A methodology for multiobjective evaluation of precipitation products for extreme weather (In a data-scarce environment)

Lu, Sha; ten Veldhuis, Marie-claire; van de Giesen, Nick

DOI
10.1175/JHM-D-19-0157.1

Publication date
2020

Document Version
Final published version

Published in
Journal of Hydrometeorology

A Methodology for Multiobjective Evaluation of Precipitation Products for Extreme Weather (in a Data-Scarce Environment)

SHA LU, MARIE-CLAIRE TEN VELDHUIS, AND NICK VAN DE GIESEN

Department of Water Management, Delft University of Technology, Delft, Netherlands

(Manuscript received 21 July 2019, in final form 30 March 2020)

ABSTRACT

In this paper, a methodology is proposed to quantitatively evaluate precipitation products for multiple purposes. Evaluation mainly focuses on rainfall characteristics relevant to hydrological or agricultural applications: spatial distribution pattern, effect of aggregation over time, the capture of small-scale variability and seasonality, detection of dry spells and wet spells, and timing and volume of heavy rainfall events. Verification statistics were modified and metrics were reported for extreme weather performance, such as flood and drought monitoring. The analysis was performed for different rainfall categories, over regions dominated by different weather systems or with different topographical structures. The latest versions of seven commonly available, high-resolution rainfall estimates have been evaluated by the method against daily data from 16 rain gauge stations over Tanzania, during 1998–2006. They were TRMM 3B42, CHIRPS, TAMSAT, CMORPH_RAW, CMORPH_BLD, WFDEI_CRU, and CPCU. All products, except for CMORPH_BLD and CPCU, were poorly correlated to gauge data at daily time scale with correlation coefficients < 0.5. Five-day aggregation was the minimum time scale that can be used for the products to reach an accuracy better than monthly-mean of gauge data. Their performance varied across different climatic or topographical regions and different rainfall seasons. Timing of precipitation was inaccurately estimated by all products, particularly for heavy rains, with less than 40% hits. The results of the evaluation procedure allow discrimination between available products and better selection of the product to be used for a specific application, such as crop insurance or flood early warning, under particular climatic conditions.

1. Introduction

Accurate rainfall observations at high resolution are an urgent priority for development of weather services (Pendergrass et al. 2017), such as flood prediction (IPCC 2012), crop yield management (Rowhani et al. 2011), and weather-index-based insurance (Collier et al. 2009). Quantitative precipitation estimates (QPEs) based on satellite imagery and/or numerical weather prediction (NWP) model outputs are being widely used to seek evidence of climate change in precipitation, to improve the understanding of the hydrologic cycle, or as input to precipitation-related models (such as crop yield model or flood warning system). Thorough and quantitative evaluation of QPEs across a range of rainfall features for multiple purposes is critical for selection of datasets, decision-making, and risk assessment. However, it is a challenge when using existing evaluation approaches, especially over regions with limited gauge availability. Unfortunately, the network of functional weather stations has been deteriorating progressively and significantly since the 1980s (Shiklomanov et al. 2002; Stokstad 1999), and Africa in particular has the lowest reporting rate of any region in the world (Washington et al. 2006). An exception is the fast growing Trans-African Hydrometeorological Observatory (TAHMO) network, but this is a very recent development (van de Giesen et al. 2014). Uneven and sparse distribution of gauges combined with high spatial variability of (daily) rainfall limits the representativeness of gauge data, particularly for regions of complex topography (Flitcroft et al. 1989).

Recently, an evaluation of 22 global datasets was conducted by Beck et al. (2017a) using both gauge data and hydrological modeling. One of their conclusions stated that notably for Africa, global precipitation products performed poorly, and the limited availability and quality of rain gauge and flow data makes it difficult to provide reliable recommendations for the use of

Denotes content that is immediately available upon publication as open access.

Corresponding author: Sha Lu, s.lu-1@tudelft.nl, s.lu_86@yahoo.com

DOI: 10.1175/JHM-D-19-0157.1

precipitation products in this region. Most of the evaluation work from past studies focused on particular aspects, such as the ability to deal with the complex topography (Dinku et al. 2007, 2010), the performance during a specific season (Jobard et al. 2011), or for a certain application (Cohen Liechti et al. 2012; Gosset et al. 2013; Li et al. 2013; Thiemig et al. 2012). They looked at the general rainfall characteristics relevant to hydrological uses, such as spatial distribution, seasonality, and number of dry/wet days. Few focused on extreme weather conditions, such as length of dry spells and wet spells, and heavy rainfall intensities. These issues are critical for many applications, but are difficult to be evaluated with existing methods that are mostly conducted using basic performance scores, such as Pearson correlation coefficient (PCC), Nash–Sutcliffe efficiency score (NSE), root-mean-square error (RMSE), bias in the annual rain amount (Bias), and bias in the frequency of rainy days (FBias). These performance scores, however, have implicit shortcomings. For instance, PCC and NSE are insensitive to short-scale variability (precipitation pattern) in a climate with large seasonal variability. RMSE, Bias, and FBias represent the degree of error, and bias in rain rate or rain frequency, but give no information on timing of heavy rainfall and dry/wet spells. To better understand the sources of error or use a product properly, evaluation is also required of the performance of the QPEs for different rainfall regimes, different seasons, and the timing of rainfall with respect to wet spells and heavy rains. In addition, local climatology should be taken into account for the classification of wet seasons and dry seasons, instead of applying a uniform classification over the whole region of study, especially for gauge-scarce regions.

In this study, we propose a methodology that combines evaluation of multiple performance aspects of rainfall products to identify minimum conditions under which QPEs can be used or, conversely, to quantify performance given a certain application context. Specifically, we seek implications of performance for application for extreme weather services such as flood prediction, drought assessment, and weather-index-based insurance. The method aims to better understand the sources of error (e.g., data inputs or estimation algorithms), the impact of different factors (e.g., local climatology or topography) on their performance, and their ability to capture extreme weather, in order to identify pathways for improvement. The evaluation method for extreme weather is inspired by the fundamental definition of extreme weather conditions (Karl et al. 1999; Schmidli and Frei 2005). We apply the method to evaluate the performance of seven widely used high-resolution datasets for Tanzania, where high-quality gauge data are available. The application for the Tanzania datasets demonstrates how the method works in a challenging, data-scarce environment. Section 2 briefly describes the geography and climate of the study region, and the datasets to be evaluated. Section 3 presents the methodology of the evaluation, including the definition of a range of performance metrics. Section 4 analyzes the results and gives an in-depth discussion. Conclusions are drawn including the summary for individual dataset in section 5.

2. Datasets

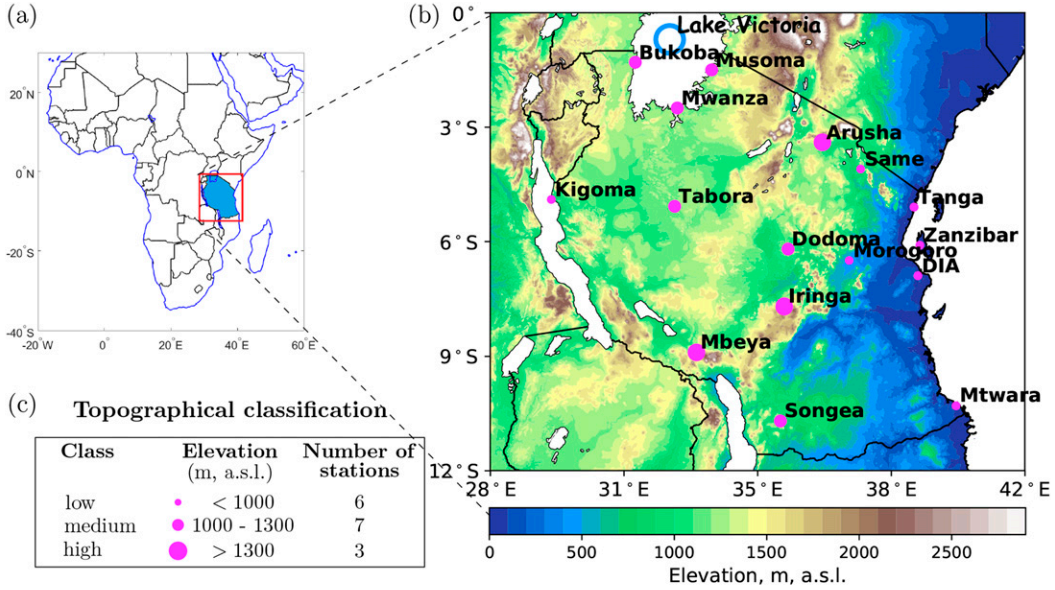
a. *Selection of study region and its rainfall climate*

For this study, we selected Tanzania as study region, it being representative of data-scarce environments, typical of Africa and other tropical regions, and including a complex topography of mountains and coastline. Moreover, a unique dataset of 36 years and good quality was available from 16 rain gauges across the country. Tanzania lies between the water bodies of Lake Victoria in the north, Lake Tanganyika to the west, Lake Nyasa to the south and the Indian Ocean to the east. Its mainland is divided into a central plateau, highlands along the north and south, and coastal plains. Northeast Tanzania is mountainous and includes Mount Meru, Mount Kilimanjaro (the highest point in Africa, 5950 m above mean sea level), and the Usambara and Pare mountain ranges. West of those mountains lies the Gregory Rift, which is the eastern arm of the Great Rift Valley. The center of Tanzania is a large plateau, which is part of the East African Plateau. Most of Tanzania, except the eastern coastline lies above 200 m above mean sea level (MSL) as shown in Fig. 1b.

Tanzania has a tropical climate but has regional variations influenced by its location with respect to the equator (latitude), the impact of the Indian Ocean, topography (elevation), and proximity to large water bodies. Seasonal rainfall is driven mainly by the migration of the intertropical convergence zone (ITCZ). The ITCZ migrates southward through Tanzania in October to December, reaching the south of the country in January and February, and returns northward in March, April, and May. This causes the north and east of Tanzania to experience two distinct wet periods, while other parts of the country have only one (Okoola 1999).

b. *Rain gauge data*

Daily rain gauge data of high quality from Tanzania Meteorological Agency (TMA) were used as reference ground observations. The data were quality controlled based on a standardized quality control procedure (USAID 2016). They were made available for this research, but are not available openly. The country-level records can contain more data than those that are publicly available.



Topographical classification

Class	Elevation (m, a.s.l.)		Number of stations
low	•	< 1000	6
medium	●	1000 - 1300	7
high	⬤	> 1300	3

FIG. 1. Topographical information of Tanzania and the TMA stations with (a) geographical location of Tanzania; (b) locations of TMA stations and topographical map of Tanzania with major rivers and lakes (white regions), where Lake Victoria is the white area marked by the blue circle; and (c) topographical classes of the stations.

The dataset covers the years from 1970 to 2006, which were used for the computation of critical thresholds such as 90th and 99th percentiles of precipitation as explained in section 3c. The period of study is from 1998 to 2006, since some satellite-based QPE products are available only from 1998 onward.

c. *Satellite-based and model-based rainfall estimates*

Seven datasets are evaluated in this study, since they performed relatively well over Africa according to past studies, and are representative of a variety of rainfall estimation algorithms. Five of them are satellite-based products that combine multiple microwave, infrared and radar sensors: the Tropical Rainfall Measuring Mission Multisatellite Precipitation Analysis (TMPA) 3B42, version 7.0 (TRMM 3B42; Huffman et al. 2010; Huffman and Bolvin 2015); Tropical Applications of Meteorology Using Satellite and Ground Based Observations, version 3.0 (TAMSAT; Tarnavsky et al. 2014; Maidment et al. 2014, 2017); the Climate Hazards Group Infrared Precipitation with Stations, version 2.0 (CHIRPS; Funk et al. 2015); a satellite-only product created using the Climate Prediction Center morphing method (CMORPH), version 1.0 (CMORPH_RAW; Joyce et al. 2004); and a satellite–gauge blended product created using CMORPH, version 1.0 (CMORPH_BLD; Xie and Xiong 2011). The sixth is a model-based analysis dataset, WATCH Forcing Data methodology applied to ERA-Interim reanalysis data with bias correction from Climatic Research Unit (CRU) time series (WFDEI_CRU; Dee et al. 2011; Weedon et al. 2014). A gauge-based gridded dataset, NOAA Climate Prediction Center Unified daily gauge analysis (CPCU; Xie et al. 2007, 2010), is used to demonstrate the degree of dependence of the TMA gauge data, whose performance will not be as thoroughly diagnosed as the other six datasets. A brief description of the seven datasets is given in Table 1.

The five satellite datasets have different input sources. CMORPH_RAW is the only product that does not ingest gauge data. TAMSAT is the only IR-based product, while others use a combination of IR and PMV for rainfall estimation.

d. *Data preprocessing*

Analysis was conducted at five time scales, which were daily, pentadal (5-day), 5-day moving average, dekadal (10-day), and monthly basis. We defined pentads and dekads such that every calendar month contains six pentads and three dekads, respectively. The following data preprocessing procedures were performed. Days with missing observations, pentadal and dekadal periods with at least one day missing data, and months with at least three days missing data were removed from the datasets for the four time scales, respectively.

For the calculations of moving average and number of consecutive dry or wet days (required in section 3c), infilling of the missing data was needed. Each dataset was checked for five or more days of missing data in dekads and 30 or more days of missing data in one year, the maximal period deemed acceptable for interpolation.

The datasets in the period of the study do not have missing data more than these thresholds, except for TAMSAT, which contains nine pentads of missing data in this 9-yr period of study. Missing data were filled in by taking the value of the temporally closest good data. Given the small number of data gaps, this is not expected to decrease the reliability of the evaluation.

TABLE 1. Description of the seven rainfall estimates.

Name	Data input	Spatial resolution	Temporal resolution	Spatial coverage	Temporal coverage	Reference	Archive source[a]
TRMM 3B42 v7.0	IR, MW, satellite–radar, gauges	0.25°	3-hourly	50°N–50°S	1998–2015	Huffman and Bolvin (2015)	A
TAMSAT v3.0	IR, gauges	0.0375°	Daily	Africa	1983–2016	Maidment et al. (2017)	B
CHIRPS v2.0	IR, gauges CFS v2 TRMM 3B42 CHPClim	0.05° 0.25°	6-hourly Daily Dekadal Monthly	Africa 50°N–50°S	1981–present	Funk et al. (2015)	C
CMORPH_RAW v1.0	IR, MW NEXRAD	0.25° 8 km	3-hourly Daily, 30 min	60°N–60°S	1998–present	Joyce et al. (2004)	D
CMORPH_BLD v1.0	CMORPH_CRT v1.0, gauges	0.25°	Daily,	Global	1998–present	Xie and Xiong (2011)	E
WFDEI_CRU	ERA-Interim, CRU	0.5°	3-hourly Daily	Global	1979–2016	Weedon et al. (2014)	F
CPCU	Gauges	0.5°	Daily	Global	1979–present	Xie et al. (2010)	G

[a] Archive sources: A: https://pmm.nasa.gov/data-access/downloads/trmm; B: https://researchdata.reading.ac.uk/112/; C: ftp://ftp.chg.ucsb.edu/pub/org/chg/products/CHIRPS-2.0; D: ftp://ftp.cpc.ncep.noaa.gov/precip/CMORPH_V1.0/; E: ftp://ftp.cpc.ncep.noaa.gov/precip/CMORPH_V1.0/; F: ftp://rfdata:forceDATA@ftp.iiasa.ac.at/WFDEI_CRU/Rainf_daily_WFDEI_CRU/; G: ftp://ftp.cdc.noaa.gov/Datasets/cpc_global_precip/.

3. Multiobjective evaluation methodology

In addition to the conventional validation statistics (summarized in Table 2) such as PCC, RMSE, NSE, and FBias, the new methodology proposes a range of performance characteristics that are relevant for specific applications, such as flood warnings and crop insurances. The performance characteristics are (i) QPEs' ability to capture variability at small time scale without influence of seasonality; (ii) QPEs' detection of intensities and timing of extreme weather conditions in terms of dry/wet spell and moderate/heavy rainfall; (iii) QPEs' performance over different rainfall-only related climate zones (characterized by yearly totals, minimal monthly totals, and maximal monthly totals of precipitation) and different rainfall seasons associated with climate zones (characterized by monthly means instead of the four calendar seasons); and (iv) QPEs' performance on extreme weather detection over different regions characterized by climate patterns (e.g., climate zones) and topography (e.g., elevation). The last two aspects aim to make full use of the gauge data to achieve a thorough evaluation when gauges are scarce, and to improve the understanding of error sources (climate and topography) in the algorithms of QPEs. Code and tutorial of the implementation of this evaluation method can be found at https://surfdrive.surf.nl/files/index.php/s/cESQ5pjY6p8XSgi.

a. Validation statistics

Table 2 summarizes the verification statistics or performance scores, which were used in the point-to-pixel analysis of QPEs. RMSE, Bias, and PCC were used for comparison with existing studies. Bias gives the tendency of a specific QPE to overestimate or underestimate the rainfall rates. PCC reflects the linear correlation between gauge data and QPEs. Zero precipitation accumulations were included in these quantitative comparisons to account for both the evaluation on rainfall occurrence and on rain rates, although this could bias the results. Categorical statistics, probability of detection (POD), success rate (SR), frequency bias (FBias), and threat score (TS), were introduced to measure the ability of QPE to detect occurrence of rainfall based on a contingency matrix (Table 3) defined by the hits, misses, false alarms (FAs), and correct negatives (CNs). FBias gives

TABLE 2. Performance scores: O and $\overline{O}$ represent gauge (observation) data and mean of gauge data, respectively; E and $\overline{E}$ represent rainfall estimates and mean of rainfall estimates, respectively; and FA represents false alarm. Parameters l and m in re_PCC and re_NSE are indices for time steps at different temporal scales and station locations.

Performance scores	Formula	Unit	Range	Perfect value	Measure
Pearson correlation coefficient (PCC)	$\frac{\sum(O_i-\overline{O})(E_i-\overline{E})}{\sqrt{\sum(O_i-\overline{O})^2}\sqrt{\sum(E_i-\overline{E})^2}}$	—	$[-1, 1]$	1	Correlation between two datasets
Nash–Sutcliffe efficiency (NSE)	$1-\frac{\sum(E_i-O_i)^2}{\sum(O_i-\overline{O})^2}$	—	$(-\infty, 1]$	1	Agreement between two datasets
Root-mean-square error (RMSE)	$\sqrt{\frac{1}{N}\sum(O_i-E_i)^2}$	mm	$[0, +\infty)$	0	Magnitude of error
Multiplicative bias (Bias)	$\frac{\overline{E}}{\overline{O}}$	—	$[0, +\infty)$	1	Bias in rainfall intensities
Probability of detection (POD)	$\frac{\text{hits}}{\text{hits}+\text{misses}}$	—	$[0, 1]$	1	Accurate detection rate
Success ratio (SR)	$\frac{\text{hits}}{\text{hits}+\text{FAs}}$	—	$[0, 1]$	1	Accurate detection rate
Frequency bias (FBias)	$\frac{\text{hits}+\text{misses}}{\text{hits}+\text{FAs}}$	—	$[0, +\infty)$	1	Bias in rainfall frequency
Threat score (TS)	$\frac{\text{hits}}{\text{hits}+\text{FAs}+\text{misses}}$	—	$[0, 1]$	1	Accurate detection rate
Relative PCC (re_PCC)	$\frac{\sum_{l=1}^{L}\sum_{m=1}^{M}(O_{i,l,m}-\overline{O}_{i,l,m})(E_{i,l,m}-\overline{E}_{i,l,m})}{\sqrt{\sum_{l=1}^{L}\sum_{m=1}^{M}(O_{i,l,m}-\overline{O}_{l,m})^2}\sqrt{\sum_{l=1}^{L}\sum_{m=1}^{M}(E_{i,l,m}-\overline{E}_{l,m})^2}}$	—	$[-1, 1]$	1	Correlation between two datasets
Relative NSE (re_NSE)	$1-\frac{\sum_{l=1}^{L}\sum_{m=1}^{M}(O_{i,l,m}-E_{i,l,m})^2}{\sum_{l=1}^{L}\sum_{m=1}^{M}(O_{i,l,m}-\overline{O}_{l,m})^2}$	—	$(-\infty, 1]$	1	Agreement between two datasets

TABLE 3. Rainfall contingency table. A threshold value of 0.2 mm day^{-1} is chosen to separate rain from no-rain events. "Observed" represents gauge data and "estimated" represents rainfall estimates.

	Observed	
Estimated	Rain	No rain
Rain	Hit	False alarm
No rain	Miss	Correct negative

the bias of the frequency of detected rainfall occurrences for QPEs. POD, SR, and TS can be viewed as alternative measures of detection performance that are more sensitive to the timing of the QPE with respect to the reference measurement.

A Taylor diagram (Taylor 2001) was utilized to visualize three performance scores measuring the errors in the estimates of rain rates, RMSE, PCC, and standard deviation (SD), in a single diagram, following their relationship:

$$\mathrm{RMSE}_E^2 = \mathrm{SD}_O^2 + \mathrm{SD}_E^2 - 2\mathrm{SD}_O\mathrm{SD}_E\mathrm{PCC}_{O,E}, \qquad (1)$$

where O represents the reference (gauge) data, and E represents the estimates (QPEs).

A Roebber performance diagram (Roebber 2009) was used to visualize four performance scores measuring performance related to rainfall occurrences: SR, POD, TS, and FBias. These statistics can be combined in one diagram, since TS and FBias can be related to SR and POD by

$$\mathrm{TS} = \frac{1}{\dfrac{1}{\mathrm{SR}} + \dfrac{1}{\mathrm{POD}} - 1}, \quad \text{and} \qquad (2)$$

$$\mathrm{FBias} = \frac{\mathrm{POD}}{\mathrm{SR}}. \qquad (3)$$

b. Spatial and temporal representativeness

For comparison of spatial rainfall patterns, all datasets were regridded to 0.25° resolution, where WFDEI_CRU and CPCU data were downscaled and TMA station data were interpolated, both by inverse distance weighted method. CHIRPS and TAMSAT were simply aggregated, and others remained at their original resolution. Since no validation statistics were computed over these gridded data, more sophisticated interpolation methods (such as kriging method) were not deemed necessary.

Rainfall over Tanzania shows strong seasonality, in which case, one QPE could yield a good NSE or PCC score when it explains the seasonality, even it fails to reproduce smaller time-scale fluctuations. The same applies when the rainfall shows strong spatial variability. To account for spatial variability and seasonality, "zero-skill" models were established with respect to a season- and spatial-regime-neutral reference. These were the time series of dekadal, monthly, 3-monthly, 6-monthly, and yearly rainfall means, averaged over the period of study at each station. To measure the efficiency or accuracy relative to the zero-skill models, we revised NSE and PCC as follows:

$$\mathrm{re\_NSE} = 1 - \frac{\sum_{l=1}^{L}\sum_{m=1}^{M}(O_{i,l,m} - E_{i,l,m})^2}{\sum_{l=1}^{L}\sum_{m=1}^{M}(O_{i,l,m} - \overline{O}_{l,m})^2}, \qquad (4)$$

$$\mathrm{re\_PCC} = \frac{\sum_{l=1}^{L}\sum_{m=1}^{M}(O_{i,l,m} - \overline{O}_{l,m})(E_{i,l,m} - \overline{E}_{l,m})}{\sqrt{\sum_{l=1}^{L}\sum_{m=1}^{M}(O_{i,l,m} - \overline{O}_{l,m})^2}\sqrt{\sum_{l=1}^{L}\sum_{m=1}^{M}(E_{i,l,m} - \overline{E}_{l,m})^2}}, \qquad (5)$$

where l and m represent time step in zero-skill models and station location, respectively; $L = 12, 4, 2, 1$ for monthly, 3-monthly, 6-monthly, and yearly zero-skill model, respectively; and $\overline{O}_{l,m}$ and $\overline{E}_{l,m}$ are the monthly, 3-monthly, 6-monthly, or yearly mean values at time step l, location m. For instance, for monthly zero-skill model, $\overline{O}_{1,1}$ represents monthly mean of January averaged over 1998–2006, for TMA station "Arusha." The relative NSE (re_NSE) shows the improvement (value > 0) or degradation (value < 0) of QPEs over the zero-skill models. The relative PCC (re_PCC) shows the correlation between QPEs and gauge data accounted for the seasonality produced by zero-skill models.

c. Extreme weather detection

The skill of the QPEs to estimate the frequency, intensity, and severity of extreme weather is particularly relevant for applications related to drought or

TABLE 4. Description of the stations and data from Tanzania Meteorological Agency in the period of 1970–2006. The first four columns from the left are geographic information of the stations, and the other columns are precipitation statistics from the historical data. Definition of precipitation patterns can be referred to in Table 5.

Name	Longitude	Latitude	Elevation (m MSL)	Precipitation pattern	Annual total of precipitation (mm)	90P (mm day^{-1})	99P (mm day^{-1})	Gap_7%
Arusha	34°S	36.7°E	1372	Bimodal	716	5.8	37.4	11.4
Bukoba	1.3°S	31.8°E	1144	Monsoon	1856	18.8	52	9.4
DIA	6.9°S	39.2°E	53	Bimodal	1058	8.6	48.4	12.1
Dodoma	6.2°S	35.8°E	1120	Arid	559	1.6	38.4	10.3
Iringa	7.7°S	35.7°E	1428	Arid	610	4	30.8	12.9
Kigoma	4.9°S	29.6°E	885	Winter-dry	917	8.4	39.8	3.6
Mbeya	8.9°S	33.4°E	1759	Winter-dry	923	8.4	35	14.7
Morogoro	6.5°S	37.4°E	579	Bimodal	795	6.4	37.4	2.8
Mtwara	10.3°S	40.2°E	113	Winter-dry	1082	8.2	47.4	8.3
Musoma	1.5°S	33.8°E	1147	Bimodal	897	6.6	37	13.9
Mwanza	2.5°S	32.9°E	1140	Winter-dry	1091	9	43.4	2.8
Same	4.1°S	37.7°E	860	Bimodal	544	3	30.4	10.7
Songea	10.7°S	35.6°E	1067	Winter-dry	1008	9.2	45.2	7.1
Tabora	5.08°S	32.8°E	1265	Winter-dry	967	8	40.2	11.7
Tanga	5.1°S	39.1°E	9	Monsoon	1157	9.2	54.4	16.7
Zanzibar	6.1°S	39.2°E	18	Monsoon	1641	13.8	64	24.2

flood warning. The approach was to first determine the occurrence of extreme weather, and then use categorical statistics, such as FBias, TS, POD, and SR, to give a quantitative evaluation. Two types of conditions can be used to determine the occurrence: 1) count of consecutive days with rainfall rates above or below certain thresholds; 2) precipitation sums in a day or a period (such as 5-day or 10-day) above or below fixed thresholds.

First, the occurrences of consecutive dry day (CDD) and consecutive wet day (CWD) were defined. The minimum length of 10 days was chosen to be agriculturally (Sivakumar 1992; Barron 2004) and hydrologically critical, and length of 20 days was used to investigate the influence of time windows on the evaluation results. Threshold of 0.1 mm day^{-1} was used for dry days with respect to the usual precision of rain gauges, and 1 mm day^{-1} for wet days on the assumption that rainfall less than this amount is evaporated off directly (Mathugama and Peiris 2011). For the determination of CWD, a 5-day moving average was used instead of daily rainfall accumulation, which allows to better account for severe prolonged wet spells. The 5-day moving average of the current day is the mean of the previous 5 days in this study. The indicators of dry spell and wet spell are given by

CDD10/CDD20: the occurrences are counted of days in a period where there are at least 10 or 20 consecutive dry days with precipitation less than 0.1 mm;

CWD10/CWD20: the occurrences are counted of days in a period of 5-day moving average where there are at least 10 or 20 consecutive wet days with precipitation larger than 1 mm.

The four indicators are defined for nonoverlapping periods. Take CDD10 for example, if there is a period of 15 days without rainfall, then the 15 days all count as the occurrence of CDD10. If there is an intermittent wet day, the count starts anew.

Second, heavy rains were defined as precipitation above a certain precipitation percentile. This percentile definition was guided by the official recommendation of the World Meteorological Organization for extreme weather analysis (Klein Tank et al. 2009). The 90th and 99th percentiles (90P and 99P, respectively) were calculated for each station from historical TMA daily observations in the period of 1970–2006. The values of 90P and 99P are shown in Table 4, and can be viewed as thresholds for moderate rain and heavy rain, respectively.

For applications where the timing of the very heavy rainfall (99P) is not a big issue, such as for crop models to estimate potential yields, a lag of 3 days between the occurrence in observations and in the QPEs was allowed. A reference period of one week was applied around the occurrence in gauge data, with 3 days before and 3 days after this occurrence. The steps of the implementation of the ‘‘lagged detection’’ are 1) check the occurrences in the time series of a single gauge; 2) when an occurrence is detected, search the QPE data for an occurrence within its reference period; 3) pick the nearest one and move it to the day when occurrence happened in the gauge data; 4) repeat steps 1–3 for all gauges. In this way, values of FBias do not change, and duplicated detection for the occurrence (the hits) in QPE will be avoided, which could happen if simply counting the occurrence of a QPE in the reference

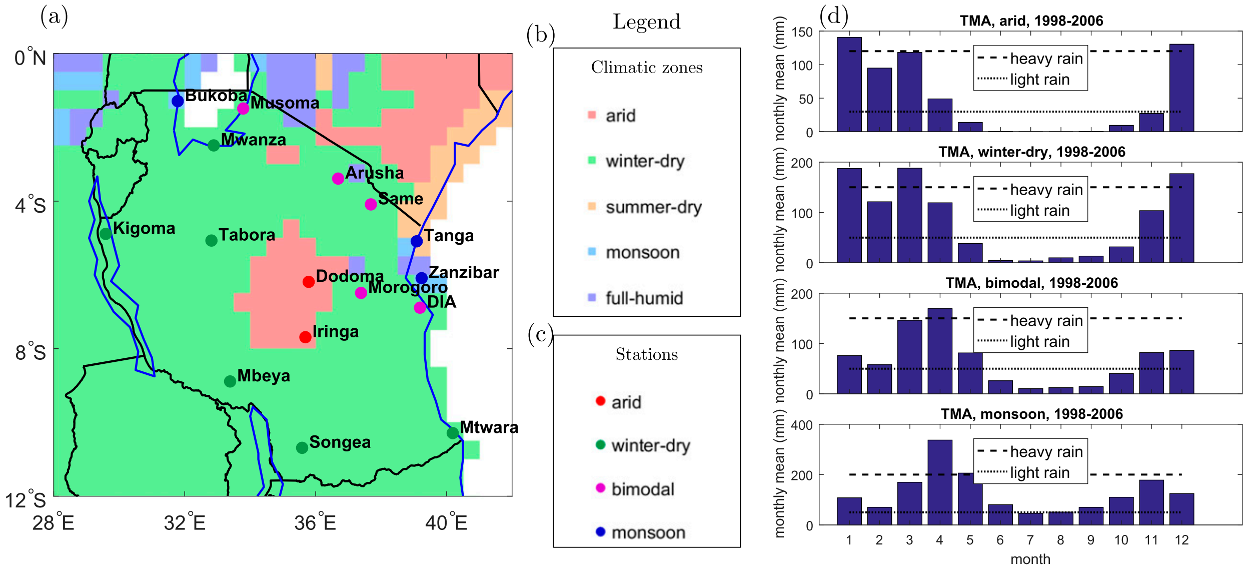


FIG. 2. Climatological information of Tanzania and the stations with (a) climate map of Tanzania, (b) legend of the climatic zones, (c) climatic classes of the stations, and (d) histogram of the monthly mean averaged over the stations in the class and over the period of 1998–2006.

period as a hit. The lag value could be chosen such that only a limited number of occurrences in the QPEs will be moved back and forth. To select an appropriate lag value, a criterion, Gap_n with $n = 2 \times \text{lag} + 1$, was defined to be the percentage of multiple occurrences of 99P within a period of n days with respect to all 99P events, with 0 being the perfect value. A lag of 3 days was selected for our case as most occurrences of 99P in the observation dataset have a gap or intermission time larger than 7 days based on the values of Gap_7 in Table 4.

d. Regional and seasonal classification

To evaluate the performance of QPEs under various morpho-climatic conditions, a classification was applied to distinguish between different seasons, climatological and topographical regions that potentially influence the rain-producing mechanism and the quality of the QPE products. A climate zone classification and seasonal classification related to precipitation were developed, which is more specific than the common way of applying Köppen–Geiger method (see the following paragraph) for regional classifications and using the four calendar seasons for seasonal classification. Rainfall seasons should be classified in relation to climate zones, instead of using a uniform classification over the whole region. For instance, a dry season should be defined in summer over regions with a summer-dry climate and in winter with a winter-dry climate.

1) CLIMATOLOGICAL CLASSIFICATION

The domain was divided into five climatic zones according to the Köppen–Geiger classification (Peel et al. 2007). This method uses three parameters to characterize climate zones: main climate, precipitation pattern and temperature pattern (optional). Main climate is more commonly used in previous studies (Beck et al. 2017), however, the precipitation pattern is more closely related to evaluating the performance of QPEs or analyzing the sources of errors for QPEs. To decrease the number of zones for a gauge-sparse environment, in this paper a classification based only on precipitation pattern was used. Figure 2a shows the Tanzania climate map consisting of five climatic zones: "arid," "winter-dry," "summer-dry," "monsoon," and "full-humid." The map was generated using the dataset from digital world map of Köppen–Geiger classification at 0.5° resolution covering 1976–2000, and each zone in the map includes multiple of Köppen climate classes. For instance, the winter-dry zone includes Köppen climates of Cwa, Cwb, and Aw. None of the stations are located in summer-dry or full-humid zone and most of the stations are located into the winter-dry zone. Figure 2d shows the histograms of the monthly mean averaged over stations in the same climatic zone over the period of study (1998–2006). Based on the precipitation patterns, stations classified as winter-dry were further divided into a winter-dry pattern with one continuous rainy season during November until May in a year, and a "bimodal" pattern with two rainy seasons in a year. As a result, there are four climatic classes for stations and five over the whole domain as depicted by Figs. 2c and 2b.

2) SEASONAL CLASSIFICATION

First, two main seasons over the whole domain were determined independent of local climatological behavior, in order to assess the overall performance for the spatial

TABLE 5. Seasonal classification.

Climate zone	No. of stations	Threshold for dry|light rain (mm month^{-1})	Threshold for light|heavy rain (mm month^{-1})	Dry months	Light rain months	Heavy rain months
Arid	2	30	120	May–Nov	Feb–Apr	Jan and Dec
Winter-dry	6	50	150	May–Oct	Feb, Apr, and Nov	Jan, Mar, and Dec
Bimodal	5	50	150	Jun–Oct	Jan–Mar, May, Nov, and Dec	Apr
Monsoon	3	50	200	Jul	Jan–Mar, Jun, and Aug–Dec	Apr–May

distribution of rainfall in different seasons. The wet season covers the months from November to April, and dry season covers from May to October. This classification is generated according to the monthly mean precipitation pattern over all stations.

A more detailed seasonal classification was made based on the rainfall patterns in different climate zones for further investigation of QPE sensitivity to rainfall seasonality. Three seasons: dry season, light rain season, and heavy rain season, were distinguished based on thresholds varying between climate zones, as shown in Fig. 2d. Details including the number of stations in each climate zone, threshold values for the classification, and months included in each seasonal class are given by Table 5.

3) TOPOGRAPHICAL CLASSIFICATION

To investigate the topographical influence on the QPEs and the capability of QPEs to capture orographic rainfall, the stations were divided into three classes: low altitude (<1000 m MSL) mainly located in the coastal plain, medium altitude (1000–1300 m MSL) located on the central plateau, and high altitude (>1300 m MSL) located in the north–south highlands, as illustrated in Figs. 1b and 1c by dots with different sizes.

4. Results and discussion

This section illustrates the use of the proposed method to evaluate, compare, and select QPEs with six satellite-based and reanalysis datasets as case study, with a gauge-based product as additional reference to TMA gauge data. We discuss how to use, modify and improve the evaluation method, with recommendations of the six QPEs, for different applications in sections 4a–4f. Factors that influence the reliability of evaluation results, and sources of errors in the estimation algorithms of the QPEs are discussed in section 4g.

a. *Overall performance*

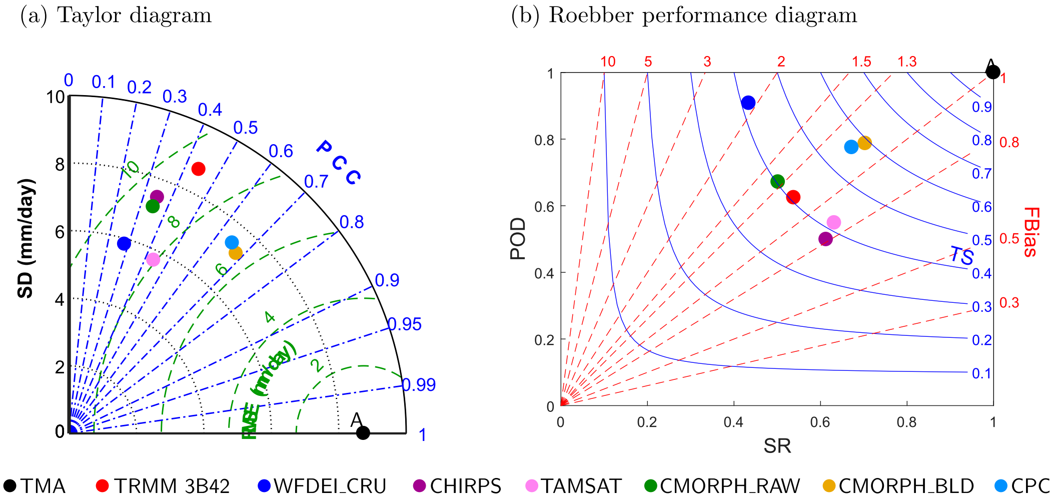
Taylor and Roebber performance diagrams provide a quick overview of skills of QPEs at rain rates and rainfall occurrences, respectively, over the whole domain. In the two diagrams, "A" (black dot) is the observation (gauge) data being treated as the reference QPE. Hence, its RMSE = 0 mm day^{-1}, PCC = 1, SR = 1, POD = 1, TS = 1, FBias = 1, and its SD = 8.73 mm day^{-1}. The Taylor diagram in Fig. 3a illustrates the values of RMSE, PCC and SD for the six QPEs. In the diagram, PCC is related to the polar angle (blue line), SD is proportional to the radial distance from the origin (black arc centered at origin) and RMSE is proportional to the radial distance from A (green arc centered at A). PCC < 0.5 implies low correlation between QPEs and gauge data. SD represents the variability of rainfall in a given QPE. The Roebber performance diagram in Fig. 3b visualizes values of SR, POD, TS and FBias. In the diagram, SR and POD are represented on the x axis and y axis, respectively, FBias is represented by the slope value in red dashed lines and TS by the blue solid contours centered at A. SR < 0.75, POD < 0.75, and TS < 0.7 can be viewed as bad timing with too many false alarms, misses and inaccurate rainfall occurrences in a QPE, respectively. FBias > 1.25 or <0.8 represents large over or underestimation of rainfall frequencies, respectively.

Observed from Fig. 3, all QPEs perform poorly at daily rainfall estimation of both the intensities and occurrence. Figure 3a shows all datasets except CMORPH_BLD and CPCU are poorly correlated to gauge data, with PCC scores < 0.45. All QPEs except TRMM 3B42 underestimate rainfall variability, as indicated by SD values lower than SD of the observation dataset (A). Figure 3b shows all the QPEs have difficulties to determine the timing of rainfall, with best TS scores given by CMORPH_BLD to 0.59 and other TS scores < 0.45. In addition, WFDEI_CRU and CMORPH_RAW significantly overestimate the frequency of rainfall, where WFDEI_CRU estimates occurrences of precipitation more than twice as high as observed by the gauges.

It can be noticed that CMORPH_BLD and CPCU have very similar performance for the seven performance scores, while CMORPH_BLD is slightly better for all aspects.

b. *Spatial pattern of precipitation*

First, we check whether the climatology obtained by different datasets are spatially consistent to the gauge-based rainfall fields (TMA and CPCU), or to each other in this


(a) Taylor diagram

(b) Roebber performance diagram

FIG. 3. Diagrams to visualize the performance scores with (a) Taylor diagram to illustrate RMSE in the green curve centered at ''A'' (black dot), PCC in the blue line, and SD in the black curve centered at the origin, where RMSE measures the magnitude of error, PCC the linear correlation, and SD the rainfall variability and (b) Roebber performance diagram to illustrate SR on the *x* axis, POD on the *y* axis, TS in the blue curve centered at ''A,'' and FBias in the red dashed line, where SR, POD, and TS measure the ratio of accurate detection of rainy days with respect to misses, false alarms, and both, respectively, and FBias is the bias in the frequency of rainfall.

gauge-sparse environment with complex topography. Second, critical regions can be identified where the QPEs differ the most from each other to provide advisory information for optimal location of new gauges, which could be used for both calibrating rainfall models and assessing QPEs.

Figure 4 shows spatial precipitation patterns of the regridded QPEs at 0.25° resolution over Tanzania, for mean annual and seasonal (wet/dry) precipitation. All datasets capture the general northeast–southwest precipitation gradient over Tanzania, which is caused by the north–south movement of the ITCZ and the influence of the highlands, mainly the Tanzania southern highlands. Compared to TMA and CPCU, the other six QPEs overestimate rainfall over the west to south Tanzania in wet seasons (see the third and bottom rows of Fig. 4). However, the rainfall intensities of different datasets vary significantly spatially and no pair of QPEs show a similar pattern.

Critical regions where QPEs show large differences are areas along coastlines and around Lake Victoria (blue circle) in the dry seasons (see the second and the fifth rows of Fig. 4): northeast of the southern highlands, the Lake Victoria region, and the southwest part of the central plateau in wet seasons (see the third and bottom rows of Fig. 4). From maps of yearly totals (the top and fourth rows of Fig. 4) we can see peaks around Mount Kilimanjaro covered by snow (green circle) and Lake Eyasi (cyan circle), over which the error of different rainfall estimation algorithms should be investigated.

Note that rainfall is a phenomenon characterized by high variability both in space and time. Number and distribution of gauges, spatial correlation, and interpolation methods (IDW method for TMA and topography-based method for CPCU) can impact the resulting rainfall fields. This may be the reason for the dependency between spatial rainfall patterns in the interpolated TMA data and in CPCU, which makes it difficult to determine which QPE performs better at nongauge locations.

c. Performance dependency on temporal scale

To identify at what temporal scale a QPE is usable, how it performs compared to zero-skill models defined in section 3b, and what their ability is to capture small-scale variability and seasonality (e.g., temporal scales larger than a month), re_NSE and re_PCC scores are developed with their values given by the six datasets shown in Fig. 5. In the figure, *x* axis represents temporal scales of zero-skill models, and *y* axis represents temporal aggregation scales for QPEs. re_NSE < 0 means the performance of a QPE at *y*-axis scale is not better than using gauge data at *x*-axis scale, and re_NSE > 0 means sufficient quality of the QPE with respect to both correlation and magnitude of error. re_PCC in a column can be used to diagnose the influence of aggregation scales on QPE performance, and re_PCC in a row to

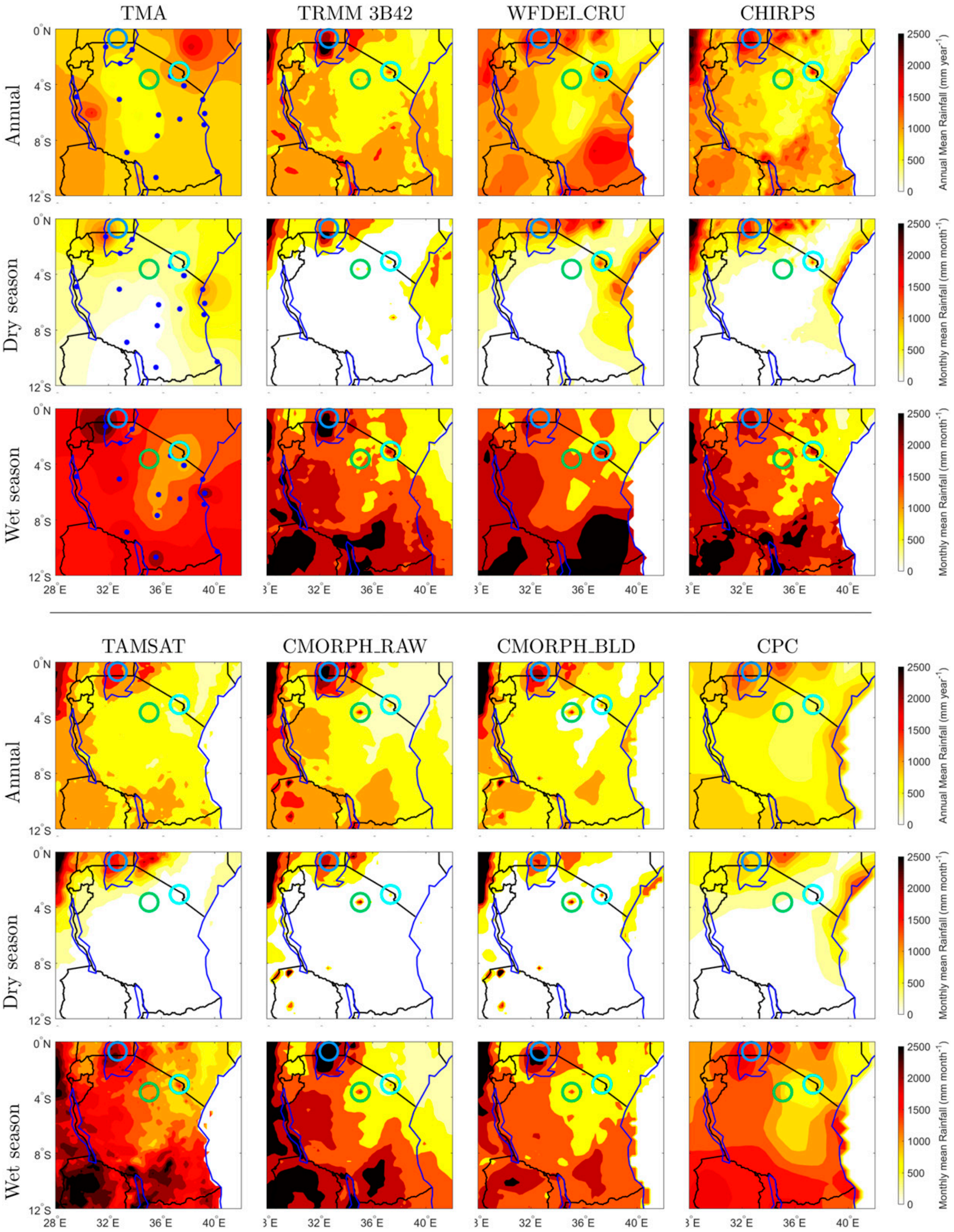


FIG. 4. Spatial distribution of the mean annual rainfall, monthly rainfall over the dry season (May–October), and monthly rainfall over the wet season (November–April) for the eight datasets, averaged over 1998–2006, regridded to 0.25° resolution. Blue, green, and cyan circles represent locations of Lake Victoria, Mt. Kilimanjaro, and Lake Eyasi, respectively.

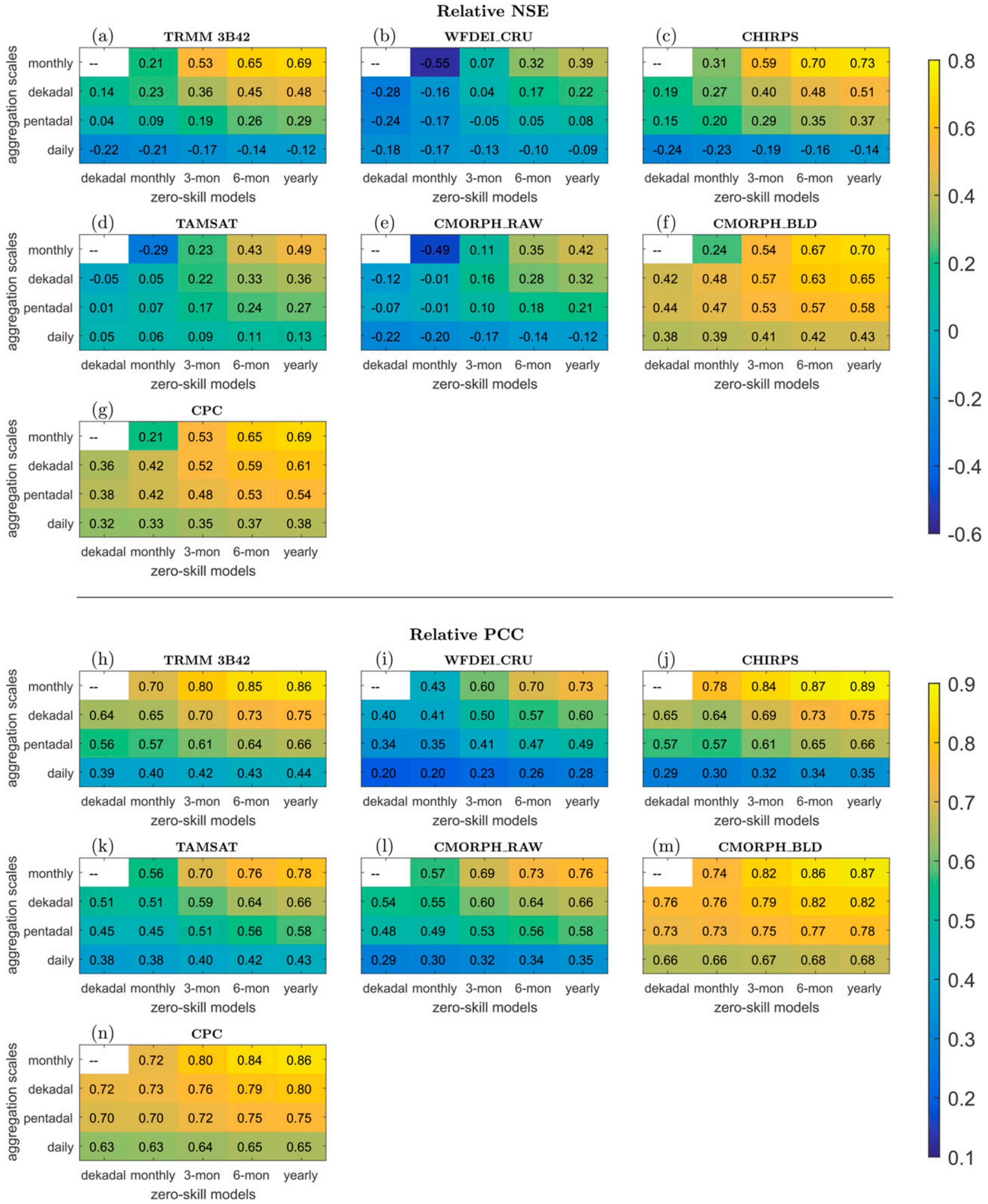


FIG. 5. Re_NSE and re_PCC scores of the seven QPEs compared to TMA data on a daily, pentadal, dekadal, and monthly basis, with respect to climatology (zero-skill models) at dekadal, monthly, 3-monthly, 6-monthly, and yearly scales. Values of performance scores show the QPE performance to capture rainfall variability at *y*-axis scales by removing the influence of seasonality at *x*-axis scales.

diagnose how the variance in model error compared to seasonality.

From Fig. 5 we can observe, in general, that the performance of QPEs increases with aggregation over larger time periods and/or at larger scales. CMORPH_BLD shows the best agreement of gauge data at all temporal scales. By looking at rows of table we can observe that, at daily scale, re_NSE values are negative for four out of the six QPEs, that is, these estimates are worse than using monthly or even yearly mean of gauge data. At 5-day to monthly aggregation scale, QPEs are slightly better, but two QPEs are still worse than monthly gauge rainfall. Correlation improves for all products with the removal of the influence of larger-scale variability, but the improvement is larger at larger aggregation scale, as indicated by rows of re_PCC values. As shown by values in the bottom rows, comparing daily re_PCC corresponding to monthly versus yearly zero-skill models, skill improves only little, which means the QPES can hardly capture small (daily) scale variability, and the uncertainty is too large to effectively represent seasonality.

A modest requirement is that both re_NSE > 0 and re_PCC > 0.5 for a product to be viewed as acceptable. In this case, all seven QPEs except, CMORPH BLD and CPCU, perform poorly at daily scale. Even at pentadal scale, only TRMM 3B42, CHIRPS, CMORPH BLD and CPCU are usable. In addition, the small re_NSE values (<0.5) of CPCU at daily scale indicate high level of independence of the TMA data as a reference dataset.

d. *Performance related to climate regime*

Rainfall-related seasonal classification in relation to climate regions can be used to understand sources of error, and help to objectively select QPEs for a certain region or season, and give instructive information on adjustment of model parameters in QPE algorithms. Figure 6 shows the performance statistics in different regional and seasonal categories by the spider charts for PCC, Bias, FBias, and TS for rainfall at daily scale. POD and SR can also be used for certain applications where misses and false alarms play an important role, and RMSE can be used in combination with Bias to diagnose the amount of error due to bias in the datasets, which are not included in this section.

In Fig. 6 the asymmetry in the spider charts indicates unbalanced scores across climate zones, while star-shaped patterns indicate poor seasonal performance. This uneven distribution of values in different categories indicates significant influence of climatology and seasonal factors on the quality of QPEs. We can observe that regional and seasonal performance of CMORPH_BLD and CPCU closely resemble each other. Figure 6a shows that CMORPH_RAW and TAMSAT perform better over the winter-dry, bimodal, and monsoon regions, while CHIRPS shows better performance for wet seasons. TRMM 3B42, CMORPH_BLD, CPCU, and WFDEI_CRU tend to have a more uniform behavior regardless of the impact of climate and seasons.

Figures 6b and 6c illustrate the bias in the estimation of rainfall intensity and frequency, respectively. Generally, the seasonal dependencies are stronger than the differences across climate zones. In dry seasons, TAMSAT and CHIRPS tend to underestimate the frequency and/or intensity of rainfall over the whole domain, while TRMM 3B42, CPCU, and CMORPH products overestimate them over winter-dry and arid regions. For wet seasons, CMORPH products and CPCU underestimate intensity and overestimate the frequency of rainfall. TAMSAT and CHIRPS have a better performance on rainfall intensity estimation but uneven performance on frequency estimation over different regions. TRMM 3B42 and WFDEI give a better estimation in heavy rain seasons than in light rain seasons. WFDEI_CRU overestimates the frequency of rainfall by more than a factor of 2, while its bias in rainfall intensity over the wet seasons is the smallest. This implies that smaller rainfall events are produced more frequently in wet seasons compared to gauge observations.

For the timing of rainfall illustrated by Fig. 6d, performance in dry seasons is much worse than that in the two wet seasons over each climate zone for all QPEs. Combined with the FBias values, this indicates that the ability of the QPEs to capture rainfall occurrences improves for seasons with higher rainfall frequency.

The seasonal classification is also useful for agricultural applications, with crop grown mainly in the light rain and heavy rain seasons.

e. *Detection of dry and wet spells*

Dry and wet spell defined by CDD10, CDD20, CWD10, and CWD20 can be important weather conditions for many applications, for instance, in the context of weather index insurance for agriculture where payouts to smallholder farmers are determined by a proxy of crop yield loss, such as precipitation. The evaluation of QPEs on dry/wet spell detection over different regions, classified by climate patterns or topography, can be helpful to select proper models as input to crop yield models for different purposes. Dry spells can be used for drought detection, which is a frequently occurring phenomenon in Tanzania. Wet spell is critical to monitor and predict the crop growing process.

Performance of QPEs averaged over the domain at daily scale is expressed in terms of POD, SR, FBias, and TS in the Roebber performance diagram, and performance over different regions classified by climate and by

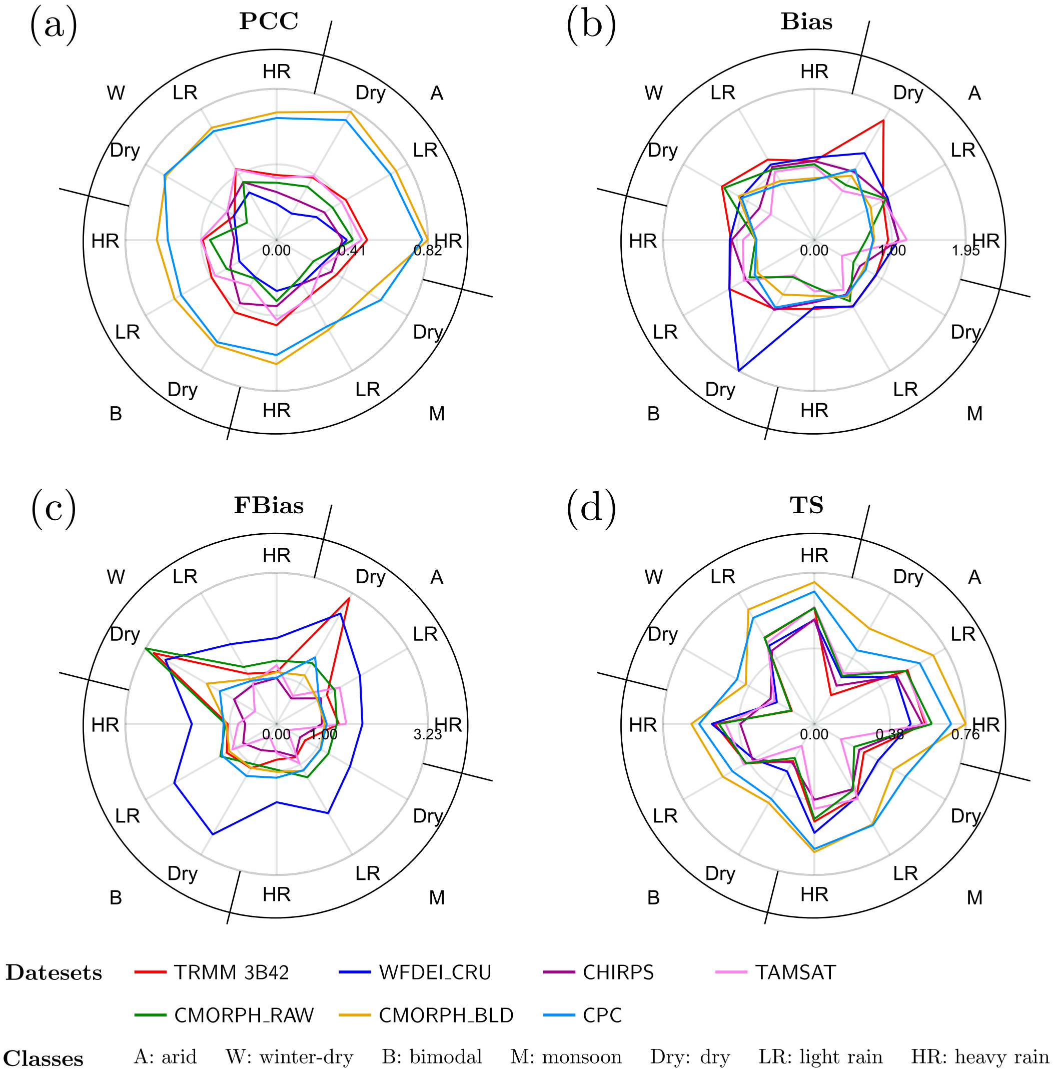


FIG. 6. (a)–(d) Spider charts of the validation statistics for datasets at daily scale in different seasons represented by Dry, LR, and HR in the inner circle, which are defined over different climatic zones represented by A, W, B, and M in the outer circle.

elevation is shown in spider charts in Fig. 7. This figure shows that TAMSAT and CHIPRS have a bias to overestimate the number of CDD10 events, while the other products tend to underestimate. The same applies for CDD20 events, with only CHIRPS and CMORPH_RAW having a slight improvement on the frequency of events and others becoming worse. Figures 7b and 7c show that all datasets except for TRMM 3B42 tend to perform best (with best FBias and TS scores) on CDD10 over arid region categorized by very little rainfall in the dry period. For topography classes, all satellite-based QPEs except TRMM 3B42 tend to perform better for higher altitudes, based on the FBias and TS scores. If one selects a QPE which achieves a minimum TS of 0.7 with 70% of droughts being identified, for 10-day dry spell detection with acceptable FBias of 0.8–1.25. Then only CMORPH_BLD and CPCU can be possibly used in all regions and climate zones. CHIRPS and TAMSAT are good for arid and winter-dry, while some of the other products are restricted to use over the arid zone. Figures 7e and 7f show a decrease of the QPE's performance on the timing of CDD20 events compared to CDD10 over monsoon region, especially for TAMSAT and CHIRPS since they produce a significant increase of overestimation (by around 1 and 0.5 times, respectively).

For CWD10 detection, Fig. 7g shows that CMORPH_BLD and CPCU underestimates the occurrence, while TRMM 3B42, TAMSAT, CHIRPS, and especially WFDEI_CRU overestimate. Figures 7h–i show that the performance of TRMM 3B42, CHIRPS, TAMSAT and CMORPH_RAW

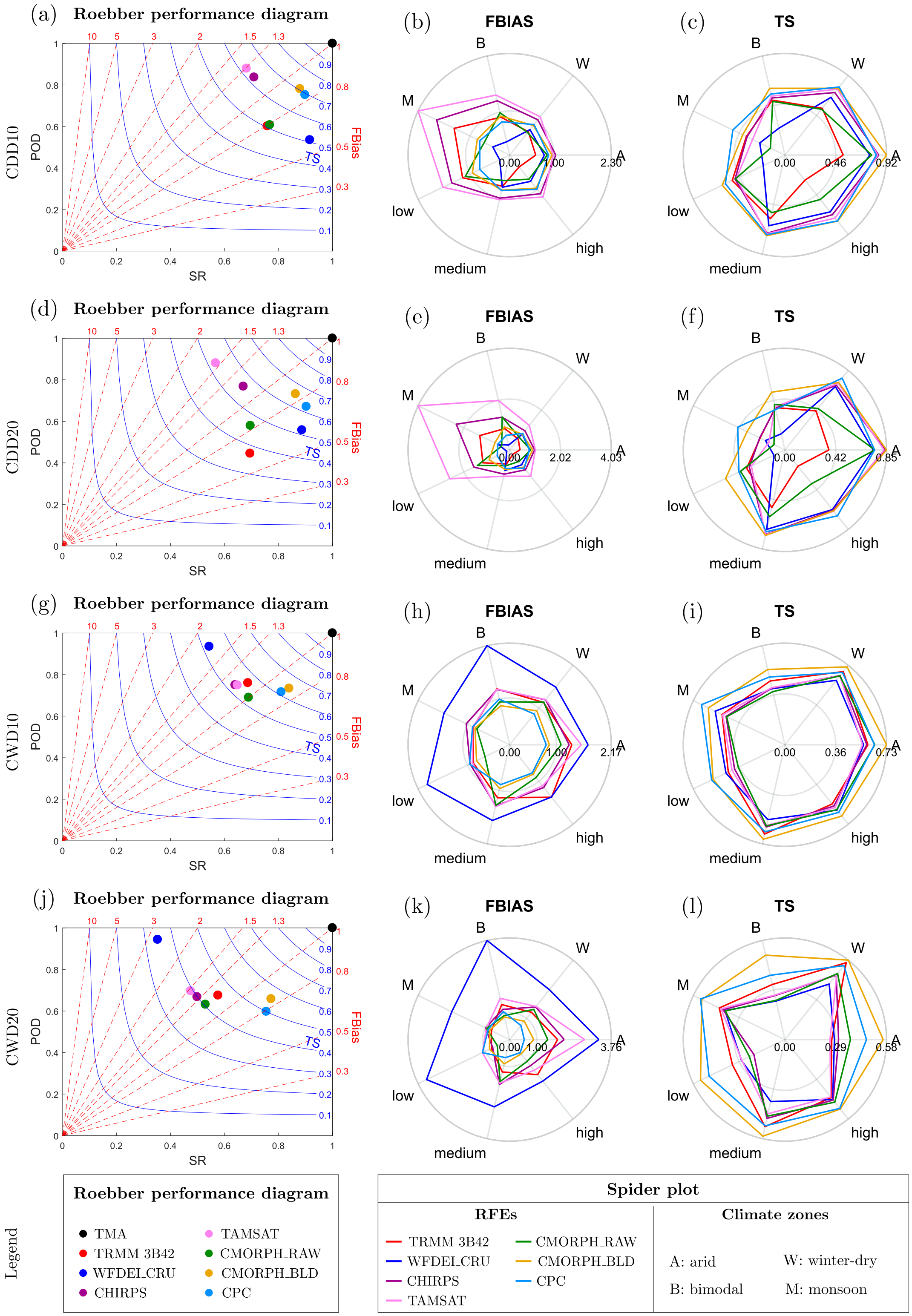


FIG. 7. Roebber performance diagrams and spider charts for the detection of (a)–(c) CDD10, (d)–(f) CDD20 (dry spell), (g)–(i) CWD10, and (j)–(l) CWD20 events (wet spells) at daily scale by the seven products. Spider charts show their performance over different climates zones represented by A, W, B, and M, and over different topographical regions represented by low, medium, and high.

are at the same level. Compared to CDD10 detection, performance is less influenced by climate, as shown by the more evenly distributed scores over climate zones in Figs. 7e and 7f, with slight underestimation of the frequency over monsoon regions and overestimation over arid regions. All satellite-based QPEs tend to give the largest FBias and TS values at medium altitudes. With TS > 0.7 and $0.8 <$ FBias < 1.25, only CMORH_BLD and CPCU can be used over arid and winter-dry regions. Figures 7k and 7l show that for CWD20 detection, all data except CMORPH_BLD and CPCU has an increase of overestimation compared to CWD10 detection and consequently a decrease of timing, over winter-dry and especially arid regions.

In general, the datasets perform worse with increase of time window of consecutive dry/wet days, with increased overestimation or underestimation. This means they have a tendency to exhibit persistence of dry and wet days.

The evaluation approach can be easily adapted to evaluate and select the QPEs for various applications, by determining and modifying the thresholds for minimum acceptable performance. Threshold of drought, for example, the length of consecutive dry days, can be flexibly adjusted based on drought vulnerability of a given crop and local soil characteristics. One could choose to focus the evaluation only over growing seasons of that crop instead of the whole year. In addition, high SR means few false alarms in drought detection, which implies the insurance company will have a limited economic loss due to paying farmers for false alarms of drought. High POD implies farmers have a good probability of avoiding loss. SR can be translated into a risk of loss for the insurers, and POD for the farmers. With respect to wet spell detection, both the minimum length of CWDs and the threshold for wet-or-dry day determination could be adjusted depending on sensitivity of the growing phase of a certain crop. In the definition of wet spell, we used 5-day moving average, where consecutive wet days and a heavily rainy day followed by several dry days are assumed to make not much difference on a crop growing process. The use of n-day moving average could also be modified according to the user's needs.

f. Detection of moderate and heavy rainfall

Moderate and heavy rainfall, characterized by 90th and 99th percentiles of daily rainfall in this study, are crucial in flood-related applications. Threshold values are different for each station as shown in Table 4. The numbers of days with rain rates above the thresholds (number of 90Ps and 99Ps), however, are similar.

Figures 8a–c illustrate the ability of QPEs to the capture of moderate rains (90th percentile). All QPEs overestimate the number of moderate rains, except the two CMORPH products. The timing of QPEs is poor, except CMORPH_BLD, with more than 50% misses and more than 50% false alarms, indicated by POD ≤ 0.5 and SR ≤ 0.5, respectively. FBias and TS values increase with smaller 90th percentile values for a given climate zone. For instance, all QPEs, except CMORPH_BLD, overestimate the number of 90Ps over arid zones but underestimate that over monsoon zones. With respect to the topographical impact, TRMM 3B42, TAMSAT, and CHIRPS perform the best at low altitude, while CMORPH_BLD and WFDEI_CRU respectively underestimate and overestimate for all altitudes.

The second and third rows of Fig. 8 exhibit the detection and lagged detection of 99Ps. All datasets except for TRMM 3B42 significantly underestimate the frequency of heavy rains, and the performance varies strongly across the climatological and topographical regions, indicated by spikes in the spider charts. The TS scores of the lagged detection (Fig. 8i) are 2–3 times better than regular detection (Fig. 8f) for all datasets, except for CMORPH_BLD which shows less improvement on the timing. This indicates that most of the 99P occurrences in the QPE products do not coincide with gauge observations, but are located within a one-week period around them. The absolute mean values in Table 6 show that CMORPH_BLD and TAMSAT have smaller lag time than the other products. The mean values of lag time in Table 6 show that satellite-based QPEs tend to give delayed detection of 99Ps while the reanalysis dataset (WFDEI_CRU) produces earlier detection.

In general, none of the QPEs are recommended if the selection criteria are $0.8 <$ FBias < 1.25, TS > 0.6, and POD > 0.7 (e.g., less than 30% misses), with CMORPH_BLD acceptable only over arid regions.

For flood warning systems, the thresholds of heavy rainfall can be adjusted for flood triggering over a specific region at different levels, instead of using 90th or 99th percentiles. An adjustment in the temporal scale could also be made, for instance, instead of daily rainfall, one could use rainfall accumulation over several days as the thresholds. This would limit applications to large systems characterized by long response times. For urban drainage systems, QPE performance needs to be satisfactory at subdaily scales. Upper and lower boundary of the lag time for lagged detection could also be adapted, for instance, to the range of $[-3, 0]$ to exclude QPE which would trigger delayed flood forecast. Acceptable FBias, SR or POD score ranges can be determined by the user's inclination to less false alarms or less misses when choosing among the QPEs.

g. Evaluation reliability and uncertainty in the QPEs

The outcomes of this study are similar to that from other literature. For instance, QPEs perform poorly

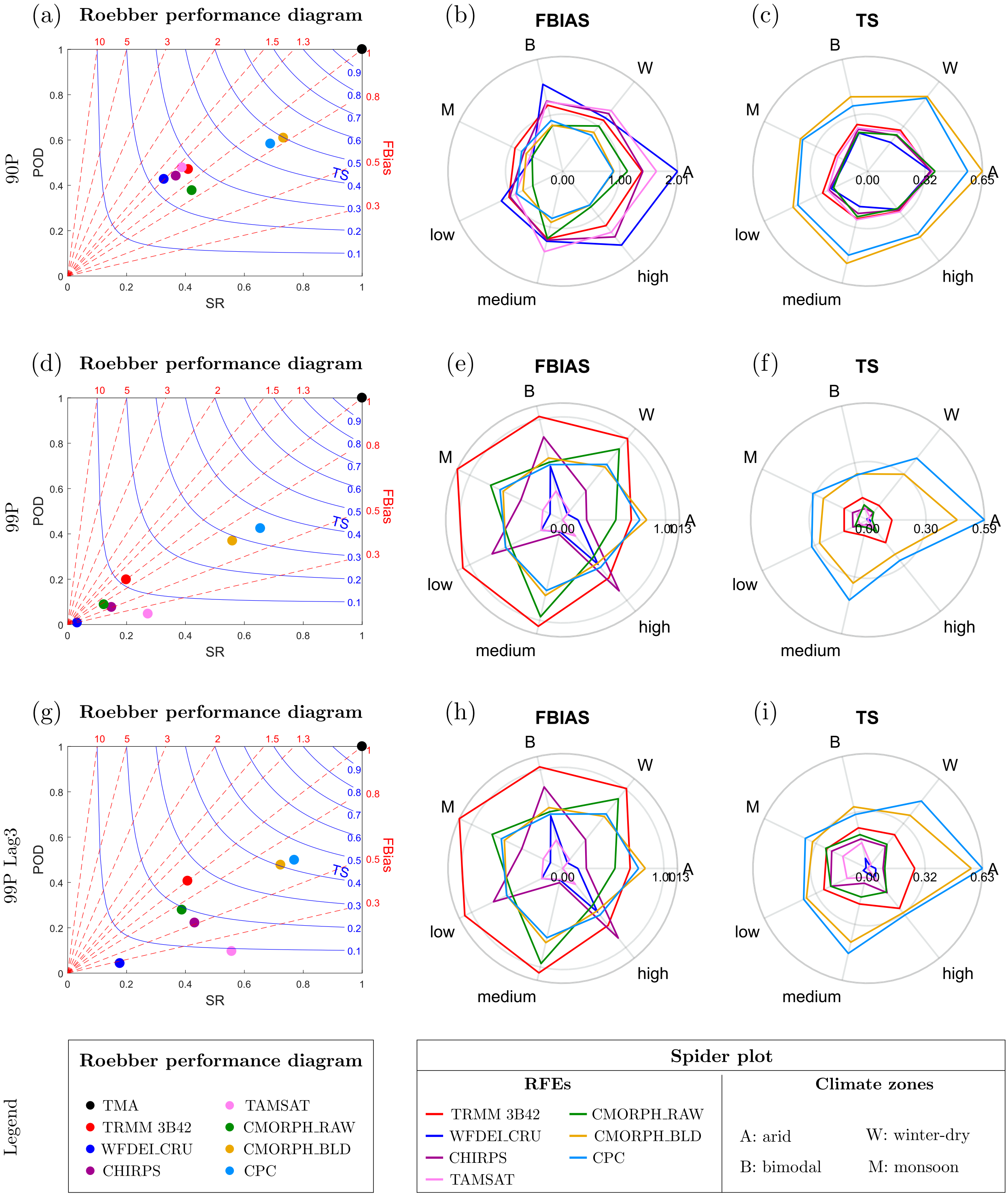


FIG. 8. Roebber performance diagrams and spider charts for the detection of daily rain rates above the thresholds of (a)–(c) 90th percentile, (e)–(f) 99th percentile, and (g)–(i) lagged detection for 99th percentile by the seven products. Spider charts show their performance over different climates zones represented by A, W, B, and M, and over different topographical regions represented by low, medium, and high.

TABLE 6. Lag time of the detection for 99P events over different regions classified by climate and topography as well as the whole region of study.

		Arid	Winter-dry	Bimodal	Monsoon	Low	Medium	High	Tanzania
Mean (day)	TRMM 3B42	0	0.30	−0.08	0.33	0.22	0.21	0.08	0.15
	WFDEI_CRU	1	−0.75	0.21	−2	0	1	−0.4	−0.11
	CHIRPS	1	0.10	0.36	0	0.17	0.23	0.29	0.26
	TAMSAT	0	−0.6	0.3	0.3	0.82	−0.13	−0.05	0.06
	CMORPH_RAW	1	0.2	0.15	0.58	0.44	0.04	0.34	0.38
	CMORPH_BLD	0.15	0.13	−0.03	0.14	−0.04	0.09	0.14	0.1
	CPC	0.05	0.35	0.52	0.40	0.46	0.26	0.4	0.39
Absolute mean (day)	TRMM 3B42	0.71	0.90	0.66	0.56	0.57	0.55	0.88	0.73
	WFDEI_CRU	1	1.75	1.5	2	0	1.67	1.47	0.89
	CHIRPS	1.5	0.87	0.93	0.70	0.74	1	1.03	0.90
	TAMSAT	0	0.6	0.57	0.9	1.36	0.38	0.37	0.45
	CMORPG_RAW	1	1.08	0.8	1.0833	0.76	0.92	1.09	0.91
	CMORPH_BLD	0.21	0.27	0.42	0.26	0.25	0.44	0.29	0.36
	CPC	0.05	0.35	0.52	0.40	0.46	0.26	0.4	0.39

over regions of Africa. Similar correlation values were shown by Dinku et al. (2007) over East Africa's complex topography compared to gauge data averaged over 0.25°–2.5° grid boxes at dekadal and monthly scales, and by Maidment et al. (2013) over Uganda (near Tanzania), who compared kriged averaged gauge data to satellite grids at dekadal scale. Beck et al. (2017) evaluated 22 QPEs globally, where 13 nongauge QPEs resulted in low PCC values over Africa when compared to gauge data at 3-day, monthly and 6-monthly scales, and the other 9 gauge-corrected QPEs gave low NSE values over Africa when compared to catchment data using hydrological modeling at 3-day scale. Another example is that QPEs tend to underestimate heavy rainfall events, in terms of frequencies of 99Ps in our study, and in terms of maximal rainfall intensities per year shown by Thiemig et al. (2012) over basin regions.

There are several factors that influence the reliability of the evaluation.

(i) Due to the scarcity or the lack of gauges in critical regions, such as Mount Kilimanjaro, Lake Eyasi, and the coastal plain, the spatial distribution of precipitation over these regions is poorly represented by the interpolated gauge observations. The abrupt change of the terrain structure over these regions increases the spatial variability of rainfall patterns, which makes it problematic to draw reliable conclusions using the current station network.

(ii) The reference gauge data are not totally independent from the datasets that are evaluated, therefore, all satellite products may not be consistently assessed in this study. The weight and amount of the gauge data used in these products influences the evaluation results, but not necessarily the accuracy of the products.

(iii) The scale mismatch, spatial mismatch, and temporal mismatch were not accounted for in the point-to-pixel analysis. The spatial integration of satellite estimates could smooth extremes. Averaging or interpolating point gauge data to QPE grids, or downscaling QPEs to smaller scales can be a solution, however, this may introduce sampling errors.

We have chosen to use the point scale as reference for the evaluation analysis, since the scale of many user applications is closer to the point than satellite pixel scale (e.g., crop field, city neighborhoods). The mismatches of the start of day between the gauges and datasets, and diurnal rainfall are not accounted for in this study. There is a consistent time differences between the definitions of ''day'' for the various datasets: TMA gauge data are at UTC + 0 h, TRMM 3B42 at UTC + 1:30 h, WFDEI_CRU at UTC − 3 h, CHIRPS at UTC + 0 h, TAMSAT at UTC + 6:15 h, CMORPH products at UTC + 6 h, and CPCU at UTC + 0 h. The mismatches between independent gauges and QPEs could lead to unreliable evaluation conclusions, especially over regions with frequent and heavy rainfall at subdaily time scales. For gauge data that are merged into QPEs, impact of diurnal rainfall is already included in the gauge-satellite calibration process.

There are several sources of error that can be found in reanalysis data.

(i) The areal-average effect can be found in WFDEI_CRU and probably in other reanalysis datasets. Its physical model, ECMWF, uses a mass-flux scheme to propagate convection, cloud, and resolved variables. The average (resolved) of temperature, humid, and cloudiness over the large grid box (e.g., 0.5° for WFDEI_CRU), might lead to the underestimation of rain rates in most rainfall events or, in this study, the underestimation of the frequency of 99P events (that are defined based on historical gauge records).

(ii) The lack of ground observations to calibrate the parameters of the convection schemes in the tropics and the lack of other rain-production processes in the NWP model may cause the low skill of WFDEI_CRU.

(iii) Another cause of the poor performance of WFDEI_CRU lies in its initialization strategies. It assimilates the factors that influence the formation of rainfall, such as humidity, temperature, and total column water vapor, instead of the rainfall itself. This results in relatively low correlation with daily rainfall observations. Therefore, although WFDEI_CRU, or its precursors—ECMWF and ERA-Interim—perform well in some gauge-dense countries, for instance, some of the European countries (Kidd et al. 2012; de Leeuw et al. 2015), indicating a good parameterization over these regions, this parameterization is not suited everywhere and requires adjustment for other regions.

The performance of the satellite-based QPEs is not related to their spatial resolutions, but related to input sources, the algorithms for the retrieval of precipitation-related geophysical parameters, and the algorithms for the precipitation estimation from these parameters (some algorithms retrieve precipitation directly from satellite inputs).

(i) Input sources can have a great influence on estimation performance. For instance, without merging gauge data, CMORPH_RAW is intrinsically less correlated to TMA data than the other satellite-based QPEs (Fig. 5). CMORPH_BLD resembles CPCU in all aspects, probably due to its merging of CPCU for bias correction (Xie et al. 2017). Low gauge density can lead to inaccurate calibration of satellite-rainfall models. It either produces statistically unreliable models if the model is calibrated by regional gauges, or results in a small weight over that region if the model is calibrated by worldwide gauges. The sensitivities and uncertainties of rainfall with respect to different sensors (IR or MW) can vary to a large extent for different topographical characteristics and rain types. This effect is visible in Fig. 4: IR-only data (TAMSAT) are partly influenced by snow-covered regions (Mount Kilimanjaro) and not influenced by waterbodies (Lake Victoria and Lake Eyasi); MW-calibrated IR data, depending more on IR (CHIRPS), are influenced by snow, and partly influenced by waterbodies; and IR-assisted MW data depending more on MW (CMORPH and TRMM 3B42) are influenced by both snow and waterbodies.

(ii) The discrimination of precipitating from nonprecipitating scenes is problematic for both IR and MW retrievals. Most products estimate rainfall occurrences by threshold methods. The formation of rain in the real world is complex and has more randomness, therefore, threshold of a certain cloud feature lacks complexity for modeling rainfall statistically.

(iii) Aims of a product and techniques used in their estimation also introduce errors. For instance, the drought monitoring products TAMSAT and CHIRPS tend to underestimate both occurrences and intensities of rainfall during dry seasons and, thereby, overestimate droughts. The overestimation of the occurrences of consecutive dry/wet days in TAMSAT and CHIRPS lies in the fact that they downscale pentadal rainfall fields to obtain daily data, so they tend to underestimate temporal variability, resulting in underestimation of intermittency during dry or wet periods. CMORPH utilizes a snow-screening process (Joyce et al. 2004) and input source (retrievals from AMSU instruments) that remove estimates over the snow/ice covered regions. This could probably result in underestimation of rainfall over Mount Kilimanjaro. Note that the operational CMORPH constituent algorithms have recently been enhanced for snowfall detection.

(iv) The representativeness error and inherent error in the QPEs are difficult to be removed or taken into account, which may also lead to the discrepancy between gauge data and QPEs. Gauges provide point measurements over continuous periods of time. QPEs are derived from remote sensing signals averaged or interpolated over the period between passage times (from 3-hourly to daily) or even longer periods (pentadal to monthly) to give homogeneous rainfall over a grid cell. Therefore, the timing of rainfall is not well determined, which is illustrated by lagged detection of 99P showing better TS scores than regular detection (Fig. 8).

5. Conclusions

A methodology for evaluation of rainfall products (e.g., QPEs) was developed to better understand their performance under a wide variety of aspects, and to select the best products for different applications. In addition to the classic performance scores, such as PCC, NSE, RMSE, Bias, FBias, and TS, to assess the overall performance on the estimation of rainfall intensity and occurrence, the method can also evaluate the detection of extreme weather events over different rainfall regimes and seasons characterized by monthly rainfall.

Seasonally adjusted PCC and NSE were developed to assess QPEs' ability to capture rainfall variability at small time scale, which were less sensitive to large seasonality than regular PCC and NSE. The extreme weather conditions related to agricultural and hydrologic applications included dry/wet spell, moderate and heavy rainfall, which were defined as consecutive 10 or 20 dry/wet days, and 90th and 99th percentiles for each location, respectively. They were transformed to binary events, and the estimation of their intensity and timing by the QPEs were evaluated by categorical scores of FBias, TS, POD, and SR. Sources of errors were analyzed over regions and seasons, where the domain was partitioned into climate zones (e.g., rainfall regimes) and topographical regions, and a year was partitioned into ''dry,'' ''light rain,'' and ''heavy rain'' seasons by thresholds of monthly mean over each climate zone.

Seven rainfall estimates (QPEs) were then evaluated using the method, against rain gauge data provided by Tanzania Meteorological Agency over the total land surface of Tanzania during 1998 to 2006: TRMM 3B42 v7.0, TAMSAT v3.0, CHIRPS v2.0, CMORPH_RAW v1.0, CMORPH_BLD v1.0, WFDEI_CRU, and CPCU. CPCU was used to demonstrate the degree of coincidence between the reference gauge data and gauge data merged in the QPEs and detailed diagnoses for CPCU were not given. In general, for the example region the chosen datasets perform poorly at daily scale, the replication of monthly to yearly climatology, timing of rainfall, and extreme weather detection.

In fact, using QPE at smaller than pentadal time scale, one is not better off than simply using a monthly mean of gauge data. Or conversely, to reach a quality of rainfall estimates better than gauge monthly rainfall, the minimum time scale that can be used for QPE is 5-day aggregations. Variability smaller than this scale can simply not be represented by QPE products.

With respect to extreme weather, given the types and sources of the input data they used, and the algorithms they employed to produce rainfall estimates, the products showed strengths and weaknesses for different applications.

(i) For dry spell detection, CHIRPS and CMORPH_BLD perform better than others, where CHIRPS overestimates drought over monsoon regions by around 100% where CMORPH_BLD underestimates by 20%. Since CHIRPS only overestimates the occurrence of dry days in the dry season in monsoon regions, which is not a growing season for crops, CHIRPS is a better option for drought detection for agricultural use.

(ii) For wet spell detection, CMORPH_BLD and TRMM 3B42 perform well, where CMOPRH_BLD slightly underestimates the occurrences of wet spells by 5% but has the best estimation of timing with 65% hits, and TRMM 3B42 slightly overestimates by around 15%.

(iii) Moderate rains and heavy rains are poorly detected by all products. All QPEs except the two CMORPH products overestimate the occurrences of moderate rains. CMORPH_BLD has the best skill at timing of heavy rains with 49% hits, while other products have ≥50% misses and ≥50% false alarms. For heavy rainfall detection, all products except TRMM 3B42 underestimate the occurrences, where TRMM 3B42 shows the best performance on the estimation of the frequency with FBias = 1 and CMORPH_BLD of the timing with 40% hits.

The regions influenced by a combination effect of complex terrain and climate show large spatial and temporal rainfall variability, which requires more gauges to improve the understanding of the rainfall characteristic and uncertainty, and to achieve more reliable evaluation results.

The methodology can be used for, and easily adapted to, agricultural and hydrological applications. Seasonally adjusted PCC and NSE can identify the minimal temporal scale at which a product is useful for models that take rainfall as input (e.g., crop models), without the need to run the actual models. Dry/wet spells are crucial for crop growth processes, and the detection of them is important for risk assessment and decision-making. Capturing of moderate and heavy rainfall is critical for flood prediction and warning. The thresholds used in the definitions of these extreme weather conditions can be adjusted to agricultural droughts and floods for a given crop, or to thresholds that trigger floods in a certain region. Besides, the rainfall zones and rainfall seasons can be classified for different crop types and their growing phases for agricultural use, or by runoff systems for flood warning.

Acknowledgments. This study is funded by European Commission's Horizon 2020 research and innovation programme. It is part of project Oasis H2020_Insurance, with project number SC5-01-2016-2017, under work package 2.2.2 for Agricultural micro-Insurance, Tanzania: https://h2020insurance.oasishub.co/wp222-agricultural-micro-insurance-tanzania. We thank Tanzania Meteorological Agency for making the high-quality rain gauge data available, and Christoph Gornott from PIK for sharing these data. We also thank the editor and reviewers for their thoughtful comments and efforts towards improving the manuscript.

REFERENCES

Barron, J., 2004: Dry spell mitigation to upgrade semi-arid rainfed agriculture: Water harvesting and soil nutrient management for smallholder maize cultivation in Machakos, Kenya. Ph.D. thesis, Stockholm University, 38 pp., http://www.diva-portal.org/smash/record.jsf?pid=diva2%3A200763&dswid=6451.

Beck, H. E., and Coauthors, 2017: Global-scale evaluation of 22 precipitation datasets using gauge observations and hydrological modeling. *Hydrol. Earth Syst. Sci.*, **21**, 6201–6217, https://doi.org/10.5194/hess-21-6201-2017.

Cohen Liechti, T., J. P. Matos, J.-L. Boillat, and A. J. Schleiss, 2012: Comparison and evaluation of satellite derived precipitation products for hydrological modeling of the Zambezi River Basin. *Hydrol. Earth Syst. Sci.*, **16**, 489–500, https://doi.org/10.5194/hess-16-489-2012.

Collier, B., J. Skees, and B. Barnett, 2009: Weather index insurance and climate change: Opportunities and challenges in lower income countries. *Geneva Pap. Risk Insur. Issues Pract.*, **34**, 401–424, https://doi.org/10.1057/gpp.2009.11.

Dee, D. P., and Coauthors, 2011: The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quart. J. Roy. Meteor. Soc.*, **137**, 553–597, https://doi.org/10.1002/qj.828.

de Leeuw, J., J. Methven, and M. Blackburn, 2015: Evaluation of ERA-Interim reanalysis precipitation products using England and Wales observations. *Quart. J. Roy. Meteor. Soc.*, **141**, 798–806, https://doi.org/10.1002/qj.2395.

Dinku, T., P. Ceccato, E. Grover-Kopec, M. Lemma, S. J. Connor, and C. F. Ropelewski, 2007: Validation of satellite rainfall products over East Africa's complex topography. *Int. J. Remote Sens.*, **28**, 1503–1526, https://doi.org/10.1080/01431160600954688.

——, S. Connor, and P. Ceccato, 2010: Comparison of CMORPH and TRMM-3B42 over mountainous regions of Africa and South America. *Satellite Rainfall Applications for Surface Hydrology*, M. Gebremichael and F. Hossain, Eds., Springer, 193–204, https://doi.org/10.1007/978-90-481-2915-7_11.

Flitcroft, I. D., J. R. Milford, and G. Dugdale, 1989: Relating point to area average rainfall in semiarid West Africa and the implications for rainfall estimates derived from satellite data. *J. Appl. Meteor.*, **28**, 252–266, https://doi.org/10.1175/1520-0450(1989)028<0252:RPTAAR>2.0.CO;2.

Funk, C., and Coauthors, 2015: The climate hazards infrared precipitation with stations–A new environmental record for monitoring extremes. *Sci. Data*, **2**, 150066, https://doi.org/10.1038/sdata.2015.66.

Gosset, M., J. Viarre, G. Quantin, and M. Alcoba, 2013: Evaluation of several rainfall products used for hydrological applications over West Africa using two high-resolution gauge networks. *Quart. J. Roy. Meteor. Soc.*, **139**, 923–940, https://doi.org/10.1002/qj.2130.

Huffman, G. J., and D. T. Bolvin, 2015: TRMM and other data precipitation data set documentation. NASA TRMM Doc., 44 pp., https://pmm.nasa.gov/sites/default/files/document_files/3B42_3B43_doc_V7.pdf.

——, R. Adler, D. Bolvin, and E. Nelkin, 2010: The TRMM Multi-Satellite Precipitation Analysis (TMPA). *Satellite Rainfall Applications for Surface Hydrology*, M. Gebremichael and F. Hossain, Eds., Springer, 3–22, https://doi.org/10.1007/978-90-481-2915-7-1.

IPCC, 2012: *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation*. Cambridge University Press, 582 pp.

Jobard, I., F. Chopin, J. C. Berges, and R. Roca, 2011: An intercomparison of 10-day satellite precipitation products during West African monsoon. *Int. J. Remote Sens.*, **32**, 2353–2376, https://doi.org/10.1080/01431161003698286.

Joyce, R. J., J. E. Janowiak, P. A. Arkin, and P. Xie, 2004: CMORPH: A method that produces global precipitation estimates from passive microwave and infrared data at high spatial and temporal resolution. *J. Hydrometeor.*, **5**, 487–503, https://doi.org/10.1175/1525-7541(2004)005<0487:CAMTPG>2.0.CO;2.

Karl, T., N. Nicholls, and A. Ghazi, 1999: CLIVAR/GCOS/WMO workshop on indices and indicators for climate extremes workshop summary. *Weather and Climate Extremes*, Springer, 3–7, https://doi.org/10.1007/978-94-015-9265-9_2.

Kidd, C., P. Bauer, J. Turk, G. J. Huffman, R. Joyce, K. L. Hsu, and D. Braithwaite, 2012: Intercomparison of high-resolution precipitation products over Northwest Europe. *J. Hydrometeor.*, **13**, 67–83, https://doi.org/10.1175/JHM-D-11-042.1.

Klein Tank, A. M., F. W. Zwiers, and X. Zhang, 2009: Guidelines on analysis of extremes in a changing climate in support of informed decisions for adaptation. WCDMP-72, WMO/TD-1500, 56 pp., www.wmo.int/datastat/documents/WCDMP_72_TD_1500_en_1_1.pdf.

Li, L., C. S. Ngongondo, C.-Y. Xu, and L. Gong, 2013: Comparison of the global TRMM and WFD precipitation datasets in driving a large-scale hydrological model in southern Africa. *Hydrol. Res.*, **44**, 770–788, https://doi.org/10.2166/nh.2012.175.

Maidment, R. I., D. I. F. Grimes, R. P. Allan, H. Greatrex, O. Rojas, and O. Leo, 2013: Evaluation of satellite-based and model re-analysis rainfall estimates for Uganda. *Meteor. Appl.*, **20**, 308–317, https://doi.org/10.1002/met.1283.

——, D. Grimes, R. P. Allan, E. Tarnavsky, M. Stringer, T. Hewison, R. Roebeling, and E. Black, 2014: The 30 year TAMSAT African rainfall climatology and time series (TARCAT) data set. *J. Geophys. Res. Atmos.*, **119**, 10 619–10 644, https://doi.org/10.1002/2014jd021927.

——, and Coauthors, 2017: A new, long-term daily satellite-based rainfall dataset for operational monitoring in Africa. *Sci. Data*, **4**, 170063, https://doi.org/10.1038/sdata.2017.63.

Mathugama, S., and T. Peiris, 2011: Critical evaluation of dry spell research. *Int. J. Basic Appl. Sci.*, **11**, 153–160.

Okoola, R. E., 1999: A diagnostic study of the eastern Africa monsoon circulation during the northern hemisphere spring season. *Int. J. Climatol.*, **19**, 143–168, https://doi.org/10.1002/(SICI)1097-0088(199902)19:2<143::AID-JOC342>3.0.CO;2-U.

Peel, M. C., B. L. Finlayson, and T. A. McMahon, 2007: Updated world map of the Köppen-Geiger climate classification. *Hydrol. Earth Syst. Sci.*, **11**, 1633–1644, https://doi.org/10.5194/hess-11-1633-2007.

Pendergrass, A. G., R. Knutti, F. Lehner, C. Deser, and B. M. Sanderson, 2017: Precipitation variability increases in a warmer climate. *Sci. Rep.*, **7**, 17966, https://doi.org/10.1038/s41598-017-17966-y.

Roebber, P. J., 2009: Visualizing multiple measures of forecast quality. *Wea. Forecasting*, **24**, 601–608, https://doi.org/10.1175/2008WAF2222159.1.

Rowhani, P., D. B. Lobell, M. Linderman, and N. Ramankutty, 2011: Climate variability and crop production in Tanzania. *Agric. For. Meteor.*, **151**, 449–460, https://doi.org/10.1016/j.agrformet.2010.12.002.

Schmidli, J., and C. Frei, 2005: Trends of heavy precipitation and wet and dry spells in Switzerland during the 20th century. *Int. J. Climatol.*, **25**, 753–771, https://doi.org/10.1002/joc.1179.

Shiklomanov, A. I., R. B. Lammers, and C. J. Vörösmarty, 2002: Widespread decline in hydrological monitoring threatens Pan-Arctic Research. *Eos, Trans. Amer. Geophys. Union*, **83**, 13–17, https://doi.org/10.1029/2002EO000007.

Sivakumar, M., 1992: Empirical analysis of dry spells for agricultural applications in West Africa. *J. Climate*, **5**, 532–539, https://doi.org/10.1175/1520-0442(1992)005<0532:EAODSF>2.0.CO;2.

Stokstad, E., 1999: Scarcity of rain, stream gages threatens forecasts. *Science*, **285**, 1199–1200, https://doi.org/10.1126/science.285.5431.1199.

Tarnavsky, E., D. Grimes, R. Maidment, E. Black, R. P. Allan, M. Stringer, R. Chadwick, and F. Kayitakire, 2014: Extension of the TAMSAT satellite-based rainfall monitoring over Africa and from 1983 to present. *J. Appl. Meteor. Climatol.*, **53**, 2805–2822, https://doi.org/10.1175/JAMC-D-14-0016.1.

Taylor, K. E., 2001: Summarizing multiple aspects of model performance in a single diagram. *J. Geophys. Res.*, **106**, 7183–7192, https://doi.org/10.1029/2000JD900719.

Thiemig, V., R. Rojas, M. Zambrano-Bigiarini, V. Levizzani, and A. D. Roo, 2012: Validation of satellite-based precipitation products over sparsely gauged African river basins. *J. Hydrometeor.*, **13**, 1760–1783, https://doi.org/10.1175/JHM-D-12-032.1.

USAID, 2016: Tanzania meteorological agency climate data rescue pilot project report. 183 pp., https://www.climatelinks.org/sites/default/files/asset/document/2016_USAID-PREPARED_Tanzania-Meteorological-Agency-Climate-Rescue-Pilot-Project.pdf.

van de Giesen, N., R. Hut, and J. Selker, 2014: The Trans-African Hydro-Meteorological Observatory (TAHMO). *Wiley Interdiscip. Rev.: Water*, **1**, 341–348, https://doi.org/10.1002/wat2.1034.

Washington, R., and Coauthors, 2006: African climate change: Taking the shorter route. *Bull. Amer. Meteor. Soc.*, **87**, 1355–1366, https://doi.org/10.1175/BAMS-87-10-1355.

Weedon, G. P., G. Balsamo, N. Bellouin, S. Gomes, M. J. Best, and P. Viterbo, 2014: The WFDEI meteorological forcing data set: WATCH forcing data methodology applied to ERA-Interim reanalysis data. *Water Resour. Res.*, **50**, 7505–7514, https://doi.org/10.1002/2014WR015638.

Xie, P., and A.-Y. Xiong, 2011: A conceptual model for constructing high-resolution gauge-satellite merged precipitation analyses. *J. Geophys. Res.*, **116**, D21106, https://doi.org/10.1029/2011jd016118.

——, A. Yatagai, M. Chen, T. Hayasaka, Y. Fukushima, C. Liu, and S. Yang, 2007: A gauge-based analysis of daily precipitation over East Asia. *J. Hydrometeor.*, **8**, 607–626, https://doi.org/10.1175/JHM583.1.

——, M. Chen, and W. Shi, 2010: CPC unified gauge-based analysis of global daily precipitation. *24th Conf. on Hydrology*, Atlanta, GA, Amer. Meteor. Soc., 2.3A, https://ams.confex.com/ams/90annual/techprogram/paper_163676.htm.

——, R. Joyce, S. Wu, S.-H. Yoo, Y. Yarosh, F. Sun, and R. Lin, 2017: Reprocessed, bias-corrected CMORPH global high-resolution precipitation estimates from 1998. *J. Hydrometeor.*, **18**, 1617–1641, https://doi.org/10.1175/JHM-D-16-0168.1.