Suitability of different precipitation data sources for hydrological analysis: a study from Western Ghats, India

Beeram Satya Narayana Reddy ⓘ ·
Shahanas P. V. · S. K. Pramada

Abstract For hydrological analysis, it is essential to have continuous and long-term precipitation data. However, the precipitation data from rain gauge stations are often insufficient and not continuous. At present, ground-based gridded data and satellite-based gridded data are often used as an alternative. However, these data sets have to be evaluated for their suitability in hydrological studies. The current study compared three different rainfall data sources with the observed station data for the Kallada River basin of Kerala, India. The ground-based gridded rainfall data from the India Meteorological Department (IMD), the high-resolution satellite product Tropical Rainfall Measuring Mission (TRMM 3B43, version 7), and the reanalysis data Modern-Era Retrospective Analysis for Research and Applications (MERRA) are used in the analysis. The correlation coefficient, normalized root mean square error, Nash–Sutcliffe efficiency, modified index of agreement, and volumetric efficiency are used as performance indicators. The performance indicator's weights are based on the entropy method. The multi-criteria decision-making techniques like compromise programming and Preference Ranking Organization Method (PROMETHEE II) are used for ranking the precipitation data sources. It is found that IMD ground-based gridded data is ranked 1 among the three data sets. The IMD ground-based gridded data are not homogeneous based on the absolute homogeneity test, even though they had the highest rank. The IMD gridded data are further corrected based on double mass curve analysis. The corrected data were analyzed using the precipitation concentration index (*PCI*) to assess the temporal variation in precipitation, and it was found that the location falls under a uniform distribution zone.

Keywords Entropy · Compromise programming · PROMETHEE II · Performance indicators · River basin · Precipitation concentration index

Introduction

Accurate and reliable precipitation data are essential for modeling studies in hydrology. Precipitation is one of the most important components of the hydrological cycle and cannot be ignored in hydrological modeling (Pascale et al., 2015). For hydro-meteorological analysis in many regions, long-term data sets for rain gauge observations are unavailable (Salman et al., 2018). Remote sensing satellite-based precipitation can provide homogeneous, continuous information in space and time over a region (Roca, 2019). At present, various gridded data sets like ground- and satellite-based data sets are increasingly used in hydrological studies. Even though many data sources for precipitation are available, these data sets may not be consistent due to

B. S. N. Reddy (✉) · S. P. V. · S. K. Pramada
Department of Civil Engineering, National Institute
of Technology Calicut, Kozhikode, Kerala, India
e-mail: satya2243@gmail.com

different sources and estimation procedures (Tapiador et al., 2017). Hence, it is very imperative to evaluate the suitability of these time series data for hydrological modeling studies. Few studies have been carried out recently that compare various gridded data sources (Beck et al., 2017; Cattani et al., 2016; Hu et al., 2016; Prakash et al., 2016a, 2016b; Salman et al., 2018; Sharannya et al., 2020; Sireesha et al., 2020; Sun et al., 2018). These methods can be classified under two groups, namely (1) comparison of the statistical indices using a reference data set and (2) evaluation of its use for a specific application.

Sharannya et al. (2020) performed analysis of gridded precipitation data, specifically Tropical Rainfall Measuring Mission (TRMM) and the Climate Hazards Group Infra-Red Station Precipitation (CHIRPS) data sets, for the an Indian river catchment named Gurupura. They have used Soil and Water Assessment Tool (SWAT) in order to simulate stream flows and compare them with flows generated by the India Meteorological Department (IMD). The TRMM outperformed CHIRPS in terms of rainfall estimation, based on the statistical results. Sireesha et al. (2020) evaluated the performance of gridded precipitation data sets, namely, Global Precipitation Climatology Centre (GPCC), TRMM, and Modern-Era Retrospective Analysis for Research and Applications (MERRA) in the Sina basin, India. The statistical indicators, percentage bias (*PBias*), normalized root-mean-square error (*NRMSE*), Nash–Sutcliffe efficiency (*NSE*), modified index of agreement (*MD*), and volumetric efficiency (*VE*) were used to check the suitability of the gridded data sets. The selected gridded precipitation data sets were ranked using compromise programming (CP). TRMM occupies the first position, followed by MERRA.

In the present study, three gridded data sets, namely India Meteorological Department gridded data (IMD gridded), TRMM, and MERRA were used. Among these data sets, IMD gridded data are prepared using Shepard's interpolation using the daily precipitation data from 6995 observed stations over India after controlling the quality of observed station data (Pai et al., 2014). TRMM data are based on remote sensing, while MERRA data are based on reanalysis. In India, recently, many studies have been carried out by taking IMD gridded data as the standard data for evaluation of satellite-based data and many other hydrological studies (Sharannya

et al., 2020; Venkatesh et al., 2020). There are only limited studies on the assessment of IMD gridded rainfall for its suitability in hydrological applications (Chowdhury et al., 2021; Subash et al., 2020). Subash et al. (2020) carried out a study to assess the error characteristics of gauge-only gridded product (IMD gridded data) and multi-satellite gridded precipitation data (TRMM-TMPA-3B42). As the reference data set, rain gauge observation data from the Kabini drainage basin in southern India were selected. Multiple visual and statistical metrics were used for evaluating the gridded data set. The results show that IMD gridded data outperforms the multi-satellite precipitation product. Chowdhury et al. (2021) analyzed the various gridded precipitation data of the Satluj River basin in India using compromise programming. The Technique for Order Preference by Similarity to an Ideal Solution in Fuzzy Field (f-TOPSIS) was applied to get the weight of the selected performance indicators. The APHRODITE (Asian Precipitation-Highly-Resolved Observational Data Integration towards Evaluation) got the highest rank, followed by IMD gridded data and ERA interim.

Few studies have ranked global climate models (GCMs) using multi-criteria decision-making (MCDM) techniques. Raju et al. (2017) ranked thirty-six GCMs based on the simulations of maximum and minimum temperatures of India across 40 grid data points. The correlation coefficient, skill score, and normalized root mean square error are used as performance indicators for assessing GCMs. The entropy method was adopted to compute the indicator weights, and the compromise programming technique is adopted for GCM ranking. Raju and Kumar (2014) adopted MCDM for evaluating 11 GCM data sets in India based on precipitation data. Raju and Kumar (2014) used Preference Ranking Organization Method (PROMETHEE II) to rank eleven GCMs for the climate variable precipitation based on five performance indicators, considering equal and varying weights. To calculate the weights, they employed the entropy method.

The precipitation data sets have to be checked for homogeneities because measurement techniques and observational procedures are different for different data sources; environment characteristics may be different. In the case of gauge-based measurements, the location of stations may be different. For the

detection of non-homogeneities in data series, there are various methods. The absolute homogeneity tests are a combination of the Pettitt test, standard normal homogeneity test (SNHT), Buishand range test (BR), and von Neumann ratio (VNR) test (Che Ros et al., 2016). In these tests, the results are classified into homogeneous, doubtful, and change point (suspect) based on the number of tests accepting the alternative hypothesis.

The precipitation concentration index (*PCI*) is an essential index for understanding the temporal variation of rainfall and for evaluating changes in the seasonal rainfall pattern (Ghorbani et al., 2021; Zhang et al., 2019). *PCI* is a valuable indicator for predicting hydrological risks like droughts and floods (Gocic et al., 2016).

The present study focuses on the Kallada River basin, Kerala, India, originating from the Western Ghats Mountains. Even though a few studies have already been carried out related to the adequacy of gridded data sets, there are only limited studies on the ranking of precipitation data sources. In addition, these techniques have not been applied to the Kerala basin. In the present study, the main objectives are (a) to examine and to assess the suitability of gridded precipitation data sets using multi-criteria decision-making methods such as the compromise programming technique and PROMETHEE II, (b) to examine the homogeneity of data sets, and (c) to rectify the top-ranked data for inconsistency, if present, and to evaluate the temporal variability in precipitation using the precipitation concentration index.
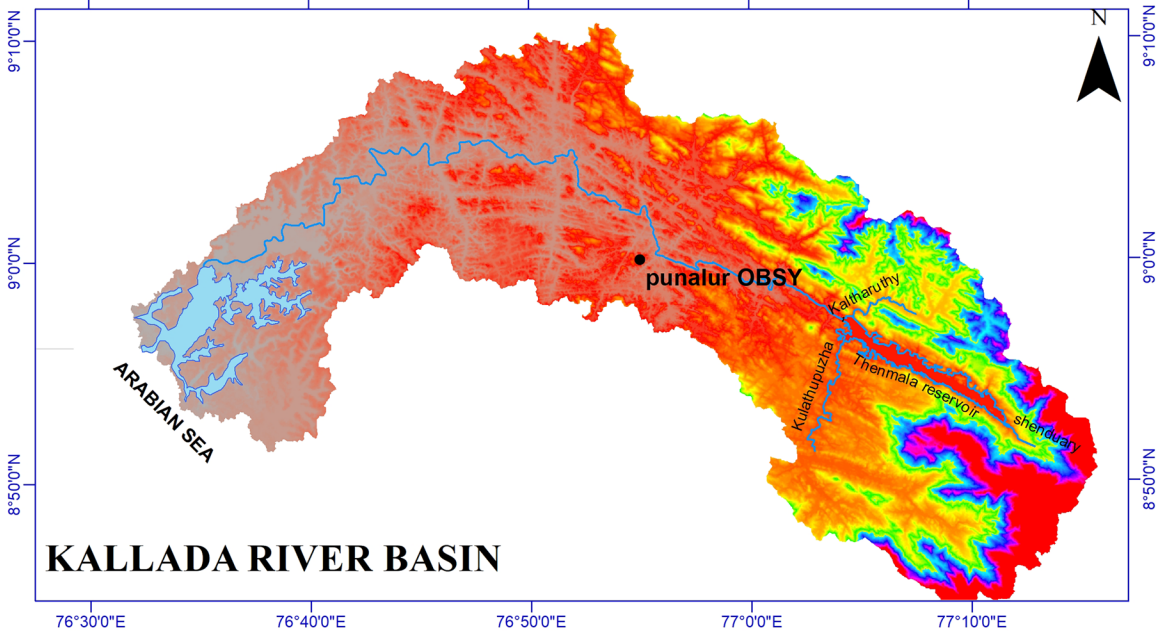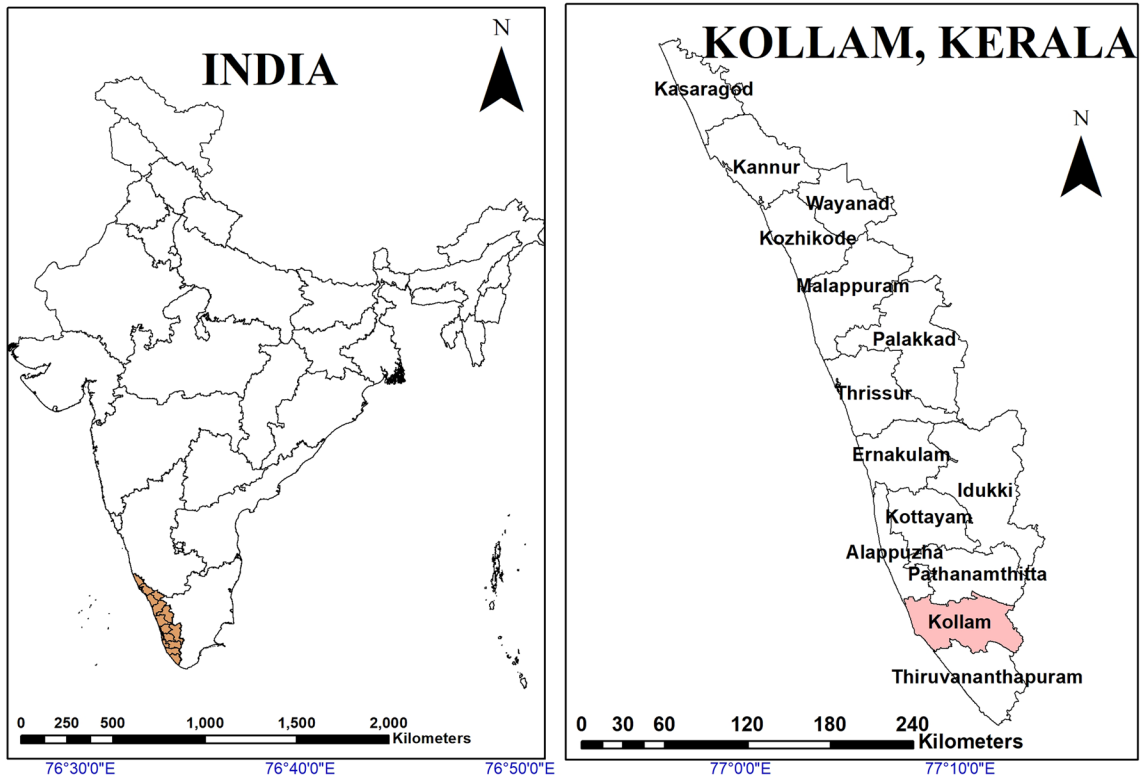
Materials and methods

Study area

The Western Ghats, also known as the Sahyadri, is a mountain range that extends 1600 km along the western coast of Peninsular India and passes through Tamil Nadu, Kerala, Karnataka, Maharashtra, and Gujarat. These mountains forms Kerala's most crucial topographic feature, significantly affecting the state's climate, vegetation, and river hydrology. In Kerala, there are 44 rivers, out of which 41 are west flowing with the origin from the Western Ghats and flows westwards towards the Arabian Sea or into the Backwaters of Kerala. The Kallada River basin was chosen as the study area, with its origin in Kulathupuzha at 1750 m. It passes through Punalur, Pathanapuram, Kunnathur, Puthoor, and Kallada for 121 km until it reaches Ashtamudi Lake. The catchment area of the river basin is 1699 km^2. The Kallada River is significant to south Kerala as a source of irrigation, electricity generation, and aquaculture (Satya narayana reddy et al., 2021). Kallada River is positioned between 8° 49′–9° 17′ north latitudes and 76° 24′–77° 16′ east longitudes (Fig. 1). This river is the confluence of three major rivers, named Kulathupuzha, Shenduary, and Kalthuruthi, which join near Parappar in Thenmala. Major rain gauge stations are located at Punalur, Kollam, and Aryankavu. The landscape is classified into four physiographic zones: lowland (0–30 m), midland (30–200 m), foothill zones (200–600 m), and highland (above 600 m) (GSI, 2005).

The climate of the Kallada River is generally tropical with marked warm and humid seasons and seasonal precipitation. In summer and winter, maximum and minimum temperatures of 34 °C and 21 °C were recorded. The mean yearly precipitation is 2600 mm. Moisture is about 90% in the wet season. The catchment receives precipitation from two types of monsoon, i.e., one is southwest (June to September) and the other is northeast (October to November) with 51.8% and 24% of yearly precipitation, respectively. The remaining 24.2% of precipitation is received during the non-monsoon season.

Data sets

Monthly observed rainfall data for Punalur station in the Kallada basin and ground-based gridded data with 0.25°×0.25° resolution is obtained from IMD Pune. The observed precipitation data and the ground-based gridded data of IMD were available from 1981 to 2013. The satellite-based gridded precipitation data TRMM is a joint mission between Japan Aerospace Exploration Agency (JAXA) and the National Aeronautics and Space Administration (NASA). Monthly TRMM 3B43, version 7 product with a spatial resolution of 0.25°×0.25° during the years 1998–2013 were downloaded from the GES DISC-NASA website in NetCDF format. MERRA-2 is a global atmospheric reanalysis data developed by the NASA Global Modeling and Assimilation Office (GMAO). It supplies a regularly gridded and homogeneous data set,

INDIA

KOLLAM, KERALA

Kasaragod
Kannur
Wayanad
Kozhikode
Malappuram
Palakkad
Thrissur
Ernakulam
Idukki
Kottayam
Alappuzha
Pathanamthitta
Kollam
Thiruvananthapuram

KALLADA RIVER BASIN

ARABIAN SEA

punalur OBSY

Kulathupuzha
Kaltharuthy
Thenmala reservoir
shenduary

Legend

- Rain gauge station
- Thenmala_dam
- Thenmala_resevoir
- Kallada_river
- Shenduary_tributary
- Asthamudi Lake
- Kulathupuzhy_tributary
- kaltharuthy_tributary
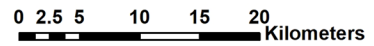
◄**Fig. 1** Study area

with $0.5° \times 0.625°$ spatial resolution for 1981–2013. MERRA data are generally classified as modern reanalysis systems with a higher spatial resolution that applies advanced numerical models and assimilation schemes to combine observations from multiple sources. Table 1 provides the details of the data.

Methodology

The inverse distance interpolation technique is used to resample the three gridded precipitation data sets to the position of the observed data point. Statistical parameters are utilized as performance indicators to compare the different gridded data with the observed data. The entropy method is adopted to determine indicator weights. Ranking of selected data sources was done using MCDM techniques, namely compromise programming (CP) and PROMETHEE II. A homogeneity test was also performed for all the data sets. The highest-ranked data set is corrected based on double mass curve analysis. The corrected data set is examined for the variability of precipitation

rate using *PCI*. The step-by-step adopted procedure is shown in Fig. 2.

Statistical indices

The statistical indices, namely correlation coefficient (*R*), normalized root mean square error (*NRMSE*), Nash–Sutcliffe efficiency (*NSE*), modified index of agreement (*MD*), and volumetric efficiency (*VE*), are determined in comparison with IMD observed station data. The statistical indices were normalized before analysis.

Correlation coefficient (R)

It is used to find how robust a relationship exists between different sets of data. If the value of $R = 1$, there is a positive correlation; if $R = -1$, then there is a negative correlation; and if $R = 0$, then there is no correlation between the data sets. It is evaluated as:

$$R = \frac{n(\sum pq) - (\sum p)(\sum q)}{\sqrt{\left[n \sum p^2 - \left(\sum p\right)^2\right]\left[n \sum q^2 - (\sum q)^2\right]}} \quad (1)$$

Table 1 Precipitation data set information

Data sets	Spatial resolution/format	Temporal resolution	Coverage	Source
IMD observed station data	-	Monthly (1981–2013)	Kallada River basin	India Meteorological Department (IMD)
IMD (gridded)	$0.25° \times 0.25°$	Monthly (1981–2013)	Regional	India Meteorological Department (IMD) (https://www.imdpune.gov.in/Clim_Pred_LRF_New/Grided_Data_Download.html)
TRMM (gridded)	$0.25° \times 0.25°$ NetCDF, ASCII	Monthly (1998–2013)	Global, land only	Tropical Precipitation Measuring Mission (TRMM) Multi-satellite Precipitation Analysis (TMPA), TRMM-3B43 V.7 (https://disc.gsfc.nasa.gov/datasets/TRMM)
MERRA (gridded)	$0.5° \times 0.625°$ NetCDF, ASCII	Monthly (1981–2013)	Global, land only	Modern-Era Retrospective Analysis for Research and Applications, MERRA V5.2.0 2D (https://disc.gsfc.nasa.gov/datasets/merra)

Fig. 2 Methodology
flowchart
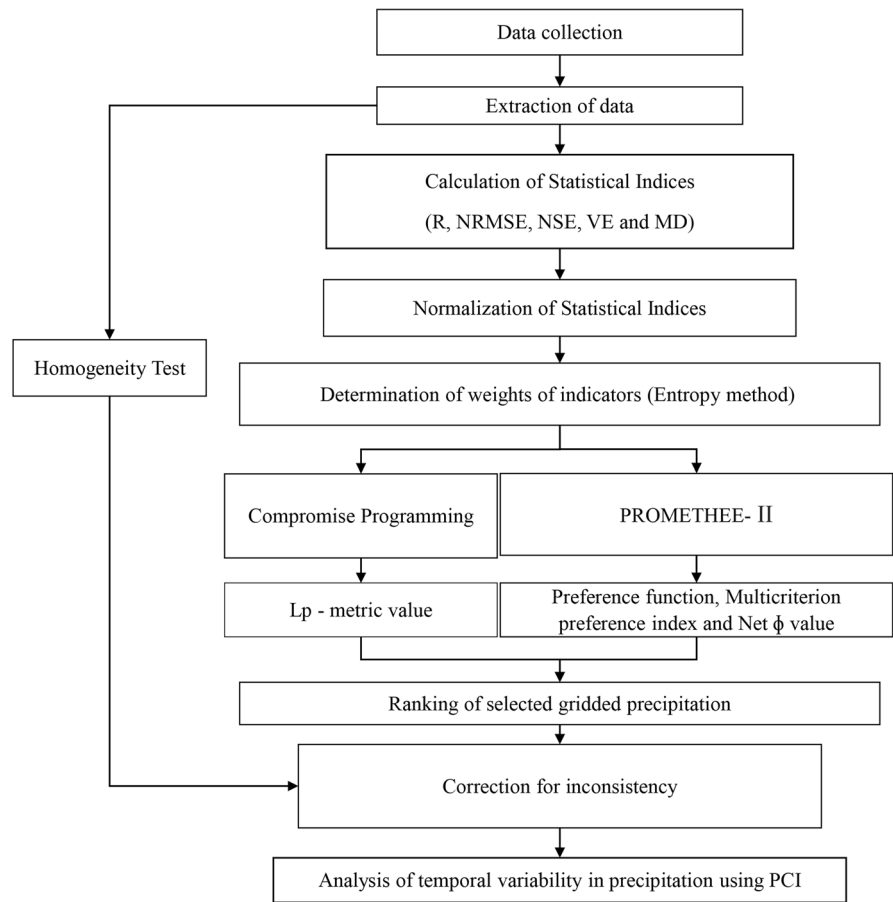
Normalized root mean square error (NRMSE)

NRMSE of 0 value means a perfect fit for the data.
NRMSE is calculated as:

$$NRMSE = \frac{\left[\frac{1}{n}\sum_{i=1}^{n}(X_{simi} - X_{obsi})^2\right]^{\frac{1}{2}}}{\frac{1}{n}\sum_{i=1}^{n}X_{obsi}} \qquad (2)$$

Nash-Sutcliffe efficiency (NSE)

This index is a normalized static measure and is calculated as:

$$NSE = 1 - \frac{\sum_{i=1}^{n}(X_{obsi} - X_{simi})^2}{\sum_{i=1}^{n}\left(X_{obsi} - \overline{X}_{obsi}\right)^2} \qquad (3)$$

A positive value shows that the estimation is good, a negative value shows that the estimation ability is poor, and 1 indicates the best model.

Modified index of agreement (MD)

MD varies between 0 and 1. A value of 1 means perfect agreement.
MD is calculated as:

$$MD = 1 - \frac{\sum_{i=1}^{n}|X_{sim} - X_{obs}|}{\sum_{i=1}^{n}(\left|X_{obs} - \overline{X}_{obs}\right| + \left|X_{sim} - \overline{X}_{obs}\right|)} \qquad (4)$$

Volumetric efficiency (VE)

It measures the ratio between observed and model precipitation volumes over a period of time. A value of 1 indicates the ideal condition.

VE is calculated as:

$$VE = 1 - \frac{\sum_{i=1}^{n}(X_{simi} - X_{obsi})}{\sum_{i=1}^{n} X_{obsi}} \quad (5)$$

Here, X_{obs} = observed precipitation, X_{sim} = gridded precipitation, and $\overline{X}_{obs}$ = mean observed precipitation.

Ranking of gridded data sets

Determination of weights of indicators using the entropy method

Various indicators' weights are determined by employing the entropy method (Raju and Kumar 2014). The weight of indicators for each gridded data set was assessed using a formulated payoff matrix (Pomerol and Romero 2000). The weights for each indicator are determined without decision maker intervention, which is the main advantage of this method, which eliminates the excessive bias against the indicator. The indicator weights can be computed as follows:

For the given normalized payoff matrix P_{ij}, the entropy E_j for the indicators j for the set of gridded precipitation data sets are computed as follows:

$$E_j = -\frac{1}{\ln(P_{ij})} \sum_{i=1}^{N} P_{ij}\ln(P_{ij}) \; for \; j = 1\ldots\ldots,J \quad (6)$$

where $i = 1, \ldots\ldots, N$ is the number of gridded precipitation data sets and j is the number of indicators.

The degree of diversification, D_j, for the information given by outcomes of indicator j is

$$D_j = 1 - E_j \; for \; j = 1\ldots\ldots,J \quad (7)$$

Normalized indicator weights are estimated as

$$W_j = \frac{D_j}{\sum_{i=1}^{J} D_j} \quad (8)$$

Compromise programming

Compromise programming is a multi-criteria approach to decision-making, based on the principle that a solution to an acceptable "distance" solution is as "similar" as possible (Raju et al., 2017; Zeleny, 2011). The L_p metric family is used as a distance measure for CP and expressed as

$$L_p(a) = \left[\sum_{j=1}^{J} W_J^{\;P} \left| f_j^* - f_j(a) \right|^p \right]^{\frac{1}{p}} \quad (9)$$

where a represents a particular precipitation data set; j is the performance indicator, $j = 1, 2, \ldots, n$; W_j is the weight of each indicator; f_j^* is the normalized ideal value for indicator j; $f_j(a)$ is the normalized value of the indicator j for the precipitation data a; and p is the metric parameter ($p = 1$ for linear measure and $p = 2$ for Euclidian squared distance measure). The precipitation data set having the least L_p metric value is considered the best.

PROMETHEE II PROMETHEE II, a multi-criteria decision-making approach (MCDM), is formulated according to the preference function approach (Brans, et al., 1986). The preference function $P_j(x, y)$ represents the degree of preference of a particular precipitation data set "x" with regard to precipitation data set "y," for a given performance indicator j and generalized criterion function. Different types of criterion functions are available, but the usual criterion was adopted in the present study, in which the preference function depends on a small positive difference $d_j(x, y)$.

The definition of preference function is as follows:

$$P_j(x, y) = \begin{bmatrix} 0 & \text{if } d_j(x, y) \leq 0 \\ 1 & \text{if } d_j(x, y) > 0 \end{bmatrix} \quad (10)$$

Then, the multi-criteria preference index $\pi(x, y)$ is the weighted average of preference function $P_j(x, y)$ defined as:

$$\pi(x, y) = \frac{\sum_{j=1}^{J} W_j P_j(x, y)}{\sum_{j=1}^{J} W_j} \quad (11)$$

Here, W_j is the weight that is assigned for each indicator, based on the entropy method. J is the performance indicator.

$$\phi^+(x) = \frac{\sum_n \pi(x, y)}{n - 1} \quad (12)$$

$$\phi^-(x) = \frac{\sum_n \pi(y, x)}{n - 1} \quad (13)$$

$$\phi(x) = \phi^+(x) - \phi^-(x) \quad (14)$$

Here, $\phi^+(x)$ and $\phi^-(x)$ are the outranking index of the gridded precipitation data set "x" in the total data set "n," and $\phi(x)$ is the overall ranking of the gridded data set "x." The gridded data set having the highest $\phi(x)$ value is considered to be the most suitable precipitation data set.

Homogeneity tests

The homogeneity test was used to check whether the given data is homogeneous over time. In other words, if there exists any significant change point in a time series, then it is classified as non-homogeneous. This test is used to identify and adjust the variation of non-climatic parameters caused due to the differences in observation procedures, time, and relocation of the gauging site (Peterson, et al., 1998). This inhomogeneity in the historical data has a high impact on the outcome of data analysis and forecast. Data homogeneity is an integral part of historical data archival. There are several methods and tools available for testing homogeneity. The most common types of these tests are the Pettitt test, standard normal homogeneity test (SNHT), Buishand test, and von Neumann ratio (VNR) test. The combination of all four tests together is called as absolute homogeneity test.

Pettitt test

It is a non-parametric ranking method widely used for continuous climate series or hydrological series data to capture a single point of change. The steps for non-parametric statistic are as follows (Pettitt, 1979):

1. Ranking of the observations (x) in increasing order (i.e., $x_1, x_2 \dots x_n$).
2. The estimation of $V_{i,n}$ is as follows:

$$V_i = n + 1 - 2r_i \, for \, i = 1, 2, 3, \dots \dots n \tag{15}$$

Here, r_i is the rank of x_i.

3. The estimation of U_i is as follows:

$$U_i = U_{i-1} + V_i \tag{16}$$

4. The value of K_n is obtained from:

$$K_n = \max_{1 \le i \le n} |U_i| \tag{17}$$

5. Finally, the estimation of P is as follows:

$$P = 2e^{\left(-\frac{6K_n^2}{n^3 + n^2} \right)} \tag{18}$$

The null hypothesis rejects if the P value is smaller than "α," whereas "α" is the level of significance.

Buishand test

This test is a parametric test that is more susceptible to deviations in the center of the data set (Costa & Soares, 2009). This test is based on the adjusted partial sum with the total deviation from the average value.

Calculation of adjusted partial sum is as follows:

$$S_0^* = 0; S_k^* = \sum_{i=1}^{k} (X_i - \overline{X}) \, for \, k = 1, 2, \dots .N \tag{19}$$

Here, $\overline{X}$ is the average of the observations in a data set ($X_1, X_2 \dots \dots X_N$) and k is the observation number where the change point has occurred.

The rescaled adjusted partial sum is calculated as:

$$S_k^{**} = \frac{S_k^*}{D_X} \, for \, k = 1, 2, \dots \dots .N \tag{20}$$

$$D_X = \sqrt{\frac{\sum_{i=1}^{N} (x_i - x_n)^2}{N}} \tag{21}$$

The statistic Q used to test homogeneity is given by:

$$Q = \max_{0 \le k \le N} |S_k^{**}| \tag{22}$$

The null hypothesis will be accepted if the $\frac{Q}{\sqrt{N}}$ value is less than the standard critical values.

Standard normal homogeneity test

In the study of climatic variations, SNHT is the most widespread homogeneity tests. SNHT is more susceptible to detecting the change points at the start and end of the series.

The statistic $T(k)$ is computed as:

$$T(k) = k\bar{z}_1^2 + (n-k)\bar{z}_2^2 \tag{23}$$

$$\bar{z}_1 = \frac{1}{k} \frac{\sum_{i=1}^{k} \left(Y_i - \overline{Y} \right)}{s} \tag{24}$$

$$\bar{z}_2 = \frac{1}{n-k} \frac{\sum_{i=k+1}^{n} \left(Y_i - \overline{Y} \right)}{s} \tag{25}$$

If there exists a change point in the data set, $T(k)$ hits the peak value during the kth year. Then, T_0 is computed as:

$$T_0 = \max(T(k)) \text{ for } 1 \le k \le n \tag{26}$$

Von Neumann ratio (VNR) test

This test detects the change point according to the statistics of N (von Neumann, 1941), which is given by:

$$N = \frac{\sum_{i=1}^{n-1} \left(Y_i - Y_{i+1} \right)^2}{\sum_{i=1}^{n} \left(Y_i - \overline{Y} \right)^2} \tag{27}$$

If the value of $N = 2$, it states that the data set is a homogeneous series, whereas if there is a change point in the data set, then the value of $N < 2$ (Buishand, 1982). The critical values of N are taken from Buishand (1982).

Precipitation concentration index (PCI)

This index is helpful to assess the variation of precipitation in annual, seasonal, and supra-seasonal scales (Michiels et al., 1992; Oliver, 1980). Based on *PCI*, the classification of precipitation distribution is

Table 2 Significance of *PCI* values

Significance	Uniform	Moderate	Irregular	Strong irregular
PCI value	$PCI \le 10$	$10 < PCI \le 15$	$15 < PCI \le 20$	$PCI > 20$

shown in Table 2 (EE et al., 2017). The *PCI* at annual scale is calculated as follows:

$$PCI_{annual} = \frac{\sum_{i=1}^{12} P_i^2}{\left(\sum_{i=1}^{12} p_i \right)^2} \times 100 \tag{28}$$

P_i = annual rainfall in an ith month.

Seasonal *PCI*s for winter (December–February), summer (Mar–May), SW monsoon (June–September), and NE monsoon (October–Nov) and supra-seasonal *PCI* for dry season (December–May) and wet season (June–November) are as follows:

$$PCI_{winter,summer} = \frac{\sum_{i=1}^{3} P_i^2}{\left(\sum_{i=1}^{3} p_i \right)^2} \times 25;$$

$$PCI_{SWmonsoon} = \frac{\sum_{i=1}^{4} P_i^2}{\left(\sum_{i=1}^{4} p_i \right)^2} \times 33;$$

$$PCI_{NEmonsoon} = \frac{\sum_{i=1}^{2} P_i^2}{\left(\sum_{i=1}^{2} p_i \right)^2} \times 17 \tag{29}$$

$$PCI_{supra-seasonal} = \frac{\sum_{i=1}^{6} P_i^2}{\left(\sum_{i=1}^{6} p_i \right)^2} \times 50 \tag{30}$$

PCI values theoretically lie between 8.3 (uniform) and 100 (extreme) distributions. Based on the values of Table 2, the type of precipitation distribution is classified.

Results and discussion

In this study, the three gridded data sets, namely IMD gridded data, TRMM, and MERRA data, are evaluated on a monthly scale with reference to the observed rain gauge data. The gridded precipitation data were resampled to the observation data point using inverse distance interpolation. The statistical indices obtained by comparing different gridded data sets with the observed data are given in Table 3. From the present study, it was observed that TRMM is showing the best performance based on the indices *R* and *NRMSE*. These results agree

Table 3 Statistical parameters

Statistical parameters	Data set (gridded precipitation)		
	IMD	TRMM	MERRA
Correlation coefficient	0.82	0.85	0.64
NRMSE	0.54	0.53	0.71
NSE	0.62	0.58	0.27
VE	0.8	0.75	0.86
MD	0.9	0.85	0.77

with the finding of Sireesha et al. (2020). Prakash et al. (2015) stated that most satellite data have difficulties in representing rainfall over orographic regions including the Western Ghats Mountains, Northeast India, and the Himalayan foothills. In terms of *NSE* and *MD*, IMD performs better. MERRA is the best based on *VE*. Thus, it is evident that one cannot finalize the performance of the gridded data sets purely based on these statistical indices alone.

The visual comparison is also carried out using the Taylor diagram, box plots, cumulative distribution function, time series plots, and scatter plots. The TRMM shows a higher correlation and less standard deviation (Fig. 3). The interquartile range of TRMM is small compared to those of the other gridded data sets for the annual series (Fig. 4).

Fig. 3 Taylor diagram of gridded precipitation data sets
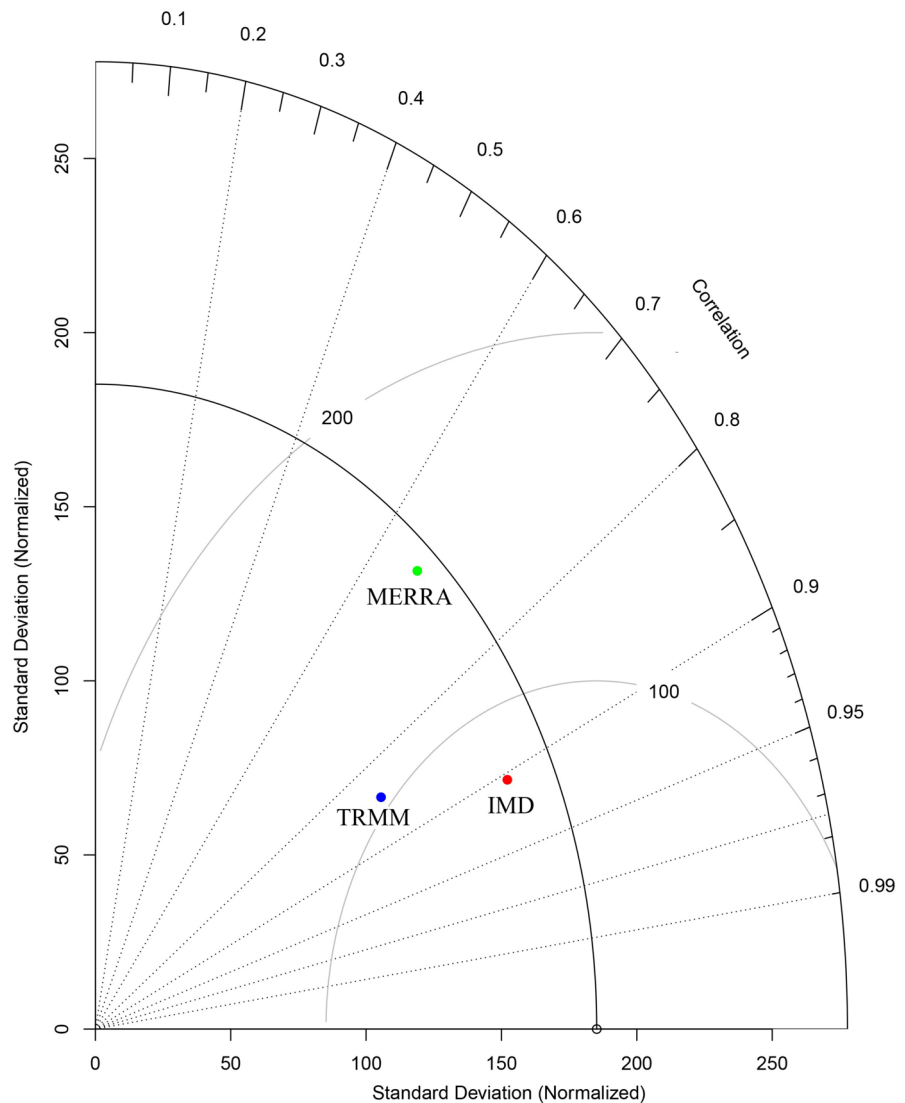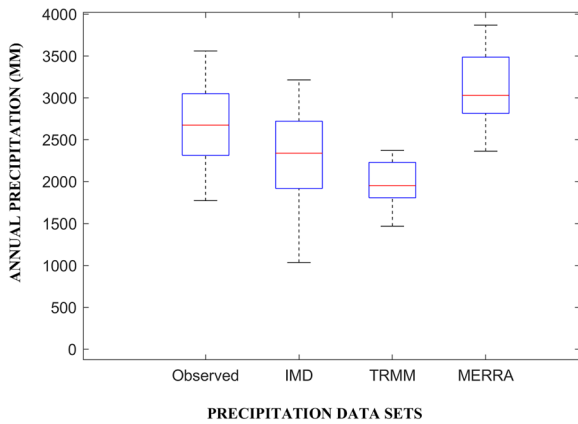
Fig. 4 Box plot of annual gridded precipitation data sets

This tallies with the study of Subash et al. (2020). This is true during the northeast monsoon season also (Fig. 5). The monthly average values of IMD gridded, TRMM, and MERRA were 193 mm, 166 mm, and 258 mm, respectively, against 221 mm for the observed data. The cumulative distribution plot of IMD gridded data compares well with the station rain gauge data (Fig. 6). Based on these figures, it is clear that the gridded IMD data are very close to the observed data; however, the MERRA data is overestimated, and the TRMM data is underestimated. Almeida et al. (2020)

stated that for a river basin in Brazil, TRMM satellite data overestimated precipitation during the wet season while it was underestimated during the dry season. Prakash et al. (2016a, 2016b) also stated that TRMM (TMPA-3B42RT) overestimates over most part of India during monsoon season. In this study, it was found that during all the months, the TRMM data is underestimated. The gridded IMD data set underestimates during the initial period, which is evident from the time series plot (Fig. 7). From the scatter plot, it is clear that TRMM is more diagonally oriented than MERRA and IMD gridded data sets (Figs. 8, 9, and 10).

Ranking of gridded data sets

Compromise programming and PROMETHEE II are applied to rank the three gridded data sets. Before applying these techniques to rank the data sets, the indicator weights are found using the entropy method. Because the performance of the gridded data set was different based on different indices, all the statistical indices were normalized before applying the entropy method. Table 4 presents the total entropy E_j, degree of diversification D_j, and normalized weights of indicators W_j. These are computed using Eqs. (6), (7), and (8), respectively. Among all five indicators, *NSE*
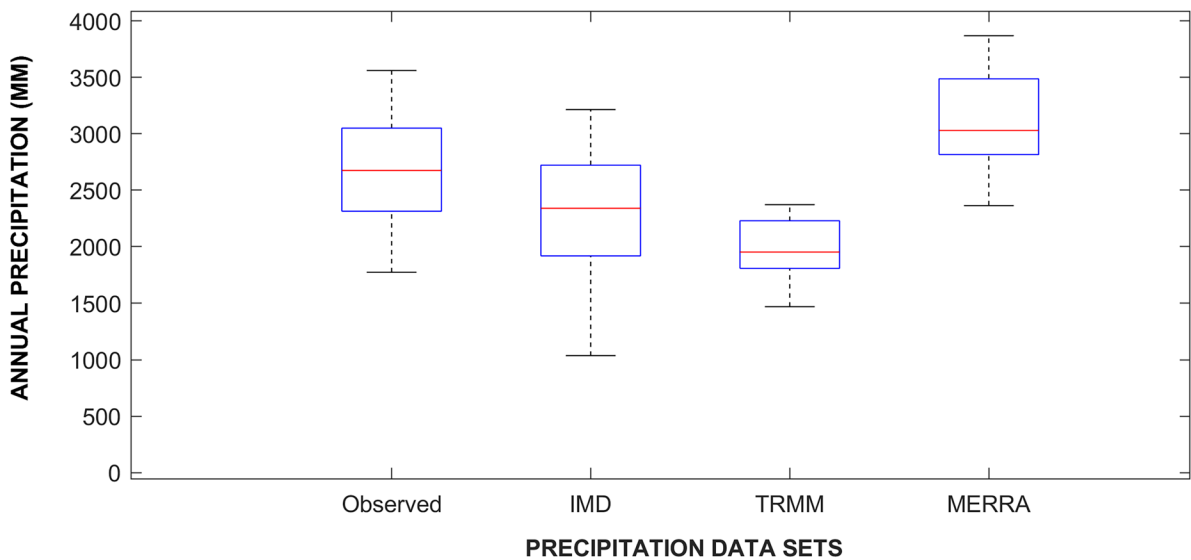

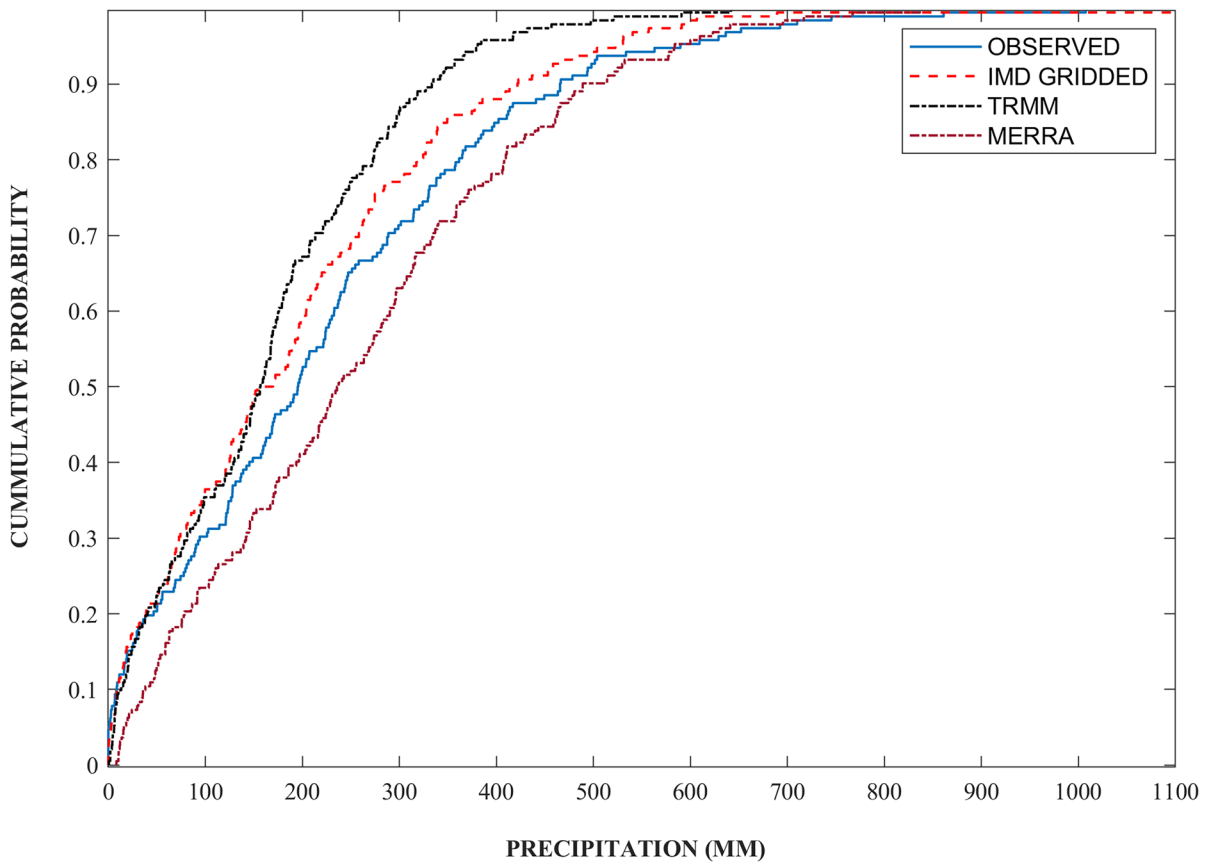**Fig. 5** Box plot of monthly gridded precipitation data sets

Fig. 6 Comparison of the non-exceedance probability of data sources

appears to have a higher significance value of 41%, indicating that its impact on the ranking of the precipitation data set is significant, whereas R, VE, and MD's total contribution is less than 20%, and $NRMSE$ contributes 26%.
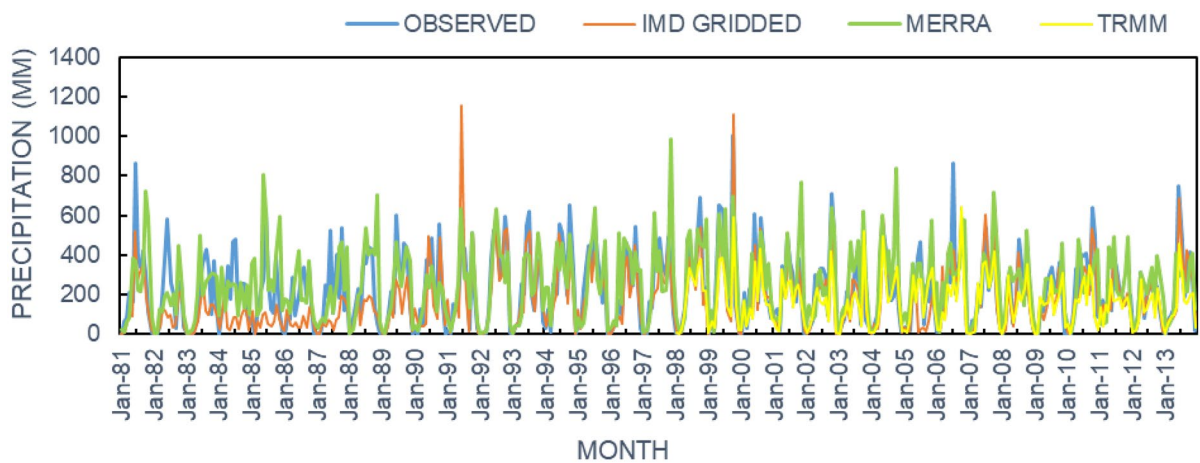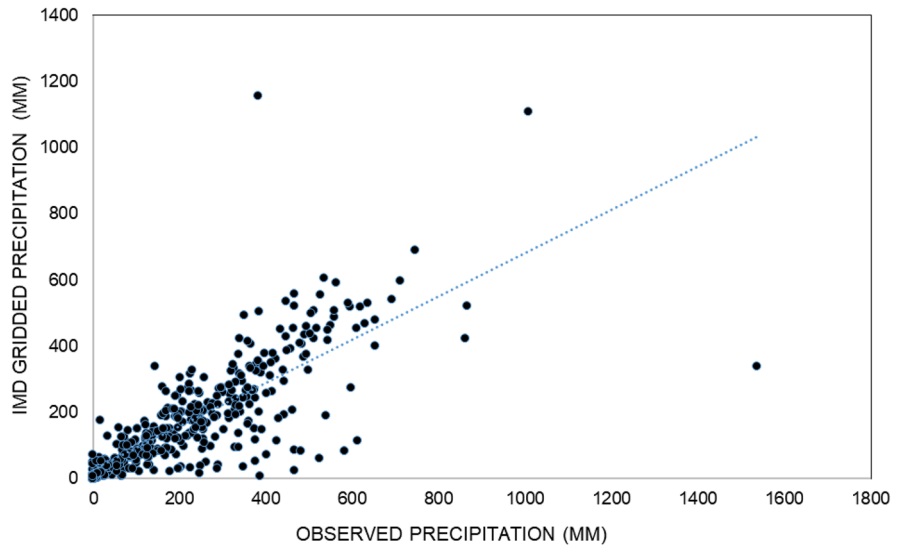

Fig. 7 Monthly rainfall time series of Punalur station and gridded data

Fig. 8 Observed precipitation vs IMD gridded data set scatter plot

Ranking using compromise programming

The CP technique is used to rank data sets, which calculates the deviation between ideal and data values (Sireesha et al., 2020; Ghorbani et al., 2021). For *R*, *NSE*, *VE*, and *MD*, the highest value are taken as the ideal, whereas for *NRMSE*, the lowest value is taken. The ideal values of statistical indicators, namely *R*, *NRMSE*, *NSE*, *VE*, and *MD*, are found to be 0.24, 0.15, 0.17, 0.26, and 0.24. The L_p metric was calculated using Eq. (9) and is given in Table 5. The L_p metric of IMD gridded data is the lowest value out of the three data sets of gridded precipitation. So IMD gridded is ranked 1, followed by TRMM and MERRA.

Ranking using PROMETHEE II

The function of the usual criterion of Brans et al. (1986) was considered in this study. According to this function, the preference of elements is either 0 or 1. In Table 4, the difference between the correlation coefficient values of IMD and TRMM is $0.22 - 0.24 = -0.02$, and so the equivalent value of preference function is 0 as per Eq. (10) (as $-0.02 < 0$) (Raju & Kumar, 2014). Likewise, the difference between the correlation coefficient values of TRMM and IMD for *R* is 0.02, and the equivalent value of preference function is 1 as per Eq. (10) (as $0.02 > 0$). Likewise, all the indicator difference function of the pairs of gridded precipitation is estimated. The

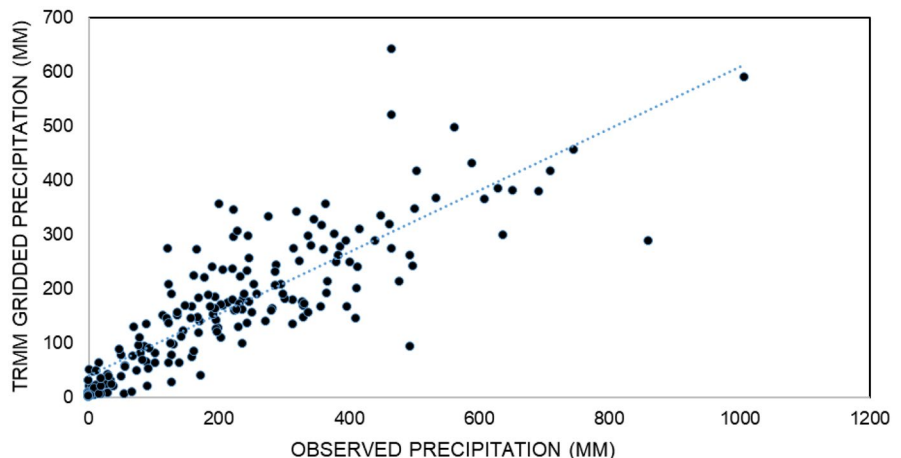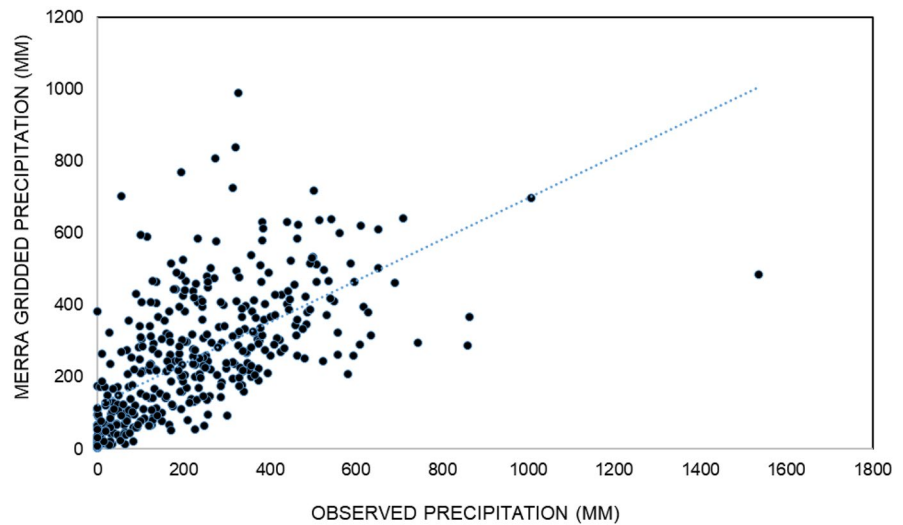**Fig. 9** Observed precipitation vs TRMM gridded data set scatter plot

Fig. 10 Observed precipitation vs MERRA gridded data set scatter plot

preference function weightage is calculated using the weights estimated by the entropy method, i.e., the multi-criteria preference index using Eq. (12), and is given in Table 6. ϕ^+, ϕ^-, and ϕ values and ranking corresponding to each data set are given in Table 7. The values in Table 7 are computed based on Eqs. (12), (13), and (14). In the case of IMD, the sum of all elements in the row from Table 6 / (number of elements -1) $= (0+0.53+0.62)$ / $(3-1) = 0.58$ (ϕ^+) (Eq. (12)). Similarly, the summation of the elements in the column / (no. of elements -1) $= (0+0.13+0.38)$ / $(3-1) = 0.26$ (ϕ^-) (Eq. (13)). The ϕ value according to Eq. (14) is 0.32 for IMD. The gridded data set having the highest value of ϕ is considered the best. Table 7 shows that, based on the ϕ value, the IMD gridded data are rated as the best data set (rank 1), and the TRMM is rated as the second-best (rank 2), followed by MERRA (rank 3) with 0.32, -0.08, and -0.24, respectively.

Table 4 Weights of indices obtained from entropy method

Data	R	$NRMSE$	NSE	VE	MD
IMD	0.22	0.15	0.17	0.22	0.24
TRMM	0.24	0.15	0.16	0.21	0.24
MERRA	0.2	0.22	0.08	0.26	0.23
Total entropy	0.91	0.82	0.72	0.92	0.94
Degree of diversification	0.09	0.18	0.28	0.08	0.06
Normalized weights	0.13	0.26	0.41	0.12	0.08

Homogeneity test

The four available precipitation monthly data sets at the Punalur location are tested for homogeneity

Table 5 Compromise programming results

Statistical index	IMD gridded	TRMM	MERRA
R	0.22	0.24	0.2
$NRMSE$	0.15	0.15	0.22
NSE	0.17	0.16	0.08
VE	0.22	0.21	0.26
MD	0.24	0.24	0.23
L_p metric	0.007	0.01	0.06

Table 6 Multi-criteria preference index values

Gridded data	IMD	TRMM	MERRA
IMD	0	0.53	0.62
TRMM	0.13	0	0.62
MERRA	0.38	0.38	0

Table 7 Values of ϕ^+, ϕ^-, and ϕ ranks of gridded precipitation data sets

	ϕ^+	ϕ^-	ϕ	Rank
IMD	0.58	0.26	0.32	1
TRMM	0.38	0.46	-0.08	2
MERRA	0.38	0.62	-0.24	3

Table 8 Homogeneity test results

Data	Pettitt test			SNHT			Buishand test			VNR			Class
	K_N	P	Trend	T_0	P	Trend	Q	P	Trend	N	P	Trend	
Observed	4028	0.73	H_0	3.6	0.72	H_0	16.96	0.42	H_0	1.23	<0.0001	H_a	Useful
IMD gridded	238	0.99	H_0	34.7	0.005	H_a	10.43	0.98	H_0	1.2	<0.0001	H_a	Doubtful
TRMM gridded	980	0.45	H_0	3.63	0.48	H_0	11.26	0.26	H_0	1.3	<0.0001	H_a	Useful
MERRA gridded	1353	0.81	H_0	5.38	0.4	H_0	9.62	0.6	H_0	1.26	<0.0001	H_a	Useful

using Pettitt, SNHT, Buishand, and von Neumann ratio tests. Test results are tested at 5% significant level. The data set is rated as non-homogeneous when the P-value is less than 5% significant level. The results are tabulated in Table 8, and the results show that all data excluding IMD gridded precipitation were homogeneous for Pettitt, SNHT, and Buishand tests, whereas for all data sets, the VNR test showed inhomogeneity characteristics. Pettitt and Buishand tests show that the IMD gridded data sets are homogeneous, while SNHT and VNR tests show that the data series is inhomogeneous. The rainfall data sets are classified into homogeneous, doubtful, and existence of change point (suspect) based on the absolute homogeneity test.

The data set is graded to be homogeneous when it rejects one or none null hypothesis, doubtful when it rejects two tests out of the four tests, and is said to be suspect when it rejects three or all tests under 5% significant level. Based on this, the results are shown in Table 8. Except for IMD gridded data set, all the data were found to be valid.

The SNHT test result graph and a double mass curve are drawn for IMD gridded data set and are given in Figs. 11 and 12. During the period 1985–1990, there exists an inconsistency, which is evident from the figures. The IMD gridded data is corrected based on the slope of the double mass curve and is given in Fig. 13.
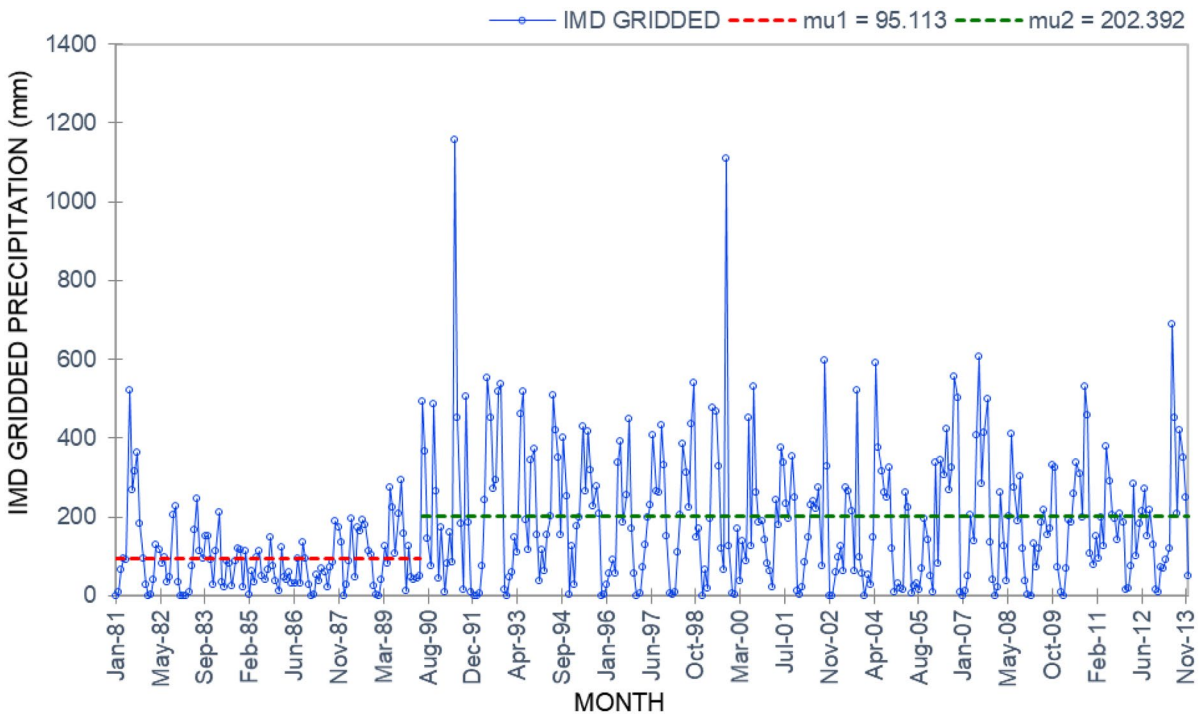

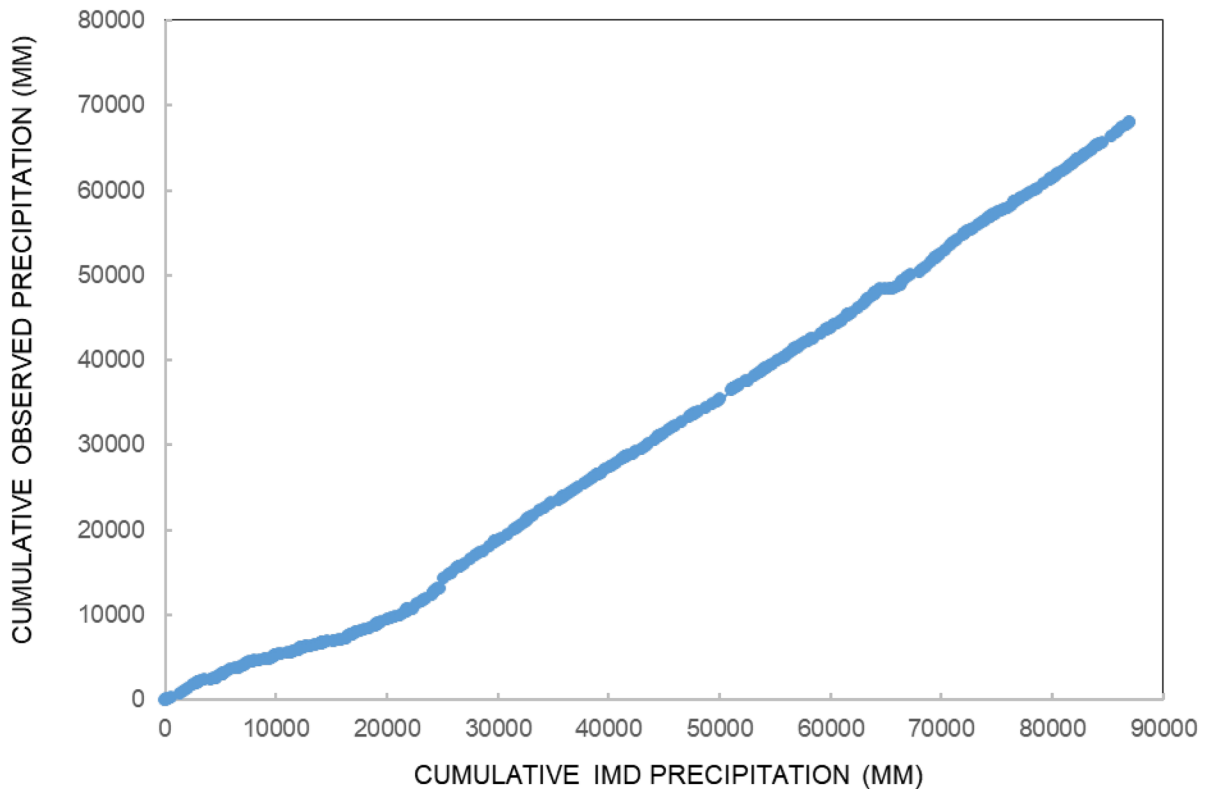**Fig. 11** SNHT test for IMD gridded data during the period of 1981–2013

Fig. 12 Double mass curve during the period of 1981–2013

Assessment of IMD gridded precipitation data set using PCI

The IMD gridded precipitation data set after correction is analyzed for the variation of the precipitation rate on a temporal scale. The estimation of *PCI* uses Eq. (28) for annual, Eq. (29) for seasonal, and Eq. (30) for supra-seasonal. The results are tabulated in Table 9 based on the significance criteria of *PCI* values. The annual *PCI* of the precipitation data set ranges from 9.80 (2011) to 20.96 (1999). Further analysis of annual *PCI* shows that 81.82% falls under the zone of moderate precipitation, whereas 12.12% in the zone of irregular precipitation, 3.03% in the zone of the strong irregularity of precipitation, and 3.03% in the zone of uniform precipitation distribution out

of 33 years of available data. The graphical representation of the yearly annual *PCI* showed in Fig. 14a.

Similarly, on a seasonal basis, *PCI* was calculated for winter, summer, and SW (southwest) and NE (northeast) monsoon seasons. In Fig. 14b, the graphical plot of calculated values is shown for seasonal variation. From Table 9, it is clear that the mean values of seasonal *PCI* show that the type distribution is strong irregular (winter) and uniform (summer, SW and NE monsoons). On the supra-seasonal basis, i.e., dry and wet seasons, *PCI* is calculated and represented in Fig. 14c. From Table 9, it is clear that for the dry season, 60.61% falls under the zone of moderate precipitation, and for the wet season, 66.67% falls under the zone of uniform precipitation.
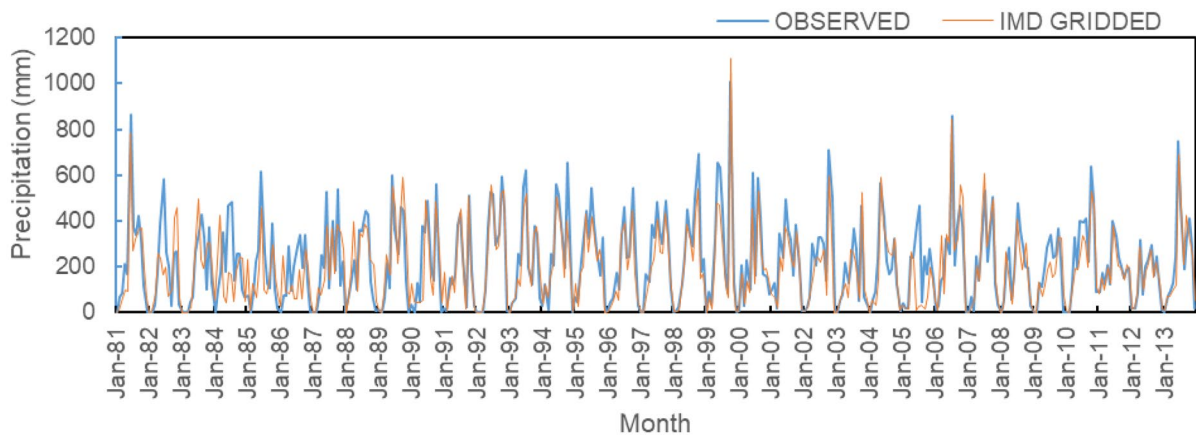
Fig. 13 Comparison of observed data with corrected IMD gridded data set

Table 9 *PCI* results of IMD gridded data set for the period 1981–2013

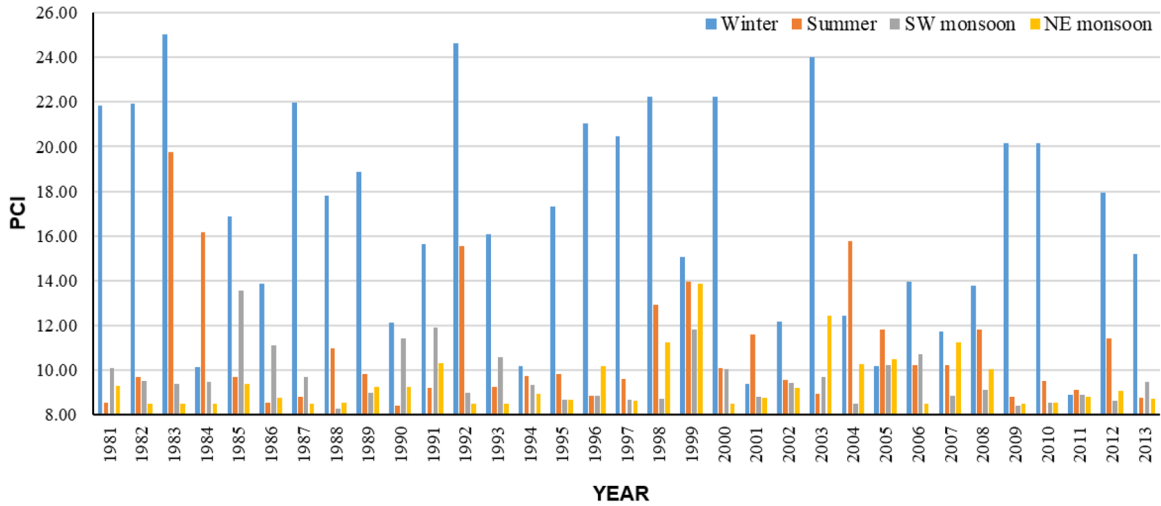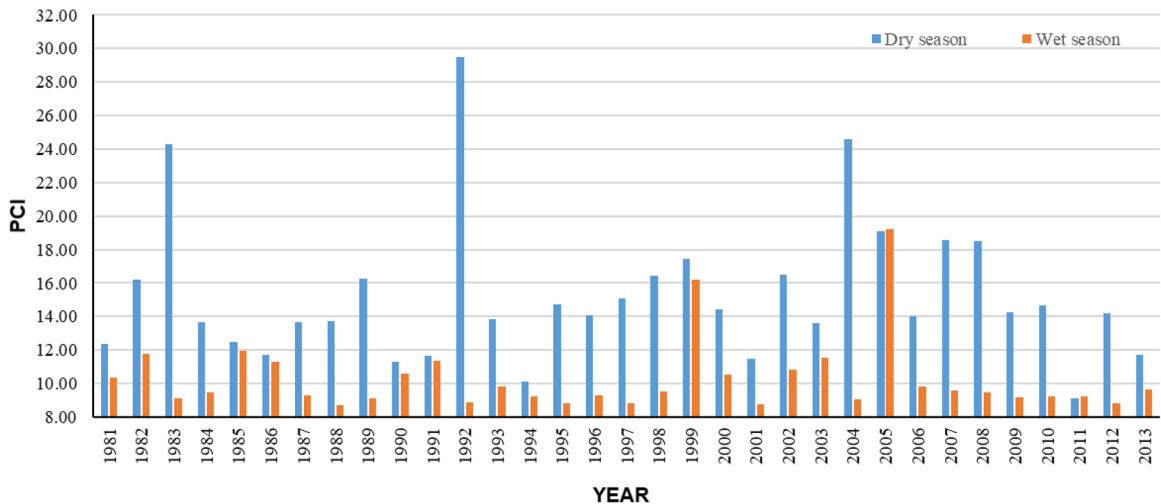|	Annual	Seasonal				Supra-seasonal	
		Winter	Summer	SW monsoon	NE monsoon	Dry season	Wet season
Uniform	3.03%	6.06%	**57.58%**	**69.70%**	**72.73%**	3.03%	**66.67%**
Moderate	**81.82%**	30.30%	30.30%	30.30%	27.27%	**60.61%**	27.27%
Irregular	12.12%	27.27%	12.12%	0.00%	0.00%	27.27%	6.06%
Strong irregular	3.03%	**36.36%**	0.00%	0.00%	0.00%	9.09%	0.00%
Mean *PCI*	13.63 (M)	16.82 (IR)	10.82 (M)	9.65 (U)	9.41 (U)	15.25 (IR)	10.26 (M)
Above average	14	17	11	12	9	11	11
Below average	19	16	22	21	24	22	22
Minimum *PCI*	9.80 (2011)	8.88 (2011)	8.42 (1990)	8.27 (1988)	8.50 (1983, 1984, 1992, 2000, 2009)	9.14 (2011)	8.72 (1988)
Maximum *PCI*	20.96 (1999)	25 (1983)	19.77 (1983)	13.54 (1985)	13.87 (1999)	29.49 (1992)	19.24 (2005)

U uniform, *M* moderate, *IR* irregular

(a)

ANNUAL PCI
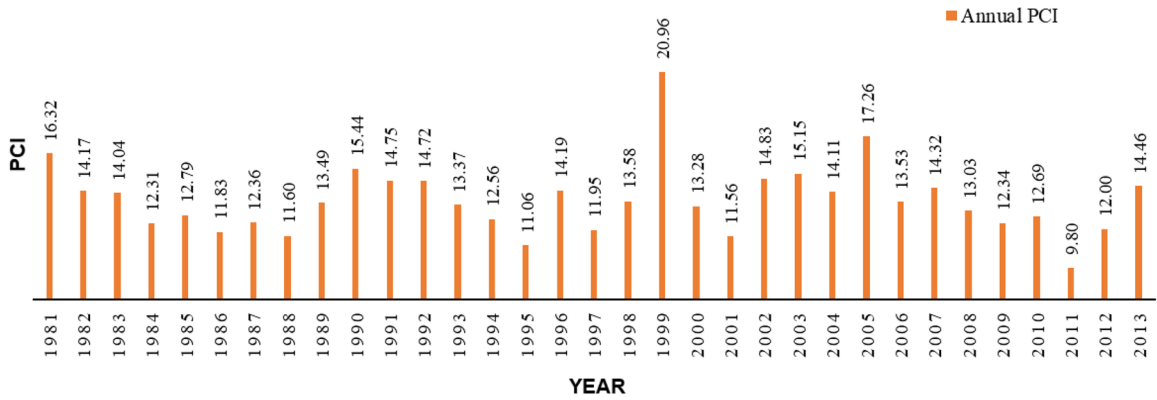
(b)

SEASONAL PCI

(c)

SUPRA-SEASONAL PCI

Conclusions

Accurate precipitation data has a significant role to play in river basin level planning and management. In the present study, a suitable data source was selected from IMD, TRMM, and MERRA gridded precipitation data sets by comparing with the IMD observed data. Multi-criteria decision-making techniques, namely compromise programming and PROMETHEE II, were employed to select the best data set. The data set ranked 1 is selected and corrected for inconsistency. The *PCI* was estimated for the corrected data set to characterize the temporal patterns of precipitation in the catchment area. One of the critical findings from the present study is that the gridded data set, whether it is gauge-based/satellite-based data set, should not be directly used for hydrological studies. A suitable correction has to be applied before its use.

The key findings are as follows:

- Based on CP and PROMETHEE II, IMD gridded data set ranked 1, followed by TRMM and MERRA.
- Gridded IMD precipitation data fails the homogeneity test. The homogeneity test and the double mass curve show that gridded IMD data have inconsistency during the periods 1985 and 1990. The gridded IMD data is corrected for inconsistency.
- The *PCI* average values of corrected gridded IMD data set for the period 1981–2013 state that the location falls in the zone of moderate for annual precipitation and uniform for summer, SW monsoon, and NE monsoon.

Author contribution Beeram Satya Narayana Reddy: Formal assessment, conceptualization, data collection, framework of methodology, model development, software, initial draft writing, review and editing. Shahanas P V: Formal assessment, conceptualization, data collection, framework of methodology, model development, software, and initial draft writing. S K Pramada: Formal assessment, conceptualization, methodology, resources, supervision, and validation.

References

Beck, H. E., Van Dijk, A. I. J. M., Levizzani, V., Schellekens, J., & Miralles, D. G. (2017). MSWEP: 3-hourly 0.25° global gridded precipitation (1979–2015) by merging gauge, satellite, and reanalysis data. *Hydrology and Earth System Sciences, 21*, 589–615. https://doi.org/10.5194/hess-21-589-2017

Brans, J. P., Vincke, P., & Mareschal, B. (1986). How to select and how to rank projects: The PROMETHEE method. *European Journal of Operational Research, 24*, 228–238.

Buishand, T. A. (1982). Homogeneity of rainfall records. *Journal of Hydrology, 58*(2), 11–27.

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

Che Ros, F., Tosaka, H., Sidek, L. M., & Basri, H. (2016). Homogeneity and trends in long-term rainfall data, Kelantan River Basin, Malaysia. *International Journal of River Basin Management, 14*(2), 151–163. https://doi.org/10.1080/15715124.2015.1105233

Chowdhury, B., Goel, N. K., & Arora, M. (2021). Evaluation and ranking of different gridded precipitation datasets for Satluj River basin using compromise programming and f-TOPSIS. *Theoretical and Applied Climatology, 143*, 101–114.

Costa, A. C., & Soares, A. (2009). Homogenization of climate data : Review and new perspectives using geostatistics. *Mathematical Geosciences, 41*, 291–305. https://doi.org/10.1007/s11004-008-9203-3

De Almeida, K. N., Antônio, J., & Buarque, D. C. (2020). Performance analysis of TRMM satellite in precipitation estimation for the Itapemirim River basin, Espirito Santo state, Brazil. *Theoretical and Applied Climatology, 141*, 791–802.

Ezenwaji EE, Nzoiwu CP, Chima GN (2017) Analysis of Precipitation Concentration Index (PCI) for Awka Urban Area, Nigeria. Hydrol Current Res, *08*(04), 4–9. https://doi.org/10.4172/2157-7587.1000287

Ghorbani, M. A., Kahya, E., & Roshni, T. (2021). Entropy analysis and pattern recognition in rainfall data, north Algeria. *Theoretical and Applied Climatology, 144*, 317–326.

Gocic, M., Shamshirband, S., Razak, Z., T, D. P., Ch, S., & Trajkovic, S. (2016). Long-term precipitation analysis and estimation of precipitation concentration index using three

support vector machine methods. *Advances in Meteorology*, 11. https://doi.org/10.1155/2016/7912357

GSI. (2005). Geology and mineral resources of the states of India part IX – Kerala. *Miscellaneous Publication, 211*(30), 2–5.

Hu, Z., Hu, Q., Zhang, C., Chen, X., & Li, Q. (2016). Evaluation of reanalysis, spatially interpolated and satellite remotely sensed precipitation data sets in central Asia. *Journal of Geophysical Research: Atmospheres, 121*, 5648–5663.

Michiels, P., Gabriels, D., & Hartmann, R. (1992). Using the seasonal and temporal Precipitation concentration index for characterizing the monthly rainfall distribution in Spain. *CATENA, 19*(1), 43–58. https://doi.org/10.1016/0341-8162(92)90016-5

Oliver, J. E. (1980). Monthly precipitation distribution: A comparative index. *Professional Geographer, 32*(3), 300–309. https://doi.org/10.1111/j.0033-0124.1980.00300.x

Pai, D. S., Sridhar, L., Rajeevan, M., Sreejith, O. P., Satbhai, N. S., & Mukhopadhyay, B. (2014). Development of a new high spatial resolution (0.25° × 0.25°) long period (1901–2010) daily gridded rainfall data set over India and its comparison with existing data sets over the region D. *Mausam, 65*(1), 1–18.

Pascale, S., Lucarini, V., & Feng, X. (2015). Analysis of rainfall seasonality from observations and climate models. *Climate Dynamics, 44*, 3281–3301. https://doi.org/10.1007/s00382-014-2278-2

Peterson, T. C., Easterling, D. R., Karl, T. R., Groisman, P., Nicholls, N., Plummer, N., & Parker, D. (1998). Homogeneity adjustments of in situ atmospheric climate data: A review. *International Journal of Climatology, 18*, 1493–1517.

Pettitt, A. N. (1979). A non-parametric approach to the change-point problem. *Journal of the Royal Statistical Society, 28*(2), 126–135. https://doi.org/10.1016/j.epsl.2008.06.016

Pomerol, J. C., & Romero, S. B. (2000). Multicriterion decision in management: principles and practice. Kluwer Academic, Netherlands.

Prakash, S., Mitra, A. K., Aghakouchak, A., Liu, Z., Norouzi, H., & Pai, D. S. (2016a). A preliminary assessment of GPM-based multi-satellite precipitation estimates over a monsoon dominated region. *Journal of Hydrology*, 1–12. https://doi.org/10.1016/j.jhydrol.2016.01.029

Prakash, S., Mitra, A. K., Momin, I. M., Rajagopal, E. N., Basu, S., Collins, M., & Ashok, K. (2015). Seasonal inter-comparison of observational rainfall datasets over India during the southwest monsoon season. *International Journal of Climatology, 35*, 2326–2338. https://doi.org/10.1002/joc.4129

Prakash, S., Mitra, A. K., Rajagopal, E. N., & Pai, D. S. (2016b). Assessment of TRMM-based TMPA-3B42 and GSMaP precipitation products over India for the peak southwest. *International Journal of Climatology, 36*, 1614–1631. https://doi.org/10.1002/joc.4446

Raju, K. S., & Kumar, D. N. (2014). Ranking of global climate models for India using multicriterion analysis. *Climate Research, 60*, 103–117. https://doi.org/10.3354/cr01222

Raju, K. S., Sonali, P., & Kumar, D. N. (2017). Ranking of CMIP5-based global climate models for India using compromise programming. *Theoretical and Applied Climatology, 128*, 563–574. https://doi.org/10.1007/s00704-015-1721-6

Roca, R. (2019). Estimation of extreme daily precipitation thermodynamic scaling using gridded satellite precipitation products over tropical land. *Environmental Research Letters, 14*(095009).

Salman, S. A., Shahid, S., Ismail, T., Al-abadi, A. M., Wang, X., & Chung, E. (2018). Selection of gridded precipitation data for Iraq using compromise programming. *Measurement*. https://doi.org/10.1016/j.measurement.2018.09.047

Satya narayana reddy, B., Pramada, S. K., & Roshni, T. (2021). Monthly surface runoff prediction using artificial intelligence : A study from a tropical climate river basin. *Journal of Earth System Science, 130*(35), 1–15. https://doi.org/10.1007/s12040-020-01508-8

Sharannya, T. M., Al-Ansari, N., Barma, S. D., & Mahesha, A. (2020). Evaluation of satellite precipitation products in simulating streamflow in a humid tropical catchment of India using a semi-distributed hydrological model. *Water, 12*(9), 2400. https://doi.org/10.3390/w12092400

Sireesha, C., Roshni, T., & Jha, M. K. (2020). Insight into the precipitation behavior of gridded precipitation data in the Sina basin. *Environmental Monitoring and Assessment, 192*(729).

Subash, Y., Teegavarapu, R. S. V., & Muddu, S. (2020). Evaluation and bias corrections of gridded precipitation data for hydrologic modelling support in Kabini River basin, India. *Theoretical and Applied Climatology, 140*, 1495–1513.

Sun, Q., Miao, C., Duan, Q., Ashouri, H., Sorooshian, S., & Hsu, K.-L. (2018). A review of global precipitation data sets: Data sources, estimation, and intercomparisons. *Reviews of Geophysics, 56*, 79–107. https://doi.org/10.1002/2017RG000574

Tapiador, F. J., Navarro, A., Levizzani, V., GArcia-Ortega, E., Huffman, G. J., Kidd, C., … Turk, F. J. (2017). Global precipitation measurements for validating climate models. *Atmospheric Research, 197*, 1–20 https://doi.org/10.1016/j.atmosres.2017.06.021

Venkatesh, K., Krakauer, N. Y., Sharifi, E., & Ramesh, H. (2020). Evaluating the performance of secondary precipitation products through statistical and hydrological modeling in a mountainous tropical basin of India. *Advances in Meteorology*, 23.

von Neumann, J. (1941). Distribution of the ratio of the mean square successive difference to the variance. *The Annals of Mathematical Statistics, 12*(4), 367–395.

Zeleny, M. (2011). Multiple criteria decision making (MCDM): From paradigm lost to paradigm regained ?†. *Journal of Multi-Criteria Decision Analysis, 89*, 77–89. https://doi.org/10.1002/mcda

Zhang, K., Yao, Y., Qian, X., & Wang, J. (2019). Various characteristics of precipitation concentration index and its cause analysis in China between 1960 and 2016. *International Journal of Climatology, 39*(12), 4648–4658. https://doi.org/10.1002/joc.6092