JOURNAL OF GEOPHYSICAL RESEARCH, VOL. 106, NO. D11, PAGES 11,775–11,784, JUNE 16, 2001

Scale issues in verification of precipitation forecasts

Ben Tustison, Daniel Harris, and Efi Foufoula-Georgiou

St. Anthony Falls Hydraulic Laboratory, Department of Civil Engineering, University of Minnesota
Minneapolis, Minnesota

Abstract. Precipitation forecasts from numerical weather prediction models are often compared to rain gauge observations to make inferences as to model performance and the "best" resolution needed to accurately capture the structure of observed precipitation. A common approach to quantitative precipitation forecast (QPF) verification is to interpolate the model-predicted areal averages (typically assigned to the center point of the model grid boxes) to the observation sites and compare observed and predicted point values using statistical scores such as bias and RMSE. In such an approach, the fact that the interpolated values and their uncertainty depend on the scale (model resolution) of the values from which the interpolation was done is typically ignored. This interpolation error, which comes from scale effects, is referred to here as the "representativeness error." It is a nonzero scale-dependent error even for the case of a perfect model and thus can be seen as independent of model performance. The scale dependency of the representativeness error can have a significant effect on model verification, especially when model performance is judged as a function of grid resolution. An alternative method is to upscale the gauge observations to areal averages and compare at the scale of the model output. Issues of scale arise here too, with a different scale dependency in the representativeness error. This paper examines the merits and limitations of both verification methods (area-to-point and point-to-area) in view of the pronounced spatial variability of precipitation fields and the inherent scale dependency of the representativeness error in each of the verification procedures. A composite method combining the two procedures is introduced and shown to diminish the scale dependency of the representativeness error.

1. Introduction and Problem Statement

Modeling or forecasting of precipitation and other atmospheric and hydrologic variables is necessary for many applications over a wide range of space and time scales. These include flash flood forecasting over small basins, assessment of interseasonal to decadal climate variability at the continental, regional, and basin scale, and assessment of global impacts of climatic anomalies. Precipitation forecasts are also produced operationally as, for example, in the United States by the National Centers for Environmental Prediction (NCEP) at scales (pixel size) of the order of 20-30 km or by the Center for the Advanced Prediction of Storms (CAPS) at scales of the order of 3-32 km. An obvious concern of all these modeling efforts is the assessment of how well precipitation fields predicted by the model compare to the observed precipitation fields. Efforts in model verification methodologies are generally lagging behind those invested in model development (e.g., see discussion in a recent paper by *Zepeda-Arce et al.* [2000] where an effort was presented to complement typical verification measures, such as threat score and bias score, with multiscale statistical measures of performance).

The present paper focuses on a very fundamental issue of model verification, namely the comparison of model outputs (which consist of areal averages over the model grid boxes) to point observations available through rain gauges. Such comparisons are commonplace in the literature (e.g., see *Xie and Arkin* [1996] and *Colle et al.* [1999] for some recent studies) and serve several purposes. Among these are (1) comparison of the performance of different forecast models, (2) assessment of model improvements when new parameterizations are introduced, (3) assessment of model performance when the resolution of the model is changed, and (4) estimation of errors of reanalysis products when model outputs are merged with observations to produce gridded fields for model initialization and other uses.

The major problem arising when spatial averages are compared to point values, is that of discrepancy of scales. A spatially averaged field always has lower variability than point values, and the degree to which the variability is reduced depends on the scale of the spatial averaging and the inherent inhomogeneity of the original field. The change in variability with scale is illustrated in Figure 1, which displays the reduction of the standard deviation of a radar observed precipitation field when the field is viewed at different scales, i.e., when averages at increasing scales are considered from the smallest scale of 2 km up to a scale of 64 km.

In order to compare the model output and the rain gauge observations, one set of data (either the observations or the model output) must be transformed to the scale of the other with some form of spatial interpolation or spatial averaging. One way in which this is done is by an area-to-point (A-P) conversion. That is, each forecast output (an areal average) is assigned to a point in the center of each grid block, and these

Paper number 2001JD900066.
0148-0227/01/2001JD900066$09.00

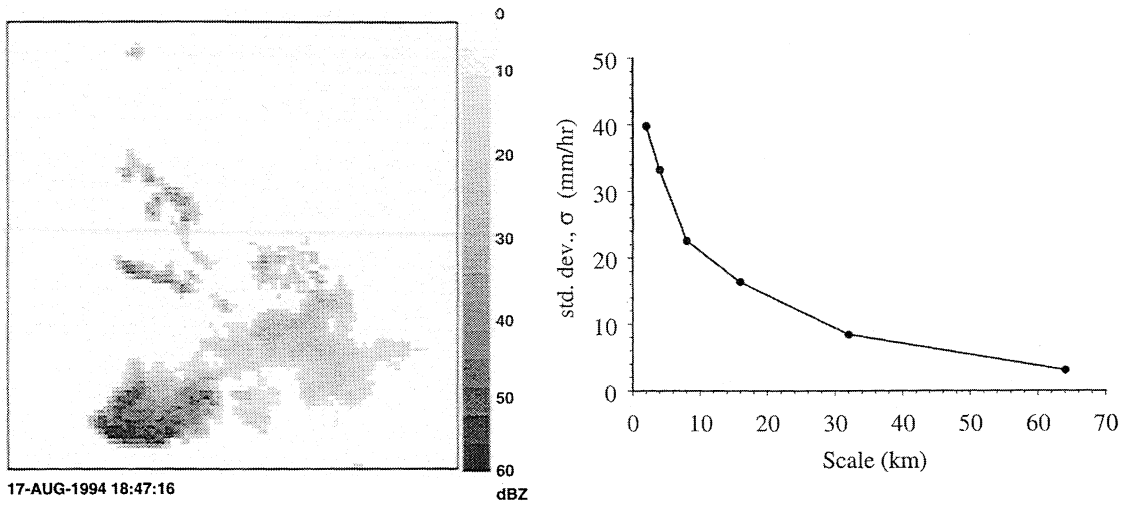

Figure 1. (left) Radar observed precipitation from the NEXRAD KICT radar on August 17, 1994, and (right) the standard deviation of nonzero precipitation as a function of grid scale. A trend of decreasing variability of the field is observed with increasing scale. Note that it is exactly this difference in the respective variabilities of the point observations and the grid-average values that is the driving force for the scale dependent representativeness error term in comparing model output with rain gauges.

point values are interpolated to get point values at the location of the observations. In turn, these interpolated values are compared to the observations to compute statistical scores quantifying the error in the quantitative precipitation forecast (QPF). Another type of verification method involves a point-to-area (P-A) conversion. This is done by taking the rain gauge observations, interpolating them to a regular grid, and computing average block values (the grid points may be seen as the corner points of blocks and the average is taken of the four corners) which are then compared to the grid block values from the forecast to compute the error in the QPF.

A potential problem with these verification methods is that changing the scale of the point observations to match the scale of the model output (P-A conversion) or vice-versa (A-P conversion) imposes a "representativeness error" which is independent of the error in the model forecast. The representativeness error is the error in representing data (i.e., either model output or observations) at a scale other than their own inherent scale. For example, taking an output grid block from a QPF run at 32 km and assigning its value to the center of the block as a point value imposes a representativeness error because the inherent scale of the output block is 32 km rather than the point scale. In general, there may be a significant representativeness error in going from one areal grid size to a different areal grid size as one may have, for example, in comparing large-scale model output to smaller-scale radar observations. In this study, one of the scales is fixed (the point scale) so that the representativeness error represents the error in either a P-A or A-P conversion.

The representativeness error would not be so problematic if it were the same across all scales. However, it is not the same, and thus it becomes a problem when trying to assess changes in model performance when the grid resolution of the model is changed. This can be illustrated with the following example. Consider two model runs, one at 8 km and one at 64 km resolution. If an area-to-point verification scheme is used, we take the forecast output grid block values and assign them to the center points of their respective blocks. Then we interpolate these values to the locations of the point observations, and the error in the QPF is computed by comparison of the interpolated and observed values. However, as previously mentioned (see also Figure 1), the output block values will have less variability than the point observations. Thus when the grid block values are assigned as point values and interpolated, they will never be able to reproduce the variability of the point observations. Also, the output blocks from the model run at 64 km have less variability than those from the model run at 8 km, so the 8 km model is better able to reproduce the variability of the point observations. In other words, the verification of the 8 km model always gives less representativeness error, following an A-P conversion, than the verification of the 64 km model. This improvement in error however, is solely due to statistical considerations and must be separated from the improvement in error that comes from truly better physical performance of the 8 km model relative to the 64 km model.

Many verification schemes currently in use ignore this representativeness error and its separation from the total error.

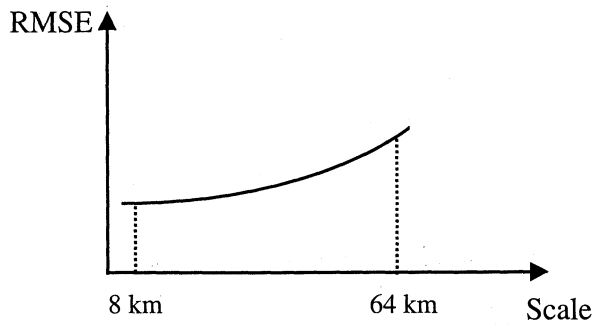

Figure 2. Schematic illustration of typical RMSE (total error) versus model resolution plots constructed in QPF validation. An important question arises: Is all of the RMSE reduction from 64 to 8 km due to model improvements? This question may not be answered without some indication of the representativeness error at these scales imposed by the verification methodology.

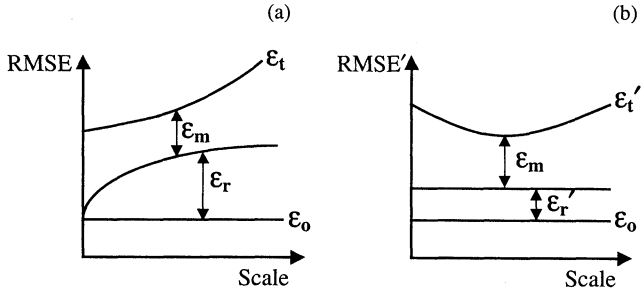

Figure 3. Schematic illustration of the breakdown of the total error, ε_t, as the sum of the observational (ε_o), representativeness (ε_r), and model error (ε_m). (a) The area-to-point verification procedure and the danger in using the total error as a measure of model performance due to the fact that the minimum total error and minimum model error might not occur at the same scale is illustrated. (b) An alternative verification procedure which results in a scale independent representativeness error (ε_r') would make decisions on model performance more straightforward as here the change in total error with scale could solely be attributed to the change in model error with scale.

For example, plots such as the schematic shown in Figure 2 are common in the literature and the reduction of the RMSE with scale is attributed to better model performance as the resolution of the model is increased. However, in any verification study which uses spatial interpolation to transform data to different scales, the total error ε_t is what is computed. This is the sum of observational error, ε_o, model error, ε_m, and representativeness error, ε_r. For models run at different scales, the observational error for the A-P verification procedure is the same for all scales because it refers to the error in the point precipitation observations (assuming the same point data are used for the verification of both models). The total error in the example of Figure 2 increases with scale; however, the representativeness error also increases with scale, as argued above, for an A-P verification. Thus the difficulty lies in trying to separate the model error, whose quantification is the target in QPF verification, from the representativeness error (see the schematic in Figure 3a). This may only be done if the representativeness error and its scale dependence are known a priori or are altogether negligible. As an alternative, one could introduce new verification procedures which result in scale-independent representativeness error such that the change in total error with scale could all be attributed to model error, thus allowing direct model performance inferences with scale (see Figure 3b). The aim of this paper is twofold; (1) to characterize the representativeness error in the area-to-point and point-to-area methodologies as a function of scale and spatial structure of the underlying precipitation field and (2) to propose a simple modification to existing procedures, which results in a scale independent representativeness error.

2. Theoretical Considerations on Representativeness Error

In general, the representativeness error depends on (1) the statistical structure of the underlying field, (2) the geometrical configuration or location of the sampled point observations, (3) the interpolation scheme used to estimate unknown point values from surrounding known point values, and (4) the scale of averaging (which in verification applications would be the model resolution). The following are some features of representativeness errors, which are argued here based on theoretical considerations but are quantified later in the paper via numerical experiments:

1. The higher the density of point observations and therefore the smaller the mean gauge spacing, L_g, the smaller the representativeness error in the point-to-area method ($\varepsilon_r^{P\text{-}A}$), since the accuracy of areal averages will strongly depend on how many point values were available to compute these averages. By contrast, L_g is not expected to significantly affect the representativeness error in the area-to-point method ($\varepsilon_r^{A\text{-}P}$), since the starting information is always in the form of the same areal averages, and it does not matter how many point estimates are computed from the same starting information.

2. The representativeness error in the point-to-area method $\varepsilon_r^{P\text{-}A}$ will decrease with increasing scale due to the decreasing variability in the difference between areal averages of the underlying field and the areal estimates obtained from averaging of the interpolated point observations. At small scales, the areal estimates, coming from the interpolation and averaging (both variability reducing operations) of the point observations, will not be variable enough to accurately estimate the true areal averages. As the scale of averaging becomes larger, the true areal averages themselves will become less variable and will be more accurately estimated by the averages coming from the point observations. By contrast, $\varepsilon_r^{A\text{-}P}$ increases with scale because the increase in scale creates a larger difference between the point estimates from areal averages and the sampled point observations. This difference arises because as the scale of the areal averages increases, they become less variable and therefore less able to capture the extremes of the point observations, which have much larger variability.

3. The more accurate the interpolation scheme, the smaller the representativeness error (regardless of the method). In other words, in the limit of an interpolation scheme which is based on complete knowledge of the multiscale statistical structure of the underlying field, the interpolation would result in a minimized and almost scale-independent representativeness error.

4. The smoother the field (which, as discussed later, manifests itself as a faster decay in the Fourier power spectrum of the field), the smaller the representativeness error for both methods. On the contrary, the less smooth or noisier the field (slower decaying spectrum), the harder it is to accurately estimate unknown values via interpolation and the larger the representativeness error.

The Fourier power spectrum is a readily computable measure for characterizing the multiscale structure of a field. For fields such as rainfall for which evidence exists for power-law spectra, i.e., log-log linearity in the spectrum [e.g., *Lovejoy and Schertzer*, 1995; *Georgakakos et al.*, 1994; *Harris et al.*, 1996, 1997, 2001], the spectral slope, β, of the power spectrum on a log-log plot can be used as an indicator of smoothness, with high spectral slopes characteristic of a smoother field [e.g., *Davis et al.*, 1996; *Harris et al.*, 1996]. As already mentioned, it is expected that for fields with small spectral slope the representativeness error will be large. To illustrate this consider the extreme case of $\beta = 0$ indicating that the field behaves as (fully uncorrelated) white noise. In this case it is extremely hard to estimate point values from areal averages and vice versa resulting in high representativeness errors. It is also worth considering that typical interpolation methods applied to highly variable fields (such as rainfall) always result in smoothed representations of the underlying field. Thus the

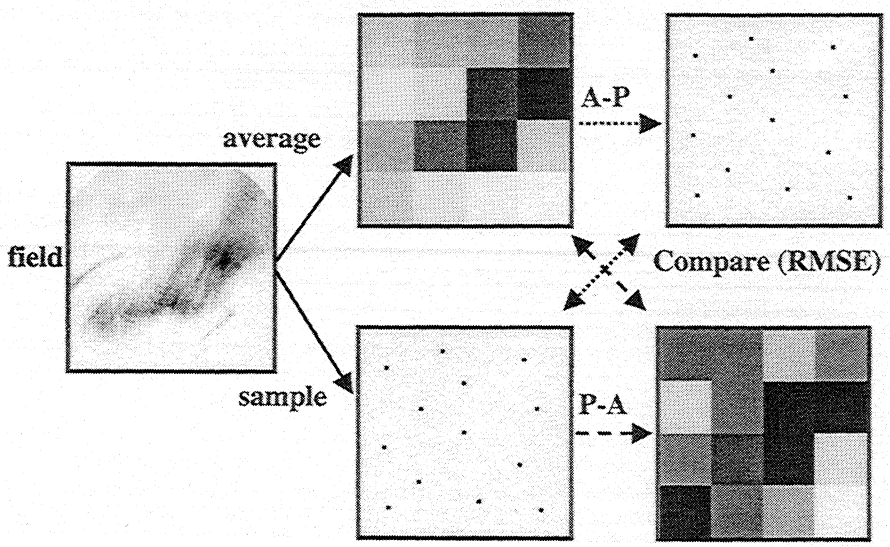

Figure 4. Schematic showing the methodology for the numerical experiment. It illustrates how perfect (no error) model outputs and perfect observations are created to be used in the numerical experiment, as these are averages and samples, respectively, of the same underlying field. Long dashed lines represent the point-to-area verification procedure, and short dashed lines represent the area-to-point verification procedure.

smoother the true underlying field (higher β), the less "harm" is done by the smoothing nature of the typical interpolation methods, therefore resulting in smaller representativeness errors.

Having these in mind it is desirable to construct a controlled experiment by which the representativeness error, and its features intuitively discussed above, can be quantified. This experiment is described in section 3.

3. Numerical Experiment to Quantify Representativeness Error

In order to quantify the representativeness error, it first has to be isolated from the total error. For this purpose, a numerical experiment was constructed in which both the observational and model errors were zero by design. This was accomplished by starting with a perfectly known field, sampling the field to represent point observations (zero observational error) and taking spatial averages of the field to represent model output at different scales (zero model error). In essence, a perfect model and perfect observations were created, leaving only the representativeness error term in the total comparison error. Figure 4 shows a schematic of the methodology followed in the numerical experiment. For the area-to-point (A-P) verification, the original field was averaged at different scales to represent perfect model outputs at varying grid resolutions. These averages were assigned to the centers of the respective boxes and were used to interpolate to the locations of the point observations (sampled points). The RMSE between the interpolated point values and the sampled point observations was computed and was taken to be the representativeness error in the A-P method. In the point-to-area (P-A) verification, the sampled point observations were used to estimate, via interpolation, the field at each of the four corner pixels of each block. These four corner estimates were then arithmetically averaged to obtain the estimated block average values. The RMSE between these estimated block values and the true block values (found again by averaging the underlying field to varying grid resolutions) was computed and taken to be the representativeness error in the P-A method.

Since the RMSE in the estimation of precipitation increases with increasing rain rate [*Huffman*, 1997], the representativeness error is dependent on the particular storms or more specifically their particular rain rates. In order to alleviate this problem and have our results applicable to hourly precipitation fields beyond those used in this study, the representativeness error was made dimensionless by dividing it by the spatial conditional mean (i.e., mean of the nonzero values) computed from fields at 2 km resolution. In the rest of the paper, only dimensionless representativeness errors will be considered, and they will be referred to simply as representativeness errors (i.e., dimensionless will be implied). We will use $\varepsilon_r^{A\text{-}P}$ to designate the dimensionless representativeness error arising from the A-P method and $\varepsilon_r^{P\text{-}A}$ that arising from the P-A method.

In this study, interpolations were performed using the inverse-distance Barnes weighting scheme [*Barnes*, 1964], commonly used in meteorology [e.g., *Krishnamurti and Bounoua*, 1996]. The Barnes method is similar to the Cressman method, which is often implemented for model verification studies [e.g., *Colle et al.*, 1999]. Typically in the A-P verification method, an estimate is obtained at each of the gauge locations from a weighted linear combination of the four nearest gridded values (i.e., represented as points at the centers of the model output boxes). In an effort to incorporate all the information that would enhance the accuracy of the interpolation, rather than just the four nearest values, any value falling within a specified radius of influence, L_c, was used in the interpolation for both methods. If no observation was found within the radius of influence, the nearest known value was found and given a weight of one for the interpolation (i.e., nearest neighbor approach). With the Barnes method, the weight w_i each observation receives is defined as

$$w_i = \exp\left(-4d_i^2 / L_c^2\right), \tag{1}$$

where d_i is the distance from the observation to the estimated

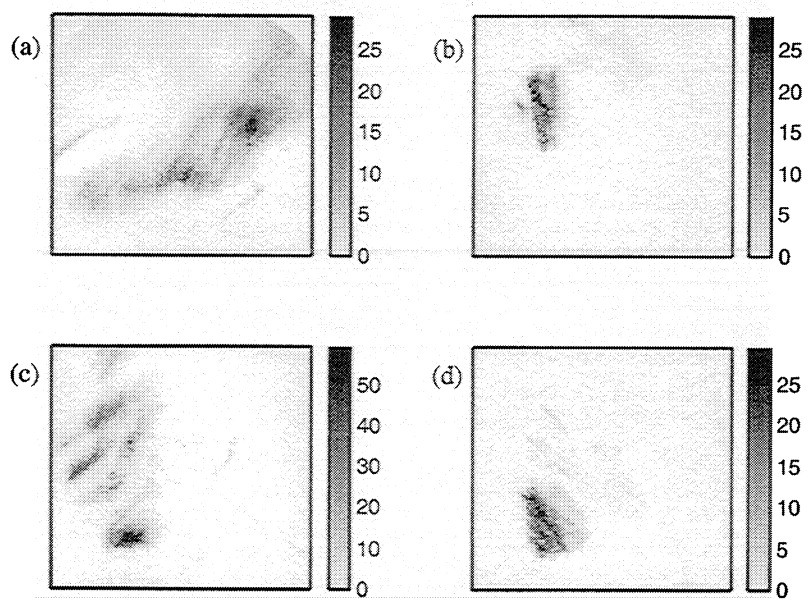

Figure 5. Plot of the four radar derived rainfall accumulation fields used in this study. Each image is a 1-hour accumulation of precipitation (in millimeters) converted from radar-observed reflectivities. Images are (a) KEAX (July 4, 1995), (b) KTLX (August 17, 1994), (c) KAMA (June 3, 1999), and (d) KICT (August 17, 1994).

point. The Barnes scheme estimates a value P_x at location, x, based on the N values falling within that point's radius of influence as

$$P_x = \sum_{i=1}^{N} P_i w_i \Big/ \sum_{i=1}^{N} w_i \,. \qquad (2)$$

The value of the radius of influence, L_c, is typically set to the distance at which values in a field begin to lose significant correlation with one another. In this study, it was chosen to be the distance (lag) at which the isotropic spatial autocorrelation function dropped to 0.35 (which is approximately equal to the value of 1/e theoretically applicable for exponentially decaying autocorrelation functions [e.g., *Bras and Rodriguez-Itrube*, 1993]).

To avoid having the results dependent on a specific gauge network, an ensemble of 100 gauge networks was created for each gauge density, quantified by the mean gauge spacing, L_g, within each of the fields. By using a variety of gauge densities, this study addresses the issue of sampling variability, which is usually ignored in forecast verification [*Murphy and Wilks*, 1998]. The mean and variance of the representativeness error over the ensemble was then computed. The network of gauges was assumed randomly uniform. Clustered networks (e.g., having most stations at the corner of a block) were not considered as these would require case-dependent solutions and would divert attention from the main focus on scale.

4. Analysis and Results

Radar derived precipitation patterns were used as the underlying fields on which the numerical experiment depicted in Figure 4 was applied. These fields were at 2 km resolution and represented hourly accumulations of precipitation converted from radar-observed base-scan reflectivities for four different radars in the United States. Reflectivity maps were converted to rain rate images using a Z-R relationship of the form $R = aZ^b$, with $a = 0.017$, $b = 0.714$ [*Smith et al.*, 1996], and where Z is in mm^6/m^3 and R is in mm/hr. Hourly accumulations were estimated as simply the sum of the instantaneous rain rate images multiplied by the time between scans (5 min). The accumulations from four individual radar sites are shown in Figure 5. Note that the fields of Figures 5b and 5d relate to each other as they are from the same storm at different times.

As already mentioned in section 3, the spectral slope was used to quantify the underlying spatial variability structure of each rainfall hourly accumulation field. It can be seen in Figure 6 that the power spectra of these fields are well approximated as log-log linear over a finite range of scales, with β, the slope of the power spectrum, being a quantifier of their

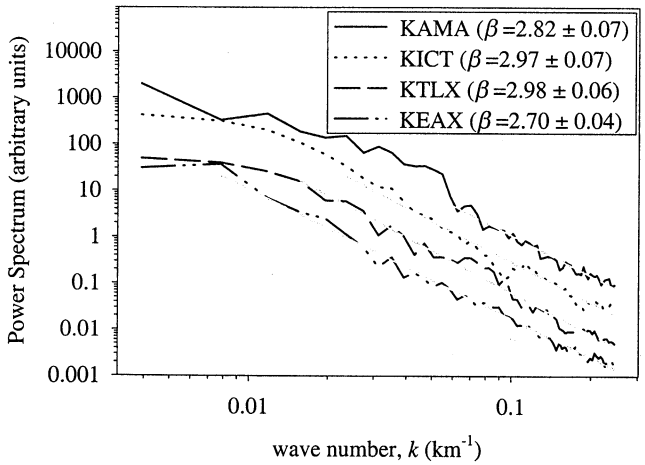

Figure 6. Log-log plot of the power spectrum as a function of frequency for the four rainfall images with individual spectra shifted for clarity. Notice that each of the images exhibits approximate log-log linear behavior over a finite range, with the spectral slope, β, characterizing the field smoothness. The length of the regression lines indicates the range of scaling for each field.

Table 1. L_c Values for the Four Radar Fields.

Radar Field	L_c, km
KAMA	15.5
KICT	21.7
KTLX	18.6
KEAX	30.7

L_c is the distance at which the spatial autocorrelation coefficient drops to 0.35

smoothness [*Davis et al.*, 1996; *Harris et al.*, 1996]. Estimates of β values ranged from 2.77 for the KEAX image (the least smooth of the fields) to 2.98 for the KTLX image (the most smooth).

Both the P-A and A-P representativeness errors were computed for all four fields for scales of 4, 8, 16, and 32 km, which correspond to the typical range of resolutions at which current mesoscale models are run. Owing to the limited size of the radar images, it was not possible to perform the analysis at scales larger than 32 km, as there was not a large enough sample size of blocks to yield statistically meaningful errors for the P-A method. The radii of influence, L_c, used in the interpolation scheme (equation (1)) are listed in Table 1 and range from 15.5 km for the KAMA image to 30.7 for the KEAX image. In addition, representativeness errors were computed for varying gauge densities corresponding to average distance between gauges, L_g, equal to 25, 50, and 75 km. In order to place these average gauge spacings in the proper context, one can consider that the Oklahoma Mesonet has an L_g value of 40.6 km [*Brock et al.*, 1995]. So the L_g values chosen for this study range from extremely dense networks (25 km) to networks closer to those typically found for the United States (75 km). The results of these analyses are shown in Figures 7-9 and exhibit all of the expected features mentioned in section 3, as is now discussed.

For the A-P verification procedure, the representativeness error increases with scale as seen in Figure 7a, which can be attributed to the fact that the error in assigning an areal average (i.e., block) value to a point becomes larger as the block area is increased. In other words, there is a larger error in representing a 32-km average as a point observation than there is in representing an 8-km average as a point observation, owing to the larger difference in variability between a 32-km averaged field and point observations than an 8-km averaged field and point observations (see Figure 1). Recall that the reported representativeness error is normalized by the conditional mean (i.e., mean of the nonzero values) computed at 2 km resolution. Thus values of $\varepsilon_r^{A\text{-}P}$ equal to 0.5 correspond to a representativeness error equal to 50% of the mean, which is a considerable error coming only from scale effects.

The opposite trend was found for the $\varepsilon_r^{P\text{-}A}$ versus scale curves in Figure 7b where the representativeness error is shown to decrease with model scale. This is due to the fact that there is a reduction in the variability of precipitation with increasing scale. The rainfall values become smoother at larger scales owing to the averaging process and therefore comparatively easier to estimate. It is important to recall that the simulated model blocks are being compared with the mean of the four interpolated values, which themselves have reduced variability inherited through the interpolation procedure (deterministic distance-weighted interpolation methods cannot return a value which is outside the range of the values used in the interpolation).

The uncertainty in the representativeness errors for the A-P and P-A verification methods, taken to be the standard deviation of the 100 ensemble members, was also computed for each of the model scales analyzed. The uncertainty for the A-P method can be seen in the error bars of Figure 8a which shows the dimensionless representativeness error plot for the KEAX radar image extracted from Figure 7a. This uncertainty ranges from 0.116 at the 4 km model scale to 0.220 at the 32 km model scale. From this plot, it is easy to see that not only the mean representativeness error but also its uncertainty increases with model scale. Furthermore, the increase in uncertainty is roughly proportional to the increase in the magnitude of the representativeness error. Similarly, the uncertainty in the P-A method can be seen in the error bars of Figure 8b which shows the P-A dimensionless representativeness error plot for the KEAX radar image extracted from Figure 7b. This plot shows that the uncertainty in the P-A method decreases with increasing scale. Additionally, like for the A-P method, the uncertainty in the representativeness error appears to be proportional to the magnitude of the representativeness error itself (a minimum value of 0.118 at 32 km model scale and a maximum of 0.149 at 4 km model scale). The same observations as above were extracted from the plots of the other stations but are not presented here for lack of space.

By varying the average spacing between point observations, the effect of gauge density on the representativeness error was investigated. The results for only the KEAX radar image are plotted in Figure 9. The other radar images show similar trends. As seen in Figure 9a, the gauge density had

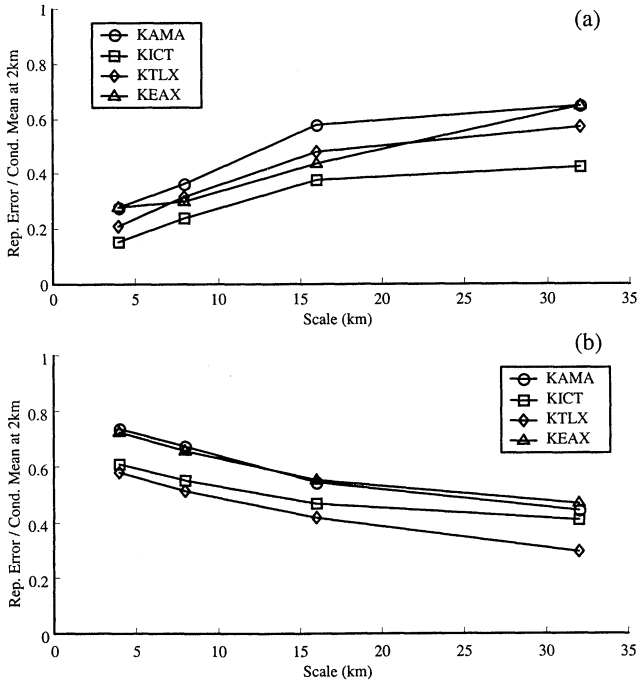

Figure 7. Normalized representativeness error versus model scale curves for the four storms. Each curve represents an ensemble average of 100 different gauge networks. The average gauge spacing was 50 km for each network. (a) Area-to-point error $\varepsilon_r^{A\text{-}P}$; notice the trend of increasing representativeness error with increasing scale for all four images. (b) Point-to-area error $\varepsilon_r^{P\text{-}A}$; notice the trend of decreasing representativeness error with increasing scale for all four images.

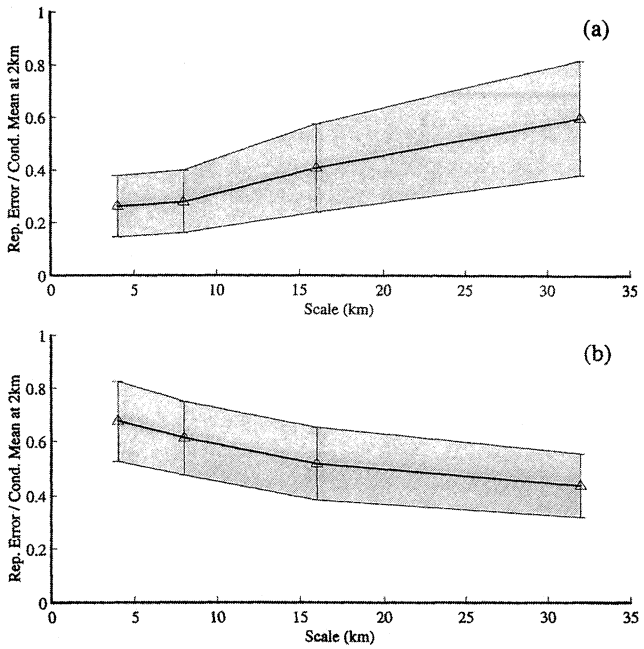

Figure 8. Normalized representativeness error versus model scale curve for the KEAX storm. These are the same KEAX curves as seen in Figures 7a and 7b but with uncertainty bars (± one standard deviation) added to represent the uncertainty in the mean of the 100 member ensemble. Recall that the average gauge spacing was 50 km for each network. (a) Area-to-point error $\varepsilon_r^{A\text{-}P}$; notice the increasing uncertainty with increasing scale. (b) Point-to-area error $\varepsilon_r^{P\text{-}A}$; notice the (slightly) decreasing trend with increasing scale.

minimal effect on $\varepsilon_r^{A\text{-}P}$ because additional gauges did not provide more information but simply gave additional sites on which to verify the model. In other words, for the A-P method, the values used for the interpolation come from the blocks, and changing the gauge density does not affect this information. However, from Figure 9b it can be seen that increasing the point observation density lowers $\varepsilon_r^{P\text{-}A}$. This is due to the fact that the block corners were able to obtain better estimates with higher gauge density because each block corner had more information at closer spatial locations to utilize for estimation.

5. Removal of Scale Dependency in the Representativeness Error

Earlier in this paper, we discussed how in the presence of a nonnegligible representativeness error, an ideal verification procedure would be one that would rely on a representativeness error which is constant with scale. Only then would it be possible to attribute a reduction in the total error with scale solely to a reduction in the model error, since the observational and representativeness errors would not change with scale. In this section we attempt to introduce a verification procedure which results in a scale independent representativeness error.

As intuitively argued in section 2 and quantified in section 4, the curves of $\varepsilon_r^{A\text{-}P}$ and $\varepsilon_r^{P\text{-}A}$ versus scale have opposite trends. That is, the P-A curve had its largest value at the smallest model scale and decreased with increasing model scale, while the A-P curve had its smallest value at the smallest model scale and increased with increasing model scale. With this in mind, the average of the two curves at different model scales was examined as a candidate for a verification measure which gives a scale independent representativeness error. In other words, verification could be performed by evaluating the composite root mean square error,

$$\mathrm{RMSE}_{\mathrm{comp}} = \frac{1}{2}\left(\mathrm{RMSE}_{P\text{-}A} + \mathrm{RMSE}_{A\text{-}P}\right). \tag{3}$$

In this study where the model error and observation errors are zero, $\mathrm{RMSE}_{\mathrm{comp}}$ is equal to the composite representativeness error, $\varepsilon_r^{\mathrm{comp}} = (\varepsilon_r^{A\text{-}P} + \varepsilon_r^{P\text{-}A})/2$.

Implementation of the A-P/P-A composite verification method, resulted in representativeness error versus model scale curves which were relatively flat. This can be seen in Figure 10a, which plots $\varepsilon_r^{\mathrm{comp}}$ for the case of L_g = 50 km. The uncertainty in the representativeness errors for the composite verification method, taken to be the standard deviation of the $\varepsilon_r^{\mathrm{comp}}$ values from the 100 ensemble members, was also computed for each of the model scales analyzed. This uncertainty can be seen in the error bars of Figure 10b, which shows the composite representativeness error plot for the KEAX radar image. This uncertainty ranges from 0.107 at the 4 km model scale to 0.137 at the 32 km model scale. It is observed that not only the mean composite representativeness error but also its uncertainty appear to change very little with model scale. Apart from the visual inspection discussed above, a more rigorous statistical test was also used to determine if $\varepsilon_r^{\mathrm{comp}}$ was, in fact, scale invariant as opposed to the other measures, $\varepsilon_r^{A\text{-}P}$ and $\varepsilon_r^{P\text{-}A}$. The hypothesis testing was based on performing weighted least squares linear regressions to the $\varepsilon_r^{A\text{-}P}$, $\varepsilon_r^{P\text{-}A}$

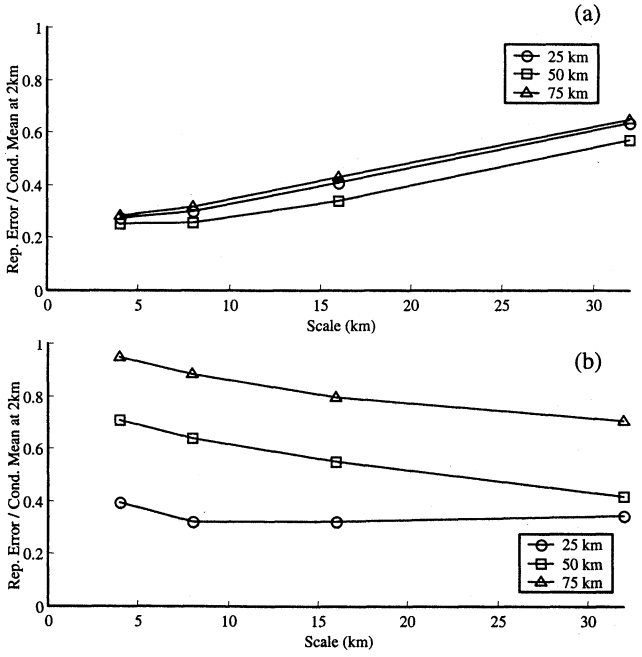

Figure 9. Normalized representativeness error versus model scale curves for the different gauge densities for the KEAX storm. Each curve represents an ensemble average of 100 different gauge networks. (a) Area-to-point error $\varepsilon_r^{A\text{-}P}$; notice that changing the gauge density has relatively little effect on the representativeness error. (b) Point-to-area error $\varepsilon_r^{P\text{-}A}$; notice that increasing the gauge density lowers the representativeness error because more information becomes available to estimate the areal averages.

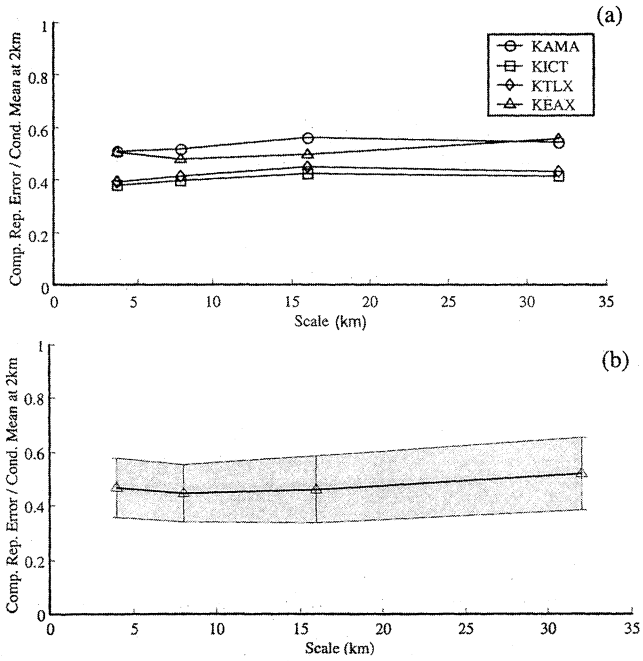

Figure 10. (a) Composite normalized representativeness error, ε_r^{comp} versus model scale curves for the four storms. Each curve represents an ensemble average of 100 different gauge networks. The average gauge spacing was 50 km for each network. Notice the relative flatness of the curves compared with those computed for A-P and P-A validation methods. (b) ε_r^{comp} versus model scale curve for the KEAX storm only with error bars added to represent the uncertainty in the ensemble mean. Notice that the uncertainty increases only slightly with increasing model scale and not as drastically as for the A-P method.

and also ε_r^{comp} curves with scale and testing whether the slope of the regression was statistically significant from zero at a desired confidence level. Weighted least squares (LS), as opposed to simple LS regression, was used to acknowledge the fact that the variance of ε_r was also changing with scale. The hypothesis tests (using a t test statistic) [e.g., *Draper and Smith*, 1981] were performed at the 95% confidence level. It was determined that the zero slope hypothesis was rejected for the $\varepsilon_r^{A\text{-}P}$ and $\varepsilon_r^{P\text{-}A}$ curves and accepted (failed to be rejected) for the ε_r^{comp} curve. The results of this test further confirm the scale invariance of ε_r^{comp} while discounting the possibility that $\varepsilon_r^{A\text{-}P}$ and $\varepsilon_r^{P\text{-}A}$ are also of this nature.

The composite representativeness error, ε_r^{comp}, as a function of model scale was examined for 25, 50, and 75 km average gauge spacings for the KEAX storm as shown in Figure 11. Since the composite measure is formed from an average of the A-P and P-A errors and the representativeness error changed very little for different gauge spacings in the A-P method, one observes as expected, that ε_r^{comp} shows similar trends to $\varepsilon_r^{P\text{-}A}$ of Figure 9b for the different gauge spacings, i.e., increasing the gauge density lowers ε_r^{comp}. In all observation densities, however, it is seen from Figure 11 (and was also confirmed with statistical testing) that ε_r^{comp} remains scale invariant.

6. Relation of the Representativeness Error to the Underlying Structure of Precipitation

As was intuitively discussed in section 2 and quantitatively verified in section 3, the smoother the precipitation field, the smaller the representativeness error for both A-P and P-A methods and therefore the smaller the composite representativeness error ε_r^{comp}. In this section an attempt is made to quantify this relationship by relating a measure of smoothness (spectral slope β) to the scale-independent value of ε_r^{comp}. Notice that prior to the introduction of the almost scale-independent ε_r^{comp}, such a quantification had to necessarily have a particular scale attached to it (owing to the scale dependency of $\varepsilon_r^{A\text{-}P}$ and $\varepsilon_r^{P\text{-}A}$) limiting thus the generality and insight obtained from this relationship. Figure 12 shows ε_r^{comp} versus β for nine precipitation fields. For this plot, five additional radar images were used to enlarge the range over which this relationship was characterized. The extra rainfall images were constructed in the same manner as previously mentioned and come from the same KEAX storm of July 4, 1995, at different, nonoverlapping times. The vertical error bars represent the uncertainties in the scale averaged ε_r^{comp}, and the horizontal error bars represent the uncertainty in the estimates of β. Recall that high values of β correspond to smoother precipitation fields. The plot of Figure 12 quantifies how the underlying statistical structure of a field relates to the representativeness error and is of significant practical value. For example, if for a given storm β was estimated to be around 2.8, then Figure 12 suggests that the composite representativeness error is considerable and is almost 60% of the storm spatial average precipitation. However, for a much smoother field for which β is ~ 3.0, the representativeness error is only ~ 25% of the spatial average precipitation.

7. Summary and Conclusions

Most current methods of QPF verification do not explicitly account for the fact that the variability and estimation uncertainty of precipitation fields depend on scale. Changing the scale of the observations to match the scale of the model output (point-to-area conversion) or vice-versa (area-to-point conversion) imposes a "representativeness error" which is nonzero even in the case of a perfect model and moreover is dependent on scale. (The representativeness error is the error in representing data, i.e., either model output or observations, at a scale other than their own inherent scale.) Especially in verification studies in which model performance is assessed as a function of model resolution, ignoring or mischaracterizing the scale dependent representativeness error can significantly

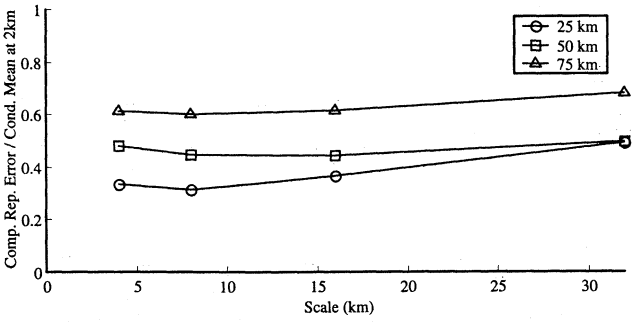

Figure 11. The ε_r^{comp} versus model scale curves for the different gauge spacings of the KEAX storm. Each curve represents an ensemble average of 100 different gauge networks. Notice that increasing the gauge spacing (lowering density) lowers the representativeness error but not to the degree seen for the P-A method.

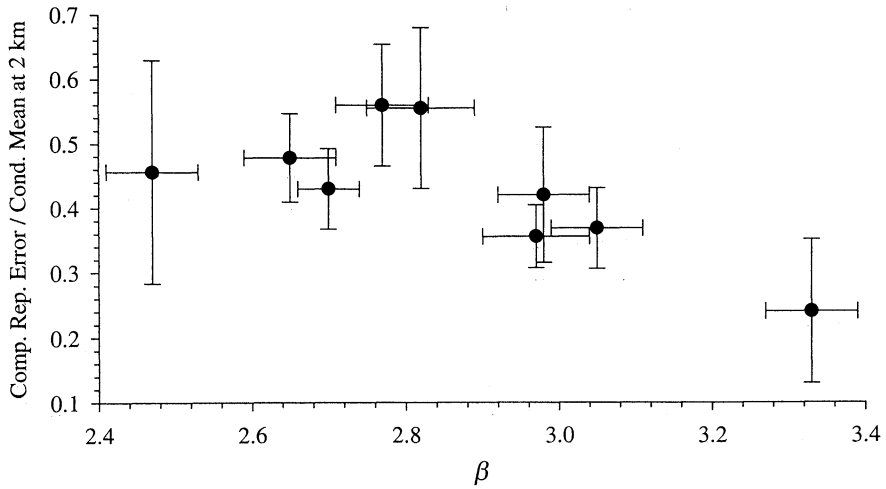

Figure 12. Scatter plot of ε_r^{comp}, averaged across all model scales, versus spectral slope β for nine storms. Each point represents an ensemble average of 100 different gauge networks. The vertical error bars represent the uncertainties in the scale averaged ε_r^{comp}, and the horizontal error bars represent the uncertainties in the estimate of β. The average gauge spacing was 75 km for each network. Notice the tendency of ε_r^{comp} to decrease with increasing β.

affect inferences about the model performance as a function of scale (e.g., which resolution gives the smallest model error). This is because the total error is composed of the observational error, the representativeness error, and the model error, and thus mispecifying the representativeness error directly mispecifies the model error and any inferences about model performance as a function of resolution.

Via a numerical experiment specifically constructed to isolate the representativeness error from the total error, this study quantified typical representativeness error versus scale curves for hourly precipitation accumulations using two basic types of verification procedures: the area-to-point (A-P) and point-to-area (P-A). It was shown that the magnitude of the representativeness error in each method is significant (it can be up to 50% of the spatial average of the precipitation field), and it has considerable scale dependency within the typical mesoscale ranges of 5 - 50 km. Additionally, key factors that affect the representativeness error have been identified, such as the smoothness of the underlying field (quantified with the spectral slope, β) and the average gauge spacing, and their influence for a range of model resolutions has been quantified.

The scale dependence of the representativeness error must be accounted for appropriately, either by characterization, in a fashion similar to this study, and subsequent subtraction from the total verification error or by choosing a verification procedure that minimizes the scale dependence of the representativeness error. The latter alternative was the motivation for introducing the A-P/P-A composite method as this method typically gives representativeness errors which vary very little across scales of interest and thus can be considered scale independent. The scale-independent nature of the representativeness error in this composite verification method is a desirable characteristic if one does not wish to precisely quantify the representativeness error and only wants to compare several models at varying grid resolutions. This is because any difference in the RMSE verification error between the models would be due to model error since observational and representativeness errors would be the same across all model scales.

Although in this paper we presented a simple method aimed at reducing the scale dependency in model verification using a composite of deterministic distance-based estimators, we believe that there is need to develop a rigorous verification methodology which can properly and explicitly account for scale effects. Toward this objective we propose a new direction which involves the use of a stochastic multiscale filtering methodology [*Chou et al.*, 1994; *Kumar*, 1999] which can optimally merge observations at different scales while explicitly accounting for their uncertainties and the variability of the process at all scales. Such a technique would result in a minimal representativeness error and the best (unbiased and minimum variance) conditional precipitation estimates at any desired model scale. Since these conditional estimates can be obtained at the same scale as the model forecasts, traditional verification measures, such as RMSE, can then be used to judge the model's performance. Furthermore, by repeating the filtering procedure using model output from runs at different resolutions, inferences can be made about which model resolutions produce statistically superior predictions, conditional on the available observations and a priori knowledge of how the variability of the natural process changes with scale. Such a framework is under investigation, and the results will be reported in the near future.

As a final remark, it is worth noting that although this study has QPF verification in mind, another important problem, which can be cast in the exact same framework, is that of validating satellite estimates rather than model estimates of precipitation. Missions such as the Tropical Rainfall Measuring Mission (TRMM) [e.g., see *Simpson et al.*, 1996] involve numerous ground based field campaigns and the widespread use of rain gauge networks to validate rainfall estimates from the spaceborne TRMM Microwave Imager (TMI) and Precipitation Radar (PR). For validation of remote-sensing estimates of precipitation, the use of the stochastic multiscale filtering methodology mentioned above offers additional advantages, as this method provides estimates of the mean as well as the error variance as a function of scale. Estimates of the mean and error variance of rainfall products are

both of paramount importance for assimilating remotely sensed precipitation estimates into numerical climate models and for assessing systematic biases of rainfall retrieval algorithms.

Acknowledgments. This work has been supported by the U.S. Weather Research Program (under NSF grant ATM-9714387) and partially by the NASA-TRMM program (grant NAG5-7715) and NASA/NOAA-GCIP (under grant NAG8-1519). In addition, the authors wish to thank the University of Minnesota Supercomputing Institute, without whose resources and technical assistance this project would not be possible.

References

Barnes, S. L., A technique for maximizing details in numerical weather map analysis, *J. Appl. Meteor.*, *3*, 396-409, 1964.

Bras, R. L., and I. Rodriguez-Iturbe, *Random Functions and Hydrology*, Dover, Mineola, N.Y., 1993.

Brock, F. V., K. C. Crawford, R. L. Elliot, G. W. Cuperus, S. J. Stadler, H. L. Johnson, and M. D. Ellis, The Oklahoma mesonet: A technical overview, *J. Atmos. Oceanic, Technol.*, *12*, 5-19, 1995.

Chou, K. C., A. S. Willsky, and A. Benveniste, Multiscale recursive estimation, data fusion, and regularization, *IEEE Trans. Autom. Control.*, *39*, 464-478, 1994.

Colle, B. A., K. J. Westrick, and C. F. Mass, Evaluation of MM5 and Eta-10 precipitation forecasts over the Pacific Northwest during the cool season, *Weather Forecasting*, *14*, 137-154, 1999.

Davis, A., A. Marshak, W. Wiscombe, and R. Cahalan, Scale invariance of liquid water distributions in marine stratocumulus, Part I, Spectral properties and stationarity issues, *J. Atmos. Sci.*, *53*(11), 1538-1558, 1996.

Draper, N. R., and H. Smith, *Applied Regression Analysis*, John Wiley, New York, 1981.

Georgakakos, K. P., A. A. Carsteanu, P. L. Sturdevant, and J. A. Cramer, Observation and analysis of midwestern rain rates, *J. Appl. Meteorol.*, *33*, 1433-1444, 1994.

Harris, D., M. Menabde, A. Seed, and G. L. Austin, Multifractal characterization of rainfields with a strong orographic influence, *J. Geophys. Res.*, *101*(D21), 26,405-26,414, 1996.

Harris, D., A. W. Seed, M. Menabde, and G. L. Austin, Factors affecting multiscaling analysis of rainfall time series, *Nonlinear Processes Geophys.*, *4*(3), 137-156, 1997.

Harris, D. H., E. Foufoula-Georgiou, K. K. Droegemeier, and J. J. Levit, Multi-scale statistical properties of a high-resolution precipitation forecast, *J. Hydrometorol.*, in press, 2001.

Huffman, G. J., Estimates of root-mean-square random error for finite samples of estimated precipitation, *J. Appl. Meteorol.*, *36*, 1191-1201, 1997.

Krishnamurti, T. N., and L. Bounoua, *An Introduction to Numerical Weather Prediction Techniques*, CRC Press, Boca Raton, Fla., 1996.

Kumar, P., A multiple scale state-space model for characterizing subgrid scale variability of near-surface soil moisture, *IEEE Trans. Geosci. Remote Sens.*, *37*, 182-97, 1999.

Lovejoy, S., and D. Schertzer, Multifractals and rain, in *New Uncertainty Concepts in Hydrology and Water Resources*, edited by A. W. Kundzewicz, pp. 61-103, Cambridge Univ. Press, New York, 1995.

Murphy, A. H., and D. S. Wilks, A case study of the use of statistical models in forecast verification: Precipitation probability forecastsl, *Weather Forecasting*, *13*, 795-810, 1998.

Simpson, J., C. Kummerow, W.-K. Tao, and R. F. Adler, On the Tropical Rainfall Measuring Mission (TRMM), *Meteorol. Atmos. Phys.*, *60*, 19-36, 1996.

Smith, J. A., D. J. Seo, M. L. Baeck, and M.D. Hudlow, An intercomparison study of NEXRAD precipitation estimates, *Water Resour. Res.*, *32*(7), 2035-2045, 1996.

Xie, P., and P. A. Arkin, Analyses of global monthly precipitation using gauge observations, satellite estimates, and numerical model predictions, *J. Clim.*, *9*, 840-58, 1996.

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

E. Foufoula-Georgiou, D. Harris, and B. Tustison, St. Anthony Falls Hydraulic Laboratory, Department of Civil Engineering, University of Minnesota, Mississippi River at 3rd Avenue, SE, Minneapolis, MN 55414. (efi@tc.umn.edu; harri127@tc.umn.edu; tustison@msi.umn.edu)

(Received September 29, 2000; revised January 16, 2001; accepted January 23, 2001.)