Journal of
Water & Climate Change

IWA PUBLISHING

Journal of Water and Climate Change Vol 14 No 11, 3950 doi: 10.2166/wcc.2023.097

Performance evaluation of multi-satellite rainfall products in the Gidabo catchment, Rift Valley Basin, Ethiopia

Kehase Neway Gebretsadkan [a,*], Melsew Berihun Tamrie [a] and Haile Belay Desta [a,b]

[a] School of Hydraulic and Water Resources Engineering, Dilla University, Dilla, Ethiopia
[b] Africa Centre of Excellence for Water Management, Addis Ababa University, Addis Ababa, Ethiopia
*Corresponding author. E-mail: kehaseneway12@gmail.com

KNG, 0000-0003-1611-8524; MBT, 0000-0003-2071-007X; HBD, 0009-0000-0757-1437

ABSTRACT

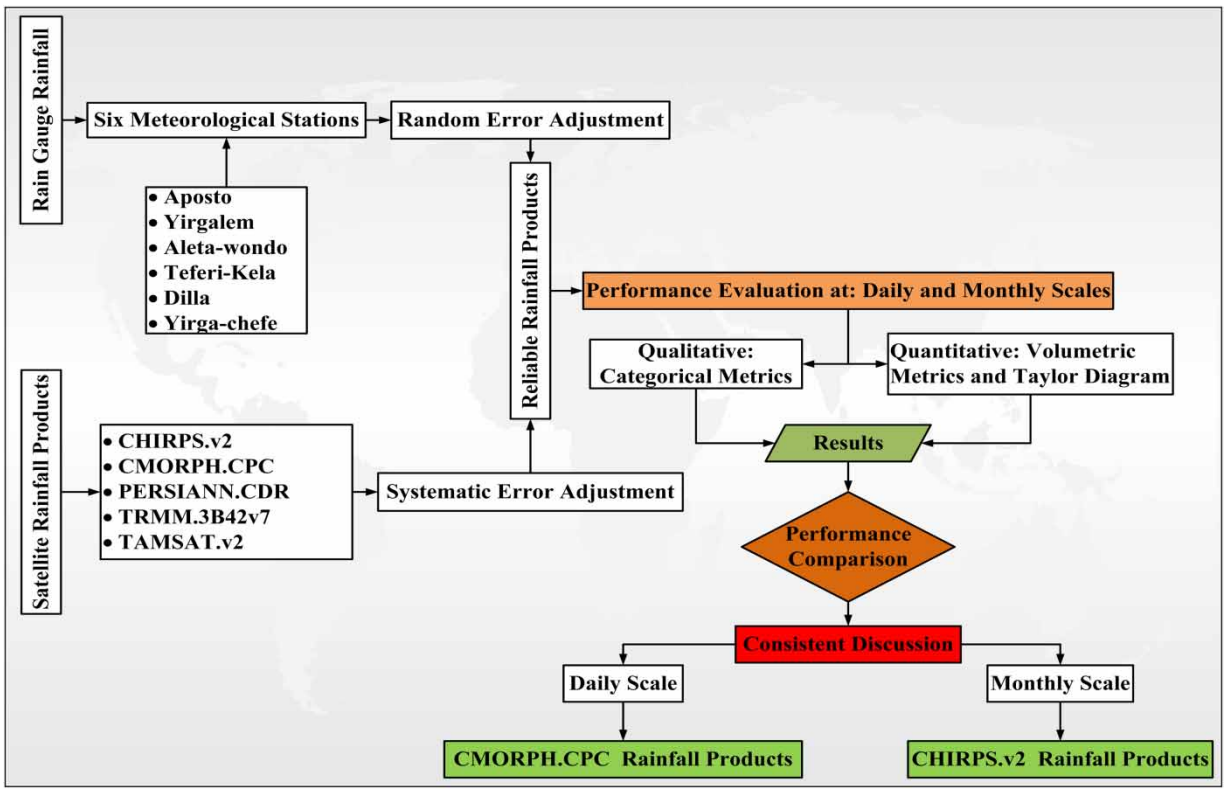
Satellite rainfalls are good options to overcome shorter records, record challenges and inconsistencies with rain-gauges. However, satellites' rainfall retrieval algorithms are region-time scale specific; hence, the key concern is selection of appropriate satellite products. Accordingly, this study evaluates the performance of five high-resolution satellites' rainfall using multiple-metrics at daily and monthly scales. The result showed that Climate Prediction Center (CPC) Morphing Algorithm (CMORPH.CPC) performed better by scoring: qualitatively; Critical Success Index (CSI = 0.856), Probability of Detection (POD = 0.911), Frequency Bias Index (FBI = 0.974), and quantitatively; correlation coefficient (CC = 0.375), Root Mean Square Error (RMSE ≈ 575), and Volumetric Critical Success Index (VCSI = 0.958) at a daily scale. At a monthly scale, Climate Hazards Group Infrared Precipitation with Stations (CHIRPS.v2) performed better by scoring CSI = 0.983, POD = 1 and FBI = 0.975 qualitatively, and quantitatively, CC = 0.836 with strong VCSI = 0.981 and better RMSE (≈125) than daily. These satellites' daily rainfall needs value-improving techniques before using in place of Gidabo's rain-gauge rainfall, while at monthly scale CHIRPS.v2's rainfall can be an alternative source of rainfall data. Finally, it ensured that for Gidabo catchment, the performance of satellite rainfall was more effective at monthly than daily scale.

Key words: multiple metrics, qualitative performance, quantitative performance, rain gauge rainfall, satellite rainfall

HIGHLIGHTS

- Performances of CHIRPS.v2, CMORPH.CPC, PERSIANN.CDR, TAMSAT.v2, and TRMM.3B42v7 rainfall were evaluated against rain gauges using qualitative and quantitative metrics.
- At daily scale, rainfall products of CMORPH.CPC performed better for both evaluation perspectives.
- At monthly scale, rainfall products from CHIRPS.v2, followed by TRMM.3B42v7, performed better.

GRAPHICAL ABSTRACT



1. INTRODUCTION

Rainfall is a vital resource to life (Sharifi *et al.* 2016) and it plays a substantial role in the livelihood of many agro-climatology regions of the world, mainly African countries that depend on rainfed agriculture (Belay *et al.* 2019). Rainfall is a decisive meteorological element in multi-water cycle practices, including hydrological modeling, hydraulic modeling, flood forecasting, water resources management and any climate change-related studies (Anie & Brema 2018; Ghorbanian *et al.* 2022). Due to different factors, such as climate change, rainfall variability influences the water system by changing the occurrence, distribution and location of water (Belay & Melesse 2022). Therefore, the nowcasts, forecasts and post-event casting investigations of rainfall pattern in a reasonable tempo-spatial scale are influential for understanding how climate influences the earth's water system (Ayehu *et al.* 2018; Liemohn *et al.* 2021). For the successful accomplishment of these practices, tempo-spatially consistent and reliable rainfall data need to be available. Due to its direct physical measurement, rain gauge rainfall data is the most accurate to represent rainfall on the earth's surface and is used for decision-level studies (Ghorbanian *et al.* 2022). However, it is impractical to get a tempo-spatially consistent record of rainfall data at rain gauge stations that are unevenly distributed, limited in number, temporally inconsistent and located remotely in the complex topography, especially in developing countries, including Ethiopia (Hordofa *et al.* 2021a; Kidie & Teklay 2022; Beles *et al.* 2023).

To overcome these limitations, the recently developed tempo-spatially high-resolution datasets – Global Climate Models (GCMs) and satellite-based grid rainfall products – are good options (Ayehu *et al.* 2018; Fenta *et al.* 2018; Hordofa *et al.* 2021a). Hordofa *et al.* (2021a) indicated that the use of GCMs rainfall needs longer rain gauge rainfall than satellite-based rainfall for downscaling and bias correction processes. Based on the application of remote sensing technology, several satellites considering their own retrieval algorithm have been developed to detect and estimate rainfall products with more inclusive tempo-spatial scales than rain gauges (Xu *et al.* 2019; Hordofa *et al.* 2021a). Satellite rainfall retrieval algorithms can be visible light and infrared (VIS/IR), passive microwave (PMW), active microwave (AMW), and multi-sensor precipitation estimation (MPE) (Hu *et al.* 2019). The VIS/IR retrieval algorithm estimates surface rainfall by geostationary optical sensors, and detects continuous rainfall intensity, but over records false alarms (Belay & Melesse 2022). The PMW

algorithm is based on the microwave radiometer carried by polar-orbiting satellites. PMW is more direct and effective than VIS/IR; however, still it underestimates heavy rainfall (Belay & Melesse 2022). The development of AMW overcomes the demerits of PMW and VIS/IR (Hu *et al.* 2019). To combine the advantages of PMW and VIS/IR and overcome the limitations in AMW, MPE was developed, and currently, it is the most common retrieval algorithm. Some of the satellites are Tropical Rainfall Measuring Mission (TRMM).3B42v7 satellite precipitation analysis, TMPA (Cattani *et al.* 2016); The precipitation Estimation from Remote Sensing Information Using Artificial Neuron Networks (PERSIANN.CDR); Climate Prediction Center (CPC) Morphing Algorithm (CMORPH) (Ghorbanian *et al.* 2022); Tropical Applications of Meteorology using Satellite (TAMSAT.v2) (Maidment *et al.* 2017); and the Climate Hazards Group Infrared Precipitation with Stations (CHIRPS.v2) (Dinku *et al.* 2018). However, it should be noted that, as per their indirect detection and estimation techniques, satellite rainfall products possess uncertainties in measurements, types of retrieval algorithms that are affected by weather variability and topographic complexity, and time scale (Kimani *et al.* 2017; Wedajo *et al.* 2021).

Evaluation of various satellite rainfall with rain gauge rainfall products has been conducted globally. For example, Hordofa *et al.* (2021a) evaluated Global Precipitation Measurement Integrated Multi-Satellite Retrieval (GPM-IMERG) and CHIRPS at monthly and seasonal scales in Ziway Lake of Ethiopia. They concluded that CHIRPS has outperformed slightly and can optionally be used for any agro-hydrological studies of the area. Bayissa *et al.* (2017) used CHIRPS.v2.0, PERSIANN, African Rainfall Climatology and Time Series (TARCAT) v2.0, TRMM and African Rainfall Estimate Climatology version 2 (ARC 2.0) at decadal, monthly and seasonal scales, to assess the spatial and temporal variability of meteorological drought in Upper Blue Nile Basin, Ethiopia. Their results showed that all satellite rainfall products have a good agreement with the rain gauge observation in the order CHIRPS.v2, TARCAT.v2, TRMM and PERSIAN, at all scales. Hence, with confidence, they concluded that CHIRPS.v2 products can be a source of alternative information for meteorological drought monitoring up to developing early warning systems at a monthly scale. CHIRPS.v2 and Multi-Source Weighted-Ensemble Precipitation (MSWEP) v2 were evaluated with reference to rainfall measured *in situ* by Taye *et al.* (2020) in monthly scale for monthly meteorological drought analysis in Upper Blue Nile Basin. Based on their result, CHIRPS.v2 performed better. Five satellites (ARC v2, TAMSAT, TRMM.3B43v7, CMORPH, CHIRPS.v2) and two reanalysis rainfall products, Climate Forecast System Reanalysis (CFSR) and the European Center for Medium Range Weather Forecast Reanalysis (ERA-Interim), were compared by Lemma *et al.* (2019) at monthly and seasonal scales. According to their conclusion, CHIRPS.v2 was superior across all rainfall regimes. Sahlu *et al.* (2017) evaluated TMPA, CMORPH, PERSIANN, European Center for Medium range Weather Forecast (ECMWF) ERA-Interim Reanalysis and MSWEP at a daily scale; hence, they implied that CMORPH exhibited the best performance. Dinku *et al.* (2018) validated CHIRP and CHIRP combined with ground observations (CHIRPS), ARC2 and TAMSAT at daily, decadal, and monthly scales for East Africa. Their result indicated that both CHIRP and CHIRPS performed well at a monthly scale; however, TAMSAT was best at a daily time scale. As a result, researchers have suggested different satellite rainfall products for the respective hydrologic region and time scale, particularly CHIRPS for east Africa (Dinku *et al.* 2018) at a monthly scale, and concluded that satellite rainfall products are region-specific (Kimani *et al.* 2017; Dinku *et al.* 2018; Wedajo *et al.* 2021), and should be carefully evaluated and ranked with the appropriate metrics before using them for any simulation (Abiola *et al.* 2013; Kimani *et al.* 2017; Liemohn *et al.* 2021).

Ghorbanian *et al.* (2022) indicated that no single evaluation metric, not all metrics for every hydrologic region (Abiola *et al.* 2013) and no single evaluation perspective could provide a decision-level evaluation of satellite rainfall products (Liemohn *et al.* 2021). It is understood that each metric was designed for a specific evaluation task; hence, the application of fewer metrics limits the findings and decision of the study. Liemohn *et al.* (2021) mentioned that metrics can be categorized based on various ways; however, only groupings and categories that are important to select the best metric are discussed. Based on the evaluation perspective of metrics, grouping was clustered into continuous and discrete, which are quantitative and qualitative, respectively. Quantitative metrics are fit performances, they are based on the exact values of rain gauge-satellite rainfall values, and the most common ones are correlation coefficient (CC), Root Mean Square Error (RMSE) and volumetric metrics, like Volumetric Critical Success Index (VCSI) (Aghakouchak & Mehran (2013); each of them measures the association, accuracy and volumetric accuracy, respectively. Qualitative metrics are based on event detections under a given threshold, and the commons of this category are critical CSI, Probability of Detection (POD), and False Alarm Ratio (FAR) to show accuracy, reliability and discrimination, respectively. Categories of metrics were based on the closeness of reference to modeled data under the criteria of accuracy, bias, precision, association, discrimination, reliability and extremes. In Ethiopia, that is outside of the Gidabo catchment, several studies have been conducted on the valuation and ranking of satellite rainfall products with rain gauges (Sahlu *et al.* 2017; Lemma *et al.* 2019; Hordofa *et al.* 2021a; Kidie & Teklay 2022). In these studies, the widely used methods

to measure the performance of satellite rainfall products were single evaluation metrics, mostly continuous, at a single performance perspective (only, quantitative) and single time scale. However, as evidenced by Abiola *et al.* (2013), continuous statistical metrics were unrealistic for all weather regions of the globe, specifically for the regions where orographic type of rainfall occurs, such as in Ethiopia.

Therefore, to fill this methodological gap, this study used the recommended multiple metrics for our catchment under pairwise evaluation perspectives; qualitative (categorical) and hybrid quantitative (volumetric and Taylor Diagram) metrics (Taylor 2001; Aghakouchak & Mehran 2013; Ayehu *et al.* 2018; Botchkarev 2019; Xu & Han 2020). The strong side of this study is its implementation of multiple performance evaluation metrics for multi-satellite rainfall products at dual analysis scales to provide temporally representative satellite rainfall products in place of the rain gauges for the specific catchment. Consequently, the objective of this study was to evaluate and compare the qualitative and quantitative performance of satellite rainfall products with rain-gauge rainfall products of the Gidabo catchment using multiple metrics and hybrid-evaluation perspectives at daily and monthly scales. For this purpose, five high-resolution satellite rainfall products: CHIRPS.v2, CMORPH.CPC, PERSIANN.CDR, TAMSAT.v2, and TRMM.3B42v7, were selected. To accomplish this, the raised research question was (1) which satellite rainfall product would have superior catchment scale temporal qualitative and quantitative performance at daily and monthly scales?

2. STUDY AREA DESCRIPTION

a. Location

The Gidado catchment is found in the Rift Valley River Basin of Ethiopia, and it lies between 6°10′0″ and 6°50′0′ N latitude and 38° 0′0″ and 38°40′0″ E longitudes with an approximate area of 3,390 km^2. It originates in the central-eastern part of the main Ethiopian Rift Valley escarpment and ends before meeting Lake Abaya; specifically, it lies in Sidama and Gedeo Zones of Southern Nation Nationalities and People of Regional State, and the Borena zone of Oromia regional state (Aragaw *et al.* 2021). In Figure 1, we tried to show the location of the catchment, the available rain gauge stations and the terrain variabilities of the area.

b. Topography

As stated by Cattani *et al.* (2016), the terrain characteristics of the Gidabo catchment are complex. The landscape of the catchment can be broadly categorized into the edge of the eastern high plateau, the large eastern escarpment of the Rift Valley and the floor of the Rift Valley, with a wide range of elevation difference from 1,147 m above mean sea level (a.m.s.l.) at Lake Abaya in the west to about 3,213 m a.m.s.l. in the north-east.

c. Climate

Climate in the Gidabo catchment ranges from semi-arid in the rift floor to humid in the plateau of the escarpment. The catchment is characterized by a bi-modal pattern of rainfall, which is two big rainy seasons (June–September) and (February–May), respectively, separated by a dry season (October–January) (Aragaw *et al.* 2021). Lemma *et al.* (2019) indicated that due to its topographic complexity and geophysical locations, the climate of Ethiopian rainfall was tempo-spatially variable and complex to quantify. The catchment has a mean annual rainfall of 1,191 mm and mean annual maximum and minimum temperatures of 26.11 and 13.55 °C, respectively. Generally, climatic description of the Gidabo catchment indicates that accessing any tempo-spatial consistent hydro-meteorologic time series data, either from the ground station or remotely from satellites, was highly challenging.

3. DATASETS AND METHODS

3.1. Datasets

i. Rain gauge data

For this study, the available daily rainfall data series of 1998–2019 from the six rain gauges (Figure 1) were obtained from Ethiopian Meteorological Agency (EMA 2021). The spatial areal coverage based on Theissen's polygon for Aposto, Yirgalem, Teferi-Kela, Aleta-wondo, Dilla and Yirga-chefe was found to be 464, 681, 327, 404, 1,079 and 435 square kilometers, respectively. However, for appropriate measurement at complex topography, like in Ethiopia of the Gidabo catchment, the spatial areal coverage of a rain gauge station should be 41.66 square kilometers (Lopez *et al.* 2015); therefore, the rain gauges of this study are sparsely distributed.

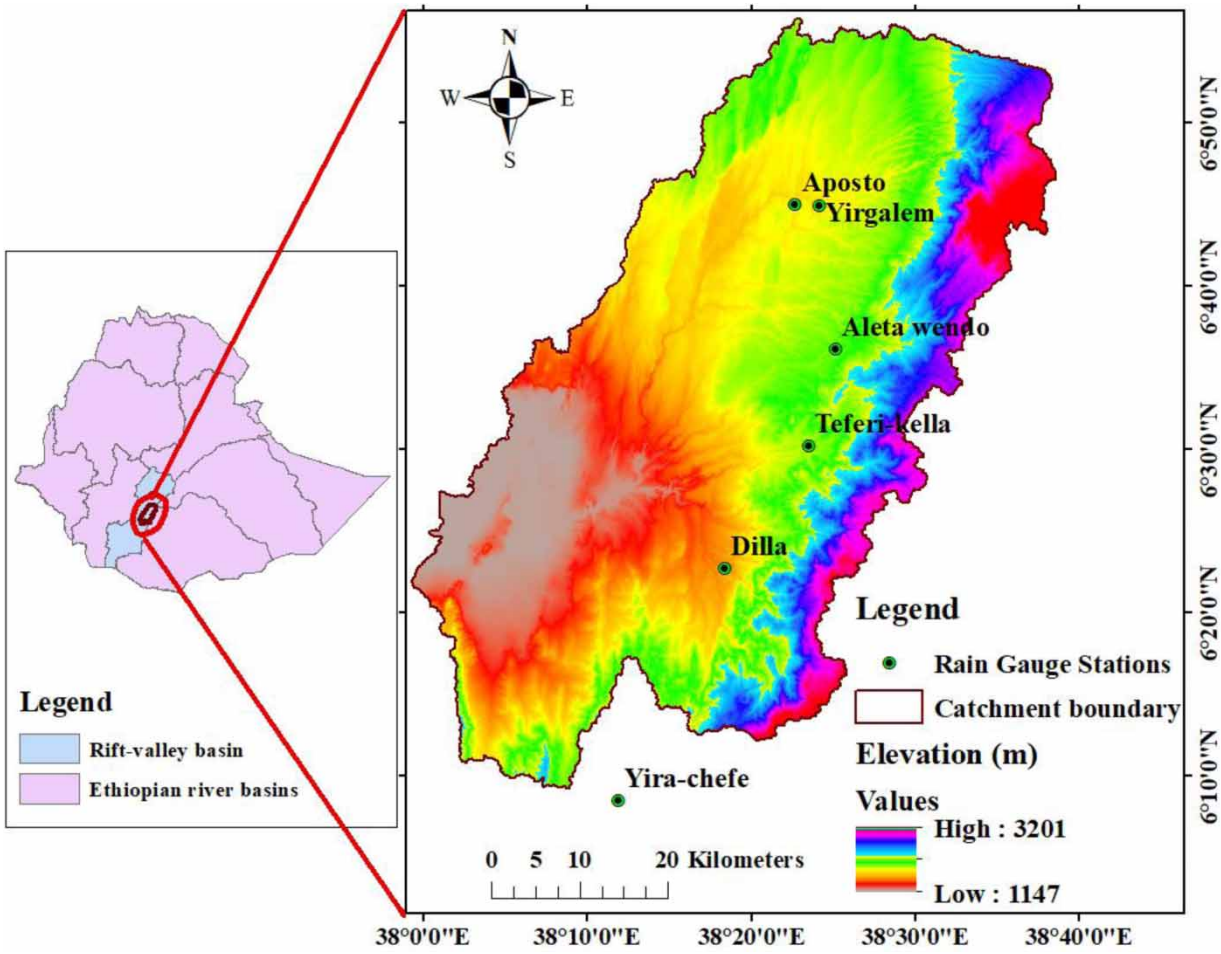


Figure 1 | Location map of the Gidabo catchment and available rain gauges.

ii. **Satellite rainfall products**

In this study, the selection of the five high-resolution satellite rainfall products was decided from the comprehensive review reports, such as tempo-spatial scales, their logical region specifications, availability and wide use in previous research studies (Lemma *et al.* 2019; Nwachukwu *et al.* 2020). Their multiple-basic descriptions are provided in Table 1.

The complementary detailed descriptions of these satellite rainfall products are available in (Dinku *et al.* 2018) for CHIRPS. v2 and TAMSAT.v2, (Cattani *et al.* 2016; Lemma *et al.* 2019; Xiao *et al.* 2020) for CMORPH.CPC, (Cattani *et al.* 2016; Xiao *et al.* 2020; Ghorbanian *et al.* 2022) for PERSIANN.CDR, and Lemma *et al.* 2019; Xiao *et al.* 2020; Ghorbanian *et al.* 2022) for TRMM.3B42v7.

Table 1 | Descriptions of the five satellite rainfall products

Datasets	Spatial/temporal resolution	Spatial coverage	Temporal span	Web source/accessed date
CHIRPS.v 2	0.25°/Daily	60°N–60°S	1981–Now	https://App.Climateengine.Com/Climateengine/August13, 2022
CMORPHv1 CPC	0.25°/Daily	60°N–60°S	1998–Now	
PERSIANN.CDR	0.25°/Daily	60°N–60°S	1983–2021	
TRMM.3B42v7	0.25°/Daily	60°N–60°S	1998–2019	
TAMSAT.v2	0.04°/Daily	60°N–60°S	1983–Now	

3.2. Research design

The overall conceptual workflow of this study is provided in Figure 2.

a. **Data processing**

To evaluate the temporal average detection and estimation performance of satellite rainfall products, the availability of temporally reliable rain gauge rainfall data is vital (Funk *et al.* 2015). However, in developing countries like Ethiopia, rain gauges have temporal limitations (Funk *et al.* 2015; Bayissa *et al.* 2017; Fenta *et al.* 2018; Musie *et al.* 2020; Hordofa *et al.* 2021a; Wedajo *et al.* 2021), as well as sensitivity to weather disruption and topographic complexity (Cattani *et al.* 2016; Kimani *et al.* 2017), which causes difficulties to measure consistently (Lemma *et al.* 2019). In addition, due to topographic and weather complexities, satellites' retrieval algorithms (Dinku *et al.* 2018; Wedajo *et al.* 2021) also faced systematic uncertainties. Therefore, before any performance analysis steps, we began to prepare reliable data using the well-known data quality assessment tests: adequacy, consistency and normalization to the rain gauge data, and normalization to the satellite rainfall datasets.

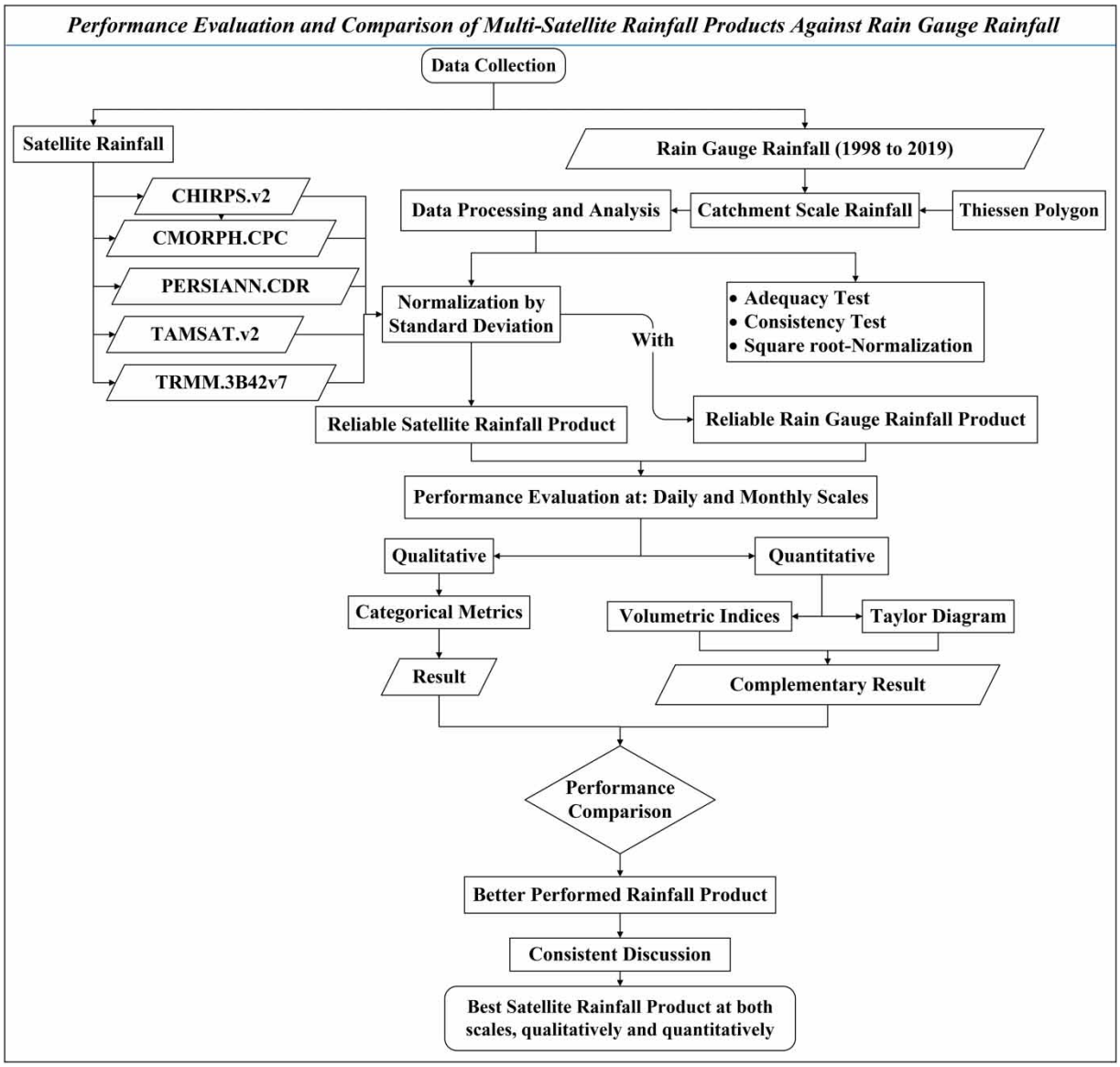


Figure 2 | Theoretical conceptual workflow of the study.

The adequacy of the rain gauge rainfall data was checked by calculating the Relative Standard Error (δ_e), which determines whether the amount of error within the data series was acceptable or not (Wijesekera & Perera 2012). Consistency refers to the slope proportionality plot of the nearby stations during the same period of record to create homogeneous data (Wijesekera & Perera 2012), and its common method was a double mass curve. The result indicated that the amount of error within the data series and consistency level was acceptable; hence, the rain gauge datasets were ready for further steps.

As per the findings of Taylor (2001), as well as Aghakouchak & Mehran (2013), Taylor diagram and volumetric analysis were the best techniques to validate the quantitative performance of satellite rainfall products. As mentioned by Abiola *et al.* (2013) and Nadeem *et al.* (2022), categorical analysis was the best-performed technique to evaluate the yes/no rain event detection capabilities of satellites. The frequency distribution of rainfall in complex terrains was skewed (Abiola *et al.* 2013). Therefore, to use the mentioned techniques, rainfall of both sources must first be normally distributed (Taylor 2001; Abiola *et al.* 2013). The tempo-spatial skewness within datasets was removed by normalization techniques (Aghakouchak & Mehran 2013; Woldemeskel *et al.* 2013; Botchkarev 2019). There exist lots of normalization methods. However, in this study, the selected performance analysis techniques and their corresponding indices were the Taylor diagram, Pearson Correlation Coefficient (CC), Normalized Root Mean Square Error (NRMSE) and Normalized Standard Deviation (δ_N), as well as volumetric and categorical elements (Hit, Miss, False Alarm and Null). Therefore, to improve these indices, square root normalization was recommended by Woldemeskel *et al.* (2013) for the daily and monthly rain gauge rainfall products. However, due to the need to minimize the dataset variability across the rain gauge and satellite rainfall products, we normalized the satellite rainfall products using the ratio of the minimum standard deviation with the normalized rain gauge data (Botchkarev 2019; Liemohn *et al.* 2021).

3.3. Performance evaluation and comparison of satellite rainfall products

The logic on precision and consistency of satellite rainfall products with rain gauge rainfall products is dynamic with time and space (Abiola *et al.* 2013; Liemohn *et al.* 2021); henceforth, the selection of best evaluation and comparison metrics is a very critical stage. In this study, details of each selected method are provided in the next sections.

i. **Categorical metrics**

Categorical metrics were used to determine the qualitative performance of satellite rainfall products by identifying rainy/dry days of rain gauge records. Based on the principle of contingency table and precipitation threshold (Ayehu *et al.* 2018; Liemohn *et al.* 2021), categorical statistical metrics were qualitative indicators of consistencies between satellite and rain gauge rainfall products. According to Ghorbanian *et al.* (2022), the precipitation threshold was invented to detect yes/no rainfall days, to compensate for the uncertainties in light precipitations and due to the watershed humidity, and recommendations from (Peinó *et al.* 2022) which satisfy the mean observation of the respected time scale, 1 mm/day and 100.28 mm/month, were fixed for both time scales of this study. The contingency table is a threshold sensitive 2×2 matrix which summarizes the four combinations of both rainfall datasets and is summarized by categorical elements: Hit, Miss, False alarm and Null. Hit represents the rain event detected by both the satellite algorithms and rain gauge observations; Miss denotes the real rainfall events, but not detected by the satellite; False alarm represents rain events detected by the satellite, but not confirmed in the rain gauge; and Null indicates the correctly detected non-rain events, including rains below the threshold. For this study, these elements were estimated from IF, COUNTIF, IF-OR and IF-AND automatic logical scenarios of spreadsheet application to calculate categorical metrics.

For this study, we used the most common categorical metrics (Abiola *et al.* 2013); POD, FAR, Critical Success Index (CSI) and Frequency Bias Index (FBI). POD measures the percentage of rain gauge rainfall events that are detected by satellites. FAR indicates the fraction of rainfall events detected by the satellite but not confirmed in the rain gauge. CSI measures the overall qualitative performance of satellite rainfall products against the rain gauge rainfall observations. The FBI is the ratio of satellite rainfall to rain gauge rainfall to measure the over-detected or under-detected satellite precipitation occurrences (Xu *et al.* 2019). The details of each metric are provided in Table 2.

ii. **Volumetric metrics**

As stated by Taylor (2001), Aghakouchak & Mehran (2013), and Ayehu *et al.* (2018), categorical metrics can only measure the qualitative features of satellite rainfall products; hence, there was a gap in the quantitative simulation of the satellite products. Therefore, Aghakouchak & Mehran (2013) developed a new approach by extending categorical metrics to volumetric statistical metrics, to examine the volumetric performance of satellite rainfall products, which is suitable for gridded datasets,

Table 2 | Summary of categorical, volumetric, and Taylor diagram evaluation indices[a]

Statistical indices		Equation	Value range	Best score
Categorical indices	POD	$\frac{\text{Hits}}{\text{Hits} + \text{Misses}}$	[0,1]	1
	FAR	$\frac{\text{False Alarms}}{\text{Hits} + \text{False Alarms}}$	[0,1]	0
	CSI	$\frac{\text{Hits}}{\text{Hits} + \text{Misses} + \text{False Ararm}}$	[0,1]	1
	FBI	$\frac{\text{Hits} + \text{False Alarms}}{\text{Hits} + \text{Misses}}$	[0, ∞]	1
Volumetric indices	VHI	$\frac{\sum_{i=1}^{n}(S_i \mid (S_i > t \,\&\, G_i > t))}{\sum_{i=1}^{n}(S_i \mid (S_i > t \,\&\, G_i > t)) + \sum_{i=1}^{n}(G_i \mid (S_i \le t \,\&\, G_i > t))}$	[0,1]	1
	VFAR	$\frac{\sum_{i=1}^{n}(S_i \mid (S_i > t \,\&\, G_i \le t))}{\sum_{i=1}^{n}(S_i \mid (S_i > t \,\&\, G_i > t)) + \sum_{i=1}^{n}(S_i \mid (S_i > t \,\&\, G_i \le t))}$	[0,1]	0
	VCSI	$\frac{\sum_{i=1}^{n}(S_i \mid (S_i > t \,\&\, G_i \le t))}{\sum_{i=1}^{n}((S_i \mid (S_i > t \,\&\, G_i > t)) + (G_i \mid (S_i \le t \,\&\, G_i > t)) + (S_i \mid (S_i > t \,\&\, G_i \le t)))}$	[0,1]	1
	VMI	$\frac{\sum_{i=1}^{n}(G_i \mid (S_i \le t \,\&G_i > t))}{\sum_{i=1}^{n}(S_i \mid (S_i > t \,\&\, G_i > t)) + \sum_{i=1}^{n}(G_i \mid (S_i \le t \,\&\, G_i > t))}$	[0,1]	0
Taylor Diagram	CC	$\frac{\sum((S_i - S_{avg})(G_i - G_{avg})}{\sqrt{\sum((S_i - S_{avg})^2 \sum(G_i - G_{avg})}}$	[−1,1]	1
	RMSE	$\sqrt{\frac{1}{N-d}\sum_{t=1}^{N}(S_i - G_i)^2}$	[0,∞]	0

[a]S is the satellite rainfall estimate, G is the rain gauge rainfall, n is the sample size, t is the rainfall threshold, S_i is the individual satellite rainfall, G_i is the individual rain gauge rainfall, N is the total number of pairs in both datasets, and d is the degree of linear freedom, $d = 2$.

and was used in this study. The parameters of volumetric statistical indices are Volumetric Hit Index (VHI), Volumetric False Alarm Ratio (VFAR), Volumetric Critical Success Index (VCSI) and Volumetric Miss Index (VMI). VHI is the measure of the volume of correctly detected rainfall by the satellites in reference to the volume of correctly detected satellites and missed rain gauge rainfall observations. VFAR is the volume of false rainfall products by the satellite relative to the sum of rainfall by the satellite. VMI represents the volumetric fraction of missed observations relative to the volume of practically detected simulations and missed observations, and VCSI is defined as the overall measure of volumetric performance. The details of each metric are provided in Table 2.

iii. **Taylor diagram metrics**

Quantitative performance of satellite rainfall products in reference to rain gauge rainfall products can be evaluated using different statistical metrics. However, the issue is in selecting appropriate metrics that can build complete and decision-level results (Liemohn *et al.* 2021). Taylor (2001) has developed a comprehensive statistical-visual plot, a quantitative performance evaluation performance score-based one-way and less vague 2D-diagrammatic summary (Xu *et al.* 2019), which is the Taylor diagram, to summarize the quantitative degree of correspondence between satellite and rain gauge rainfall products, in terms of three most common statistical metrics (Xu *et al.* 2016): Pearson CC, Root Mean Square Error (RMSE) and Normalized Standard Deviation (δ_N). CC provides the degree of linear correlation between the rain gauge and satellite datasets as a function of time; however, RMSE represents the overall error level or accuracy and δ_N indicates the scattering of both data sets from their respective mean to represent percent bias (Xu *et al.* 2016).

The Taylor diagram can be constructed using normalized input data (Nadeem *et al.* 2022) in different free and commercial software products like GrADS, IDL, and others (Taylor 2001); however, due to its easiness, having active users' community and accessibility, in this paper, it was drawn by R-programming software. A Taylor diagram has three geometric components (Xu *et al.* 2016): the radii sides, the isoline curves that represent RMSE and the quartile circle that represents CC. These three statistical metrics are related to each other with the concept of error propagation formula derived from the law of cosine and detailed in Table 2.

4. RESULTS

4.1. Performance and comparison of satellite rainfall products at a daily scale

i. **Categorical metrics**

Table 3 summarizes the catchment scale temporal qualitative performance (rainfall frequency) of the five satellite rainfall products. As per the result, CMORPH.CPC performed better over the rest of the satellite rainfall products and their average indices; the highest POD, reasonable FAR, and highest SCI and FBI scores over the others were 0.911, 0.063, 0.856 and 0.974, respectively. When considering only the FAR values, TAMSAT.v2, CHIRPS.v2 and TRMM.3B42v7 scored 0.013, 0.021 and 0.044, respectively, and performed in the decreasing order of their rainfall frequency detection performance with fewer false records. The FBI values of all satellites were less than the perfect score 1; hence, this revealed that all the satellites were under-detected by the confirmed daily rain gauge rainfall events of the catchment. Overall, from the categorical analysis at a daily scale, the average temporal qualitative performance of the satellites is ordered in the last column of Table 3.

ii. **Volumetric metrics**

The volumetric performance (rainfall intensity) evaluation of the five satellite rainfall products at a daily scale is provided in Table 4. The result evidenced that CMORPH.CPC exhibited the leading volumetric performance as the highest VHI, lowest VFAR, highest VCSI and lowest VMI values, represented by 0.987, 0.022, 0.958 and 0.029, respectively. This means CMORPH.CPC has a 98.7% skill of correctly estimating the reference volume; however, it still reflected a weakness for about 2.2% false rainfall volume simulations and 2.9% volumetric fraction of missed values in reference to the rain gauge rainfall records.

Table 3 | Average categorical performance summary of satellites at a daily scale (1998–2019)

	Categorical metrics				
Datasets	**POD**	**FAR**	**CSI**	**FBI**	**Rainfall detection efficiency**
CHIRPS.v2	0.443	0.021	0.445	0.451	CMORPH.CPC
CMORPH.CPC	0.911	0.063	0.856	0.974	PERSIANN.CDR
PERSIANN.CDR	0.811	0.060	0.768	0.859	TRMM.3B42v7
TRMM.3B42v7	0.793	0.044	0.765	0.831	TAMSAT.v2
TAMSAT.v2	0.637	0.013	0.631	0.646	CHIRPS.v2

Table 4 | Average volumetric performance summary of satellites at a daily time scale (1998–2019)

	Volumetric metrics				
Datasets	**VHI**	**VFAR**	**VCSI**	**VMI**	**Rainfall estimation efficiency**
CHIRPS.v2	0.573	0.022	0.562	0.432	CMORPH.CPC
CMORPH.CPC	0.987	0.025	0.958	0.029	TRMM.3B42v7
PERSIANN.CDR	0.841	0.041	0.821	0.156	PERSIANN.CDR
TRMM.3B42v7	0.851	0.027	0.834	0.150	TAMSAT.v2
TAMSAT.v2	0.754	0.043	0.749	0.235	CHIRPS.v2

The rainfall products of TRMM.3B42v7 followed by PERSIANN.CDR were the next outperformed products to replace rain gauge stations of the Gidabo catchment. When we compare TAMSAT.v2 and CHIRPS.v2, it seems to have some confusion; the VHI value of TAMSAT.v2 was 0.754, which was better than that of CHIRPS.v2, 0.573; however, the VFAR of TAMSAT.v2 was much more in error than the VFAR of CHIRPS.v2, and as a metric, it was more dangerous than VHI. Therefore, in order to select the accurate satellite rainfall product, we must depend on the overall volumetric comparison index, VCSI; hence, their volumetric efficiency is provided in the last column of Table 4.

iii. **Taylor diagram metrics**

Figure 3 summarizes the quantitative visual and comprehensive statistical comparison in terms of δ_N, CC and NRMSE to define the percent bias, association and accuracy of the five satellites' rainfall products with the gauges separately with in a single plot. As indicated by Taylor (2001) and Ghorbanian *et al.* (2022), satellites that have closer δ_N, CC and RMSE values to the gauged rainfall have the best accuracy. As Taylor (2001) indicated, CC and RMSE are the most used association and accuracy measurement metrics, and as per the result of Figure 3, all the satellite datasets have a poor linear agreement with the ground datasets by scoring CC values less than 0.7 (reference value for good correlation) approximately as 0.290, 0.353, 0.345, 0.152 and 0.375 for CHIRPS.v2, TRMM.3B42, TAMSAT.v2, PERSIANN.CDR and CMORPH.CPC, respectively. Therefore, from the value of the CC, CMORPH.CPC, TRMM.3B42v7 and TAMSAT.v2 were the first three outclassed satellites in the Gidabo catchment.

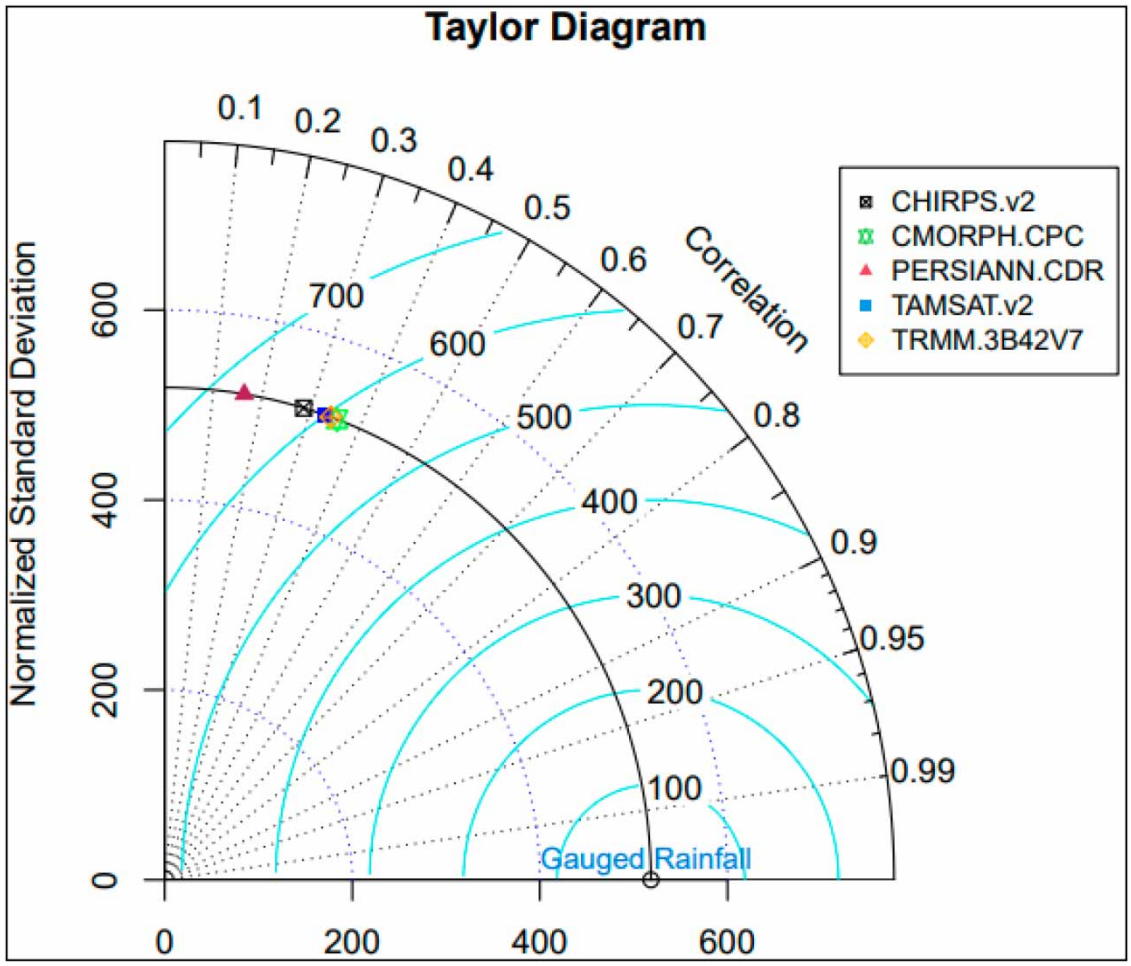


Figure 3 | The Taylor diagram plot showing the average quantitative performance of the five satellites' rainfall product by CC, RMSE, and δN at grid to catchment, and daily time scales (1998–2019).

Based on the δ_N result, the value of the observed datasets lies between 400 and 600 units, which is represented by the curved solid line extended across the radii. The δ_N of all the satellite rainfall products are aligned with the rain gauge values as displayed in Figure 3. This indicates that all the satellite datasets have independently the same pattern with the gauged rainfall to their average value, which is evidence for the consistency of the satellites' data to be a reliable input for comparison. In case of the RMSE concept, the perfect performance of a satellite is indicated by its zero value and by detecting the overall error levels; however, in Figure 3, the values were from little less than 600 up to ~675 units; as a result of this, CMORPH.CPC tailed by TRMM.3B42 and TAMSAT.v2 were the three better outshined satellites. Taylor (2001) explained that both CC and RMSE showed complementary correspondence details; so then to draw complete performance information, a normalized standard deviation was an additional parameter. The result in Figure 3 also shows the same agreement with this fact; the rank of satellite products from the normalized standard deviation is the same as the rain gauge rainfall; hence, the drawn rank from CC and RMSE was supported with this third parameter for its best decision.

4.2. Performance and comparison of satellite rainfall products at a monthly scale

i. Categorical metrics

Table 5 shows the average categorical rainfall performance of the satellites in terms of POD, FAR, SCI and FBI at a threshold of 100.28 mm/month on a monthly time scale, from 1998 to 2019. Considering POD, CHIRPS.v2 has scored the highest level 1, TRMM.3B42 v7 scored 0.964, PERSIANN.CDR 0.955, TAMSAT.v2 0.875, and CMORPH.CPC scored 0.867. From this, we concluded that at a monthly scale, CHIRPS.v2 followed by TRMM.3B42 v7 and PERSIAN CDR were the first three best satellites to detect the real rain event confirmed by the rain gauge stations. The rainfall-based overall qualitative accuracy (SCI) of CHIRPS.v2, CMORPH.CPC, PERSIANN.CDR, TRMM.3B42 v7 and TAMSAT.v2 was 0.983, 0.850, 0.909, 0.923 and 0.857, respectively; hence, from this score, CHIRPS.v2 exhibited the highest performance.

As per the concept of FBI, the best score for the perfect performance of satellites is 1, under/over-detections for below/above 1, respectively, while as illustrated in Table 5, they are 0.975, 0.850, 1.009, 1.012 and 0.868, respectively, according to satellites' order in column 1. Therefore, PERSIANN.CDR over detects the rain gauge rain events to a degree of 0.009 and TRMM.3B42 v7 over detects the reference data to a degree of 0.012. This indicates that both satellites have over-detections of light precipitation that was confirmed at the rain gauge, and relatively, PERSIANN.CDR possesses a lower degree of over-detections. On the contrary, CHIRPS.v2, CMORPH.CPC and TAMSAT.v2 under detect the ground rain event with the degree of 0.026, 0.15 and 0.132, respectively; hence, their qualitative performance was ordered decreasingly as CHIRPS.v2, TAMSAT.v2, and CMORPH.CPC. When we compare the satellites having the nature of over-detections and under-detections, the basic issue is their degree of factor from the best score, 1 (their distance from the reference). For instance, in Table 5, PERSIANN.CDR over detects the reference data by 0.009 and CMORPH.CPC under detects by 0.15; so then, PERSIANN.CDR possesses a smaller distance and leads to better performance. Then, accordingly, the monthly overall rainfall detection efficiency of all satellites is ordered in the last column of Table 5.

ii. Volumetric indices

The volumetric performance of the five satellite rainfall products at a monthly time scale is provided in Table 6. As per the evaluation, the volumetric hit rate simulation of each satellite was measured by comparing their VHI and VCSI values, at least, nearest to 1, and their VFAR and VMI values nearest to 0, to be perfect ground station rainfall volume estimators. The result, Table 6 evidenced that CHIRPS.v2 showed the leading average volumetric performance by scoring the reasonable

Table 5 | Average temporal categorical performance summary of satellites at a monthly scale from 1998 to 2019

	Categorical metrics				
Datasets	**POD**	**FAR**	**CSI**	**FBI**	**Rainfall detection efficiency**
CHIRPS.v2	1	0.010	0.983	0.975	CHIRPS.v2
CMORPH.CPC	0.867	0.014	0.850	0.850	TRMM.3B42v7
PERSIANN.CDR	0.955	0.015	0.909	1.009	PERSIANN.CDR
TRMM.3B42v7	0.964	0.016	0.923	1.012	TAMSAT.v2
TAMSAT.v2	0.875	0.013	0.857	0.868	CMORPH.CPC

Table 6 | Average volumetric performance summary of satellites at a monthly time scale (1998–2019)

	Volumetric metrics				
Datasets	**VHI**	**VFAR**	**VCSI**	**VMI**	**Rainfall estimation efficiency**
CHIRPS.v2	0.989	0.000	0.981	0.016	CHIRPS.v2
CMORPH.CPC	0.924	0.003	0.929	0.059	TRMM.3B42v7
PERSIANN.CDR	0.884	0.006	0.887	0.019	TAMSAT.v2
TRMM.3B42v7	0.978	0.001	0.971	0.024	CMORPH.CPC
TAMSAT.v2	0.947	0.002	0.952	0.038	PERSIANN.CDR

values of VHI, VFAR, VCSI and VMI as 0.982, 0.000, 0.981 and 0.016, respectively. This indicated that CHIRPS.v2 has 98.2% skill of correctly estimating the rain gauge volume with zero percent of false alarm, and the rainfall products of TRMM.3B42v7 and TAMSAT.v2 were the next outperformed products to replace rain gauge stations of the Gidabo catchment at a monthly time scale.

iii. **Taylor diagram metrics**

Figure 4 illustrates the monthly comprehensive statistical result of CHIRPS.v2, CMORPH.CPC, PERSIANN.CDR, TRMM.3B42v7 and TAMSAT.v2 rainfall products at the catchment level. Except for PERSIANN.CDR (CC < 0.7), the rest

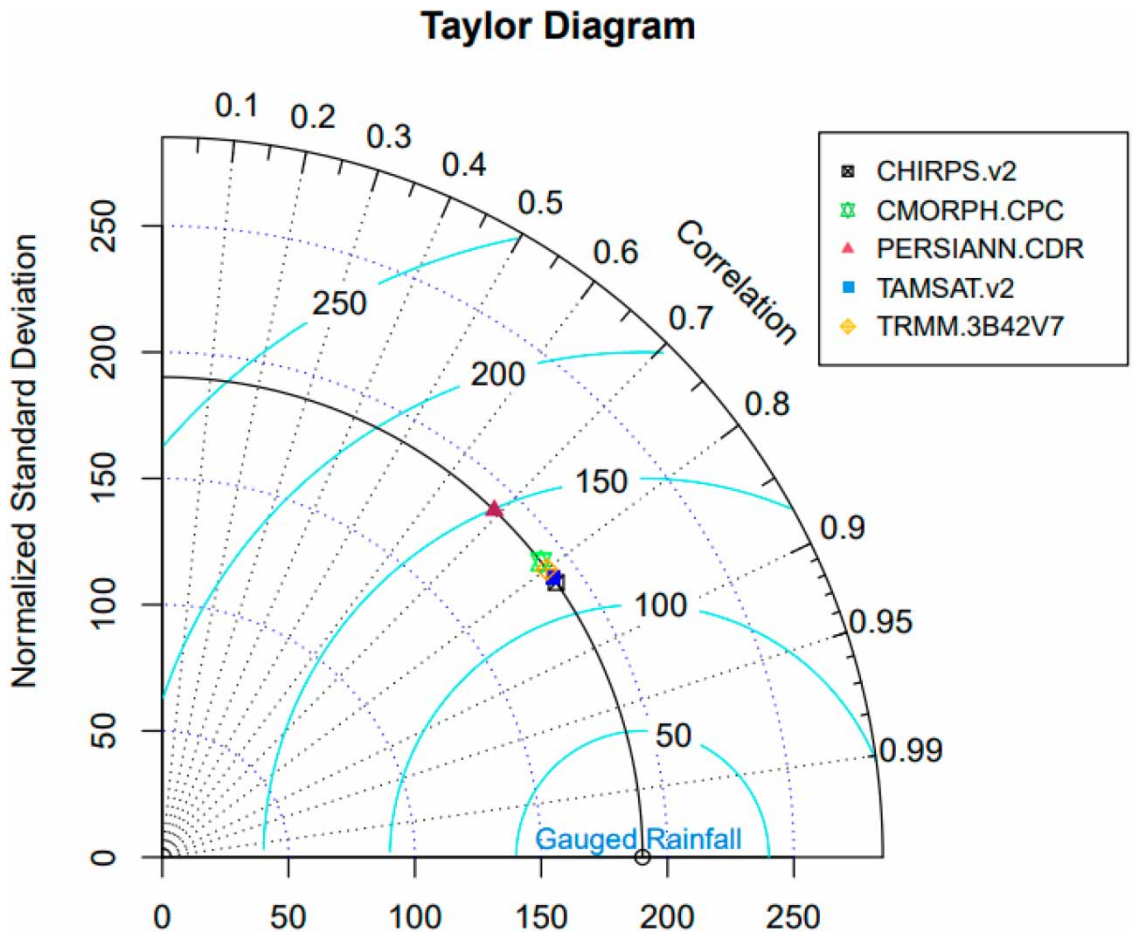


Figure 4 | The Taylor diagram plot showing the average quantitative performance of the five satellites' rainfall product by CC, RMSE, and δN at grid to catchment, and monthly scales (1998–2019).

of the satellites' rainfall dataset has acceptable linear agreement with the rain gauge rainfall datasets by scoring CC values approximately 0.836, 0.825, 0.800 and 0.793 for CHIRPS.v2, TRMM.3B42, TAMSAT.v2, and CMORPH.CPC, respectively.

The overall quantitative error level (RMSE) of each satellite was labeled from some greater than 100 up to a little less than 150 units. Therefore, CHIRPS.v2, followed by TRMM.3B42 and TAMSAT.v2, were the best three performing satellites. For this ranking, the normalized standard deviation was supportive by indicating that each satellite rainfall datasets have the same pattern with the reference data to their average value.

5. DISCUSSION

The application of multiple metrics improves the decision level of satellite to rain gauge evaluations (Liemohn *et al.* 2021). The best way to select an appropriate evaluation metric was the insight consideration of groupings and categories (Liemohn *et al.* 2021). Groupings of metrics provide specifications for the use of continuous or discrete metrics, and/or both of them are quantitative and qualitative evaluation perspectives. Categories of metrics are mechanisms to fix the number of metrics that would be used in each grouping based on its multiple criteria: accuracy, bias, precision, association, skill, extremes, discrimination and reliability. In this study, we used qualitative and quantitative performance evaluation perspectives; hence, both continuous and discrete assessments were applied. It is understood that both assessments have various metrics subjected to different sets of analysis (Aghakouchak & Mehran 2013); however, to select the appropriate metrics for our study, we considered the category criteria: (1) for qualitative; reliability, discrimination, accuracy and precision of the satellites measured by POD, FAR, CSI and FBI, respectively; (2) in the case of quantitative, (a) volumetrically, we used VHI, VMI, VCSI and VFAR to measure the corresponding volumetric metrics mentioned in order in (1). Since the application of satellites' quantitative fit to represent tempo-spatially consistent rain gauge rainfall products is vital in various hydro-meteorologic, hydraulic, climate change studies and policy making, we add the most common and recommended quantity measurement metrics (Taylor 2001; Xu *et al.* 2016; Liemohn *et al.* 2021), (b) CC, RMSE and δ_N to measure the association, accuracy and bias of the satellites, respectively, to boost our analysis leads to provide complementary and comprehensive rainfall intensity results. The Taylor Diagram, developed by (Taylor 2001), was the best, quick, and not only evaluation but also comparison method, to use these metrics effectively (Xu & Han 2020) at a single plot; hence, we used it. Therefore, this study was the result of qualitative evaluation from five categorical metrics and quantitative evaluation of hybrid analysis under seven metrics, generally, the result of appropriate comparison methods.

At a daily scale, all the satellites under-detected the number of rain gauge rainfall events (Table 3; POD, FBI & CSI < 1, FAR > 0) and underestimated the quantities of rain gauge rainfall records (Table 4, VHI & VCSI < 1, and VFAR and VMI > 0) and in Figure 3 (CC < 1, RMSE > 0). This was an expected result at complex hydrological regions where retrieval of orographic rainfall is challenging (Kimani *et al.* 2017), like in the Gidabo catchment. As concluded by Nadeem *et al.* (2022) and Dinku *et al.* (2018), the retrieval algorithm of satellites has detection and estimation limitations at a finer temporal scale; hence, the result of this study was consistent. Even though CMORPH.CPC showed poor performance in reference to the standard score, since the comparison is relative and its better performance score values – detection metrics (Table 3) and volumetric metrics (Table 4), and δ_N, RMSE and CC (0.375) in Figure 3 – have made it to be the better-performing satellite. The value of RMSE (little less than 600) was far from its perfect score 0, especially at the finest time scale, but, still, it was the leading score; in Liemohn *et al.* (2021), it was mentioned that while the RMSE was less than the standard deviation of both rainfall datasets, it can be considered as a good comparison. Volumetric metrics were developed from categorical metrics from input data having the same normal distribution under a combination of the same contingency table categorical elements to measure quantitative performance (Aghakouchak & Mehran 2013). As a result, when CMORPH.CPC categorically performs better, it has also the probability to outperform volumetrically. Therefore, the best performance of CMORPH.CPC qualitatively and quantitatively at a daily scale was the consistent result. The truthfulness of this result proved that CMORPH.CPC is among the finest tempo-spatial resolution satellites, which makes it to own an improved source of rainfall retrieval algorithm over complex hydrological regions at a daily scale (Xiao *et al.* 2020), like in Ethiopia (Cattani *et al.* 2016; Hordofa *et al.* 2021b; Kidie & Teklay 2022). On daily scales, CHIRPS.v2 derives its rainfall products using VIS/IR retrieval algorithm, which relies on a cold cloud duration (CCD), which made it to characterize with lower performance (Dinku *et al.* 2018) and its limitations on penetrating clouds to detect rainfall information that leads to under detection and estimation (Belay & Melesse 2022). Overall, these results were appropriate and consistent with the findings of previous studies, including Sahlu *et al.* (2017) and Dinku *et al.* (2018).

In the case of monthly scales, CHIRPS.v2 had scored the best qualitative and quantitative performance over the rest of the satellites, and the acceptable scores with the standards are summarized in Table 5 (POD = 1, FAR = 0.01, CSI = 0.983 and FBI = 0.975), and Table 6 and Figure 4, respectively. In these results, CHIRPS.v2 rainfall products was the best performer at a monthly scale due to (1) its high spatial purpose of development; CHIRPS.v2 was earlier developed to support African Rainfall Climatology (Novella & Thiaw 2013); hence, its monthly rainfall dataset includes gauge information with a larger number of gauges from East Africa, particularly 50 of them were from Ethiopia (Funk *et al.* 2015; Taye *et al.* 2020; Hordofa *et al.* 2021b), and (2) its capability of retrieval algorithm in developing its monthly rainfall product from three reliable inputs: (a) Climate Hazards Group Precipitation Climatology (CHPclim) at 0.05° resolution based on station data, average satellite observations, elevation, latitude and longitude, (b) VIS/IR of satellite observations and measures monthly gauge data for bias correction and (c) *in situ*-rain gauge measurements (Ayehu *et al.* 2018; Dinku *et al.* 2018; Hordofa *et al.* 2021b; Wedajo *et al.* 2021; Ray *et al.* 2022). So, after this was reviewed at the early stage of this study and from conclusions of some researchers on better satellite rainfall products for East Africa, we hypothesized that CHIRPS.v2 would be an expected better result at the monthly scale of this study. Furthermore, as it was proved by previous studies in Ethiopia, such as Bayissa *et al.* (2017); Dinku *et al.* (2018); Ayehu *et al.* (2018); Lemma *et al.* (2019);Wedajo *et al.* (2021); Hordofa *et al.* (2021a); Taye *et al.* (2020), and globally, by Ray *et al.* (2022); Morsy *et al.* (2021); and Tramblay *et al.* (2016), the monthly result of CHIRPS.v2 of this study has significant consistency.

The result of this study showed that the detection and estimation performance of satellites were critically affected by variations of temporal scales from daily to monthly and the selection of satellites' retrieval algorithms. Regardless of the change of the outperformed satellite rainfall products from CMORPH.CPC-daily to CHIRPS.v2-monthly, their corresponding degree of performance across all evaluation metrics improved at a monthly scale; for example, qualitatively, CSI was improved from 0.856 to 0.983, and quantitatively, VCSI was improved from 0.958 to 0.981, CC was improved from 0.375 to 0.836, and RMSE was improved from little less than 600 to little greater than 100. For regions having the same descriptions like topography and weather variability with Gidabo catchment, CMORPH.CPC and CHIRPS.v2 were better for qualitative and quantitative performances at daily and monthly scales, respectively. As per the decisions of Anjum *et al.* (2018); Dinku *et al.* (2018); Xiao *et al.* (2020); Kidie & Teklay (2022); Peinó *et al.* (2022); and Nadeem *et al.* (2022), it was the consistent conclusion on satellite rainfall products that they have lower performance at a finer temporal scale.

In the Gidabo catchment, no related studies have been conducted before; hence, this study was conducted from the recent literature in a pairwise evaluation perspective at multiple metrics because its results will be essential and preferably appropriate inputs for any decision-level simulations and modelings to water resource systems and climate analysis in the catchment. However, it should be noted that this study was faced with limitations sourced from (1) the use of temporally short data records from the spatially unrepresentative rain gauges' rainfall as input data, (2) the used input data were not conducted for uncertainty evaluation, and (3) it considers only single meteorologic variable (rainfall) to evaluate and compare the performance of satellites. As stated in the datasets section of this study, the temporal record of the rain gauges was only 22 years from sparse networks, and rainfall data from sparse rain gauges were less performed than CHIRPS2 and IMERG6 at Dhidhessa River Basin, Ethiopia (Wedajo *et al.* 2021). Therefore, rainfall from these gauges was inappropriate to be the reference data in performance evaluation of satellites' rainfall products. As indicated by Lemma *et al.* (2019), satellites' rainfall products perform better at the well rain gauge networked catchments. Therefore, to improve the spatial coverage of the rain gauges that leads to getting rain gauge representative rainfall data from the satellites, the responsible organizations should carefully plan to install appropriate rain gauge networks. Regardless of error minimization capabilities of all the conducted metrics in this study, the used input data were not considered – uncertainty estimations which may cause inadequate similarity patterns and under-decision-level results. Hence, to minimize these limitations, future researchers may use the Generalized Likelihood Uncertainty Estimations (GLUE) (Yuan *et al.* 2019) before conducting any analysis metrics. To draw the overall meteorologic performance of satellites, the performance of satellites worked at rainfall should be checked for the other meteorologic variables. To this end, the next research needs to test the performance of satellites on multi- meteorologic variables at hybrid sets of evaluations using reasonable multiple metrics for at least two-time scales.

6. CONCLUSIONS

The rain gauge stations in the Gidabo catchment were with impractical networks and short temporal rainfall data records, which is challenging for conducting any decision-level studies in the region. To solve this, five high-resolution satellites were evaluated and compared, based on the recent techniques. The generated conclusions were as follows:

- On the daily scale: due to their low algorithm capabilities at a daily scale (Nadeem *et al.* 2022) for complex hydrological regions (Kimani *et al.* 2017; Dinku *et al.* 2018), all the satellite rainfall products used in this study under-detected the rain gauge rainfall events and underestimated the rain gauge rainfall products, and relatively, CMORPH.CPC had better accuracy in both performance evaluation perspectives. Therefore, before value-improving techniques, they cannot be an alternative rainfall product in place of the rain gauge rainfall products in the Gidabo catchment and elsewhere with similar terrain complexity and climate variability.
- On the monthly scale: CHIRPS.v2, TRMM.3B42v7, TAMSAT.v2, and CMORPH.CPC showed the highest detection and estimation ability of the rain gauge rainfall products up to their rainfall and can be an alternative rainfall source in place of the rain gauge rainfall of the Gidabo catchment. However, in order to get a perfect agreement, they still need value improvement techniques, such as ensembling and merging with other reanalysis products.

Overall, this study ensured that satellite rainfall products at the Gidabo catchment and elsewhere with similar catchment characteristics were more effective at a monthly scale than a daily scale. The findings of this study can be useful for researchers and policymakers in making informed decisions about the use of satellite rainfall products for various applications, mainly climate change trend assessment, agro-hydrological modeling, hydrological simulations, any type of drought analysis, hydrological dam breach analysis and flood inundation forecasting.

AUTHOR CONTRIBUTIONS

Early motivation, title suggestion, first proof reading, critical review, rain gauge rainfall data collection, satellite data website suggestion, journal and software selection were performed by H.B.D; early introduction structure and conceptualization and early proof reading were done by M.B.T; introduction review and re-structuring, re-conceptualization, satellite data accessing, methodology design, data processing and analyses, result writing and interpretation, discussions, visualization, and preparation of all rounds of the revised manuscript were handled by K.N.G.

FUNDING

This research has not received any external funding and it is the not-for-profit sector.

ACKNOWLEDGEMENTS

We forward special thanks to satellites' retrieval algorithm and R-programing software developers, and EMA for supplying us the rain gauge rainfall data for free.

DATA AVAILABILITY STATEMENT

Data cannot be made publicly available; readers should contact the corresponding author for details.

CONFLICT OF INTEREST

The authors declare there is no conflict.

REFERENCES

Abiola, S., Mohd-Mokhtar, R., Ismail, W., Mohamad, N. & Mandeep, J. S. 2013 Categorical statistical approach to satellite retrieved rainfall data analysis in Nigeria. *Scientific Research and Essays* **8** (43), 2123–2137. https://doi.org/10.5897/SRE2013.5512.

Aghakouchak, A. & Mehran, A. 2013 Extended contingency table: performance metrics for satellite observations and climate model simulations. *Water Resources Research* **49** (10), 7144–7149. https://doi.org/10.1002/wrcr.20498.

Anie John, S. & Brema, J. 2018 Rainfall trend analysis by Mann-Kendall test for Vamanapuram river basin, Kerala. *International Journal of Civil Engineering and Technology* **9** (13), 1549–1556.

Anjum, M. N., Ding, Y., Shangguan, D., Ahmad, I., Ijaz, M. W., Farid, H. U., Yagoub, Y. E., Zaman, M. & Adnan, M. 2018 Performance evaluation of latest integrated multi-satellite retrievals for Global Precipitation Measurement (IMERG) over the northern highlands of Pakistan. *Atmospheric Research* **205**, 134–146. https://doi.org/10.1016/j.atmosres.2018.02.010.

Aragaw, H. M., Goel, M. K. & Mishra, S. K. 2021 Hydrological responses to human induced land use/land cover changes in the Gidabo River basin, Ethiopia. *Hydrological Sciences Journal* **66** (4), 640–655. doi:10.1080/02626667.2021.1890328.

Ayehu, G. T., Tadesse, T., Gessesse, B. & Dinku, T. 2018 Validation of new satellite rainfall products over the Upper Blue Nile Basin, Ethiopia. *Atmospheric Measurement Techniques* **11** (4), 1921–1936. https://doi.org/10.5194/amt-11-1921-2018.

Bayissa, Y., Tadesse, T., Demisse, G. & Shiferaw, A. 2017 Evaluation of satellite-based rainfall estimates and application to monitor meteorological drought for the Upper Blue Nile Basin, Ethiopia. *Remote Sensing* **9** (7). https://doi.org/10.3390/rs9070669.

Belay, H. & Melesse, A. M. 2022 Merging satellite products and rain-gauge observations to improve hydrological simulation: a review. *Earth* **3** (4), 1275–1289.

Belay, A. S., Fenta, A. A., Yenehun, A., Nigate, F., Tilahun, S. A., Moges, M. M., Dessie, M., Adgo, E., Nyssen, J., Chen, M., Van Griensven, A. & Walraevens, K. 2019 Evaluation and application of multi-source satellite rainfall product CHIRPS to assess spatio-temporal rainfall variability on data-sparse western margins of Ethiopian highlands. *Remote Sensing* **11** (22). https://doi.org/10.3390/rs11222688.

Beles, M., Blue, U., Abbay, N., Belay, A., Wondmageghu, H., Zimale, F. A. & Endalew, A. 2023 Evaluation of multiple bias correction methods with different satellite rainfall products. *Journal of Water and Climate Change* **14** (1), 156–174. https://doi.org/10.2166/wcc.2022.244.

Botchkarev, A. 2019 Performance metrics (Error measures) in machine learning regression, forecasting and prognostics: properties and typology. *Journal of Hydrometeorology* **2018**, 1–37. Available from: http://arxiv.org/abs/1809.03006.

Cattani, E., Merino, A. & Levizzani, V. 2016 Evaluation of monthly satellite-derived precipitation products over East Africa. *Journal of Hydrometeorology* **17** (10), 2555–2573. https://doi.org/10.1175/JHM-D-15-0042.1.

Dinku, T., Funk, C., Peterson, P., Maidment, R., Tadesse, T., Gadain, H. & Ceccato, P. 2018 Validation of the CHIRPS satellite rainfall estimates over Eastern Africa. *Quarterly Journal of the Royal Meteorological Society* **144**, 292–312. https://doi.org/10.1002/qj.3244.

EMA 2021 National Meteorology Agency, Data Service: meteorological station information. Ethiopian Meteorological Agency, Addis Ababa, Ethiopia.

Fenta, A. A., Yasuda, H., Shimizu, K., Ibaraki, Y., Haregeweyn, N., Kawai, T., Belay, A. S., Sultan, D. & Ebabu, K. 2018 Evaluation of satellite rainfall estimates over the Lake Tana basin at the source region of the Blue Nile River. *Atmospheric Research* **212**, 43–53. https://doi.org/10.1016/j.atmosres.2018.05.009.

Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., Husak, G., Rowland, J., Harrison, L., Hoell, A. & Michaelsen, J. 2015 The climate hazards infrared precipitation with stations – a new environmental record for monitoring extremes. *Scientific Data* **2**. https://doi.org/10.1038/sdata.2015.66.

Ghorbanian, A., Mohammadzadeh, A., Jamali, S. & Duan, Z. 2022 Performance evaluation of six gridded precipitation products throughout Iran using ground observations over the last two decades (2000–2020). *Remote Sensing* **14** (15). https://doi.org/10.3390/rs14153783.

Hordofa, A. T., Leta, O. T., Alamirew, T., Kawo, N. S. & Chukalla, A. D. 2021a Performance evaluation and comparison of satellite-derived rainfall datasets over the Ziway lake basin, Ethiopia. *Climate* **9** (7). https://doi.org/10.3390/cli9070113.

Hordofa, A. T., Leta, O. T., Alamirew, T. & Chukalla, A. D. 2021b Spatiotemporal trend analysis of temperature and rainfall over Ziway Lake Basin, Ethiopia. *Hydrology* **9** (1). https://doi.org/10.3390/hydrology9010002.

Hu, Q., Li, Z., Wang, L., Huang, Y., Wang, Y. & Li, L. 2019 Rainfall spatial estimations: a review from spatial interpolation to multi-source data merging. *Water* **11** (3), 579. https://doi.org/10.3390/w11030579.

Kidie, M. A. & Teklay, A. 2022 *Spatio-temporal Validation of Multi-Source Satellite Rainfall Products in the Upper Tekeze-Atbara Basin, Ethiopia*. https://doi.org/10.21203/rs.3.rs-1604785/v1.

Kimani, M. W., Hoedjes, J. C. B. & Su, Z. 2017 An assessment of satellite-derived rainfall products relative to ground observations over East Africa. *Remote Sensing* **9** (5). https://doi.org/10.3390/rs9050430.

Lemma, E., Upadhyaya, S. & Ramsankaran, R. A. A. J. 2019 Investigating the performance of satellite and reanalysis rainfall products at monthly timescales across different rainfall regimes of Ethiopia. *International Journal of Remote Sensing* **40** (10), 4019–4042. https://doi.org/10.1080/01431161.2018.1558373.

Liemohn, M. W., Shane, A. D., Azari, A. R., Petersen, A. K., Swiger, B. M. & Mukhopadhyay, A. 2021 RMSE is not enough: guidelines to robust data-model comparisons for magnetospheric physics. *Journal of Atmospheric and Solar-Terrestrial Physics* **218**, 105624.

Lopez, M. G., Wennerström, H., Nordén, L.-Å. & Seibert, J. 2015 Location and density of rain gauges for the estimation of spatial varying. *Source: Geografiska Annaler. Series A* **97** (1), 167–179.

Maidment, R. I., Grimes, D., Black, E., Tarnavsky, E., Young, M., Greatrex, H., Allan, R. P., Stein, T., Nkonde, E., Senkunda, S. & Alcántara, E. M. U. 2017 A new, long-term daily satellite-based rainfall dataset for operational monitoring in Africa. *Scientific Data* **4**. https://doi.org/10.1038/sdata.2017.63.

Morsy, M., Scholten, T., Michaelides, S., Borg, E., Sherief, Y. & Dietrich, P. 2021 Comparative analysis of TMPA and IMERG precipitation datasets in the arid environment of El-Qaa plain, Sinai. *Remote Sensing* **13** (4), 1–19. https://doi.org/10.3390/rs13040588.

Musie, M., Sen, S. & Srivastava, P. 2020 Application of CORDEX-AFRICA and NEX-GDDP datasets for hydrologic projections under climate change in Lake Ziway sub-basin, Ethiopia. *Journal of Hydrology: Regional Studies* **31**, 100721.

Nadeem, M. U., Ghanim, A. A. J., Anjum, M. N., Shangguan, D., Rasool, G., Irfan, M., Niazi, U. M. & Hassan, S. 2022 Multiscale ground validation of satellite and reanalysis precipitation products over diverse climatic and topographic conditions. *Remote Sensing* **14** (18), 4680. https://doi.org/10.3390/rs14184680.

Novella, N. S. & Thiaw, W. M. 2013 African rainfall climatology version 2 for famine early warning systems. *Journal of Applied Meteorology and Climatology* **52** (3), 588–606. https://doi.org/10.1175/JAMC-D-11-0238.1.

Nwachukwu, P. N., Satgé, F., El Yacoubi, S., Pinel, S., Bonnet, M.-P., Satge, F. & Pinel, S. 2020 How reliable are satellite-based precipitation data across Nigeria. **12** (23). https://doi.org/10.3390/rs12233964ï.

Peinó, E., Bech, J. & Udina, M. 2022 Performance assessment of GPM IMERG products at different time resolutions, climatic areas and topographic conditions in Catalonia. *Remote Sensing* **14** (20), 5085. https://doi.org/10.3390/rs14205085.

Ray, R. L., Sishodia, R. P. & Tefera, G. W. 2022 Evaluation of gridded precipitation data for hydrologic modeling in North-Central Texas. *Remote Sensing* **14** (16). https://doi.org/10.3390/rs14163860.

Sahlu, D., Moges, S. A., Nikolopoulos, E. I., Anagnostou, E. N. & Hailu, D. 2017 Evaluation of high-resolution multisatellite and reanalysis rainfall products over east Africa. *Advances in Meteorology* **2017**. https://doi.org/10.1155/2017/4957960.

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). https://doi.org/10.3390/rs8020135.

Taye, M., Sahlu, D., Zaitchik, B. F. & Neka, M. 2020 Evaluation of satellite rainfall estimates for meteorological drought analysis over the upper blue Nile basin, Ethiopia. *Geosciences (Switzerland)* **10** (9), 1–22. https://doi.org/10.3390/geosciences10090352.

Taylor, K. E. 2001 Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research Atmospheres* **106** (D7), 7183–7192. https://doi.org/10.1029/2000JD900719.

Tramblay, Y., Thiemig, V., Dezetter, A. & Hanich, L. 2016 Evaluation of satellite-based rainfall products for hydrological modelling in Morocco. *Hydrological Sciences Journal* **61** (14), 2509–2519. https://doi.org/10.1080/02626667.2016.1154149.

Wedajo, G. K., Kebede Muleta, M. & Gessesse Awoke, B. 2021 Performance evaluation of multiple satellite rainfall products for Dhidhessa River Basin (DRB), Ethiopia. *Atmospheric Measurement Techniques* **14** (3), 2299–2316. https://doi.org/10.5194/amt-14-2299-2021.

Wijesekera, N. T. & Perera, L. R. 2012 *Study on Key Issues of Data and Data Checking for Hydrological Analyses (Case Study of Attanagalu Oya Basin of Sri Lanka)*. Vol. xxxxv. The Institution of Engineers, Colombo, Sri Lanka, p. 2.

Woldemeskel, F. M., Sivakumar, B. & Sharma, A. 2013 Merging gauge and satellite rainfall with specification of associated uncertainty across Australia. *Journal of Hydrology* **499**, 167–176.

Xiao, S., Xia, J. & Zou, L. 2020 Evaluation of multi-satellite precipitation products and their ability in capturing the characteristics of extreme climate events over the Yangtze River Basin, China. *Water (Switzerland)* **12** (4). https://doi.org/10.3390/W12041179.

Xu, Z. & Han, Y. 2020 Short communication comments on 'DISO: a rethink of Taylor diagram'. *International Journal of Climatology* **40** (4), 2506–2510. https://doi.org/10.1002/joc.6359.

Xu, Z., Hou, Z., Han, Y. & Guo, W. 2016 A diagram for evaluating multiple aspects of model performance in simulating vector fields. *Geoscientific Model Development* **9** (12), 4365–4380. https://doi.org/10.5194/gmd-9-4365-2016.

Xu, J., Ma, Z., Tang, G., Ji, Q., Min, X., Wan, W. & Shi, Z. 2019 Quantitative evaluations and error source analysis of Fengyun-2-based and GPM-based precipitation products over mainland China in summer, 2018. *Remote Sensing* **11** (24). https://doi.org/10.3390/rs11242992.

Yuan, F., Zhang, L., Soe, K. M. W., Ren, L., Zhao, C., Zhu, Y., Jiang, S. & Liu, Y. 2019 Applications of TRMM- and GPM-era multiple- satellite precipitation products for flood simulations at sub-daily scales in a sparsely gauged watershed in Myanmar. *Remote Sensing* **11** (2). https://doi.org/10.3390/rs11020140.

First received 29 January 2023; accepted in revised form 12 August 2023. Available online 25 August 2023