

# VERIFYING SATELLITE PRECIPITATION ESTIMATES FOR WEATHER AND HYDROLOGICAL APPLICATIONS

Elizabeth E. Ebert

Bureau of Meteorology Research Centre  
GPO Box 1289K, Melbourne, Victoria, 3001, Australia

## ABSTRACT

Near real time satellite precipitation estimates are becoming increasingly available to the wider community. These precipitation estimates are potentially very useful for applications such as NWP data assimilation, nowcasting and flash flood warning, tropical rainfall potential, and water resources monitoring, to name a few. As with any observational data, it is important to understand their accuracy and limitations. This is done by validating the satellite estimates against independent data from rain gauges and radars.

Familiar measures such as bias, correlation, and RMS error have been very useful in quantifying the errors in climate-scale satellite precipitation estimates. However, users of near real time precipitation estimates often require more specific information on expected errors in rain location, type, mean and maximum intensities. Diagnostic validation approaches can reveal additional information about the nature of the errors.

This paper discusses validation methods that give useful information to (a) help algorithm developers to improve their products, and (b) help users of short-period satellite precipitation estimates to understand the accuracy and limitations of those products.

## 1. VALIDATION OF SHORT-TERM VS. CLIMATE-SCALE PRECIPITATION

A great deal of work has been done to validate climate-scale precipitation estimates against gauge data over land and for island stations. The quantity of interest in that case is mean rain amount integrated over space and time (2.5° monthly rainfall, for example), and the statistics used to measure the accuracy of the estimates are usually the bias, correlation coefficient, and RMS error. For climate-scale rainfall the requirement is that the algorithm provide a good estimate *on average*, and so errors on shorter time and space scales are unimportant. It does not matter, for example, that the GOES Precipitation Index (GPI) erroneously associates rain with cirrus clouds and fails to detect rain from warm clouds because these errors cancel out over large space and time scales.

Users of short-term precipitation estimates cannot afford to tolerate such errors. Hydrologists need accurate estimates of rain *volume* at the catchment scale. In NWP data assimilation of satellite rainfall, experience suggests that detecting the correct rain *location* and *type* may be more important than getting the correct *amount* (Xiao et al., 2000). For flash flood warning and tropical rainfall potential, it is not only important to correctly detect the rainfall, it is also important to be able to estimate the *maximum rain rates* (e.g., Kidder et al. 2001).

The methods used to validate near real time precipitation estimates must give additional information not provided by the simpler scores. Validating short-term satellite rainfall estimates is akin to verifying\* quantitative precipitation forecasts (QPFs) from numerical weather prediction (NWP) models. Indeed, some of the methods described in this paper were developed for QPF verification.

## 2. VALIDATION DATA

Some preliminary discussion of the validation data, or ground truth, is warranted because if they are used inappropriately then the validation results will be invalid!

Due to the high spatial and temporal variability of rainfall, it is very difficult to make accurate measurements of precipitation at the scales required to validate short-term satellite precipitation estimates. Measurement and sampling errors in the observations increase the *apparent* error of the satellite estimates; this is often neglected in short-period rainfall verification. We can get away with this as long as the observational error is random (unbiased) and is much smaller than the algorithm error. Note that such imperfect “truth” data can be reliably used to *intercompare* estimates from different algorithms.

The main sources of rainfall “truth” data for satellite algorithm validation are observations from rain gauges, radar rainfall estimates, and objective analyses of one or both types of observations. Each has distinct advantages and disadvantages.

Rain gauges are the only instrument to give *direct* measurements of rain accumulation. However, because they are point measurements, they are likely to be unrepresentative of the aerial value estimated by the satellite. The time scale is generally also unrepresentative, i.e., a gauge accumulation over several minutes to several hours versus a satellite “snapshot”. Averaging over a long period of time is required to remove the representativeness errors. If satellite estimates are validated directly against individual gauge observations that are irregularly distributed in space (for example, having higher sampling density in more populous regions), then the results will be biased toward the better sampled regions. Rain gauge observations are limited to land regions and islands and atolls, and are therefore not as useful for validating oceanic satellite precipitation estimates.

Radar observations are similar to satellite estimates in that they give “snapshots” in time and aerial values in space. They have the advantage of high spatial and temporal resolution (~1 km and 5-10 minutes), which means that some of the random error will be averaged out when scaling up to the (usually) lower satellite resolution. The disadvantage with using radar data is that they are themselves indirect estimates of rainfall, and are prone to errors of calibration, attenuation, anomalous propagation, reflectivity-to-rainrate conversion, etc. Careful quality control of the data and bias correction using nearby rain gauge observations can correct much of the error in radar rainfall estimates. Like gauges, they are confined to land and near-coastal regions.

Objective analyses of rain gauges and/or radar data combine the observations onto a spatial grid. The merging of data within grid boxes reduces some of the random noise and regularizes the spatial

---

\* The terms **validate** and **verify** are taken to mean the same thing, “to determine or test the truth or accuracy of”. The former is preferred by the satellite community, while the latter is preferred by the NWP modelling community. They are used interchangeably in this paper.

distribution. When used to validate satellite precipitation estimates, the satellite data must be mapped onto the same grid as the analysis. The disadvantage of this approach is that spatial detail on the satellite estimate is then lost and maximum rain rates are somewhat diminished. The quantitative verification results depend on the scale of the analysis grid, with coarser grids generally producing more favorable results.

The most appropriate choice of validation data depends on the availability of the data, the resolution of the satellite estimates, and the needs of the user. For instantaneous and high temporal and spatial resolution estimates, gauge-corrected radar estimates or analyses are generally preferable to gauge observations. On larger space and time scales (6h to daily), rain gauge analyses or combined gauge/radar analyses are more accurate and should be used in preference to raw gauge or radar observations.

### 3. STANDARD VERIFICATION METHODS

The first step in validating satellite precipitation estimates should always be to check whether they *look right*. The most effective way to do this is to plot the estimates alongside the observations on a map or time series, and visually compare the two (sometimes called “eyeball” verification). The human mind has an acute ability to discern and interpret differences, that scientists are still struggling to objectify in a meaningful way. Some of the diagnostic verification methods discussed in the next section are attempts to do just that.

Most standard objective verification methods and statistics give quantitative measures of accuracy or skill. They will be presented only briefly here, as excellent discussions can be found in the textbook of Wilks (1995) or the WMO report of Stanski et al. (1989). The advantages of the standard statistics are (a) they are familiar and easily understood by most people, (b) they are simple to compute, and (c) they are useful for intercomparing the skill of different algorithms. Each statistic gives only one piece of information about the error, and so it is necessary to examine a number of statistics *in combination* in order to get a more complete picture.

#### a. Continuous verification statistics

Continuous verification statistics measure the accuracy of a continuous variable such as rain amount or intensity. These are the most commonly used statistics in the validation of satellite estimates. In the equations to follow  $Y_i$  indicates the estimated value at point  $i$ , and  $O_i$  indicates the observed value.

The mean error (bias) measures the average difference between the estimated and observed values. The mean absolute error (*MAE*) measures the average magnitude of the error. The root mean square error (*RMSE*) also measures the average error magnitude but gives greater weight to the larger errors.

$$\text{Mean Error} = \frac{1}{N} \sum_{i=1}^N (Y_i - O_i) \quad \text{MAE} = \frac{1}{N} \sum_{i=1}^N |Y_i - O_i| \quad \text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (Y_i - O_i)^2}$$

The linear error in probability space (*LEPS*) (Potts et al. 1996) is used more frequently in verifying seasonal climate predictions, but could be useful for short-term precipitation verification as well. Instead of measuring error in the units of the variable itself, it measures the error in *probability space*. For example, if the satellite estimate was 90 mm hr<sup>-1</sup> and the observation was 100 mm hr<sup>-1</sup>, the difference would be 10 mm hr<sup>-1</sup>, which is a large value by most standards. However, since high intensities are relatively rare most people would be happy with such an estimate. The LEPS score is designed to reflect that by measuring the difference in the climatological frequency of the event.

$$LEPS = \frac{1}{N} \sum_{i=1}^N |CDF_o(Y_i) - CDF_o(O_i)|$$

where  $CDF_o()$  is the climatological cumulative density function of the observations. (A simple approach that achieves much the same effect, at least for precipitation estimates where the expected frequency generally decreases monotonically with increasing intensity, is to express the *MAE* or *RMSE* as a percentage of the observed value.)

The correlation coefficient  $r$  measures the degree of correspondence between the estimated and observed distributions. It is independent of absolute or conditional bias, however, and therefore should be used along with other measures when validating satellite estimates.

$$r = \frac{\sum(Y - \bar{Y})(O - \bar{O})}{\sqrt{\sum(Y - \bar{Y})^2} \sqrt{\sum(O - \bar{O})^2}} \quad Skill\ score = \frac{score_{estimate} - score_{reference}}{score_{perfect} - score_{reference}}$$

Any of the error statistics can be used to construct a skill score that measures the degree of improvement over a reference estimate. The most frequently used scores are the *MAE* and the mean squared error. The reference estimate is usually climatology or persistence (the most recent set of observations), but it could also be an estimate from another algorithm.

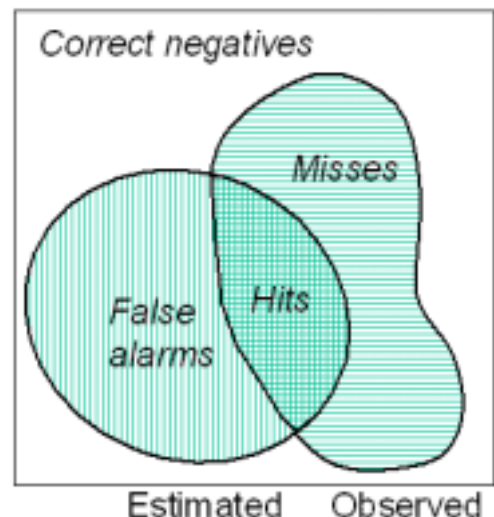
### b. Categorical verification statistics

Categorical verification statistics measure the correspondence between the estimated and observed *occurrence* of events, such as rain exceeding  $0.2\text{ mm hr}^{-1}$ . Most are based on a 2x2 contingency table of yes/no events, shown in Table 1. The elements in the table (*hits*, *misses*, etc.) give the conditional distribution of events, while the elements below and to the right (*estimated yes*, *estimated no*, etc.) are called the marginal distributions. Figure 1 shows the elements in the contingency table for a schematic map of satellite estimates and rainfall observations.

		Estimated		
		yes	no	
Observed	yes	<i>hits</i>	<i>misses</i>	<i>observed yes</i>
	no	<i>false alarms</i>	<i>correct negatives</i>	<i>observed no</i>
		<i>estimated yes</i>	<i>estimated no</i>	<i>Total</i>

**Table 1. 2x2 contingency table. The off-diagonal elements show the nature of the errors.**

The bias score gives the ratio of the estimated rain area (frequency) to the observed rain area (frequency), regardless of how well the rain patterns correspond with each other. The probability of detection (*POD*) measures the fraction of observed events that were correctly diagnosed, and is sometimes called the “hit rate”. The false alarm ratio (*FAR*) gives the fraction of diagnosed events that were actually non-events.



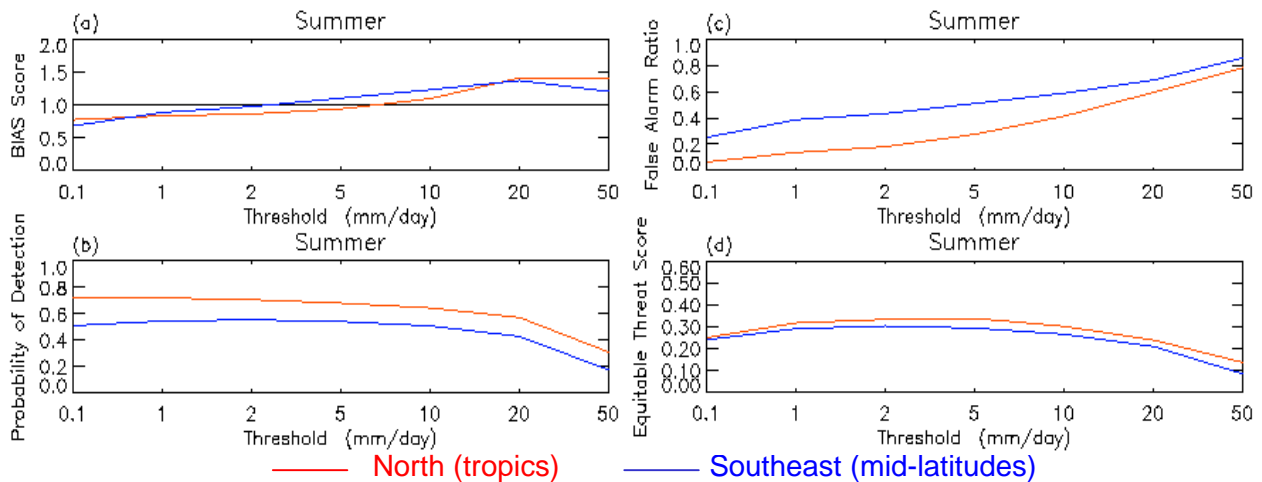
**Figure 1. Schematic map of satellite estimates and rainfall observations showing elements of the contingency table.**

$$BIAS = \frac{hits + false\ alarms}{hits + misses} \quad POD = \frac{hits}{hits + misses} \quad FAR = \frac{false\ alarms}{hits + false\ alarms}$$

The threat score (*TS*), also known as the critical success index (*CSI*), measures the fraction of all events estimated and/or observed that were correctly diagnosed. Since this score is naturally higher in wet regimes, a modified version known as the equitable threat score (*ETS*) was devised to account for the hits that would occur purely due to random chance. The *ETS*, though not a true skill score, is often interpreted that way since it has a value of 1 for perfect correspondence and 0 for no skill. It penalizes misses and false alarms equally, and for this reason it is commonly used in NWP QPF verification.

$$TS = CSI = \frac{hits}{hits + misses + false\ alarms} \quad ETS = \frac{hits - hits_{random}}{hits + misses + false\ alarms - hits_{random}}$$

To measure algorithm performance for increasingly heavier rain it is useful to plot the categorical scores as a function of an increasing rain threshold. Figure 2 shows an example for the GPCP 1-degree daily estimates (Huffman et al., 2001). An interpretation of these results is given below.



**Figure 2. Categorical verification of daily satellite precipitation estimates from the GPCP 1DD algorithm during summer 2000-01 over southeastern and tropical northern Australia.**

The bias plot (Fig. 2a) shows that the algorithm underestimated the frequency of light rain and overestimated the frequency of heavy rain. The probability of detection (Fig. 2b) was higher in the tropics than in mid-latitudes, and fell between 0.6 to 0.7 (60-70% of observed rain was detected) for all rainfall except the heaviest. The algorithm had particular difficulty with false alarms in mid-latitudes (Fig. 2c), where 50% of all rain detections above 5 mm d<sup>-1</sup> were incorrect, and 75% of rain detections above 20 mm d<sup>-1</sup> were incorrect. The equitable threat score (Fig. 2d) shows that the maximum detection skill was achieved for rain exceeding 2-5 mm d<sup>-1</sup> in both mid-latitudes and tropics, with values of 0.30 and 0.34, respectively.

Verification can also be done across multiple categories, for example, for rain rate in *K* intensity ranges. The *KxK* contingency table provides greater detail about the nature of the estimation errors than does the simple yes/no table. The marginal distribution of the estimates can also be compared to that for the observations to check for bias and skewness in the distribution, as demonstrated by Brooks and Doswell (1996).

The most frequently computed statistic for multiple categories is the Heidke skill score (*HSS*),

$$HSS = \frac{\sum_{k=1}^K P(Y_k, O_k) - \sum_{k=1}^K P(Y_k)P(O_k)}{1 - \sum_{k=1}^K P(Y_k)P(O_k)}$$

which measures the skill of the estimate relative to random chance. In the equation  $P(Y_k)$  is the marginal probability of the estimate falling in the  $k$ th category,  $P(O_k)$  is the marginal probability of the observation falling in the  $k$ th category, and  $P(O_k, Y_k)$  are the diagonal elements, i.e., the probability of a correct match.

#### 4. DIAGNOSTIC VERIFICATION METHODS

Diagnostic, or “scientific”, verification methods allow the user to explore the errors in greater detail than is usually done using the standard techniques. These more interpretive methods consider the spatial arrangement of pixels or grid boxes into recognizable patterns or entities, rather than simply matched pairs of independent estimates and observations. They attempt to provide clues as to *why* the estimates contain certain types of errors, and thus are appealing to algorithm developers. Because diagnostic methods tend to be more complex than standard methods, they are more likely to be used in research mode as opposed to operational mode.

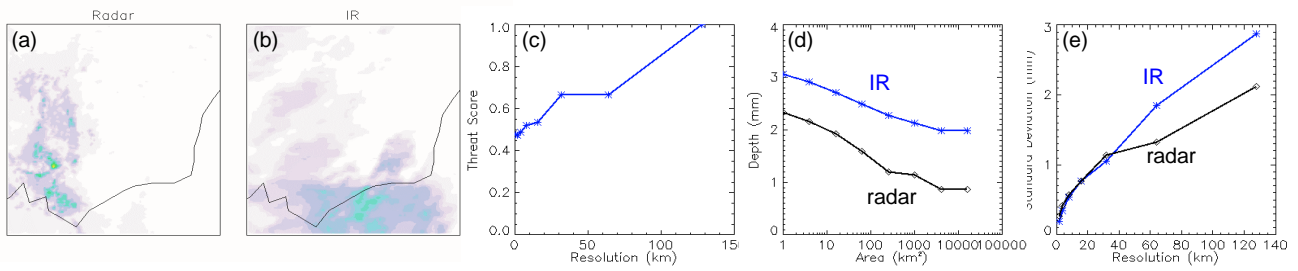
This section will give describe some verification techniques from two classes of diagnostic methods, scale decomposition methods and entity-based methods.

##### a. Scale decomposition methods

Scale decomposition methods measure the correspondence of the estimates and observations at different spatial scales. The goal is to discover which scales are best represented, and which scales are poorly represented by the estimates. Ideally, the algorithm developer would then put some effort into improving the algorithm at those poorly represented scales; if this is not possible or practical then it may be desirable to filter those scales out of the final product. Implicit in this strategy is the assumption that the estimates and observations are available at high spatial resolution.

Early scale decomposition methods used 2D Fourier transforms. In recent years discrete wavelet decomposition has received greater attention, as it has some attractive properties. Wavelets are locally defined functions characterized by a location and a spatial scale, and do not have periodic properties as does Fourier decomposition. Briggs and Levine (1997) were among the first to use wavelets as a verification tool for forecasts of 500 hPa geopotential height fields. Casati et al (2002) have developed a new verification method that first removes the bias component of the forecast (estimate) by histogram-matching recalibration against the observations, then generates binary rain images by thresholding. A wavelet decomposition is performed on the (binary) error field to obtain the structure of the errors as a function of scale. Categorical verification statistics are then computed as a function of scale to measure the contribution of each scale to the total error.

A simple scale decomposition method is the upscaling approach proposed by Zepeda-Arce et al. (2000), in which verification scores and Spatial variability measures are recomputed as the estimates are averaged over increasingly larger grid boxes. The rate of improvement with scale is a measure of the quality of the estimate. Their motivation was to give credit to high resolution model forecasts that looked good in a broad sense, but did not match the observations very well in a point-to-point sense. The same issue is valid for short-term satellite precipitation estimates.



**Figure 3. (a) Hourly radar rainfall and (b) IR power law rainrate estimate for 2230 UTC on 16 September 2002 in southeastern Australia. The verification shows (c) the threat score for rain  $\geq 0.2$  mm hr<sup>-1</sup>, (d) conditional rain intensity, and (e) spatial variability as a function of scale.**

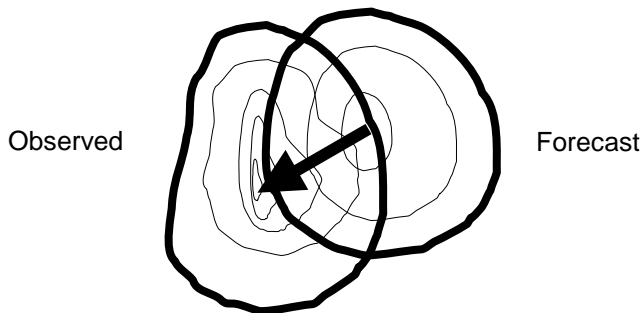
An example of the upscaling approach is shown in Figure 3 for a case of wintertime frontal rainfall in southeastern Australia, estimated by a geostationary IR power law algorithm (Vicente et al. 1998). The satellite estimate gives rain rates that are similar to those observed by the radar, but the satellite rain extends too far in the southeastly direction, possibly a result of misdiagnosed cirrus. The threat score improves with increasing scale, but achieves a perfect value of 1 only at the coarsest scale. The conditional rain rates of the satellite estimate are much higher for the radar. The spatial variability of the satellite and radar estimates are nearly identical for all but the largest scales.

### b. Entity-based methods

An alternative approach is the evaluation of forecast rain entities, or contiguous rain areas (CRAs) ("blobs" on a rainfall map), in which the bulk properties of the satellite estimated rain entity are verified against the bulk properties of the observed rain entity (Ebert and McBride 2000). The verification quantities of interest include the position error, the difference between the estimated and observed rain area, volume, and mean and maximum rain rates, and the correlation between the position-corrected estimate and observations.

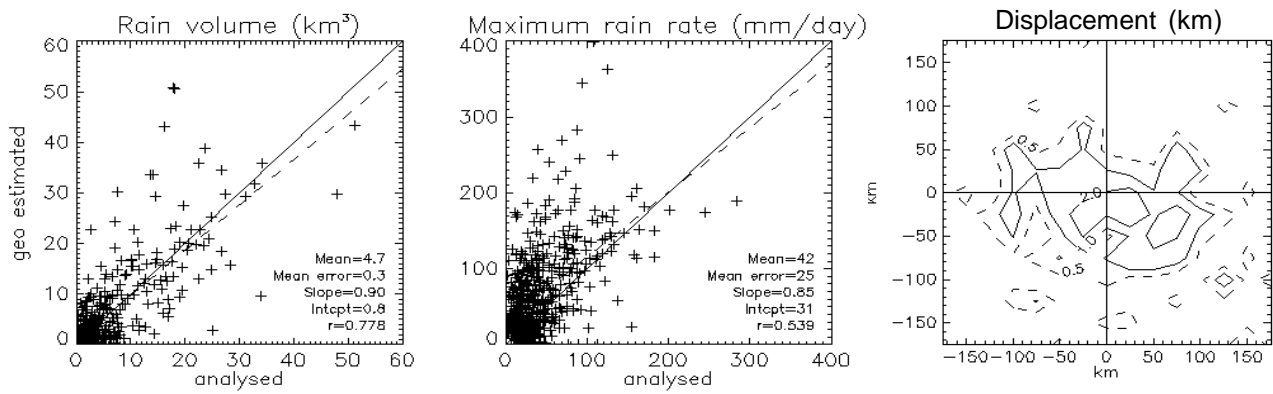
The approach is shown schematically in Fig. 4. Entities are defined by a threshold isohyet, which may be assigned a very low value to capture all rainfall, or a high value if heavy rainfall is the focus. The position error is determined (to the nearest gridpoint) by pattern matching, where the estimate is horizontally translated over the observations until a best fit criterion is satisfied. Possible best fit criteria include minimization of the total squared error, maximization of the spatial correlation coefficient, or maximum overlap; the correlation criterion appears to give the most satisfactory fit for satellite precipitation estimates. The disadvantage of entity-based verification is that if the estimate does not resemble the observations sufficiently, it may not be possible to objectively associate two entities, i.e., to decide which estimated blob goes with which observed blob.

CRA verification lends itself well to characterizing systematic errors when a large number of rain systems are verified. Figure 5 shows scatter plots of estimated versus observed rain volume and maximum rain rate, as well as a



frequency plot of location errors, for 24-hour estimates from the NRL experimental geostationary algorithm during April 2001-March 2002. The algorithm produced good estimates of rain volume within systems, although it had a tendency to overestimate their intensity. There appears to be a very slight location bias of estimates toward the south.

**Figure 4. Schematic diagram of a CRA, with the arrow showing the optimum shift of the forecast.**



**Figure 5. Verification of (a) rain volume, (b) maximum rain rate, and (c) location for 289 CRAs from the NRL experimental geostationary algorithm during April 2001-March 2002.**

When the displacement is determined using squared error minimization then it is possible to write the mean squared error as the sum of three contributions:

$$MSE_{total} = MSE_{displacement} + MSE_{volume} + MSE_{pattern}$$

(see Ebert and McBride, 2000, for details). The difference between the mean square error before and after translation is the contribution to total error due to *displacement*,

$$MSE_{displacement} = MSE_{total} - MSE_{shifted}$$

The error component due to *volume* represents the bias in mean intensity,

$$MSE_{volume} = (\bar{F} - \bar{X})^2$$

where  $F$  and  $X$  are the CRA mean estimated and observed values after the shift. The *pattern error* accounts for differences in the fine structure of the forecast and observed fields,

$$MSE_{pattern} = MSE_{shifted} - MSE_{volume}$$

For the daily rainfall estimates shown in Fig. 5 the primary source of error is related to the fine scale structure.

## 5. FINAL REMARKS

Users of short-term satellite precipitation estimates need to understand the expected accuracy and error associated with the estimates. Satellite algorithm developers also need to validate their algorithms. The standard continuous and categorical verification methods give quantitative measures of the accuracy of the satellite estimates in diagnosing rain amount and occurrence. There is no one score that summarizes the ability of an algorithm to make correct rainfall estimates. Rather, it is necessary to examine several scores in order to characterize both the skill and the errors. The newer diagnostic verification methods give more detailed information on the nature of the errors, and are therefore appealing to algorithm developers.

One of the major problems in validating satellite precipitation estimates is the imperfect “truth” data that is available for doing the validation. Although we cannot expect the situation to improve enormously in the next few decades, polarimetric radar technology and improved gauge-radar analysis methods will lead to better quality validation data. Since we know that the “truth” data do contain error, methods be developed and put in place to quantify the uncertainty in the validation results that arise from uncertainty in the observations.

## 6. REFERENCES

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

Brooks, H.E. and C.A. Doswell III, 1996: A comparison of measures-oriented and distributions-oriented approaches to forecast verification. *Wea. Forecasting*, **11**, 288-303.

Casati, B., D.B. Stephenson, and G. Ross, 2002: New methods for verifying spatial precipitation forecasts. *WWRP Intl. Conf. on Quantitative Precipitation Forecasting (QPF)*, Univ. Reading, UK, 2-6 September 2002.

Ebert, E.E. and J.L. McBride, 2000: Verification of precipitation in weather systems: Determination of systematic errors. *J. Hydrology*, **239**, 179-202.

Kidder, S.Q., J.A. Knaff, and S.J. Kusselson, 2001: Using AMSU data to forecast precipitation from landfalling hurricanes. *Symposium on Precipitation Extremes: Prediction, Impacts, and Responses*, Amer. Met. Soc., Albuquerque, NM, 14-19 January 2001, 344-347.

Huffman, G.J., R.F. Adler, M.M. Morrissey, D.T. Bolvin, S. Curtis, R. Joyce, B. McGavock, and J. Susskind, 2000: Global precipitation at one-degree daily resolution from multisatellite observations. *J. Hydrometeorology*, **2**, 36-50.

Potts, J.M., C.K. Folland, I.T. Jolliffe, and D. Sexton, 1996: Revised "LEPS" scores for assessing climate model simulations and long-range forecasts. *J. Climate*, **9**, 34-53.

Stanski, H.R., L.J. Wilson, and W.R. Burrows, 1989: *Survey of common verification methods in meteorology*. World Weather Watch Tech. Rept. No.8, WMO/TD No.358, WMO, Geneva, 114 pp.

Vicente, G.A., R.A. Scofield, and W.P. Menzel, 1998: The operational GOES infrared rainfall estimation technique. *Bull. Amer. Met. Soc.*, **79**, 1883-1898.

Wilks, D.S., 1995: *Statistical Methods in the Atmospheric Sciences. An Introduction*. Academic Press, San Diego, 467 pp.

Xiao, Q., X. Zou, and Y.-H. Kuo, 2000: Incorporating the SSM/I-derived precipitable water and rainfall rate into a numerical model: a case study for the ERICA IOP-4 cyclone. *Mon. Wea. Rev.*, **128**, 87-108.

Zepeda-Arce, J., E. Foufoula-Georgiou, and K.K. Droegemeier, 2000: Space-time rainfall organization and its role in validating quantitative precipitation forecasts. *J. Geophys. Res.*, **105 (D8)**, 10,129-10,146.