# **Cross-evaluation of Uncertainties in Extreme Precipitation Events using Multiplicative Triple Collocation**

## ABSTRACT

## INTRODUCTION

Extreme rainfall events associated with flash floods and landslides are resulting in tremendous damages, including properties, fatalities, etc. According to (Mazzoglio, Laio et al. 2019), extreme weather condition tends to intensify and become more frequent. It is thus a great concern to face these calamities. Among them, tropical cyclone is one of the most excessive rainfall producers because it carries a considerable amount of water moisture from costal towards the inland. Hurricane Harvey is termed as category four hurricane that made landfall on Texas and impacted Louisiana, Oklahoma, with devastating urban flooding and deaths in August 2017. With data delivered by Texas state (Emanuel 2017), it produced the most abundant rainfall of hurricane in the recorded history of the United States and caused at least 70 casualties, and economic lost beyond 150 million (Emanuel 2017, Omranian, Sharif et al. 2018). Harvey brings the maximum wind of 115 to 130 mph, which damaged a significant amount of trees, fences and power poles in the Metropolitan area. More recently, tropical cyclone Imelda made landfall in Texas on 17 September, with similar impacted regions as Harvey. It produced 1096 mm total amount of rainfall in Texas, which is ranked as the fourth-highest rainfall records in the history (contributors). Damages in Texas were reported to be beyond $3 million in this disaster, while $1 billion over the whole country (News 2019). The other two tropical cyclones Bill and Cindy, influenced the same region and also bringing in copious rainfall in 2015 and 2017, respectively.

Traditionally, the rainfall rate was captured by gauges as a direct measurement but only at point sampling. Even though gauge data are often treated as the reference, it is still not impeccable as it suffers from splash-out with heavy rainfall, lack of sensitivity to light rain rates, under-catching by wind effect, and evaporations (Molini, Lanza et al. 2005, Hong and Gourley 2014). Especially in heavy rain events, (Berlamont 2001, Molini, Lanza et al. 2005) demonstrated that the error caused by these factors is not trivial. (Berlamont 2001) investigated the disadvantages of tipping bucket rain gauges during extreme weather conditions and found they underestimated rainfall volumes due to loss of water during the tipping action of the device. When merging point samplings to spatial coverage, it further introduces uncertainties coming from interpolation that may not be representative of rainfall variability depending upon the gauge quality and density. The convective nature of rainfall can also increase gauge-interpolation uncertainties for the increased standard deviation values (Stampoulis and Anagnostou 2012). Wind as an essential factor in tropical cyclones would substantially affect the performance of rain gauges, with the relative bias ranging from 5 percent to 80 percent according to the speed (Pollock, Dutton et al. 2010), but no clear consensus on the percentage of undercatching for various wind speed categories (Medlin, Kimball et al. 2007). This may be problematic when one applies gauge correction for other products e.g., radar, satellite in that event as gauge itself underestimates rainfall with up to 80 percent of bias.

More recently, with the emerging of radar technology after world war two, it has been applied in metrology to measure rain rate by emitting and receiving electromagnetic signals in the continental scale. The most prominent advantage of radar over rain gauges is that radar provides a more refined spatial-temporal scale and larger area coverage (He, Sonnenborg et al. 2013). However, since it is an indirect measurement, radar itself produces and propagates errors in the end products. Its uncertainties can be categorized as incorrect calibration, sampling representativeness, non-weather echoes, and errors in Z-R relations (Medlin, Kimball et al. 2007, Ryzhkov, Diederich et al. 2014, Kirstetter, Gourley et al. 2015). These fundamental systematic errors are challenging to mitigate for radar-only products. Thus, many researchers tend to couple radar with gauges and satellites to improve the radar performance e.g. Kriging with External Drift (KED) (Jewell and Gaussiat 2015, Cecinati, Moreno-Ródenas et al. 2018), Mean Field Bias Correction (MFB) (Yoo, Park et al. 2014). (Kidd, Bauer et al. 2011) compared ground radar QPE with gauge data in Germany, and it brings up overestimation in convective rainfall regimes. (Medlin, Kimball et al. 2007) has evaluated National Weather Service (NWS) Weather Surveillance Radar-1988 Doppler (WSR-88D) during hurricane Danny. They concluded both radar and rain gauge seriously underestimated event rainfall.

Another commonly used data source for retrieving precipitation is satellite attributing to the advantage of broad spatial coverage and the top view. Satellite data utilize the information provided by Infrared/visible channels from geostationary (GEO) satellites, passive microwave sensors, and spaceborne radar from low orbiting (LEO) satellites. With analogy to radar measurements, satellite rainfall retrieval is again an indirect measurement of rainfall. It is certain that the relationship between radiances retrieved by satellite and rain rate is less solid as compared to reflectivity by radar measurements (Scofield and Kuligowski 2003). Many dedicated works (Omranian, Sharif et al. 2018, Huang, Hu et al. 2019) stated that satellite data would underestimate the magnitude of rain rate. (Hong, Hsu et al. 2006) demonstrated satellite performance decreased with the increase of rain rate but higher normalized bias in the light rain rates. (Chen, Hong et al. 2013) compared four satellite QPEs and ground-based radar with gauge references for Typhoon Morakot, and they found satellite QPEs underestimate extreme rainfall possibly due to gauge smoother, missed precipitation outside of PMW overpasses, ice areas. (Omranian, Sharif et al. 2018) evaluated IMERG V05 final gauge adjusted products with The National Weather Service/National Centers for Environmental Prediction (NWS/NCEP) River Forecast Center (RFC) Stage-IV Quantitative Precipitation Estimates (QPEs) as the reference in hurricane Harvey event. It revealed that IMERG could detect the spatial variability of the rainfall field and overestimated rain rainfall to some degree, but it remains unclear whether NCEP stage IV radar QPE is able to be representative as the reference during Harvey event.

Given this context that no either source of a product can be trustable, researchers performed some stochastic approach to analyze these uncertainties with collocated data sources (Tian and Peters-Lidard 2010). Triple Collocation (TC) has been proven to be a powerful statistical approach to estimate uncertainties within each of three independent products (Stoffelen 1998, McColl, Vogelzang et al. 2014, Massari, Crow et al. 2017, Li, Tang et al. 2018). The concept behind is that three independent products are digested and to estimate relative error without knowing the "truth". TC was firstly applied to evaluate ocean surface wind variability by inputting different wind products (Stoffelen 1998). After that, it has been extended to measure errors of sea surface temperature (Gentemann 2014), sea surface salinity (Ratheesh, Mankad et al. 2013), wave height (Caires 2003), leaf area index and soil moisture. (Roebeling, Wolters et al. 2012) was the first to apply TC in hydrology by incorporating remote sensing, weather radar and rain gauges in Europe. (Massari, Crow et al. 2017) compared the performance of five satellite QPEs over the U.S., and deduced the results of the correlation coefficient towards the globe. (Alemohammad, McColl et al. 2015) introduced the multiplicative triple collocation method (MTC), suggesting its appropriateness in rainfall error evaluation and then decomposed the error term in order to investigate the violation of assumptions. (Li, Tang et al. 2018) used TC to perform uncertainty analysis over ungauged regions in Tibetan in China after validated TC with traditional statistics. To the best of our knowledge, TC has not been utilized to evaluate the uncertainties in extreme weather conditions in which they contain more uncertainties than usual cases. It is even challenging to validate TC without references. To investigate the validity of TC and interpret the uncertainties with three common QPEs -gauge, radar, satellite, our objectives are designed as three-folds: 1. Compare differences and performances of three independent products during multiple events; 2. Perform uncertainty estimation over multiple extreme events with particular emphasis on hurricane Harvey; 3. Evaluate the applicability of TC during extreme events.

We organize this article into four sections: section 1 will introduce the study domain and briefly review the three datasets been used in this study; section 2 will describe in detail the formula to derive Root Mean Squared Error (RMSE) and Correlation Coefficient (CC) from TC method; section 3 and 4 will follow up with performed results and conclusions from this study. The structure of this article will start from a broad overview (multiple events) and then dive into a specific event (Harvey) to scrutinize the differences.

## STUDY AREA and DATASETS

### ***Study domain***

The area we are of interest in Southern America is where it endures several events recently and historically. It is one of the most frequently impacted areas by hurricanes, tropical cyclones, experienced massive disasters like hurricane Harvey, and storms e.g., Imelda, Bill, Cindy. **Figure 1** illustrates the relative location of the impacted area, containing the states of Texas, Oklahoma, Louisiana, Arkansas, Tennessee, Mississippi, and Alabama. They account for almost 10 percent of the total areas of the United States. Storm tracks, including Harvey (2017), Bill (2015), Cindy (2017), and Imelda (2019), are also shown up in **Figure 1**. These events share similar tracks, starting from the Gulf of Mexico and bending towards North West, except for Imelda, which was dissipated in the region of Texas. **Figure 1** as well depicted the accumulative rainfall derived from MRMS (Multi-Radar Multi-Sensor) QPE, with maximum recorded as 2636 mm for four concatenated events. **Table 1** here listed out details about four collected events, including their durations, amount of rainfall falling in that period. To be comparative, we concatenated Bill, Cindy, and Imelda so that the total duration and also rainfall amount are similar to Harvey. After doing so, hurricane Harvey even observed a slightly higher amount than the other three added together, which illustrates the severity of that case.

***Datasets description***

Because of the requirement of independence, three products digested by TC method should not be coupled or integrated. Therefore, we chose National Centers for Environmental Prediction (NCEP) 4 km gridded hourly gauge only data (Lin 2011), Multi-Radar Multi-Sensors (MRMS) radar only data (Zhang, Howard et al. 2016), and Integrated Multi-satellitE Retrievals for GPM (IMERG) final uncalibrated data (Huffman, F. et al. 2019) as the triplet for evaluation.

#### NCEP gridded gauge only QPE

NCEP gridded gauge only product thereafter denoted as "NCEP" is an operational product (Lin 2011) covering the conterminous US (CONUS) and parts of Puerto Rico. It is automatically derived from approximately 3000 operational hourly rain gauge observations across 48 states to produce a 4km/hour rainfall field. The interpolation techniques behind are described by (Seo 1998) who introduced Double Optimal estimation (DO) and Single Optimal estimation (SO) to gain a conditional expectation of rainfall estimation. This technique accounts for fractional coverage of rainfall due to sparse gauge networks. (Gourley, Hong et al. 2009) performed inter-comparisons of NCEP gauge only QPE, NCEP stage IV radar QPE and PERSIANN-CSS satellite QPE. It revealed that NCEP gauge only QPE delivers the best performance within a longer time scale e.g., seasonal, daily. However, it encounters underperformance in the finer temporal scale (1 hour), especially for the storm. NCEP gridded gauge only QPE is downloadable at https://data.eol.ucar.edu/dataset/21.088.

#### MRMS radar only QPE

MRMS has around 180 integrated operational radars, including 146 S-band and 30 C-band radars, creating seamless 3D radar mosaic across the CONUS and Southern Canada at 1km/2min resolution (Zhang, Howard et al. 2016). It produces both radar-based QPE and radar gauge calibrated QPE to improve performance. It is selected because of the strict quality control, involving filtering out non-hydrometeor signals, corrections for anomalous propagation, beam blockage, VPR (Vertical Profile Reflectivity) correction, adaptive Z-R relations (Zhang, Howard et al. 2016). Despite the adoption of these quality control steps, it still suffers from uncertainties of common issues. (Gao, Zhang et al. 2018) evaluated the performance of MRMS QPE during hurricane Harvey, they envisioned MRMS QPE underestimated total accumulated rainfall by a small factor and overestimated very light precipitation. For our study, 1km/2min radar-based QPE is retrieved and processed to upscale by average and then aggregated to 4km/hourly to be compatible with NCEP gauge only QPE. Historical MRMS radar only QPE can be downloaded at http://mtarchive.geol.iastate.edu/.

#### IMERG satellite QPE

IMERG satellite precipitation product V06 (Huffman, 2019) is integrated from its core satellite (GPM Core Observatory), microwave constellations, Infrared, and additional constellations, aiming at providing global coverage of rainfall field (90N-S from V06 onward) beyond its predecessor Tropical Rainfall Measurement Mission (TRMM). GPM Core Observatory has additional channels of dual-frequency precipitation radar (DPR) and GPM Microwave Imager (GMI), which are capable of detecting very light precipitation and falling snow (Skofronick-Jackson, Petersen et al. 2017). It produces three stages: early run, late run, and final run with 4 hours latency, 12 hours latency, and three months latency respectively at a half-hour and 0.1-degree scale. The early run provides near real-time brief observations with inter-calibrated satellite products primarily for operational forecast, and the late run adds up the late coming high-quality PMW data and climatological calibration to serve for agricultural purposes. The final run compares the late run product with the Global Precipitation Climatology Project (GPCP), and adjust the factor to compensate for under/over-estimation (Huffman, F. et al. 2019). To account for independence, the final run without calibration/late run with calibration is utilized in this study. In order to perform pixel-wise analysis, IMERG data needs to be downscaled by the nearest neighbor and then accumulated to 1 hour. Current IMERG final product V06 can be accessed at https://disc.gsfc.nasa.gov/datasets/GPM\_3IMERGHH\_06/summary?keywords=IMERG.

## METHODOLOGY

### ***Assumptions***

Before applying TC, it is of great significance to bear in mind the assumptions that determine the reliability and applicability of TC method. In TC, it is assumed that 1) the three forcing data should be independent, for instance, they are derived from different instruments in our dataset; 2) The errors of three independent products should be independent or unrelated (,) which refers as zero cross-correlation. 3) The expectation of error is treated as zero known as the unbiasedness assumption, which is often described in geostatistical analysis e.g., kriging.

The three sensor products we selected gauge only, radar only, satellites only meet the above criteria well because they are not able to interact with each other. In fact, three sensors are the natural testbed for TC in rainfall uncertainty estimation.

### ***Expressions***

The basics of TC method is to treat three independent products as equally important, and thus no bias is produced in between. Since no ground truth value assumed, it then uses a linear combination of three products and affine transformed error model to derive root mean square error and correlation coefficient (Zwieback, Scipal et al. 2012).

|  |  |  |
| --- | --- | --- |
|  |  | (1) |

Where indicates each of the independent source data, is the “relative truth”, is the weights and biases to adjust, and represents error for each product.

(Tian, Huffman et al. 2013) then proposed a way to transform the additive error model to multiplicative by logarithmic transformation, and it is proved to be more appropriate in rainfall error estimation (Tian, Huffman et al. 2013, Alemohammad, McColl et al. 2015, Li, Tang et al. 2018). Hence, the error model can be reformed as:

|  |  |  |
| --- | --- | --- |
|  |  | (2) |

From that, we can derive rain rate and error model by transforming into linear combination so that it fits into TC method.

|  |  |  |
| --- | --- | --- |
|  |  | (3) |

Where demonstrating the multiplicative error, indicating the residual error, and as the deformation error. Hence, after we transform the model to be multiplicative form, these parameters along with error should be also in logarithmic form. In the analysis of (Alemohammad, McColl et al. 2015), they later transformed back error to linear scale by Taylor series expansion as first order approximation.

|  |  |  |
| --- | --- | --- |
|  |  | (4) |
|  |  | (5) |

here represents the mean of field. In doing so, the error field could be identified as linear scale, meaning the same unit of mm/hour as rain rate. From linear equation (3), we are able to derive RMSE in the following set of equations based on the covariance of triples (McColl, Vogelzang et al. 2014):

|  |  |
| --- | --- |
|  | (6) |

(McColl, Vogelzang et al. 2014) introduced a way to evaluate correlation coefficient (CC) from manipulating covariance matrices with called “ETC” method in which CC is formed shown below as a set of equations.

|  |  |
| --- | --- |
|  | (7) |

Since both RMSE and CC are derived from covariance between the three products, they infer the relative error, which is treated as uncertainties. Therefore, the less uncertain product or the best performance in the following sections is the one that has the lowest RMSE and highest CC. Likewise, the most uncertain is associated with the highest RMSE and lowest CC.

### ***Data preparation and model setup***

The rainfall data retrieved from each source are the high-level preprocessed product, but containing different extent. To make them comparable, IMERG and MRMS data are scaled to the same degree of NCEP gauge only data with the same array dimensions. Because we need to transform data into a logarithmic scale, it is of importance to filter out zero-rain and NAN data. The treatment of zero values is varying, either by merely removing zero values (Alemohammad, McColl et al. 2015, Massari, Crow et al. 2017) or replacing with near-zero values(Roebeling, Wolters et al. 2012). (Li, Tang et al. 2018) tested the sensitivity by reassigning , to non-rain samples. Removing zero values may shorten the samples especially for events based evaluation. Thus, we dropped NAN values and treat zero to be . To obtain more robust results and filter out noise, we utilized bootstrap with 500 trials for evaluation at each pixel and stored the mean value of RMSE and CC.

### ***Statistical evaluation metrics***

A list of evaluation indicators are summarized in **(Table 2)**. Due to the lack of reference data in extreme events, we will term "error" as "difference" between two products e.g., RMS difference. Quantitative difference, including Correlation Coefficient (CC), Root Mean Squared (RMS) difference, are the 1st and second-order evaluations, respectively. They are computed in a domain at each pixel (4km) for each pair. For categorical differences, we considered the Probability of Detection (POD), False Alarm Rate (FAR), Critical Success Index (CSI). The "reference" data in the denominator is chosen by the least uncertain product provided by TC. For instance, TC predicted MRMS QPE to be less uncertain, which is quantified by RMSE and CC compared to others. Then we put radar as the relative reference data for the categorical difference.

Since it is more insightful to compute these metrics in a conditional form during extreme events (Sukovich, Ralph et al. 2014), We further condition these metrics at Harvey core to validate the consistency between TC metrics and traditional skills e.g. . This condition is based on different accumulated rainfall observed for each pixel but still adopts the similar formula in **Table 2**. We consider conditions when hourly rainfall exceeds 50 percentile, 75 percentile, and 99 percentile based on the estimate itself. For instance, when evaluating the NCEP gauge performance at Harvey core, we filter out pixel time series of rain rate to be less than 50 percentile, 75 percentile, and 95 percentile of gauge rain rate at the collocated point, and use the remaining to compute conditional metrics.

### ***Harvey core separation***

In order to specifically focus on extreme rainfall, then the core, non-core regions are identified. According to the MRMS radar observation, 95 percentiles of rainfall falling outside of the core region, which corresponds to 400 mm rainfall isoline, depicted in **Figure 8**. We further used that line as a separation between the core and non-core regions. In core regions, we computed the conditional metrics described above and selected several representative pixels to plot the time series.

## RESULTS

### ***Overall behaviors***

The accumulative rainfall for all events, Harvey and non-Harvey in **Figure 2** shows that MRMS product always captures the most rainfall amount, especially with more significant portion inside Texas. IMERG satellite data measure the second largest, and the least is NCEP gauge data. Harvey causes more intensive rainfall near the boundary of Texas and Louisiana state, while other events scattered like a squall line. We can observe the lumped patches-like pattern for NCEP data, meaning the quality of gauge during extreme events is varying and will be investigated further in the Harvey events. Nevertheless, for MRMS and IMERG, they are stratified well, which shows the spatial continuity of the rainfall field.

The first-hand overview of applying TC to four events concatenated together as "All", Harvey-only as "Harvey" and non-Harvey as "other" are illustrated in **Figure 3** for CC**, Figure 4** for RMSE. Firstly, both CC and RMSE demonstrate consistent behaviour in which higher CC and lower RMSE. All three cases identified the same result that the ordered product uncertainties (from high to low) are MRMS, IMERG, NCEP according to the median value and metrics distribution at the whole domain. The median value is more insightful than the mean value because there are noises that tend to shift up mean values, especially in places where rainfall samples are limited. NCEP gauge data tends to have higher CC value in the western regions in which a smaller amount of rainfall observed compared to core regions. MRMS is behaving more uniform that higher CC is widespread over the domain. IMERG concentrates its higher value in the western region as well as NCEP, but cover more regions than NCEP. The violin plot inside each ax indicates the distribution of CC. They are more skewed to higher values (above 0.5) but still have some samples that close to zero, especially for NCEP data. These values typically appear outside of the core where we set non-rain samples to be . When the covariance of the estimated product, e.g., NCEP with either other two products, e.g., MRMS and NCEP becomes 0, it will impose the estimated CC to be zero, meaning no correlation with the other two products. These low values will be eliminated when we only consider core regions. We would like to point out that, in event Harvey, CCs of three product is more divergent compared to either All or Other. This indicates the three products agree more with each other for the non-Harvey case, and also push us to investigate Harvey with special attention.

**Figure 4** depicted the RMSE value from TC results, which reveals the other side of the CC value. In **Figure 4**, we emphasized two remarkable consistent behaviors. One is for NCEP data that "All" and "Harvey" both experience higher RMSE value in the Houston region. This needs our further scrutinization by inspecting its time series in the detailed Harvey analysis. The other for MRMS and IMERG shows anomalous signals at the boundary of Louisiana and Mississippi. The corresponding time series plot of rainfall at one location is selected and shown up in the small axes. The overall trend is well followed by three products except that MRMS radar has some anomalous impulse that disagrees with the other two data sources. This anomalous echo may not be coming from hydrometeors. It is probably caused by debris rolled up by strong wind that was recorded of 108 mph.

### ***Detailed Harvey Analysis***

**Table 3** listed both quantitative intensity differences and the categorical difference in Harvey with three different percentiles. NCEP vs. MRMS, NCEP vs. IMERG, IMERG vs. MRMS are taken into account, and the latter product is considered as "reference" because of lower uncertainty in the previous TC results. What this result conveys is with an analogy to the TC results. IMERG vs. MRMS comparison has the highest correlation coefficient for three percentiles, even though they are slightly lower than NCEP vs. IMERG in terms of RMS difference in lower percentiles. For categorical differences, all results suggest IMERG and MRMS behave relatively close while NCEP deviates from both MRMS and IMERG. It again validates the results that TC is capable of conveying some consistent results with traditional methods.

The CC and RMSE from TC are further split based on accumulative rainfall by MRMS data at each interval of 50 mm in **Figure 5.** In doing so, we could identify the range that is suitable for each product to be having less uncertainty. As an overview, MRMS data is substantially robust and stable as rainfall increases, but NCEP gauge data generally acquire more uncertainties than IMERG and MRMS at all ranges. Besides, NCEP obviously aggravates its behavior at the higher range of the rainfall i.e., above 1200 mm. IMERG data is behaving worse at lower tail i.e., below 150 mm. That indicates satellite data under-performs at lower tail probably due to the sensor sensitivity, signal attenuation, and smoothing effect of large size of the footprint. It has been reported in the literature that IMERG data typically tends to overestimate light rain and underestimate heavy rainfall (Guo, Chen et al. 2016, Sharifi, Steinacker et al. 2016, O, Foelsche et al. 2017). (Omranian, Sharif et al. 2018) also concluded that the IMERG final product generally has better performance with higher precipitation rates compared to lower rates in the case of hurricane Harvey.

**Figure 6** depicts TC metrics CC and RMSE for three grouped regions – whole, core, and non-core. After thresholding, we clearly witnessed the improved performance for both MRMS and IMERG – higher CC and lower RMSE – in terms of median value and uncertainty bound. However, the RMSE for NCEP even remained unchanged. This points out the noise removal for NCEP data inside the core is superseded by the degradation of performance, probably due to the impact of intensive wind or out-of-splash. While for the performance of each product, the results are still remained, which again validated the consistency of the TC method. The whole region's performance sits in between core and non-core because it neutralizes the two tails. Looking at the distribution of rainfall for each product in **Figure 7**, It associates with the characteristics of the product itself. For NCEP gauges, It is likely to underestimate total rainfall because of splash-out-of-water, wind under-catching, and not representative of rainfall variability. It is thus apparent for NCEP data to concentrate in the range of 400 to 600 mm. Even though IMERG data is more widespread than NCEP, it still cuts off at 1100 mm. The reasons behind are myriad e.g., the sensitivity of sensors, type of sensors (IR, PMW, DPR), but most importantly, since IMERG has a resolution of 0.1 degrees (around 10 km), it acts like a smoother that takes the average of the grid. Hence, it is challenging to capture fine-scale rainfall characteristics. To be noted, MRMS data could also suffer from overestimation since we observed some abnormal impulses during this event, but we tend to believe that MRMS radar QPE is more or less close to "reference" as TC results suggested.

By inspecting the time series at selected representative points in **Figure 8**, we can unravel the reason that NCEP data behaved relatively worse. Points 1,3, 4, and 5 are picked based on the RMSE spatial map in which NCEP data had the highest value or maximum rainfall are captured. In the corresponding time series plots, the grey window mentioned that in that specified period, NCEP data either recorded as zero value, but the other two captured intensive rainfall (point 1,3) or stopped recording any data (point 4,5). This anomaly could be caused by wind that blows raindrops out of the range or mechanical misfunctioning but more likely due to uncertainties introduced by interpolation. Because of the variability of the rainfall field, a sparse gauge network (e.g., 3000 in total around the U.S.) is not able to capture this variability. Point 2 is selected as IMERG data recorded the maximum amount of rainfall in this event. The horizontal blue line marked the maximum rain rate that GMI (GPM Microwave Imager) can record due to sensor sensitivity (Skofronick-Jackson, Petersen et al. 2017). In other words, the IMERG product will cut off any rain rate larger than 60 mm/h, except places swept by DPR (dual-frequency precipitation radar). Nevertheless, DPR has limited swath and coverage. Point 4 is where MRMS has relatively low CC value inside the core. We indeed found that MRMS sometimes jumped up instantaneously while NCEP and IMERG showed some agreements such as the highlighted red window. It probably attributes to non-weather echoes or applying incorrect Z-R relations.

**Figure 9** has depicted the conditional inter-comparison results for each pair. As expected, those metrics are getting worse when larger percentile of data exceeded because larger rain rate is associated with more difference. Except for RMS difference, all other statistics suggest IMERG and MRMS are more or less similar in whatever kinds of extreme rainfall conditions, which is again the same signature in Table 3. It once again proved that NCEP data inside Harvey core has a certain degree of degradation. Researchers need to be cautious when deciding to use this data set to evaluate extreme weather conditions, and additional justification should be provided.

## CONCLUSIONS AND FUTURE WORKS

In this experiment, we tried to explore the applicability of the TC method in extreme events, and interpret the results from it. It is an ill-posed statement because we cannot find valuable reference data to absolutely validate the performance of TC method even though a considerable amount of previous works has proven that TC is a powerful statistical tool to analyze the uncertainties of three independent precipitation products. However, instead of directly compare TC with traditional methods for given reference data, we could evaluate the consistency of TC in the following ways:

1. Applying TC in the overall extreme events combined and separately perform it with individuals. The results in **(Figure 3 and 4)** have suggested that MRMS data are always providing the lowest uncertainties and then follows IMERG satellite QPE. NCEP data are associated with the largest amount of uncertainties, which may not be an appropriate product in evaluating rainfall at extremes.
2. The traditional evaluation metrics e.g., RMS difference, Correlation Coefficient, POD, FAR, CSI, and also their conditioned values, are well in line with the TC results. They indicate that NCEP data have more massive differences with MRMS and IMERG.
3. The results to separate Harvey core and non-core (Figure 5) also demonstrated that NCEP data performance is not substantially improved inside the core while the other two get different amount of increment. This points out that NCEP data may be subject to degradation in Harvey's core.
4. The time series for collocated pixels where gauge has more uncertainties inside the core showed that a large portion of NCEP data did not record rainfall while the other two products indeed observed some amount. We believe that gauge sparsity plays the first-order role, and then it may also subject to other effects e.g., wind under-catching, splash-out-of-water, mal-functioning.
5. IMERG data has the second-largest uncertainty because of 1. Its large resolution smooths out rainfall variability so that it does not capture enough rainfall (Figure 8); 2. The sensitivity for IMERG data is not trivial in extreme conditions (Figure 7) as it cuts off rain rate beyond a certain threshold; 3. Signal attenuation may also lower the rain rate captured by satellite QPEs; 4. The gauge may also smooth out extreme values for gauge adjusted satellite QPEs.

The TC results also provide us insightful considerations. Firstly, it can assess the quality of each product. In our extreme rainfall cases, we found NCEP gauge based QPE encountered some problems that may be due to the uncertainty introduced by interpolation or systematic error e.g., wind under-catching, splash-out-of-water, etc. IMERG satellite QPE sits in between NCEP and MRMS. MRMS radar QPE is proved to be the most reliable dataset in capturing extreme rainfall. Secondly, it tells the story about limiting the boundaries of each product. NCEP gauge data increase its uncertainties as to the accumulative rainfall amount increases. One can infer that gauge-like QPE is more susceptible to a higher rainfall rate due to its deficiencies mentioned above. IMERG data have higher uncertainties at the lower tail, meaning light rainfall could cumber its performance. MRMS radar data are behaving more robust and stable at either high rain amount or low rain amount. It suggests that MRMS radar QPE is more appropriate to evaluate precipitation during extreme events.

This paper serves as the first order overview of the quality of each product during extreme events as each suffers from its deficiencies. However, it is still unclear which systematic error plays the first role for radar and satellite QPE. It motivates us to explore more detailed error decomposition in future researches. Beyond that, we believe that the gauge corrected products i.e., radar and satellite, will not get improved performance in extreme events (Huang, Hu et al. 2019). Further comparisons regarding this topic could also help developers to adopt more intelligent algorithms for quantitative precipitation estimation.

## ACKNOWLEDGEMENTS

## REFERENCE

Alemohammad, S. H., K. A. McColl, A. G. Konings, D. Entekhabi and A. Stoffelen (2015). "Characterization of precipitation product errors across the United States using multiplicative triple collocation." Hydrology and Earth System Sciences **19**(8): 3489-3503.

Berlamont, G. L. a. J. (2001). "Simplified method to correet rainfall measurements from tipping bucket rain gauges ".

Caires, S. (2003). "Validation of ocean wind and wave data using triple collocation." Journal of Geophysical Research **108**(C3).

Cecinati, F., A. Moreno-Ródenas, M. Rico-Ramirez, M.-c. ten Veldhuis and J. Langeveld (2018). "Considering Rain Gauge Uncertainty Using Kriging for Uncertain Data." Atmosphere **9**(11).

Chen, S., Y. Hong, Q. Cao, P.-E. Kirstetter, J. J. Gourley, Y. Qi, J. Zhang, K. Howard, J. Hu and J. Wang (2013). "Performance evaluation of radar and satellite rainfalls for Typhoon Morakot over Taiwan: Are remote-sensing products ready for gauge denial scenario of extreme events?" Journal of Hydrology **506**: 4-13.

contributors, W. (9 October 2019 19:58 UTC). " Tropical Storm Imelda." Retrieved 9 October 2019 20:04 UTC, 2019, from <https://en.wikipedia.org/w/index.php?title=Tropical_Storm_Imelda&oldid=920438568>.

Emanuel, K. (2017). "Assessing the present and future probability of Hurricane Harvey's rainfall." Proc Natl Acad Sci U S A **114**(48): 12681-12684.

Gao, S., J. Zhang, D. Li, H. Jiang and N. Z. Fang (2018). Evaluation of Multi-Radar Multi-Sensor (MRMS) and Stage IV Gauge-adjusted Quantitative Precipitation Estimate (QPE) During Hurricane Harvey.

Gentemann, C. L. (2014). "Three way validation of MODIS and AMSR-E sea surface

temperatures." Journal of Geophysical Research: Oceans **119**: 2583-2598.

Gourley, J. J., Y. Hong, Z. L. Flamig, L. Li and J. Wang (2009). "Intercomparison of Rainfall Estimates from Radar, Satellite, Gauge, and Combinations for a Season of Record Rainfall." Journal of Applied Meteorology and Climatology **49**(3): 437-452.

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 Research **176-177**: 121-133.

He, X., T. O. Sonnenborg, J. C. Refsgaard, F. Vejen and K. H. Jensen (2013). "Evaluation of the value of radar QPE data and rain gauge data for hydrological modeling." Water Resources Research **49**(9): 5989-6005.

Hong, Y. and J. Gourley (2014). Radar Hydrology: Principles, Models, and Applications.

Hong, Y., K.-l. Hsu, H. Moradkhani and S. Sorooshian (2006). "Uncertainty quantification of satellite precipitation estimation and Monte Carlo assessment of the error propagation into hydrologic response." Water Resources Research **42**(8).

Huang, C., J. Hu, S. Chen, A. Zhang, Z. Liang, X. Tong, L. Xiao, C. Min and Z. Zhang (2019). "How Well Can IMERG Products Capture Typhoon Extreme Precipitation Events over Southern China? ." Remote Sensing **11**(70): 1-22.

Huffman, G. J., S. E. F., B. D. T., N. E. J and T. Jackson (2019). GPM IMERG Final Precipitation L3 Half Hourly 0.1 degree x 0.1 degree V06. Greenbelt, MD, Goddard Earth Sciences Data and Information Services Center (GES DISC).

Jewell, S. A. and N. Gaussiat (2015). "An assessment of kriging-based rain-gauge-radar merging techniques." Quarterly Journal of the Royal Meteorological Society **141**(691): 2300-2313.

Kidd, C., P. Bauer, J. Turk, G. J. Huffman, R. Joyce, K. L. Hsu and D. Braithwaite (2011). "Intercomparison of High-Resolution Precipitation Products over Northwest Europe." Journal of Hydrometeorology **13**(1): 67-83.

Kirstetter, P.-E., J. J. Gourley, Y. Hong, J. Zhang, S. Moazamigoodarzi, C. Langston and A. Arthur (2015). "Probabilistic precipitation rate estimates with ground-based radar networks." Water Resources Research **51**(3): 1422-1442.

Li, C., G. Tang and Y. Hong (2018). "Cross-evaluation of ground-based, multi-satellite and reanalysis precipitation products: Applicability of the Triple Collocation method across Mainland China." Journal of Hydrology **562**: 71-83.

Lin, Y. (2011). NCEP/EMC U.S. Gridded Gage Only Hourly Precipitation. Y. Lin, UCAR/NCAR - Earth Observing Laboratory.

Massari, C., W. Crow and L. Brocca (2017). "An assessment of the performance of global rainfall estimates without ground-based observations." Hydrology and Earth System Sciences **21**(9): 4347-4361.

Mazzoglio, P., F. Laio, S. Balbo, P. Boccardo and F. Disabato (2019). "Improving an Extreme Rainfall Detection System with GPM IMERG data." Remote Sensing **11**(6).

McColl, K. A., J. Vogelzang, A. G. Konings, D. Entekhabi, M. Piles and A. Stoffelen (2014). "Extended triple collocation: Estimating errors and correlation coefficients with respect to an unknown target." Geophysical Research Letters **41**(17): 6229-6236.

Medlin, J. M., S. K. Kimball and K. G. Blackwell (2007). "Radar and Rain Gauge Analysis of the Extreme Rainfall during Hurricane Danny’s (1997) Landfall." Monthly Weather Review **135**(5): 1869-1888.

Molini, A., L. G. Lanza and P. La Barbera (2005). "The impact of tipping-bucket raingauge measurement errors on design rainfall for urban-scale applications." Hydrological Processes **19**(5): 1073-1088.

News (2019). Why can't I sign up for FEMA assistance after Imelda?

O, S., U. Foelsche, G. Kirchengast, J. Fuchsberger, J. Tan and W. A. Petersen (2017). "Evaluation of GPM IMERG Early, Late, and Final rainfall estimates using WegenerNet gauge data in southeastern Austria." Hydrology and Earth System Sciences **21**(12): 6559-6572.

Omranian, E., H. Sharif and A. Tavakoly (2018). "How Well Can Global Precipitation Measurement (GPM) Capture Hurricanes? Case Study: Hurricane Harvey." Remote Sensing **10**(7).

Pollock, M., M. Dutton, P. Quinn, E. O’Connell, M. Wilkinson and M. Colli (2010). "Accurate Rainfall Measurement: The Neglected Achiles Heel of Hydro-Meteorology."

Ratheesh, S., B. Mankad, S. Basu, R. Kumar and R. Sharma (2013). "Assessment of Satellite-Derived Sea Surface Salinity in the Indian Ocean." IEEE Geoscience and Remote Sensing Letters **10**(3): 428-431.

Roebeling, R. A., E. L. A. Wolters, J. F. Meirink and H. Leijnse (2012). "Triple Collocation of Summer Precipitation Retrievals from SEVIRI over Europe with Gridded Rain Gauge and Weather Radar Data." Journal of Hydrometeorology **13**(5): 1552-1566.

Ryzhkov, A., M. Diederich, P. Zhang and C. Simmer (2014). "Potential Utilization of Specific Attenuation for Rainfall Estimation, Mitigation of Partial Beam Blockage, and Radar Networking." Journal of Atmospheric and Oceanic Technology **31**(3): 599-619.

Scofield, R. A. and R. J. Kuligowski (2003). "Status and Outlook of Operational Satellite Precipitation Algorithms for Extreme-Precipitation Events." Weather and Forecasting **18**(6): 1037-1051.

Seo, D.-J. (1998). "Real-time estimation of rainfall fields using rain gage data under fractional coverage conditions." Journal of Hydrology **208**: 25-36.

Sharifi, E., R. Steinacker and B. Saghafian (2016). "Assessment of GPM-IMERG and Other Precipitation Products against Gauge Data under Different Topographic and Climatic Conditions in Iran: Preliminary Results." Remote Sensing **8**(2).

Skofronick-Jackson, G., W. A. Petersen, W. Berg, C. Kidd, E. F. Stocker, D. B. Kirschbaum, R. Kakar, S. A. Braun, G. J. Huffman, T. Iguchi, P. E. Kirstetter, C. Kummerow, R. Meneghini, R. Oki, W. S. Olson, Y. N. Takayabu, K. Furukawa and T. Wilheit (2017). "The Global Precipitation Measurement (Gpm) Mission for Science and Society." Bull Am Meteorol Soc **98**(8): 1679-1695.

Stampoulis, D. and E. N. Anagnostou (2012). "Evaluation of Global Satellite Rainfall Products over Continental Europe." Journal of Hydrometeorology **13**(2): 588-603.

Stoffelen, A. (1998). "Toward the true near-surface wind speed: Error modeling and calibration using triple collocation." Journal of Geophysical Research: Oceans **103**(C4): 7755-7766.

Sukovich, E. M., F. M. Ralph, F. E. Barthold, D. W. Reynolds and D. R. Novak (2014). "Extreme Quantitative Precipitation Forecast Performance at the Weather Prediction Center from 2001 to 2011." Weather and Forecasting **29**(4): 894-911.

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?" Geophysical Research Letters **40**(10): 2060-2065.

Tian, Y. and C. D. Peters-Lidard (2010). "A global map of uncertainties in satellite-based precipitation measurements." Geophysical Research Letters **37**(24): n/a-n/a.

Yoo, C., C. Park, J. Yoon and J. Kim (2014). "Interpretation of mean-field bias correction of radar rain rate using the concept of linear regression." Hydrological Processes **28**(19): 5081-5092.

Zhang, J., K. Howard, C. Langston, B. Kaney, Y. Qi, L. Tang, H. Grams, Y. Wang, S. Cocks, S. Martinaitis, A. Arthur, K. Cooper, J. Brogden and D. Kitzmiller (2016). "Multi-Radar Multi-Sensor (MRMS) Quantitative Precipitation Estimation: Initial Operating Capabilities." Bulletin of the American Meteorological Society **97**(4): 621-638.

Zwieback, S., K. Scipal, W. Dorigo and W. Wagner (2012). "Structural and statistical properties of the collocation technique for error characterization." Nonlinear Processes in Geophysics **19**(1): 69-80.