Investigating Error Metrics for Satellite Rainfall Data at Hydrologically Relevant Scales

FAISAL HOSSAIN

Department of Civil and Environmental Engineering, Tennessee Technological University, Cookeville, Tennessee

GEORGE J. HUFFMAN

Science Systems and Applications, Inc., and Laboratory for Atmospheres, NASA Goddard Space Flight Center, Greenbelt, Maryland

ABSTRACT

This paper addresses the following open question: What set of error metrics for satellite rainfall data can advance the hydrologic application of new-generation, high-resolution rainfall products over land? The authors' primary aim is to initiate a framework for building metrics that are mutually interpretable by hydrologists (users) and algorithm developers (data producers) and to provide more insightful information on the quality of the satellite estimates. In addition, hydrologists can use the framework to develop a space–time error model for simulating stochastic realizations of satellite estimates for quantification of the implication on hydrologic simulation uncertainty. First, the authors conceptualize the error metrics in three general dimensions: 1) spatial (how does the error vary in space?); 2) retrieval (how "off" is each rainfall estimate from the true value over rainy areas?); and 3) temporal (how does the error vary in time?). They suggest formulations for error metrics specific to each dimension, in addition to ones that are already widely used by the community. They then investigate the behavior of these metrics as a function of spatial scale ranging from 0.04° to 1.0° for the Precipitation Estimation from Remotely Sensed Information Using Artificial Neural Networks (PERSIANN) geostationary infrared-based algorithm. It is observed that moving to finer space–time scales for satellite rainfall estimation requires explicitly probabilistic measures that are mathematically amenable to space–time stochastic simulation of satellite rainfall data. The probability of detection of rain as a function of ground validation rainfall magnitude is found to be most sensitive to scale followed by the correlation length for detection of rain. Conventional metrics such as the correlation coefficient, frequency bias, false alarm ratio, and equitable threat score are found to be modestly sensitive to scales smaller than 0.24° latitude/longitude. Error metrics that account for an algorithm's ability to capture rainfall intermittency as a function of space appear useful in identifying the useful spatial scales of application for the hydrologist. It is shown that metrics evolving from the proposed conceptual framework can identify seasonal and regional differences in reliability of four global satellite rainfall products over the United States more clearly than conventional metrics. The proposed framework for building such error metrics can lay a foundation for better interaction between the data-producing community and hydrologists in shaping the new generation of satellite-based, high-resolution rainfall products, including those being developed for the planned Global Precipitation Measurement (GPM) mission.

---

1. Introduction

Rainfall is a critical input for hydrologic models that predict the makeup of the hydrologic state over land. Because rainfall is intermittent, accurate modeling of the dynamic surface hydrologic state requires accurate rainfall data at the highest possible resolution. However, as in situ networks for rainfall measurements continue to decline worldwide (Stokstad 1999; Shiklomanov et al. 2002), spaceborne global observations are the only viable means to promote our understanding of terrestrial hydrology over the vast regions that are ungauged (Hossain and Lettenmaier 2006).

The global importance of satellite-derived rainfall has led to the development and accuracy assessment of an increasing number of satellite-based rainfall products to meet the needs of various users. Anagnostou

Corresponding author address: Faisal Hossain, Department of Civil and Environmental Engineering, Tennessee Technological University, 1020 Stadium Drive, Cookeville, TN 38505-0001.
E-mail: fhossain@tntech.edu

(2004) provides a detailed synopsis of the evolution of current satellite-estimation techniques over land, while Ebert et al. (2007) summarize several "high-resolution rainfall products" that are currently available via the Internet. Generally, the satellite data and hydrologic communities tend to characterize the accuracy of rainfall data using metrics such as bias, correlation coefficient, and standard deviation of "error." Additional measures, such as critical success index (CSI), Heidke skill score (HSS; Heidke 1926), equitable threat score (ETS), and false alarm ratio (FAR; Ebert et al. 2007) have seen use in the meteorological community engaged in forecasting (e.g., the National Weather Service or the European Centre for Medium-Range Weather Forecasts). These measures have proved useful in assessing satellite rainfall algorithms at scales pertinent for climate modeling, weather prediction, or even large-scale water management studies. However, with the planned Global Precipitation Measurement (GPM) mission (Smith et al. 2007) and the continued shift toward hydrologically more relevant scales (5–10 km and hourly), there is an urgent need to investigate metrics that can more effectively advance the use of satellite algorithms for hydrology over land, among other uses (Huffman et al. 2004; Lee and Anagnostou 2004). Hossain and Lettenmaier (2006) have argued that a shift in paradigm is needed to properly assess estimates of rainfall from satellite sensors for modeling of dynamic hydrologic phenomenon such as flood prediction. Among the many issues that require the exercise of caution, one that bears critical importance is the uncertainties in satellite-estimated rainfall that cascade nonlinearly through the simulation of the terrestrial hydrologic processes (Nijssen and Lettenmaier 2004). This nonlinear effect is difficult to model because of the prominent discontinuities of the rainfall process in space and time that are observed as scales become smaller.

Recognizing the need for assessment of uncertainty for the new generation of high-resolution precipitation products (HRPP), several recent studies have compared the accuracy of various satellite rainfall products over land. For example, Ebert et al. (2007), as a contribution to the International Precipitation Working Group (IPWG), assessed six widely available HRPP using an array of error metrics currently used by the community. Hong et al. (2006) have evaluated an infrared satellite-estimation technique for hydrologic applications using error conceptualizations initiated by North and Nakamoto (1989) and subsequently formalized by Steiner et al. (2003). Other examples of evaluating satellite rainfall uncertainty include McCollum et al. (2002) on the assessment of bias, Gebremichael and Krajewski (2005, 2004) on sampling errors, and Ali et al. (2005) on satellite error functions for the Sahel region.

While these and other studies of satellite rainfall uncertainty have advanced the application of HRPP in terrestrial hydrology to some extent, some issues continue to remain open. For example, many studies treat error as a unidimensional measure and use power-law-type relationships or models for estimating this aggregate error as a function of spatial and temporal sampling parameters (Moradkhani et al. 2006; Hong et al. 2006; Steiner et al. 2003). Such frameworks are acceptable for estimating the average error over an areal domain, but they do not have explicit representation of the space–time covariance structure of the estimation error, which can have significant implications in the simulation of the terrestrial hydrologic processes (Hossain and Anagnostou 2005). Also, most studies, such as that of Ebert et al. (2007), have typically addressed uncertainty at daily or larger time scales, which are somewhat coarse for resolving the evolution of the dynamic hydrologic state over land (e.g., for floods and soil moisture).

In this paper, we address the following open question: What set of error metrics for satellite rainfall data can advance the hydrologic application of new-generation HRPP over land? The satellite rainfall data producing community have long recognized that information on the reliability of satellite rainfall estimates is valuable to a wide range of users. Yet the definition of acceptable skill in the satellite data is relative to the nature of the application. Ebert et al. (2007, p. 49) provide a lucid perspective on the diverse accuracy requirements. Our initial question leads us to pose a set of additional questions: What should be the characteristics of error metrics at hydrologically relevant scales? How should they be designed so that they are conveniently interpretable by both data producing and hydrologic communities? How should these metrics be packaged into standard satellite data products for best use in hydrologic modeling and decision making?

Clearly, these questions require error expressions that capture mean behavior accounting for space–time correlations and intermittency in the estimated rainfall fields. Hence, for the hydrologist, error should be defined in terms of the rainfall and tagged to a given space and time scale. We therefore conceptualize that the error metrics should be associated, at a minimum, with three general dimensions: 1) spatial (how does the error vary in space?); 2) retrieval (how "off" is the rainfall estimate from the true value over rainy areas?); and 3) temporal (how does the error vary in time?).

As with any modeling exercise, there is probably no unique way of representing error completely. But, we

note that studies of uncertainty in hydrologic prediction have usually evolved independently of efforts to characterize uncertainty in remote sensing estimates of rainfall. In this study, our aim is therefore to initiate a common framework for building error metrics. In particular, we are motivated to build such a framework comprising multidimensional metrics that can be mathematically transformed into a model for simulating stochastic realizations of satellite rainfall data for a given satellite rainfall algorithm. There are several mathematical error models today (e.g., Steiner 1996; Steiner et al. 2003; Hong et al. 2006) that can yield such stochastic realizations of simulated rainfall. However, there seems to be none, to the best of our knowledge, that remain conceptually flexible enough to investigate inclusion of additional, or modification of, error metrics for a given application.

We also emphasize that our framework is intended to augment the commonly used metrics (e.g., correlation, bias, RMSE, etc.) in order to provide a better assessment of algorithms at hydrologically relevant scales. In section 2, we introduce a set of error metrics that was first formalized by Hossain and Anagnostou (2006). Overviews of the study region and rainfall datasets (reference and satellite) are provided in section 3. In section 4, we present our error assessment across hydrologically relevant spatial scales ranging from 0.04° to 1.0° of latitude/longitude for a particular satellite-based set of rainfall estimates. The implications of these results on data use are discussed, along with the challenges ahead in developing more robust metrics for operational data products. Finally, in section 5, we summarize the major findings and recommend future work. While rainfall is our primary focus, the techniques that are described here are general enough to be applied to the broader spectrum comprising precipitation.

2. Error metrics for satellite rainfall

In section 1 we hypothesized that error metrics should quantify, at a minimum, three specific dimensions related to rainfall intermittency. We address this concept using the error-modeling approach first outlined by Hossain and Anagnostou (2006, hereafter HA06). First, we note that the error structure necessary to capture the rainfall intermittency at hydrologically relevant scales arises from the physical issues associated with satellite rainfall estimation. Satellite-derived estimates are typically instantaneous, area-averaged rainfall. Since rainfall is an intermittent process, each satellite gridbox value will be classified by a rainfall algorithm as rainy or nonrainy (as discussed above, "rain" is used here as a shorthand for "precipitation"). When

compared to the corresponding ground validation rainfall data (hereafter referred to as "reference"), a satellite estimate may fall into one of four possible outcomes:

1) Satellite successfully detects rain (successful rain detection, or "hit").
2) Satellite fails to detect rain (unsuccessful rain detection, or "miss").
3) Satellite successfully detects the no-rain case (successful no-rain detection).
4) Satellite fails to detect the no-rain case (unsuccessful no-rain detection, or "false alarm").

For the data-producing community, there are already accepted metrics in use that can quantify these notions of hits, misses, and false alarms. Some examples are, frequency bias (FB), FAR, and ETS. Ebert et al. (2007, p. 52) provides an introductory background on the formulation of these metrics that are often tagged with the satellite estimates by the data producers during algorithm comparisons (see also appendix B). Each of these conventional metrics typically refers to a particular aspect of rain estimation considered important for evaluating algorithms at large space–time scales. For example, the metric FB indicates the tendency of an algorithm to overestimate or underestimate the aerial extent of rainy areas (>1 for overestimation; <1 for underestimation). However, an issue that has remained unclear is the use of these metrics to generate stochastic realizations of satellite rainfall for assessing their hydrologic implications. It is not clear how one would use the FB measure in an error model to simulate satellite-like rainy areas with coherent space–time structures. Hence, an inherent limitation associated with some of these conventional metrics is the difficulty in mathematically modeling the property they represent for simulation of stochastic realizations of satellite rainfall data. Many of the currently accepted metrics therefore have diagnostic power (i.e., they tell us the level of error for an actual algorithm). But most lack prognostic qualities for hydrologic error propagation experiments (i.e., they do not tell us how to use it a step further to generate stochastic realizations of satellite rainfall data). This fact motivates our proposed framework on additional and hydrologically more relevant metrics.

In Fig. 1, we outline the layout of the HA06 error metrics and describe the logic behind formulation of the metrics hereafter. For satellite grid boxes that are correctly detected as rainy (e.g., a rainy area HIT), the probability of successful detection likely depends on the magnitude of the rainfall rate. This comes from our experience that satellites are less likely to miss areas that are raining more heavily than others. Our first metric numbered 1 in Fig. 1, probability of detection of rain
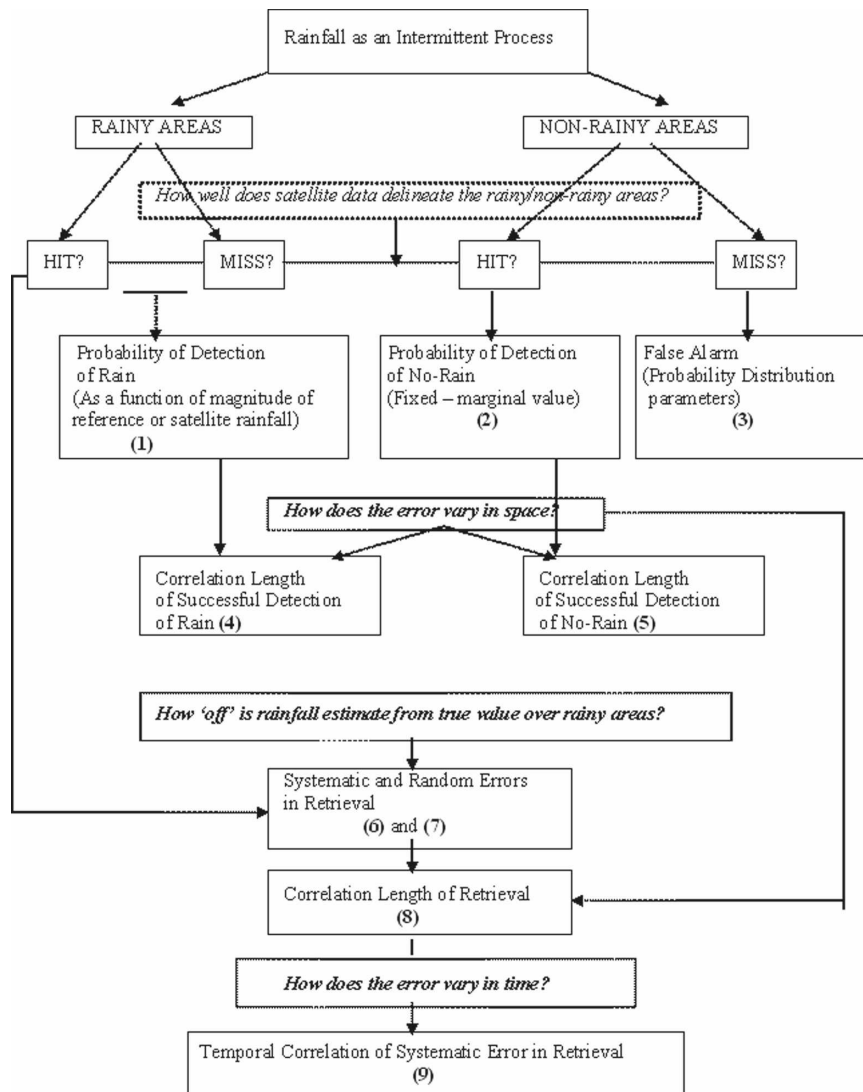
FIG. 1. The logic of the conceptual framework for hydrologically relevant error metrics.
Numbers in parentheses denote the metrics as discussed in section 2.

(POD_{RAIN}), is used to quantify this property. Here the POD_{RAIN} for a given threshold rain rate τ is defined statistically as follows:

$$\text{POD}_{\text{RAIN}} = \text{Prob}\{R_{\text{SAT}} > 0 | R_{\text{REF}} > \tau\}, \qquad (1)$$

where subscripts SAT and REF refer to satellite and reference values for rain rate, respectively.

The functional form of POD_{RAIN} may be based on either the reference [ground validation (GV)] or the estimated rain rate. For example, the traditional hydrologist users would likely be interested in the probability of rain detection benchmarked with respect to ground data. On the other, the data producers may find it almost impossible to tag the probability of detection of the satellite estimates to the GV data on an opera-

tional basis due to lack of global-scale ground validation data, so they would choose to index errors on satellite estimates instead.

To maintain consistency with the HA06 metric formulation, the τ value is varied from the lowest threshold (>0 mm h^{-1}) to an arbitrarily high τ value (>15 mm h^{-1}). The POD_{RAIN} is then regressed against the τ value to identify the functional form for $\text{POD}_{\text{RAIN}} = f(\tau)$. From our experience, the functional form has most often been a logistic equation. As an example, if the POD_{RAIN} value is 0.72 for $\tau = 5$ mm h^{-1}, then this means that the satellite algorithm has a 72% chance of successfully detecting rainy grid boxes with reference value exceeding 5 mm h^{-1}. Note that "detection" is defined as any nonzero rain value. Such an error con-

ceptualization for rain detection makes the POD_{RAIN} relatively easy to simulate using the concept of Bernoulli trials in a stochastic error model (HA06).

Next, for grid boxes that are detected correctly as nonrainy (a nonrainy HIT), the algorithm can be characterized by a marginal probability of no rain as POD_{NORAIN}. This measure is the ratio of the number of satellite grid boxes correctly classified as nonrainy to the total number of grid boxes that are actually non-rainy according to reference data (metric number 2 in Fig. 1). Grid boxes that are unsuccessfully detected as nonrainy—that is, the satellite detecting rain when there is none (rainy MISS)—can be classified as "false alarm" grid boxes. For these grid boxes, it is more convenient for us to quantify and simulate the observed probability distribution of the false alarm rain rates. Hence, the first and second moments of the distribution can account for an algorithm's tendency to produce false alarms (in Fig. 1, these two metrics are collectively numbered 3), which can then be easily simulated in a stochastic error model.

For satellite grid boxes that are detected correctly as rainy or nonrainy (all HITs in Fig. 1), the detection's spatial pattern may exhibit a clear covariance structure—the probability of successful detection of a satellite grid box as rainy or nonrainy may be a function of its proximity to another successfully detected grid box in the neighborhood. One measure to quantify this spatial structure is the correlation length, which can be used in a stochastic error model for the generation of correlated random fields (Deutsch and Journel 1998). Accordingly, we quantify the spatial structure of successful detection of rainy and nonrainy area with two metrics—CL_{RAIN} and CL_{NORAIN}, respectively (the metrics numbered 4 and 5 in Fig. 1; CL stands for correlation length). The metrics CL_{NORAIN} and CL_{RAIN} specifically refer to the spatial dimension of error when considered in combination with the POD metrics.

Finally, when compared to the reference value, the grid boxes that are successfully detected as rainy may exhibit three additional properties: (i) they will have a spatial structure; (ii) they may be "off" from the true value; and (iii) they will have temporal persistence. These three properties pertain to the three dimensions of error metrics highlighted in section 1. The first and second moments of retrieval error (bias and standard deviation) can collectively quantify the error property for the retrieval dimension; in Fig. 1, these two metrics are numbered 6 and 7. To account for the spatial structure of the retrieval error, we introduce a metric CL_{RET}—correlation length of retrieval—that can be used to simulate correlated random fields. This metric is numbered 8 in Fig. 1.

We address the temporal dimension of the satellite-estimation error with a relatively simple representation: We assume that only the mean-field bias (systematic error) of retrieval error is correlated in time in an Eulerian (surface based) frame of reference. Hence, the lag 1 correlation of the mean-field bias is used as the metric to quantify the temporal dimension of error (metric number 8 in Fig. 1). The temporal persistence of satellite-estimated rainfall probably arises from a mixture of the true spatial and temporal correlations of the rain system in its Lagrangian (system following) frame of reference, and the advection speed of that frame of reference. The temporal persistence of satellite-estimation error will therefore arise from the above combination of correlations of the rain system. At this stage, an Eulerian simplification makes our framework tractable for generating stochastic ensembles of satellite rainfall data, although a more sophisticated approach may be needed in the future.

Collecting all these components, it appears that one possible set of error metrics is as follows: 1) Probability of rain detection (and as a function of rainfall magnitude), POD_{RAIN}; 2) Probability of no-rain detection, POD_{NORAIN}; 3) First- and second-order moments of the probability distribution during false alarms; 4) Correlation lengths for the detection of rain, CL_{RAIN}, and 5) no rain, CL_{NORAIN}; 6) Conditional systematic retrieval error or mean-field bias (when reference rain > 0); 7) Conditional random retrieval error or error variance; 8) Correlation length for the retrieval error (conditional, when rain > 0.0), CL_{RET}; and 9) Lag 1 autocorrelation of the mean-field bias. The mathematical formulations of these nine error metrics are reasonably straightforward and are provided in appendix A. Readers should refer to HA06 for a more complete description on how these nine error metrics can be used in a stochastic error model to simulate realizations of satellite rainfall data.

It is not clear whether these nine metrics completely describe the error structure of satellite rainfall estimation at hydrologically relevant scales. The needs of particular users and applications will necessarily drive the evolution to the best representation of these error structure parameters.

3. Data and study region

We choose the U.S. National Weather Service (NWS) stage II rainfall data as the ground validation rainfall dataset for illustrating the nine error metrics. This dataset uses the NWS Weather Surveillance Radar-1988 Doppler (WSR-88D) estimates with real-time adjustments based on mean-field radar–rain gauge hourly accumulation comparisons (Fulton et al. 1998;
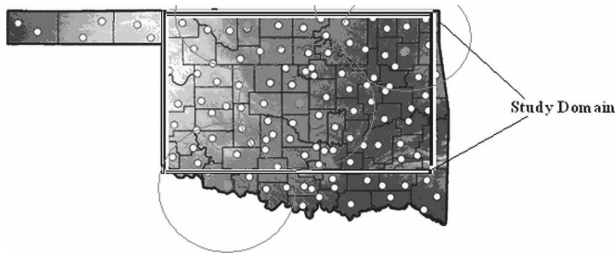
FIG. 2. Study region over Oklahoma bounded by 37°–34°N, 100°–95°W. The light dots show the location of the Oklahoma Mesonet meteorological stations with gauge rainfall data for WSR-88D radar calibration. Dark dots indicate the location of WSR-88Ds inside Oklahoma; circles approximately indicate coverage with 100-km radius.

Seo et al. 1999, 2000). The original resolution is 4 km, hourly.

For a representative satellite rainfall algorithm, we use data from a recent version of the Precipitation Estimation from Remotely Sensed Information Using Artificial Neural Networks (PERSIANN; Sorooshian et al. 2000). The original PERSIANN is a satellite IR-based algorithm with calibration by passive microwave rainfall estimates in a neural net framework that produces global estimates of rainfall at $0.25° \times 0.25°$, half-hourly resolution (Hong et al. 2005a). The revised version used in this study includes a self-organizing nonlinear output (SONO) neural network for cloud-patch-based rainfall estimation. The revised PERSIANN algorithm estimates 0.04°, half-hourly rainfall and is available over the Internet (http://hydis8.eng.uci.edu/GCCS/) (Hong et al. 2005b). The fine submicrowave footprint scale is achieved by using the Climate Prediction Center Merged IR Dataset (Janowiak et al. 2001) at full resolution to guide disaggregating the microwave estimates from the original PERSIANN grid of 0.12° before use in training the neural network scheme.

To minimize effects due to complex terrain and radar range, the error computation exercise is performed over the region of Oklahoma bounded by 37°–34°N, 100°–95°W (Fig. 2), which is relatively flat and well covered by radars and the PERSIANN data. We selected a one-month period for this study, 1 May 2002 to 30 May 2002, which contains 720 hourly time steps representing the set of grids closest to each nominal hour.

4. Methodology and results

We assessed the nine error metrics at seven spatial scales: (i) 0.04° (original); (ii) 0.08°; (iii) 0.12°; (iv) 0.16°; (v) 0.24°; (vi) 0.48°; and (vii) 1.0°. The lower end of this range is considered more relevant to hydrologic modeling, while the higher end is typical of many long-term satellite rainfall products (i.e., the GPCP products;

Huffman et al. 2001) and evaluations (Ebert et al. 2007). Note, however, that a statistically significant sample for spatial correlation lengths is not possible at the two largest scales due to the size of the study region ($5.0° \times 3.0°$), and hence, these values have not been reported. In addition, three other commonly used diagnostic metrics were also evaluated: FB, FAR, and ETS. See appendix B for their mathematical formulation.

Using a very simple cropping technique, the stage II data are remapped to the 0.04° PERSIANN grid to allow consistent comparisons. We verified that the cropping-based interpolation had no effect on the statistics of the stage II data in this study. The temporal resolution was kept fixed at hourly. The spatial scales of aggregation are chosen to be integer multiples of the original 0.04° grid to avoid spatial interpolation errors, which were found to be problematic in our preliminary investigation. This choice allows us to focus on the scaling behavior of error parameters purely as a function of aggregation. We seek to identify how each of the nine error metrics responds to spatial scaling and whether there exists some minimum scale at which some or most of the error parameters remain "acceptable" for the hydrologist user.

Ordinarily, one would expect the data producer to use error metrics that are robust to changes in scale for the sake of consistency. However, such an approach may not provide the best insight into applying satellite rainfall data at hydrologically relevant scales. As an example, consider the case when the spatial scale for a data product decreases from 0.24° to 0.16° or 0.12° as a result of, say, spatial downscaling. The correlation coefficient or systematic bias metrics may register a change with scale that is considerably more modest than changes in the algorithm's ability to correctly delineate the rainy or nonrainy areas, given the intermittency of the rainfall process. This is because marginal measures such as correlation coefficient are parameters that reflect essentially the aggregate effect of the algorithm's ability to retrieve rainfall over the study area (discussed later). But the intermittency has important implications for hydrologic simulation of the terrestrial water cycle and must be considered in evaluating the use of satellite data.

In Table 1, we summarize results for the correlation coefficient, RMSE, FB, FAR, and ETS metrics. The conditional correlation refers to the cases when both reference and satellite rain is nonzero. In general, we observe that the response to spatial scale is similar for all these conventional metrics, often appearing insensitive to scales smaller than 0.24°. For example, Fig. 3 shows the correlation coefficient as a function of scale. The sensitivity of correlation coefficient appears mod-

TABLE 1. Some commonly used error metrics as a function of spatial scales (degrees of lat/lon).

	0.04°	0.08°	0.12°	0.16°	0.24°	0.48°	1.0°
Correlation (unconditional)	0.386	0.383	0.393	0.401	0.418	0.469	0.569
Correlation (conditional)	0.272	0.298	0.319	0.334	0.361	0.437	0.547
Root-mean-square error (mm h^{-1})	4.708	3.933	3.530	3.230	2.776	1.98	1.25
FB	1.524	1.405	1.423	1.419	1.460	1.548	1.677
FAR	0.686	0.634	0.619	0.601	0.5804	0.5370	0.5066
ETS	0.205	0.235	0.245	0.255	0.271	0.302	0.306

est and it remains difficult to use this metric in a stochastic error model.

An interesting pattern emerges in Fig. 4 for our proposed metric on probability of rain detection (POD$_{RAIN}$) as a function of reference threshold rain rate. With spatial aggregation, both the maximum POD$_{RAIN}$ (at reference rain rates greater than 15 mm h^{-1}) and the gradient of probability detection as a function of reference threshold rainfall rate increase noticeably, as expected (the highest being for 1.0°). As explained in section 2, this means that the algorithm is highly unlikely to have zero rain in areas raining at a rate greater than 15 mm h^{-1}, although the satellite-estimated rain rate may not necessarily match closely with the ground validation data. The POD$_{NORAIN}$ exhibits even stronger scale dependence (Fig. 5). The probabilities for both rain and no-rain detection respond definitively as spatial scales decrease below 0.48°. Hence, these two metrics that account for rainfall intermittency add information to the traditional list of metrics for exploring satellite rainfall data application in hydrologic models.

In Fig. 6, we show the spatiotemporal structure for metrics on conditional retrieval error (random; upper left panel), mean-field bias (temporal autocorrelation function; upper right panel), and the spatial correlation functions for rain detection (lower left panel) and no-rain detection (lower right panel). Even though the distinction in spatial scaling for these spatiotemporal error parameters is weak at the scales considered, HA06 show that these metrics can be used coherently in an error model to simulate realistic space–time covariance structure of satellite rainfall data. If correlation lengths in space and time are computed (assuming that an exponential model is appropriate; see appendix A), a more informative picture emerges (Fig. 7). We observe a clear sensitivity of the correlation length to spatial scales, with the correlation length for rain detection being the most sensitive. The lag 1 (hourly) temporal correlation of mean-field bias however remains insensitive to scale as would be expected since the domain's area is the same for all grid sizes. These results demonstrate that the suggested error metrics on rain detec-

tion/delineation correlation lengths can offer insight into the useful spatial scales for applying satellite-based rainfall in hydrologic models. While the assessment of the direct implications of these metrics on hydrologic modeling is beyond the scope of this study, Hossain and Anagnostou (2004) have shown that an improvement in the probability of rain detection can yield substantial improvements in flood prediction at the 0.1° scale for saturation-excess watersheds in the Alps. Intuitively, the same can be expected of Hortonian watersheds where spatial pattern of the rainy areas along with the rain rate and soil's infiltration capacity dictate the propensity of a region to produce direct runoff.

Finally, in order to demonstrate the value of our proposed error framework in distinguishing the strengths and weaknesses of existing HRPP algorithms, we look at four global satellite rainfall products. These are 1) Tropical Rainfall Measuring Mission (TRMM) Multisatellite Precipitation Analysis (TMPA) microwave-calibrated infrared (IR) rainfall product 3B41RT (Huffman et al. 2007); 2) TMPA merged microwave-IR rainfall product 3B42RT (Huffman et al. 2007); 3) NOAA CPC morphed passive microwave rainfall product (CMORPH; Joyce et al. 2004), and 4) PERSIANN. The error analysis is performed at the native scale of algorithms for the year 2004 over two regions in the
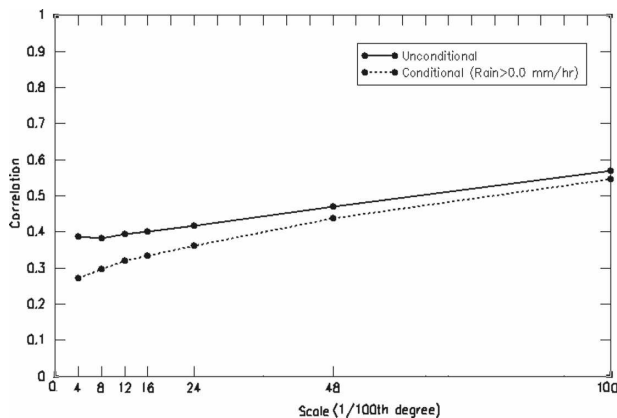

FIG. 3. Correlation coefficient of satellite rainfall data with stage II rainfall as a function of spatial scale.
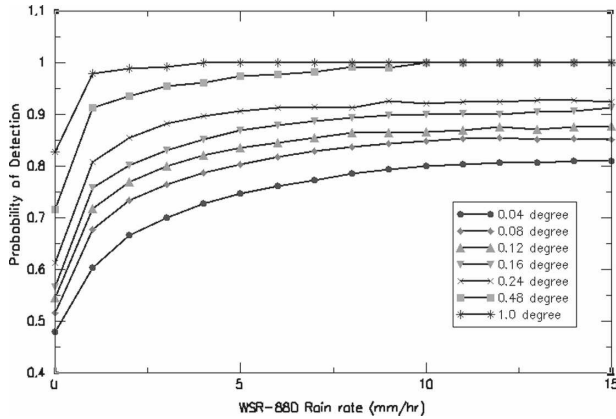
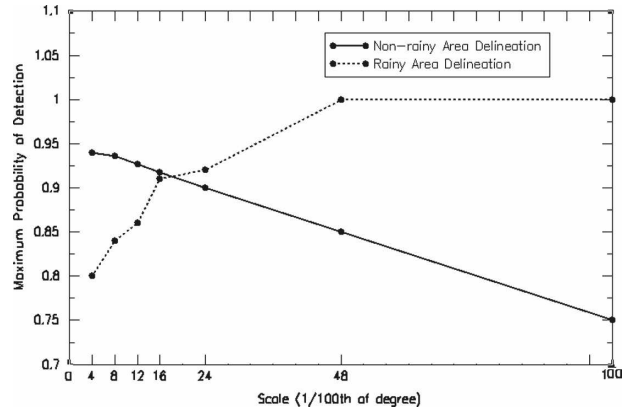FIG. 4. Probability of rain detection as a function of reference (stage II) rainfall rate and spatial scales.

FIG. 5. Comparison of scaling behavior of maximum probability for rain (at rain rates > 15 mm h^{-1}) and marginal probability of no-rain detection.

United States known to have a distinct hydroclimatology: (i) the Midwest (a semiarid zone) and (ii) Florida (subtropical zone modulated by coastal effects). WSR-88D stage-II rainfall data are used as a reference for ground validation data. The main purpose of this exercise is to assess whether our proposed framework of

metrics can clarify regional and seasonal differences for a given algorithm when used in combination with conventional metrics, without undertaking a comparative assessment of the various algorithms.

Table 2 shows a summary of conventional metrics (correlation and standard deviation of error) and the
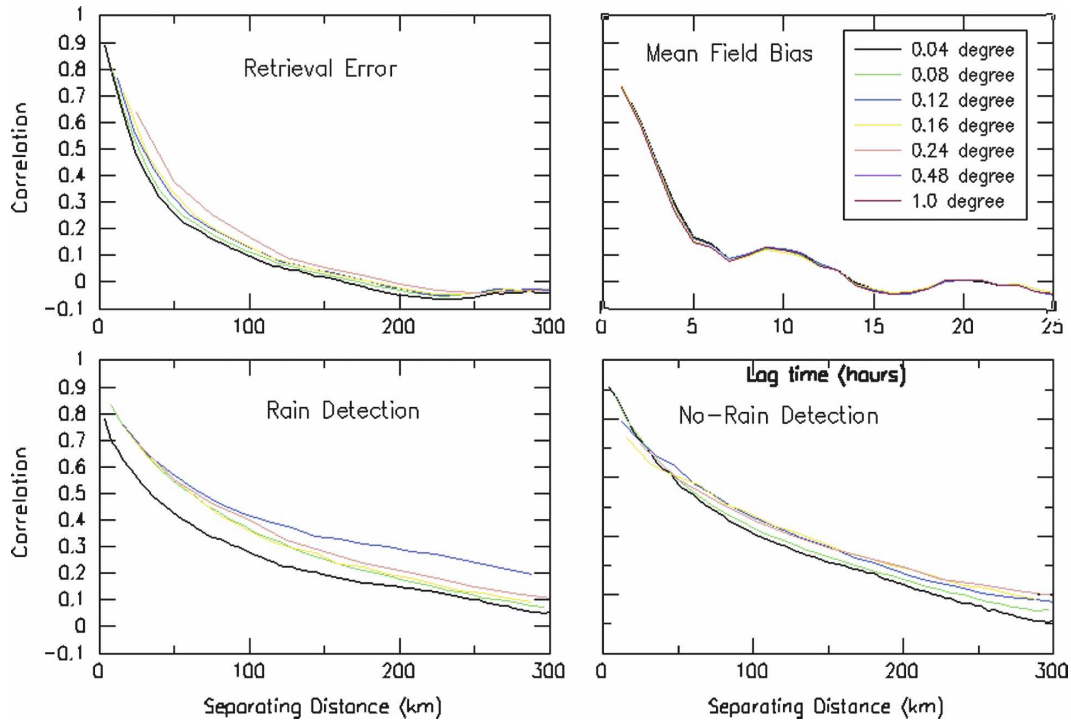

FIG. 6. Spatiotemporal metrics as a function of spatial scales. (top left) Spatial correlation function for conditional retrieval error (rain > 0); (top right) temporal correlation function for mean-field bias; (bottom left) spatial correlation function for rain detection; and (bottom right) spatial correlation function for no-rain detection. Note that spatial correlation functions are not reported at scales 0.48° and 1.0° due to the small spatial sample available over the study region. The separating distance (km) is reported by approximating 1° of latitude or longitude as 100 km.
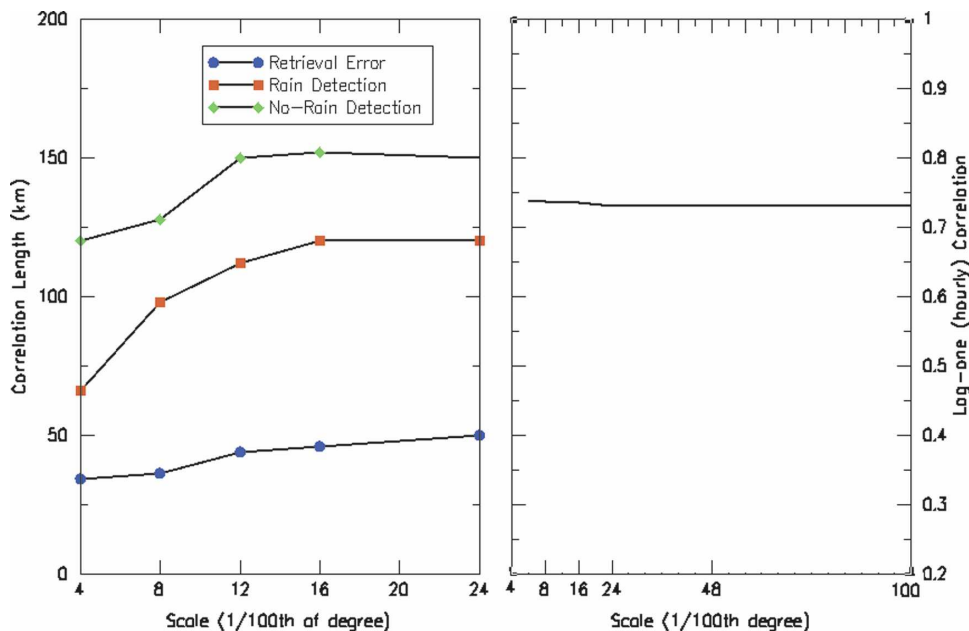
FIG. 7. Correlation lengths and lag 1 autocorrelation as a function of scale (assuming an exponential model is appropriate to describe the correlation function in space). (Note that spatial correlation lengths are not reported at scales 0.48° and 1.0° due to the small spatial sample available for the study region.)

metric of the lag 1 autocorrelation of mean-field bias. In Figs. 8 and 9, the variation of two other proposed metrics—POD_{RAIN} and POD_{NORAIN}—during the winter (January and February) and summer (June, July, and August) are also shown. It is clear from Table 2 that the correlation metric fails to highlight any major differences in each algorithm as a function of region and season. For example, the seasonal variation of conditional correlation (i.e., when reference rainfall > 0 mm h^{-1}) for CMORPH ranges between 0.226 and 0.251 and 0.195 and 0.153 for Oklahoma and Florida, respectively. Similarly, the standard deviation of conditional retrieval error varies between the 2.1 and 2.5 mm h^{-1} range. Normally, such differences may be considered statistically insignificant to elucidate any seasonal or regional sensitivity of an algorithm at hydrologically relevant scales (i.e., subdaily and < 25 km^2). On the other hand, the lag 1 autocorrelation and the POD_{RAIN} and POD_{NORAIN} show higher sensitivity within an algorithm across regions and seasons. CMORPH registers highest seasonal sensitivity for POD_{NORAIN}. When used with conventional metrics, these proposed metrics can enhance our search for the physical implications behind the use of an algorithm for a particular hydrologic application. For example, the consistently high values of POD_{NORAIN} for TMPA algorithms (3B41RT and 3B42RT; Fig. 9) would indicate, as a first cut to a hydrologist, that TMPA data could be more ideally suited for drought and agricultural applications, while

CMORPH could be more appropriate for forecasting of floods caused by short-duration storm events.

5. Conclusions and future needs

Representing the error structure of satellite rainfall as a function of scale against quality-controlled ground

TABLE 2. Summary of some error metrics over Oklahoma and Florida during winter and summer of 2004.

| HRPP	Correlation (conditional)	Error std dev (conditional) (mm h−1)	Lag 1 correlation
Oklahoma winter			
CMORPH	0.226	2.19	0.540
3B41RT	0.106	1.89	0.855
3B42RT	0.180	1.95	0.467
PERSIANN	0.268	3.16	0.710
Oklahoma summer			
CMORPH	0.251	2.47	0.642
3B41RT	0.274	2.31	0.911
3B42RT	0.121	2.18	0.564
PERSIANN	0.303	3.44	0.689
Florida winter			
CMORPH	0.195	2.41	0.628
3B41RT	0.102	2.18	0.960
3B42RT	0.121	2.18	0.564
PERSIANN	0.159	3.59	0.721
Florida summer			
CMORPH	0.153	2.52	0.428
3B41RT	0.219	2.38	0.896
3B42RT	0.220	2.29	0.371
PERSIANN	0.248	3.68	0.711
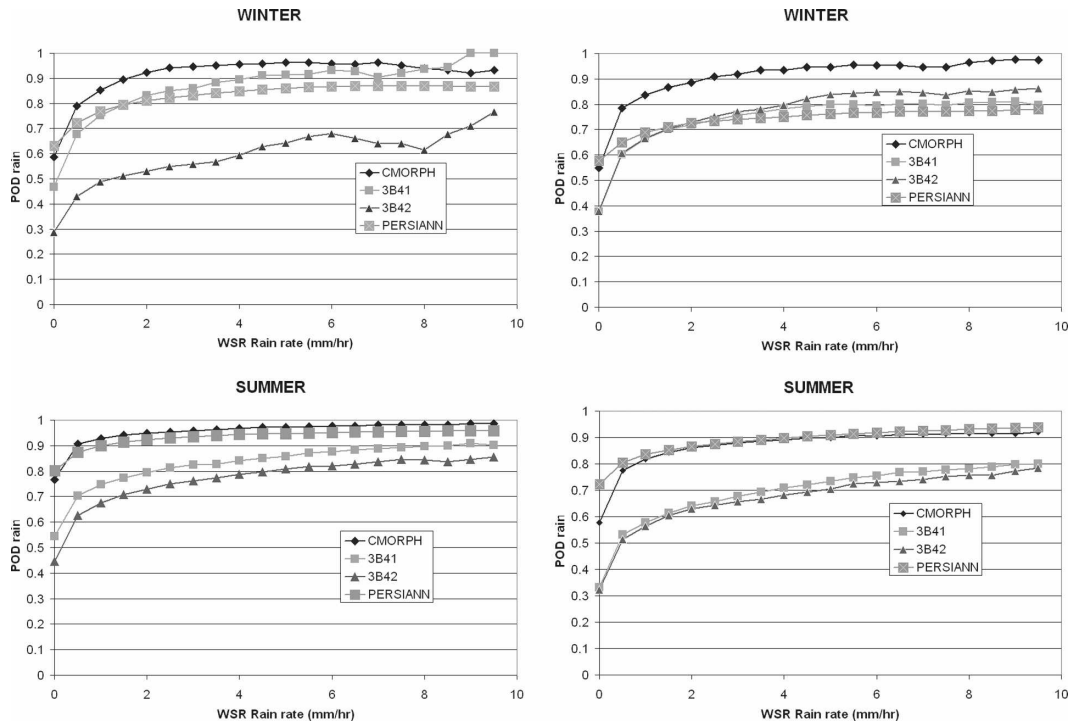
FIG. 8. POD$_{RAIN}$ as a function of season, region, algorithm, and reference (GV) rain rate at the native scales. (left) Oklahoma (semiarid hydrology) and (right) Florida (coastal hydrology).

validation datasets remains a critical research problem. Hydrologists and other users, to varying degrees, need to know the errors of the satellite rainfall datasets across the range of time–space scales over the whole domain of the dataset. In this study, we investigated the behavior of a suggested set of error metrics that were linked primarily to rainfall intermittency for a microwave-calibrated geostationary infrared-based algorithm. In general, the conventional error metrics such as correlation coefficient, frequency bias, false alarm ratio, and equitable threat score appeared to have similar levels of sensitivity to scale. However, the use of these common metrics for simulating stochastic realizations of satellite rainfall with realistic space–time covariance structures does not seem feasible. In our opinion, this limits the value of the metrics to the hydrologist who may choose to probabilistically quantify the implications of each metric for overland hydrologic
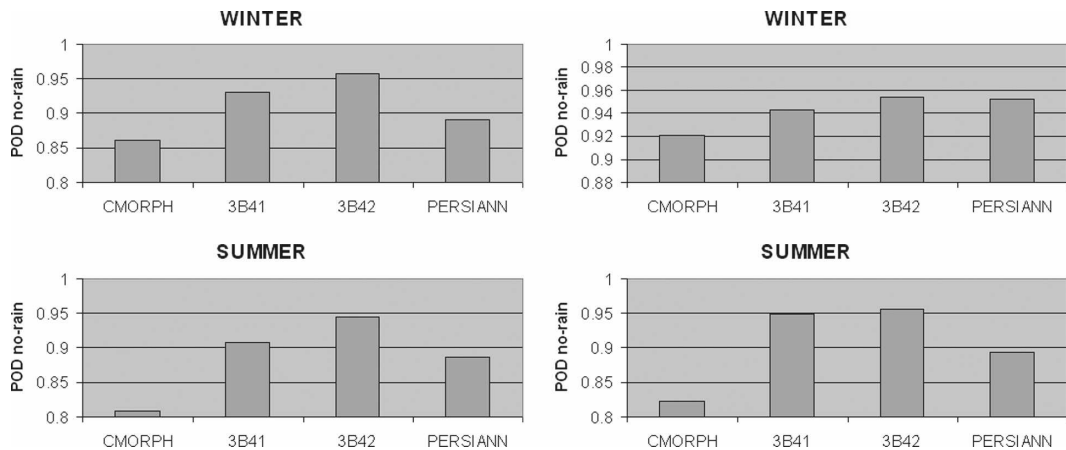

FIG. 9. POD$_{NORAIN}$ for algorithms across regions and seasons. (left) Oklahoma and (right) Florida.

simulations. The probability of detection of rain and its functional relationship to ground validation rainfall were found to be the most sensitive to scale followed by the correlation length for detection of rain. These specific error metrics appear informative for identifying the useful scales for data integration over land and can also be used in a stochastic error model.

The new-generation HRPP datasets pose significant opportunities for hydrologists, but with associated challenges. Specifically, effective assessment frameworks and metrics for satellite rainfall data must be developed that enhance the HRPPs' utility for land surface hydrology and are jointly defined by the hydrologic and data-producing communities. While we have shown some tangible examples of error metrics and their potential value in gauging utility, more work is needed to define hydrologically relevant metrics that can connect more directly to the physics and geometries of satellite rainfall estimation.

The practicality of the approach presented in this paper can be questioned because the details likely vary by region and season, and because many regions lack the necessary ground validation data to develop region-specific error representations. However, it appears conceptually feasible to build on work already accomplished on global classification of rainfall systems (Petersen and Rutledge 2002). In addition, it is possible to use the TRMM Precipitation Radar (PR) as the reference for the spatial domain (e.g., Hossain and Anagnostou 2004) and apply a recent approach suggested by

Bellerby and Sun (2005) based on transfer of probability distribution functions for the temporal domain. Another approach could be the use of geostatistical simulation techniques such as ordinary or indicator kriging (Deutsch and Journel 1998) to transfer error metrics from a ground validation site to ungauged regions under the assumption of stationarity of the metric values.

A challenge that remains is to choose a small set of error parameters that enable practical use of the uncertainty information and capture the time–space structure of the uncertainties. At this level of complexity it might be best to establish functional forms for the error metrics and supply coefficients for large regions and seasons. In some cases, average or "climatological" coefficients might suffice, while in other cases routine updates as a function of time might be required. Given this information a user could easily estimate the errors that correspond to the time–space scale of their application. In particular, a hydrologist could identify the necessary scale of aggregation of satellite rainfall data to achieve a specified level of accuracy that would minimize error propagation in a hydrologic model for his/her intended application.

APPENDIX A

Formulation of Error Metrics

Consider first, the following 2×2 validation (contingency) matrix shown in Table A1 for hits and misses associated with satellite rainfall estimates:

$$\text{Probability of detection for rain (POD}_{\text{RAIN}}): \frac{N_A}{N_A + N_B}, \quad \text{(A1)}$$

$$\text{Probability of detection for no rain (POD}_{\text{NORAIN}}): \frac{N_D}{N_D + N_C}. \quad \text{(A2)}$$

We also define the (successful) rain detection probability, POD_{RAIN}, as a function of rainfall magnitude of either the reference rainfall or satellite estimate. The functional form is usually identified through calibration with actual data (HA06) The $\text{POD}_{\text{NORAIN}}$, is the unitary probability that satellite retrieval is zero when reference rainfall is zero, which is also determined on the basis of actual data.

A probability density function (D_{false}) is defined to characterize the probability distribution of the satellite estimates when there are misses over nonrainy areas. This function is also identified through calibration on the basis of actual sensor data. HA06 have reported that this D_{false} probability density function typically

tends to appear exponential. Hence, both the moments (first and second) can be defined using only one parameter of the distribution, λ. This can be computed using the chi-squared or maximum likelihood method.

To identify the correlation lengths of error (i.e., how does the error vary in space) a simple exponential-type

TABLE A1. Validation (contingency) matrix for hits and misses of satellite rainfall data.

Satellite estimates	Truth/reference	
	Rainy grid boxes	Nonrainy grid boxes
Rainy grid boxes	N_A (HIT)	N_B (MISS)
Nonrainy grid boxes	N_C (MISS)	N_D (HIT)

autocovariance function is assumed. The correlation length (the separation distance at which correlation = $1/e = 0.3678$) is then determined on the basis of calibration with actual data over a large domain (the size of Oklahoma in this study). For identifying the spatial correlation length of rain detection, CL_{RAIN} (or, no-rain detection–CL_{NORAIN}) from data, all successfully detected rainy (nonrainy) pixels are assigned a value of 1.0 while the rest has a value of 0.0. The empirical semivariogram is then computed as follows:

$$\gamma(h) = \frac{1}{2n(h)} \sum_{i=1}^{n(h)} [z(x_i) - z(x_i + h)]^2, \quad (A3)$$

where $z(x_i)$ and $z(x_i + h)$ are the binary pixel values (0 or 1) at distance x_i and $x_i + h$, respectively, and h is the lag in kilometers; n represents the number of data points at a separation distance of h. The term $\gamma(h)$ is the semivariance at separation distance h. Assuming that the empirical variogram is best represented by an exponential model, the functional parameters describing the spatial variability can be fitted as follows:

$$\gamma(h) = c_0 + c\,(1 - e^{-h/CL}), \quad (A4)$$

where c_0 represents the nugget variance, c is the sill variance, and CL is the distance parameter known as correlation length. Conversely, the correlation function is modeled as $C = \exp(-h/CL)$, where C is the correlation.

For identifying the correlation length for retrieval error, CL_{RET}, a similar set of steps are adopted as above for rain/no-rain detection, with the exception that the binary values (0–1) are no longer pertinent. Instead, one computes the correlation length in terms of retrieval error defined as the difference between reference and satellite estimate as described below.

The conditional (i.e., reference rainfall > 0 unit) nonzero satellite rain rates, R_{SAT}, can be statistically related to corresponding conditional reference rain rates, R_{REF}, as

$$R_{SAT} = R_{REF}\,\varepsilon_S, \quad (A5)$$

where the satellite retrieval error parameter, ε_s, is assumed to be lognormally distributed. It is up to the data producers to verify the assumption of lognormality. The advantage of such an assumption is that a log transformation $[\log(R_{SAT}) - \log(R_{REF})]$ of Eq. (A5) transforms the ε_s to a Gaussian $N(\mu, \sigma)$ deviate, ε, where μ and σ are the mean and standard deviation of retrieval error, respectively.

The retrieval error parameter ε is both spatially and temporally autocorrelated. The spatial aspect has already been discussed earlier in this appendix. For tem-

poral correlation, a lag 1 autocorrelation function is used to identify the temporal variability of μ (i.e., conditional satellite rainfall bias).

APPENDIX B

Formulation of Some Common Error Metrics

Using the terminology adopted by Ebert et al. (2007), a grid box can be classified as a hit (H, observed rain is correctly detected), miss (M, observed rain is not detected), false alarm (F, rain detected but not observed), or null (no rain observed or detected).

The frequency bias is defined as

$$FB = \frac{(H + F)}{(H + M)}. \quad (B1)$$

The false alarm ratio is defined as

$$FAR = \frac{F}{(H + F)}. \quad (B2)$$

The equitable threat score is defined as

$$ETS = \frac{H - H_e}{(H + M + F - H_e)}, \quad (B3)$$

where, $H_e = (H + M)(H + F)/N$ and N = the total number of grid boxes.

REFERENCES

Ali, A., T. Lebhl, and A. Amani, 2005: Rainfall estimation in the Sahel. Part I: Error function. *J. Appl. Meteor.,* **44,** 1691–1706.

Anagnostou, E. N., 2004: Overview of overland satellite rainfall estimation for hydro-meteorological applications. *Surv. Geophys.,* **25,** 511–537.

Bellerby, T., and J. Sun, 2005: Probabilistic and ensemble representations of the uncertainty in IR/microwave rainfall product. *J. Hydrometeor.,* **6,** 1032–1044.

Deutsch, C. V., and A. Journel, 1998: *GSLIB: Geostatistical Software Library and User's Guide.* Oxford University Press, 369 pp.

Ebert, E. E., J. E. Janowiak, and C. Kidd, 2007: Comparison of near-real-time rainfall estimates from satellite observations and numerical models. *Bull. Amer. Meteor. Soc.,* **88,** 47–64.

Fulton, R. A., J. P. Breidenbach, D.-J. Seo, D. A. Miller, and T. O'Bannon, 1998: The WSR-88D rainfall algorithm. *Wea. Forecasting,* **13,** 377–395.

Gebremichael, M., and W. F. Krajewski, 2004: Characterization of the temporal sampling error in space-time-averaged rainfall estimates from satellites. *J. Geophys. Res.,* **109,** D11110, doi:10.1029/2004JD004509.

——, and ——, 2005: Modeling distribution of temporal sampling errors in area-time-averaged satellite rainfall estimates. *Atmos. Res.,* **73,** 243–259.

Heidke, P., 1926: Berechnung des Erfolges und der Gute der Windstarkevorhersagen im Sturmwarnungsdienst (in German). *Geogr. Ann.,* **8,** 301–349.

Hong, Y., K.-L. Hsu, S. Sorooshian, and X. Gao, 2005a: Self-organizing nonlinear output (SONO): A neural network suitable for cloud patch–based rainfall estimation from satellite imagery at small scales. *Water Resour. Res.,* **41,** W03008, doi:10.1029/2004WR003142.

——, ——, ——, and ——, 2005b: Improved representation of diurnal variability of rainfall from the Tropical Rainfall Measuring Mission Imager adjusted Rainfall Estimation from Remotely sensed Information Using Artificial Neural Networks (PERSIANN) system. *J. Geophys. Res.,* **110,** D06102, doi:10.1029/JD2004JD005301.

——, ——, H. Moradkhani, and S. Sorooshian, 2006: Uncertainty quantification of satellite rainfall estimation and Monte Carlo assessment of the error propagation into hydrologic response. *Water Resour. Res.,* **42,** W08421, doi:10.1029/2005WR004398.

Hossain, F., and E. N. Anagnostou, 2004: Assessment of current passive-microwave- and infrared-based satellite rainfall remote sensing for flood prediction. *J. Geophys. Res.,* **109,** D07102, doi:10.1029/2003JD003986.

——, and ——, 2005: Using a multi-dimensional satellite rainfall error model to characterize uncertainty in soil moisture fields simulated by an offline land surface model. *Geophys. Res. Lett.,* **32,** L15402, doi:10.1029/2005GL023122.

——, and ——, 2006: A two-dimensional satellite rainfall error model. *IEEE Trans. Geosci. Remote Sens.,* **44,** 1511–1522, doi:10.1109/TGRS.2005.863866.

——, and D. P. Lettenmaier, 2006: Flood prediction in the future: Recognizing hydrologic issues in anticipation of the Global Precipitation Measurement mission—Opinion paper. *Water Resour. Res.,* **44,** W11301, doi:10.1029/2006WR005202.

Huffman, G. J., R. F. Adler, M. Morrissey, D. T. Bolvin, S. Curtis, R. Joyce, B. McGavock, and J. Susskind, 2001: Global precipitation at one-degree daily resolution from multisatellite observations. *J. Hydrometeor.,* **2,** 36–50.

——, ——, D. T. Bolvin, and E. J. Nelkin, 2004: The status of the multi-satellite rainfall analysis and insights gained from adding new data sources. *Second Int. Precipitation Working Group Workshop,* Monterey, CA, World Meteorological Organization, 123–127.

——, and Coauthors, 2007: The TRMM multisatellite rainfall analysis: Quasi-global, multiyear, combined-sensor rainfall estimates at fine scale. *J. Hydrometeor.,* **8,** 38–55.

Janowiak, J., R. J. Joyce, and Y. Yarosh, 2001: A real-time global half-hourly pixel-resolution infrared dataset and its application. *Bull. Amer. Meteor. Soc.,* **82,** 205–217.

Joyce, R. L., J. E. Janowiak, P. A. Arkin, and P. Xie, 2004: CMORPH: A method that produces global precipitation estimates from passive microwave and infrared data at high spatial and temporal resolution. *J. Hydrometeor.,* **5,** 487–503.

Lee, K. H., and E. N. Anagnostou, 2004: Investigation of the nonlinear hydrologic response to rainfall forcing in physically based land surface modeling. *Can. J. Remote Sens.,* **30,** 706–716.

McCollum, J. R., W. F. Krajewski, R. R. Ferraro, and M. B. Ba, 2002: Evaluation of biases of satellite rainfall estimation algorithm over the continental United States. *J. Appl. Meteor.,* **41,** 1065–1080.

Moradkhani, H., K. Hsu, Y. Hong, and S. Sorooshian, 2006: Investigating the impact of remotely sensed rainfall and hydrologic model uncertainties on the ensemble streamflow forecasting. *Geophys. Res. Lett.,* **33,** L12401, doi:10.1029/2006GL026855.

Nijssen, B., and D. P. Lettenmaier, 2004: Effect of rainfall sampling error on simulated hydrological fluxes and states: Anticipating the Global Precipitation Measurement satellites. *J. Geophys. Res.,* **109,** D02103, doi:10.1029/2003JD003497.

North, G. R., and S. Nakamoto, 1989: Formalism for comparing rain estimation designs. *J. Atmos. Oceanic Technol.,* **6,** 985–992.

Petersen, W. A., and S. A. Rutledge, 2002: Regional variability in tropical convection: Observations from TRMM. *J. Climate,* **14,** 3566–3586.

Seo, D.-J., J. P. Breidenbach, and E. R. Johnson, 1999: Real-time estimation of mean field bias in radar rainfall data. *J. Hydrol.,* **223,** 131–147.

——, ——, R. Fulton, D. Miller, and T. O'Bannon, 2000: Real-time adjustment of range-dependent biases in WSR-88D rainfall estimates due to nonuniform vertical profile of reflectivity. *J. Hydrometeor.,* **1,** 222–240.

Shiklomanov, A. I., R. B. Lammers, and C. J. Vörösmarty, 2002: Widespread decline in hydrological monitoring threatens pan-Arctic research. *Eos, Trans. Amer. Geophys. Union,* **83,** 16–17.

Smith, E., and Coauthors, 2007: The international Global Precipitation Measurement (GPM) program and mission: An overview. *Measuring Rainfall from Space: EURAINSAT and the Future,* V. Levizzani and F. J. Turk, Eds., Kluwer Academic, 722 pp.

Sorooshian, S., K.-L. Hsu, X. Gao, H. V. Gupta, B. Imam, and D. Braithwaite, 2000: Evaluation of PERSIANN system satellite-based estimates of tropical rainfall. *Bull. Amer. Meteor. Soc.,* **81,** 2035–2046.

Steiner, M., 1996: Uncertainty of estimates of monthly areal rainfall for temporally sparse remote observations. *Water Resour. Res.,* **32,** 373–388.

——, T. L. Bell, Y. Zhang, and E. F. Wood, 2003: Comparison of two methods for estimating the sampling-related uncertainty of satellite rainfall averages based on a large radar dataset. *J. Climate,* **16,** 3759–3778.

Stokstad, E., 1999: Scarcity of rain, stream gages threatens forecasts. *Science,* **285,** 1199.