Performance of IMERG as a Function of Spatiotemporal Scale

Jackson Tan*

Universities Space Research Association, and NASA Goddard Space Flight Center, Greenbelt, Maryland

Walter A. Petersen

Earth Sciences Office, ZP-11, NASA Marshall Space Flight Center, Huntsville, Alabama

Pierre-Emmanuel Kirstetter

School of Civil Engineering and Environmental Sciences, University of Oklahoma, Norman, Oklahoma, and NOAA/National Severe Storms Laboratory, Norman, Oklahoma

Yudong Tian

Earth System Sciences Interdisciplinary Center, University of Maryland, College Park, College Park, Maryland, and NASA Goddard Space Flight Center, Greenbelt, Maryland

**Corresponding author address:* Jackson Tan, NASA Goddard Space Flight Center, Code 613 Building 33 Room C327, 8800 Greenbelt Road, Greenbelt, MD 20771.

E-mail: jackson.tan@nasa.gov

ABSTRACT

The Integrated Multi-satellitE Retrievals for GPM (IMERG), a global high-resolution gridded precipitation data set, will enable a wide range of applications, ranging from studies on precipitation characteristics to applications in hydrology to evaluation of weather and climate models. These applications focus on different spatial and temporal scale and thus average the precipitation estimates to coarser resolutions. Such a modification of scale will impact the reliability of IMERG. In this study, the performance of the Final run of IMERG is evaluated against ground-based measurements as a function of increasing spatial resolution (from 0.1° to 2.5°) and accumulation periods (from 0.5 h to 24 h) over a region in the southeastern US. For ground reference, a product derived from the Multi-Radar/Multi-Sensor suite, a radar- and gauge-based operational precipitation dataset, is used. The TRMM Multisatellite Precipitation Analysis (TMPA) is also included as a benchmark. In general, both IMERG and TMPA improve when scaled up to larger areas and longer time periods, with better identification of rain occurrences and consistent improvements in systematic and random errors of rain rates. Between the two satellite estimates, IMERG is slightly better than TMPA most of the time. These results will inform users on the reliability of IMERG over the scales relevant to their studies.

1. Introduction

Satellite retrievals of precipitation are instrumental in understanding the distribution of precipitation around the globe. In regions with sparse measurements, such as mountainous areas and oceans, these remotely sensed estimates help to bridge gaps and constrain the errors in ground-based data. This is typically achieved through the use of gridded high resolution precipitation datasets, such as the Integrated Multi-satellitE Retrievals for GPM (IMERG; Huffman et al. 2015), the TRMM Multisatellite Precipitation Analysis (TMPA; Huffman et al. 2007), Climate Prediction Center morphing algorithm (CMORPH; Joyce et al. 2004; Joyce and Xie 2011), and Precipitation Estimation from Remotely Sensed Information using Artificial Neural Networks Cloud Classification Scheme (PERSIANN-CCS; Hong et al. 2004). These gridded precipitation datasets use a blend of data from various sources with advanced techniques to provide a near-global coverage with high spatial and temporal resolution.

However, to understand and benchmark the performances of these datasets, they need to be evaluated against ground measurements. To this end, a whole range of ground validation efforts have been undertaken to evaluate these datasets based on different criteria. Some studies focus on different rain systems (e.g. Ebert et al. 2007; Habib et al. 2009; Roca et al. 2010; Mei et al. 2014). Some studies analyze the performance by terrain or surface (e.g. Tian and Peters-Lidard 2007; Kubota et al. 2009; Stampoulis and Anagnostou 2012; Chen et al. 2013b; Liu 2016). Some studies investigate the downstream impact of the estimates on hydrologic modeling (e.g. Gottschalck et al. 2005; Xue et al. 2013; Falck et al. 2015; Tang et al. 2016b). Some studies focus on a better understanding of the errors in these datasets themselves (e.g. Maggioni et al. 2014; Tang et al. 2015; Tan et al. 2016).

The aim of this study is to quantify the performance of IMERG as a function of spatial and temporal scale. Similar analyses have been performed for other products. For example, Tian et al. (2007) compared TMPA and CMORPH at daily, seasonal and annual time scales against ground radar and gauges, finding that CMORPH is better at daily resolution while TMPA is superior at the longer time scales. On the other hand, Hossain and Huffman (2008) examined the sensitivity of various metrics to spatial and temporal scale in PERSIANN-CCS against rain gauges, and found that the probability of detection of rain is most sensitive to scale, followed by correlation length. Gourley et al. (2010) evaluated TMPA and PERSIANN-CCS against a radar-based product as a function of spatial scale, temporal scale and intensity, showing that TMPA is better than PERSIANN-CCS, though both had reduced skill at higher intensities. Habib et al. (2012) investigated the performance of CMORPH against gauges and radar across a range of spatial and temporal scales, with the conclusion that random error decreases with increasing scale. Sarachi et al. (2015) proposed a statistical model to quantify the uncertainties in gridded satellite estimates by deriving parameters to a generalized normal distribution as a function of scale.

In this study, we build on these studies and evaluate the IMERG Final run on its ability to identify rain occurrences and rain rates over a range of spatial and temporal scales against a ground-based dataset derived from the Multi-Radar/Multi-Sensor product over a region in the United States. Our goal is to examine how various aspects of IMERG change as it is averaged over larger areas and longer periods. For example, it is expected that random errors would decrease with more averaging; indeed, our study will show that averaging the estimates in a 0.1° grid box from 0.5 h to 24 h will reduce the normalized root-mean-square error from 1.7 to 1.0. Hence, our results also provide users with quantitative information on the performance of IMERG at a scale suitable to their purposes.

2. Data

a. IMERG

The Integrated Multi-satellitE Retrievals for GPM (IMERG) is a gridded precipitation product that merges measurements from a network of satellites in the GPM constellation (Huffman et al. 2015). IMERG uses the GPM Core Observatory satellite, which has a dual-frequency precipitation radar and a 13-channel passive microwave imager, as a reference standard to intercalibrate and merge precipitation estimates from individual passive microwave (PMW) satellites in the constellation (Hou et al. 2014). Lagrangian time-interpolation is then applied to these estimates using displacement vectors derived from infrared (IR) measurements on geosynchronous satellites to produce gridded high resolution estimates of rainfall. This process, known as morphing, was first introduced as the central component in CMORPH (Joyce et al. 2004; Joyce and Xie 2011). This gridded estimate is further supplemented via a Kalman filter with microwave-calibrated rainfall estimates calculated directly from IR measurements following the PERSIANN-CCS algorithm (Hong et al. 2004). The final satellite estimate is then calibrated, either directly for the post-real-time product or indirectly for the near-real-time products, using gauge data from the Global Precipitation Climatology Centre monthly precipitation dataset following the approach employed in TMPA (Huffman et al. 2007).

IMERG has a high resolution of 0.1° every half-hour covering up to $\pm 60^\circ$ latitudes. Three choices of IMERG runs are available depending on user requirements. The Early run, available at a 6-hour delay for real-time applications such as for hazard predictions, is limited to rainfall morphing only forward in time. The Late run, with a 18-hour delay for purposes such as crop forecasting, employs morphing both forward and backward in time. The Final run is at a 4-month delay for research applications. Both the Early and Late runs have climatological gauge adjustment while the Final run uses monthly gauge adjustments to reduce bias. Moreover, runs with longer delays will use more PMW estimates due to latency in data delivery. Note that these delays will eventually be reduced towards the targets of 4-hour, 12-hour and 2-month respectively. This study focuses on the calibrated estimate from Final run of IMERG, which is available from Apr 2014 onwards.

Currently, IMERG ingests data from Version 3 of GPM, which uses algorithms implemented at the launch of the GPM Core Observatory in Feb 2014 and is thus subject to further improvements as measurements are collected. The release of an updated IMERG using Version 4 of the GPM products is imminent and may involve potential improvements. We do not expect this new version of IMERG to introduce major changes to the results of our study; however, should any significant difference arise, we will address the changes in a follow-up paper. IMERG can be downloaded from http://pmm.nasa.gov/data-access.

b. TMPA

The TRMM Multisatellite Precipitation Analysis (TMPA; also known as TRMM 3B42) is the gridded precipitation product from the TRMM project. Just as with IMERG, TMPA uses the TRMM satellite to calibrate and combine PMW estimates from different platforms. Estimates derived from geosynchronous IR measurements calibrated against PMW estimates on a monthly basis are used to fill in the gaps in the PMW field.

TMPA is available at a resolution of $0.25°$ every 3-hour covering up to $\pm 50°$ latitudes. Two different products of TMPA are available: the real-time product (with a 9-hour delay) and the research product. This study uses the research product, which is available beginning 1998. The research product utilizes the TRMM Precipitation Radar on board the satellite for calibration and has the additional monthly gauge adjustment step. TMPA can be downloaded from http://pmm.nasa.gov/data-access.

Due to the decommissioning of the TRMM satellite, the TMPA research product switches, in Oct 2014, from calibration with the Precipitation Radar to a climatological calibration modified from the real time product. While this change may introduce a discontinuity from Sep to Oct 2014, the use of gauge adjustment should minimize, if not eliminate, artifacts for estimates over land (Bolvin and Huffman 2015).

c. Reference

The Multi-Radar/Multi-Sensor (MRMS; formerly National Mosaic and Multi-Sensor QPE) system is a gridded product by NOAA/NSSL based primarily on the US WSR-88D network (Zhang et al. 2011b). Reflectivity data are mosaicked onto a 3D grid over the United States with quality control for beam blockages and bright band. From the reflectivity structure and environmental field at each grid point, a precipitation regime (e.g. snow, stratiform rain, convective rain) is determined using physically-based heuristic rules and a corresponding reflectivity-precipitation relationship is applied to estimate the surface precipitation rate. These precipitation rates are bias-corrected using gauge data from the Hydrometeorological Automated Data System[1] and regional rain gauge networks. A radar quality index (RQI) is produced alongside each precipitation estimate in MRMS (Zhang et al. 2011a), providing a numerical value that reflects sampling and estimation uncertainty, such as beam issues relating to orography and bright bands. Evaluation of MRMS shows better performances with the gauge correction and the quantitative benefit of the RQI filter (Chen et al. 2013a; Kirstetter et al. 2015a).

[1] More information on the Hydrometeorological Automated Data System is available at http://www.nws.noaa.gov/oh/hads/WhatIsHADS.html.

For the analysis herein, we use a reference dataset processed from the MRMS suite in support of the GPM mission for ground validation, available from Jun 2014 onwards (Kirstetter et al. 2012, 2014, 2015b; Gebregiorgis et al. 2016). This product aggregates the MRMS rain rates to produce a half-hourly accumulated rain rates over the conterminous United States (20°–55°N, 130°–60°W) with a high spatial resolution of 0.01°. For this reference product, the RQI ranges from 0 (lowest quality) to 100 (highest quality). We mask pixels with RQI less than 100, thus keeping only perfect-RQI pixels in computing the areal averages. A perfect RQI indicates an absence of blockage and a radar beam below the bright band. We also exclude all pixels in which frozen precipitation is identified. Thus, this study focuses only on the most reliable estimates of liquid precipitation.

3. Approach

We restrict our analysis to 30.0–41.5°N, 93.5–83.5°W, a region within which the reference is highly reliable due to good radar coverage, high density of gauges and absence of significant orography. The RQI in this region is generally high (Fig. 1). This flat topography, together with a lack of frozen surfaces at most times of the year, also means that satellite retrievals are generally more accurate, though the reliance on ice scattering in retrievals over land will lead to challenges in the estimation of warm rain. Within this region, we randomly sample an ensemble of 100 square boxes of length 0.1° and extract the IMERG and reference precipitation rates in each of these boxes over the period of 19 months (Jun 2014 to Dec 2015). We then do the same for square boxes of length 0.2° (i.e. 2×2 IMERG grid boxes), repeating it at 0.1° increments up to and including 2.5°. From these rates as a function of spatial scale, we average them to get rates over periods of 1 h, 3 h, 6 h, 12 h and 24 h. This is also done separately for TMPA and the reference, at increments of 0.25° to 2.50° and periods of 3 h, 6 h, 12 h and 24 h. Therefore, for each spatial and temporal scale, we have 100 sets of precipitation rates between IMERG and the reference as well as TMPA and the reference, from which we can derive the statistics for each pair of rain rates and take the average across the ensemble to reduce sampling bias. Note that we are working with precipitation rate and not accumulated precipitation; in other words, the units of the precipitation are mm / h over 1 hr, 3 hr, ... , 24 h instead of mm.

The period of this analysis covers 19 months over 2014 and 2015 without a distinction between different seasons. Additional analyses for the warm season (Apr 2015 to Sep 2015) and the cold season (Oct 2014 to Mar 2014) show that the difference is generally an offset in the performance of IMERG, with the warm season slightly better than the cold season as consistent with previous studies (Guo et al. 2016; Liu 2016). However, as the behavior of the performance as a function of scale is generally similar between the two seasons, we will not distinguish between the two seasons in the following sections. Instead, readers interested in the results for each season can refer to the Supplementary Material.

We evaluate IMERG and TMPA against the reference on two aspects: (i) rain occurrences, i.e. if they agree that it is raining above a certain threshold or not; and (ii) rain rates, i.e. when both are raining, the degree to which the rates are similar. This follows the approach advocated in Tang et al. (2015). As such, our analyses may depend considerably on the chosen threshold. This presents an immediate challenge as rain rates are a function of scale, a situation well exemplified in Fig. 2, which shows better agreement between IMERG and the reference at longer and larger scales. While we expect rain rates to decrease with increasing scale due to coarsening, the fraction of raining events actually increase, as demonstrated in Fig. 3 through a fixed threshold of 0.2 mm / h. This will have a bearing on the results because many aspects of rainfall evaluation, such as the probability of detecting rain, are a function of the number of raining events.

Instead of using a fixed threshold at all scales, we reduce the threshold with increasing scale. Since the purpose of a threshold is to account for measurement uncertainty, this uncertainty and thus the threshold should decline as we consider more grid boxes. In the limit of a very large scale, measurement uncertainty should be infinitesimally small. This then leads to the next question of how the threshold should decline with scale. To resolve this, we draw our inspiration from the Central Limit Theorem (Wilks 2011), whereby the standard deviation of a sample mean is the population standard deviation divided by $\sqrt{N}$, where N is the number of samples. In our case, we set our threshold at box length l and time period t as $T(l,t) = T(0.1°, 0.5\text{ h})/\sqrt{N}$, where N is the number of grid boxes and time steps that we averaged over. This leads to

$$T(l,t) = \frac{T(0.1°, 0.5\text{ h})}{\sqrt{\frac{l}{0.1°} \times \frac{l}{0.1°} \times \frac{t}{0.5\text{ h}}}}. \quad (1)$$

We set $T(0.1, 0.5\text{ h}) = 0.2$ mm / h, which is the minimum nonzero value of IMERG rain rates prior to gauge adjustment (personal comm., G. Huffman, 2014). Fig. 4 shows the thresholds as a function of scale calculated in this way. In the Supplementary Material, we provide an alternative set of figures, showing values calculated using a constant threshold of 0.2 mm / h.

With a scale-consistent set of thresholds, we consider an estimate to be raining if the precipitation rate is at least that of the threshold and not raining if it is below the threshold. This approach allows us to construct a contingency matrix (hits, misses, false alarms, and correct negatives) for each ensemble member of every scale, from which we can calculate the probability of detection, false alarm ratio, bias in detection, and Heidke skill score (Wilks 2011). The probability of detection is the fraction of actual rain occurrences that the estimate detected; a perfect score is 1. The false alarm ratio is the fraction of rain occurrences in the estimates that are wrong; a perfect score is 0. The bias in detection quantifies the tendency for the estimate to overestimate (> 1) or underestimate (< 1) the number of rain occurrences; a perfect score is 1. Bias in detection, also known as bias ratio (Wilks 2011), should not be confused with "bias", which is a measure of rain rate. The Heidke skill score is a generalized skill score than quantifies whether the estimate is worse (< 0) or better (> 0) than random chance; a perfect score is 1. Then, for the subset of the hits, we calculate the correlation, normalized mean error, normalized mean absolute error and root-mean-squared error, as well as parameters used in the multiplicative error model of Tian et al. (2013). These quantities are defined in Appendix. In the following sections, we will present these quantities as a function of scale, averaged over all ensemble members. Note that as we are using square boxes, an increase in spatial scale correspond to a squared increase in the actual area (e.g. double the box length from $0.1°$ to $0.2°$ increases the area by a factor of 4).

4. Evaluation of Rain Occurrences

We begin our evaluation by examining the ability of the satellite estimates to identify the rain occurrences. Fig. 5 gives the average percentages of hits, misses, false alarms and correct negatives between IMERG/TMPA and the reference. The percentage of hits increases monotonically with increasing scale for IMERG and TMPA, which is expected since there are more rain occurrences even with a constant threshold (Fig. 3), much less for a threshold that decreases with scale. For the same reason, the percentage of correct negatives decreases monotonically for both IMERG and TMPA. The percentage of misses (false negatives) in IMERG increases with scale but converge to between 8% and 9% at $2.5°$. The increase itself may be a consequence of the lower threshold at coarser scales, but the fact that the percentage of misses approaches a common value may be an indication of the merit of Eq. (1). On the other hand, for TMPA, whether the percentage of misses increases with spatial scale depends on the temporal scale, and vice versa. For example, the percentage of misses at 3 h increases with spatial scale while that at 24 h decreases with spatial scale. Interestingly, IMERG at 24 h also exhibits a similar behavior at coarser spatial scales, though with a more muted decline. Finally, for false alarms (false positives), the percentage in IMERG increases with scale, though remaining below 8% over the range of scales considered. Likewise, the percentage of false alarms for TMPA increase with scale, though with larger magnitudes and at a faster rate. The percentage of false alarms is higher in the cold season than in the warm season (not shown).

From the rain occurrences, we can calculate the probability of detection, false alarm ratio, bias in detection and Heidke skill score as a function of scale (Fig. 6). The probabilities of detection for both IMERG and TMPA rise monotonically with scale. This means that both datasets are better at identifying rain occurrences at coarser scales. Between IMERG and TMPA, the former is better at finer scales, but the probability of detection for TMPA increases more rapidly with spatial scale and outperforms IMERG after 1.0° to 1.5°. At 24 h and 2.5°, the probability of detection is 0.87 for IMERG and 0.90 for TMPA. The probability of detection remains above 0.5 at all scales.

The false alarm ratios for IMERG decline rapidly with scale, but the improvement diminishes at coarser scales (Fig. 6). This means that, of all the occurrences which the estimates classify as raining, the fraction that are false positives decreases as IMERG estimates are averaged over larger areas and longer periods. For TMPA, the false alarm ratios remain roughly constant with spatial scales, but is lower at longer periods. This behavior of constant performance with spatial scale is due to the decreasing thresholds; when we use a constant threshold of 0.2 mm / h, the false alarm ratios for TMPA decrease with spatial scale just like in IMERG (Supplementary Material). Regardless of the threshold or scale, IMERG has consistently lower false alarm ratios than TMPA. Taking together the fact that TMPA has higher probability of detection but also higher false alarm ratios than IMERG, it suggests the possibility that TMPA identifies more rain events than IMERG.

The bias in detection of IMERG remains below one for the range of scales considered here (Fig. 6). This means that IMERG is underestimating the number of rain occurrences, though there is a gradual increase towards one with increasing grid box size. For TMPA, the bias in detection does not differ between different temporal scales, but it increases sharply with the size of the box, overshooting the ideal value of one at about 1.0°. Therefore, on the number of rain occurrences, TMPA underestimates in grid boxes smaller than 1.0° but overestimates in grid boxes larger than 1.0°. The behavior of the bias in detection in both IMERG and TMPA reflect the asymmetry in how the percentages of misses and false alarms change (Fig. 5). Since the bias in detection has false alarms in the numerator and misses in the denominator (see Appendix), the greater increase in misses than in false alarms meant that bias in detection will increase. Using a constant threshold of 0.2 mm / h, the bias in detection of both IMERG and TMPA are roughly constant with scale, with TMPA being closer to one than IMERG (Supplementary Material).

Finally, the Heidke skill scores for IMERG and TMPA are well above zero for all scales (Fig. 6), with IMERG consistently outperforming TMPA. This means that both datasets are better at identifying rain occurrences than random chance. For IMERG, the scores generally increase with spatial and temporal scale, though reaching an asymptotic value of about 0.70. However, for TMPA, the Heidke skill score either remains constant or declines with scale, though this is primarily due to the decreasing threshold: using a constant threshold of 0.2 mm / h results in an improvement in scale similar to IMERG (Supplementary Material).

In summary, Figs. 5 and 6 evaluate the performance of IMERG and TMPA in identifying rain occurrences. They showed that IMERG is in general better at identifying rain occurrences at larger spatial scale and longer temporal scale, though this improvement is not always monotonic. TMPA, on the other hand, provides mixed results with increasing scale. Between IMERG and TMPA, the former is generally better, primarily due to the lower percentage of false alarms. However, these results are strongly affected by the thresholds (Fig. 4) as alternative figures for a constant threshold of 0.2 mm / h have shown (Supplementary Material). Therefore, even though we see that the aggregation of rainfall estimates over longer periods and larger areas improve the performance, results on rain occurrences are sensitive to the chosen threshold. Because of this, we also provide, in the Supplementary Material, the data computed in this section over a range of thresholds (i.e. instead of fixing the threshold, we have three dependence variables on top of spatial and temporal scale).

5. Evaluation of Rain Rates

The previous section evaluated the ability of IMERG and TMPA to identify rain occurrences. In this section, we select the subset of hits, i.e. cases in which both the satellite estimate and the ground reference are equal or above the thresholds, and further investigate how well the satellite-retrieved rain rates match those from ground measurements. We begin by examining the correlation coefficient between IMERG/TMPA and the reference (Fig. 7). On this measure, both IMERG and TMPA shows a clearly increasing correlation with increasing scale though with diminishing returns at coarser scales. Notably, IMERG has significantly higher correlations than TMPA at the same scale. For example, at 3 h and 0.5°, IMERG has a correlation of 0.68 whereas TMPA has a correlation of only 0.56. In fact, even the 1 h IMERG correlations are better than the 3 h TMPA correlations.

A similar improvement in the rain rates as a function of scale is also present in the three errors calculated (Fig. 8). All three errors generally decrease at coarser scales. For normalized mean error, with the exception of IMERG at 0.5 h, the errors decline with increasing spatial scale but rapidly levels off at about zero after 1.0°. This implies that some spatial aggregation of IMERG and TMPA will remove most of the systematic error. For IMERG at 0.5 h, the normalized mean error becomes negative in grid boxes larger than 0.3°, but this underestimation is largely due to the decreasing thresholds with scale as negative normalized mean errors is not present when a constant threshold is used (Supplementary Material). Regardless, it should be noted that the magnitudes of normalized mean errors are small, being mostly below ± 0.1 as compared to mostly above +0.5 in the normalized mean absolute error. This lower value in the normalized mean error is expected due to the cancellation of positive and negative errors in a dataset that has been gauge-adjusted for systematic error. What is also shown in Fig. 8 that averaging over larger spatial scales further reduces the systematic error in general.

Both normalized mean absolute error and normalized root-mean-square error show comparable behavior. Both errors have higher magnitudes than normalized mean error. Since they are more strongly influenced by random error, the reduction of the two errors with a greater degree of averaging is not surprising. One puzzling observation in Fig. 8 is how the two errors for 0.5 h declines with scale faster than for 1 h and 3 h, such that the 0.5 h estimates actually have lower errors than the 1 h and 3 h estimates; the reason for this is unclear. One salient distinction between the two errors is that IMERG is better than TMPA in normalized mean absolute error whereas the reverse is true for normalized root-mean-square error. Since normalized root-mean-square error is affected by outliers to a greater degree, this suggests that IMERG has more outliers and/or the outliers have larger magnitudes. One plausible explanation for this is the fact that IMERG uses a pre-launch GPM database (Version 3); it is likely that the transition to a full GPM database will improve the accuracy of IMERG.

One drawback of correlations and the errors employed thus far is the assumptions of additive errors and Gaussian distribution that underpin their formulation. As rain rates are not normally distributed, such assumptions may not adequately represent the statistics of rainfall, resulting in problems such as a changing variance with rain rate and the failure to properly distinguish between systematic and random errors (Tian et al. 2013, 2016). As such, here we adopt the multiplicative error model, a framework that has greater validity for rainfall. This approach fits the estimate and the reference in a power-law relationship, with two parameters α and β expressing the systematic error and the parameter σ representing the bias-adjusted random error (see Appendix for more details).

The three parameters of the multiplicative error model have different responses to increasing spatial and temporal scales (Fig. 9). At the finest scales, α is positive but rapidly becomes negative with just a slight increase in scale, both spatially and temporally. While there is some improvement at the coarsest scale, α remains negative throughout. On the other hand, β shows a more expected response consistent with the normalized mean error: a gradual increase with spatial and temporal scale towards the perfect value of 1. In fact, IMERG has a β of one at 24 h and 2.5°. To interpret the combined behavior of α and β, we must bear in mind that α represents a multiplicative offset while β represents the dynamic range (see Fig. A1). In this light, what our results suggest is that, with upscale averaging, IMERG and TMPA are better able to capture the actual range of the rain rates, but this comes at a cost of a bias towards lower values on the whole.

As for the bias-adjusted random error, σ clearly decreases with longer temporal scale as expected, but its behavior with spatial scale is inconsistent with what we have observed in normalized mean absolute error and root-mean-square error. Instead of a monotonic decline, σ actually rises sharply until about 0.5° before falling very gradually. This bizarre behavior in σ is apparently due to how our thresholds are chosen in Eq. (1). Indeed, when we use a fixed threshold of 0.2 mm / h, σ decreases with coarser scales similar to normalized root-mean-square error (Supplementary Material).

In summary, Figs. 7, 8 and 9 evaluate the performance of IMERG and TMPA in identifying rain rates of raining events. They showed that both satellite estimates generally have improved performance at larger spatial scale and longer temporal scale, both for systematic and random errors. The decomposition using the more relevant multiplicative error model, however, suggests that the improvement is more subtle: upscaling improves the range of rain rates in the estimates as compared to the reference, but it also adds an overall bias towards lower values. In general, IMERG is better than TMPA. The impact of our chosen thresholds is lower for rain rates than for rain occurrences, with its effect only evident for σ. Just as with the quantities calculated in Sec. 4, the Supplementary Material contains data for the quantities in this section over a range of thresholds.

6. Conclusion

In this study, we evaluated IMERG, the gridded satellite rainfall product from GPM, against a ground-based reference dataset derived from MRMS as a function of spatial and temporal scale, using TMPA as a benchmark. The motivation behind this study is to acquaint users of IMERG with its performance at a scale that is relevant to their purpose. This evaluation is performed over a region where the reference is reliable due to dense radar coverage and general absence of significant orography. We examined IMERG based on two aspects: (i) whether it can identify rain occurrences above a specified threshold, and (ii) whether it can capture the correct rain rates when it correctly identifies rain occurrences.

In general, both IMERG and TMPA improve when scaled up to larger areas and longer time periods. In terms of identifying rain occurrences, there is an increase in misses and false alarms at coarser scales due to our threshold definition, but the four skill scores demonstrate that IMERG is on average better able to identify rain occurrences at coarser scales than TMPA. However, these results on rain occurrences are sensitive to the chosen rain/no-rain threshold. In terms of the rain rates, there are consistent improvements in correlations and both systematic error and random error. This reduction in random error with scale is also reported in similar studies (e.g. Roca et al. 2010; Habib et al. 2012). However, results from multiplicative error model suggest that these improvements may have subtle compensating changes. Between the two products, IMERG is slightly better than TMPA at identifying rain occurrences and estimating rain rates. This is consistent with early studies on IMERG, finding that it has generally comparable or better performance than TMPA (Guo et al. 2016; Tang et al. 2016a,b).

Our results provide a reference for IMERG users on its performance specific to their purpose. For example, in an evaluation of daily precipitation in a climate model with resolution of 1.0°, our results show that IMERG can correctly identify whether it is raining or not (at a threshold of 0.004 mm / h) 85% of the time with a Heidke skill score of 0.68, and the rain rates have a normalized root-mean-square error of 0.9. Alternatively, if IMERG were to be used for hydrological modeling over a basin of area equivalent to $2.5^\circ \times 2.5^\circ$ at hourly resolution, it will miss 8.5% of the rain occurrences ($\geq$ 0.008 mm / h), falsely identify a positive 5.5% of the time, and have a correlation of 0.78 on its rain rates.

While the results in this study are restricted to land and over a limited range of latitudes, the relative performance between different scales should be applicable to all regions. Furthermore, the values in this study may be "transferred" to other regions according to our understanding of how satellite retrievals of rain rates perform over different regions. For example, for regions that are similar to our area of study, i.e. land surfaces in the low to mid-latitude with some vegetation cover and no significant orography, our results should be directly applicable. Over oceans, it is likely that the performance of IMERG will be better due to better microwave retrieval over ocean. On the other hand, we would expect IMERG to perform poorer over mountainous areas, so the results here may indicate a likely upper bound. In a similar way, since we do not expect the Early and Late runs of IMERG to be better than the Final runs, the results here set an upper limit for the performance of these estimates. As such, with the knowledge of the relative performance of microwave retrievals between the region of interest and the region considered here, the results herein will be useful for IMERG users in better understanding the performance of the dataset.

Acknowledgments. We are grateful to George Huffman, David Bolvin and Ali Tokay for discussions on the direction of this study, as well as three anonymous reviewers for their detailed suggestions on improving the manuscript. JT is supported by an appointment to the NASA Postdoctoral Program at Goddard Space Flight Center, administered by Universities Space Research Association through a contract with NASA. WAP acknowledges support from the GPM Mission (Project Scientist, Gail S.-Jackson, and GV Systems Manager, Mathew Schwaller) and also PMM Science Team funding provided by Dr. Ramesh Kakar. YT is supported by the National Aeronautics and Space Administration Precipitation Science Program under solicitation NNH09ZDA001N. The IMERG and TMPA data were provided by the NASA/Goddard Space Flight Center's PMM and PPS teams, which develop and compute IMERG and TMPA as a contribution to GPM and TRMM respectively, and archived at the NASA GES DISC. All codes used in this analysis are freely available at [URL to be provided upon publication].

APPENDIX

Definition of Metrics, Errors and the Multiplicative Error Model

We evaluate the satellite estimate against the ground reference based on its ability to identify (i) rain occurrences and (ii) rain rates of the hits. To evaluate rain occurrences, we count the number of hits (both estimate and reference are raining), misses (estimate is below threshold while reference passes the threshold), false alarms (estimate passes the threshold when reference is below threshold), and correct negatives (both estimate and reference are below threshold). We denote these as H, M, F, and C respectively. We remind readers that our threshold varies with scale (Fig.

4). Then, we can calculate the probability of detection, false alarm ratio and bias in detection, defined as,

$$\text{probability of detection} = \frac{H}{H+M}, \tag{A1}$$

$$\text{false alarm ratio} = \frac{F}{H+F}, \tag{A2}$$

$$\text{bias in detection} = \frac{H+F}{H+M}, \tag{A3}$$

$$\text{Heidke skill score} = \frac{H+C-H_e}{N-H_e}, \tag{A4}$$

where

$$H_e = \text{no. of correct rain occurrences by chance} = \frac{1}{N}\left((H+M)(H+F)+(C+M)(C+F)\right), \tag{A5}$$

and N is the sample size (Wilks 2011). It may help to recall that $H+M$ is the number of rain events according to the reference while $H+F$ is the number of rain events according to the estimate. Probability of detection is also sometimes called hit rate; bias in detection is also known as bias ratio and should not be confused with rain rate bias.

The perfect value for probability of detection, bias in detection and Heidke skill score is one; the perfect value for false alarm ratio is zero. We compute these scores for each ensemble member, and then average across the ensemble to obtain the mean scores as a function of scale.

For the hits, we can further evaluate their rain rates using normalized mean error, normalized mean absolute error and root-mean-square error, define as,

$$\text{normalized mean error} = \frac{\frac{1}{n}\sum_i (y_i - x_i)}{\overline{x}}, \tag{A6}$$

$$\text{normalized mean absolute error} = \frac{\frac{1}{n}\sum_i |y_i - x_i|}{\overline{x}}, \tag{A7}$$

$$\text{root-mean-square error} = \frac{\sqrt{\frac{1}{n}\sum_i (y_i - x_i)^2}}{\overline{x}}, \tag{A8}$$

where x_i and y_i are the reference and estimate respectively, $\bar{x} = \frac{1}{n}\sum_i x_i$ is the mean of the reference, and n is the number of hits. Perfect values are zero. Note that normalized mean error is sometimes also defined as "bias", but we avoid this terminology due to potential confusion with bias in detection.

We can also examine the rain rates of the hits using the multiplicative error model (Tian et al. 2013), which expresses the estimate and the reference through the relationship,

$$y_i = e^{\alpha} x_i^{\beta} e^{\varepsilon_i}, \tag{A9}$$

where α and β characterize the systematic errors and ε_i represents the bias-corrected random error with a normal distribution of mean 0 and standard deviation σ. With a logarithmic transformation, this relationship becomes

$$\log(y_i) = \alpha + \beta \log(x_i) + \varepsilon_i, \tag{A10}$$

which can be fitted using ordinary least squares. The perfect value of α is zero; the perfect value of β is one; and the perfect value of σ is zero.

One way to visualize this is via Fig. A1, which shows the effects of α and β on linear axes for x and y. α quantifies the "tilt" from the one-to-one line: with a perfect β, the deterministic part of the model becomes $y = e^{\alpha}x$, with α determining the gradient of the relationship. β characterizes the departure from linearity: with a perfect α, the deterministic part of the model becomes $y = x^{\beta}$, with β being the exponent in the power-law relationship. With a logarithmic transformation, the model becomes a straight line in log-log axes, with β being the slope and α being the intercept at x = 1. σ, on the other hand, quantifies the stochastic component in the model, representing the spread of the points from the best fit curve of $y = e^{\alpha}x^{\beta}$. As such, it can be considered as the spread of the points after removing any systematic errors.

References

Bolvin, D. T., and G. J. Huffman, 2015: Transition of 3B42/3B43 Research Product from Monthly to Climatological Calibration/Adjustment. https://pmm.nasa.gov/sites/default/files/document_files/3B42_3B43_TMPA_restart.pdf.

Chen, S., and Coauthors, 2013a: Evaluation and Uncertainty Estimation of NOAA/NSSL Next-Generation National Mosaic Quantitative Precipitation Estimation Product (Q2) over the Continental United States. *J. Hydrometeorol.*, **14 (4)**, 1308–1322, doi:10.1175/JHM-D-12-0150.1.

Chen, S., and Coauthors, 2013b: Similarity and difference of the two successive V6 and V7 TRMM multisatellite precipitation analysis performance over China. *J. Geophys. Res. Atmospheres*, **118 (23)**, 13,060–13,074, doi:10.1002/2013JD019964.

Ebert, E. E., J. E. Janowiak, and C. Kidd, 2007: Comparison of Near-Real-Time Precipitation Estimates from Satellite Observations and Numerical Models. *Bull. Am. Meteorol. Soc.*, **88 (1)**, 47–64, doi:10.1175/BAMS-88-1-47.

Falck, A. S., V. Maggioni, J. Tomasella, D. A. Vila, and F. L. Diniz, 2015: Propagation of satellite precipitation uncertainties through a distributed hydrologic model: A case study in the Tocantins–Araguaia basin in Brazil. *J. Hydrol.*, **527**, 943–957, doi:10.1016/j.jhydrol.2015.05.042.

Gebregiorgis, A., P.-E. Kirstetter, Y. Hong, N. Carr, J. J. Gourley, and Y. Zheng, 2016: Understanding Overland Multi-Sensor Satellite Precipitation Error in TRMM TMPA-RT Products. *J. Hydrometeorol.*, In Revision.

Gottschalck, J., J. Meng, M. Rodell, and P. Houser, 2005: Analysis of multiple precipitation products and preliminary assessment of their impact on global land data assimilation system land surface states. *J. Hydrometeorol.*, **6 (5)**, 573–598, doi:10.1175/JHM437.1.

Gourley, J. J., Y. Hong, Z. L. Flamig, L. Li, and J. Wang, 2010: Intercomparison of Rainfall Estimates from Radar, Satellite, Gauge, and Combinations for a Season of Record Rainfall. *J. Appl. Meteorol. Climatol.*, **49 (3)**, 437–452, doi:10.1175/2009JAMC2302.1.

Guo, H., S. Chen, A. Bao, A. Behrangi, Y. Hong, F. Ndayisaba, J. Hu, and P. M. Stepanian, 2016: Early assessment of Integrated Multi-satellite Retrievals for Global Precipitation Measurement over China. *Atmospheric Res.*, **176-177**, 121–133, doi:10.1016/j.atmosres.2016.02.020.

Habib, E., A. T. Haile, Y. Tian, and R. J. Joyce, 2012: Evaluation of the High-Resolution CMORPH Satellite Rainfall Product Using Dense Rain Gauge Observations and Radar-Based Estimates. *J. Hydrometeorol.*, **13 (6)**, 1784–1798, doi:10.1175/JHM-D-12-017.1.

Habib, E., A. Henschke, and R. F. Adler, 2009: Evaluation of TMPA satellite-based research and real-time rainfall estimates during six tropical-related heavy rainfall events over Louisiana, USA. *Atmospheric Res.*, **94 (3)**, 373–388, doi:10.1016/j.atmosres.2009.06.015.

Hong, Y., K.-L. Hsu, S. Sorooshian, and X. Gao, 2004: Precipitation estimation from remotely sensed imagery using an artificial neural network cloud classification system. *J. Appl. Meteorol.*, **43 (12)**, 1834–1853.

Hossain, F., and G. J. Huffman, 2008: Investigating Error Metrics for Satellite Rainfall Data at Hydrologically Relevant Scales. *J. Hydrometeorol.*, **9 (3)**, 563–575, doi:10.1175/2007JHM925.1.

Hou, A. Y., and Coauthors, 2014: The Global Precipitation Measurement Mission. *Bull. Am. Meteorol. Soc.*, **95 (5)**, 701–722, doi:10.1175/BAMS-D-13-00164.1.

Huffman, G. J., D. T. Bolvin, D. Braithwaite, K. Hsu, R. Joyce, C. Kidd, E. J. Nelkin, and P. Xie, 2015: Algorithm Theoretical Basis Document (ATBD) Version 4.5. NASA Global Precipitation Measurement (GPM) Integrated Multi-satellitE Retrievals for GPM (IMERG). NASA.

Huffman, G. J., and Coauthors, 2007: The TRMM Multisatellite Precipitation Analysis (TMPA): Quasi-Global, Multiyear, Combined-Sensor Precipitation Estimates at Fine Scales. *J. Hydrometeorol.*, **8 (1)**, 38–55, doi:10.1175/JHM560.1.

Joyce, R. J., 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. Hydrometeorol.*, **5 (3)**, 487–503, doi:10.1175/1525-7541(2004)005⟨0487:CAMTPG⟩2.0.CO;2.

Joyce, R. J., and P. Xie, 2011: Kalman Filter–Based CMORPH. *J. Hydrometeorol.*, **12 (6)**, 1547–1563, doi:10.1175/JHM-D-11-022.1.

Kirstetter, P.-E., J. J. Gourley, Y. Hong, J. Zhang, S. Moazamigoodarzi, C. Langston, and A. Arthur, 2015a: Probabilistic precipitation rate estimates with ground-based radar networks. *Water Resour. Res.*, **51**, 1422–1442, doi:10.1002/2014WR015672.

Kirstetter, P.-E., Y. Hong, J. J. Gourley, Q. Cao, M. Schwaller, and W. Petersen, 2014: Research Framework to Bridge from the Global Precipitation Measurement Mission Core Satellite to the Constellation Sensors Using Ground-Radar-Based National Mosaic QPE. *Geophysical Monograph Series*, V. Lakshmi, D. Alsdorf, M. Anderson, S. Biancamaria, M. Cosh, J. Entin, G. Huffman, W. Kustas, P. van Oevelen, T. Painter, J. Parajka, M. Rodell, and C. Rüdiger, Eds., John Wiley & Sons, Inc, Hoboken, NJ, 61–79.

Kirstetter, P.-E., Y. Hong, J. J. Gourley, M. Schwaller, W. Petersen, and Q. Cao, 2015b: Impact of sub-pixel rainfall variability on spaceborne precipitation estimation: Evaluating the TRMM 2A25 product: Impact of Sub-Pixel Rainfall Variability on TRMM 2A25. *Q. J. R. Meteorol. Soc.*, **141 (688)**, 953–966, doi:10.1002/qj.2416.

Kirstetter, P.-E., and Coauthors, 2012: Toward a Framework for Systematic Error Modeling of Spaceborne Precipitation Radar with NOAA/NSSL Ground Radar–Based National Mosaic QPE. *J. Hydrometeorol.*, **13 (4)**, 1285–1300, doi:10.1175/JHM-D-11-0139.1.

Kubota, T., T. Ushio, S. Shige, S. Kida, M. Kachi, and K. ’ichi Okamoto, 2009: Verification of High-Resolution Satellite-Based Rainfall Estimates around Japan Using a Gauge-Calibrated Ground-Radar Dataset. *J. Meteorol. Soc. Jpn.*, **87A**, 203–222, doi:10.2151/jmsj.87A.203.

Liu, Z., 2016: Comparison of Integrated Multisatellite Retrievals for GPM (IMERG) and TRMM Multisatellite Precipitation Analysis (TMPA) Monthly Precipitation Products: Initial Results. *J. Hydrometeorol.*, **17 (3)**, 777–790, doi:10.1175/JHM-D-15-0068.1.

Maggioni, V., M. R. P. Sapiano, R. F. Adler, Y. Tian, and G. J. Huffman, 2014: An Error Model for Uncertainty Quantification in High-Time-Resolution Precipitation Products. *J. Hydrometeorol.*, **15 (3)**, 1274–1292, doi:10.1175/JHM-D-13-0112.1.

Mei, Y., E. N. Anagnostou, E. I. Nikolopoulos, and M. Borga, 2014: Error Analysis of Satellite Precipitation Products in Mountainous Basins. *J. Hydrometeorol.*, **15 (5)**, 1778–1793, doi:10.1175/JHM-D-13-0194.1.

Roca, R., P. Chambon, I. Jobard, P.-E. Kirstetter, M. Gosset, and J. C. Bergès, 2010: Comparing Satellite and Surface Rainfall Products over West Africa at Meteorologically Relevant Scales during the AMMA Campaign Using Error Estimates. *J. Appl. Meteorol. Climatol.*, **49 (4)**, 715–731, doi:10.1175/2009JAMC2318.1.

Sarachi, S., K.-l. Hsu, and S. Sorooshian, 2015: A Statistical Model for the Uncertainty Analysis of Satellite Precipitation Products. *J. Hydrometeorol.*, **16 (5)**, 2101–2117, doi:10.1175/JHM-D-15-0028.1.

Stampoulis, D., and E. N. Anagnostou, 2012: Evaluation of Global Satellite Rainfall Products over Continental Europe. *J. Hydrometeorol.*, **13 (2)**, 588–603, doi:10.1175/JHM-D-11-086.1.

Tan, J., W. A. Petersen, and A. Tokay, 2016: A Novel Approach to Identify Sources of Errors in IMERG for GPM Ground Validation. *J. Hydrometeorol.*, **17 (9)**, 2477–2491, doi:10.1175/JHM-D-16-0079.1.

Tang, G., Y. Ma, D. Long, L. Zhong, and Y. Hong, 2016a: Evaluation of GPM Day-1 IMERG and TMPA Version-7 legacy products over Mainland China at multiple spatiotemporal scales. *J. Hydrol.*, **533**, 152–167, doi:10.1016/j.jhydrol.2015.12.008.

Tang, G., Z. Zeng, D. Long, X. Guo, B. Yong, W. Zhang, and Y. Hong, 2016b: Statistical and Hydrological Comparisons between TRMM and GPM Level-3 Products over a Midlatitude Basin: Is Day-1 IMERG a Good Successor for TMPA 3B42V7? *J. Hydrometeorol.*, **17 (1)**, 121–137, doi:10.1175/JHM-D-15-0059.1.

Tang, L., Y. Tian, F. Yan, and E. Habib, 2015: An improved procedure for the validation of satellite-based precipitation estimates. *Atmospheric Res.*, **163**, 61–73, doi:10.1016/j.atmosres.2014.12.016.

Tian, Y., G. J. Huffman, R. F. Adler, L. Tang, M. Sapiano, V. Maggioni, and H. Wu, 2013: Modeling errors in daily precipitation measurements: Additive or multiplicative? *Geophys. Res. Lett.*, **40 (10)**, 2060–2065, doi:10.1002/grl.50320.

Tian, Y., G. S. Nearing, C. D. Peters-Lidard, K. W. Harrison, and L. Tang, 2016: Performance Metrics, Error Modeling, and Uncertainty Quantification. *Mon. Weather Rev.*, **144 (2)**, 607–613, doi:10.1175/MWR-D-15-0087.1.

Tian, Y., and C. D. Peters-Lidard, 2007: Systematic anomalies over inland water bodies in satellite-based precipitation estimates. *Geophys. Res. Lett.*, **34 (14)**, L14 403, doi:10.1029/2007GL030787.

Tian, Y., C. D. Peters-Lidard, B. J. Choudhury, and M. Garcia, 2007: Multitemporal Analysis of TRMM-Based Satellite Precipitation Products for Land Data Assimilation Applications. *J. Hydrometeorol.*, **8 (6)**, 1165–1183, doi:10.1175/2007JHM859.1.

Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences*. 3rd ed., No. 100, International geophysics series, Elsevier/Acad. Press, Amsterdam.

Xue, X., Y. Hong, A. S. Limaye, J. J. Gourley, G. J. Huffman, S. I. Khan, C. Dorji, and S. Chen, 2013: Statistical and hydrological evaluation of TRMM-based Multi-satellite Precipitation Analysis over the Wangchu Basin of Bhutan: Are the latest satellite precipitation products 3B42V7 ready for use in ungauged basins? *J. Hydrol.*, **499**, 91–99, doi:10.1016/j.jhydrol.2013.06.042.

Zhang, J., Y. Qi, K. Howard, C. Langston, and B. Kaney, 2011a: Radar quality index (RQI)—A combined measure of beam blockage and VPR effects in a national network. *Proc. Eighth Int. Symp. on Weather Radar and Hydrology*, 388–393.

Zhang, J., and Coauthors, 2011b: National Mosaic and Multi-Sensor QPE (NMQ) System: Description, Results, and Future Plans. *Bull. Am. Meteorol. Soc.*, **92 (10)**, 1321–1338, doi: 10.1175/2011BAMS-D-11-00047.1.

LIST OF FIGURES

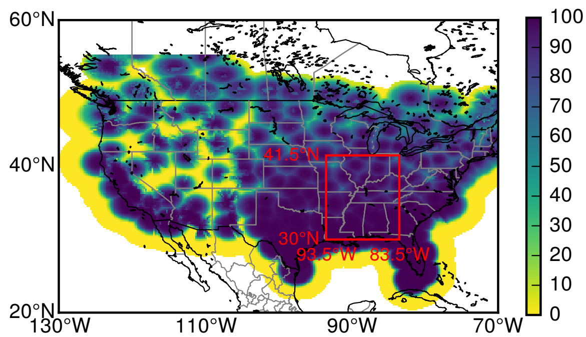


FIG. 1. A map of the average RQI for 2015. The red box shows our region of analysis: 30.0–41.5°N, 93.5–83.5°W.

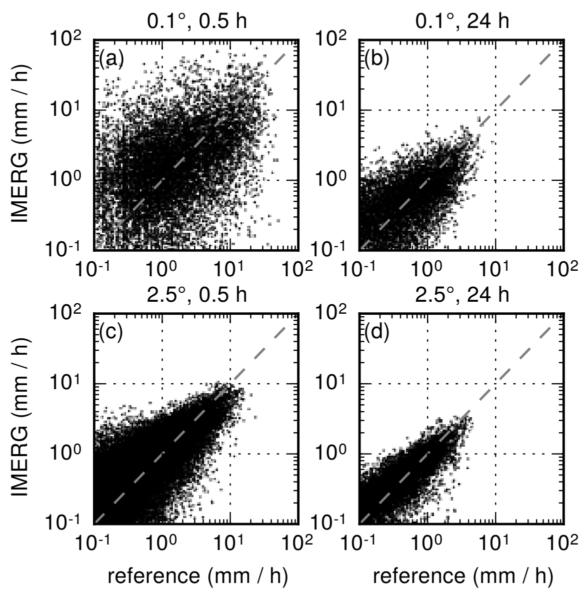
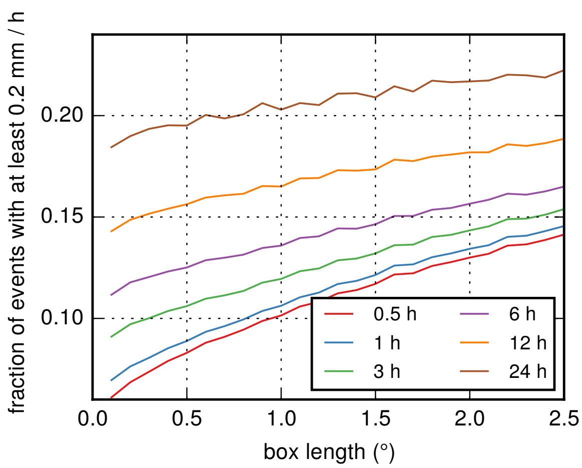


FIG. 2. A scatter diagram between IMERG and the reference at different scales: (a) $0.1^\circ \times 0.1^\circ$ grid box at 0.5 h, (b) $0.1^\circ \times 0.1^\circ$ grid box at 24 h, (c) $2.5^\circ \times 2.5^\circ$ grid box at 0.5 h, and (d) $2.5^\circ \times 2.5^\circ$ grid box at 24 h.



612 FIG. 3. Fraction of occurrences for which the reference is at least 0.2 mm / h. These fractions are obtained by
613 sampling different spatial and temporal scales a hundred times.

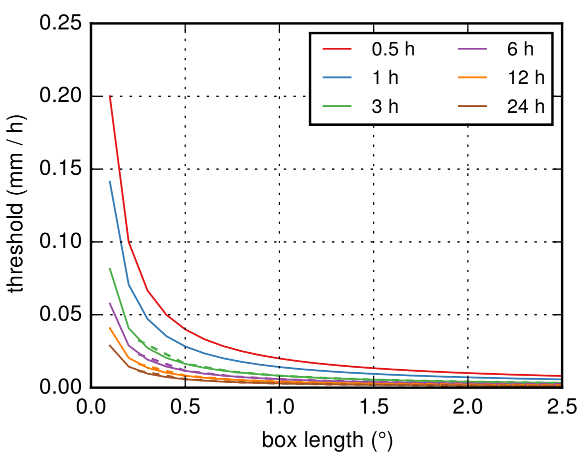
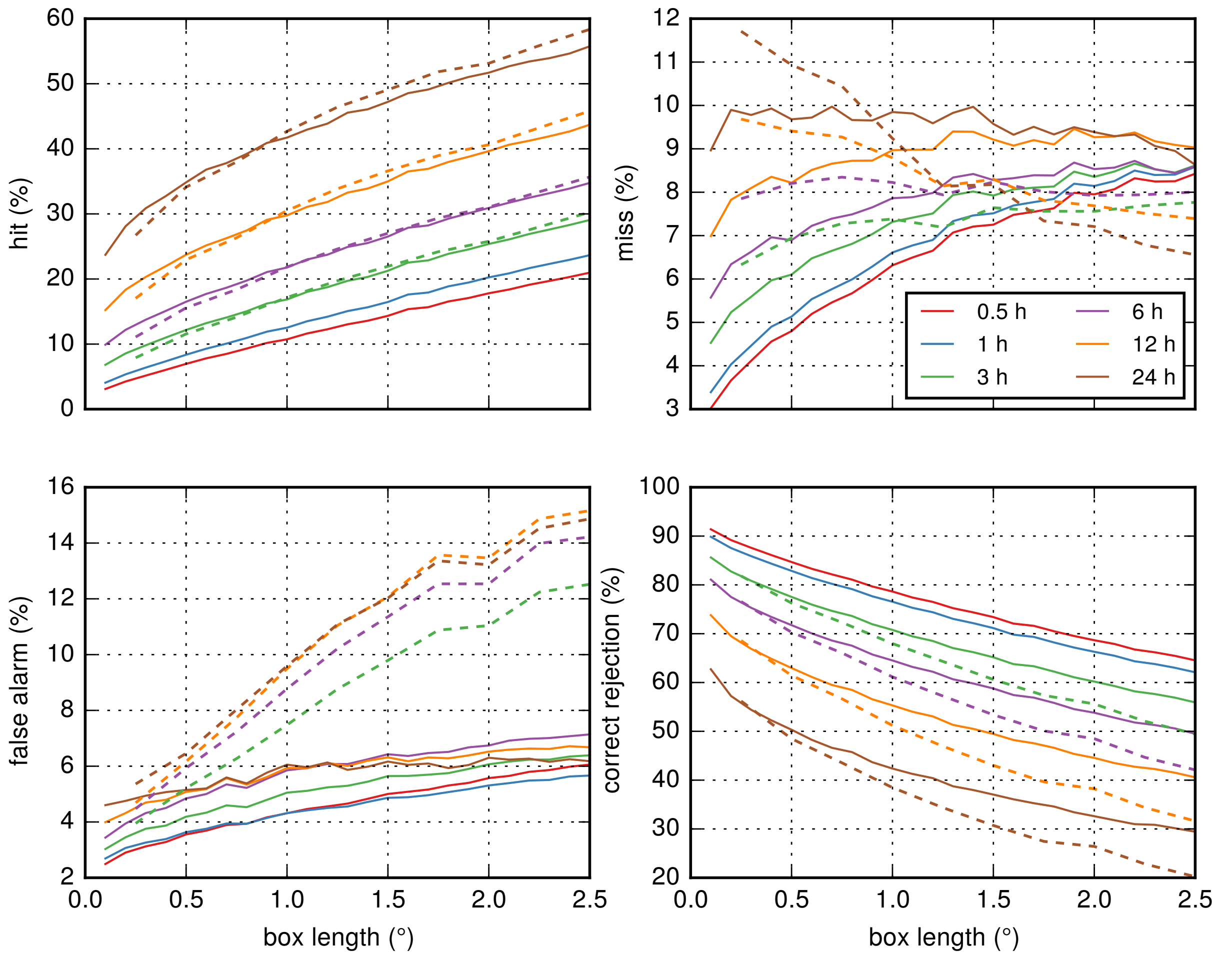
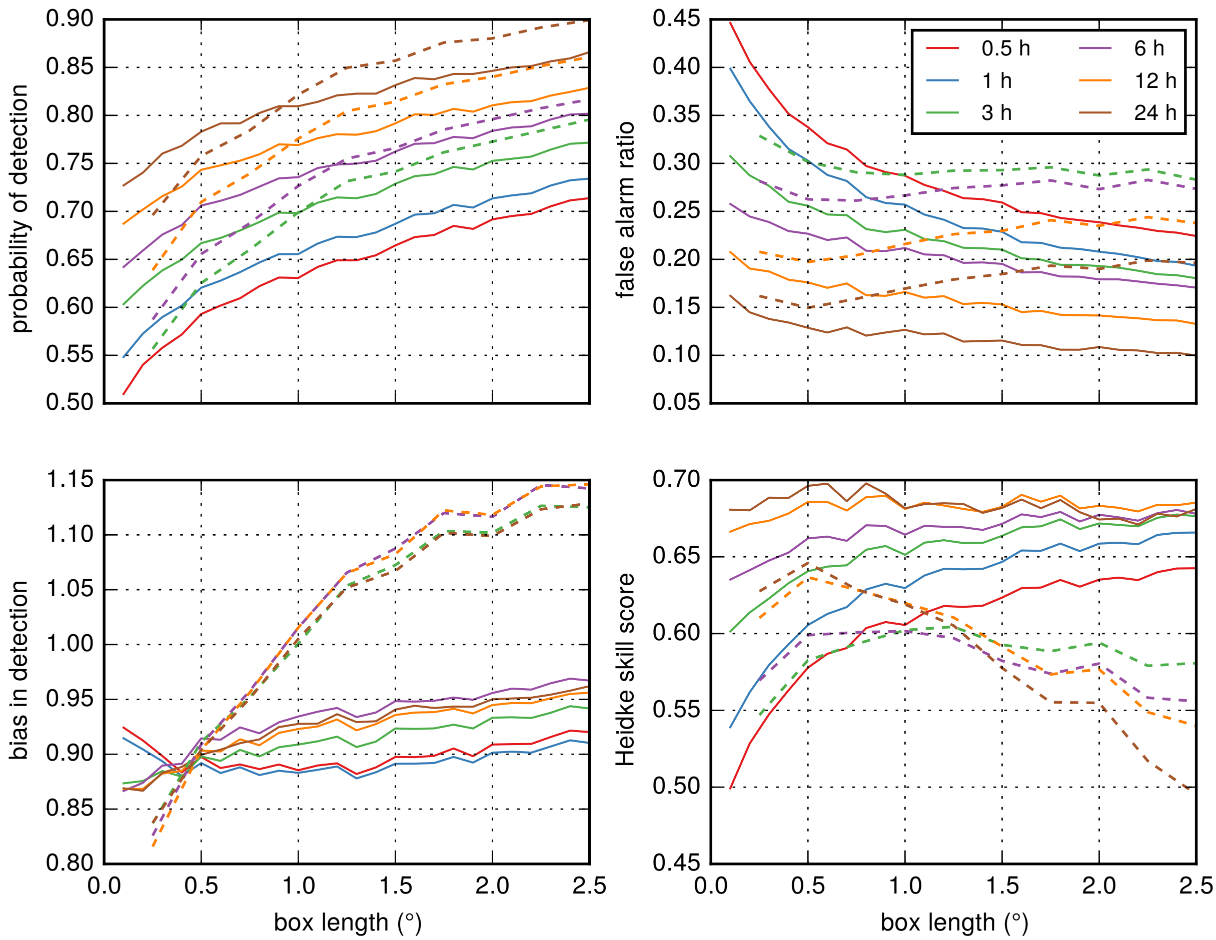


FIG. 4. Thresholds for raining events as a function of scale. Solid lines are for IMERG comparisons while dashed lines are for TMPA comparisons.



616 FIG. 5. Hits, misses, false alarms and correct rejections in IMERG (solid lines) and in TMPA (dashed lines)
617 as a function of scale.



618 FIG. 6. Probability of detection, false alarm ratio, bias in detection, and Heidke skill score of IMERG (solid
619 lines) and of TMPA (dashed lines) as a function of scale.

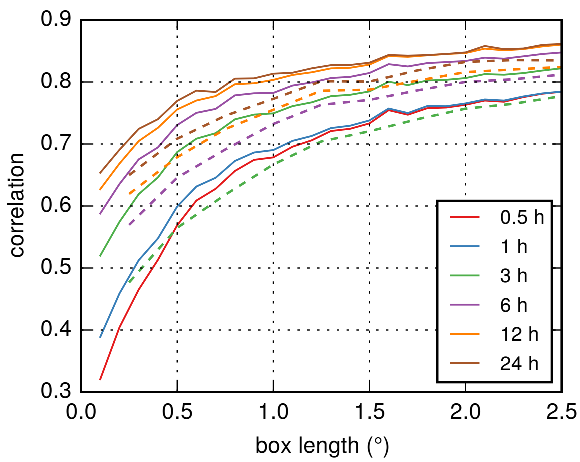
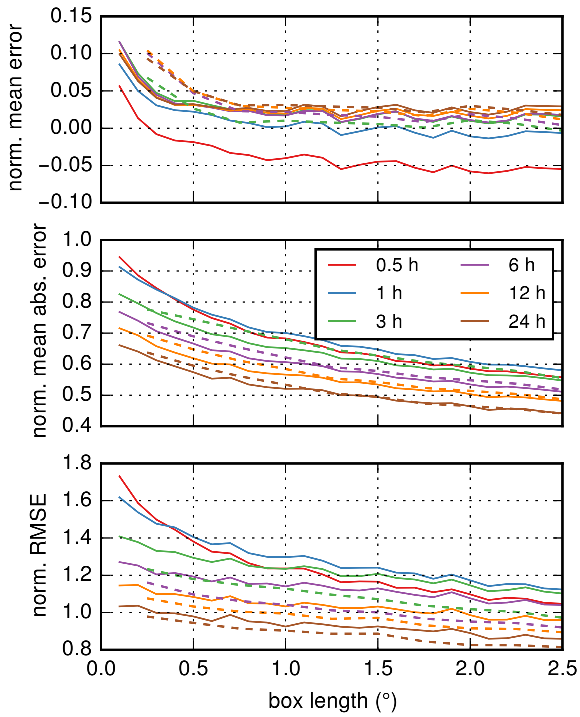


FIG. 7. Correlations of the hits between IMERG and the reference (solid lines), and TMPA and the reference (dashed lines) as a function of scale.



622 FIG. 8. Normalized mean errors, normalized mean absolute errors and normalized root-mean-square errors (RMSE) of the hits in IMERG (solid lines) and in TMPA (dashed lines) as a function of scale.

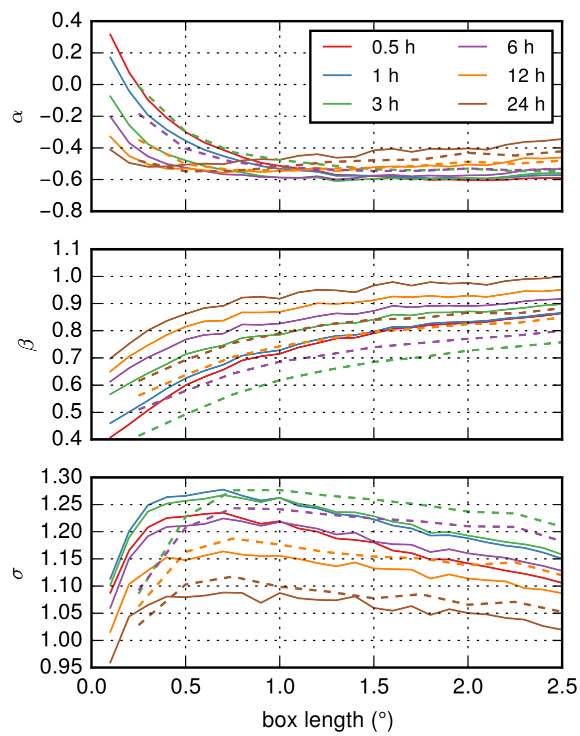
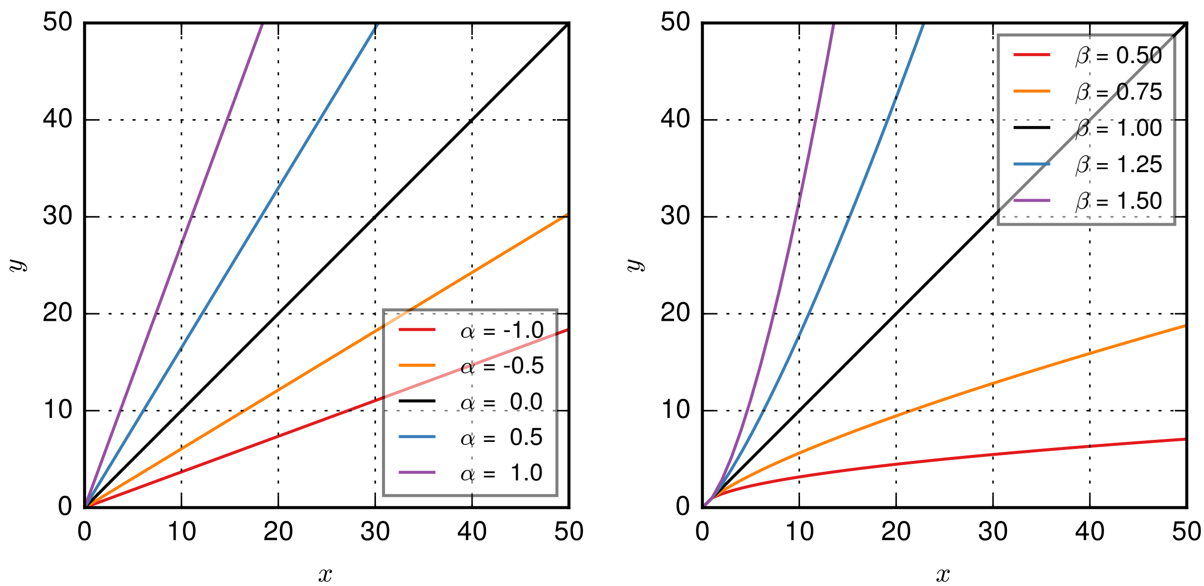


FIG. 9. Multiplicative error model parameters of the hits in IMERG (solid lines) and in TMPA (dashed lines) as a function of scale.



626 Fig. A1. The effects of α with $\beta = 1$ (left) and β with $\alpha = 0$ (right) from the multiplicative error model on a

627 linear axes.