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

## ABSTRACT

## INTRODUCTION

Extreme rainfall events associated with flash floods (Smith et al., 2011; Villarini and Smith 2010), landslides (Kirschbaum et al., 2012), debris flow (Dong et al., 2010) often lead to tremendous damages, including properties, fatalities, and etc. According to Cerveny et al. (2007) and Mazzoglio et al. (2019), extreme rainfall events tend to be more severe and frequent. It is thus a great concern for local residents and weather forecasters to face these calamities (Gao et al., 2009). Among these extreme rainfall events, tropical cyclone is one of the most excessive rainfall producers because it carries a considerable amount of water moisture from oceans towards inland (Dare et al., 2012). Rainfall from tropical cyclones is the predominant cause of deaths in the United States and disruption to transportations, utilities, communications, and agricultures (Elsberry 2002; Knight and Davis 2007; Gao et al., 2009; Dare et al., 2012). Hurricane Harvey is termed as a category four hurricane that made landfall on Texas and impacted the states including Texas, Oklahoma, Louisiana, Arkansas, with devastating urban flooding and fatalities in August 2017. Based on the survey by Emanuel (2017), Hurricane Harvey had reached the heaviest rainfall record in the history of the United States and caused at least 70 casualties, and economic lost beyond 150 million (Emanuel, 2017; Omranian et al., 2018). In Fall 2019, tropical cyclone Imelda made landfall in Texas, with recorded 1096 mm of total amount of rainfall in Jefferson County, which is ranked as the fourth-highest rainfall records in the history (Wikipedia, 2019). It now represents the fifth 500-year flood that impacted a portion of Southeast Texas (NOAA, 2019). The other two tropical cyclones Bill and Cindy, influenced the Southeast Texas and also brought in copious rainfall in 2015 and 2017, respectively.

Traditionally, the rainfall rate was captured by gauges as a direct measurement but only at a point scale (Sarachi et al., 2015). Even though gauge data are often treated as the “ground truth” for rainfall measurement, they are still not impeccable due to splash-out during heavy rainfall, lack of sensitivity to light rain rates, under-catching by wind effect, and evaporations (Luyckx & Berlamont, 2001; Molini et al., 2005; Hong & Gourley, 2014; Dai et al., 2018). Especially in heavy rain events, multiple studies have demonstrated that the error caused by these factors is not trivial (Luyckx & Berlamont, 2001; Molini et al., 2005). Luyckx (2001) investigated the disadvantages of tipping bucket rain gauges during extreme weather conditions and found there was underestimation of rainfall volumes due to loss of water during the tipping action from the comparison of 24 well calibrated gauges. Molini et al. (2015) quantified the bias as an underestimation from 60% to 100% for the 1h design rainfall and return periods from 20 to 200 years. The wind from tropical cyclones would substantially affect the performance of rain gauges as well, with the relative bias ranging from 5 percent to 80 percent (Medlin et al.,2007; Pollock et al., 2010). Besides the systematic error, when interpolating point samples to be spatially representative and comparable with other data sources, errors caused by interpolation accounts for 50% to 80% of the total difference depending upon the gauge quality and density (Ciach and Krajewski 1999; Dai et al., 2018). Stampoulis and Anagnostou (2012) discovered that the convective nature of rainfall can also increase gauge-interpolation uncertainties. It may be problematic when one applies gauge correction for other products e.g., radar QPE, satellite precipitation estimation in the event.

In the past few decades, the emerging radar technology has been applied in Meteorology to measure rain rate by emitting and receiving electromagnetic signals. The most prominent advantage of radar over rain gauges is that radar provides a more refined spatiotemporal scale and a larger area coverage (He et al., 2013). However, since it is an indirect measurement of rain rate, radar technology itself produces and propagates errors in the end products. The uncertainties of radar-based precipitation estimation can be categorized as incorrect calibration, sampling representativeness, non-weather echoes, and errors in Z-R relations (Gourley et. al., 2007; Medlin et al., 2007; Zhang et. al., 2008; Ryzhkov et al., 2014; Kirstetter et al., 2015; Cao et al., 2018; Dai et al., 2018). These inherent systematic errors are challenging to mitigate by improving radar technology alone. Thus, many researchers tend to blend radar with rain gauges and satellites to improve the performance e.g., Kriging with External Drift (KED; Jewell & Gaussiat, 2015; Cecinati et al., 2018), Mean Field Bias Correction (MFB; Yoo et al., 2014). A few studies investigated the bias of radar rainfall product in excessive rainfall events (Medlin et al., 2007; Kidd et al., 2011; Chen et al., 2013; Cao et al., 2018). Kidd et al. (2011) compared ground radar Quantitative Precipitation Estimation (QPE) with gauge data in Germany, and the results indicated radar overestimation in convective rainfall regimes. Medlin et al. (2007) evaluated National Weather Service (NWS) Weather Surveillance Radar-1988 Doppler (WSR-88D) during Hurricane Danny. They concluded that both radar and rain gauge seriously underestimated rainfall. Cao et al. (2018) evaluated the performance of S-band dual-polarized radar in Hurricane Irma in which the radar showed nearly 50% of underestimation. Gao et al., (2018) evaluated the performance of Multi-Radar Multi-Sensor (MRMS) QPE during hurricane Harvey, they envisioned MRMS QPE underestimated total accumulated rainfall by a small factor and overestimated very light precipitation.

Another commonly used data source is remote sensing satellite, attributing to the advantages of broad spatial coverage and uninhibition by complex terrains (Mei et al., 2014; Sarachi et al., 2015). Satellite-based precipitation products utilize the information provided by visible-infrared (VIS-IR) channels from geostationary (GEO) satellites, passive microwave (PMW) sensors, and spaceborne radars from low orbiting (LEO) satellites. Just like weather radar QPE, satellite rainfall retrieval is also an indirect measurement of rainfall. It is certain that the relationship between the satellite radiances retrieval with rain rate is less robust compared to reflectivity by radar measurements (Scofield & Kuligowski, 2003). Many dedicated works (Mei et al., 2014; Omranian et al., 2018; Huang et al., 2019) stated that satellite data would underestimate the magnitude of rain rate. Hong et al. (2006) stated that satellite performance decreased with the increase of rain rate but higher normalized bias in the light rain. Chen et al. (2013) compared four satellite precipitation products (SPPs) and ground-based radar with gauge references for Typhoon Morakot, and they found that SPPs underestimate extreme rainfall possibly due to gauge smoother, missed precipitation outside of PMW overpasses, ice areas. Omranian et al. (2018) evaluated Integrated Multi-satellitE Retrievals for GPM (IMERG) Version 5 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. they revealed that IMERG is able to detect the spatial variability of the rainfall field but overestimated heavy rain rate. Notwithstanding, it remains unclear whether NCEP stage IV radar QPE is able to be representative as the reference.

Given this context that every data source comes with its own deficiencies, researchers performed some stochastic approaches to analyze these uncertainties (Tian & Peters-Lidard, 2010; Sarachi et al., 2015). Triple Collocation (TC) has been proven to be a powerful statistical tool to estimate uncertainties within each of three independent products (Stoffelen, 1998; Caires, 2003; Zwieback et al., 2012; Roebeling et al. , 2012; Ratheesh et al., 2013; Gentemann, 2014; McColl et al., 2014; Alemohammad et al., 2015; Massari et al., 2017; Li et al., 2018). TC was firstly applied to evaluate ocean surface wind variability by inputting different wind products (Stoffelen, 1998). After that, it has been extended to determine errors of sea surface temperature (Gentemann, 2014), sea surface salinity (Ratheesh et al., 2013), wave height (Caires, 2003). Roebeling et al. (2012) were the first to apply TC in hydrometeorology to intercompare remote sensing, weather radar and rain gauges in Europe. Massari et al. (2017) compared the performance of five satellite precipitation products over the U.S., and applied TC method to evaluate Correlation Coefficient worldwide. Alemohammad et al. (2015) introduced the multiplicative triple collocation method (MTC), suggesting its appropriateness in rainfall error estimation and proposed a way to decomposed the error term in order to investigate the violation of assumptions. Li et al. (2018) used TC method to perform uncertainty analysis for five SPPs, reanalysis data, and gridded gauge data over ungauged regions in Tibetan Plateau in China after validating TC with traditional statistics. To the best of our knowledge, we are the one of the first to apply TC method in extreme events in which they arguably contain more uncertainties than usual cases, and also evaluate its applicability. To investigate the validity of TC and interpret the uncertainties with three rainfall products −gridded gauge-only rainfall field, radar QPE and satellite precipitation estimation, we design the objectives of this study as follows: (1) To intercompare differences and uncertainties of three independent products with both traditional metrics and TC method during multiple events; (2) To evaluate the applicability of TC during extreme events.

We organize the following article into four sections: Section 2 will introduce the study domain and briefly review the three datasets been used in this study; Section 3 will describe in detail the formulas to derive Root Mean Squared Error (RMSE) and Correlation Coefficient (CC) from TC method; in Section 4, we demonstrate the results of this study from a broad overview and further dive into a specific event (Harvey) and Harvey core region to dissect the differences. Section 5 will follow up with the conclusions and the future works.

## STUDY AREA and DATASETS

### ***Study domain***

The area we are of interest is in the Gulf Coast of North America, where it endures several events recently and historically. It is one of the most frequently impacted areas by hurricanes, such as Hurricane Harvey. In other instances, tropical cyclones, Imelda, Bill, Cindy also affected this area in the recent five years. **Figure 1** illustrates the storm tracks and accumulative precipitation amount of Tropical Storm Bill (2015), Cindy (2017), and Imelda (2019) as well as Hurricane Harvey from 2017. The impacted area of the aforementioned events contains the states of Texas, Oklahoma, Louisiana, Arkansas, Tennessee, Mississippi, and Alabama, which accounts for 10 percent of the conterminous United States (CONUS). These events had similar moving patterns that they approached inland from the Gulf of Mexico and then bending towards Northwest after landing, except for Imelda which dissipated shortly after making the landfall. **Table 1** here lists the details about four individual events, including their durations, and amount of rainfall falling in that period. Due to the unprecedented nature of Hurricane Harvey, three other tropical storms were selected for analysis to improve the genericity of the study findings. All three tropical storms in total would match the duration of Hurricane Harvey, which gave the identical data sample sizes for two study groups: Harvey and non-Harvey.

***Datasets description***

NCEP gridded gauge-only hourly precipitation product (NCEP) thereafter denoted as "NCEP" is an operational product (Lin, 2011) covering the 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, using the interpolation method from (Seo 1998), which introduced Double Optimal estimation (DO) and Single Optimal estimation (SO) to gain a conditional expectation of the rainfall estimation. This technique accounts for fractional coverage of rainfall due to sparse gauge networks. Gourley et al. (2009) performed inter-comparisons of NCEP, NCEP stage IV radar QPE and PERSIANN-CSS SPP, and demonstrated that NCEP delivers the best performance at longer time scale e.g., seasonal, daily. However, it encounters underperformance at shorter time scale i.e., 1 hour, especially for the short-living storm. NCEP were obtained from National Center for Atmospheric Research/Earth Observing Laboratory (NCAR/EOL): https://data.eol.ucar.edu/dataset/21.088.

MRMS radar-only product (hereafter 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 et al., 2016). It is selected because of the strict quality control, involving filtering out non-hydrometeor signals, corrections for anomalous propagation, beam blockage, vertical profile reflectivity (VPR) correction, adaptive Z-R relations (Zhang et al., 2016). Despite the adoption of these rigorous quality control steps, it still suffers from uncertainties of common issues mentioned before. For our study, 1km/2min radar-based QPE were retrieved and aggregated by averaging to 4km/hourly, in order to be compatible with NCEP. Historical MRMS radar-only QPE was downloaded at http://mtarchive.geol.iastate.edu/.

IMERG satellite precipitation final product V06 (hereafter IMERG; Huffman, 2019) is an integrated SPP from its core satellite GPM Core Observatory (GPM CO), 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 CO 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 et al., 2017). It produces three stages: early run, late run, and final run with 4 hours latency, 12 hours latency, and 3.5 months latency respectively to accommodate different purposes. 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 et al., 2019). To account for the independence, we selected the final run without calibration in this study. In order to perform pixel-wise analysis, we further downscaled IMERG data by the nearest neighbors and then accumulated it to 1 hour. Current IMERG final run product V06 can be accessed at https://disc.gsfc.nasa.gov/datasets/GPM\_3IMERGHH\_06/summary?keywords=IMERG.

## METHODOLOGY

### ***Assumptions***

A few assumptions need to comply for TC method: (1) the three forcing data should be independent, for instance, they are derived from different instruments; (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. Yilmaz and Crow (2014) conducted experiments on the TC errors due to the relevance of three products, and the results revealed that more independent they are, the less TC induced error. It is essential to consider the relevance of the inputs in order to make TC method more reliable (Li et al., 2018). The three precipitation products we selected, i.e. gauge-only, radar-only, and satellites-only meet the above criteria well because they are not able to interact with each other.

### ***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 values are assumed, TC method then uses a linear combination of three products and affine transformed error model to derive RMSE and CC (Zwieback et al., 2012).

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

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 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 et al., 2013; Alemohammad et al., 2015; Li et al., 2018). Hence, the error model can be reformed as:

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

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

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

Where is the logarithmic form of rain rate , demonstrates the multiplicative error, indicates the residual error, and as the deformation error.

From linear equation (3), we are able to derive RMSE in the following set of equations based on the covariance of triples (McColl et al., 2014):

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

Because we transformed the model to be in additive form, these parameters along with error should be also in logarithmic form. In the analysis of Alemohammad et al. (2015), they later transformed back error to linear scale by Taylor series expansion as first order approximation.

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

here represents the RMSE in linear form, logarithmic form, and 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.

(McColl 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.

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

To clarify the TC calculated RMSE with traditional evaluation, thereafter we only use RMSE that refers to TC results and RMSD for traditional evaluation. Since both RMSE and CC are derived from covariance between the three products, they reveal 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 highly preprocessed products, but containing different spatial extents. To make them comparable, IMERG and MRMS are downscaled and aggregated to the same resolution of NCEP, with 4km spatial resolution and 1 hour temporal resolution. Because we need to transform data into a logarithmic scale, it is important to filter out zero-rain and NAN data. The treatment of zero values can be done either by simply removing zero values (Alemohammad et al., 2015; Massari et al., 2017) or replacing with near-zero values (Roebeling et al., 2012). Li et al. (2018) tested the sensitivity by reassigning , to non-rain samples. Removing zero values may reduce 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 bootstrapping with 500 trials for evaluation at each pixel and stored the mean values of RMSE and CC.

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

A list of evaluation metrics is summarized in **Table 2**. As we assumed there was no ground truth in the extreme events, we will term "error" as "difference" between two products e.g., Root Mean Squared Difference (RMSD). Continuous difference, including CC, RMSD, are the second-order evaluations. They are all computed in a domain at each pixel (4km by 4km) for each pair. For categorical differences, we consider the Probability of Detection (POD), False Alarm Rate (FAR), Critical Success Index (CSI). The "reference" data in the denominator is chosen by the less uncertain product calculated by TC method.

Since it is more insightful to compute these metrics in a conditional way during extreme events (Sukovich et al., 2014), We further condition these metrics at Harvey core to validate the consistency between unconditioned results and conditioned, i.e. . This condition is based on what the percentiles of rainfall rate the product exceeds with similar formula in **Table 2**. We consider conditions when hourly rainfall exceeds 50 percentiles, 75 percentiles, and 95 percentiles based on the estimate itself. For instance, when evaluating the NCEP performance at Harvey core, we filter out pixels with time series of rain rates to be less than 50 percentiles, 75 percentiles, and 95 percentiles, and use the remaining to compute conditional metrics.

### ***Hierarchical evaluation***

We perform the analysis in three layers: in layer one, we combined all events and compared the TC results for three products in high-level view; in layer two, we split into Harvey and non-Harvey cases for testing the stability and consistency of TC results; in the last layer, Harvey core is separated in the presence of highly impacted region by torrential rainfall. We specifically focus on Hurricane Harvey due to its high socioeconomic impacts and this choice is justified by TC results because it is more diverging

According to the MRMS, 95 percent of rainfall falls outside of the core region, which corresponds to 400 mm rainfall isoline, depicted in **Figure 8**. We delineated that line as a boundary between the core and non-core regions. The reasons behind are two-folds: (1) To investigate the error and rainfall distribution over high and low impacted areas; (2) Filter out noise due to limited rainfall samples. In core regions, we computed the conditional metrics described above and selected several representative pixels to investigate the time series. These pixels from 1 to 5 are: (1) Houston metropolitan area where NCEP is highly uncertain for both combined events and Harvey; (2) The maximum amount of rainfall observed by IMERG; (3) The maximum amount of rainfall observed by MRMS; (4) and (5) are the highly uncertain places for IMERG and MRMS.

## RESULTS

### ***Overall behaviors***

The accumulative rainfall for all events, Harvey and non-Harvey in **Figure 2** shows that The highest observed rainfall occurs in the coastline of Texas as each product depicted, which suggests that the spatial variability of rainfall field for each product is similar. However, MRMS always captures the maximum rainfall amount of 2876 mm, 1625 mm, and 1451 mm respectively for all, Harvey and Other. IMERG measures the second largest rainfall of 1749 mm, 1116 mm, and 853 mm, and the least is captured by NCEP as 1366 mm, 978 mm, and 685.5 mm. Harvey produced more concentrated rainfall especially near the boundary of Texas and Louisiana, while the other events are more scattered. We can observe the lumped patches-like pattern for NCEP and it will be investigated later in the Detailed Harvey Analysis. Nevertheless, for MRMS and IMERG, they are stratified well, which shows the spatial continuity of the rainfall field. From the rainfall distribution in the small panels, the other events are accumulated in the light rain range i.e. 0 to 200 mm, but Harvey tends to behave more uniform than the other events. It indicates Harvey impacted the whole domain more severely.

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 in space corresponds to lower RMSE. 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. All three cases identify the same result that the ordered product uncertainties (from low to high) are MRMS, IMERG, NCEP according to the median value and metrics distribution at the whole domain. Secondly, NCEP tends to have higher CC value in the western regions in which a smaller amount of rainfall observed compared to core regions. MRMS behaves more uniform that higher CC is widespread over the domain. IMERG has its higher values concentrated in the western region, similar to NCEP, but cover greater areas than NCEP. Analogously, IMERG encounters some problems in the east of the domain. The violin plot inside each panel 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. These values typically appear outside of the core where we set non-rain samples to be . When the covariance of the estimated product, i.e. NCEP with either two products becomes 0, it will cause 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 Harvey, CCs of three product are 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 motivates us to investigate Harvey with special attention. In **Figure 4**, we emphasize a remarkable consistent behaviour for NCEP. "All" and "Harvey" both experience higher RMSE value in the Houston region (in the red sphere). This needs our further dissections by inspecting its time series in the detailed Harvey analysis.

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

The separation of Harvey directs us to deal with more severe events as in Harvey, more rainfall amounts are observed and both CC and RMSE are more diverging. **Table 3** lists both continuous differences and the categorical differences in Harvey for the inter-comparison with three different percentiles (50th percentile, 75th percentile, and 95th percentile). NCEP vs. MRMS, NCEP vs. IMERG, and IMERG vs. MRMS are taken into account, and the latter product in each pair is considered as "reference" because of lower uncertainty in the previous TC results. What this result conveys is analogous to the TC results. IMERG and MRMS pair has the highest correlation coefficient for three percentiles with 0.2 higher than both NCEP vs. MRMS and NCEP vs. IMERG, even though they are slightly lower than NCEP vs. IMERG in terms of RMS difference in lower percentiles with differences only in 1/100. For categorical differences, all results suggest IMERG and MRMS behave relatively close while NCEP deviates from both MRMS and IMERG. It thus conveys two messages: Firstly, NCEP may not be able to classify rain, no-rain well because of the higher categorical differences. Secondly, it corroborates the results that TC is capable of delivering some consistent results with traditional methods.

The CC and RMSE from TC are further conditioned based on accumulative rainfall by MRMS, binned with intervals of 50 mm as shown in **Figure 5.** In doing so, we could identify the range that is suitable for each product to have less uncertainty. As an overview, MRMS is substantially more robust and stable as rainfall increases, and NCEP generally bears more uncertainties than IMERG and MRMS at all ranges. What’s more, NCEP obviously has higher uncertainties at the higher range of the rainfall i.e., above 1200 mm while IMERG data performs worse at lower tail i.e., below 150 mm. That indicates satellite data under-performs at lower tail possibly due to the sensor sensitivity, signal attenuation, and smoothing effect of large size of the footprint. Similar findings have been reported in the literature that IMERG data typically tends to overestimate light (Guo et al., 2016; Sharifi et al., 2016; O et al., 2017). Omranian 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.

### ***Harvey core***

**Figure 6** depicts TC results of CC and RMSE for three grouped regions – whole, core, and non-core. After thresholding, the uncertainty ranking are still remained the same i.e. MRMS is still the lowest, and then IMERG, NCEP, which again validates the stability of the TC method. Moreover, the results inside core regions are generally better than non-core and whole regions in terms of the median value and the uncertainty bound. Non-core has the worst results because the log transformed light rain represented in negative values and limited rainy samples further propagate to bias the TC results. The whole region sits in between core and non-core as expected because it averages out the good and bad results. Notwithstanding, the exception of RMSE for NCEP is taken because it remains almost the same in core and non-core regions. It possibly attributes to the wind under-catching and splash of water in the core regions. The result of NCEP points out the noise removal for NCEP data inside the core, similar to MRMS and IMERG is superseded by the degradation of performance for the listed possible reasons.

The distribution of rainfall for each product in **Figure 7** associates with the characteristics of the product. For NCEP, It is likely to underestimate total rainfall because of splash of water, wind under-catching, and not representative of rainfall variability. It thus possibly results in the rainfall amount for NCEP to concentrate in the range of 400 to 600 mm. Even though the spread of the distribution for IMERG is wider than NCEP, The rainfall amount 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 likely, 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 for IMERG to capture fine-scale rainfall characteristics. To be noted, MRMS may suffer from overestimation since we observed some abnormal impulses during this event. However, we tend to believe that MRMS is more or less close to "truth" as TC results suggested.

By inspecting the time series at selected representative points in **Figure 8**, we can unravel why NCEP performed the worst among the three. Points 1,3, 4, and 5 refer to the place where the RMSE values are high for NCEP. For the corresponding time series plots, the grey window mentions in the specified period, NCEP data either showed zero value (point 1, 3) or stopped recording any data (point 4, 5) but the other two data sources captured intensive rainfall. This anomaly could be caused by wind under-catching, interpolation for zero rainfall, and mechanical misfunctioning for none values. 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. These phenomena can explain TC results for NCEP is the worst because it deviates from the other two sensors the most, and it also explains the categorical difference for NCEP is the worst because it mostly recorded zero or none values while the other observed some rainfall. Point 2 is selected as IMERG data recorded the maximum amount of rainfall in this event. The horizontal blue line marks the maximum rain rate of 60 mm/hour that GMI can record due to its sensor sensitivity (Skofronick-Jackson et al., 2017). In other words, the IMERG product has certain threshold that will cut off if it exceeds due to its inherent sensor sensitivity. Point 1, 4, 5 highlight that the red windows coincide with timestamps where MRMS rises up instantaneously while NCEP and IMERG showed some agreements on lower rain rates. This radar overestimation is possibly due to non-weather echoes (Gourley et. al., 2007) and errors caused by VPR correction (Zhang et. al., 2010).

**Figure 9** depicts the conditional inter-comparison results for each pair. As expected, the RMSD is getting worse when larger percentile of data exceeded because larger rain rates are associated with more RMS differences. For categorical differences, it means the higher rain rate is more challenging to detect. Except for RMSD, all the other statistics suggest IMERG and MRMS are more or less similar in whatsoever rain rate percentiles, which is again the same information from **Table 3**. It once again proves 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 study, we utilized TC method in extreme events with three rainfall datasets, i.e., MRMS, IMERG and NCEP. Because of no trustable reference data in these events, we intercompared these products with traditional metrics. The conclusions are organized from three layers:

1. In layer one, the consistency of TC results (uncertainties ranking from high to low: MRMS, IMERG, NCEP) indicates the stability of TC method in the extreme rainfall events. we later indirectly corroborate that TC is valid from the inter-comparison results showing that NCEP indeed deviates from the other two in terms of continuous differences and categorical differences.
2. In layer two, the in-depth Harvey analysis reveals that NCEP is more susceptible to heavy rainfall (Gourley et al., 2009) i.e. beyond 1400 mm, while IMERG is more susceptible to light rain (Omranian et al., 2018), i.e., below 200mm from the exposed TC results. MRMS is the most stable and robust product among them during different range of accumulative rainfall amounts.
3. In layer three, only the improvement of RMSE for NCEP in core regions is superseded by the deficiencies itself possibly due to gauge misfunctioning, wind under-catching and splash of water in that extreme event. IMERG and MRMS get improved after filter out non-core noises. However, MRMS may encounter overestimation because of the abnormal signals and IMERG may underestimate because of the sample size and sensor sensitivity. The traditional conditioned inter-comparison results are analogous to the results from layer one, validating that NCEP is away from MRMS and IMERG in extreme conditions.

The TC method is proven to be a powerful statistical tool even in extreme rainfall events according to its own stability for different events and revealed same results with traditional evaluation metrics. From this analysis, MRMS is proven to be the best product among IMERG and NCEP to be used in extreme rainfall events.

This paper serves as the first-hand overview of the quality of each product during extreme events as each has its own deficiencies. However, it is still unclear which systematic error plays the most significant role for radar and satellite QPE. It motivates us to explore more detailed error decomposition in the future researches. Beyond that, we hypothesize that the gauge corrected products i.e. radar and satellite may not get improved performance in extreme events. Further comparisons regarding this topic could also help developers to adopt more robust algorithms considering the extreme situations for quantitative precipitation estimation.

## ACKNOWLEDGEMENTS

## REFERENCE

Alemohammad, S. H., McColl, K. A., Konings, A. G., Entekhabi, D., & Stoffelen, A. (2015). Characterization of precipitation product errors across the United States using multiplicative triple collocation. *Hydrology and Earth System Sciences, 19*(8), 3489-3503. doi:10.5194/hess-19-3489-2015

Caires, S. (2003). Validation of ocean wind and wave data using triple collocation. *Journal of Geophysical Research, 108*(C3). doi:10.1029/2002jc001491

Cao, Q., Knight M., & Qi, Y., "Dual-pol radar measurements of Hurricane Irma and comparison of radar QPE to rain gauge data," 2018 IEEE Radar Conference (RadarConf18), Oklahoma City, OK, 2018, pp. 0496-0501.doi: 10.1109/RADAR.2018.8378609

Cecinati, F., Moreno-Ródenas, A., Rico-Ramirez, M., ten Veldhuis, M.-c., & Langeveld, J. (2018). Considering Rain Gauge Uncertainty Using Kriging for Uncertain Data. *Atmosphere, 9*(11). doi:10.3390/atmos9110446

Cerveny, R.S., J. Lawrimore, R. Edwards, and C. Landsea, 2007: Extreme Weather Records. Bull. Amer. Meteor. Soc., 88, 853–860, https://doi.org/10.1175/BAMS-88-6-853

Chen, S., Hong, Y., Cao, Q., Kirstetter, P.-E., Gourley, J. J., Qi, Y., . . . Wang, J. (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. doi:10.1016/j.jhydrol.2012.12.026

Ciach, G. J., and W. F. Krajewski, 1999a: On the estimation of radar rainfall error variance. Adv. Water Resour., 22, 585–595,

Dai, Q., Q. Yang, J. Zhang, and S. Zhang, 2018: Impact of Gauge Representative Error on a Radar Rainfall Uncertainty Model. J. Appl. Meteor. Climatol., 57, 2769–2787, https://doi.org/10.1175/JAMC-D-17-0272.1

Dare, R.A., N.E. Davidson, and J.L. McBride, 2012: Tropical Cyclone Contribution to Rainfall over Australia. Mon. Wea. Rev., 140, 3606–3619, https://doi.org/10.1175/MWR-D-11-00340.1

Dong, M., L. Chen, Y. Li, and C. Lu, 2010: Rainfall Reinforcement Associated with Landfalling Tropical Cyclones. J. Atmos. Sci., 67, 3541–3558, <https://doi.org/10.1175/2010JAS3268.1>

Elsberry, R. L., 2002: Predicting hurricane landfall precipitation: Optimistic and pessimistic views from the symposium on precipitation extremes. Bull. Amer. Meteor. Soc., 83, 1333–1339.

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. doi:10.1073/pnas.1716222114

Gao, S., Zhang, J., Li, D., Jiang, H., & Fang, N. Z. (2018, 2018/12/1). *Evaluation of Multi-Radar Multi-Sensor (MRMS) and Stage IV Gauge-adjusted Quantitative Precipitation Estimate (QPE) During Hurricane Harvey*.

Gao, S., Z. Meng, F. Zhang, and L.F. Bosart, 2009: Observational Analysis of Heavy Rainfall Mechanisms Associated with Severe Tropical Storm Bilis (2006) after Its Landfall. Mon. Wea. Rev., 137, 1881–1897, https://doi.org/10.1175/2008MWR2669.1

Gentemann, C. L. (2014). Three way validation of MODIS and AMSR-E sea surface temperatures. *Journal of Geophysical Research: Oceans, 119*, 2583-2598. doi:10.1002/2013JC009716

Gourley, J.J., P. Tabary, and J. Parent du Chatelet, 2007: A Fuzzy Logic Algorithm for the Separation of Precipitating from Nonprecipitating Echoes Using Polarimetric Radar Observations. J. Atmos. Oceanic Technol., 24, 1439–1451, https://doi.org/10.1175/JTECH2035.1

Gourley, J. J., Hong, Y., Flamig, Z. L., Li, L., & Wang, J. (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. doi:10.1175/2009JAMC2302.1

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

He, X., Sonnenborg, T. O., Refsgaard, J. C., Vejen, F., & Jensen, K. H. (2013). Evaluation of the value of radar QPE data and rain gauge data for hydrological modeling. *Water Resources Research, 49*(9), 5989-6005. doi:10.1002/wrcr.20471

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

Hong, Y., Hsu, K.-l., Moradkhani, H., & Sorooshian, S. (2006). Uncertainty quantification of satellite precipitation estimation and Monte Carlo assessment of the error propagation into hydrologic response. *Water Resources Research, 42*(8). doi:10.1029/2005wr004398

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

Huffman, G.J., E.F. Stocker, D.T. Bolvin, E.J. Nelkin, Jackson Tan (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), Accessed:11/04/2019, 10.5067/GPM/IMERG/3B-HH/06Jewell, S. A., & Gaussiat, N. (2015). An assessment of kriging-based rain-gauge-radar merging techniques. *Quarterly Journal of the Royal Meteorological Society, 141*(691), 2300-2313. doi:10.1002/qj.2522

Kidd, C., Bauer, P., Turk, J., Huffman, G. J., Joyce, R., Hsu, K. L., & Braithwaite, D. (2011). Intercomparison of High-Resolution Precipitation Products over Northwest Europe. *Journal of Hydrometeorology, 13*(1), 67-83. doi:10.1175/JHM-D-11-042.1

Kirschbaum, D., R. Adler, D. Adler, C. Peters-Lidard, and G. Huffman, 2012: Global Distribution of Extreme Precipitation and High-Impact Landslides in 2010 Relative to Previous Years. J. Hydrometeor., 13, 1536–1551, https://doi.org/10.1175/JHM-D-12-02.1

Kirstetter, P.-E., Gourley, J. J., Hong, Y., Zhang, J., Moazamigoodarzi, S., Langston, C., & Arthur, A. (2015). Probabilistic precipitation rate estimates with ground-based radar networks. *Water Resources Research, 51*(3), 1422-1442. doi:10.1002/2014wr015672

Knight, D. B., and R. E. Davis, 2007: Climatology of tropical cyclone rainfall in the Southeastern United States. Phys. Geogr., 28, 126–147.

Li, C., Tang, G., & Hong, Y. (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. doi:10.1016/j.jhydrol.2018.04.039

Lin, Y. (2011). *NCEP/EMC U.S. Gridded Gage Only Hourly Precipitation*. Retrieved from: <https://data.eol.ucar.edu/dataset/21.088>

Luyckx, Gert & Berlamont, J.. (2001). Simplified Method to Correct Rainfall Measurements from Tipping Bucket Rain Gauges. 767-776. 10.1061/40583(275)72.

Massari, C., Crow, W., & Brocca, L. (2017). An assessment of the performance of global rainfall estimates without ground-based observations. *Hydrology and Earth System Sciences, 21*(9), 4347-4361. doi:10.5194/hess-21-4347-2017

Mazzoglio, P., Laio, F., Balbo, S., Boccardo, P., & Disabato, F. (2019). Improving an Extreme Rainfall Detection System with GPM IMERG data. *Remote Sensing, 11*(6). doi:10.3390/rs11060677

McColl, K. A., Vogelzang, J., Konings, A. G., Entekhabi, D., Piles, M., & Stoffelen, A. (2014). Extended triple collocation: Estimating errors and correlation coefficients with respect to an unknown target. *Geophysical Research Letters, 41*(17), 6229-6236. doi:10.1002/2014gl061322

Medlin, J. M., Kimball, S. K., & Blackwell, K. G. (2007). Radar and Rain Gauge Analysis of the Extreme Rainfall during Hurricane Danny’s (1997) Landfall. *Monthly Weather Review, 135*(5), 1869-1888. doi:10.1175/mwr3368.1

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

Molini, A., Lanza, L. G., & La Barbera, P. (2005). The impact of tipping-bucket raingauge measurement errors on design rainfall for urban-scale applications. *Hydrological Processes, 19*(5), 1073-1088. doi:10.1002/hyp.5646

NOAA National Centers for Environmental Information (NCEI) U.S. Billion-Dollar Weather and Climate Disasters (2019). https://www.ncdc.noaa.gov/billions/

O, S., Foelsche, U., Kirchengast, G., Fuchsberger, J., Tan, J., & Petersen, W. A. (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. doi:10.5194/hess-21-6559-2017

Omranian, E., Sharif, H., & Tavakoly, A. (2018). How Well Can Global Precipitation Measurement (GPM) Capture Hurricanes? Case Study: Hurricane Harvey. *Remote Sensing, 10*(7). doi:10.3390/rs10071150

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

Ratheesh, S., Mankad, B., Basu, S., Kumar, R., & Sharma, R. (2013). Assessment of Satellite-Derived Sea Surface Salinity in the Indian Ocean. *IEEE Geoscience and Remote Sensing Letters, 10*(3), 428-431. doi:10.1109/lgrs.2012.2207943

Roebeling, R. A., Wolters, E. L. A., Meirink, J. F., & Leijnse, H. (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. doi:10.1175/jhm-d-11-089.1

Ryzhkov, A., Diederich, M., Zhang, P., & Simmer, C. (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. doi:10.1175/jtech-d-13-00038.1

Sarachi, S., K. Hsu, and S. Sorooshian, 2015: A Statistical Model for the Uncertainty Analysis of Satellite Precipitation Products. J. Hydrometeor., 16, 2101–2117, https://doi.org/10.1175/JHM-D-15-0028.1

Scofield, R. A., & Kuligowski, R. J. (2003). Status and Outlook of Operational Satellite Precipitation Algorithms for Extreme-Precipitation Events. *Weather and Forecasting, 18*(6), 1037-1051. doi:10.1175/1520-0434(2003)018<1037:SAOOOS>2.0.CO;2

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., Steinacker, R., & Saghafian, B. (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). doi:10.3390/rs8020135

Skofronick-Jackson, G., Petersen, W. A., Berg, W., Kidd, C., Stocker, E. F., Kirschbaum, D. B., . . . Wilheit, T. (2017). The Global Precipitation Measurement (Gpm) Mission for Science and Society. *Bull Am Meteorol Soc, 98*(8), 1679-1695. doi:10.1175/BAMS-D-15-00306.1

Smith, J. A., G. Villarini, and M. L. Baeck, 2011: Mixture distributions and the hydroclimatology of extreme rainfall and flooding in the eastern United States. J. Hydrometeor., 12, 294–309, doi:https://doi.org/10.1175/2010JHM1242.1.

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

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. doi:10.1029/97jc03180

Sukovich, E. M., Ralph, F. M., Barthold, F. E., Reynolds, D. W., & Novak, D. R. (2014). Extreme Quantitative Precipitation Forecast Performance at the Weather Prediction Center from 2001 to 2011. *Weather and Forecasting, 29*(4), 894-911. doi:10.1175/waf-d-13-00061.1

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

Tian, Y., & Peters-Lidard, C. D. (2010). A global map of uncertainties in satellite-based precipitation measurements. *Geophysical Research Letters, 37*(24), n/a-n/a. doi:10.1029/2010gl046008

Villarini, G., and J. A. Smith, 2010: Flood peak distributions for the eastern United States. Water Resour. Res., 46, W06504, doi:https://doi.org/10.1029/2009WR008395.

Wikipedia contributors. (2019, October 9). Tropical Storm Imelda. In Wikipedia, The Free Encyclopedia. Retrieved 02:50, December 4, 2019, from <https://en.wikipedia.org/w/index.php?title=Tropical_Storm_Imelda&oldid=920438568>

Yilmaz, M.T. and W.T. Crow, 2014: Evaluation of Assumptions in Soil Moisture Triple Collocation Analysis. J. Hydrometeor., 15, 1293–1302, https://doi.org/10.1175/JHM-D-13-0158.1

Yoo, C., Park, C., Yoon, J., & Kim, J. (2014). Interpretation of mean-field bias correction of radar rain rate using the concept of linear regression. *Hydrological Processes, 28*(19), 5081-5092. doi:10.1002/hyp.9972

Zhang, J. and Y. Qi, 2010: A Real-Time Algorithm for the Correction of Brightband Effects in Radar-Derived QPE. J. Hydrometeor., 11, 1157–1171, https://doi.org/10.1175/2010JHM1201.1

Zhang, J., Howard, K., Langston, C., Kaney, B., Qi, Y., Tang, L., . . . Kitzmiller, D. (2016). Multi-Radar Multi-Sensor (MRMS) Quantitative Precipitation Estimation: Initial Operating Capabilities. *Bulletin of the American Meteorological Society, 97*(4), 621-638. doi:10.1175/bams-d-14-00174.1

Zwieback, S., Scipal, K., Dorigo, W., & Wagner, W. (2012). Structural and statistical properties of the collocation technique for error characterization. *Nonlinear Processes in Geophysics, 19*(1), 69-80. doi:10.5194/npg-19-69-2012