Validation of new satellite rainfall products over the Upper Blue Nile Basin, Ethiopia

Getachew Tesfaye[1,3], Tsegaye Tadesse[2], Berhan Gessesse[1], Tufa Dinku[4]

[1]Earth observation research division, Entoto Observatory & Research Center, Addis Ababa, 33679, Ethiopia
5   [2]National Drought Mitigation Center, School of Natural Resources, University of Nebraska-Lincoln, Lincoln, 830988, USA
[3]Geospatial Data & Technology Center, Institute of Land Administration, Bahir Dar University, Bahir Dar, 79, Ethiopia
[4]International Research Institute for Climate and Society, The Earth Institute at Columbia University,
10   Palisades NY 10964, USA

Correspondence to: Getachew Tesfaye (sget2006@yahoo.com)

Abstract. Accurate measurement of rainfall is vital to analyze the spatial and temporal patterns of precipitation at various scales. However, the conventional rain gauge observations in many parts of the world such as Ethiopia are sparse and
15   unevenly distributed. An alternative to traditional rain gauge observations could be satellite-based rainfall estimates. Satellite rainfall estimates could be used as a sole product (e.g. in areas with no (poor) ground observations) or through integrating with rain gauge measurements. In this study, the potential of a newly available Climate Hazards Group Infrared Precipitation with Stations (CHIRPS) rainfall product has been evaluated in comparison to rain gauge data over the Upper Blue Nile basin in Ethiopia for the period of 2000 to 2015. Besides, the Tropical Applications of Meteorology using SATellite and ground-
20   based observations (TAMSAT 3) and the African Rainfall Climatology (ARC 2) products have been used as a benchmark and compared with CHIRPS. From the overall analysis at dekadal and monthly temporal scale, CHIRPS exhibited the better performance in comparison to TAMSAT 3 and ARC 2 products. An evaluation based on categorical/volumetric and continuous statistics indicated that CHIRPS has the greatest skills in detecting rainfall events (POD = 0.99, 1.00) and measure of volumetric rainfall (VHI = 1.00, 1.00), the highest correlation coefficients (r = 0.81, 0.88), the better bias values
25   (0.96, 0.96), and the lowest RMSE (28.45 mm/dekad, 59.03 mm/month) than TAMSAT 3 and ARC 2 products at dekadal and monthly analysis, respectively. CHIRPS overestimates the frequency of rainfall occurrence (up to 31% at dekadal scale), although the volume of rainfall recorded during those events was very small. Indeed, TAMSAT 3 has shown a comparable performance with that of CHIRPS product, mainly with regard to bias. The ARC 2 product was found to have the weakest performance underestimating rain gauge observed rainfall by about 24%. In addition, the skill of CHIRPS is less affected by
30   variation in elevation in comparison to TAMSAT 3 and ARC 2 products. CHIRPS resulted an average biases of 1.11, 0.99 and 1.00 at lower (< 1000 m a.s.l.), medium (1000 to 2000 m a.s.l.), and higher elevation (> 2000 m a.s.l.), respectively. Overall, the finding of this validation study shows the potentials of CHIRPS product to be used for various operational applications such as rainfall pattern and variability study in the Upper Blue Nile basin in Ethiopia.

1 Introduction

Rainfall is a major component of the climate system and plays a key role in the Earth's hydrological cycle and energy balance. Rainfall variability in its rate, amount, and distribution substantially determine the earth's ecosystem, water cycle and climate (Huang and Van den Dool, 1993; Stillman et al., 2014). Thus, accurate measurement of rainfall is vital to analyze the spatial and temporal patterns of precipitation at various scales and advance our understanding of the effect of rainfall on agriculture, hydrology, and climatology. Conventionally, the rain gauge is a primary source of rainfall data, which has been the most accurate and reliable approach for rainfall measurement. However, ground rainfall stations in many parts of the world and most parts of Ethiopia are very sparse and unevenly distributed. As a result, analysis using rain gauge observation is significantly limited to point based particular location. Because of this scattered distribution of weather stations, the dependability of rain gauge data to estimate areal rain and spatial distribution of rainfall over large areas of Ethiopia is considerably reduced. However, advances in remote sensing science have provided an opportunity to estimate rainfall from satellite observations and are becoming an important source of rainfall data.

Satellite-derived rainfall estimates (SREs) are widely available from Thermal Infrared Radiation (TIR) and Passive Microwave (PMW) Channels, from geostationary and Low-Earth-orbiting satellites, respectively. The TIR-based approaches use an indirect relationship to estimate rainfall from cloud top brightness temperatures. The TIR-based rainfall estimates have some uncertainties because of misidentification of rain-producing clouds such as cirrus clouds while warm clouds might generate a considerable amount of rain (Trejo et al., 2016). On the contrary, PMW approach is based on the direct measurements of atmospheric liquid water content and rainfall intensity by penetrating clouds and as a result would give more accurate rainfall estimates (Kummerow et al., 2001; Young et al., 2014). However, observations from PMW are less frequent due to a relatively low temporal resolution from low Earth-orbiting satellites. Combining TIR and PMW has been the recent approach to estimate rainfall from satellites nowadays.

Techniques for satellite rainfall estimates have limitation and embedded uncertainties because satellites do not measure rainfall by itself and should be related to precipitations based on one or multiple surrogate variables (Wu et al., 2012; Toté et al., 2015). The uncertainties, therefore, may instigate in the processes of temporal samplings, error from algorithms and satellite instruments itself (Gebremichal et al., 2005). These may affect the accuracy of satellite-derived rainfall products and may result in a significant error when they are used for various purposes such as rainfall pattern and variability study. The issue of accuracy has received substantial attention to the extent that satellite-derived rainfall products are concerned. In this respect, stringent validation is essential to verify the performance of the product in a diverse physiographic setting and use for the intended applications.

Several studies have been conducted in Ethiopia (e.g., Dinku et al, 2007, 2008 and 2011a; Hirpa et al., 2010; Romilly and Gebremichael, 2011; Young et al., 2014 and Gebre et al., 2015) and specifically to the Upper Blue Nile basin (e.g., Dinku et al. 2011b; Gebremichael et al., 2014; Fenta et al., 2014 and Worqlul et al., 2014) to validate the performance of satellite-based rainfall products. These studies validated mainly the skills of Tropical Applications of Meteorology using SATellite

and ground-based observations (TAMSAT, Grimes et al., 1999; Thorne et al., 2001, Maidment et al., 2014; Tarnavsky et al., 2014), Africa Rainfall Climatology version 2 (ARC 2, Novella and Thiaw, 2013), Tropical Rainfall Measuring Mission (TRMM, Huffman et al., 2007), Climate Prediction Centre (CPC) morphing technique (CMORPH, Joyce et al., 2004), and Precipitation Estimation from Remotely Sensed Information Using Artificial Neural Networks (PERSIANN, Hsu and

5   Sorooshian, 2008) precipitation products at different spatial and temporal scale and topographic patterns. The results of these studies indicate that the skills of SREs vary with the characteristics of local climate, topography, and seasonal distributions of rainfall and have shown low to moderately high skills. There is now a newly available satellite rainfall product called the Climate Hazards Group Infrared Precipitation with Stations (CHIRPS, Funk et al., 2015) with a relatively high spatial and temporal resolution (i.e. 5 km resolution at daily temporal scale) and quasi-global coverage. So far, however, there has been

10   a very little work on the performance of CHIRPS satellite rainfall estimates over Ethiopia as well as other countries in Africa. That might be because CHIRPS is a relatively new dataset. The works of Toté et al. (2015) in Mozambique can be mentioned here as the first validation work we are aware of that reveals the potential applications of CHIRPS in Africa. Maidment et al. (2017) have also validated the performance of satellite rainfall products (including CHIRPS V 2.0) in four countries in Africa (Mozambique, Nigeria, Uganda, and Zambia). In general, these few validation works have shown the

15   promising skills of CHIRPS in Africa and its potentials for various working applications in the continent. Nevertheless, it is important to note that for better exploitation of a relatively new CHIRPS rainfall product more validation work needs to be done at different spatial and temporal scales in the region.

For this validation study, the Upper Blue Nile basin in Ethiopia was selected because of a relatively good density of rain gauge stations, varied topography, and high spatial and temporal variability of precipitation (Taye and Willems, 2013). The

20   aim of this study was, therefore, to compare and validate the performance of CHIRPS with rain gauge observations that were collected from thirty-two weather stations from 2000 to 2015. CHIRPS performance was also compared against TAMSAT 3, TAMSAT 2, and ARC 2 satellite rainfall products. In the course of this analysis, both the TAMSAT and ARC 2 products have been validated as well. In addition, this study has also been paid some courtesy to compare TAMSAT 2 and TAMSAT 3 products to assess the improvements made with the new version (TAMSAT 3, Maidment et al., 2017). The analyses used

25   dekadal (10 days) and monthly time scale rainfall data both from the satellite products and rain gauge observations.

The paper is structured as follows. Section 2 provides the site descriptions of the study area, followed by dataset used in the study (Section 3). In section 4 detailed descriptions of the methodology used in this study are provided. The results and discussions are given in Section 5. Finally, Section 6 presents our conclusions.

2 Site descriptions

30   The Upper Blue Nile (UBN) basin is located in the north-western part of Ethiopia with latitude between 7°45' and 12°45' N and longitude between 34°30' and 39°45' E (Figure 1). The Blue Nile River is originated from Lake Tana in Ethiopia and travels all the way to the Sudanese border to finally meet the White Nile at Khartoum. The UBN basin is a primary source of

the Nile River, and it contributes about 60% of the annual flow of the Nile (Conway, 2005; Degefu, 2003). The basin has an approximate drainage area of 176,000 km^2 (Conway, 2000). The basin is characterized by a complex topography with elevation ranging from 4261 m a.s.l. at the north-eastern part of the basin to 500 m a.s.l. at the western part of the basin near the Ethiopian-Sudan border (Figure 1). The incessantly changing topography of the basin leads to varying agro-ecology within short distances. The climate of the UBN basin ranges from humid to semi-arid. The main rainfall season (known as "Kiremt") occurs from June to September. The dry season runs from October to January followed by a short rainy season (called "Belg") from February to May. According to Kim et al. (2008), about 70% of the annual precipitation in the study area (UBN basin) is observed during the Kiremt season. The UBN basin receives up to 2,200 mm of annual rainfall. The annual mean rainfall varies between 1200-1800 mm (Conway, 2000) with an increasing trend from northeast to southwest (Kim et al., 2008). However, the basin is characterized by large temporal fluctuations in rainfall (Conway, 2000; Taye and Willems, 2013) both in intra-annual and inter-annual scale. As a result, the hydrological processes in the basin are quite complex and highly variable in space and time. The impact of rainfall variability in the basin described by severe and regular climatic and hydrological extremes, such as floods and droughts and ensuing low rate of food production and poverty (Taye and Willems, 2012). Although quite a diverse land use systems are common, the livelihoods of the majority of the populations in the basin are highly dependent on rain-fed agriculture.

3 Data set

Rainfall data for this study were collected from ground based weather stations and remote sensing satellite estimates.

3.1 Station data

Rain gauge observed daily rainfall data from thirty-two 1st and 2nd class stations from 2000 to 2015 were collected from the National Meteorological Agency (NMA) of Ethiopia. First class stations (synoptic stations) are those stations where all meteorological parameters are recorded every hour. While for second class stations observations are taken every three hours. Since the SREs evaluated here incorporate rain gauge data, the available rain gauge datasets were compared with the station archives (data source for the generation of CHIRPS, TAMSAT, and ARC2) and those datasets used for the generation of SREs were removed from the analysis to guarantee the complete independence of the validation datasets. Therefore, a total of 3460 complete dekadal observations that were not used for the generation/calibration of SREs were retained for the validation over the thirty-two stations.

3.2 Satellite rainfall data

High resolution satellite rainfall products selected for this study are CHIRPS v.2.0 (a relatively new satellite rainfall product), TAMSAT 3, TAMSAT 2, and ARC 2. The TAMSAT 2 product was used in this study mainly to assess the improvements made by the recent version TAMSAT 3. These rainfall products were selected because they: (i) have a

relatively high spatial resolution, (ii) relatively with long time series, and (iii) are freely available. Brief descriptions of these datasets are given below.

3.2.1 CHIRPS, v 2.0

CHIRPS is a quasi-global (50° S-50° N) gridded products available from 1981 to near-present at 0.05° spatial resolution
5 (~5.3 km) and at daily, pentadal, dekadal, and monthly temporal resolution (Funk et al., 2015). The CHIRPS dataset is developed by the U.S. Geological Survey (USGS) and the Climate Hazards Group (CHG) at the University of California (Knapp et al., 2011; Funk et al., 2015). The development of CHIRPS products entails three major input datasets and processes. First, Infrared Precipitation (IRP) pentad (5-day) rainfall estimates are created from two TIR satellite observations archives (i.e., Globally Gridded Satellite (GriSat) and NOAA Climate Prediction Center dataset (CPC TIR)) using cold cloud
10 durations (CCD) and calibrated using the Tropical Rainfall Measuring Mission Multi-Satellite Precipitation Analysis (TMPA 3B42) precipitation pentads. Then, the IRP pentads were divided by its long-term IRP mean values to present as percent of normal. Second, the percent of normal IRP pentad is then multiplied by the corresponding Climate Hazards Precipitation Climatology (CHPClim) pentad to produce an unbiased gridded estimate, with units of millimetres per pentad, called the CHG IR Precipitation (CHIRP). In the third part of the process, the final product CHIRPS has been produced through
15 blending stations with the CHIRP data sets. Details of CHIRPS satellite rainfall products can be found in Funk et al. (2015).

3.2.2 TAMSAT

TAMSAT product is developed by the University of Reading based on Meteosat TIR (Thermal Infrared) channel. TAMSAT rainfall estimation method (Dugdale et al., 1991; Grimes et al., 1999; Thorne et al., 2001; Maidment et al., 2014; Tarnavsky et al., 2014) assumes that rainfall is produced from convective clouds that lead to cold cloud tops, and rainfall and CCD are
20 linearly correlated. The retrieval algorithm is calibrated using local gauge records. TAMSAT products are available from 1983 onwards at 0.0375° spatial resolution (~4 km) and at dekadal, monthly, and seasonal temporal resolution. This validation study has considered the recent version of TAMSAT product (TAMSAT 3) for the comparison to CHIRPS product. However, the previous version (TAMSAT 2) was also incorporated to further confirm the improvements made by the recent version TAMSAT 3. The principle of the TAMSAT method is still the same for TAMSAT 2 and TAMSAT 3.
25 However, there are some improvements on the calibration procedures and approaches. Details on the main difference between the recent version (TAMSAT 3) and the previous version (TAMSAT 2) have been provided by (Maidment et al., 2017).

3.2.3 ARC 2

ARC 2 is the revised version of the first version ARC 1 (Novella and Thiaw, 2013). The ARC 2 satellite rainfall estimates
30 was produced from two primary input data sources: 1) 3-hourly geostationary infrared (IR) data centred over Africa from the European Organization for the Exploitation of Meteorological Satellites (EUMETSAT) and 2) quality controlled Global

Telecommunication System (GTS) gauge observations reporting 24-hour rainfall accumulations over Africa (Novella and Thiaw, 2013). The ARC 2 dataset is available at daily time scale with a grid resolution of 0.1° X 0.1°and with a spatial domain of 40° S-40° N and 20° W-55° E, encompassing the African continent from 1983 to the present.

4 Methodology

5   This study has evaluated the performance of CHIRPS satellite rainfall estimates at dekadal and monthly temporal scales against thirty-two rain gauge observations and compared with TAMSAT 3, TAMSAT 2, and ARC 2 products for the period of 2000 to 2015. The dekadal and monthly data were further classified to validate the satellite products per elevations and for each month over the UBN basin, respectively. The double mass curve techniques and correlation coefficient analysis (similar to Gebere et al., 2015) confirmed the consistency and homogeneity of rain gauge observations, respectively. The dekadal and

10   monthly data were created from the aggregates of daily rain gauge observations and TAMSAT 3 and ARC 2 rainfall values, while CHIRPS and TAMSAT 2 satellite products are available at dekadal and monthly time scale. The comparison between gridded satellite rainfall estimates and ground rainfall observations can be made using either grid-to-grid or point-to-grid comparison methods. However, an attempt made to convert point ground observations to gridded interpolated dataset lead to poor result due to uneven geospatial distributions of gauge stations. Thus, this study has used point-to-grid comparison

15   approaches. For each validation station, the grid values of satellite rainfall products containing the stations were extracted and pair-wise comparisons with rain gauge values have been undertaken.

4.1 Performance analysis

The performances of satellite rainfall estimates were analyzed using categorical and volumetric indices and the continuous statistical measures. The most common form of categorical indices is a 2 x 2 contingency table which reports the number of

20   hit (H), miss (M), false alarm (F) and true null events. To describe whether there is rain or no rain events, a threshold value of 1.0 mm/dekad or month was used in evaluating the skills of the satellite products.

4.1.1 Categorical validation indices

This section summarizes the categorical indices used to assess the intensity of rainfall estimated by satellite products with respect to gauge observation. These include the Probability of Detection (POD), the False Alarm Ratio (FAR), and the

25   Critical Success Index (CSI). The POD score is defined as $H/(H+M)$, and it describes the fraction of the gauge observations detected correctly by the satellite, while the false alarm ratio, FAR=$F/(H+F)$, corresponds to the portion of events identified by the satellite but not confirmed by gauge observations. The Critical Success Index, CSI=$H/(H+M+F)$, combines different aspects of the POD and FAR, describing the overall skill of the satellite products relative to gauge observation. All these categorical validation indices have a score values ranging from 0 to 1; and 1 indicates the perfect skill, except for FAR the

30   perfect score is 0.

4.1.2 Volumetric validation indices

Since the contingency table metrics do not provide information regarding the volume of correctly (incorrectly) detected rainfall by the satellite products relative to rain gauge observations, recently AghaKouchak and Mehran (2013) suggested an extension of categorical table indices known as "volumetric indices". In this study, therefore, the volumetric indices that include: (a) Volumetric Hit Index (VHI), (b) Volumetric False Alarm Ratio (VFAR), and (c) the Volumetric Critical Success Index (VCSI) that were proposed by AghaKouchak and Mehran (2013) have been adopted to evaluate the volumetric performance of the selected satellite rainfall products.

$$\text{VHI} = \frac{\sum_{i=1}^{n}(S_i|(S_i>t \ \& \ G_i>t))}{\sum_{i=1}^{n}(S_i|(S_i>t \ \& \ G_i>t))+\sum_{i=1}^{n}(G_i|(S_i\leq t \ \& \ G_i>t))} \ ,$$

(1)

VHI: volume of correctly detected rainfall by the satellites relative to the volume of the correctly detected satellites and missed gauge observations

$$\text{FAR} = \frac{\sum_{i=1}^{n}(S_i|(S_i>t \ \& \ G_i\leq t))}{\sum_{i=1}^{n}(S_i|(S_i>t \ \& \ G_i>t))+\sum_{i=1}^{n}(S_i|(S_i>t \ \& \ G_i\leq t))} \ ,$$

(2)

VFAR: volume of false rainfall by the satellites relative to the sum of rainfall by the satellites

$$\text{VCSI} = \frac{\sum_{i=1}^{n}(S_i|(S_i>t \ \& \ G_i>t))}{\sum_{i=1}^{n}(S_i|(S_i>t \ \& \ G_i>t))+\sum_{i=1}^{n}(G_i|(S_i\leq t \ \& \ G_i>t))+\sum_{i=1}^{n}(S_i|(S_i>t \ \& \ G_i\leq t))} \ ,$$

(3)

VCSI: overall measure of volumetric performance.

Here S is satellite rainfall estimates, G is gauge observations, $i=1$ to n and n is the sample size, and t is the threshold values ($t=1$ mm in this study).

4.1.3 Continuous statistical tools

In addition, the continuous statistical measures were used to quantify the overall performance of the satellite rainfall products.

$$r = \frac{\sum(G-\bar{G})(S-\overline{S})}{\sqrt{\sum(G-\bar{G})^2}\sqrt{\sum(S-\overline{S})^2}} \ ,$$

(4)

Pearson correlation (*r*) is used to evaluate the goodness of fit of the relation. The value of 1 is the perfect score.

$$\text{RMSE} = \sqrt{\frac{\sum(G-S)^2}{n}} \ ,$$

(5)

The root mean square error (RMSE) measures the absolute mean difference between two data sets. The value of 0 is the perfect score.

$$\text{Bias} = \frac{\sum S}{\sum G},$$ (6)

Bias is a measure of how the average satellite rainfall magnitude compares to the ground rainfall observation. The value of 1 is the perfect score. A bias value above (below) 1 indicates an aggregate satellite overestimation (underestimation) of the ground precipitation amounts.

Here G is gauge rainfall observations, S is satellite rainfall estimates, $\bar{G}$ is average gauge rainfall observations, $\bar{S}$ is the average satellite rainfall estimates, and n is the number of data pairs.

5 Results and Discussions

The performances of satellite rainfall estimates were evaluated using the categorical indices (i.e., POD, FAR and CSI), volumetric index (i.e., VHI, VFAR, and VCSI), and a set of continuous statistics (i.e., correlation coefficient (r), bias and RMSE) at dekadal and monthly temporal scale. High values of POD, VHI, CSI, VCSI, and r, and small values of FAR, VFAR, and RMSE, and bias values of one (or near to one) indicate a good performance of the satellite rainfall products.

5.1 Spatial rainfall patterns of satellite products

Figure 2 provides the 16-years mean rainy season (June to September) and the annual rainfall of TAMSAT 2, ARC 2, TAMSAT 3, and CHIRPS satellite rainfall products over the UBN basin in Ethiopia for the period of 2000 to 2015. The wet/kiremt season (June to September) produced the majority of the total annual precipitation. Therefore, both the rainy season (Figure 2 (a)) and annual estimates (Figure 2 (b)) generated by the satellite products have shown a similar rainfall patterns. However, TAMSAT 2 and ARC 2 showed a decreasing trend of rainfall from west to the east region (or from low to high elevation areas) of the basin, while TAMSAT 3 and CHIRPS show a significant amount of rainfall in the central and southwest regions. The large discrepancy in TAMSAT 2 and ARC 2 rainfall pattern in the west and east areas could be attributed to the orographic effect on rainfall.

5.2 Dekadal comparison

The dekadal comparisons were made using i) all dekadal values from rain gauge observation and satellite products and ii) classified the dekadal values, for further validation, per elevation of the UBN basin.

5.2.1 Overall validation at dekadal temporal scale

Table 1 gives an overall comparison between the satellite products and rain gauge observation from 2000 to 2015 at a dekadal temporal scale. Besides, Figure 3 and Figure 4 provide the cumulative distribution function (CDF) and the scatter plot, respectively.

The overall evaluation and comparison summary, shown in Table 1, indicate that CHIRPS scored relatively higher POD, VHI, and VCSI values followed by TAMSAT 3 and TAMSAT 2. It is apparent from the same table that, both TAMSAT products have shown a similar skill and have scored almost similar POD, VHI and VCSI values. Given these results, it is possible to conclude that the improvement made by TAMSAT 3 over the previous version TAMSAT 2 on the skills of detecting the frequency of a rainfall event is very insignificant. On the other hand, ARC 2 scored relatively lower POD, VHI and VCSI values.

However, ARC 2, TAMSAT 2, and TAMSAT 3 scored lower FAR and higher CSI values than CHIRPS. The CHIRPS product resulted in the highest FAR (0.31) and lowest CSI (0.68) values. Similarly, a FAR value of 0.29 (close to 0.31 of this study) for CHIRPS has been obtained by Tote et al. (2015) from the dekadal product validation in Mozambique. This means that TAMSAT (hereafter refers both version 2 and version 3) and ARC 2 products are better than CHIRPS in detecting the relative frequency of rain events. The overestimation of rainy days by CHIRPS might be related to the process of translating infrared (IR) CCD values into estimates of precipitation using the 0.25° grid cell TMPA datasets, that may result in the formation of too much light rain (Funk et al., 2015). Nevertheless, from the volumetric indices, VFAR values (0.06) of CHIRPS is much reduced, and its overall performance (VCSI=0.94) is improved and even better than TAMSAT and ARC 2 products. Since the volumes of rainfall detected by CHIRPS during false events were tiny, it had a minimal contribution to the total amounts of rainfall.

Further Table 1 showed that CHIRPS has a better agreement with the rain gauge observations than TAMSAT and ARC 2 on most continuous statistical assessments, results in the highest correlation coefficient (r), better bias value and the lowest RMSE. The two likely explanations for CHIRPS good performance might be the use of CHPClim and the inclusion of stations data in the CHIRPS datasets (Funk et al., 2015). Indeed, TAMSAT 3 has scored very comparable values with CHIRPS product, particularly to the bias ratio. Both CHIRPS and TAMSAT 3 have managed to reproduce the rainfall amount observed by rain gauge stations reasonably well (with an overall bias of 0.96 (i.e., underestimate only by 4%) and 1.04 (overestimate only by 4%), respectively), while TAMSAT 2 and ARC 2 showed a substantial underestimation of rain gauge observation by 31% and 24%, respectively. The underestimations of ARC 2 and TAMSAT 2 might be attributed to the complex topography of the validation site (possibly dominated by warm rain processes) that may reduce the ability to identify rainy clouds (Dinku et al., 2007; Funk et al., 2015; Maidment et al., 2014) and the calibration process using gauge stations. However, the statistical analysis in Table 1 reveals that the recent version TAMSAT 3 has well addressed the problem of underestimation of rainfall by TAMSAT 2 and it significantly improved the bias ratios. Thus, the overall dekadal

validation and comparison indicated that CHIRPS has a high level of correspondence with rain gauge observations and may have a useful skill for various functions in the study area.

In Figure 3, the CDF of dekadal rainfall between the satellite products and the rain gauge observation have presented to validate how often the satellite products occur below or above the rain gauge observation values. As can be seen in Figure 3 (a), TAMSAT 3 has shown a better performance (followed by CHIRPS) in detecting dekadal maximum values observed by rain gauge stations. The result may show the significant improvements made by TAMSAT 3 in comparison to the previous version TAMSAT 2 product. The plot in Figure 3 (b) further reveals that CHIRPS and TAMSAT 3 are very close to the rain gauge observation at all rainfall measurement values, except for little rain (< 20 mm) and rainfall between (20 to100 mm) accumulation, respectively, where they show a slight overestimation. The CHIRPS product has also demonstrated a little underestimation in high rainfall areas. A similar result for CHIRPS product has been noted by prior studies of Tote et al. (2015) and Trejo et al. (2016) in Mozambique and Venezuela, respectively. However, TAMSAT 2 and ARC 2 are well below the rain gauge observations. The comparison between SREs and rain gauge observations at 80% frequency level indicated that TAMSAT 3 and CHIRPS only varies with 5.9 mm above and 2.69 mm below, respectively, from the 71.5 mm rainfall value observed by rain gauge stations while ARC 2 and TAMSAT 2 are 13.84 mm and 16.3 mm below, respectively, at dekadal temporal scale. This shows that CHIRPS (followed by TAMSAT 3) is very close while TAMSAT 2 and ARC 3 are well below to rain gauge observed values.

In addition, a scatter plots shown in Figure 4 were used to further define the relationship between satellite rainfall products and rain gauge observations. The satellite rainfall estimates show better agreement with rain gauge observations at lower rainfall amount. The agreement slowly reduces to the higher values. However, CHIRPS and TAMSAT 3 have shown a relatively better deal with rain gauge observations (with r = 0.81 and 0.78, in their order of appearance) in comparison to TAMSAT 3 and ARC 2 at dekadal time-scale. Whereas, ARC 2 has exhibited the lowest agreement with rain gauge values (r = 0.72), compared to other SREs. The regression values are very consistent with the values presented in CDF shown in Figure 3.

5.2.2 Comparison at different elevations using the dekadal time scale data

The effect of topography on the skill of satellite rainfall products might be substantial (Hirpa et al., 2010). Stations selected in this study have a broad range of elevation from 790 to 3098 m a.s.l. This wide range of elevation and spatial variation is essential to confirm the dependence of the satellite rainfall products on topographic patterns. The dekadal time scale data classified into the 32 rain gauge stations. Thus, the skills of the satellite products at different station elevations have been validated, and the results are given in Figure 5 and 6.

Figure 5 depicts the categorical and volumetric indices of the satellite products at different elevation values during 2000 to 2015. CHIRPS has shown a more prominent skill than TAMSAT and ARC 2 products and scored POD and VHI values close to 1.00 at most elevations. However, the competencies of TAMSAT and ARC 2 products in detecting rainfall events seem to reduce with elevations.

A closer look in Figure 5 (a) and to some extent in Figure 5 (c) indicated a clear trend of decreasing the skills of TAMSAT 3 and ARC 2 with an increase in elevations. Stations with relatively low elevation values ranging from 790 to 1928 m resulted in the highest POD and CSI values for TAMSAT and ARC 2 estimates, whereas the majority of TAMSAT and ARC 2's least skills were recorded by relatively high elevation stations ranging from 2000 to 3098 m. Further analysis of the

5   correlation between the satellites skills (i.e. POD, FAR, CSI, VHI, VFAR, and VCSI) and elevations (given in Table 2) showed that the POD of TAMSAT 3, TAMSAT 2, and ARC 2 products have a substantial negative correlation with elevation. The same table (Table 2), indicates that the skill of CHIRPS has resulted in a relatively low correlation coefficient with elevations. Overall, these results could imply that the skills of CHIRPS estimate are less affected by variation in elevation in comparison to TAMSAT and ARC 2 products. However, in most other indices no clear relationships between

10   the skills of the SREs and change in elevations were observed.

From the statistical analysis presented in Figure 6 (a), the satellite products have shown correlation coefficients (r) ranging from 0.32 to 0.91 independent of variation in elevations. The lowest correlation ($r=0.32$) was scored by TAMSAT 2 at "Sirinka" rain gauge station with an elevation of 1861 m.a.s.l. Moreover, the bias ratios for TAMSAT 2 and ARC 2 seem to have elevation dependent trends (Figure 6 (b) and Table 2). The CHIRPS and TAMSAT 3 have scored the best average bias

15   ratios (1.00 and 1.07, respectively) independent of elevations, although they considerably under-/over-estimates rainfall values at some elevations. The average bias ratio among satellite products at wiser elevation range were compared and ARC 2 (TAMSAT 2) resulted in the mean bias of 1.53 (1.35), 0.86 (0.73) and 0.77 (0.66) at low (< 1000 m a.s.l.), medium (1000 to 2000 m a.s.l.), and high elevation (> 2000 m a.s.l.), respectively. On the other hand, CHIRPS dataset has scored a bias of 1.11, 0.99 and 1.00, while TAMSAT 3 gained 1.14, 1.07 and 1.07 at low, medium, and high elevation, respectively. These

20   results are in a good agreement with those presented in Table 2 and Figure 7. The results, as shown in Table 2, indicated that the bias ratio of ARC 2 and TAMSAT 2 have the modest negative correlation with elevations ($r=-0.44$ and $r=-38$), respectively, while CHIRPS and TAMSAT 3 resulted in correlation values close to zero.

The same result has been revealed by Figure 7, in which TAMSAT 2 and ARC 2 underestimate rainfall values at higher (Figure 7 (a)) and medium (Figure 7 (b)) elevations while they overestimate at lower elevation (Figure 7 (c)) stations. The

25   average dekadal values from all stations given in Figure 7 (d) further showed that TAMSAT 2 and ARC 2 consistently underestimate rain gauge values, while CHIRPS and TAMSAT 3 show very close estimation, with a better performance from CHIRPS. The relatively good performance of CHIRPS at different elevations is partly due to the inclusion of typical physiographic indicators such as elevation during the development of the datasets (Funk et al., 2015). These could make CHIRPS a relatively better satellite rainfall product that might be used in complex topographic areas, such as the UBN basin,

30   to detect the pattern and variability of precipitation.

A possible explanation for TAMSAT 2 and ARC 2 overestimations at lower elevation might be the deep convective nature of Inter Tropical Convergence Zone (ITCZ), the main rain producing mechanism in Ethiopia (Seleshi and Zanke, 2004), in the lower elevation areas that results in too deep cold clouds that may stay for a number of days. The underestimations at higher elevation could be linked to the potential evaporation of rainfall at the cloud base in high altitude areas. However,

results (in Figure 6, Figure 7 and Table 2) provide confirmatory evidence that the recent TAMSAT product (TAMSAT 3) has addressed many of the weaknesses of TAMSAT 2 at complex topographic areas, particularly to the bias ratios, and the improvement in this regard is very encouraging.

Besides, Figure 6 (c) shows that the RMSE of satellite products have no significant relationship with elevations. Nevertheless, CHIRPS and TAMSAT 3 have scored the least average RMSE (30.02 and 32.24 mm/dekad) in comparison to ARC 2 and TAMSAT 2 RMSE (38.44 and 38.13 mm/dekad, respectively).

5.3 Monthly comparison

The daily rain gauge observation and TAMSAT 3 and ARC 2 products were aggregated to monthly total rainfall, while CHIRPS and TAMSAT 2 satellite rainfall products are available at monthly time scale. The monthly comparison was made using i) all monthly values from rain gauge observation and satellite products and ii) classified the monthly values into twelve classes for further validation of the satellite products for each specific month of the UBN basin.

5.3.1 Overall comparison at monthly temporal scale

Table 3 presents the summary of the overall monthly validation results. Figure 8 shows the scatter plot between rain gauge observations and satellite rainfall estimates at monthly temporal scale. In general, the overall monthly comparisons between the four SREs and the rain gauge observations have shown a better agreement than the comparison at dekadal temporal scale. This is quite expected because errors at sub-monthly scale are showing closely symmetric characteristics and they may finally cancel each other out following the aggregation to monthly temporal scale. The monthly comparisons, shown in Table 3, indicated CHIRPS' better performance in most validation tools than TAMSAT and ARC 2. However, CHIRPS is still with high FAR values and overestimates the frequency of rainfall events by 14%, but its monthly FAR value is much improved in comparison to the dekadal time-scale analysis (FAR=0.31). Comparing the TAMSAT products, TAMSAT 2 has outperformed the new version TAMSAT 3 in the scores of POD and CSI, while they gained equal values in FAR, VFAR and VCSI values. ARC 2 exhibited the lowest categorical and volumetric values.

Additionally, from the continuous statistical analysis in Table 3 and the scatter plot in Figure 8, a good agreement was found between rain gauge observations and all the four SREs ($r>=0.80$). CHIRPS scored the highest correlation coefficient ($r=0.88$) and the least RMSE (59.03 mm month^{-1}). While ARC 2 resulted in the largest RMSE (79.21 mm month^{-1}) and the weakest but fairly good correlation coefficients ($r= 0.80$). On the other hand, CHIRPS and TAMSAT 3 satellite products resulted in bias values close to the perfect score 1.00, whereas TAMSAT 2 and ARC 2 showed poor bias ratios and underestimated monthly gauge observed rainfall by 31% and 24%, respectively. In this respect, a lot has been done in the recent version of TAMSAT 3 and has significantly improved the weak bias values of the previous version TAMSAT 2.

Overall, the skill of CHIRPS is still better than the other satellite rainfall estimates in the monthly time scale analysis as well. In fact, TAMSAT 3 has shown a comparable performance and very close scores, in the majority of validation tools, with CHIRPS, particularly to bias ratio, similar to the dekadal time scale analysis above.

5.3.2 Comparison at each month

The performances of the satellite rainfall products were also evaluated for each month of the UBN basin, where different amount of rainfall is recorded. Thus, the monthly data from all the 32 stations for the validation period of 2000-2015 (both from SREs and rain gauge observations) were categorized into 12-month classes. Months from June to September (wet months) contribute the largest proportion of annual rainfall in the study area, followed by the dry months (from October to January) and little rainfall months (from February to May). Figure 9 and 10, illustrate the performances of all the four SREs for the categorical, volumetric and continuous statistical validation tools.

The categorical and volumetric analysis, presented in Figure 9, for each month revealed that the performances of all the four satellite rainfall products are very encouraging during the wet months and have good agreement with rain gauge observations, shown in the lower semicircle of the polar plot (i.e., high POD, VHI, CSI, VCSI and low FAR and VFAR). A similar result has been obtained by Young et al. (2014) and Dinku et al. (2011b) during the wettest periods in Ethiopian highland and over the upper Nile region in Ethiopia, respectively, using TAMSAT 2, ARC 2, TRMM and CMORPH satellite rainfall products. This might be because the numbers of hit values are noticeably larger than the number of missed and false events during the wet months. Over the upper semicircle of the polar plot in the same figure (Figure 9), dominated by dry and small rainfall months, however, the satellite products have shown a relatively wider difference in their skills. CHIRPS has scored the highest POD, VHI, CSI and VCSI values in comparison to TAMSAT and ARC 2 products. A comparable finding has been reported by Tote et al. (2015) and Young et al. (2014). However, CHIRPS is still with high FAR (up to 0.4) and VFAR (0.31) values, particularly during the months of January. The increased in FAR and VFAR values of CHIRPS is because of the over-detecting of rainfall events that can perhaps be linked to its calibration with TMPA 3B42. The rather weak performance of TAMSAT and ARC 2 products during the dry and small rainfall months could be associated with low frequency of rain events ensuring to less amount of rainfall to be detected by the satellites. Overall, the result highlighted that the skill of CHIRPS is relatively better than TAMSAT and ARC 2 products and has a good agreement with rain gauge observed rainfall data both in the wet and dry seasons, although it over-predicts rainfall events particularly for dry and small rainfall months.

As can be seen from the continuous statistical validation presented in Figure 10 (a), the correlation coefficient for all four satellite rainfall products are generally low (as low as r=0.03) during the dry months shown in the upper semicircle of the polar plot, except for the month of October. However, over the little rainfall and wet months, the correlation coefficient for CHIRPS was relatively high in comparison to TAMAT and ARC2, except for the months of February and May, with values of r=0.53, r=0.82, r=0.72 and r=0.77 for the months of March, June, July, and September, respectively. TAMSAT 3 has also scored comparable correlation values during these months next to CHIRPS, while TAMSAT 2 and ARC 2 scored the weakest values. At the months of February (r=0.58) and May (r=0.79), the highest correlation coefficients were recorded by TAMSAT 2 and TAMSAT 3, respectively. Over the months of December, April and August all the four SREs scored low correlation values.

Further, in Figure 10 (b), CHIRPS has shown better bias ratio in all months, except at the months of November (0.78) and December (0.70) and a little overestimation (1.13) at the months of February, when a mainly small amount of rainfall recorded. This result is consistent with CDF in Figure 3, where CHIRPS is less above rain gauge observation towards to the little rainfall accumulations. TAMSAT 3 scored the second best bias ratio next to CHIRPS, except for the months of December to March, in which it considerably underestimates rain gauge observed rainfall. On the other hand, TAMSAT 2 and ARC 2 result in a weak bias ratio for all months, mainly to dry months indicated in the upper semicircle of the polar plot (Figure 10 (b)). Overall, the dependency of CHIRPS bias ratio on the monthly temporal pattern, particularly during the wet seasons, is very minimal in comparison to TAMSAT 3. These would appear to indicate that the potential of CHIRPS satellite rainfall estimates for hydrological functions. Following the performance of CHIRPS during months of high rainfall Trejo et al. (2016) has also suggested it for hydrological applications. For hydrological monitoring, it is vital to accurately estimate significant rain events (Dinku et al., 2007). CHIRPS scored the lowest RMSE followed by TAMSAT 3. TAMSAT 2 and ARC 2 presented the relatively largest values of RMSE (Figure 10 (c)). In fact, RMSE is higher in the wet months due to increased amounts of rainfall.

6 Conclusions

This study was set out with the aim of evaluating the performance of CHIRPS satellite rainfall estimates against 32 rain gauge observations over the Upper Blue Nile (UBN) basin in Ethiopia for the period of 2000 to 2015.Then, the performance of CHIRPS was compared with TAMSAT (TAMSAT 2 and TAMSAT 3) and ARC 2 rainfall products. In the course of the analysis, the TAMSAT and ARC 2 products were validated as well. The TAMSAT 2 rainfall estimate was used in this study mainly to assess the improvements made by the recent version of TAMSAT product (TAMSAT 3). A point-to-grid based comparison was carried out at dekadal and monthly temporal time scale using categorical, volumetric and continuous statistical validation tools. The dekadal and monthly time scale data were further utilized for the validation of the SREs at different elevations and for each particular month of the UBN basin, respectively.

From the overall validation at dekadal and monthly temporal scale, CHIRPS has shown the highest skill, the lowest RMSE, and the better bias values than TAMSAT 3 and ARC 2. Indeed, TAMSAT 3 has scored very comparable values with CHIRPS product, particularly to the bias ratio. While ARC 2 underestimates rain gauge observed rainfall by 24%. Although CHIRPS over-predict rainy days (i.e. high false alarm rate), its volumes of false alarm ratio's are much reduced, and its overall performance is significantly improved and was better than TAMSAT 3 and ARC 2. Since the volumes of rainfall detected by CHIRPS during false events were tiny, it had a minimal contribution to the total amounts of rainfall. The findings of this study, therefore, indicated that event-based analysis solely might not be enough to verify the skill of the satellite rainfall product as small rain events might lead to wrong conclusions.

Validation at different elevations indicated that all the SREs have generally good agreement with rain gauge observations and their performances are independent of elevations, except at the skills of detecting rainfall events (POD) by TAMSAT 3

and ARC 2. The POD of TAMSAT 3 and ARC 2 have a considerable negative correlation (r=-55) with elevations and their skills of detecting rainfall events reduced with an increase in elevations, while CHIPS result in a relatively small positive correlation (r=0.34). Comparing all the satellite rainfall products, CHIRPS still scores better values at most elevations. In fact, TAMSAT 3 has also scored a comparable average bias ratio (1.07), pretty close to the CHIRPS' perfect score of 1.00. Besides, the bias ratio of TAMSAT 2 and ARC 2 seems affected by variation in elevation.

Generally, the validation for each specific month of the study area indicated that the performances of SREs are better during the wet months, except for the RMSE, and has a good agreement with rain gauge observations. In fact, RMSE is higher in the wet months due to increased amounts of rainfall. The best values were scored by CHIRPS, closely followed by TAMSAT 3 particularly for the correlation coefficient and the bias ratio. However, over the majority of little rainfall and dry months, the SREs have shown their weak performance, especially for POD and VHI (TAMSAT and ARC 2), FAR and VFAR (CHIRPS) and CSI, VCSI, correlation coefficient, and bias (all the four SREs). However, the overall skill of CHIRPS is relatively good during these months as well and was better than TAMSAT 3 and ARC 2. A good performance has also been observed from TAMSAT 3 next to CHIRPS, particularly to the bias ratios.

Summing up the results, the performance of CHIRPS in the UBN basin is very encouraging and relatively better than the other satellite rainfall products (TAMSAT and ARC 2). More specifically, the reliable performance of CHIRPS at different elevations and during the wet months could make the product more appropriate for various hydrological and rainfall analysis functions in complex topographic areas, such as the UBN basin. The performance of TAMSAT 3 is very comparable to CHIRPS product and score close values to CHIRPS in many of the validation indicators, particularly to the bias ratios. This validation study has also provided confirmatory evidence that the recent version of TAMSAT product (TAMSAT 3) has well addressed many of the weaknesses of TAMSAT 2 (e.g., underestimations up to 31% in this study) at complex topographic areas, and the improvement in this regard is very encouraging. Future work will involve on validation of the product at different rainfall categories, spatial and temporal scale as well as during the drought and wet periods for complete understandings of its potential.

Acknowledgements

This research was supported in part by the NASA Interdisciplinary Research in Earth Sciences Program, NASA Greater Horn of Africa Project NNX14AD30G and Entoto Observatory & Research Center postgraduate program research fund. We are grateful to the data providers of CHIRPS, TAMSAT and ARC 2. The National Meteorological Agency of Ethiopia is also well acknowledged for providing gauge station rainfall data found in the UBN basin.

References

AghaKouchak, A. and Mehran, A.: Extended contingency table: Performance metrics for satellite observations and climate model simulations, Water Resour. Res., 49, 7144–7149, doi:10.1002/wrcr.20498, 2013.

Conway, D.: Some aspects of climate variability in the northeast Ethiopian highlands-Wollo and Tigray, Ethiopian Journal of Science, 23, 139–161, 2000.

Conway, D.: From headwater tributaries to international river: Observing and adapting to climate variability and change in the Nile basin, Global Environmental Change, 15, 99-114, doi: 10.1016/j.gloenvcha.2005.01.003, 2005.

5 Degefu, G.T.: The Nile Historical Legal and Developmental Perspectives, Trafford Publishing, St. Victoria, Canada, 2003.

Dinku, T., Ceccato, P., Grover-Kopec, E, Lemma, M., Connor, S., and Ropelewski, C.:Validation of satellite rainfall products over East Africa's complex topography, International Journal of Remote Sensing, 28, doi: 10.1080/014311606009546881503-1526, 2007.

Dinku, T., Chidzambwa, S., Ceccato, P., Connor, S., and Ropelewski, C.: Validation of high resolution satellite rainfall products over complex Terrain'. International Journal of Remote Sensing, 29, 4097–4110, doi: 10.1080/01431160701772526, 2008.

Dinku, T., Ceccato, P., and Connor, S.: Challenges of satellite rainfall estimation over mountainous and arid parts of east Africa, International Journal of Remote Sensing, 32, 5965–5979, doi: 10.1080/01431161.2010.499381, 2011a.

Dinku, T., Connor, S., and Ceccat, P.: Evaluation of satellite rainfall estimates and gridded gauge products over the Upper Blue Nile region, In Nile River Basin, Melesse AM (ed). Springer: Dordrecht, The Netherlands, Heidelberg, Germany, London, and New York, NY, 109–127, 2011b.

Dugdale, G., McDougall, V., and Milford, J.:1991, 'Rainfall estimates in the Sahel from cold cloud statistics: Accuracy and limitations of operational systems, in: Proceedings of the Niamey Workshop on Soil Water Balance in the Sudano-Sahelian Zone, IAHS Publication, February 1991, 51–67, 1991.

20 Fenta, A., Rientjes, T., Haile, A., and Reggiani, P.: Satellite rainfall products and their reliability in the Blue Nile Basin, In Nile River Basin, Melesse AM et a. (ed). Springer International Publishing Switzerland, 109–127, 2014.

Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., Husak, G., Rowland, J., Harrison, L., Hoell, A., and Michaelsen, J.: The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes, Sci. Data., 2, 1-21, doi: 10.1038/sdata.2015.66, 2015.

25 Gebere, S., Alamirew, T., Merkel, B., and Melesse, A.: Performance of High Resolution Satellite Rainfall Products over Data Scarce Parts of Eastern Ethiopia, Remote sensing,7, 11639-11663, doi:10.3390/rs70911639, 2015.

Gebremichael, M., Krajewski, W., Morrissey, M., Huffman, G., and Adler, R.: A detailed evaluation of GPCP one-degree daily rainfall estimates over the Mississippi River Basin, J. Appl. Meteorol., 44, 665–681,2005.

Gebremichael, M., Bitew, M., Hirpa, F., and Tesfaye, G.: Accuracy of satellite rainfall estimates in the Blue Nile Basin, Lowland plain versus highland mountain, Water Resour. Res., 50, 8775–8790, doi: 10.1002/2013WR014500, 2014.

30 Grimes, D., Pardo-Igzquiza, E., and Bonifacio, R.: Optimal areal rainfall estimation using rain gauges and satellite data', J. Hydrol., 222, 93–108, 1999.

Hirpa, F., Gebremichael, M., and Hopson, T.: Evaluation of High-Resolution Satellite Precipitation Products over Very Complex Terrain in Ethiopia', J.Appl. Meteorol. Clim., 49,1044-1051, doi: 10.1175/2009JAMC2298.1, 2010.

35 Hsu, K., and Sorooshian, S.: Satellite-based precipitation measurement using PERSIANN system, in Hydrological Modelling and the Water Cycle, *Springer, Berlin, Heidelberg,* 63, 27–48. 2008.

Huang, J. and van den Dool, H.: Monthly precipitation temperature relations and temperature prediction over the United States', Journal of Climate, 6, 1111–1132, 1993.

Huffman, G., Bolvin, D., Nelkin, E., Wolff, D., Adler, R., Gu, G., Hong, Y., Bowman, K., and Stocker, E. : The TRMM Multisatellite Precipitation Analysis (TMPA): Quasi- Global, Multiyear, Combined-Sensor Precipitation Estimates at Fine Scales, *J. Hydrometeorol.,* 8, 38–55, 2007.

Joyce, R., Janowiak, J., Arkin, P., and Xie, P.: CMORPH: A Method That Produces Global Precipitation Estimates from Passive Microwave and Infrared Data at High Spatial and Temporal Resolution, *Journal of Hydrometeorology*, 5, 487–503, 2004.

45 Kim, U., Kaluarachchi, J., and Smakhtin, V.: Generation of Monthly precipitation under climate change for the upper Blue Nile River Basin, Ethiopia, J.Am. Water Resources, 44,1231-1247, 2008.

Kummerow, C., Hong, Y., Olso, W., Yang, S., Adler, R., McCollum, J., Ferraro, R., Petty, G., Shin, D., and Wilheit, T.: The evolution of the Goddard Profiling Algorithm (GPROF) for rainfall estimation from passive microwave sensors, J. Appl. Meteorol., 40, 1801-1820, 2001.

50 Maidment, R., Grimes, D., Allan, R., Tarnavsky, E., Stringer, M., Hewison, T., Roebeling, R., and Black, E.: The 30 year

TAMSAT African Rainfall Climatology and Time series (TARCAT) data set, J. Geophys. Res. Atmos., 119, 619–10, doi: 10.1002/2014JD021927, 2014.

Maidment, R., Grimes, D., Black, E., Tarnavsky, E., Young, M., Greatrex, H., Allan, R., Stein, T., Nkonde, E., Senkunda, S., and Alcántara, E.: A new, long-term daily satellite-based rainfall dataset for operational monitoring in Africa, Sci. Data., 4:170063., doi: 10.1038/sdata.2017.63, 2017.

Novella, N. and Thiaw, W.: African rainfall climatology version 2 for famine early warning systems, J. Appl. Meteorol. Climatol., 52, 588–606, doi: 10.1175/JAMC-D-11-0238.1, 2013.

Romilly, T. and Gebremichael, M.: Evaluation of satellite rainfall estimates over Ethiopian river basins, Hydrol. Earth Syst. Sci., 15, 1505–1514, doi:10.5194/hess-15-1505-2011, 2011.

Seleshi, Y. and Zanke, U.: Recent changes in rainfall and rainy days in Ethiopia', Int. J. Climatol., 24, 973-983, doi: 10.1002/joc.1052, 2004.

Stillman, S., Ninneman, J., Zeng, X., Franz, T., Scott, R., Shuttleworth, W., and Cummins, K.: Summer Soil Moisture Spatiotemporal Variability in Southeastern Arizona, J. Hydrometeor., 15, 1473-1485. doi: 10.1175/JHM-D-13-0173.1, 2014.

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

Taye, M. and Willems, P.: Temporal variability of hydroclimatic extremes in the Blue Nile basin, Water Resour. Res., 48, 1-13, doi: 10.1029/2011WR011466, 2012.

Thorne, V., Coakeley, P., Grimes, D., and Dugdale, G.: Comparison of TAMSAT and CPC rainfall estimates with rain gauges for southern Africa, Int. J. Remote Sens., 22, 1951–1974, 2001.

Toté, C., Patricio, D., Boogaard, H., van der Wijngaart, R., Tarnavsky, E., and Funk, C.: Evaluation of Satellite Rainfall Estimates for Drought and Flood Monitoring in Mozambique, Remote Sensing, 7, 1758-1776, doi:10.3390/rs70201758, 2015.

Trejo, F., Barbosa, H., Murillo, M., and Farias, M.: Intercomparison of improved satellite rainfall estimation with CHIRPS gridded product and rain gauge data over Venezuela, Atmósfera, 29, 323-342, doi: 10.20937/ATM.2016.29.04.04, 2016.

Worqlul, A., Maathuis, B., Dame, A., Demissie, S., Langan, S., and Steenhuis, T.: Comparison of rainfall estimations by TRMM 3B42, MPEG and CFSR with ground-observed data for the Lake Tana basin in Ethiopia, Hydrol. Earth Syst. Sci., 18, 4871–4881, doi:10.5194/hess-18-4871-2014, 2014.

Wu, H., Adler, R., Hong, Y., Tian, Y., and Policelli, F.: Evaluation of global flood detection using satellite-based rainfall and a hydrologic model, J. Hydrometeor.,13,1268-1284, doi: 10.1175/JHM-D-11-087.1, 2012.

Young, M., Williams, C., Chiu, J., Maidment, R., and Chen, S.: Investigation of Discrepancies in Satellite Rainfall Estimates Over Ethiopia, J. Hydrometeor.,15, 2347-2369, doi: 10.1175/JHM-D-13-0111.1, 2014.
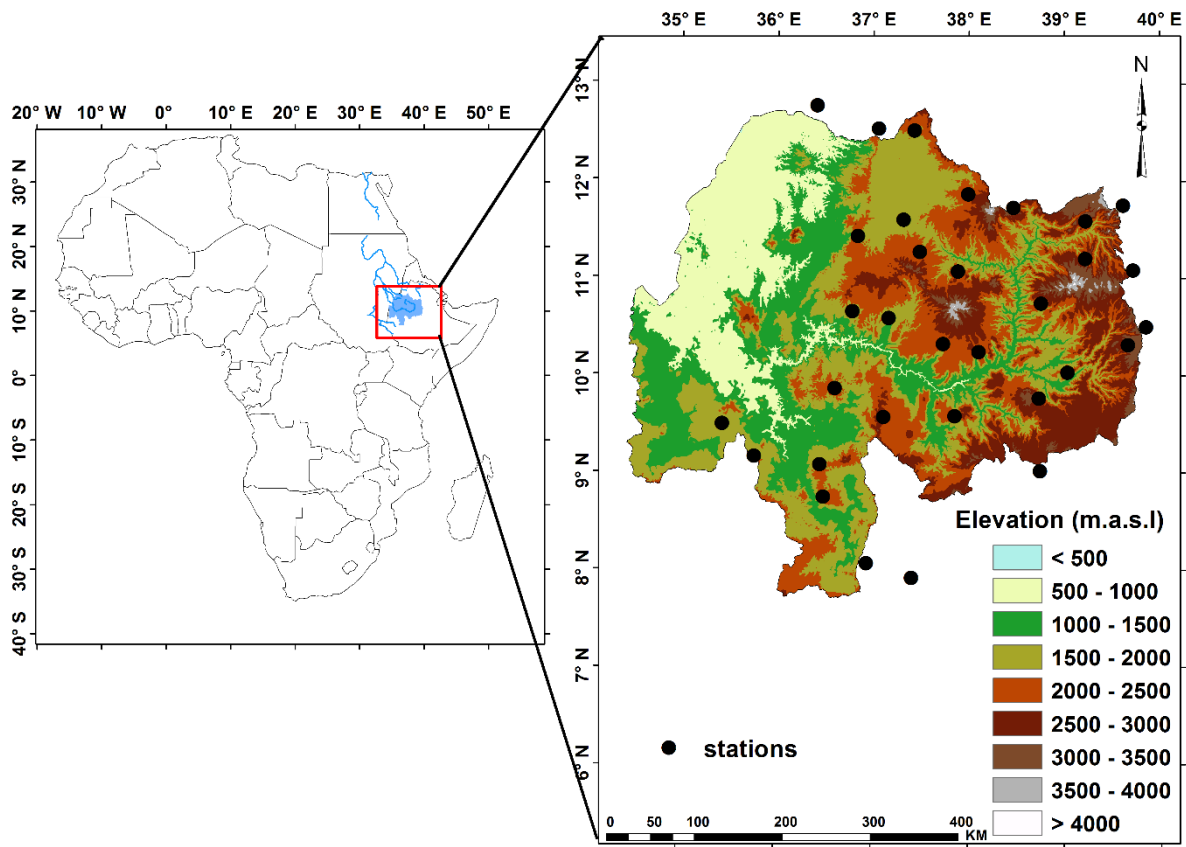
Figure 1: Elevations map of the Upper Blue Nile (UBN) basin and its location in Africa. The northeastern regions are with high elevation scenery while the northwestern regions are with low elevations.
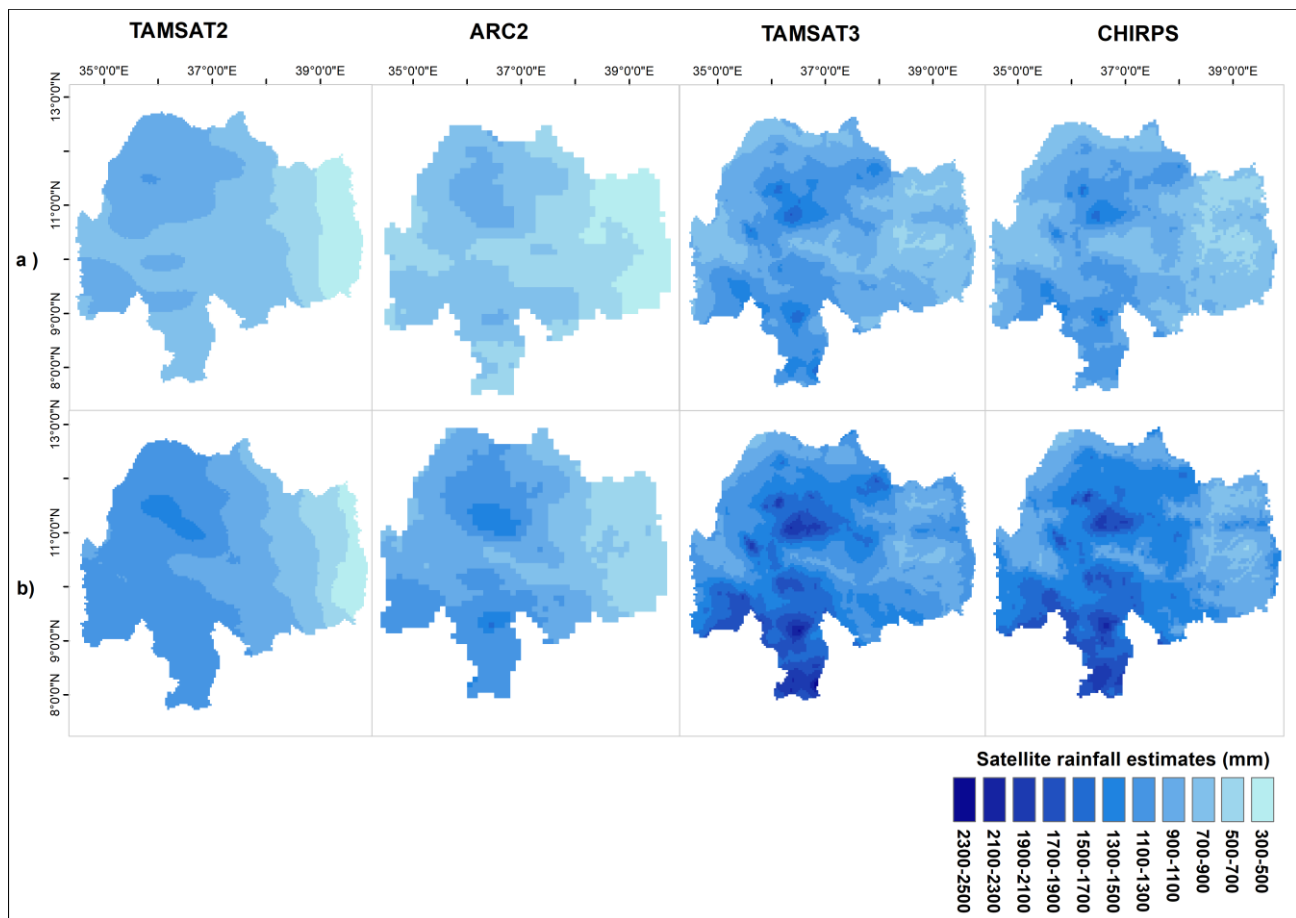
Figure 2: Comparison of mean satellite rainfall estimates for: (a) Kiremt season (June-September) and (b) annual rainfall over the Upper Blue Nile basin for the period of 2000-2015. Years with missed values were not considered in the mean analysis.

Table 1. Summary of the point-to-grid evaluation at dekadal temporal scale using categorical, volumetric and continuous statistical tools. Probability of Detection (POD), False Alarm Ratio (FAR), Critical Success Index (CSI), Volumetric Hit Index (VHI), Volumetric False Alarm Ratio (VFAR), Volumetric Critical Success Index (VCSI), correlation coefficient (*r*), bias and the Root Mean Square Error (RMSE). The RMSE values are shown in millimeters.

Datasets	POD	FAR	CSI	VHI	VFAR	VCSI	*r*	bias	RMSE
ARC 2	0.75	0.06	0.71	0.91	0.03	0.89	0.72	0.76	35.02
TAMSAT 2	0.83	0.09	0.77	0.94	0.03	0.91	0.76	0.69	34.03
TAMSAT 3	0.83	0.09	0.76	0.96	0.03	0.93	0.78	1.04	32.19
CHIRPS	0.99	0.31	0.68	1.00	0.06	0.94	0.81	0.96	28.45
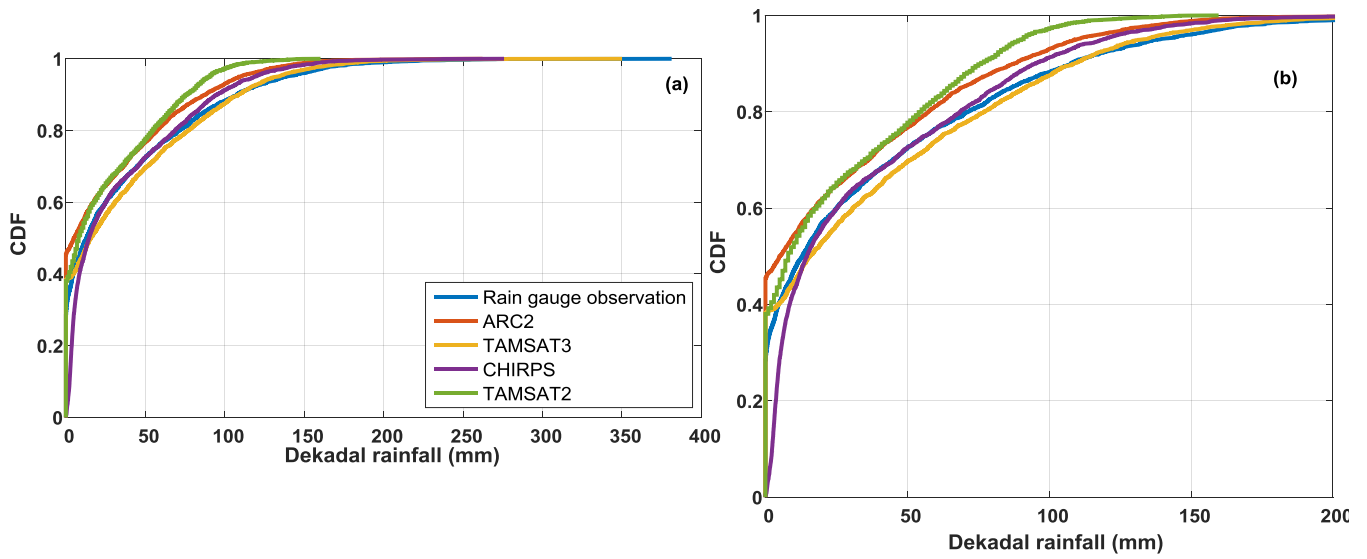
Figure 3: Cumulative distribution function (CDF) of dekadal rainfall for (a) ground rainfall observation, ARC 2, TAMSAT, and CHIRPS rainfall estimates and (b) magnified view of their CDF for 0 to 200 mm part over the Upper Blue Nile basin for the period of 2000-2015.
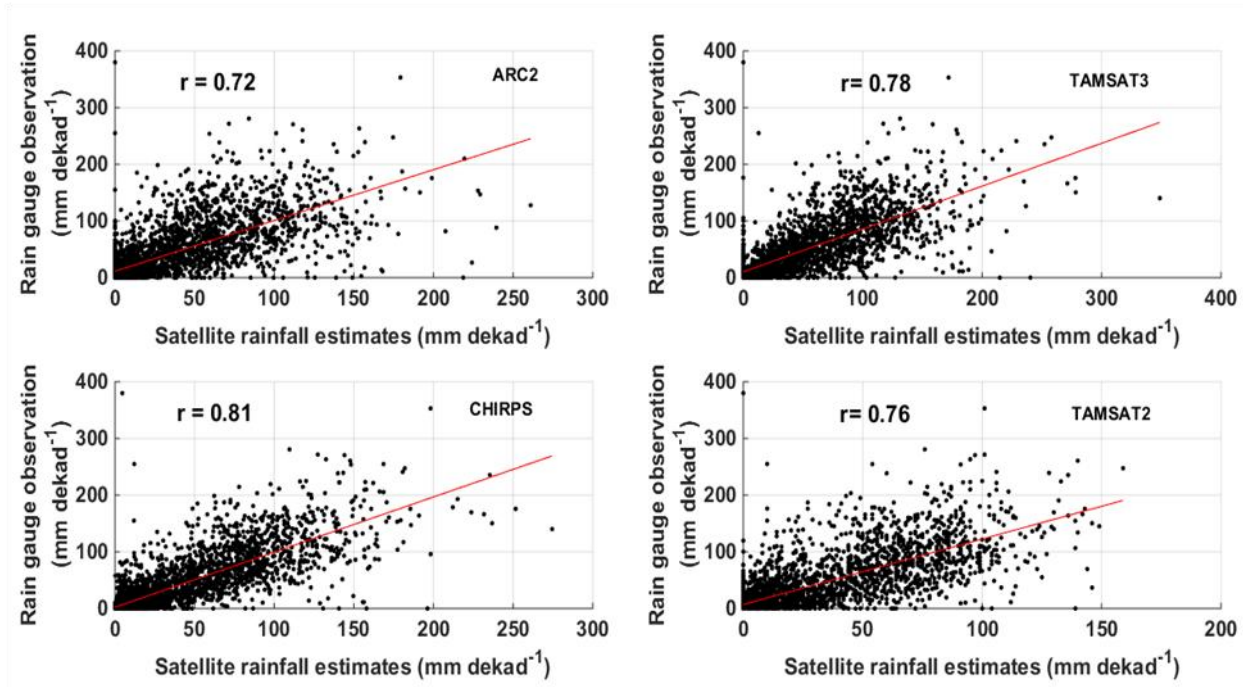
Figure 4: Scatter plot between rain gauge observations and satellite rainfall estimates at dekadal temporal scale over the Upper Blue Nile basin for the period of 2000-2015.
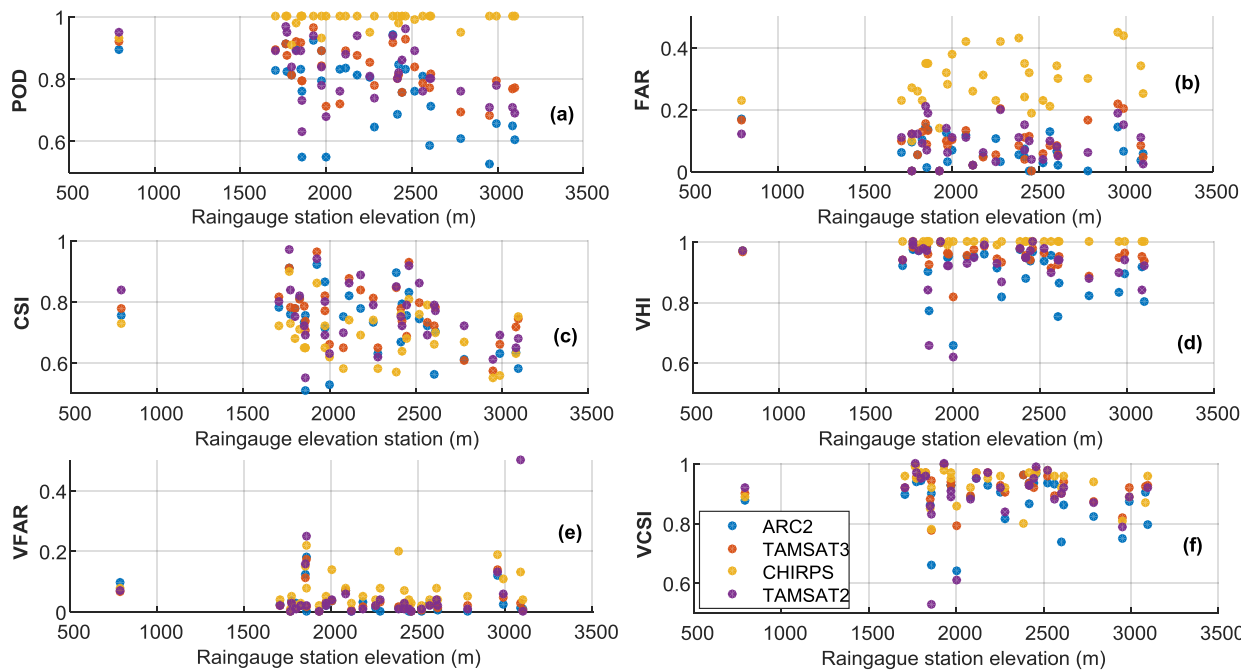
Figure 5: Categorical and volumetric indices of satellite rainfall products as a function of elevation: (a) Probability of Detection (POD), (b) False Alarm Ratio (FAR), (c) Critical Success Index (CSI), (d) Volumetric Hit Index (VHI), (e) Volumetric False Alarm Ratio (VFAR), and (f) Volumetric Critical Success Index (VCSI) over the Upper Blue Nile basin for the period of 2000-2015.

Table 2. Pearson correlation between the skills of SREs and station elevations (only important correlations are presented here). Probability of Detection (POD), Critical Success Index (CSI), and bias.

Indices	Stations elevation			
	ARC 2	TAMSAT 3	CHIRPS	TAMSAT 2
POD	-0.55	-0.55	0.34	-0.44
CSI	-0.43	-0.38	-0.26	-0.31
bias	-0.44	0.18	0.10	-0.39
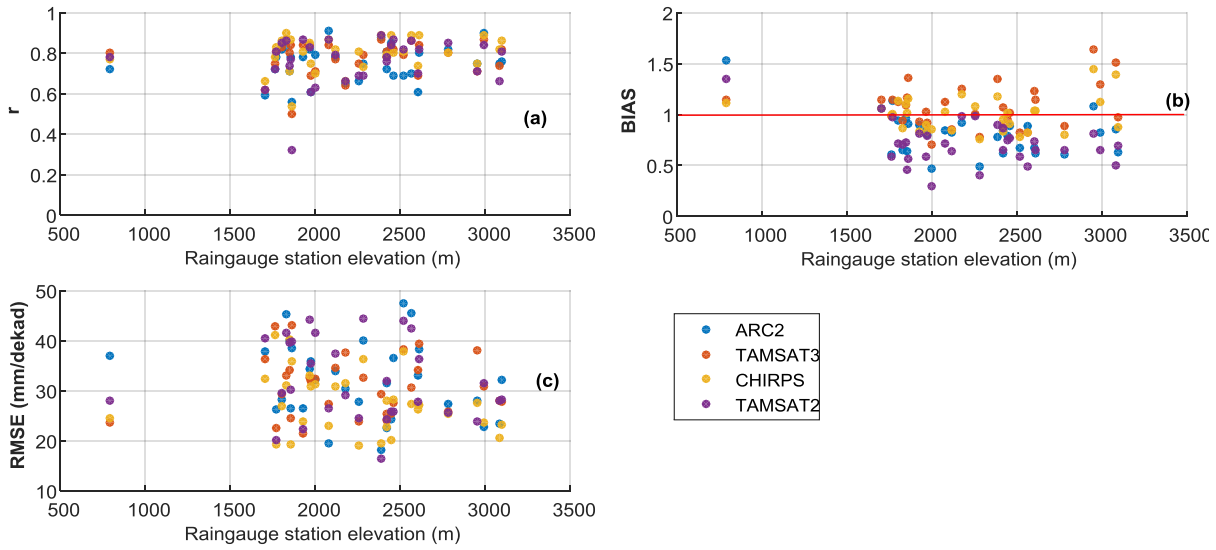
Figure 6: Statistical validation of the satellite products as a function of elevation: (a) Pearson correlation coefficient (*r*), (b) bias, and (c) the Root Mean Square Error (RMSE) over the Upper Blue Nile basin for the period of 2000-2015.
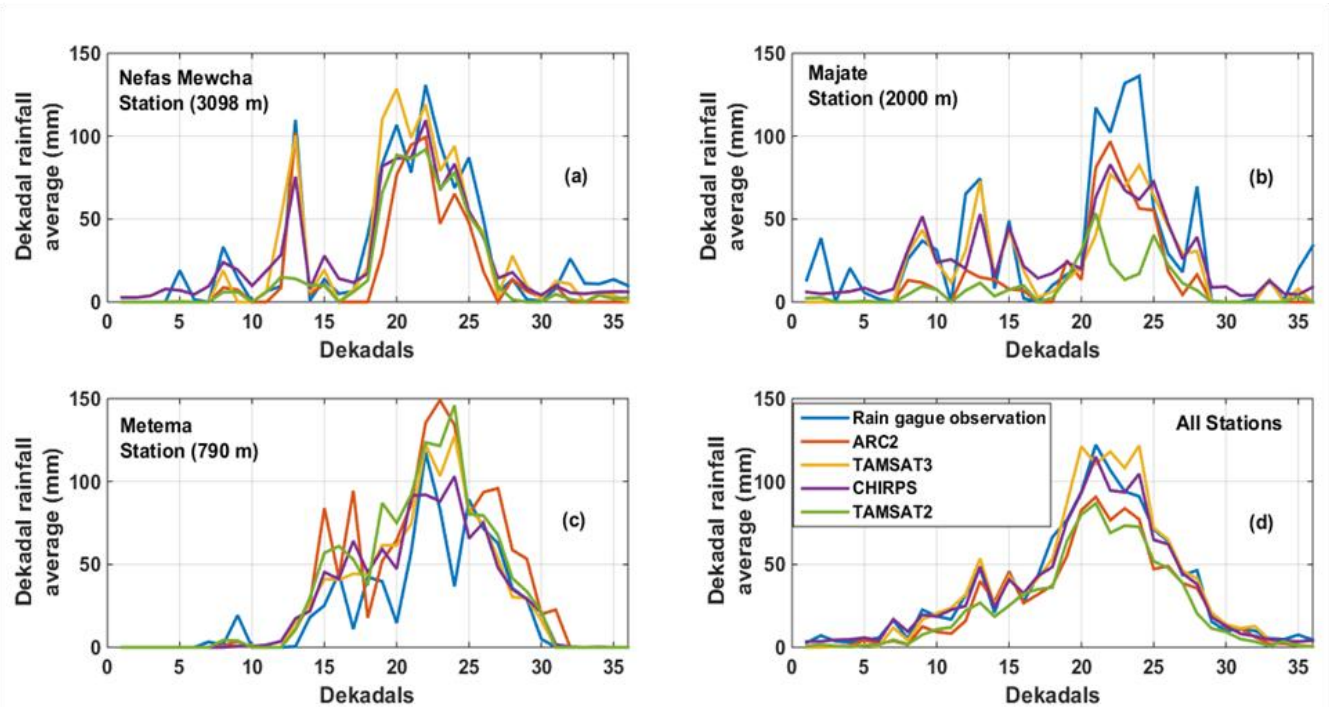
Figure 7: Comparison of the satellite products at gauge stations with wider difference in elevation values (e.g., > 2000m), based on dekadal average over the Upper Blue Nile basin for the period of 2000-2015: (a) at "Nefas Mewucha" station with an elevation of 3098 m a.s.l., (b) at "Majate" stations with an elevations of 2000 m a.s.l., (c) at "Metema" stations with an elevation of 790 m a.s.l., and (d) at dekadal rainfall average from all rain gauge stations. The x-axis represents the 36 dekadals of a year.

Table 3. Summary of the point-to-grid evaluation at monthly temporal scale using categorical, volumetric, and continuous statistical tools. Probability of Detection (POD), False Alarm Ratio (FAR), Critical Success Index (CSI), Volumetric Hit Index (VHI), Volumetric False Alarm Ratio (VFAR), Volumetric Critical Success Index (VCSI), correlation coefficient (*r*), bias and the Root Mean Square Error (RMSE). The RMSE values are shown in millimeters.

Datasets	POD	FAR	CSI	VHI	VFAR	VCSI	*r*	bias	RMSE
ARC 2	0.78	0.03	0.76	0.95	0.02	0.93	0.80	0.76	79.21
TAMSAT 2	0.86	0.04	0.83	0.97	0.01	0.96	0.83	0.69	78.65
TAMSAT 3	0.83	0.04	0.80	0.97	0.01	0.96	0.85	1.03	69.28
CHIRPS	1.00	0.14	0.86	1.00	0.02	0.98	0.88	0.96	59.03
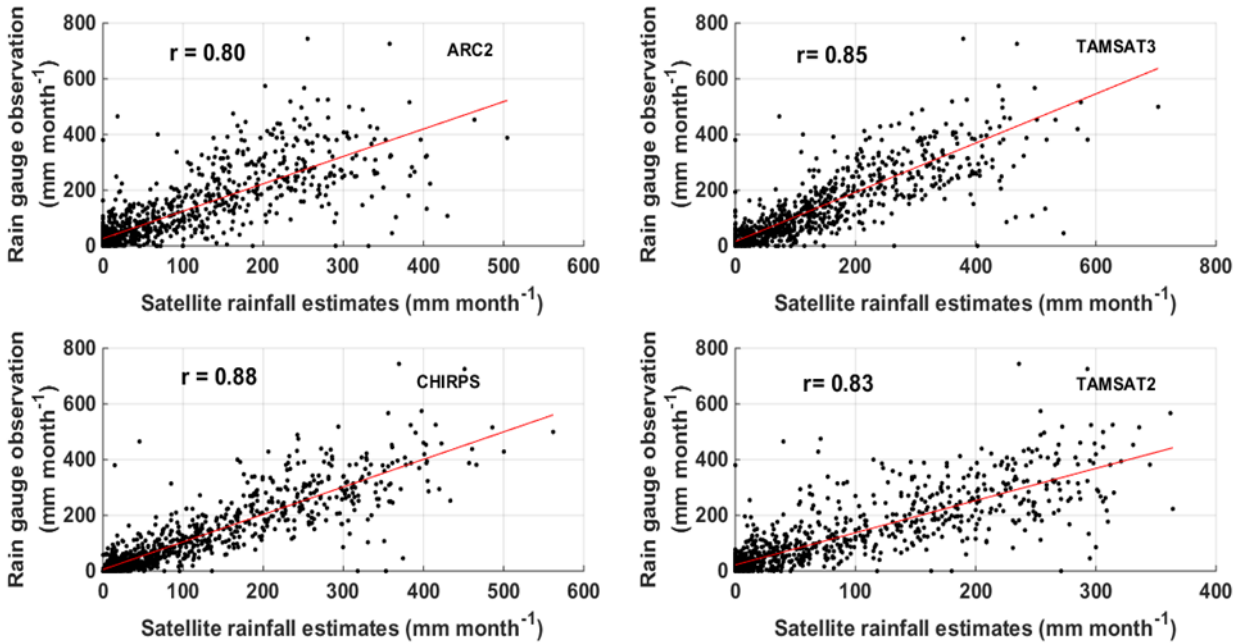
Figure 8: Scatter plot between rain gauge observations and satellite rainfall estimates at monthly temporal scale over the Upper Blue Nile basin for the period of 2000-2015.
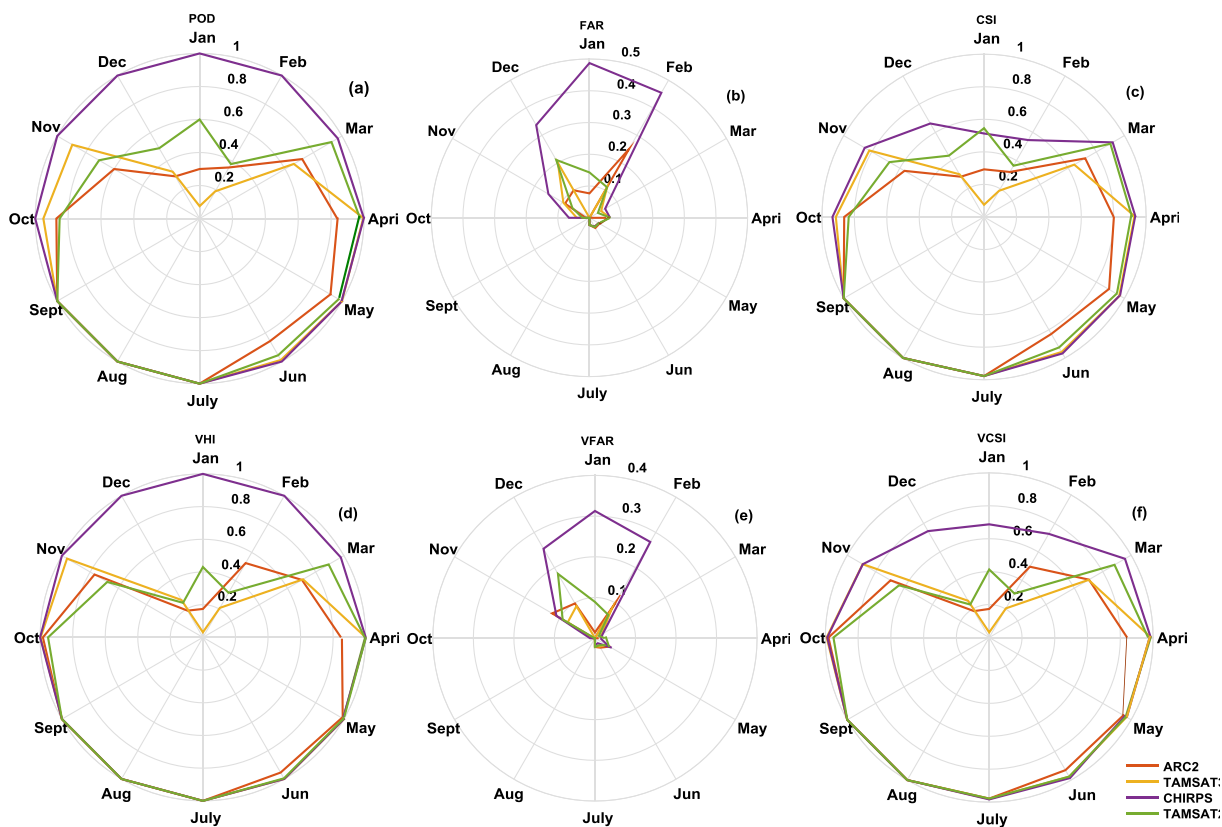
Figure 9: Categorical and volumetric validation of the satellite products for each month of the Upper Blue Nile basin for the period of 2000-2015: (a) Probability of Detection (POD), (b) False Alarm Ratio (FAR), (c) Critical Success Index (CSI), (d) Volumetric Hit Index (VHI), (e) Volumetric False Alarm Ratio (VFAR), and (f) Volumetric Critical Success Index (VCSI).
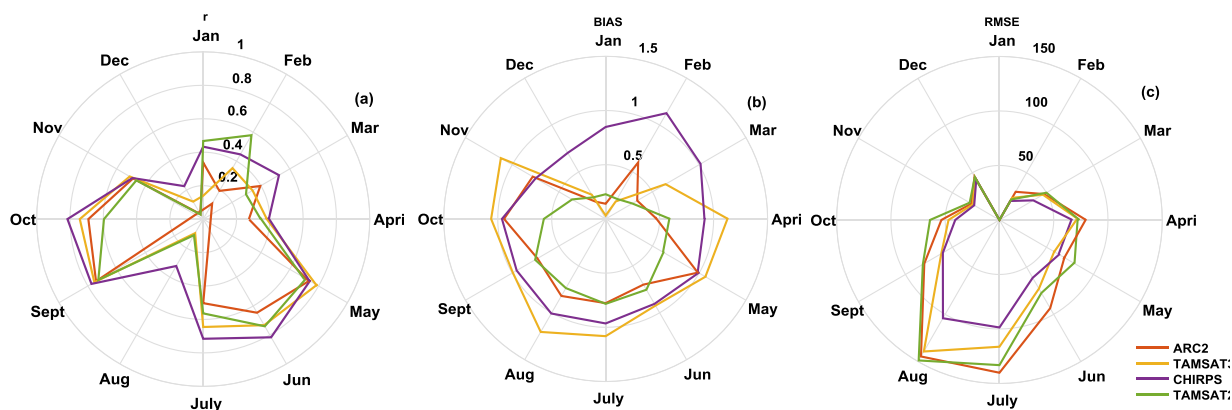
Figure 10: Statistical validation of the satellite products for each month of the Upper Blue Nile basin for the period of 2000-2015: (a) Pearson correlation coefficient (*r*), (b) bias ratio, and (c) the Root Mean Square Error (RMSE). The RMSE values are shown in millimeters.