Advances in Meteorology

Statistics of the performance of gridded precipitation datasets in Indonesia

Trinah Wati[1,2], Tri W. Hadi[1], Ardhasena Sopaheluwakan[2], and Lambok M. Hutasoit[1]

[1] Graduate Program of Earth Sciences, Faculty of Earth Sciences and Technology, Institut Teknologi Bandung, Bandung, 40132, Indonesia.
[2] Deputy Unit of Climatology, Indonesia Agency for Meteorology, Climatology, and Geophysics, Jakarta, 10610, Indonesia

Correspondence should be addressed to Trinah Wati; trinah.wati19@students.itb.ac.id

Abstract

Gridded precipitation datasets have been used as alternatives to rain gauge observations, but their applicability for a specific region should be thoroughly evaluated. This paper aims at finding the most appropriate one for climatological and hydrological applications in Indonesia, by evaluating the statistics of the performance of eight different datasets (research products) having horizontal resolutions between 0.1◦ and 0.25◦ and with a time span of data availability from 2003 to 2015. The datasets are compared against the observed daily rainfall at 133 stations using 13 statistical metrics that can be classified into three groups with different characteristics of measurements, namely, distribution, time sequence, and extreme value representations. By applying Summation of Rank (SR), it is found that MSWEP and TMPA 3B42 are the top two datasets that outperformed based on distribution and time-sequence performance metric groups. The extreme performances for all datasets are still good in 75th percentiles, however the performance decreasing for more than 75th percentiles indicating still poorly representation of daily extreme rainfall for all gridded datasets. Results of this study suggest that MSWEP (v2) is presently the best gridded precipitation datasets available for climatological and hydrological applications in Indonesia.

Introduction

Climate variability at sub-seasonal, seasonal, inter-annual, and inter-decadal time scales has potential societal impacts across the globe. In terms agricultural production, for example, roughly one third of the observed variations in global yield is caused climate variability [49]. Furthermore, climate change is causing extreme weather events and climate anomalies to increase in both frequency and intensity [1], leading to greater risks for natural and human systems [2]. The risks are even higher for countries like Indonesia that are prone to natural disasters. During 1900 to 2011, 56% of the disasters that killed almost 241,000 people, affected about 28 million population, and cost, around US$ 24 billion are of hydro-meteorological (climate-related) type [3]. Therefore, accurate estimation of hazards and risks due to historical and projected climate anomalies is essential to make development and business plans more climate-proof and adaptive to climate change.

High impacts hydro-meteorological disasters (drought, wildfire, flood, and landslide) in Indonesia are associated with the excess or deficit of rainfall and more prevalent than other types of climatological disasters such as heat wave [16]. While climate change analyses using the top-down approach have been facilitated by downscaling of projected future precipitation under the WCRP CORDEX (the Program Coordinated Regional Downscaling Experiment sponsored by World Climate Research Program) for the Southeast Asia region [5], the feasibility of further quantitative impact studies depends on the availability of observational data to calibrate the model output. In this context, the availability and quality of baseline climate data are crucial to carry out both bottom-up and top-down climate change studies [4] at regional scales. However, long-term continuous rainfall observations in Indonesia are only available at a very limited number of locations. Jakarta, for example, is an exceptional location where 130-year records, from 1983 to 2012, of rainfall and temperature data are available [6, 7]. Other than that, rainfall data vary in length, network density, quality, and consistency for different regions, making it difficult to assess climate hazards and risks associated with extreme events, even for regions with important socio-political contexts.

In recent decades, there have been efforts to develop globally gridded precipitation datasets by various research groups and institutions. Those datasets vary in terms of purpose, data origin, area coverage, record length, as well as spatial and temporal resolution [8]. In any case, the availability of such precipitation datasets is potentially helpful for coping with the lack of rain gauge observations. In fact, gridded datasets such as Tropical Rainfall Measuring Mission (TRMM)-based precipitation products have been extensively used in various studies with main concerns on large scale climatic features [50]. However, prior to their application to study climate impacts at regional scales, global precipitation datasets need to be evaluated to understand their advantages, limitations, and uncertainties [8, 9]. Moreover, Indonesia's archipelago constitutes the largest part of the Maritime Continent (MC) where spatial variations of rainfall climatology are prominent due to complex land-sea distribution, topography, and strong influence of Asia-Australian monsoons [42].

Intercomparisons of global precipitation datasets have been conducted for monsoon and MC regions involving reanalysis and TRMM precipitation products [44, 45, 46]. These studies used EOF analysis, correlation coefficient, and bias in comparing the datasets with observations, except [46] that focused on the relative differences among the products. There are also studies focusing on Indonesian regions and the validation of specific datasets such as TRMM [10] and GSMaP [11, 12] using rain gauge data. An intercomparison of four precipitation datasets, i.e., SA-OBS, APHRODITE, CMORPH, and TRMM, has also been conducted for performance evaluation against rain gauge data [13]. Another study focused on performance evaluation for a specific purpose to detect low rainfall for drought monitoring on three datasets, i.e., TMPA, 3B42RT, PERSIANN, and CMORPH [14], and another one for a specific region of Bali Island on three other different datasets i.e., GSMaP, IMERG, and CHIRPS [15]. These studies have compared different, but still limited number of datasets. Moreover, only a few performance metrics such as bias and correlation are used, except Liu et al. [15] who used more diverse metrics of continuous, categorical, and volumetric types. None of the studies compared the statistical distribution between precipitation datasets and observed data.

In this work, we performed a more comprehensive evaluation on eight precipitation products that are derived from rain gauges, satellite-based estimates, and their combinations (see Table 1). This study aims to find the most robust precipitation dataset for climatological and meteorological research and applications in Indonesia. We propose a multi-metric approach [17, 18] with a total of 13 metrics that can be classified into three groups of statistical measures against observations. The first group is for assessing data distribution: (1) Mean (g), (2) Standard Deviation (SC), (3) Coefficient of Variance (CV), (4) PDF Skill Scores (SS), and (5) Kolmogorov Smirnov test (KST). The second is for evaluating the relationships in sequential data pairs as time series: (6) Pearson correlation coefficient (r), (7) mean error (ME), (8) Root mean square error (RM), (9) Relative Change (RC), (10) T-test (TT), and (11) Z-test (ZT). The third is to address the performance on extreme events detection: (12) Fraction Skill Score (FSS), and (13) Anderson-Darling test for the 75th, 90th, and 98th. For overall performance, we apply a summation of rank (SR) to all metrics used in all groups of scores and select the top one dataset. As a reference, we employed rain gauge data of 133 meteorological stations belonging to the Indonesian Meteorological, Climatological, and Geophysical Agency (BMKG) observed from 2003 to 2015 (see Figure 1).

Materials and Methods

Gridded Precipitation Datasets

The summary of eight primary datasets used in this study is presented in Table 1. These datasets are research products with high-latency data transfer (in the order of several months) and generally derived from combination of rain gauge, satellite, and reanalysis data. It should be noted that all datasets have daily temporal resolution, but the spatial resolutions are 0.25° for five (CHIRPS, CMORPH-CDR, GFD, PERSIANN-CDR, TMPA) and 0.1° for the other three (GSMaP RNL, GPM-IMERG, MSWEP) datasets. Comparisons between gridded and observed station precipitations were performed for the station locations by interpolating the gridded data, using the "nearest neighbor" method. In addition to the eight primary datasets, we also analyzed other five gridded datasets that have different specifications (see the Discussion section).

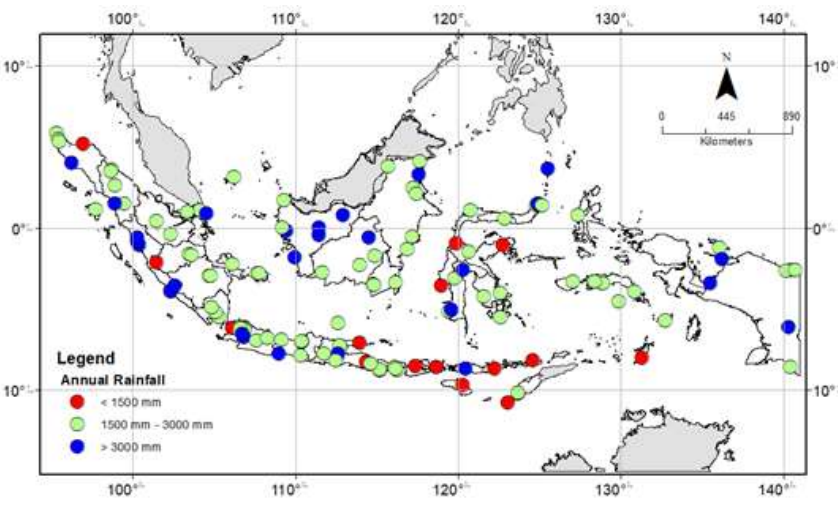

Figure 1: The location of station sites with observed mean annual rainfall distribution in the period of 2003-2015 with red (<1500 mm/yr), green (1500 – 3000 mm/yr), and blue (>3000 mm/yr).

Table 1: Summary of the eight gridded precipitation products evaluated in this study

Datasets [reference]	Input data & sensors	Periods	Spatial Resolution	Data Repository
Climate Hazards Group Infrared Precipitation Stations (CHIRPS) v2.0 [21]	CHPClim, IR NOAA: CPC, IR & NCDC B1 IR, TMPA 3B42, CFSv2, and rain gauge from NHMs, regional, GHCN, GSOD, GTS, SASSCAL	1981-present	0.25°	https://www.chc.ucsb.edu/data/chirps/
Climate Prediction Center Morphing Technique Climate Data Record (CMORPH-CDR) v1.0 [22]	DMSP 13, 14 & 15 (SSM/I), NOAA-15, 16, 17 & 18 (AMSU-B), AMSR-E and TMI from AQUA and TRMM NASA	1998-present	0.25°	https://www.ncei.noaa.gov/data/cmorph-high-resolution-global-precipitation-estimates/access/daily/0.25deg/
Princeton Global Meteorological Forcing Dataset (GFD)v3 [23]	NCEP-NCAR, CRU, daily GPCP, TRMM and NASA Langley Surface Radiation Budget	1948-2016	0.25°	http://hydrology.princeton.edu/data.pgf.php
Precipitation Estimation from Remotely Sensed Information Using Artificial Neural Networks-Climate Data Record (PERSIANN-CDR) v01 r01 [24]	Hourly NCEP stage IV from NEXRADs radar and rain gauges, GEO B1 global ISCCP (GEOS, Meteosat, GMS, FY2), GPCP v 2.2, GridSat-B1 IRWIN	1983-present	0.25°	https://www.ncei.noaa.gov/data/precipitation-persiann/access/
Tropical Rainfall Measuring Mission Multi-Satellite Precipitation Analysis (TMPA 3B42v7) [25, 26]	TMI-TRMM, DMSP-SSM/I, AQUA-AMSR-E, NOAA-AMSU-B, GEO-IR, LEO–GPI, TCI-TRMM, TRMM-PR	1998-2019	0.25°	https://disc.gsfc.nasa.gov/datasets?keywords=TRMM&page=1
Global Satellite Mapping of Precipitation Reanalysis Product (GSMaP_RNL) V06 [27, 28]	FCST, TRMM, AQUA, DMSP F13-F17, NOAA N15-N18, GMS, METEOSAT, Himawari, TMI, GMI, TRMM/PR and GPM/DPR	2000-present	0.1°	https://sharaku.eorc.jaxa.jp/GSMaP_crest/html/data.html
Global Precipitation Measurement-Integrated Multi-Satellite Retrieval for GPM (GPM-IMERG) V06 (Final Run) [29, 30]	DMSP 13, 14 & 15 (SSM/I), NOAA-15, 16, 17 & 18 (AMSU-B), AMSR-E and TMI from AQUA and TRMM NASA, TRMM PR-TMI, GPM DPR-GMI, AQUA AMSR-E, DMSP SSMI F13-F15, SSMIS F16-F19, GCOMW1-AMSR2, NOAA-AMSU 15-17, MHS 18-19, ATMS 20, METOP 1, 2, M-T SAPHIR, GEOS FP, GMS, MTSAT, HIMAWARI, Meteosat, GPCP	2000-present	0.1°	https://pmm.gsfc.nasa.gov/data-access/downloads/gpm
Multi-Source Weighted-Ensemble Precipitation (MSWEP) v2 [31]	CMORPH, ERA-Interim, GridSat, GSMaP, JRA-55, and TMPA 3B42RT, GPCC FDR, rain gauge stations GHCN, GSOD and WorldClim V.2	1979-present	0.1°	http://www.gloh2o.org

Observation Datasets

The observation datasets used as a reference employs rain gauge data from 133 meteorological stations in Indonesia from 2003 to 2015 (Figure 1). These periods overlap the years between observation and all precipitation datasets that were being compared. The rainfall data are the same observed daily precipitation dataset that was used in Supari et al. [7] up to 2012, with additional stations and time periods. The same quality control analysis described in Supari et al. [7] was applied, consisting of checking for gross errors, missing values, outliers, and overall data homogeneity. The accumulation of daily precipitation measured at 07.00 local time assigned the date of that day's precipitation data regarding the guidelines of the Indonesia Meteorological Service (BMKG) [32]. The intercomparison between precipitation datasets and observation rain gauge in this study was carried out without a day shift; please refer to Van den Besselaar et al. [13].

Metrics of Distribution-Based Performance

Mean (g)

The mean (or average) is the measure of the central tendency for both discrete and continuous data. Given μ_o: time mean, σ_o: standard deviation of the observation, and μ_m: time mean of a precipitation dataset with the same period, we define the performance metric

$$g = 1 - \frac{1}{n_g}\frac{|\mu_m - \mu_o|}{\sigma_o} \qquad (1)$$

where *ng* is a scale factor taken as equal to 1 [17, 18]. The maximum value of g = 1, and g <0 if the difference between the time-mean of rainfall dataset and observation is greater than *ng* multiplied by σ_o. The performance indices for this metric were calculated for daily (g_d), monthly (g_m), seasonal (g_DJF, g_MAM, g_JJA, and g_SON) and annual (g_a) time scales. Herein, DJF, MAM, JJA, and SON are the months of December-January-February, March-April-May, and September-October-November respectively.

Standard Deviation (SC)

Standard deviation measures the spread of data distribution. The metric SC is a normalized quantity to represent the comparison between the spreads of evaluated datasets and the referent, given by

$$SC = 1 - \frac{|\sigma_m - \sigma_o|}{\sigma_o} \qquad (2)$$

where σ_o and σ_m are the standard deviations observation of gridded precipitation datasets respectively, so that *SC* = 1 is being the perfect skill [17, 18]. As with g, the performance indices for SC were calculated for daily (SC_d), monthly (SC_m), seasonal (SC_DJF, SC_MAM, SC_JJA, and SC_SON), and annual (SC_a) time scales.

Coefficient of Variance (CV)

The metric CV is a normalized measure of dispersion. Given CV_o and CV_m are the coefficients of variation for the observation and precipitation dataset [17, 18], the performance index is calculated as

$$CV = 1 - \frac{|CV_m - CV_o|}{CV_o} \qquad (3)$$

for daily (CV_d), monthly (CV_m), seasonal (CV_DJF, CV_MAM, CV_JJA, CV_SON) and annual (CV_a) time scales.

PDF Skill Score (SS)

The Probability Density Function (PDF) skill score (SS) is calculated using samples of rain days (days with precipitation > 0.5 mm) [32]. The SS compares the PDF observed and gridded precipitation datasets by the formula [34]

$$SS = \sum_{k=1}^{nb} \min(f_m^k, f_o^k) \qquad (4)$$

with f_m^k and f_o^k are relative frequency of occurrence of a value in the k^{th} bin belonging to the histograms of the dataset and observation, whereas Nb is the number of bins used to calculate the empirical PDF. If SS = 1, the precipitation dataset perfectly simulates the observed PDF [18, 34].

Kolmogorov Smirnov Test (KST)

The KST test is like SS but for ECDF (empirical cumulative distribution function). Given $F_o(x)$ and $F_m(x)$ are ECDFs of the observed data and a precipitation dataset, the KST performance index is calculated by

$$D_{KS} = maks|F_o(x) - F_m(x)| \qquad (5)$$

$$KST = 1 - D_{KS} \qquad (6)$$

with D_{KS} is the maximum absolute difference between ECDF of two different datasets. The metric values are normalized and bounded by one as the perfect skill [17, 18].

Metrics of Time Sequence

Normalized Mean Error (ME)

Normalized mean error (ME) is to measure value-to-value differences between two time series calculated as

$$ME = 1 - \frac{\sum_{i=1}^{N} |m_i - o_i|}{\sum_{i=1}^{N} o_i} \qquad (7)$$

with m_i and O_i are the i^{th} data of the time series of the gridded dataset and observation, whereas N is the number of data records [47]. Calculations of ME are at daily (ME_d), monthly (ME_m), annual (ME_a), and seasonal (ME_DJF, ME_MAM, ME_JJA, and ME_SON) time sequences with ME = 1 means the perfect skill.

Normalized Root Mean Square Error (RM)

Root mean square error, or RMSE, is a common measure to quantify the difference between two time series. For two time series p(t) and f(t) with n data records, it can be calculated as

$$RMSE = \frac{\left[\sum_{i=1}^{n}(p_i - f_i)^2\right]^{1/2}}{n} \qquad (8)$$

Herein, we use normalized of RMSE [18] which is expressed by

$$RM = 1 - \frac{1}{n_g}\frac{RMSE_i}{{\sigma_o}^i} \qquad (9)$$

where ng and ${\sigma_o}^i$ are the scale factor and observation standard deviation relevant to the index *i* of interest related to daily (RM_d), monthly (RM_m), annual (RM_a), and seasonal (RM_DJF, RM_MAM, RM_JJA, and RM_SON) time sequences.

Relative Change (RC)

The RC metric is only applied for annual rainfall P by calculating changes in two consecutive years as

$$C^i = \frac{P_{i+1} - P_i}{P_i} \qquad (10)$$

RC index for the whole data period (years) is then calculated as the mean difference of C^i for the observation and precipitation datasets using Eq. (1) [17,18].

Pearson Correlation Coefficient (r)

The Pearson Correlation Coefficient (r) of the sequential time series for every point of observation and the corresponding grid cell of a precipitation dataset is computed using the following equation:

$$r = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{\frac{1}{2}}\left[\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2\right]^{\frac{1}{2}}} \qquad (11)$$

where x and y are the variables of observation and the precipitation dataset with the mean $\bar{x}$, and $\bar{y}$, n is the degree of freedom of the variables [18]. Calculations of r are at daily (r_d), monthly (r_m), annual (r_a), and seasonal time scales (r_DJF, r_MAM, r_JJA, and r_SON).

Z test (ZT)

Z test compares the significant difference between the mean values of observation and a precipitation dataset taking into account the difference in sample size.

$$Z = \frac{\overline{x_m} - \overline{x_o}}{\left(\frac{\sigma_m^2}{n_m} + \frac{\sigma_n^2}{n_o}\right)^{1/2}} \quad (12)$$

where $\overline{x_m}$, σ_m, n_m are the mean, standard deviation, and sampling size of the dataset, respectively; $\overline{x_o}$, σ_o, n_o are the mean, standard deviation, and sampling size of observation, respectively. The test p-value of this statistic is approximated using the standard Gaussian distribution at 95%. P-value < 0.05 means the average of the precipitation dataset is significantly different with observation. The score of ZT is the number of stations with an insignificant p-value (≥0.05) divided by the total number of stations [18]. ZT calculations are at daily (ZT_d), monthly (ZT_m), Seasonal (ZT_DJF, ZT_MAM, ZT_JJA, and ZT_SON) and annual (ZT_a) time scales.

T test (TT)

The metric TT is computed the same way as the Z-test but using Student's-t distribution [18].

Metrics of Extreme Value Representation

To evaluate performance of precipitation dataset for extreme value representation, two metrics are used: fraction skill score (FSS), and the Anderson-Darling Tests (ADT). The FSS uses the forecast verification approach to evaluate the detectability of moderate to heavy rainfall events [35], where rainfall data that exceed a threshold transform into a binary number of one, otherwise it is zero following the formula

$$FSS = 1 - \frac{FBS}{FBS_w} \quad (13)$$

FBS represents the differences of mean squares between the referent (O_i), and the precipitation dataset (F_i) on each grid computed as

$$FBS = 1 - \frac{1}{N}\sum_{i=1}^{N}(O_i - F_i)^2 \quad (14)$$

whit N is the amount of data, while FBS_w is the largest FBS that can be obtained. FSS ranges between 0 (no skill) and 1 (perfect skill). Although FSS can be calculated with absolute thresholds, in this study it is defined relatively to the 75th (p75), 90th (p90) [36], and 98th (p98) percentiles of each dataset.

A modified version of Anderson-Darling [37, 38] was applied to test the differences between the precipitation dataset distribution with reference data, using the formula:

$$A = \frac{mn}{N}\int_{+\infty}^{-\infty}\frac{(G_m - F_n)^2}{H_N}\,dH_N \quad (15)$$

$$H_N = \frac{nF_n + mG_m}{N} \quad (16)$$

A represents the distribution of daily precipitation from the dataset that reproduces the distribution of daily observations concerning moderate-to-heavy rainfall (using the same thresholds as FSS). Smaller values of A indicate a similarity between the two distributions of daily precipitation at a 95% significance level. Let X and Y be an n-and m-sample with the empirical Cumulative Distribution Function (CDF) of F and G. H denotes a measure determined by the weighted average of F and G, N = n + m. The rainfall distribution dataset is significantly different from the observed rainfall when p values <0.05. The score values were normalized value of A and bounded by one as the perfect skill (ADT).

The Summation of Rank

The metrics were applied based on point-to-grid comparisons since the observation data as referents were at point locations, which may affect the representativeness of the values being compared. However, Tan et al. [48] pointed out that results of point-to-grid comparison are substantially similar to the grid-to-grid comparison. In this work, we use two scoring schemes for indexing the precipitation dataset performance with the distribution and time sequence metrics: the "ratio" scheme, which summates the station index of more than 0.5 divided by the total number of stations; and the "mean" scheme which averages the entire index from all stations. These indices and scores measure statistical performance at daily, monthly, seasonally, and annual time scales. However, the performance scores for extreme value representation were only calculated with the "mean" scheme. The Summation of Rank (SR) [17, 18, 33] method is used to summarize and quantify the total score of all metrics.

Results

Distribution-Based Performance

Scores for five performance metrics in the data distribution group: g, SC, CV, SS, and KST, are presented as heatmaps in Figure 2. It should be noted that the scores are calculated as an aggregate of all validated points in Figure 1, with two representation schemes, i.e., "ratio" (Figure 2a) and "mean" (Figure 2b) as previously explained. The number of samples for CMORPH-CDR, GPM-IMERG-F, GSMaP_RNL, GFD, and TMPA 3B42 datasets are 133 or sampled at the 133 validating points. However, PERSIANN-CDR, MSWEP, and CHIRPS had slightly fewer samples (128, 131, and 131 respectively) because some grids do not enclose observational data.

It can be seen from Figure 2a that, based on the ratio scoring (summation) scheme, MSWEP has the highest scores followed by TMPA 3B42, and GPM-IMERG-F. The MSWEP also has the highest scores with the mean scoring scheme, followed by TMPA 3B42 and CMORPH-CDR (Figure 2b). It is of interest to note that the metric g shows relatively low scores for seasonal and annual, in comparison to daily and monthly, time scales. Seasonally, the g scores are worse for DJF and MAM than those for JJA and SON. At most stations, the JJA and DJF periods correspond to dry and rainy season respectively. The spatial distribution of SR distribution-based performance at all 133 stations in Indonesia can be seen on Figure S1 in the Supplementary Material (SM).

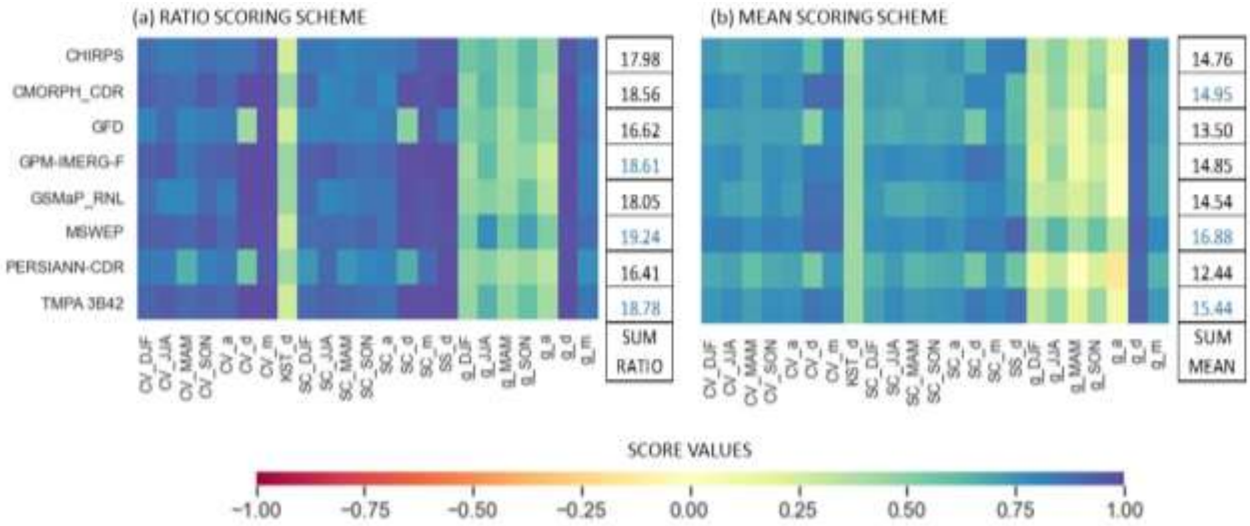

Figure 2: Heatmaps of distribution performance metrics by ratio (a), and mean (b) with corresponding SR.

Time Sequence Performance

Time sequence performance for each dataset was evaluated using six metrics: ME, RM, RC, r, TT, and ZT are presented as heatmaps in Figure 3. As in Figure 2, the heatmaps in Figure 3a and 3b correspond to the results of the ratio and mean scoring schemes. The SR scores show that the MSWEP dataset consistently appears in the topmost rank, along with the TMPA 3B42 in the second rank. On the other hand, the CHIRPS now emerges as the third-ranking dataset overperforming GPM-IMERG-F and CMORPH-CDR. In contrast to the distribution-based performance, the metrics for time sequence performance are worse with daily comparing to monthly, seasonal, and annual, time scales. Thus, in general, longer temporal aggregates improve the time sequence performance of the datasets.

It is notable from Figures 3 that metrics RM or normalized RMSE metric groups show the worst scores at all-time scales. Most datasets have RM scores less than 0.5 (0.3) for the ratio (mean) scoring schemes (Figures 3.a and 3.b). There are even negative scores, which indicate that the RMSE of datasets precipitation is larger than one standard deviation of the observation [18]. Negative values also appear in the normalized mean error (ME) scores at daily time scale, with the mean scoring scheme, except that of TMPA 3B42. In general, the mean scoring scheme tends to produce lower scores for all metrics in Figures 2 and 3. The spatial distribution of time sequence SR for all datasets at 133 stations as seen on Figure S2 in SM, showed randomly distributed and lower values than the SR of the distribution performance.

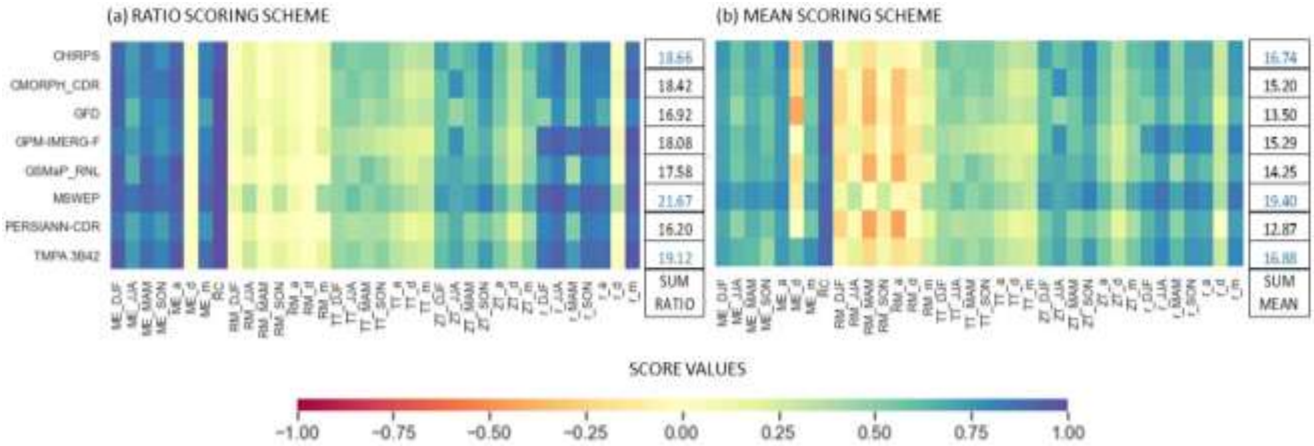

Figure 3: Heatmaps of time sequence performance metrics by ratio (a), and mean (b) with corresponding SR.

Extreme Value Representation

The two metrics for extreme value representation i.e., FSS and AD can be calculated using both percentiles and absolute thresholds. However, considering that the distributions of extreme values can be significantly different among datasets (Figure S3 in the SM), we calculated the scores only based on three percentiles i.e., p75, p90, and p98, and show the results in Figure 4. In contrast to the results of two previously discussed performance rankings, Figure 4 shows that the MSWEP and TMPA 3B42 datasets are in the bottom three with the SR scores and only perform better when compared to CMORPH CDR. In this case, the PERSIAN-CDR is in the top rank followed by GFD and CHIRPS. Figure S3 in SM showed the empirical CDF of all stations average (Figure S3a) and five stations in Sumatra, Borneo, Java, Sulawesi and Papua (from Figure S3b to S3f).

In general, all scores in this category drop drastically with higher percentiles. The ADT scores for p75 are the highest but those for p98 are the lowest among all scores. On the other hand, the scores of FSS decrease more gradually with p75, p90, and p98. These results indicate that extreme precipitations are still poorly represented in the globally gridded datasets, bearing in mind that daily precipitation values are being compared.

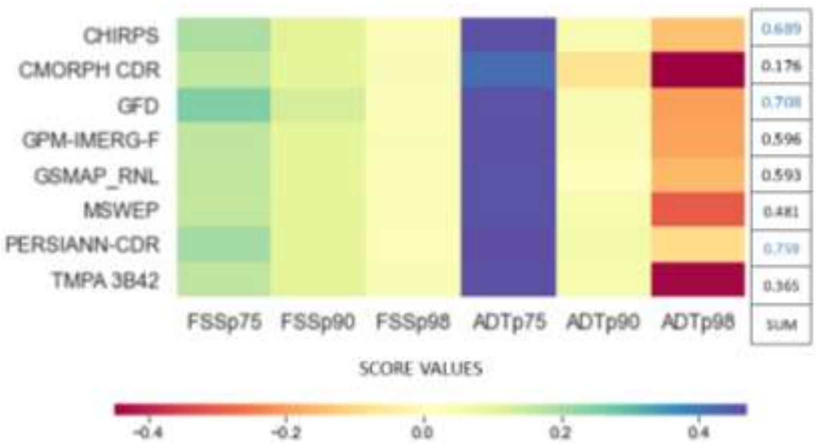

Figure 4: Heatmap of extreme value representation performance metric with SR.

Discussion

Based on SR scores, MSWEP and TMPA 3B42 are the top two datasets that consistently outperform the others for both distribution-based and time-sequence performance metric groups. It should be noted that MSWEP combines several datasets including other gridded data such as CMORPH and TMPA 3B42, and station (GHCN) data. MSWEP data processing involve bias and frequency correction, CDF matching, improvement of the peak attenuation of rain distribution, weighted combination of data sources, and refining the spatial resolution to 0.1° [31]. Results of a previous study [8] that compared 22 precipitation datasets also show that MSWEP had the best performance in temporal correlation with rain gauge observations and calibration scores for hydrological model applications. However, Figure 3 clearly shows that the scores of time-sequence performance metrics are low at daily resolution. Longer temporal aggregation is likely needed to improve the representativeness of the datasets.

In contrast, statistical performances of MSWEP and TMPA 3B42 datasets are low in extreme value representation. The averaging process when combining several data sources could have smoothing effects that eliminates real extreme values. A study by Hamada et al. [43] shows that maximum near-surface rain rate from TRMM PR data surpasses 50 mm per hour only at percentiles higher than 90. Therefore, gridded precipitation datasets may still be useful for qualitative studies of extreme events, but additional efforts should be needed in more quantitative applications.

Considering the variations in rainfall climatology of Indonesian region [41], spatial variation of performance indices might also be of user's concerns. Maps of SR scores for all 133 validating stations can be found in the SM (Figure S1). In general, the spatial distribution of SR shows random patterns. However, relatively lower scores tend to concentrate over the northern part of Sumatra Island. This area is characterized by a bimodal annual rainfall and mountainous topography.

So far, we have discussed the statistical performance of eight gridded precipitation datasets that are categorized as (high latency) research data. Some previous studies include rain-gauge based datasets, as well as near real-time (low latency) satellite products in their analysis. For more comparisons, we applied the same procedures of scoring and ranking to five additional datasets i.e., SA-OBS, APHRODITE, CMORPH-Raw, GPM-IMERG-Early Run, and GSMaP NRT (see Table S4 in the SM), and summarize the results in Table 2. By comparing the SR scores with those in Figures 2 and 3, the performance of MSWEP is comparable with that of rain-gauge based SA-OBS, whereas APHRODITE does not show good performance. On the other hand, from three low latency satellite products, GPM IMERG Early Run show the best statistical performance comparable to MSWEP in terms of SR scores.

Table 2: SR scores comparison based on datasets types

Datasets	**Distribution**		**Time Sequences**		**Extreme**
	SUM RATIO	SUM MEAN	SUM RATIO	SUM MEAN	SUM MEAN
Rain-gauge based datasets					
APHRODITE	13.06	7.85	11.40	7.72	0.32
SA-OBS	14.02	17.43	17.27	20.16	0.14
Satellite datasets (near real time)					
CMORPH-Raw	14.58	10.90	13.31	9.14	0.37

GPM-IMERG-Early Run	18.33	14.85	17.78	14.87	0.61
GSMaP_NRT	17.35	14.45	17.39	13.67	0.51

This study adopts a multimetric approach [17, 18] to evaluate climate models. We use a combination of standard continuous and categorical verification statistics [40] as quantitative measures to assess the accuracy of the rainfall estimation amounts and occurrence of the gridded precipitation dataset. We apply continuous type metrics for the performance of data distribution comprehensively and time sequences in data pairs from time to time between the rainfall of the dataset and the observed data. In addition, categorical type metrics were employed to evaluate the representation of extreme events with a threshold value. The summation of ranks diagnoses the rank of precipitation dataset performance to decide the best performance robustly. Previous studies [8, 14, 15] used fewer metrics without rank scoring and a greater focus on biases and correlation in comparison. This study compares more of the statistical distribution for all rainfall data and extreme rain days. The weakness of this study lies in not considering physical factors when comparing the rainfall of precipitation datasets and observed data, such as comparisons with altitude differences [15], while regional influences of monsoon [18] were only briefly discussed. This research is still purely statistical analysis with the quantitative evaluation of the 'value-to--to-value' between gridded precipitation datasets and point-based rain gauge stations. Nonetheless, our results could be informative for those who need to use gridded precipitation datasets, especially for climatological and hydrological applications in the Indonesian region.

Conclusions

The performance and reliability of eight gridded precipitation datasets: CHIRPS v2.0, CMORPH-CDR v1.0, GFDv3, PERSIANN-CDR v01r01, TMPA 3B42v7, GSMaP_RNL V06, GPM-IMERG V06 (Final Run), and MSWEPv2, were compared with rain gauge station observations for daily, monthly, seasonal and annual timescales in the period of 2003-2015. The findings of this study can be summarized as follows:

- A multimetric approach of 13 metrics grouped into three groups: data distribution, time sequence, and extreme value representation. The application of summation of rank deals with the ranking of all datasets for every performance metric and quantifies the scores of all metrics for diagnosing and deciding the best performing dataset.
- The results show that MSWEPv2 is the best product, followed by TMPA 3B42 for daily, monthly, seasonal, and annual precipitation in comparison with rain gauge data based on summation of rank. The extreme performance of all gridded precipitation datasets are low in more than 75th percentiles daily rainfall. This study implicates for the application of climatology and hydrology in the Indonesian region using gridded precipitation datasets.

Data Availability

The meteorological stations data are available in BMKG or Indonesia Agency for Meteorology, Climatology, and Geophysics, which can be accessed at https://dataonline.bmkg.go.id/home. Precipitation datasets used for this study are included within the paper.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Funding Statement

The authors would like to express their gratitude to the Indonesia Endowment Fund for Education (LPDP) for funding this research [Grant Number 201812210213600].

Acknowledgments

The authors would like to acknowledge the Indonesia Agency for Meteorology, Climatology, and Geophysics for facilitating the data of rainfall stations. We thank the two anonymous reviewers for their constructive comments. Author TW drafted the initial manuscript and statistical data processing. Author TWH revised and improved the overall manuscript including the storyline and discussion. Authors AS and LMH performed further revisions of the manuscript. All authors were involved in discussions during the review process and all the authors have read and approved the final manuscript. This manuscript is part of a dissertation on the improvement of methods and analysis.

Supplementary Materials

The Supplementary Materials are available at **Supplementary_doc.docx**.

References

[1] S. Seneviratne, N. Nicholls, D. Easterling, et al.,” Changes in climate extremes and their impacts on the natural physical environment.” A Special Report of Working Groups I and II of the Intergovernmental Panel on Climate Change (IPCC). 2012.

[2] IPCC. “Summary for policymakers Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part A: Global and Sectoral Aspects.” Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge, United Kingdom: Cambridge University Press, 2014.

[3] R. Djalante and F. Thomalla, “Disaster risk reduction and climate change adaptation in Indonesia: Institutional challenges and opportunities for integration.” International Journal of Disaster Resilience in the Built Environmen, vol. 3, 166-180, 2012.

[4] IPCC-TGICA. “General guidelines on the use of scenario data for climate impact and adaptation assessment (p. 66)”. Version 2. Technical report, 2007.

[5] F. Tangang, S. Supari, J. X. Chung, et al., “Future changes in annual precipitation extremes over Southeast Asia under global warming of 2°C”, APN Science Bulletin, vol. 8, no.1, 2018.

[6] S.Siswanto, G. J. van Oldenborgh,, G. van der Schrier, et al.,”Temperature, extreme precipitation, and diurnal rainfall changes in the urbanized Jakarta city during the past 130 years”. International. Journal of Climatology, 36, 3207-3225, 2016.

[7] S.Supari, F. Tangang, L. Juneng, et al., “Observed changes in extreme temperature and precipitation over Indonesia”, International. Journal of Climatology, 37, 1979-1997, 2017.

[8] H. E.Beck, N. Vergopolan, M. Pan, et al., “Global-scale evaluation of 22 precipitation datasets using gauge observations and hydrological modelling”, Hydrology and Earth System Sciences, vol. 21, no.12, 6201-6217, 2017.

[9]. Q. Sun, C. Miao, Q, Duan, et al.,.” A review of global precipitation data sets: Data sources, estimation, and intercomparisons”, Reviews of Geophysics, Vol 56, no.1, 79-107, 2018.

[10] Prasetia, R., As-syakur, A. R., & Osawa, T. Validation of TRMM Precipitation Radar satellite data over Indonesian region. Theor. Appl. Climatol. 112, 575–587, https://doi.org/10.1007/s00704-012-0756-1 (2013).

[11] F. Fatkhuroyan, T. Wati, A. Sukmana, et al., “Validation of Satellite Daily Rainfall Estimates Over Indonesia” Forum Geografi, vol 32, 170-180, 2018.

[12] Setiyoko, A., Osawa, T., & Nuarsa, I. W. Evaluation of GSMaP Precipitation Estimates Over Indonesia. Int. J. Environ. Geoscie., 3, 26-43, https://doi.org/10.24843/ijeg.2019.v03.i01.p04 (2019).

[13] E.J. Van Den Besselaar, G. Van Der Schrier, R. Cornes, et al., “SA-OBS: A daily gridded surface temperature and precipitation dataset for Southeast Asia”. Journal of Climatology, vol 30, 5151-5165, 2017.

[14] R. R. E. Vernimmen, A. Hooijer, M. Mamenun, et al., “Evaluation and bias correction of satellite rainfall data for drought monitoring in Indonesia”. Hydrology and Earth System Sciences, vol 16, 133–146, 2012.

[15] C.Y. Liu, P. Aryastana, G.R. Liu, et al., “Assessment of satellite precipitation product estimates over Bali Island”, Atmospheric Research., vol 244, 105032, 2020.

[16] R. Djalante and M. Garschagen, “A review of disaster trend and disaster risk governance in Indonesia: 1900–2015”. Disaster risk reduction in Indonesia, 21-56 , 2017.

[17] A. Ahmadalipour, A. Rana, H. Moradkhani, et al., “Multi-criteria evaluation of CMIP5 GCMs for climate change impact analysis”. Theoretical and applied climatology, vol 128, no 1-2, 71-87, 2017.

[18] R. Chhin, and S. Yoden, “Ranking CMIP5 GCMs for model ensemble selection on regional scale: case study of the Indochina Region”. Journal of Geophysical Research: Atmospheres, Vol 123 no 17, 8949-8974, 2018.

[19] A. Yatagai, O. Arakawa, K. Kamiguchi, et al., “A 44-year daily gridded precipitation dataset for Asia based on a dense network of rain gauges”. Sola, Vol 5, 137-140, 2009.

[20] A. Yatagai, K. Kamiguchi, O. Arakawa, et al., “APHRODITE: Constructing a long-term daily gridded precipitation dataset for Asia based on a dense network of rain gauges”. Bulletin of the American Meteorological Society, Vol 93, no. 9, 1401-1415,2012.

[21] C. C. Funk, P. J. Peterson, M. F. Landsfeld, et al., “A quasi-global precipitation time series for drought monitoring”. US Geological Survey data series, Vol 832, No 4, 1-12, 2014.

[22] R. J. Joyce, J. E. Janowiak, P. A. Arkin, et al., “CMORPH: A method that produces global precipitation estimates from passive microwave and infrared data at high spatial and temporal resolution”. Journal of hydrometeorology, Vol 5, 487-503, 2004.

[23] J. Sheffield, G Goteti, and E. F.Wood, “Development of a 50-year high-resolution global dataset of meteorological forcings for land surface modelling”. Jornal of Climatology, Vol 19, No 13, 3088-3111, 2006.

[24] H. Ashouri, K. L. Hsu, S. Sorooshian, et al., “PERSIANN-CDR: Daily precipitation climate data record from multisatellite observations for hydrological and climate studies”, Bulletin of the American Meteorological Society, Vol 96, no 1, 69-83, 2015.

[25] G. J. Huffman, D. T. Bolvin, E. J. Nelkin, et al., “The TRMM Multisatellite Precipitation Analysis (TMPA): Quasi-global, multiyear, combined-sensor precipitation estimates at fine scales”, Journal of hydrometeorology, Vol 8, No 1, 38-55, 2007.

[26] G. J. Huffman, R. F. Adler, D. T. Bolvin,et al., “The TRMM multi-satellite precipitation analysis (TMPA)”, In Satellite rainfall applications for surface hydrology (pp. 3-22). Springer, Dordrecht, 2010.

[27] T. Kubota, S. Shige, H. Hashizume, et al., ”Global precipitation map using satellite-borne microwave radiometers by the GSMaP project: Production and validation”, IEEE Transactions on Geoscience and Remote Sensing, Vol 45, no 7, 2259-2275, 2007.

[28] T. Kubota, K. Aonashi, T. Ushio, et al., “Global Satellite Mapping of Precipitation (GSMaP) products in the GPM era”, Satellite precipitation measurement, no.1, 355-373, 2020.

[29] G. J. Huffman, D. T. Bolvin, D. Braithwaite, et al., “NASA global precipitation measurement (GPM) integrated multi-satellite retrievals for GPM (IMERG)”, Algorithm Theoretical Basis Document (ATBD) Version, 4, 26, 2015.

[30] G. J. Huffman, D. T. Bolvin, D. Braithwaite, et al., “Integrated multi-satellite retrievals for the Global Precipitation Measurement (GPM) mission (IMERG)”, In Satellite precipitation measurement (pp. 343-353), Springer, Cham, 2020.

[31] H. E. Beck, E. F. Wood, M. Pan, et al., “MSWEP V2 global 3-hourly 0.1 precipitation: methodology and quantitative assessment”, Bulletin of the American Meteorological Society, Vol 100, no 3,473-500, 2019.

[32] BMKG, Regulation No. 04 about Observation and Management of Climate Data, 2016.

[33] W. Chen, Z. Jiang, and L. Li, “Probabilistic projections of climate change over China under the SRES A1B scenario using 28 AOGCMs”, Journal of Climatology, Vol 24, 4741-4756, 2011.

[34] S. E. Perkins, A. J. Pitman, N. J. Holbrook, et al., “Evaluation of the AR4 climate models’ simulated daily maximum temperature, minimum temperature, and precipitation over Australia using probability density functions”, Journal of Climatology, Vol 20, 4356-4376, 2007.

[35] N. M. Roberts and H. W. Lean, “Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events” Monthly Weather Review, Vol 136, no 1, 78-97, 2008.

[36] A. Ceglar, A. Toreti, G. Balsamo, et al., “Precipitation over monsoon Asia: a comparison of reanalyses and observations”, Journal of Climatology, Vol 30, 465-476, 2017.

[37] A. N. Pettitt, “A two-sample Anderson-Darling rank statistic” Biometrika,Vol 63, 161-168, 1976.

[38] F. W. Scholz, and M. A. Stephens, “K-sample Anderson–Darling tests”, Journal of the American Statistical Association, Vol 82, no 399, 918-924, 1987.

[39] J. Tan, G. J. Huffman, D. T. Bolvin, et al., “IMERG V06: Changes to the morphing algorithm. Journal of Atmospheric and Oceanic Technology, Vol 36, no 12, 2471-2482, 2019.

[40] E. E. Ebert, “Methods for Verifying Satellite Precipitation Estimates”, In Measuring precipitation from space (pp. 345-356). Springer, Dordrecht, 2007.

[41] E. Aldrian and R. D. Susanto. "Identification of three dominant rainfall regions within Indonesia and their relationship to sea surface temperature." International Journal of Climatology: A Journal of the Royal Meteorological Society, Vol 23, no. 12: 1435-1452, 2003.

[42] V. Moron, A. W. Robertson, and J.H. Qian. "Local versus regional-scale characteristics of monsoon onset and post-onset rainfall over Indonesia." Climate dynamics Vol 34, no. 2: 281-299, 2010.

[43] J. I. Hamada, D. Yamanaka, M. Matsumoto, et al. "Spatial and temporal variations of the rainy season over Indonesia and their link to ENSO." Journal of the Meteorological Society of Japan. Ser. II 80.2: 285-310, 2002.

[44] S. Prakash, A. K. Mitra, I.M. Momin, et al. "Seasonal intercomparison of observational rainfall datasets over India during the southwest monsoon season." International Journal of Climatology, Vol 35, no. 9: 2326-2338, 2015.

[45] D.Q. Huang, J. Zhu, Y. C. Zhang,et al. "Assessment of summer monsoon precipitation derived from five reanalysis datasets over East Asia". Quarterly Journal of the Royal Meteorological Society, Vol 142, no. 694: 108-119, 2016.

[46] A.B. Sekaranom, and H. Masunaga, “Comparison of TRMM-derived rainfall products for general and extreme rains over the Maritime Continent”. Journal of Applied Meteorology and Climatology, Vol 56, no.7, 1867-1881, 2017.

[47] M.T. Chuang, C. C. K. Chou, N.H.Lin, et al, “A simulation study on PM2. 5 sources and meteorological characteristics at the northern tip of Taiwan in the early stage of the Asian haze period”. Aerosol and Air Quality Research, Vol 17, no.12: 3166-3178, 2017.

[48] M. L.Tan, N. Samat, N. W. Chan, et al. "Hydro-Meteorological Assessment of Three GPM Satellite Precipitation Products in the Kelantan River Basin, Malaysia". Remote Sensing, Vol. 10, no. 7: 1011, 2018.

[49] D. K.Ray, J. S. Gerber, G. K. MacDonald, et al. "Climate variation explains a third of global crop yield variability." Nature communications, Vol 6, no. 1: 1-9, 2015.

[50] C. P. Chang, Z. Wang, J. McBride, et al. "Annual cycle of Southeast Asia—Maritime Continent rainfall and the asymmetric monsoon transition." Journal of climate, Vol 18, no. 2: 287-301, 2005.