_2) = \frac{\frac{PHD_k(R_1, R_2)}{Mean[PHD_k(R_1, Surrogates\ of\ R_1)]}}{\frac{2\mu_{R_1}\mu_{R_2}}{\mu_{R_1}{}^2 + \mu_{R_2}{}^2} \frac{2\sigma_{R_1}\sigma_{R_2}}{\sigma_{R_1}{}^2 + \sigma_{R_2}{}^2}} \quad (3)$$

where the means and standard deviations in the denominator are computed for only the nonzero pixels.

[11] The numerator is a normalized partial Hausdorff distance (PHD) between two binary fields R_1 and R_2, where R_1 is the reference or "true" field, and R_2 is the field to be compared and k is the percentile at which the computation of PHD is done (see equation (2)). *PHD* is sensitive to the percentage of nonzero pixels over the domain of observation. Therefore, if the objective is to merely compare two images at any particular time instant, a nonnormalized PHD would suffice in the numerator of FQI. However, this would not be appropriate for comparison of two fields over a time period or for comparison of two different ensemble members to the control run or observed field, because the percentage of nonzero pixels (rain-covered area) could change significantly with time, or from ensemble member to another. Thus, for consistency, one could simply normalize the *PHD* by the total number of nonzero pixels to account for the sensitivity, but that would not result in a dimensionless measure. For that reason, we normalize *PHD*(R_1, R_2) by the mean of the partial Hausdorff distance between the R_1 (reference field) and its surrogates. A surrogate field is a stochastic realization of a process (observed field, in this case) with the same probability density function (PDF) and spatial correlation structure (see *Kantz and Schreiber* [1997] for more details). Traditionally, surrogates are generated with either the same PDF or the same correlation structure as the process under study, but not both. *Schreiber and Schmitz* [1996] proposed an algorithm called iterative amplitude-adjusted Fourier transform (IAAFT) technique to generate surrogates which simultaneously preserve the correlation structure and the PDF. Thus, by taking the mean of *PHD*(R_1, *Surrogates of* R_1), we normalize by a representative value of *PHD* between R_1 and its possible stochastic realizations within the domain of observation. We use 10 surrogates to compute the mean distance. To aid the reader in visualizing the surrogates, we show a sample image and three of its surrogates in Figure 2. We do not delve into the details of how to compute surrogates, but refer the reader to *Kantz and Schreiber* [1997] for further details. (For the interested reader, we provide a URL in the Acknowledgments section for obtaining Matlab scripts to generate surrogates.) It is worth mentioning here that surrogates can be seen as a generalization of transposes and mirrors of the original image. In that respect, using a collection of transposed and mirrored images can be seen as a "poor man's" surrogates, and could potentially be used for normalization

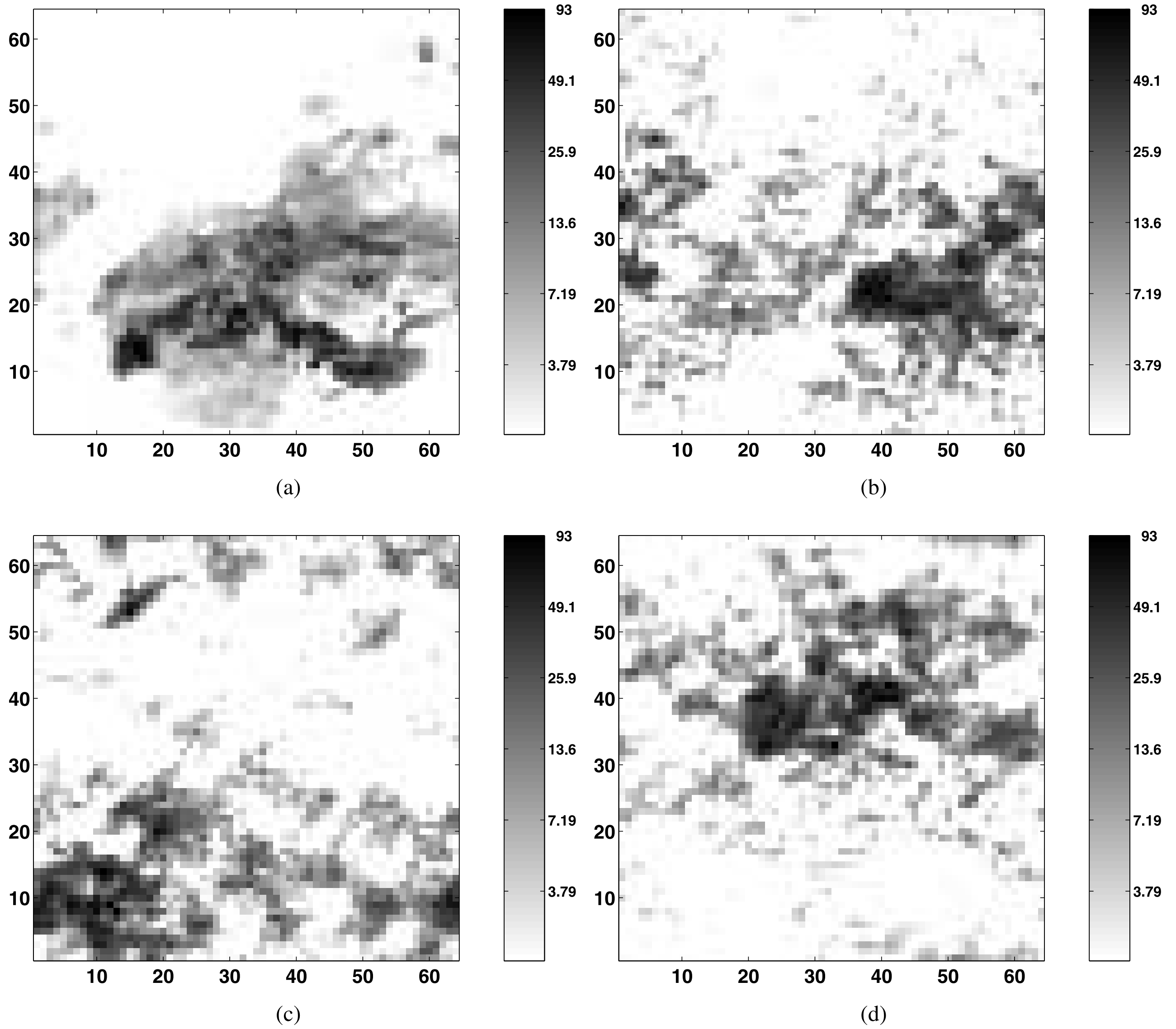

Figure 2. (a) Original field at 4 km spatial resolution; (b)–(d) surrogates of this field.

in practical applications. However, we caution that such a normalization leads to a conservative estimate of the numerator, since it is based on a restricted set of surrogates, which exclude realizations that redistribute intensities from a large connected area into smaller areas.

[12] The denominator of the proposed metric is a modified UIQI index computed from the rainfall intensity field (and not only its binary counterpart which was used in the numerator of equation (3)). As can be seen from equation (1), the covariance term of UIQI is not considered, since PHD in the numerator accounts for any discrepancies due to displacement. Note that in creating the binary images to be used in the computation of *PHD*, one could employ thresholding. For instance, this could be useful in QPF applications where the interest might be primarily in those intensities that are above a threshold (or, equivalently, when low intensities are not of significance). In that case, the means and standard deviations that appear in the denominator (modified UIQI) also need to be calculated for the thresholded image. The range of *PHD* is $[0, \infty]$, while the range of the modified UIQI is [0, 1], thus making the range of FQI, $[0, \infty]$. A value of FQI close to zero would imply that the two sets (images) being compared are close to each other.

2.3. Illustrative Examples

[13] To illustrate the merit of the proposed measure for comparing precipitation patterns, we start with a very simple example as shown in Figure 3. The field to be studied is a disc of radius 10 units, centered at (20,20) with exactly the same values across the disc. This is marked "original" in the figure and could be seen as the observed field in a QPF context. "Member 1" and "member 2" are shifted versions of the "original"; member 1 has its center at (40,40), while member 2 is centered at (70,70). In the context of QPF, the "original" could be treated as the "observed" and "members 1 and 2" could potentially represent ensemble members of a numerical weather prediction model output. For these three fields, we computed three measures, amplitude-based RMSE, distance-based EqTh [see *Miller*, 2000], and the proposed hybrid, forecast quality index (FQI). For completeness purpose, we define

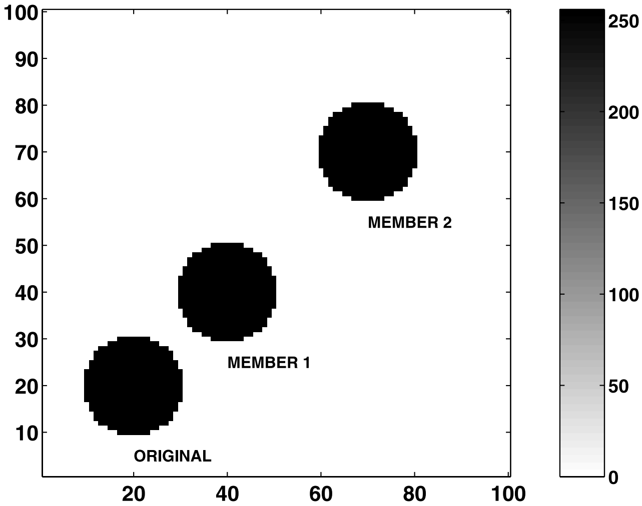

Figure 3. A simplified case study to illustrate the utility of the proposed index FQI.

here RMSE and EqTh. Given two images R_1 and R_2 of size $N \times N$

$$\mathrm{RMSE}(R1, R2) = \frac{1}{N}\sqrt{\sum_i \sum_j [R_1(i,j) - R_2(i,j)]^2}$$

$$\mathrm{EqTh}(R1, R2) = \frac{(N_i - S)}{(N_u - S)}$$

where N_i is the count of $\{NR_1 \cap NR_2\}$, N_u is the count of $\{NR_1 \cup NR_2\}$, and $S = NR_1 \times NR_2/N$, where NR_1, NR_2 represent the number of nonzero values in R_1 and R_2, respectively. Table 1 shows the computed values of these measures and the proposed FQI measure.

[14] It is obvious that in this example, the RMSE and EqTh fail to capture the difference between members 1 and 2 compared to the original field, namely fail to illustrate that member 2 is farther away from the original than member 1 is. The proposed index, however, clearly depicts this fact. The illustrative example considered above is clearly too simplistic, and not a reflection of the kinds of cases one might encounter in reality. Below, we consider another case that illustrates the shortcomings of RMSE and threat score, and the potential of the proposed measure, FQI, in a more realistic case.

[15] A radar-observed rainfall field over Houston, Texas, is considered. The original observations are at 2km spatial resolution, but for this illustrative example we have aggregated the field to 4 km spatial resolution. The area of observation is about 256 × 256 km^2. Figure 4a shows the original observations from which we constructed three rainfall images: case 1, observation field is rotated by 90° (Figure 4b); case 2, observation field is mirrored around a horizontal axis passing through the center of the field (Figure 4c); and case 3, observation field is mirrored around a vertical axis passing through the center of the field (Figure 4d). Cases 2 and 3 perhaps are most reflective of a real situation, wherein a numerical forecast has predicted the location incorrectly (in the x or y direction). The three measures, RMSE, EqTh, and FQI are calculated for a threshold of 2 mm, i.e., all observations less than 2mm are not taken into consideration (considered to be zero).

[16] Visually, one would expect that case 3 is the closest to the original field, followed by case 1 and case 2. As can be seen from Table 2, it is clear that the relative change in RMSE between the three cases is not very great, indicating that RMSE is not powerful enough to distinguish well between the three types of differences. EqTh is also not able to distinguish cases 1 and 2 (gives a value close to zero, just suggesting that they are different and no additional information); however, in case 3, EqTh is higher (i.e., different from zero) given that there is a significant overlap. The differences in FQI (relative change from case to case as well) between the three cases is indicative of the power of the proposed measure. As expected, FQI is lowest for case 3, and highest for case 2, indicating that case 3 is the closest and case 1 is the farthest from the original field. Thus FQI is able to incorporate the optimal qualities of distance- and amplitude-based measures resulting in a more effective measure (which can be used alone or in addition to other traditional measures) for image comparison purposes.

3. Fine-Scale Precipitation Forecast: A Case Study

3.1. ARPS Ensemble Forecast and Data Description

[17] We chose the 28–29 March 2000 Fort Worth, Texas tornadic thunderstorm for analysis purposes since this event has been well documented in the literature and also because high resolution numerical models have been able to successfully simulate it. Figure 5a shows a snapshot of the storm on 29 March, 2000 as observed by a radar. Below, we describe briefly the generation of the ensemble members that were used for this study. The reader is referred to *Xue et al.* [2003] and *Levit et al.* [2004] for more details relating to the storm conditions, and to *Xue et al.* [2003], *Levit et al.* [2004], and *Kong et al.* [2004] for details pertaining to the generation of the ensemble members.

[18] A three-nested domain system was used, the finest resolution being 3km. The 3-km domain was centered over Fort Worth and spanned approximately 500 × 500 km^2. The grid size of the outer two domains was 24 and 6km, respectively [see *Kong et al.*, 2004, Figure 2]. All the domains used 53 terrain-following vertical layers stretching from 20 m at the ground to approximately 800 m at the top. In this study, the SLAF (scaled lagged average forecast) method was used for ensemble generation, due to its simplicity and economy [*Ebisuzaki and Kalnay*, 1991]. For each nested domain, a 5-member ensemble (one control run and 4 perturbed members) was generated. To construct the four members, the perturbation between one previous ARPS forecast and the current analysis was scaled based on

Table 1. Three Measures Computed for the Simplified Case Study of Figure 3 to Illustrate the Utility of the Proposed Index

	RMSE	EqTh	FQI
Original and member 1	68.41	−0.02	0.39
Original and member 2	68.41	−0.02	1.15

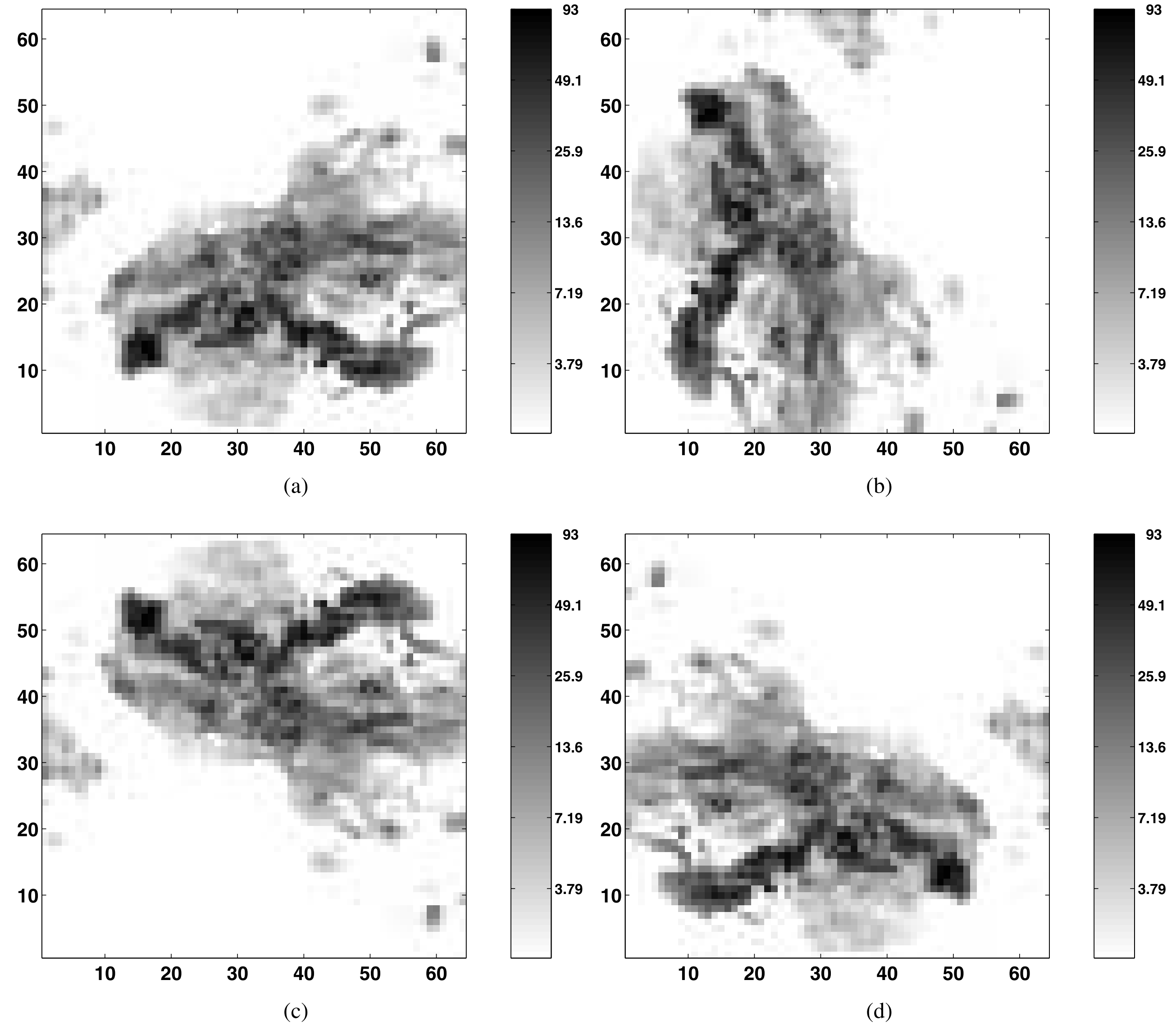

Figure 4. Illustrative example: (a) original field at 4 km spatial resolution; (b) case 1, field in Figure 4a is rotated by 90°; (c) case 2, field in Figure 4a is mirrored around a horizontal axis passing through the center of the field; and (d) case 3, field in Figure 4a is mirrored around a vertical axis passing through the center of the field.

an error growth assumption and then added to and subtracted from the analysis to form two (paired) members. For the rest of the discussion, we refer to the control as Cn, and the 4 perturbed members as S1, S2, S3, and S4 [see *Kong et al.*, 2004, Figure 4]. Radar-observed precipitation fields were available every 5 min.

3.2. Results and Discussion

[19] Two types of comparisons are performed between the 5 ensemble members of ARPS (at a spatial resolution of 3km) and radar-observed rainfall (also at the same spatial resolution re-gridded from the original 1km resolution fields): (1) The control run and members S1, S2, S3 and S4 are compared with the observed field and (2) members S1 through S4 are compared to the control run. While the first type of comparison tells us how well the model has done in reproducing the observed field, the second type of comparison gives us an idea of the kind of diversity within the ensembles (note that members S1 through S4 are the result of perturbations around the control run).

[20] Analysis of the ensemble was performed over one hour of simulation (forecast), but we report here only the analysis of a specific time instant of the storm evolution where the model appears to have mostly captured the basic statistics of the event under study such as conditional mean and standard deviation (see Figure 6 and Table 3).

Table 2. Three Measures Computed for the Three Case Studies of Figure 4 to Illustrate the Potential of the Proposed Index

	RMSE	EqTh	FQI
Case 1	14.2	0.08	0.81
Case 2	15.7	−0.02	1.5
Case 3	11.4	0.27	0.2

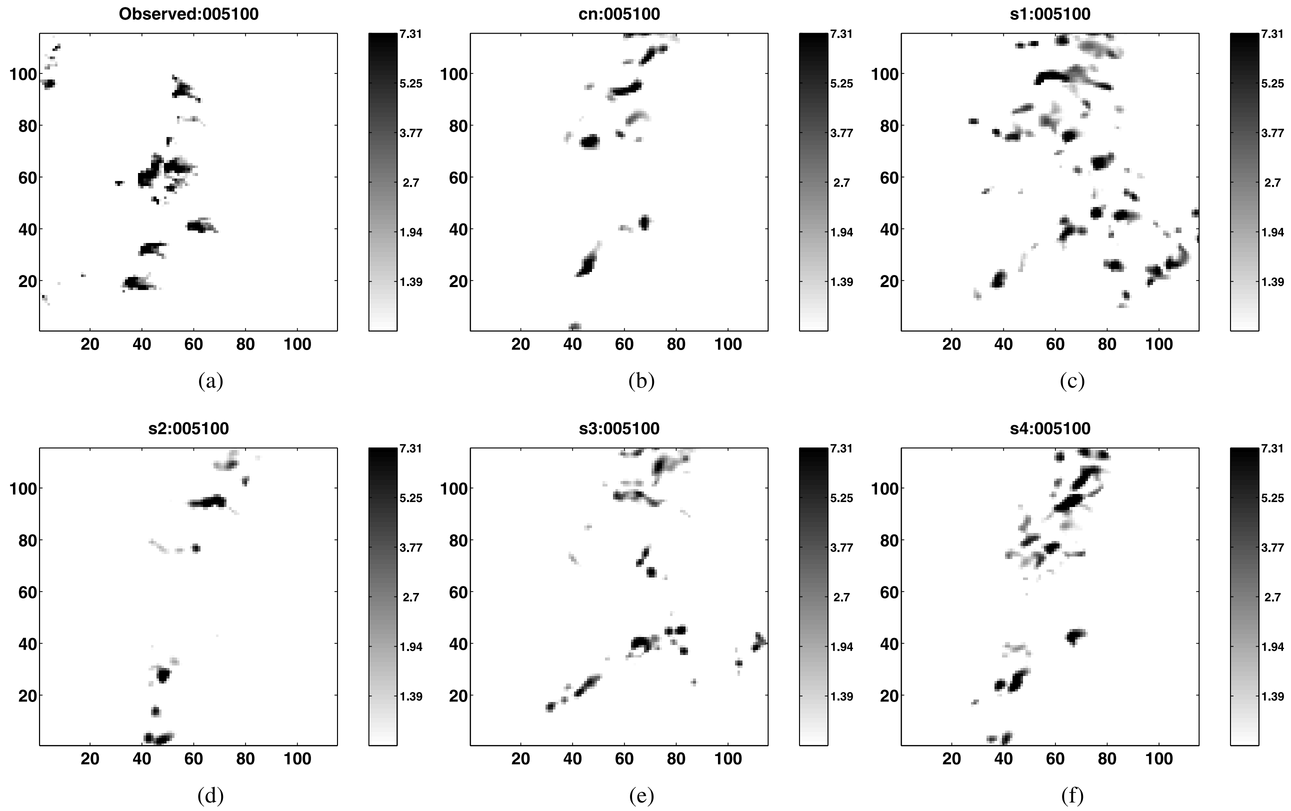

Figure 5. Comparison of snapshots at 3 km spatial scale of the radar-observed precipitation and different ensemble realizations from the model. (a) Observed, (b) control, and (c)–(f) ensemble members S1 through S4, respectively. Threshold equals 1 mm/hour.

[21] Figure 6 shows the comparison of time variation of the basic statistics of the observed and ensemble fields. The statistics chosen were the conditional (nonzero) mean and standard deviation, and percentage of rain-covered area. Apparently, the ensemble (barring S1) has underestimated the percentage of rain-covered area. However, it appears that the mean and standard deviation of the ensemble members envelope the observed mean and standard deviation. This, to an extent, is encouraging, and implies that there is diversity in the ensemble statistics around the true state of the system.

[22] Another measure of closeness is a simple comparison of the range of values in each field. Figure 7 shows scattergrams (also known as quantile-quantile plots; see *Jolliffe and Stephenson* [2003]) of the observed versus the modeled precipitation fields shown in Figure 5. The values plotted have been sorted in ascending order (to get an estimate of the range), and the dotted line indicates the perfect correspondence line. For the low intensities, S1, S3 and S4 appear to have over-estimated (above the 45° line) the intensities relative to the observed field, while the control run and S2 appear to have under-estimated (below

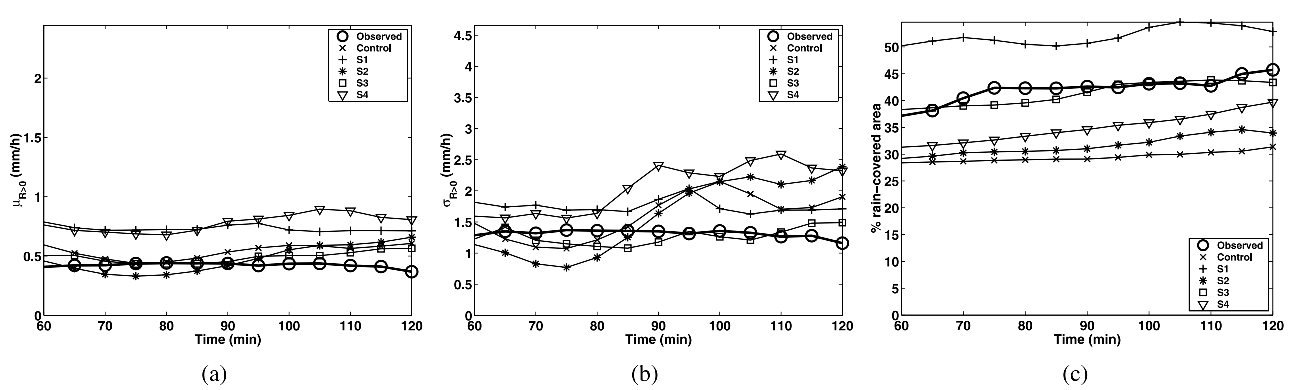

Figure 6. Comparison of the time variation of basic statistics of ARPS ensemble and observations: (a) mean conditional on rain ($\mu_{R>0}$), (b) standard deviation conditional on rain ($\sigma_{R>0}$), and (c) percentage of rain-covered area. Threshold equals 0 mm/hour.

Table 3. Conditional Statistics of the Observed Field and the Ensemble Members Shown in Figure 5

	Mean	Standard Deviation	Rainy Area, %
Observed	0.51	1.42	29.98
Control	0.53	1.77	29.09
S1	0.76	1.86	50.68
S2	0.42	1.63	31.02
S3	0.45	1.18	41.58
S4	0.79	2.41	34.60

the 45° line) the intensities. For the high intensities, while S1 over-predicted, the rest of the ensemble members under-predicted. In other words, the correspondence plot of Figure 7 seems to suggest that the ensemble members do have a spread in their values which is comparable to that of the observed fields, and is not systematically biased around the range of the observed values.

[23] Simple visual inspection of Figure 5 suggests that the control run, S2 and S4 appear to be the closest to the observed field, while S1 and S3 seem to have "additional" features present (for instance, southeast corners). It is worth repeating here that only amplitude-based or only distance-based measures would not be sufficient to illustrate important differences between the observed field and the ensemble members. Some combination of the two types (i.e., a hybrid measure) is thus necessary to be able to better gauge the performance of the model. Below, we discuss the proposed hybrid measure (FQI), and compare its performance with more traditional measures such as RMSE (amplitude-based) and EqTh (distance-based). We note here again that the value of *PHD* in the numerator of FQI in equation (2) is computed for the 75th percentile. Again, the comparison is done in relation to the observed as well as the control run.

[24] The magnitudes of FQI, EqTh and RMSE (for t = 85 min. in Figure 6) when the members of the ensemble, i.e.,

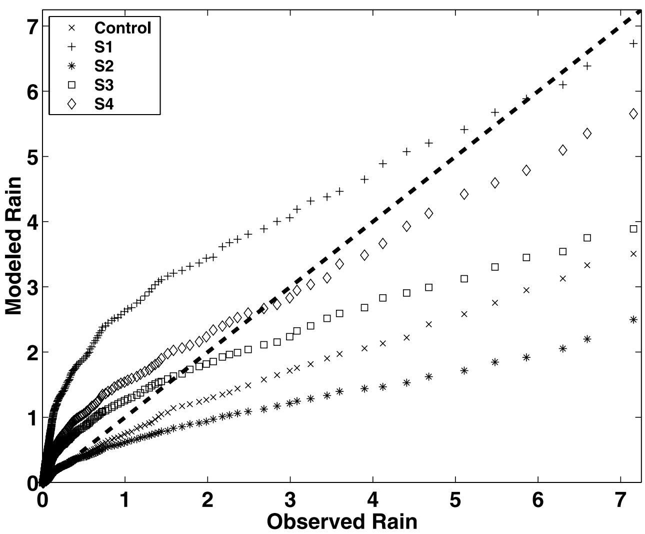

Figure 7. Scattergrams (also called quantile-quantile plots) of the observed versus modeled (control run and four ensemble members) precipitation fields (shown in Figure 5). The values that are plotted have been sorted in ascending order, and the dotted line indicates a "perfect correspondence line." Threshold equals 0 mm/hour.

Table 4. Comparison of Ensemble Members With Observed Using the Three Measures Discussed in the Text for a Threshold of 1 mm/hour[a]

	RMSE	EqTh	$Numer_{FQI}$	$Denom_{FQI}$	$FQI = \frac{Numer_{FQI}}{Denom_{FQI}}$
Control	1.31	0.05	0.52	0.89	0.58
S1	1.63	0.02	1.18	0.94	1.25
S2	1.29	0.01	0.54	0.83	0.65
S3	1.16	0.03	0.95	0.94	1.01
S4	1.73	0.02	0.36	0.80	0.45

[a]Ensemble members shown in Figures 5b–5f; observed shown in Figure 5a.

control, S1, S2, S3 and S4, are compared with the respective observed field are given in Tables 4 and 5. The comparison has been done for different thresholds and the results are given for 1 mm (Table 4) and 5 mm (Table 5). Several observations can be made from Tables 4 and 5:

[25] 1. RMSE suggests that S3 is the "closest" to the observed field. This appears to be contradictory to the visual evidence that control, and S4 (or S2) are closest.

[26] 2. EqTh has a value very close to zero. This suggests that all the ensemble member forecasts are very different from the observed, which is certainly not the case. Moreover, it does not tell us much about the visual discrepancies that we notice from the images (Figure 5). Given its low values, it does not help to look at the diversity of the ensemble.

[27] 3. The lower the value of FQI, the closer the two fields that are being compared. With that in mind, looking at the proposed hybrid measure, FQI, one could infer that control, S2 and S4 are the closest to the observed. This, in contrast to RMSE and EqTh, does indeed corroborate visual evidence (Figure 5). Furthermore, the diversity in the forecast can also be measured by the range of FQI (0.4 to 1.2). In essence, FQI, which captures both distance- and magnitude-related discrepancies between the forecasted and observed fields, definitely presents a more robust way to characterize uncertainty of a forecast.

[28] 4. While on one hand it may be convenient to have a single measure to evaluate the performance of a model forecast, on the other hand, it may be equally useful to be able to assess if the model has done "poorly" in reproducing amplitudes or capturing the location of features. This can be achieved by looking at the individual components of the proposed measure, i.e., the values of the numerator (distance-based) and denominator (amplitude-based). In other words, if the model forecast matched well in the locations of the features, it would show a value close to 0 for the numerator, while if it matched well in the intensities (irrespective of the location; i.e., the overall PDF), it would reflect in the denominator value being close to 1. Columns 4 and 5 of Table 4 respectively show the numerator and

Table 5. Same as Table 4, but for a Threshold of 5 mm/hour

	RMSE	EqTh	$Numer_{FQI}$	$Denom_{FQI}$	$FQI = \frac{Numer_{FQI}}{Denom_{FQI}}$
Control	1.23	0.002	0.38	0.29	1.34
S1	1.47	0.016	0.76	0.30	2.54
S2	1.21	0.004	0.70	0.20	3.49
S3	1.04	0.003	0.77	0.51	1.50
S4	1.63	−0.006	0.56	0.19	2.93

Table 6. Comparison of Ensemble Members With the Control Run Using the Three Measures Discussed in the Text for a Threshold of 1 mm/hour[a]

	RMSE	EqTh	FQI
S1	1.68	0.08	0.73
S2	1.13	0.18	0.17
S3	1.13	0.10	0.30
S4	1.34	0.22	0.13

[a]Ensemble members shown in Figures 5c–5f; control run shown in Figure 5b.

denominator for a threshold of 1mm. We notice that the denominator is close to 1 indicating that the magnitude of the intensities has been captured very well by the model. However, when we look at the numerator, it suggests that the forecast provided by S1 is the furthest (in relation to the other members) in terms of capturing the location of the features. This, of course, is evident visually from Figure 5c, where one notices that there are many small "cells" that are present in comparison to the observed. Furthermore, as can again be verified visually, S3 is the next "worst" in reproducing the location of the features, and confirmed by the relatively high value of the numerator.

[29] 5. The values of RMSE and EqTh appear to be insensitive to the threshold value chosen to compare the fields (see Tables 4 and 5) suggesting that the extreme precipitation intensities are forecast with the same skill as the lower precipitation intensities. Given that we have clearly seen from Figure 5 that the ensemble forecast members exhibit a much more small-scale cellular structure than the observed field, we know that this is not the case. It is interesting to note that the values of FQI for the two different thresholds differ significantly as seen from Tables 4 and 5). For a threshold of 5 mm, FQI indicates that the control run forecast is the closest to the observed, while S2 and S4 appear to have done a poor job at capturing the higher intensities (>5 mm), even though at a lower threshold (1 mm; see Table 4), the predicted fields from the control run, S2 and S4 were all "close" to the observed field. This, in turn, means that the inability of the ensemble members to reproduce the location and intensity of the high precipitation values is well depicted by the FQI metric but not by the RMSE and EqTh metrics.

[30] 6. Increasing the threshold changes the numerator and denominator significantly (see columns 4 and 5 in Table 5). The values shown suggest that all of them have done very poorly (S3 to a lesser extent) when it comes to capturing the higher intensities. While control, S1 and S3 show a decrease in the numerator (indicating a better match of the location of features with intensity larger than 5 mm than for 1 mm), S2 and S4 show an increase in the numerator suggesting a poor match of location of features with intensity larger than 5mm (in relation to a good match when a threshold of 1 mm was applied). It is noted that traditional measures like RMSE and EqTh are evaluated and reported for different threshold values, so as to assess the performance of a forecast in capturing extreme events. It is suggested that the proposed measure be also computed for different thresholds, and a curve (FQI vs. threshold) rather than a single value be reported.

[31] From the above observations, one could infer that using RMSE or EqTh alone would provide limited assessment about model performance. At the same time, the FQI seems to be a reliable and robust measure for QPF verification, which could be used alone or in addition to other measures to provide confidence in model assessment. It is worth mentioning here that distance-based measures, and consequently our proposed measure, are less successful in cases where the images to be compared have 100 percent rain coverage. However, this concern can be alleviated by employing an appropriate threshold.

[32] Apart from the comparison of the ensemble forecast members to the observed precipitation, equally important is the assessment of the "spread" within an ensemble by comparing the ensemble members to the control run. Tables 6 and 7 show RMSE, EqTh and FQI when S1 through S4 are compared with the control run with a threshold of 1 and 5 mm, respectively. On the basis of what the comparison with the observed fields shows, one could anticipate that FQI would be the least for S2 and S4 (for a threshold of 1 mm) suggesting that they are close to control (given that control, S2 and S4 are close to the observed), and S1 and S3 comparisons should yield higher FQI values. This is exactly what we see in Table 6. RMSE does in fact also show that S2 and S4 are closer to the control run than S1 and S3. The same is true with EqTh. RMSE again is insensitive to the change in threshold, while EqTh seems to corroborate what FQI suggests (S4 is closest to control; see Table 7).

[33] To illustrate the applicability of the proposed measure over larger areas, comparisons were made between the forecasts from the 1998 Storm and Mesoscale Ensemble Experiment (SAMEX '98), and the observed precipitation available from NCEP. The SAMEX multimodel ensemble consists of a total of 25 members from 4 different models: 5 ARPS, 5 Eta, 5 RSM and 10 MM5 members. The spatial and temporal resolutions of both the observed and forecasted accumulated precipitation fields are 30 $\times$ 30 km^2 and 3 hours, respectively. The spatial domain covers the continental United States (see *Hou et al.* [2001] and *Miller* [2000] for more details). The comparison we have done is for a lower threshold of 5 mm. Figure 8 shows the observed precipitation field (middle panel) and one forecasted field from each of the aforementioned four models. Table 8 shows the various measures computed for this case study. As can be clearly seen from the table, both the "traditional" measures (RMSE and EqTh) rank the bottom right (forecast 4) to be the best, although it does not capture any of the features of the observed field. However, the ranking inferred visually and that obtained from the proposed measure (FQI) (as shown in Table 8) are in good agreement.

4. Conclusions

[34] Rainfall being the result of complex atmospheric phenomena possesses a complex temporal and spatial structure. The intermittent yet organized nature of spatial rainfall not only makes quantitative precipitation forecasting

Table 7. Same as Table 6, but for a Threshold of 5 mm/hour

	RMSE	EqTh	FQI
S1	1.55	−0.01	0.63
S2	1.13	0.10	0.42
S3	1.06	0.04	0.46
S4	1.33	0.16	0.11

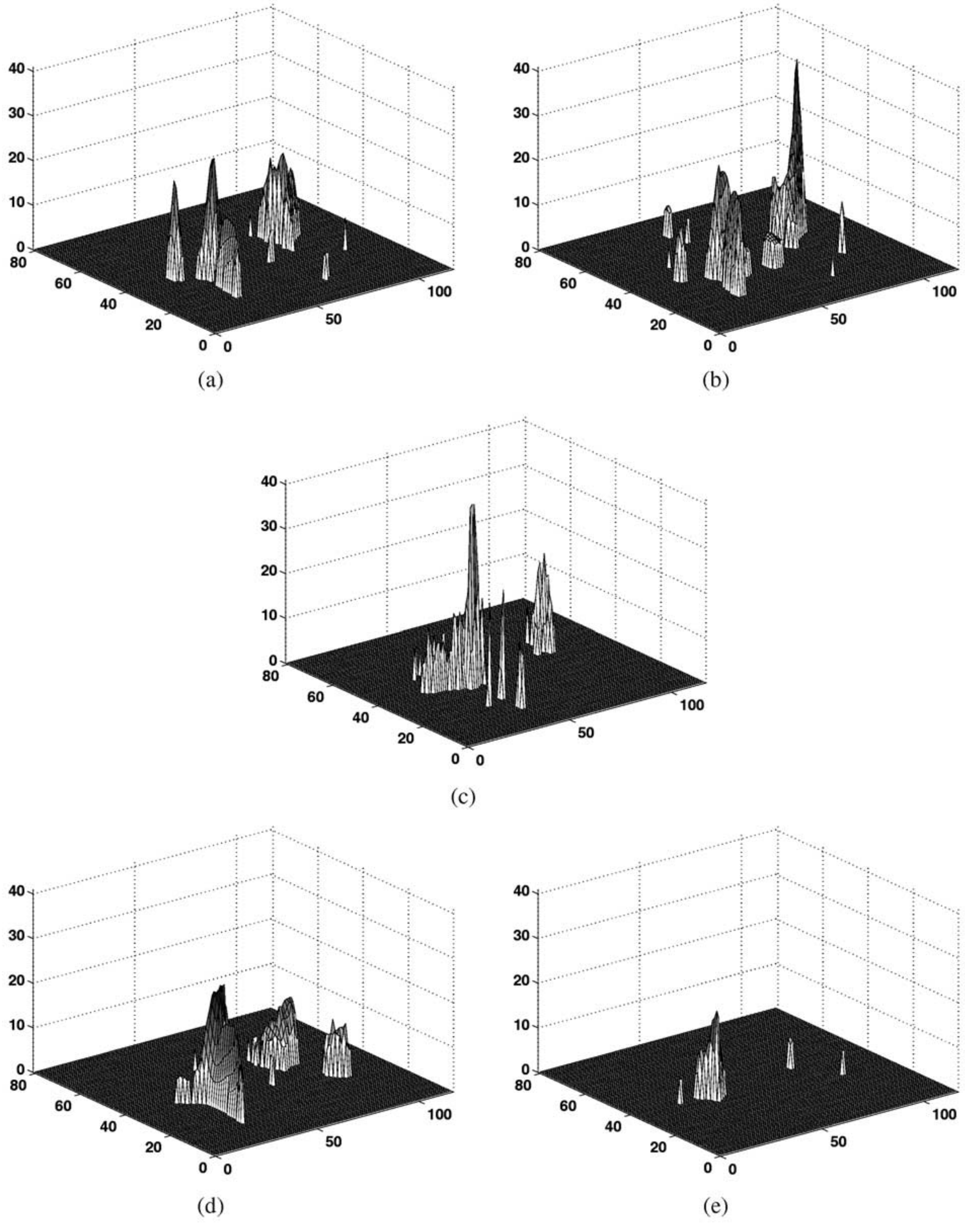

Figure 8. Comparison of (a, b, d, e) SAMEX '98 multimodel forecasts over the continental United States with (c) the observed precipitation field.

(QPF) a challenging task but also renders QPF verification nontrivial. Some of the more common measures of QPF verification include threat score, equitable threat score (EqTh), brier score and bias score. Simple coefficient of correlation and root mean squared error (RMSE) are also used quite frequently to delineate error growth curves and assess limits of predictability of precipitation. These commonly used measures can be classified under two categories: amplitude-based (e.g., RMSE) and distance-based (e.g., EqTh). In this work, we proposed a new measure

Table 8. Ranking of Four Forecasts According to Visual Inspection and the Three Measures Discussed in the Text for the SAMEX '98 Experiment

	Best			Worst
Visual	F1	F3/F2	F2/F3	F4
RMSE	F4 (1.53)	F1 (2.06)	F3 (2.33)	F2 (2.47)
EqTh	F1 (0.15)	F4 (0.15)	F2 (0.12)	F3 (0.10)
FQI	F1 (0.18)	F2 (0.20)	F3 (0.24)	F4 (0.54)

for QPF verification, called forecast quality index (FQI), which combines both types of measures (amplitude and distance). The distance-based measure we used is based on a nonlinear metric called the Hausdorff distance which was modified in our work to account for robustness to outliers. The amplitude-based measure we used is based on a newly developed metric called the universal image quality index. The proposed combined measure takes advantage of both metrics and depicts differences in both intensity and location of two rainfall intensity fields.

[35] Precipitation fields from an ARPS ensemble were compared with radar-observed fields using RMSE, EqTh and FQI, and the merit of the proposed measure in QPF verification was demonstrated. Also, we used data from the SAMEX '98 experiment to illustrate the potential applicability of the proposed measure for large-scale multimodel forecast comparison. Further analysis is needed using (1) larger ensemble sizes; (2) forecast observation pairs for several time instants; and (3) different types of events to further strengthen our conclusions and establish the utility of the proposed measure toward the purpose of QPF verification (especially in ensemble forecasts). Robust and discriminatory QPF verification metrics are expected to play an important role in ensemble forecasting research, as for example, in better understanding the nature of ensemble diversity (''dynamic'' versus ''statistical'') and also in gaining insight into the problem of the ''best ensemble member'' (i.e., as to whether a ''best'' member at an early stage of the forecast remains the ''best'' throughout the forecast period).

[36] **Acknowledgments.** This work was partially supported by NSF (grant ATM-0130394) and NASA (grants NAG5-12909 and NAG5-13639). The authors wish to thank Kelvin Droegemeier and Fanyou Kong for providing the ARPS ensemble, and also for valuable discussions during the course of this work. Fruitful discussions with Boyko Dodov on the Hausdorff distance are greatly appreciated. Computational resources were provided by the Minnesota Supercomputing Institute, Digital Technology Center, University of Minnesota. For the interested reader, Matlab™ scripts to compute the numerator (*PHD* and surrogates) and denominator of the proposed measure are available from http://home.safl.umn.edu/vvuruputur/matlabscripts/fqi.tar.gz.

References

Bright, D. R., and P. A. Nutter (2004), On identifying the ''best'' ensemble member in operational forecasting, paper presented at 84th Annual Meeting of the American Meteorology Society, Seattle, Wash.

Buizza, R. (1997), Potential forecast skill of ensemble prediction, and spread and skill distributions of the ECMWF ensemble prediction system, *Mon. Weather Rev.*, *125*, 99–119.

Buizza, R., and T. N. Palmer (1995), The singular vector structure of the atmospheric general circulation, *J. Atmos. Sci.*, *52*, 1434–1456.

Ebert, E. E., U. Damrath, W. Wergen, and M. E. Baldwin (2003), The WGNE assessment of short-term quantitative precipitation forecasts, *Bull. Am. Meteorol. Soc.*, *84*(4), 481–492.

Ebisuzaki, W., and E. Kalnay (1991), Ensemble experiments with a new lagged average forecasting scheme, *WMO Rep. 15*, World Meteorol. Org., Geneva.

Gesù, V. D., and V. Starovoitov (1999), Distance-based functions for image comparison, *Pattern Recognition Lett.*, *20*, 207–214.

Golub, G. H., and C. F. Van Loan (1996), *Matrix Computations*, 694 pp., Johns Hopkins Univ. Press, Baltimore, Md.

Hou, D., E. Kalnay, and K. K. Droegemeier (2001), Objective verification of the SAMEX '98 ensemble forecasts, *Mon. Weather Rev.*, *129*, 73–91.

Huttenlocher, D. P., R. H. Lilien, and C. F. Olson (1999), Object recognition using subspace methods, *IEEE Trans. Pattern Anal. Mach. Intell.*, *21*(9), 951–956.

Jolliffe, I. T., and D. B. Stephenson (Eds.) (2003), *Forecast Verification: A Practitioner'S Guide in Atmospheric Science*, John Wiley, Hoboken, N. J.

Kantz, H., and T. Schreiber (1997), *Nonlinear Time Series Analysis*, Cambridge Univ. Press, New York.

Kong, F., K. Droegemeier, V. Venugopal, and E. Foufoula-Georgiou (2004), Application of scale-recursive estimation to ensemble forecasts: A comparison of coarse and fine resolution simulations of a deep convective storm, paper presented at 20th Conference on Weather Analysis and Forecasting, 16th Conference on Numerical Weather Prediction, Am. Meteorol. Soc., Boston, Mass.

Krause, E. F. (1986), *Taxicab Geometry: An Adventure in Non-Euclidean Geometry*, Dover, Mineola, N. Y.

Levit, N. L., K. K. Droegemeier, and F. Kong (2004), High-resolution storm-scale ensemble forecasts of the 28 March 2000 Fort Worth tornadic storms, paper presented at 20th Conference on Weather Analysis and Forecasting, 16th Conference on Numerical Weather Prediction, Am. Meteorol. Soc., Boston, Mass.

Miller, M. (2000), Verification of precipitation and forecast usefulness for SAMEX mesoscale ensemble forecasts, M.S. thesis, 95 pp., Univ. of Okla., Norman.

Moeckel, R., and A. B. Murray (1997), Measuring the distance between time series, *Phys. D*, *102*, 187–194.

Pappas, T. N., and R. J. Safranek (2000), Perceptual criteria for image quality evaluation, in *Handbook of Image and Video Processing*, edited by A. C. Bovik, pp. 669–684, Elsevier, New York.

Roulston, M., and L. Smith (2003), Combining dynamical and statistical ensembles, *Tellus, Ser. A*, *55*, 16–30.

Rucklidge, W. (1996), *Efficient Visual Recognition Using the Hausdorff Distance*, 178 pp., Springer, New York.

Schreiber, T., and A. Schmitz (1996), Improved surrogate data for nonlinearity tests, *Phys. Rev. Lett.*, *77*, 635–638.

Toth, Z., and E. Kalnay (1993), Ensemble forecasting at NMC: The generation of perturbations, *Bull. Am. Meteorol. Soc.*, *74*, 2317–2330.

Toth, Z., and E. Kalnay (1997), Ensemble forecasting at NCEP and the breeding method, *Mon. Weather Rev.*, *125*, 3297–3319.

Wang, Z., and A. C. Bovik (2002), A universal image quality index, *IEEE Signal Process. Lett.*, *9*, 81–84.

Xue, M., D. Wang, J. Gao, K. Brewster, and K. K. Droegemeier (2003), The Advanced Regional Prediction System (ARPS), storm-scale numerical weather prediction and data assimilation, *Meteorol. Atmos. Phys.*, *76*, 143–165.

S. Basu, E. Foufoula-Georgiou, and V. Venugopal, St. Anthony Falls Laboratory, University of Minnesota, Mississippi River at 3rd Avenue SE, Minneapolis, MN 55414, USA. (basus@msi.umn.edu; efi@umn.edu; venu@msi.umn.edu)