



Application of Machine Learning and Remote Sensing for Gap-filling Daily Precipitation Data of a Sparsely Gauged Basin in East Africa

Marzie Faramarzadeh¹ · Mohammad Reza Ehsani² · Mahdi Akbari³ ·
Reyhane Rahimi⁴ · Mohammad Moghaddam² · Ali Behrangi² · Björn Klöve³ ·
Ali Torabi Haghighi³ · Mourad Oussalah¹

Received: 3 October 2022 / Accepted: 7 February 2023 / Published online: 13 February 2023
© The Author(s) 2023

Abstract

Access to spatiotemporal distribution of precipitation is needed in many hydrological applications. However, gauges often have spatiotemporal gaps. To mitigate this, we considered three main approaches: (i) using remotely sensing and reanalysis precipitation products; (ii) machine learning-based approaches; and (iii) a gap-filling software explicitly developed for filling the gaps of daily precipitation records. This study evaluated all approaches over a sparsely gauged basin in East Africa. Among the examined precipitation products, PERSIANN-CDR outperformed other satellite products in terms of root mean squared error (7.3 mm), and correlation coefficient (0.46) while having a large bias (50%) compared to the available in situ precipitation records. PERSIANN-CDR also demonstrates the highest skill in distinguishing rainy and non-rainy days. On the other hand, Random Forest outperformed all other approaches (including PERSIANN-CDR) with the least relative bias (-2%), root mean squared error (6.9 mm), and highest correlation coefficient (0.53).

Highlights

- Ways to fill in spatiotemporal gaps in gauge measurements were explored.
- Satellite and reanalysis, machine learning, and gap-filling software were investigated.
- Random Forest performed the best among all other methods to fill in gaps.

Keywords Gap-filling · Deep learning · Random forest · ReddPrec · Precipitation products · Machine learning

1 Introduction

Precipitation is probably the most important component of the hydrological cycle (Eltahir 1998; Oki and Kanae 2006; Saemian et al. 2021; Sun et al. 2018; Ghajarnia et al. 2022). In addition to being the main source of renewable water, precipitation is also critical for the socioeconomic development of nations, especially African countries that depend on rain-fed agriculture (Dinku et al. 2007; Awange et al. 2016). In recent years and due to climate change, most African regions experienced high precipitation variability that led to frequent drought and floods. A recent study (Alahacoon et al. 2021) has shown that the majority of African countries have experienced a significant variation in long-term (1983–2020) precipitation. Therefore, reliable and consistent precipitation estimates for water resources monitoring are vital in Africa (Awange et al. 2016). Nonetheless, studies evaluating precipitation products (with reference to precipitation gauges) in Africa are limited by the lack of in situ observations (Awange et al. 2016; Romilly and Gebremichael 2011; Owusu et al. 2019; Echeta et al. 2022; Logah et al. 2021; Pérez-Alarcón and Fernández-Alvarez 2022).

Accurate estimation of precipitation is challenging (Adhikari et al. 2020; Foufoula-Georgiou et al. 2020; Akbari et al. 2022) due to different sources of uncertainty associated with different estimation methods and high spatiotemporal variability in complex topographical regions (Beck et al. 2019; Adhikari and Behrangi 2021; Akbari et al. 2020). Due to the localized nature of precipitation patterns, especially for extreme events, precipitation gauges provide the most accurate precipitation measurements (Sun et al. 2018). However, they are geographically sparse, especially in remote areas (Huffman and Bolvin 2013; Akbari et al. 2019), and suffer from sample measurement, and under catch errors (Rasmussen et al. 2012; Ehsani and Behrangi 2022). Remote sensing (RS) precipitation products mitigate some of the shortcomings of the gauges by incorporating observations from thermal infrared, passive microwave, and radar instruments (Huffman and Bolvin 2013). However, remote sensing products may have significant biases due to systematic and random errors in their retrieval algorithms (Ehsani et al. 2021), inadequate spatial and temporal sampling (Sun et al. 2018), relatively poor performance over snow and ice surfaces (Ferraro et al. 2013; Rahimi et al. 2022), and relatively short data records (Sadeghi et al. 2019). Reanalysis precipitation products benefit from data assimilation systems that incorporate available observations (in situ and remotely sensed data) into numerical models (Morales-Moraga et al. 2019). Although reanalysis products enable an extended temporal estimation (e.g., over 40 years) of precipitation (Morales-Moraga et al. 2019), their reliability depends on observational constraints, which can vary significantly over space and time (Dee et al. 2016; Rahimi et al. 2021). Ground-based precipitation radars provide near real-time coverage with a high spatiotemporal resolution (Sokol et al. 2021). However, they are only available in a few countries and have some limitations such as the interference of the earth's curvature with the beam at long distances (Sebastianelli et al. 2013), and the high cost of the equipment (Sokol et al. 2021). All precipitation products including RS, radars, and reanalysis depend on gauges records for high-quality retrievals and bias adjustment (Sunt et al. 2018).

Accurate precipitation estimation is crucial for climate studies, trend analysis, water resources management, hydrological forecasting, and so on (Jiang et al. 2012; Liu et al. 2017). However, precipitation observations are qualified by the IPCC as of medium confidence (IPCC 2013). The confidence metric provides a qualitative synthesis of the IPCC expert team's judgment about the validity of a finding based on the level of agreement

and evidence (type, amount, quality, and consistency; IPCC 2013). One of the most common problems in precipitation time-series analyses is the presence of gaps with different lengths (Bellido-jiménez et al. 2021). Gaps are due to erroneous manual data entry, equipment errors during the data collection, data loss due to defective storage technologies, and so on (Tannenbaum 2009).

Gap-free time series are required for statistical and trend analysis (Farhangfar et al. 2008; Shen et al. 2015; Li et al. 2019). Gap-filling methods can be used to fill in the missing data. Three categories of gap-filling methods are investigated in this study: (i) machine learning-based; (ii) precipitation products; and (iii) daily precipitation gap-filling software. Machine learning-based methods are the most versatile approach due to the availability of powerful algorithms and improving access to more data. Also, these methods can be calibrated locally based on available records. However, they need a significant amount of observations (Soley-Bori 2013). On the other hand, many precipitation products are available globally and can be easily accessed from online data sources. However, these products cannot be calibrated locally by end users. Finally, some software are developed exclusively for the gap-filling of daily precipitation based on the geostatistical and geospatial relationship among adjacent gauges. The inputs of such software are precipitation records and the location of stations. Machine learning-based imputation models have outperformed other approaches (Bellido-jiménez et al. 2021), but their ability is often overlooked by the hydrological community (Gao et al. 2018).

The case study of this paper is Tanzania. Climate-related hazards such as droughts and floods are increasing in this country (United Republic of Tanzania 2012). Gap-free precipitation data is crucial for hydrological studies there considering the steady increase in population and limited water resources, especially for food security purposes. A proper understanding of spatiotemporal variations of precipitation is necessary to ensure sustainable water resources management (Mashingia et al. 2014). Despite the importance of gap-free precipitation time series, only a limited number of in situ observations in Africa are readily available to the Global Telecommunication Systems (GTS) global data archives (Nicholson et al. 2003). This will negatively affect the accuracy of global precipitation product (categories of gap-filling ii explained above). On the other hand, other categories for filling the gap of daily precipitation data (machine learning and software) have not been studied or compared in Tanzania.

This study investigates the performances of the three gap-filling approaches mentioned above. Random Forest (RF) and Fully Connected Deep Neural Network (FCDNN) algorithms are selected as machine learning-based methods. This is motivated by the results of previous hydrological studies (Bellido-jiménez et al. 2021; Portuguese-Maurtua et al. 2022; Kim and Ryu 2016). Also, well-known precipitation products, including Global Precipitation Climatology Centre (GPCC) V2020, Global Precipitation Climatology Project (GPCP) V1.3, Precipitation Estimation from Remotely Sensed Information using Artificial Neural Networks - Climate Data Record (PERSIANN-CDR), European Centre for Medium-Range Weather Forecasts Reanalysis V5 (ERA5), and Integrated Multi-satellitE Retrievals for GPM (IMERG) Final V6 (Table 1), are evaluated over the study area. Finally, Reconstruction of Daily Data - Precipitation (reddPrec) software (Serrano-Notivoli et al. 2017) is chosen as the representative of gap-filling software due to its acceptable performance in previous studies (Serrano-Notivoli et al. 2018; 2017; Navarro et al. 2020; Merino et al. 2021). This software enables obtaining serially complete precipitation datasets, estimating

Table 1 Evaluated precipitation products in this study over the IDB

Data set	Resolution	Coverage	Reference
GPCC-daily V2020	1.0°	Global land	Schamm et al. (2014)
GPCP V1.3	1.0°	Global	Huffman and Bolvin (2013)
PERSIANN-CDR	0.25°	60°S–60°N	Ashouri et al. (2015)
IMERG Final V6	0.1°	60°S–60°N	Huffman et al. (2022)
ERA5	0.25°	Global	Hersbach et al. (2020)

new data at ungauged locations, and/or creating regular grids of daily precipitation based on original data containing missing values or even large data gaps. In the upcoming sections, the study area, datasets, and methodology of each gap-filling approach is explained. Then, statistical metrics for evaluating the performance of the methods based on comparison against in situ records are presented. Finally, the best method among the examined ones is determined. This is derived based on the overall daily performance compared to daily precipitation observations. Also, spatial analysis on the accuracy has been carried to show how each method performs in different locations of the studied area. The best approach that fills the gap of daily precipitation data is then retained as the key outcome of the present study.

2 Study Area, Methodology, and Datasets

2.1 Study Area

In Tanzania, the extent of the ground-based precipitation network is not adequate to capture all the spatial rainfall variability (Mashingia et al. 2014). This country consists of nine main basins (Fig. 1). There are two types of rain gauges: the non-recording type gives only total rainfall that occurred during a particular time, and the recording type gives hourly rainfall. Based on the World Meteorology Organization (WMO 2008, 2017) guidelines, the minimum density for non-recording rain gauges is between 250 and 900 km² per station (varying according to physiographic properties from mountains to coastal areas). Fifty-eight non-recording rain gauges used in this study are mainly located in the Internal Drainage Basin (IDB; Fig. 1). IDB is the second biggest basin covering almost 20% of Tanzania ($\approx 154,000$ km²), so the coverage of each station in this basin is above 2,600 km² per station. The annual evapotranspiration rate over this region is 2,000 mm. The climate of the studied area is mainly Tropical Savanna, and the seasons are divided into dry (June to October) and wet (November to May). The average annual precipitation in IDB ranges from 600 to 900 mm, but the northeastern part (near the border of Kenya) comes to more than 1,000 mm. Almost all the rivers in this region are seasonal and flow from December to July, but they are often dry for the rest of the year. In the central and northeastern parts of IDB, there are volcanoes such as Mt. Hanang, Kilimanjaro (the highest mountain in Africa), and Ngorongoro crater. In the northern to the central part of IDB, several large lakes are located, such as Lake Natron, Lake Manyara, and Lake Eyasi (JICA 2008).

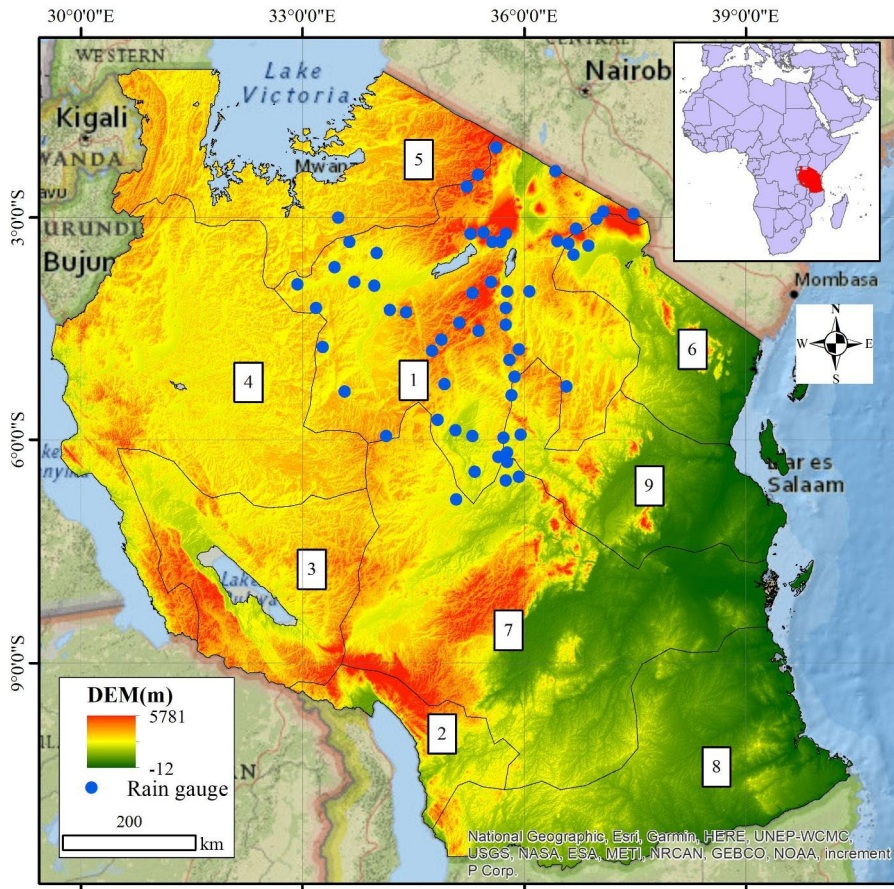


Fig. 1 Main nine sub-basins of Tanzania: 1-Internal drainage basin (IDB), 2-Lake Nyasa, 3-Lake Rukwa, 4-Lake Tanganyika, 5-Lake Victoria, 6-Pangani, 7-Rufiji, 8-Ruvuma South Coast, and 9-Wami Ruvu, with the location of rain gauges

2.2 Datasets

A total of fifty-eight daily precipitation gauge (Fig. 1) records were analyzed. Data were provided by the Ministry of Water in Tanzania. The quality control of data, as shown in Supplementary Material (SM) (Figure SM1b), was performed based on the framework suggested in Wijngaard et al. (2003) and Ghajarnia et al. (2022). Statistical tests were utilized to exclude gauges with low quality data. Two gauges were excluded (more details on tests in the SM). In addition to in situ observations, five precipitation products were used in this study (Table 1). The current study was conducted for the 2000–2010 period because it has the highest overlap with precipitation products (limited by the availability of satellite products and gauge observations; Figure SM1a).

Table 2 Meteorological parameters utilized as inputs of the gap-filling methods by machine learning

Meteorological parameters	Unit
Total precipitable liquid water	kg/m ²
2-meter eastward wind	m/s
Surface pressure	Pa
2-meter northward wind	m/s
2-meter specific humidity	kg/kg
2-meter wet bulb temperature	K

3 Methodology

Three approaches of gap-filling were examined in this study. Utilizing: (i) the FCDNN and RF methods as two machine learning techniques; (ii) well-known precipitation products available globally; (iii) the reddPrec software developed for gap-filling of daily precipitation. Precipitation products are shown in Table 1. In the following sections, FCDNN, RF, and reddPrec are explained briefly.

3.1 Gap-filling by Machine Learning Algorithms

Table 2 summarizes the daily climate variables used as the inputs/features of the machine learning models. Models are trained with 70% of the available daily precipitation records (training set), while the hyperparameters are tuned over 15% of the records (validation set), and the remaining data is used for independent evaluation of the machine learning models and other gap-filling approaches (test set). These meteorological inputs are taken from the Modern-Era Retrospective analysis for Research and Applications version 2 (MERRA-2) which has about 50 km spatial resolution.

3.1.1 Fully Connected Deep Neural Networks (FCDNN)

FCDNN is a multilayer feed-forward neural network that is the simplest (Moghaddam et al. 2022) and one of the most common neural network forms (Partal and Kişi 2007). Each layer consists of several processing units (neurons). Each neuron is connected to adjacent layers with an individual weight assigned to each interlayer link. All inputs into a single neuron are multiplied by their associated weights and summed up to form a single output. Finally, each of these outputs is subject to a nonlinear transformation referred to as the activation function. As a result, FCDNN can be represented as a nested set of functions. It is the superposition of many simple nonlinear functions that enable FCDNN to estimate non-linear functions. FCDNN is fully connected, with each node connected to every node in the next and previous layer (Gardner and Dorling 1998). The number of layers, the number of nodes in each layer, the loss function, and the learning rate are among the hyperparameters that should be tuned for the FCDNN model in this study.

3.1.2 Random Forest (RF)

Random Forest (RF) was first introduced by Breiman (2001) as a supervised learning algorithm. Random forests are a combination of predictors (i.e., trees) such that each of them depends on the values of a random vector sampled independently and with the same distri-

bution for all trees in the forest. Internal estimates measure variable importance, and also monitor error, strength, and correlation utilized to show the response to increasing the number of inputs used in the classification. The number of trees and minimum sample split are the hyperparameters tuned for the RF model.

3.2 Gap-filling by the Reconstruction of Daily Data – Precipitation (reddPrec) Software

We selected the reddPrec software (www.cran.r-project.org/web/packages/reddPrec) because: (i) it applies comprehensive quality control over original daily precipitation datasets, and flags suspicious data based on five predefined criteria; and (ii) it fills missing values in original data series by estimating precipitation values using a number of nearest observations for each day. The reddPrec creates daily reference values using all the data recorded at the nearest stations for each targeted day. Multivariate logistic regression is used to compute these reference values based on the nearest neighbors and geographic and topographic variables as covariates. A threshold parameter is integrated to set a maximum distance (in km) to the search for the nearest neighbors (Serrano-Notivoli et al. 2017).

3.3 Evaluating the Performance of Gap-filling Approaches

30% of available daily precipitation data (15,075 observations) were randomly nulled and split as the validation and test sets. The remaining data were used for training machine learning (FCDNN, RF) methods and running the reddPrec software. The validation set was used for tuning the hyperparameters of the machine learning models. Finally, gap-filling approaches were compared over the test set. For consistency, precipitation products were compared with each other with the same validation data set. The best-performing precipitation product was selected based on several evaluation metrics. Finally, all approaches were compared with each other to select the best gap-filling approach in the study area.

3.3.1 Evaluation Metrics

The Pearson correlation coefficient (CC), the relative bias (Rbias), the root mean square error (RMSE), the probability of detection (POD), the false alarm ratio (FAR), and the Heidke skill score (HSS) metrics were used to evaluate the performance of each method. More details on each metric are presented in the Supplementary Materials file.

4 Results and Discussion

Among examined precipitation products, PERSIANN-CDR outperformed the rest in some evaluation metrics (compare Fig. 2d and Figure SM2). PERSIANN-CDR has the least RMSE (7.3 mm), the highest correlation coefficient (0.46), and the highest HSS among all products. Though other precipitation products have fewer Rbias than PERSIANN-CDR (e.g., Rbias of IMERG is -32% compared to that of PERSIANN-CDR which is -50%), PERSIANN-CDR was selected as the best-performing precipitation product because of better scores in other metrics.

Comparison of daily precipitation estimates by other evaluated methods (FCDNN, RF, and reddPrec) against gauge-based observations (Fig. 2) revealed that RF has the lowest RMSE (6.9 mm), the highest correlation coefficient (0.53), and least Rbias (-2%). Therefore, RF as a machine learning-based imputation method can improve evaluation metrics considerably compared to PERSIANN-CDR. This method is less biased than all examined precipitation products (Fig. 2c and Figure SM2). reddPrec has the least skill based on its high RMSE (14.2 mm) and low correlation coefficient (0.19).

Based on the histogram of observations in the validation set (Table SM1), 85% of daily precipitation values are less than 2.3 mm. Also, 90%, 95%, and 99% of precipitation records are less than 9.2, 16.1, and 39.1 mm, respectively. The frequency of precipitation above 51 mm is less than 0.5%. Most data in the scatter plots (Fig. 2) are inclined toward the x-axis (observation) for FCDNN, RF, and PERSIANN-CDR which explains the negative values for Rbias for these methods. Based on the Rbias equation (Eq. S3 in the SM), negative Rbias means that the model underestimates compared to observations. On the other hand, the reddPrec method has a positive Rbias and the majority of data are inclined toward the y-axis (model estimate); thus, this method is overestimating daily precipitation.

The standard deviation of predicted daily precipitation records for FCDNN, reddPrec, RF, and PERSIANN-CDR is 8.1, 2.3, 13.6, 4.1, and 3.8 mm/day. FCDNN, RF and PERSIANN have lower standard deviations than the observations. Estimates by these methods are mainly less than 20 mm/day while observations show that the frequency of precipitation between 20 and 50 mm/day is considerable (x-axis in Fig. 2). The standard deviation of reddPrec is higher than the observations, and the points are more scattered toward the y-axis (Fig. 2b).

The reddPrec software has a parameter (“thres”) to search nearest stations within a specific distance. If this parameter is set to “NA”, the software will search 10 nearest observations without a distance limit (more detail in <https://cran.r-project.org/web/packages/reddPrec/index.html>). In the IDB, 92% of the stations are located 0-400 km from each other (Table SM2). Therefore, we investigated the effect of different “thres” values (i.e., “NA”, 50, 100, 200, 300, and 400 km) on the performance of the reddPrec. Based on Figure SM3, it was found that in the sparsely gauged IDB, introducing lower values for “thres” will improve reddPrec performance slightly. However, this minor improvement in the performance will compromise the number of gaps filled by reddPrec. When “thres” is defined as “NA”, the majority of the validation set is filled by reddPrec (95%). However, in low values of “thres”, less than 60% of gaps are filled. Thus, we decided to set “thres” to “NA” (results in Figs. 2, 3 and 4 are based on this assumption).

Calculation of the categorical metrics (POD, HSS, and FAR) using different precipitation thresholds (0, 0.1, 0.5, 1, 2, and 5 mm/day) for distinguishing between rainy and non-rainy days revealed critical information on the capacity of each method in detecting precipitation events (Fig. 3). Among the imputation methods, RF and FCDNN have the best POD (true hits divided by the sum of true hits and misses as shown in Eq. S4 in SM). However, in low thresholds (0-0.5 mm/day), RF has higher FAR (false alarms divided by the sum of true hits and false alarms as shown in Eq. S5 in SM) compared to other methods. As the threshold increases, the performance of the RF model improves indicating that the RF model has problems capturing low precipitation rates.

The HSS metric, aggregating the effect of FAR and POD was investigated to consider the effect of both directions in the contingency table (true hits and false alarms as well as true

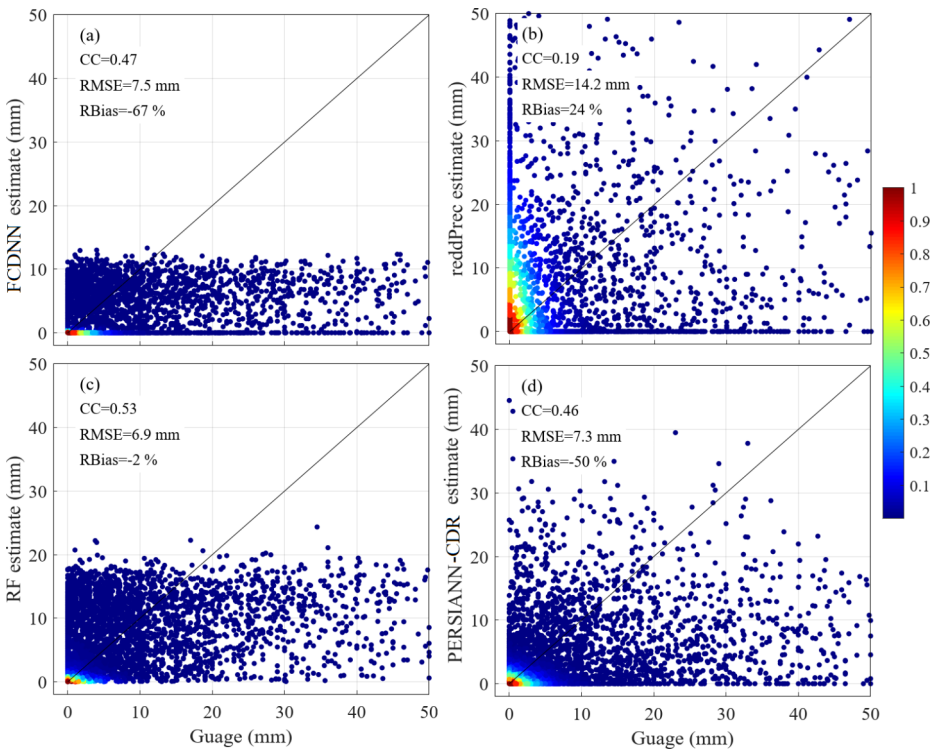


Fig. 2 Scatter plots of the observed daily precipitation (mm/day) and estimated values by (a) FCDNN, (b) reddPrec, (c) RF, and (d) PERSIANN-CDR (the color bar shows the density of points in the plot)

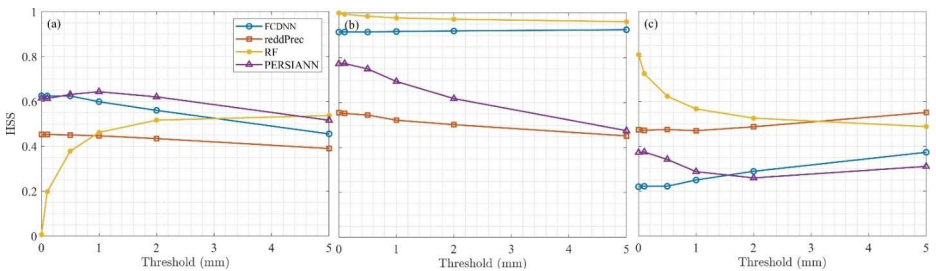


Fig. 3 Performance of different methods at different thresholds in terms of: (a) HSS, (b) POD, and (c) FAR

hits and misses). Therefore, HSS is a good score to quantify the trade-off between FAR and POD. In the low thresholds (0-0.5 mm/day), the trade-off between the high value of FAR and POD for RF has led to the lowest values of HSS for this method. Although this method is powerful in correctly detecting rainy events (high POD), RF (high FAR) reports many days wrongly as rainy. Therefore, the overall strength of RF for distinguishing between rainy and non-rainy days is very low compared to other methods in low thresholds. In-

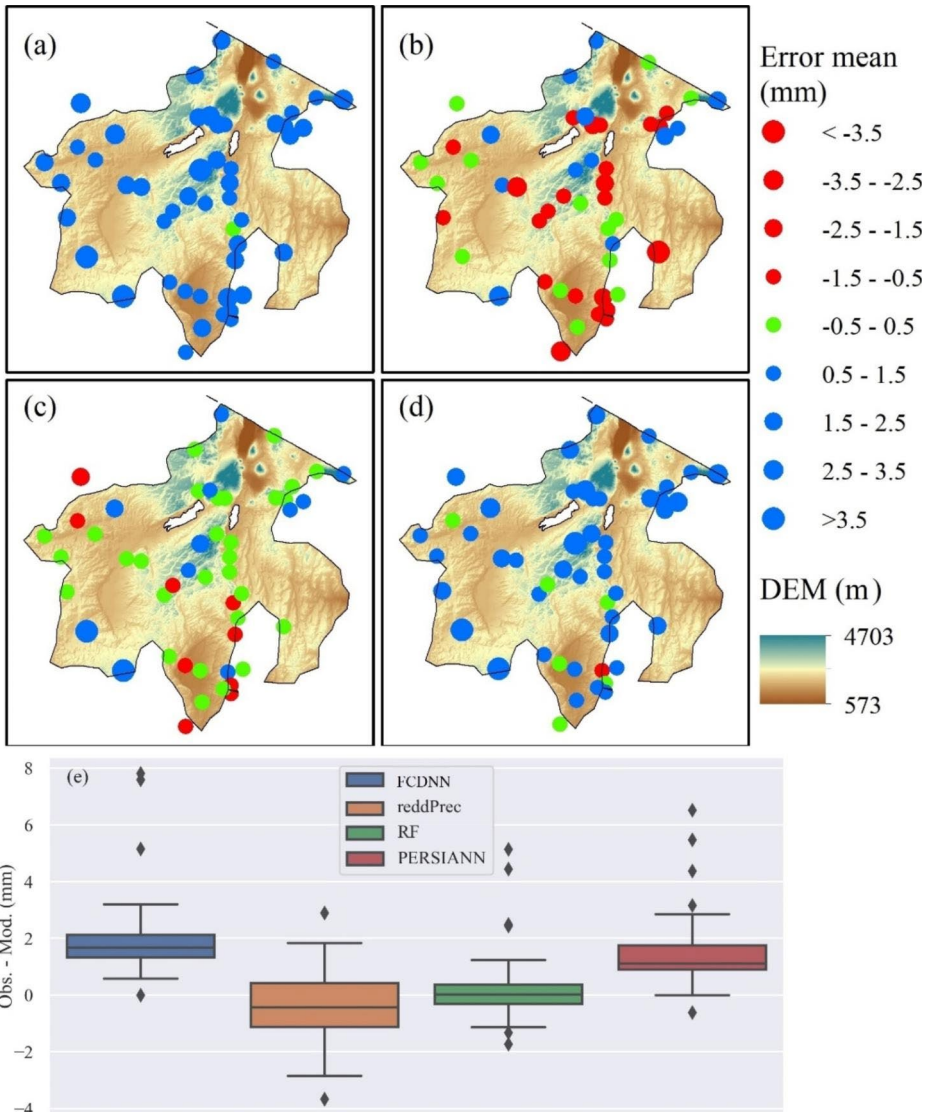


Fig. 4 Spatial distribution of mean error for the stations in the study area for different models: (a) FCDNN, (b) reddPrec, (c) RF, (d) PERSIANN-CDR, and (e) boxplots for different models based on the mean of error for each station

ing the threshold increases RF performance considerably, so this method achieves the highest HSS for a 5 mm/day threshold among all methods.

The reddPrec method has consistently low POD and high FAR. Therefore, this method has the lowest HSS among others. On the other hand, FCDNN and PERSIANN-CDR have consistently high POD and low FAR (specifically in thresholds less than 2 mm/day). Thus, these methods have the highest HSS. However, in higher thresholds, the performance of FCDNN decreases since its FAR increases considerably compared to little improvement in

POD. Based on the HSS value, the PERSIANN-CDR performs best when the threshold is set to 1 mm/day (the highest HSS is 0.64).

The spatial distribution of error for stations is displayed in Fig. 4. The majority of stations (62%) filled by the RF method (Fig. 4c) have an absolute error of less than 0.5 mm/day. After RF, reddPrec (Fig. 4b), has the highest number of stations having low error (-0.5 to 0.5 mm/day shown by green points). However, for reddPrec, overestimation in other stations (big red points in Fig. 4b) has affected overall accuracy (Fig. 2b). FCDNN and PERSIANN-CDR mainly underestimate daily precipitation although the performance of PERSIANN-CDR is better (Fig. 4a and d). It is noteworthy that the mean error for all stations is very close to zero for RF as the most accurate model.

All examined methods in this research have limitations. Although reddPrec has the least accuracy among all other methods, the spatial distribution of error in the stations reveals that this method is more successful in filling the gaps of data in many locations compared to FCDNN and PERSIANN-CDR. The “thres” parameter is very important in reddPrec as it determines which adjacent stations can be used for gap-filling. IDB is a sparsely gauged basin, so the “thres” had to be set to “NA” to fill the highest ratio of gaps. This will introduce more uncertainty/error into the reddPrec method because distant stations are allowed to be utilized in the gap-filling process. Even though the “thres” was set to “NA”, 5% of gaps in the test set were not filled by reddPrec. High error values in some stations may be attributed to what is described above.

The study area is a data-deprived region which has many gaps in the observations. In other words, we do not have a continuous time-series in gauges; hence, options for selection of machine learning algorithms are limited. Many algorithms such as convolutional neural networks could not be used because of lack of data, so we have included only a fully connected neural network in our models. Both machine learning-based methods use gridded meteorological parameters (Table 2 from MERRA-2). All precipitation products are similarly gridded data. Therefore, values of the nearest cell to each rain gauge were used to fill the gaps. Attributing the values of the whole cell to a point will introduce uncertainties/errors in estimations, especially in a low-resolution dataset as the cell values represent the entire cell, not the gauge location. Finally, it should be noted that most of the daily precipitation records are less than 1 mm/day although the range of daily precipitation reaches 100 mm/day. The high concentration of low values (less than 1 mm) in records used for training RF and FCDNN led to low variance in their predictions. Thus, FCDNN and RF could not estimate precipitation higher than 10 and 20 mm/day, respectively. This can also be attributed to the limited number of samples representing higher precipitation rates in the training set, inaccuracy of the high precipitation rates in the reference set, and/or lack of related info in the feature to represent extreme precipitation rates.

5 Conclusions

This study investigated the performance of three precipitation gap-filling approaches over a sparsely gauged region in Tanzania: (i) precipitation products including GPCP V2020, GPCP V1.3, PERSIANN-CDR, ERA5, and IMERG Final V6; (ii) machine learning-based imputation approaches such as Fully Connected Deep Neural Network and Random Forest, and (iii) daily precipitation gap-filling software namely reddPrec. Based on available

data in the study area, 2000–2010 was selected as the study period because of the highest overlap between in situ records and satellite/reanalysis data. To evaluate the performance of each approach, we utilized 30% of the available rain gauge records (15,075 observations) for validation and testing and the rest of the data was used for training of machine learning-based methods (category ii) and reddPrec software (category iii). Evaluation of precipitation products (category i) against the test set revealed that PERSIANN-CDR has the best performance compared to other examined precipitation products (the lowest RMSE, and the highest correlation coefficient). Also, PERSIANN-CDR has the best performance in detecting rainy days based on HSS among all examined gap-filling methods. However, other precipitation products are less biased than PERSIANN-CDR (based on Rbias). This study showed that machine learning-based gap-filling methods trained by meteorological data (from MERRA-2) have overall better performance compared to other methods. Random Forest has a lower bias than PERSIANN-CDR, and is the best-performing product in the study area. RF also has the lowest RMSE, highest correlation coefficient, and lowest bias among all examined methods/precipitation products. The main difference between the trained Random Forest model and global precipitation products (e.g., PERSIANN-CDR) is in the utilization of a higher number of in situ records in the training process. The accuracy of global precipitation products suffers from the lack of in situ data in the calibration process, especially in developing countries. In these countries, the contribution of in situ records in international data centers (e.g., Global Telecommunication System) is very low. Global Telecommunication System, as an example, is a source of calibration for many precipitation products. Consequently, the accuracy of precipitation products (e.g., GPCP used for bias adjustment in PERSIANN-CDR, GPCP, and IMERG) is negatively affected by the lack of in situ data.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s40710-023-00625-y>.

Author Contributions Conceptualization: MRE, MA, RR; Writing-original draft: MF, MA, MRE; Methodology: MRE, MA, RR; Data curation: MRE, MA, RR, MF; Formal analysis: MRE, MA; Visualization: MA, MRE, RR, MF; Writing-review & editing: MF, MRE, MA, RR, MM, AB, BK, ATH, MO; Supervision: ATH, MO.

Funding This work is supported by the University of Oulu, Finland.

Open Access funding provided by University of Oulu including Oulu University Hospital.

Open Access funding provided by University of Oulu including Oulu University Hospital.

Data Availability Precipitation products data were downloaded through Google Earth Engine Java Script API (Gorelick et al. 2017; <https://earthengine.google.com/>, Assessed on <https://developers.google.com/s/results/earth-engine/datasets?q=precipitaiton&text=precipitaiton>). Daily in-situ precipitation data are accessible upon request from Tanzania Ministry of Energy (<https://www.nishati.go.tz/>, accessed by personal contact) and MERRA-2 data were obtained from NASA website (<https://gmao.gsfc.nasa.gov/reanalysis/MERRA-2/>, Assessed on <https://developers.google.com/s/results/earth-engine/datasets?q=MERRA2&text=MERRA2>). All figures were made with Matplotlib (Caswell et al. 2020; Hunter 2007), seaborn (Waskom et al. 2017), and MATLAB (2019).

Declarations

Ethics Approval Not applicable.

Consent to Participate The authors declare that they are aware and consent to their participation in this paper.

Consent for Publish The authors declare that they consent to the publication of this paper.

Conflicts of Interest The authors have no relevant financial or non-financial interests to disclose.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Adhikari A, Ehsani MR, Song Y, Behrangi A (2020) Comparative assessment of snowfall retrieval from microwave humidity Sounders using machine learning methods. *Earth and Space Science* 7(11), e2020EA001357
- Adhikari A, Behrangi A (2021) Assessment of satellite precipitation products in relation with orographic enhancement over the Western United States. *Earth and Space Science Open Archive ESSOAr*
- Akbari M, Haghighi AT, Aghayi MM, Javadian M, Tajrishy M, Kløve B (2019) Assimilation of satellite-based data for hydrological mapping of precipitation and direct runoff coefficient for the Lake Urmia Basin in Iran. *Water* 11, no. 8 (2019): 1624
- Akbari M, Baubekova A, Roozbahani A, Gafurov A, Shiklomanov A, Rasouli K, Ivkina N, Kløve B, Haghighi AT (2020) Vulnerability of the Caspian Sea shoreline to changes in hydrology and climate. *Environ Res Lett* 15:115002
- Akbari M, Mirchi A, Roozbahani A, Gafurov A, Kløve B, Haghighi AT (2022) Desiccation of the Trans-boundary Hamun Lakes between Iran and Afghanistan in Response to Hydro-climatic Droughts and Anthropogenic Activities. *Journal of Great Lakes Research* 48, no. 4 (2022): 876–889
- Alahacoon N, Edirisinghe M, Simwanda M, Perera E, Nyirenda V, Ranagalage M (2021) Rainfall Variability and Trends over the African Continent Using TAMSAT Data (1983–2020): Towards Climate Change Resilience and Adaptation. *Remote Sensing* 14, no. 1 (2022): 96
- Ashouri H, Hsu K, Sorooshian S, Braithwaite D, Knapp K, Cecil D, Nelson B, Prat P (2015) *Bull Am Meteorol Soc* 96(1):69–83. <https://doi.org/10.1175/BAMS-D-13-00068.1>. PERSIANN-CDR: Daily Precipitation Climate Data Record from Multisatellite Observations for Hydrological and Climate Studies
- Awange JL, Ferreira VG, Forootan E, Khandu A, Agutu NO, He XF (2016) Uncertainties in remotely sensed precipitation data over Africa. *Int J Climatol* 36(1):303–323. <https://doi.org/10.1002/joc.4346>
- Beck H, Pan M, Roy T, Weedon G, Pappenberger F, Van Dijk A, Huffman G, Adler R, Wood E (2019) Daily evaluation of 26 precipitation datasets using Stage-IV gauge-radar data for the CONUS. *Hydrol Earth Syst Sci* 23(1):207–224. <https://doi.org/10.5194/hess-23-207-2019>
- Bellido-jiménez JA, Gualda JE, García-marín AP (2021) Assessing machine learning models for gap filling daily rainfall series in a semiarid region of Spain. *Atmosphere* 12(9). <https://doi.org/10.3390/atmos12091158>
- Breiman L (2001) Random forests. *Mach Learn* 45(1):5–32
- Caswell T, Droettboom M, Lee A, Hunter J, Firing E, Stansby D (2020) *Matplotlib v3.2.1* [Software]. Zenodo. <https://doi.org/10.5281/zenodo.3714460>
- Dee D, Fasullo J, Shea D, Walsh J, NCAR S (2016) The climate data guide: atmospheric reanalysis: overview and comparison tables. National Center for Atmospheric Research, Boulder, CO). Available at <https://climatedataguide.ucar.edu/climatedata/atmospheric-reanalysis-overview-comparison-tables>. Accessed on June 1, 2017
- Dinku T, Ceccato P, Grover-Kopec E, Lemma M, Connor S, Ropelewski C (2007) Validation of satellite rainfall products over East Africa's complex topography. *Int J Remote Sens* 28(7):1503–1526
- Echeta O, Adjei K, Andam-Akorful S, Gyamfi C, Darko D, Odai S, Kwarteng E (2022) Performance evaluation of Near-Real-Time Satellite Rainfall estimates over three distinct climatic zones in Tropical West-Africa. *Environ Processes* 9(4):59

- Ehsani M, Behrangi A (2022) A comparison of correction factors for the systematic gauge-measurement errors to improve the global land precipitation estimate. *J Hydrol* 610:127884. <https://doi.org/10.1016/j.jhydrol.2022.127884>
- Ehsani M, Behrangi A, Adhikari A, Song Y, Huffman G, Adler R, Bolvin D, Nelkin E (2021) Assessment of the Advanced Very High-Resolution Radiometer (AVHRR) for snowfall retrieval in high latitudes using CloudSat and machine learning. *Journal of Hydrometeorology* 22, no. 6 (2021): 1591–1608
- Eltahir E (1998) A soil moisture–rainfall feedback mechanism: 1. Theory and observations. *Water Resour Res* 34(4):765–776
- Farhangfar A, Kurgan L, Dy J (2008) Impact of imputation of missing values on classification error for discrete data. *Pattern Recogn* 41(12):3692–3705
- Ferraro R, Peters-Lidard C, Hernandez C, Joseph TF, Aires F, Prigent C, Lin X, Boukabara S, Furuzawa F, Gopalan K, Harrison KW, Karbouf F, Li L, Liu C, Masunaga H, Moy L, Ringerud S, Skofronick-Jackson G, Tian Y, Wang N (2013) An evaluation of microwave land surface emissivities over the continental United States to benefit GPM-Era precipitation algorithms. *IEEE Trans Geosci Remote Sens* 51(1):378–398. <https://doi.org/10.1109/TGRS.2012.2199121>
- Foufoula-Georgiou E, Guilloteau C, Nguyen P, Aghakouchak A, Hsu K, Busalacchi A, Turk F, Peters-Lidard C, Oki T, Duan Q, Krajewski W (2020) Advancing precipitation estimation, prediction, and impact studies. *Bull Am Meteorol Soc* 101(9):E1584
- Gao Y, Merz C, Lischeid G, Schneider M (2018) A review on missing hydrological data processing. *Environ Earth Sci* 77(2):1–12. <https://doi.org/10.1007/s12665-018-7228-6>
- Gardner MW, Dorling S (1998) Artificial neural networks (the multilayer perceptron) - a review of applications in the atmospheric sciences. *Atmos Environ* 32(14–15):2627–2636. [https://doi.org/10.1016/S1352-2310\(97\)00447-0](https://doi.org/10.1016/S1352-2310(97)00447-0)
- Ghajarnia N, Akbari M, Saemian P, Ehsani M, Hosseini-Moghari S, Azizian A, Kalantari Z, Behrangi A, Tourian M, Klöve B, Haghighi A (2022) Evaluating the Evolution of ECMWF Precipitation Products Using Observational Data for Iran: From ERA40 to ERA5. *Earth and Space Science*, 9(10), p.e2022EA002352
- Gorelick N, Hancher M, Dixon M, Ilyushchenko S, Thau D, Moore R (2017) Google Earth Engine: planetary-scale geospatial analysis for everyone, vol 202. *Remote Sensing of Environment*, pp 18–27
- Hersbach H, Bell B, Berrisford P, Hirahara S, Horányi A, Muñoz-Sabater J, Nicolas J, Peubey C, Radu R, Schepers D, Simmons A, Soci C, Abdalla S, Abellan X, Balsamo G, Bechtold P, Biavati G, Bidlot J, Bonavita M (2020) The ERA5 global reanalysis. *Q J R Meteorol Soc* 146(730):1999–2049. <https://doi.org/10.1002/QJ.3803>
- Huffman G, Bolvin D (2013) Version 1.2 GPCP. One-Degree Daily Precipitation Data Set Documentation
- Huffman G, Stocker E, Bolvin D, Nelkin E, Jackson T (2022) GPM IMERG final precipitation L3 half hourly 0.1 degree x 0.1 degree V06, Greenbelt, MD, Goddard Earth Sciences Data and Information Services Center (GES DISC)
- Hunter J (2007) Matplotlib: a 2d graphics environment. *Comput Sci Eng* 9(3):90–95. <https://doi.org/10.1109/MCSE.2007.55>
- IPCC (2013) The physical science basis; summary for policymakers. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change
- Jiang S, Ren L, Hong Y, Yong B, Yang X, Yuan F, Ma M (2012) Comprehensive evaluation of multi-satellite precipitation products with a dense rain gauge network and optimally merging their simulated hydrological flows using the bayesian model averaging method. *J Hydrol* 452:213–225
- JICA (2008) The study of the groundwater development and management in the internal drainage basin of Tanzania. Ministry of Water, Tanzania
- Kim J, Ryu J (2016) A heuristic gap filling method for daily precipitation series. *Water Resour Manage* 30(7):2275–2294
- Li X, Wang L, Cheng Q, Wu P, Gan W, Fang L (2019) Cloud removal in remote sensing images using non-negative matrix factorization and error correction. *ISPRS J Photogrammetry Remote Sens* 148:103–113
- Liu X, Yang T, Hsu K, Liu C, Sorooshian S (2017) Evaluating the streamflow simulation capability of PERSIANN-CDR daily rainfall products in two river basins on the Tibetan Plateau. *Hydrol Earth Syst Sci* 21(1):169–181
- Logah F, Adjei K, Obuobie E, Gyamfi C, Odai N (2021) Evaluation and comparison of Satellite Rainfall Products in the Black Volta Basin. *Environ Processes* 8:119–137
- Mashingia F, Mtalo F, Bruen M (2014) Validation of remotely sensed rainfall over major climatic regions in Northeast Tanzania. *Phys Chem Earth* 67–69. <https://doi.org/10.1016/j.pce.2013.09.013>
- MATLAB (2019) Statistics Toolbox Release 2019b, The MathWorks, Inc., Natick, Massachusetts, United States. <https://www.mathworks.com/>
- Merino A, García-Ortega E, Navarro A, Fernández-González S, Tapiador F, Sánchez J (2021) Evaluation of gridded rain-gauge-based precipitation datasets: impact of station density, spatial resolution, altitude gradient and climate. *Int J Climatol* 41(5):3027–3043

- Moghaddam M, Ferre T, Chen X, Chen K, Ehsani M (2022) Application of machine learning methods in inferring surface water groundwater exchanges using high temporal resolution temperature measurements. arXiv preprint arXiv. <https://doi.org/10.48550/arXiv.2201.00726>. :2201.00726
- Morales-Moraga D, Meza F, Miranda M, Gironás J (2019) Spatio-temporal estimation of climatic variables for gap filling and record extension using reanalysis data. *Theoret Appl Climatol* 137(1):1089–1104
- Navarro A, García-Ortega E, Merino A, Sánchez J (2020) Extreme events of precipitation over complex terrain derived from satellite data for climate applications: an evaluation of the southern slopes of the pyrenees. *Remote Sens* 12(13):2171
- Nicholson S, Some B, McCollum J, Nelkin E, Klotter D, Berte Y, Diallo B, Gaye I, Kpabebe G, Ndiaye O, Noukpozoukou J (2003) Validation of TRMM and other rainfall estimates with a high-density gauge dataset for West Africa. Part II: validation of TRMM rainfall products. *J Appl Meteorol* 42(10):1355–1368
- Oki T, Kanai S (2006) Global hydrological cycles and world water resources. *Science* 313(5790):1068–1072
- Owusu C, Adjei K, Odai S (2019) Evaluation of satellite rainfall estimates in the Pra Basin of Ghana. *Environ Processes* 6:175–190
- Partal T, Kişi Ö (2007) Wavelet and neuro-fuzzy conjunction model for precipitation forecasting. *J Hydrol* 342(1–2):199–212. <https://doi.org/10.1016/J.JHYDROL.2007.05.026>
- Pérez-Alarcón A, Fernández-Alvarez D (2022) Improving monthly rainfall forecast in a watershed by combining neural networks and autoregressive models. *Environ Process* 9
- Portuguez-Maurtua M, Arumi J, Lagos O, Stehr A, Montalvo Arquiniño N (2022) Filling gaps in daily precipitation series using regression and machine learning in Inter-Andean Watersheds. *Water* 14(11):1799
- Rahimi R, Tavakol-Davani H, Nasserli M (2021) An uncertainty-based regional comparative analysis on the performance of different bias correction methods in statistical downscaling of precipitation. *Water Resour Manage* 35(8):2503–2518
- Rahimi R, Ebtehaj A, Panegrossi G, Milani L, Ringerud S, Turk F (2022) Vulnerability of Passive Microwave Snowfall Retrievals to Physical Properties of Snowpack: a perspective from dense media radiative transfer theory. *IEEE Trans Geosci Remote Sens* 60:1–13
- Rasmussen R, Baker B, Kochendorfer J, Meyers T, Landolt S, Fischer A, Black J, Thériault J, Kucera P, Gochis D, Smith C (2012) How well are we measuring snow: the NOAA/FAA/NCAR winter precipitation test bed. *Bull Am Meteorol Soc* 93(6):811–829
- Romilly G, Gebremichael M (2011) Evaluation of satellite rainfall estimates over ethiopian river basins. *Hydrol Earth Syst Sci* 15(5):1505–1514
- Sadeghi M, Asanjan A, Faridzad M, Nguyen P, Hsu K, Sorooshian S, Braithwaite D (2019) PERSIANN-CNN: precipitation estimation from remotely sensed information using artificial neural networks–convolutional neural networks. *J Hydrometeorol* 20(12):2273–2289
- Saemian P, Hosseini-Moghari S, Fatehi I, Shoarinezhad V, Modiri E, Tourian M, Tang Q, Nowak W, Bárdossy A, Sneeuw N (2021) Comprehensive evaluation of precipitation datasets over Iran. *Journal of Hydrology*, 603, p.127054
- Schamm K, Ziese M, Becker A, Finger P, Meyer-Christoffer A, Schneider U, Schröder M, Stender P (2014) Global gridded precipitation over land: a description of the new GPCC First guess daily product. *Earth Syst Sci Data* 6(1):49–60. <https://doi.org/10.5194/ESSD-6-49-2014>
- Sebastianelli S, Russo F, Napolitano F, Baldini L (2013) On precipitation measurements collected by a weather radar and a rain gauge network. *Nat Hazards Earth Syst Sci* 13(3):605–623
- Serrano-Notivol R, de Luis M, Beguería S (2017) An R package for daily precipitation climate series reconstruction. *Environ Model Softw* 89:190–195. <https://doi.org/10.1016/j.envsoft.2016.11.005>
- Serrano-Notivol R, Martín-Vide J, Saz M, Alberto Longares L, Beguería S, Sarricolea P, Meseguer-Ruiz O, De Luis M (2018) Spatio-temporal variability of daily precipitation concentration in Spain based on a high-resolution gridded data set. *International Journal of Climatology* 38 (2018): e518–e530
- Shen H, Li X, Cheng Q, Zeng C, Yang G, Li H, Zhang L (2015) Missing information reconstruction of remote sensing data: a technical review. *IEEE Geoscience and Remote Sensing Magazine* 3(3):61–85
- Sokol Z, Szturc J, Orellana-Alvarez J, Popová J, Jurczyk A, Céleri R (2021) The role of weather radar in rainfall estimation and its application in meteorological and hydrological modelling—A. *Rev Remote Sens* 13(3):351
- Soley-Bori M (2013) Dealing with missing data: Key assumptions and methods for applied analysis. *Boston University* 4, no. 1 (2013): 1–19
- Sun Q, Miao C, Duan Q, Ashouri H, Sorooshian S, Hsu K (2018) A review of global precipitation data sets: data sources, estimation, and intercomparisons. *Rev Geophys* 56(1):79–107. <https://doi.org/10.1002/2017RG000574>
- Tannenbaum C (2009) The empirical nature and statistical treatment of missing data. University of Pennsylvania, Dissertations available from ProQuest. AAI3381876. <https://repository.upenn.edu/dissertations/AAI3381876>

- United Republic of Tanzania (2012) National Climate Change Strategy. Available at: <https://www.climate-laws.org/geographies/tanzania/policies/national-climate-change-strategy-2021-2026>, accessed at 6.2.2023
- Waskom M, Botvinnik O, O’Kane D, Hobson P, Lukauskas S, Gemperline D, Augspurger T, Halchenko Y, Cole J, Warmenhoven J, de Ruiter J (2017) Mwaskom/Seaborn: V0. 8.1 (September 2017). Zenodo. Available at: <https://zenodo.org/record/883859#.Y-ExU3ZBw2w>, accessed at 6.2.2023
- Wijngaard JB, Klein AM, Können G (2003) Homogeneity of 20th century european daily temperature and precipitation series. *Int J Climatology: J Royal Meteorological Soc* 23(6):679–692
- WMO (2008) Guide to Hydrological Practices. WMO-No. 168, ISBN 978-92-63-10168-6. Available at: <https://www.hydrology.nl/mainnews/1-latest-news/189-guide-to-hydrological-practices-new-edition-by-wmo.html>, accessed at 6.2.2023
- WMO (2017) Guidelines on the Calculation of Climate Normals. WMO-No. 1203, 1203, 18. Available at: https://library.wmo.int/index.php?lvl=notice_display&id=20130#.Y-Eyg3ZBw2w, accessed at 6.2.2023

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Authors and Affiliations

**Marzie Faramarzadeh¹ · Mohammad Reza Ehsani² · Mahdi Akbari³ ·
Reyhane Rahimi⁴ · Mohammad Moghaddam² · Ali Behrangi² · Björn Klöve³ ·
Ali Torabi Haghighi³ · Mourad Oussalah¹**

✉ Mahdi Akbari
mahdi.akbari@oulu.fi

¹ Faculty of Information Technology and Electrical Engineering, University of Oulu, Pentti Kaiteran katu 1, 90570 Oulu, Finland

² Department of Hydrology and Atmospheric Sciences, University of Arizona, 85721 Tucson, Arizona, AZ, USA

³ Water, Energy and Environmental Engineering Research Unit, Faculty of Technology, University of Oulu, Pentti Kaiteran katu 1, 90570 Oulu, Finland

⁴ Department of Civil, Environmental and Geo-Engineering, University of Minnesota, 55455 Minneapolis, MN, USA