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

## ABSTRACT

## INTRODUCTION

Extreme rainfall events associated with flash flood and landslides are resulting in tremendous damages including properties, fatalities etc. According to (Mazzoglio, Laio, Balbo, Boccardo, & Disabato, 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 great amount of water moisture from costal towards inland. Hurricane Harvey is termed as category 4 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 largest 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, & Tavakoly, 2018). Harvey brings the maximum wind of 115 to 130 mph which damaged significant amount of trees, fences and power poles in the Metropolitan area. More recently, tropical cyclone Imelda made a 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 4th highest rainfall records in the history (contributors). Damages in Texas were reported to be beyond $3 million in this disaster while $1 billion over 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, rainfall rate was captured by gauges as a direct measurement but only at point sampling. Even though gauge data are often treated as 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 (Hong & Gourley, 2014; Molini, Lanza, & La Barbera, 2005). Especially in heavy rain events, (Berlamont, 2001; Molini et al., 2005) demonstrated 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 & 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 et al., 2010), but no clear consensus on the percentage of undercatching for various wind speed categories (Medlin, Kimball, & Blackwell, 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 the 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, Refsgaard, Vejen, & Jensen, 2013). But 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 (Kirstetter et al., 2015; Medlin et al., 2007; Ryzhkov, Diederich, Zhang, & Simmer, 2014). These intrinsic 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) (Cecinati, Moreno-Ródenas, Rico-Ramirez, ten Veldhuis, & Langeveld, 2018; Jewell & Gaussiat, 2015), Mean Field Bias Correction (MFB) (Yoo, Park, Yoon, & Kim, 2014). (Kidd et al., 2011) compared ground radar QPE with gauge data in Germany, and it brings up overestimation in convective rainfall regimes. (Medlin 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 large spatial coverage and the top view. satellite data utilize the information provided by Infrared/visible channels from geostationary (GEO) satellites, passive microwave sensor 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 & Kuligowski, 2003). Many dedicated works (Huang et al., 2019; Omranian et al., 2018) stated that satellite data will underestimate the magnitude of rain rate. (Hong, Hsu, Moradkhani, & Sorooshian, 2006) demonstrated satellite performance decreased with the increase of rain rate but higher normalized bias in the light rain rates. (Chen 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 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 reference in hurricane Harvey event. It revealed that IMERG is able to 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 product can be trustable, researchers performed some stochastic approach to analyze these uncertainties with collocated data sources (Tian & Peters-Lidard, 2010). Triple Collocation (TC) has been proven to be a powerful statistical approach to estimate uncertainties within each of three independent products (Li, Tang, & Hong, 2018; Massari, Crow, & Brocca, 2017; McColl et al., 2014; Stoffelen, 1998). 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). Thereafter, it has been extended to measure errors of sea surface temperature (Gentemann, 2014), sea surface salinity (Ratheesh, Mankad, Basu, Kumar, & Sharma, 2013), wave height (Caires, 2003), leaf area index and soil moisture. (Roebeling, Wolters, Meirink, & Leijnse, 2012) was the first to apply TC in hydrology by incorporating remote sensing, weather radar and rain gauges in Europe. (Massari et al., 2017) compared performance of five satellite QPEs over US, and deduced the results of correlation coefficient towards globe. (Alemohammad, McColl, Konings, Entekhabi, & Stoffelen, 2015) introduced 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 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 normal 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 special 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 specific event (Harvey) to scrutinize the differences.

## STUDY AREA and DATASETS

### ***Study domain***

The area we are of interest in Southern America is where endures several events recently and historically. It is one of the most frequently impacted areas by hurricanes, tropical cyclones, experienced huge 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 the similar tracks, starting from 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 Imeda so that the total duration and also rainfall amount are similar to Harvey. After doing so, hurricane Harvey even observed 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 et al., 2016), and Integrated Multi-satellitE Retrievals for GPM (IMERG) final uncalibrated data (Huffman, F., T., J, & Jackson, 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 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 conditional expectation of rainfall estimation. This technique accounts for fractional coverage of rainfall due to sparse gauge networks. (Gourley, Hong, Flamig, Li, & Wang, 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 longer time scale e.g. seasonal, daily. However, it encounters underperformance in finer temporal scale (1 hour) especially for storm. NCEP gridded gauge only QPE are 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 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 et al., 2016). Despite the adoption of these quality control steps, it still suffers from uncertainties of common issues. (Gao, Zhang, Li, Jiang, & Fang, 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 are 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 et al., 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 et al., 2017). It produces three stages: early run, late run and final run with 4 hours latency, 12 hours latency and 3 months latency respectively at a half-hour and 0.1-degree scale. Early run provides near real-time brief observations with inter-calibrated satellite products primarily for operational forecast, and late run adds up the late coming high quality PMW data and climatological calibration to serve for agricultural purposes. Final run compares late run product with Global Precipitation Climatology Project (GPCP), and adjust the factor to compensate for under/over-estimation (Huffman et al., 2019). To account for independency, 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 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, satellite only meet 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.

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

The rainfall data retrieved from each source are 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 logarithmic scale, it is of importance to filter out zero-rain and NAN data. The treatment of zero values is varying, 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 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 the 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 2nd order evaluation respectively. They are computed in a domain at each pixel (4km) for each pair. For categorical difference, we considered 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 radar product to be less uncertain which is quantified by RMSE compared to satellite. 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, Barthold, Reynolds, & Novak, 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 collocated point, and use the remaining to compute conditional metrics.

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

In order to be 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 core and non-core region. In core regions, we computed the conditional metrics described above, and selected several representative pixels to plot the time series.