Applied Computing and Geosciences 20 (2023) 100145

Contents lists available at ScienceDirect

Applied Computing and Geosciences

journal homepage: www.sciencedirect.com/journal/applied-computing-and-geosciences

Evaluating Imputation Methods for rainfall data under high variability in Johor River Basin, Malaysia

Zulfaqar Sa'adi [a,b,*], Zulkifli Yusop [a,b], Nor Eliza Alias [a,b], Ming Fai Chow [c], Mohd Khairul Idlan Muhammad [d], Muhammad Wafiy Adli Ramli [e], Zafar Iqbal [f], Mohammed Sanusi Shiru [g], Faizal Immaddudin Wira Rohmat [h,i], Nur Athirah Mohamad [d], Mohamad Faizal Ahmad [j]

[a] *Centre for Environmental Sustainability and Water Security, Research Institute for Sustainable Environment, Universiti Teknologi Malaysia, 81310 UTM Skudai, Johor Bahru, Malaysia*
[b] *Department of Water and Environmental Engineering, Faculty of Civil Engineering, Universiti Teknologi Malaysia, 81310 UTM Skudai, Johor Bahru, Malaysia*
[c] *Department of Civil Engineering, School of Engineering, Monash University Malaysia, Jalan Lagoon Selatan, 47500 Bandar Sunway, Selangor, Malaysia*
[d] *Department of Water & Environmental Engineering, School of Civil Engineering, Faculty of Engineering, Universiti Teknologi Malaysia (UTM), 81310, Johor Bahru, Malaysia*
[e] *School of Humanities, Universiti Sains Malaysia, 11700, Penang, Malaysia*
[f] *NUST Institute of Civil Engineering-SCEE, National University of Sciences and Technology (NUST), H-12, Islamabad, 44000, Pakistan*
[g] *Department of Environmental Sciences, Faculty of Science, Federal University Dutse, P.M.B 7156, Dutse, Nigeria*
[h] *Water Resources Development Center, Bandung Institute of Technology, Indonesia*
[i] *Water Resources Research Group, Faculty of Civil and Environmental Engineering, Bandung Institute of Technology, Indonesia*
[j] *Faculty of Civil Engineering, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Malaysia*

ARTICLE INFO

Keywords:
Daily rainfall
Johor river basin
Missing data
Multiple imputation methods
Peninsular Malaysia
Spatiotemporal variability

ABSTRACT

Missing values in rainfall records might result in erroneous predictions and inefficient management practices with significant economic, environmental, and social consequences. This is particularly important for rainfall datasets in Peninsular Malaysia (PM) due to the high level of missingness that can affect the inherent pattern in the highly variable time series. In this work, 21 target rainfall stations in the Johor River Basin (JRB) with daily data between 1970 and 2015 were used to examine 19 different multiple imputation methods that were carried out using the Multivariate Imputation by Chained Equations (MICE) package in R. For each station, artificial missing data were added at rates of up to 5%, 10%, 20%, and 30% for different types of missingness, namely, Missing Completely At Random (MCAR), Missing At Random (MAR), and Missing Not At Random (MNAR), leaving the original missing data intact. The imputation quality was evaluated based on several statistical performance metrics, namely mean absolute error (MAE), root mean square error (RMSE), normalized root mean square error (NRMSE), Nash-Sutcliffe efficiency (NSE), modified degree of agreement (MD), coefficient of determination (R2), Kling-Gupta efficiency (KGE), and volumetric efficiency (VE), which were later ranked and aggregated by using the compromise programming index (CPI) to select the best method. The results showed that linear regression predicted values (*norm.predict*) consistently ranked the highest under all types and levels of missingness. For example, under MAR, MNAR, and MCAR, this method showed the lowest MAE values, ranging between 0.78 and 2.25, 0.93–2.57, and 0.87–2.43, respectively. It also consistently shows higher NSE and R2 values of 0.71–0.92, 0.6–0.92, and 0.66–0.91, and 0.77–0.92, 0.71–0.93, and 0.75–0.92 under MAR, MCAR, and MNAR, respectively. The methods of *mean*, *rf*, and *cart* also appear to be efficient. The incorporation of the compromise programming index (CPI) as a decision-support tool has enabled an objective assessment of the output from the multiple performance metrics for the ranking and selection of the top-performing method. During validation, the Probability Density Function (PDF) demonstrated that even with up to 30% missingness, the shape of the distribution was retained after imputation compared to the actual data. The methodology

* Corresponding author. Centre for Environmental Sustainability and Water Security, Research Institute for Sustainable Environment, Universiti Teknologi Malaysia, 81310 UTM Skudai, Johor Bahru, Malaysia.

E-mail addresses: zulfaqar@utm.my (Z. Sa'adi), zulyusop@utm.my (Z. Yusop), noreliza@utm.my (N.E. Alias), chow.mingfai@monash.edu (M.F. Chow), mohdkhairulidlan@utm.my (M.K.I. Muhammad), mwadli2@gmail.com (M.W.A. Ramli), zafar.thalvi@nice.nust.edu.pk (Z. Iqbal), shiru.sanusi@gmail.com (M.S. Shiru), faizalrohmat@itb.ac.id (F.I.W. Rohmat), nathirah75@live.utm.my (N.A. Mohamad), faizal_9273@yahoo.com (M.F. Ahmad).

https://doi.org/10.1016/j.acags.2023.100145
Received 24 July 2023; Received in revised form 22 November 2023; Accepted 1 December 2023
Available online 6 December 2023
2590-1974/© 2023 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

proposed in this study can help in choosing suitable imputation methods for other tropical rainfall datasets, leading to improved accuracy in rainfall estimation and prediction.

1. Introduction

Accurate daily rainfall data is crucial for hydro-meteorological analysis in climate research, and its absence can have a negative impact on hydro-climatological management, flood forecasting, irrigation scheduling, and water resource administration. The analysis of long-term series data allows researchers to identify patterns, trends, and anomalies that may not be apparent in shorter-term datasets. Therefore, long-term data completeness is crucial for obtaining high-quality analysis of rainfall, particularly in developing countries like Malaysia, which frequently have the issue of missing values (Kamaruzaman et al., 2017). Missing rainfall data is an unavoidable and persistent issue, resulting from various factors like extreme weather, environmental changes, observation errors, procedural modifications, station reorganization, instrument malfunctions, and human error (Burhanuddin et al., 2021). Missing data, whether sporadic or systematic, can lead to bias in estimation and predictions, leading to inconsistencies in rainfall records (Burhanuddin et al., 2021; Chiu et al., 2019b).

Excluding missing values in rainfall datasets during data preprocessing is a common approach but not recommended due to the potential discontinuity and significant loss of important information (Nor et al., 2020). The importance of imputing incomplete data using appropriate methods to ensure accurate analysis remains a central tenet of time series analysis. Understanding the reasons that cause missing data is crucial for effective imputation, as it helps establish connections to the primary causes and relationships between measured variables and data incompleteness (Chiu et al., 2019a). The rainfall records that are missing are usually categorized as Missing Completely At Random (MCAR). This is because the probability of any specific data being missing is unrelated to both observed and unobserved data, as well as any variables within the dataset (Burhanuddin et al., 2021; Hamzah et al., 2021; Nor et al., 2020). According to Kalteh and Hjorth (2009), MCAR assume that the occurrence of missing values is unrelated to any unobserved data. This implies that the probability of data being missing is independent of any observation in the dataset. For model-based methods, this presumption is necessary, and it is plausible to presume that missing rainfall data adheres to the MCAR process. According to Hanaish et al. (2013), missing values in Malaysian rainfall data are MCAR, which denotes that the cause for the missingness is either unrelated to the values that are missing or unrelated to the observed data. But it is important to take into account several types of missingness, such as Missing At Random (MAR) and Missing Not At Random (MNAR), throughout the imputation process. Given the constraints and uncertainties related to missing data that cannot be completely clarified by observed variables alone, the imputation method that takes into account all types of missingness may offer more accurate estimates and forecasts of rainfall.

The imputation of missing data at a target station from a nearby station is a common practice in the domains of hydro-climatology and related disciplines (Shaharudin et al., 2020). These missing values are approximated using a variety of methods, including function-fitting, statistical, and empirical techniques (Chiu et al., 2021; Miró et al., 2017; Nor et al., 2020), or more straightforward methods, including substituting missing numbers with the mean or median (Al-Khwarizmi et al., 2016; Nor et al., 2020). Other methods, like spatial interpolation techniques like inverse distance weighting average, normal ratio, simple arithmetic average, kriging, and co-kriging, are also widely used in a variety of geographical contexts (Martínez et al., 2019; Pinthong et al., 2022). Additionally, statistical techniques like multiple linear regression and correlation coefficient weighting offer effective substitutes for estimating missing values at the target station (Latrubesse et al., 2022; Pinthong et al., 2022).

The optimal method for imputing rainfall data must maintain the essential characteristics of the datasets and follow the unique rainfall patterns of specific locations, as stated by Nor et al. (2020). Multiple imputations have gained popularity as an effective technique for handling missing data in recent years, surpassing alternative methods in predicting missing rainfall values, as evidenced by the studies of Sattari et al. (2017), Miró et al. (2017), Jakhar et al. (2018), and Milo et al. (2019a). Multiple imputations, accounting for uncertainty and variability during the process, may yield more accurate and reliable estimation results compared to single imputations, enabling robust statistical inference (Burhanuddin et al., 2021; Enders, 2010). The R programming language offers various packages for multiple imputation approaches to handle missing data problems, including "mi", "Hmisc", "MICE", "missForest", and the "Amelia II package". In addition to multiple imputation, various machine learning-based imputation methods have been introduced, such as artificial neural networks (Canchala-Nastar et al., 2019; Norazizi and Deni, 2019a), random forests (Addi et al., 2022; Appiah-Badu et al., 2022; Chivers et al., 2020), gradient boosting (Chivers et al., 2020; Gorshenin and Martynov, 2019), bootstrapping (Addi et al., 2022; Chen et al., 2019), and bayesian (de Carvalho et al., 2017; Lai and Kuok, 2019), among others. Other imputation methods that have received traction in recent years are satellite retrieval, which uses satellite-based products such as IMERG-GPM for direct imputation (Latrubesse et al., 2022).

Even though various comparative studies have been done between these imputation methods over the years, mixed results on the performance of the methods were found depending on the performance metric used and the geographical and climate context (Addi et al., 2022). For example, a study by Norazizi and Deni (2019a) concluded that the artificial neural network was the best imputation method, followed by MICE, and bootstrapping and expectation maximization algorithm method. Carvalho et al. (2017) found that multiple imputation performs better than geo-statistical techniques such as ordinary kriging and co-kriging. Another study by Balcha et al. (2023) found that the majority of stations demonstrated good performance through multiple linear regression and multiple imputation.

The variability in results observed among different imputation methods can be attributed to the distinct assumptions and algorithms that may not always hold true for the specific dataset or context in which they are applied. Besides, the climate pattern and distribution of missing data within the dataset can greatly influence the performance of imputation methods. Certain methods may excel when handling data MCAR, while others may perform better when data is MAR or MNAR. The choice of imputation model or technique may interact differently with the underlying data characteristics, such as the presence of outliers, the degree of multicollinearity, or the complexity of relationships between variables. Additionally, the quality and quantity of available auxiliary information for imputation can vary, impacting the accuracy of the imputed values.

In the case of multiple imputation using the MICE R package, it was also notable that most of the comparative performance studies were limited to employing the default method (predictive mean matching) being provided, such as the studies by Milo et al. (2019b), Norazizi and Deni (2019a), de Carvalho et al. (2017), Addi et al. (2022), and Tasho and Zeqo (2022). Other works employed one of the selected MICE methods as part of their research work, but no comprehensive assessment was made to compare the different methods provided by the packages, such as the studies by Dewan et al. (2022), Tefera et al. (2023), Zvarevashe et al. (2019), and Worku et al. (2019). To ensure a fair and more comprehensive comparative assessment of the different

imputation approaches (single imputation, multiple imputation, machine learning-based imputation, etc.), the top-performing methods for each approach should initially be determined. Therefore, this study's aim is to evaluate different multiple imputation methods available within the MICE R package and ascertain how well they work under high rainfall variability. To do so, this research evaluates the effectiveness of multiple imputation techniques, employing daily rainfall stations with extensive records across JRB, to fill missing data (MCR, MAR, and MNAR) at varying levels (5%, 10%, 20%, and 30%) in the highly variable tropical climate.

In addition, the study introduced methodological novelty through a multi-step approach involving an initial step of performance assessment of the imputation methods, including mean absolute error (MAE), root mean square error (RMSE), normalized root mean square error (NRMSE), Nash-Sutcliffe efficiency (NSE), modified degree of agreement (MD), coefficient of determination (R2), Kling-Gupta efficiency (KGE), and volumetric efficiency (VE). This is followed by the ranking of the imputation methods using CPI to aggregate the ranked performance based on the statistical metrics used in determining the top-performing methods. Previous works have relied on the subjective evaluation of the selected metrics in determining the top-performing imputation methods (Addi et al., 2022; Balcha et al., 2023; Norazizi and Deni, 2019b). Subjectivity may introduce bias into the decision-making process, as individual preferences may influence the selection of metrics or the weighing of their importance. Additionally, it may lack transparency and reproducibility, making it challenging for others to understand or replicate the decision-making process. The omission of some metrics or the unequal weighting of others can lead to an incomplete or biased assessment of imputation methods. Furthermore, subjective evaluation does not account for the complex interplay between various performance criteria, often resulting in suboptimal or inefficient decisions. In this study, by providing a systematic and objective approach to decision-making, the CPI reduces subjectivity and bias, making it a valuable tool for making informed decisions when selecting the top-performing method. In order to examine the performance between the actual and imputed datasets, data completeness assessment using the best-performed method was carried out based on Probability Density Function (PDF) evaluation. In terms of novelty, the study's findings aid in identifying the optimal strategy for reconstructing complete rainfall datasets by imputing missing data under basin-scale high rainfall variability, with potential applications of the imputation methodological procedure in other river basins to enhance rainfall estimation and forecasting accuracy for datasets with similar characteristics.

2. Study area

The Johor River, shown in Fig. 1, originates at Mount Gemuruh in Malaysia's Johor State and flows south, covering a distance of about 122.7 km and draining an area of about 2636 km^2, before curving southwest and finally discharge into the Strait of Johor. The JRB plays a crucial role in providing water to Johor and Singapore, supporting the state's growth by supplying vital water resources for domestic, industrial, and agricultural purposes. Frequent flood events in the JRB have caused extensive infrastructure damage, economic disruptions, loss of lives, and environmental degradation, prompting numerous hydrological studies, especially on rainfall patterns and variability (Pak et al., 2021; Saudi et al., 2015; Tan et al., 2014, 2015).

According to Peel et al. (2007), the JRB and most of PM fall under the Köppen and Geiger climatic classification of Tropical Wet (*Af*), characterized by relatively uniform temperatures, high humidity, and regular rainfall, while the JRB's extensive spatial coverage and complex terrain contribute to significant spatial-scale variability in rainfall. The average annual rainfall in the region stands at 2340 mm, yet historical rainfall patterns exhibit considerable variability due to large-scale climate

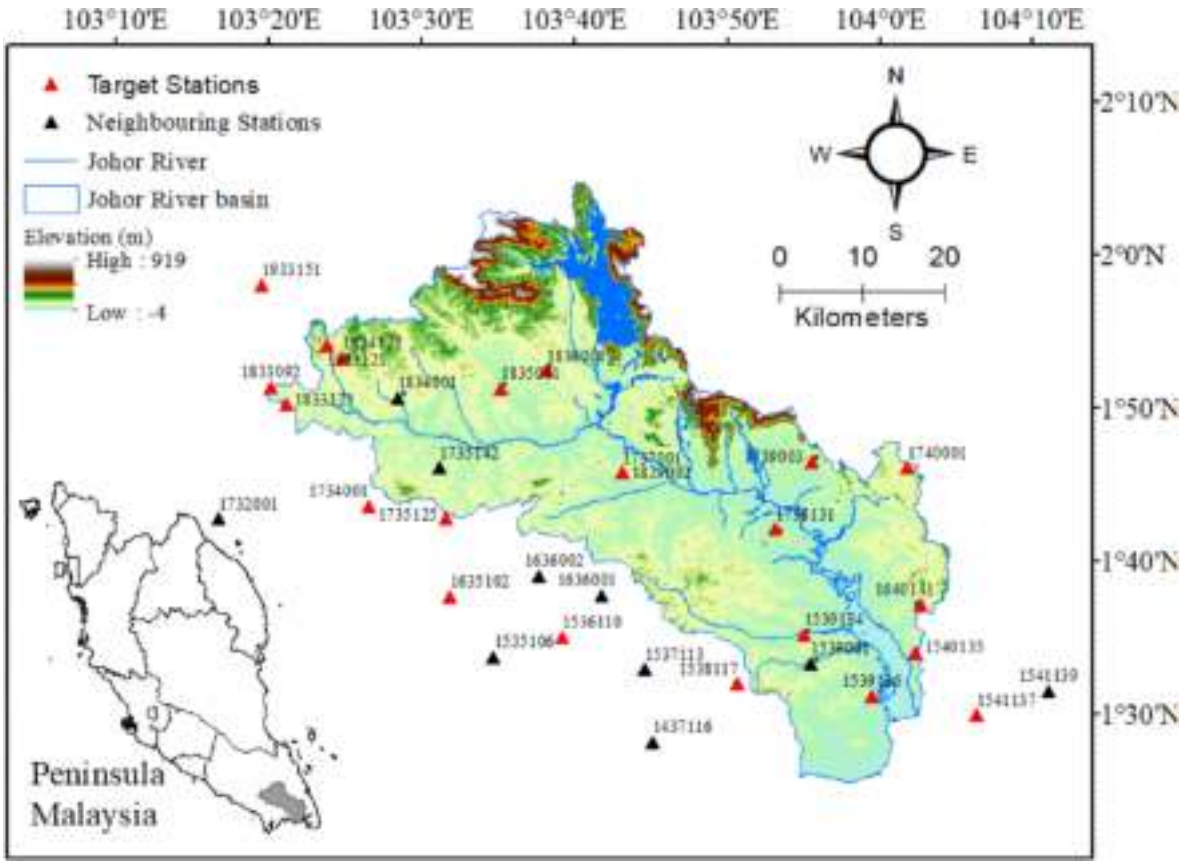

Fig. 1. The location of 32 rainfall stations (including 21 target stations) for the period of 1970–2015 in JRB used for this study.

events such as La Niña (3104 mm in 1995) and El Niño (1826 mm in 2015). The North East and South West monsoons (NEM and SWM) have a considerable impact on the seasonal patterns of weather in JRB (Wong et al., 2009). Therefore, when working with rainfall datasets in JRB, it is important to take seasonal factors into consideration and maintain the time series' original structure, conserving all relevant data, including extreme rainfall, to enable proper analysis (Burhanuddin et al., 2021).

3. Data and sources

As recommended by the World Meteorological Organization, at least 30 years of long-term climate data are recommended for climate assessment (Arguez and Vose, 2011). The daily rainfall was data provided by the Department of Irrigation and Drainage Malaysia (DID), which was obtained from rainfall stations situated within and nearby JRB for the 46-year period between 1970 and 2015. However, the northern mountainous region lacked available rainfall datasets, and certain stations were excluded from the analysis due to insufficient long-term data as the station's installation and operational dates only commerce in the early 2000s, inconsistent start and end years, and persistently missing data for extended periods. Only 21 out of the 32 identified rainfall stations (Fig. 1) were selected for the imputation process due to having a maximum of 20% missing data, with neighboring stations within a 20 km radius considered appropriate for PM regions based on a moderate effect size (Kamaruzaman et al., 2017). A 20 km radius of moderate effect size is considered suitable for selecting neighboring stations due to the prevalence of convective weather events within a $\leq$10 km scale (Suhaila et al., 2008), providing a balance between having a sufficient number of stations for accurate estimation results and avoiding increased computation time associated with a larger radius (Erdal and Karakurt, 2013). For each target station, at least 3 neighboring stations were available for the imputation process. The target and neighboring stations' locations are shown in Fig. 1, while Table 1 provides descriptions of the stations.

4. Methods

4.1. Procedure

The manuscript outlines the following procedure for the study.

1. The data quality assessment for target stations was initially performed, including missing data percentage, mean, maximum, variance, standard deviation, coefficient of variation, skewness, percentage of zero rainfall, and outlier ratio, while the Welch Two-Sample *t*-test was utilized to detect homogeneity or inhomogeneity in the daily rainfall data series. To further confirm if the inhomogeneous trend was caused by natural climate variability, the station with the irregular pattern was compared with nearby stations.
2. Then artificial missingness (MCAR, MAR, and MNAR) was introduced in increasing increments of 5%, 10%, 20%, and 30% to assess the effectiveness of each imputation method while retaining the original missing data.
3. Selected imputation methods from the MICE package were applied to the target stations with additional artificial missingness, and the imputation quality was evaluated by comparing the imputed data to the observed data using statistical assessments, namely, MAE, RMSE, NRMSE, NSE, MD, R2, KGE, and VE.
4. The statistical analysis from step 3 for each of the 21 target stations was ranked based on CPI to determine the highest-ranked imputation method. The details of the methodology for CPI can be found in Muhammad et al. (2019).
5. The imputed dataset's ability to capture extreme values under different missingness types and levels was then assessed by constructing the PDF for the best-performing method.

4.2. Missing data generation

Missing data was generated by assuming three distinct mechanisms, namely, MCAR, MAR, and MNAR, and the missingness level was generated up to 5%, 10%, 20%, and 30% without eliminating the original set of missing data (Wissler et al., 2022). A detailed description of the type of missingness can be found in Salgado et al. (2016). The number of stations available for comparison varied at each missingness level, resulting in 12, 14, 21, and 21 stations, respectively, as listed in Table 2. Generating missing data based on these three mechanisms allows a comprehensive assessment of multiple imputation methods for dealing with various types of missing data, demonstrating their reliability (Tong et al., 2020).

4.3. Imputation methods

The MICE R package has been successfully used in the past to fill in

Table 1
Data for the selected 21 rainfall target stations situated within and near the JRB from 1970 to 2015. NA (%) denotes the percentage of missing data, whereas Neighbr. St. denotes the number of nearbouring stations.

No	Target St.	St. Name	Lat	Lon	Neighbr. St
1	1933151	Ldg. Lambak	1.97	103.33	4
2	1933121	Ldg. Getah See Sun	1.90	103.4	5
3	1836001	Rancangan Ulu Seboi	1.88	103.64	5
4	1835001	Ldg. Pekan Layang Layang	1.86	103.59	6
5	1834122	Ldg. Rengam	1.89	103.42	5
6	1833123	Ldg. Benut	1.84	103.35	7
7	1833092	Ldg. Simpang Rengam	1.86	103.34	7
8	1740001	Felda Bkt. Wah Ha	1.77	104.03	3
9	1739003	Ldg. Permatang	1.78	103.93	4
10	1738131	Ldg. Getah Malaya	1.70	103.89	7
11	1737001	Sek. Men. Bkt. Besar	1.76	103.72	6
12	1735125	Ldg. Sedenak	1.71	103.53	4
13	1734001	Loji Pembersih Bkt. Batu	1.73	103.44	5
14	1640141	Felda Air tawar 1	1.62	104.04	8
15	1635102	Ldg. Kulai Young	1.63	103.53	7
16	1541137	Ldg. Sg. Papan	1.50	104.11	3
17	1540135	Ldg. Telok Sengat	1.57	104.04	6
18	1539136	Ldg. Lim Lim Bhd.	1.52	103.99	5
19	1539134	Ldg. Sg. Tiram	1.59	103.92	3
20	1538117	Ldg. Sg. Plentong	1.53	103.84	4
21	1536110	Ldg. Senai	1.58	103.65	5

Table 2
List of generated missing data (5%, 10%, 20%, 30%) for MCAR, MAR and MNAR, respectively used for comparison.

St. No.	NA (%)	NA (5%)	NA (10%)	NA (20%)	NA (30%)
1536110	16.1			✓	✓
1538117	0.7	✓	✓	✓	✓
1539134	1.6	✓	✓	✓	✓
1539136	0.6	✓	✓	✓	✓
1540135	0.5	✓	✓	✓	✓
1541137	4.9	✓	✓	✓	✓
1635102	11.7			✓	✓
1640141	9.9		✓	✓	✓
1734001	9.1		✓	✓	✓
1735125	3.6	✓	✓	✓	✓
1737001	14.2			✓	✓
1738131	0.2	✓	✓	✓	✓
1739003	13.1			✓	✓
1740001	19.2			✓	✓
1833092	1.5	✓	✓	✓	✓
1833123	1.3	✓	✓	✓	✓
1834122	1.8	✓	✓	✓	✓
1835001	17.2			✓	✓
1836001	18.2			✓	✓
1933121	1.9	✓	✓	✓	✓
1933151	2	✓	✓	✓	✓

the gaps in hydro-climatological data (Farzandi and Rezaee-Pazhand, 2021; Milo et al., 2019b; Norazizi and Deni, 2019a). This research utilized the MICE R package (van Buuren and Groothuis-Oudshoorn, 2011; White et al., 2011) for multiple imputations, which creates several imputations to handle missing data ambiguity, providing a successful and flexible approach for dealing with missing data in a multivariable setting. Numerous studies have demonstrated the effectiveness of MICE to impute missing rainfall data (Norazizi and Deni, 2019a; Poyatos et al., 2018). Table 3 lists the imputation methods used in this study using the MICE package. Various types of imputation methods, including numeric, binary, ordered, unordered, and any type of data, were selected to allow for a comprehensive evaluation of their effectiveness. Among the 23 imputation methods in the MICE package, four (*norm.boot*, *norm, norm. nob*, and *ri*) were excluded due to generating negative values, which are not valid for rainfall data.

4.4. Data completeness assessment

When utilising the best-performing approach to assess the accuracy of imputed rainfall data, evaluating data completeness is a crucial step because missing data might introduce bias and jeopardise the validity of the results. In this study, PDFs were generated for the actual and imputed data to facilitate comparison and assess the quality of the imputed data, ensuring reasonability and compatibility with the actual data distribution.

Table 3
List of the selected imputation methods from *mice* package. Generated Value of (+) and (−) means the numerical positive or negative outcome, respectively, produced by the chosen imputation method.

	Name	Type	Symbol	Generated Value
A	Bayesian linear regression	numeric	*norm*	–
B	Imputation of quadratic terms	numeric	*quadratic*	+
C	Level-1 normal heteroscedastic	numeric	*2l.norm*	+
D	Level-1 normal homoscedastic, lmer	numeric	*2l.lmer*	+
E	Level-1 normal homoscedastic, pan	numeric	*2l.pan*	+
F	Level-2 class mean	numeric	*2lonly. mean*	+
G	Level-2 class normal	numeric	*2lonly. norm*	+
H	Linear regression ignoring model error	numeric	*norm.nob*	–
I	Linear regression using bootstrap	numeric	*norm.boot*	–
J	Linear regression, predicted values	numeric	*norm. predict*	+
K	Random indicator for nonignorable data	numeric	*ri*	–
L	Unconditional mean imputation	numeric	*mean*	+
M	Level-1 logistic, glmer	binary	*2l.bin*	+
N	Logistic regression	binary	*logreg*	+
O	Logistic regression with bootstrap	binary	*logreg.boot*	+
P	Proportional odds model	ordered	*polr*	+
Q	Linear discriminant analysis	unordered	*lda*	+
R	Polytomous logistic regression	unordered	*polyreg*	+
S	Classification and regression trees	any	*cart*	+
T	Predictive mean matching	any	*pmm*	+
U	Random forest imputations	any	*rf*	+
V	Random sample from observed values	any	*sample*	+
W	Weighted predictive mean matching	any	*midastouch*	+

5. Results and discussion

5.1. Data quality assessment

Before imputation, data exploration was carried out on target stations, analyzing statistics including missing data parentage, mean, maximum values, standard deviation, variance, coefficient of variation, skewness, percentage of zero rainfall, and outlier ratio to identify potential data quality issues (Table 4). Figs. 2 and 3 show the target stations' missing data's histogram percentage, distribution, and number of intersections. The analysis showed missing values ranging from 0.2% to 19.2% for all stations. The highest daily rainfall ranged from 178 to 457.5 mm, and the mean daily rainfall varied from 5.61 to 8.00 mm, indicating the occurrence of extreme rainfall events, especially during the NEM. High variability in daily rainfall was observed, with variance, standard deviation, and coefficient of variation ranging from 12.5 to 18.3 mm^2, 156–334 mm, and 200–269%, respectively, typical of a tropical rainforest climate region, while the low percentage of zero rainfall ($\leq$0.68) suggests a high occurrence of rainy days across the basin.

The outlier ratio ranged from 8.3% to 19.3%, signifying a substantial proportion of outliers, and positive skewness (3.5–7.3) indicated a right-skewed distribution due to frequent extremely heavy downpours, especially around the NEM's peak. The Welch Two-Sample *t*-test evaluated homogeneity and identified potential shifting time series points at each station, with Table 2 showing the t-statistic and p-value results; 7 stations were found to be inhomogeneous ($p < 0.05$). The remaining stations, however, were revealed to be homogeneous. Sun et al. (2018) emphasize the importance of considering natural variability in accurately calculating long-term rainfall change, which may contribute to data inhomogeneity, as stated by Hyndman and Hyndman (2016). Consistent with Nashwan et al. (2019) findings of non-stationarity in rainfall intensities in stations over Johor's southern state, this study revealed 5 of the 7 inhomogeneous stations located in the basin's southern region, suggesting a geographic concentration that could imply higher vulnerability to large-scale climate events in this area. Suhaila and Yusop (2018) also noted that there were instances of breaks or discontinuities in temperature time series that could be caused by large-scale climate events such as El Niño and La Niña, while Che Ros et al. (2016) found that the El Niño-Southern Oscillation (ENSO) has a significant impact on the sudden increase in rainfall and long-term variability in the basin of the Kelantan River. A comparable pattern was observed in all inhomogeneous sites near the break point in the time series, with variable magnitudes of rainfall variability. It is important to note that the ENSO event and the time series' breakpoint were observed to coincided, indicating that inhomogeneity may have been cause by natural variability. The correlation bears similarities to Suhaila and Yusop (2018) previous research. Therefore, all stations in the study are suitable for further analysis without anthropogenic influence on the rainfall series.

5.2. Selection of imputation methods based on comparative assessment

Each imputation method's effectiveness was evaluated for each imputed station by initially examining it with the actual data using a variety of statistical performance metrics, namely, MAE, RMSE, NRMSE, NSE, MD, R2, KGE, and VE. By taking into account a variety of metrics, the study offers a complete evaluation of the efficacy of the imputation approach, encapsulating accuracy, precision, and goodness of fit, which enables a detailed understanding of method performance. Then, for each metric, an evaluation was made for every rainfall station and for each imputation method under different types of missingness (MCAR, MAR, and MNAR) and different levels of missingness (5%, 10%, 20%, and 30%). Therefore, there will be a large pool of output with 5% level of missingness (12 stations × 3 type of missingness × 19 imputation methods × 8 statistical metrics), 10% level of missingness (14 stations ×

Table 4
Data quality information for 21 target stations. Bold St. No., means inhomogeneous station based on the Welch Two-Sample *t*-test; NA means missing value; Std. Dev. means standard deviation; Var. means variance; CV means coefficient of variation.

St. No.	NA (%)	Mean	Max	Std. Dev.	Var.	CV	Skewness	Zero (%)	Outliers ratio	t-statistic	p-value
1536110	16.1	6.40	178	12.8	163	200	3.5	0.50	8.3	−6.1	1.5E-09
1538117	0.7	6.88	257	16.3	266	237	4.2	0.63	16.5	−3.3	8.2E-04
1539134	1.6	6.17	245.5	14.2	201	230	4.3	0.64	15.2	0.1	9.3E-01
1539136	0.6	5.88	274	15.8	250	269	5.2	0.68	19.3	−0.7	4.9E-01
1540135	0.5	6.51	282	15.4	236	237	4.9	0.62	14.2	−1.4	1.6E-01
1541137	4.9	6.62	375	16.6	277	251	6.5	0.61	14.8	−1.3	1.8E-01
1635102	11.7	8.00	230	16.7	278	209	3.6	0.54	10.8	−4.9	1.1E-06
1640141	9.9	6.56	244.5	14.3	203	218	4.4	0.54	12.0	−0.4	7.0E-01
1734001	9.1	6.39	180	13.6	185	213	3.7	0.50	12.8	−0.4	6.8E-01
1735125	3.6	6.54	250	14.3	204	218	4.5	0.55	13.1	−0.1	8.9E-01
1737001	14.2	5.61	290	13.7	188	244	5.4	0.46	12.8	0.9	3.4E-01
1738131	0.2	6.85	305	16.5	273	241	5.1	0.59	15.6	−2.0	4.2E-02
1739003	13.1	6.73	235	15.5	240	230	4.8	0.45	12.0	5.9	3.8E-09
1740001	19.2	7.09	457.5	18.3	334	258	7.3	0.50	11.4	−1.0	3.1E-01
1833092	1.5	5.97	270	13.3	176	223	4.2	0.58	14.3	−0.7	5.0E-01
1833123	1.3	5.88	210	13.1	171	223	4.0	0.61	15.5	−1.1	2.9E-01
1834122	1.8	5.67	307	12.5	156	220	4.3	0.59	14.7	0.4	6.9E-01
1835001	17.2	6.88	372	17.2	296	250	6.2	0.49	11.7	−3.7	1.9E-04
1836001	18.2	6.50	315	14.9	223	229	5.4	0.48	11.4	−0.9	3.7E-01
1933121	1.9	6.23	225	14.7	215	236	4.1	0.66	16.4	1.7	8.1E-02
1933151	2.0	5.74	308	14.1	200	245	5.7	0.58	13.9	−2.4	1.8E-02

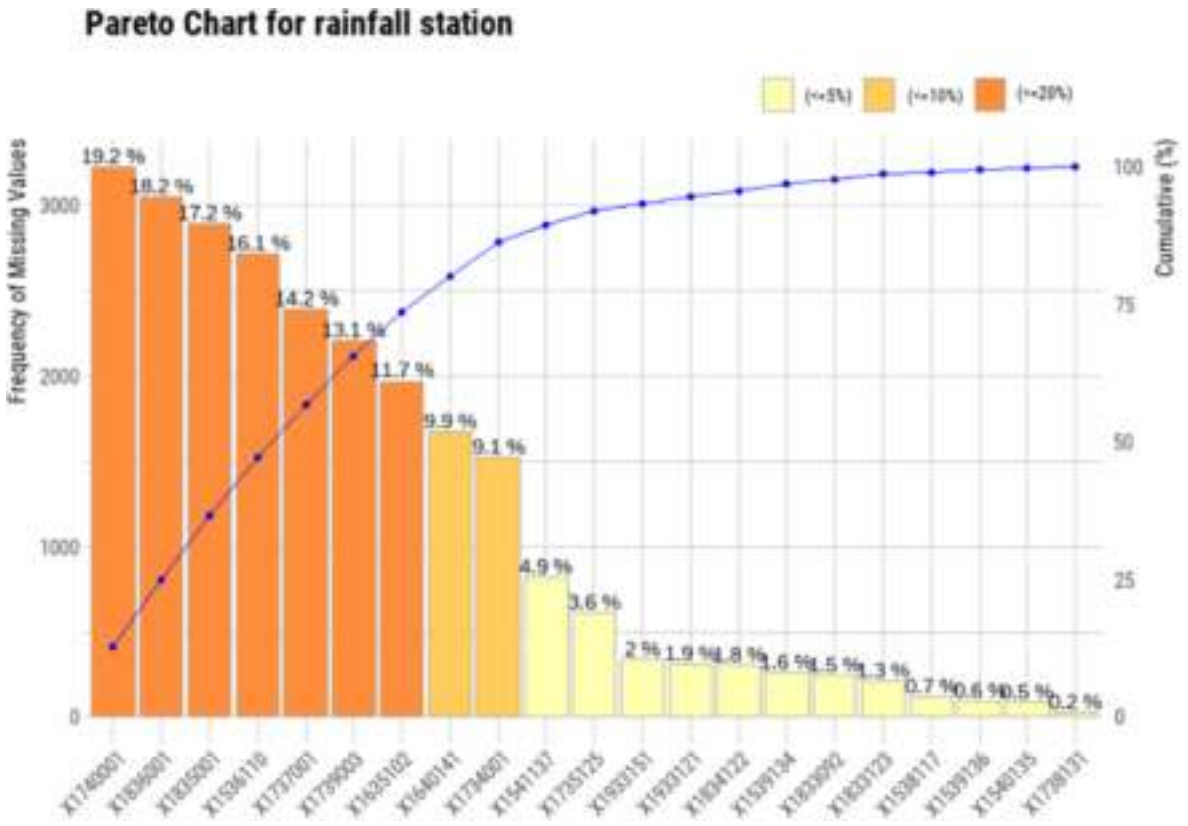

Fig. 2. Pareto chart of rainfall stations with the frequency, percentage, and cumulative information of missing data for 21 target stations from 1970 to 2015 in JRB.

3 type of missingness × 19 imputation methods × 8 statistical metrics), 20% level of missingness (21 stations × 3 type of missingness × 19 imputation methods × 8 statistical metrics), and 30% level of missingness (21 stations × 3 type of missingness × 19 imputation methods × 8 statistical metrics). The heat map, presented in Fig. 4, using MNAR with a 30% missingness level as an example, demonstrated how each method performed across the metrics. The analysis did not include the *quadratic* method because of its poor performance.

Table 5 shows the range of the performance of the employed rainfall station for each of the imputation methods based on the type of missingness under the highest 30% missingness level. The 30% missingness level was used to discuss the effectiveness of the imputation methods in handling higher degrees of missing data. Generally, the imputation methods perform well under MCAR, where data is missing randomly without any systematic pattern. This is followed by MAR, suggesting its effectiveness in imputing data when missingness is random. Conversely, MNAR shows the lowest performance, where the mechanism causing data to be missing depends on unobserved data, which poses greater

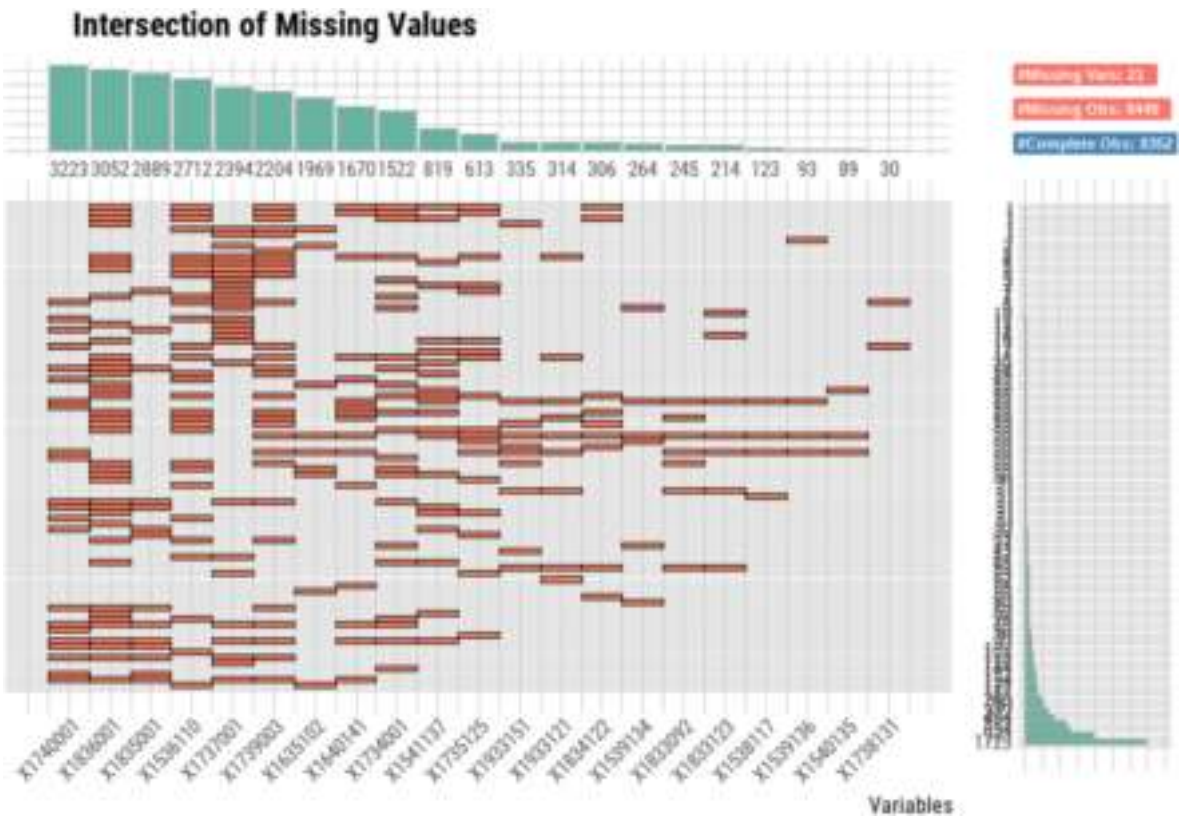

Fig. 3. Visualization of the distribution of missing data and the number of intersections of missing data for 21 target stations from 1970 to 2015 in JRB. The x-axis displays 21 selected stations, with bars atop the plot indicating the counts of missing values. The y-axis illustrates the combinations of rainfall stations and their corresponding frequencies, providing insight into the patterns of missing data from 1970 to 2015.

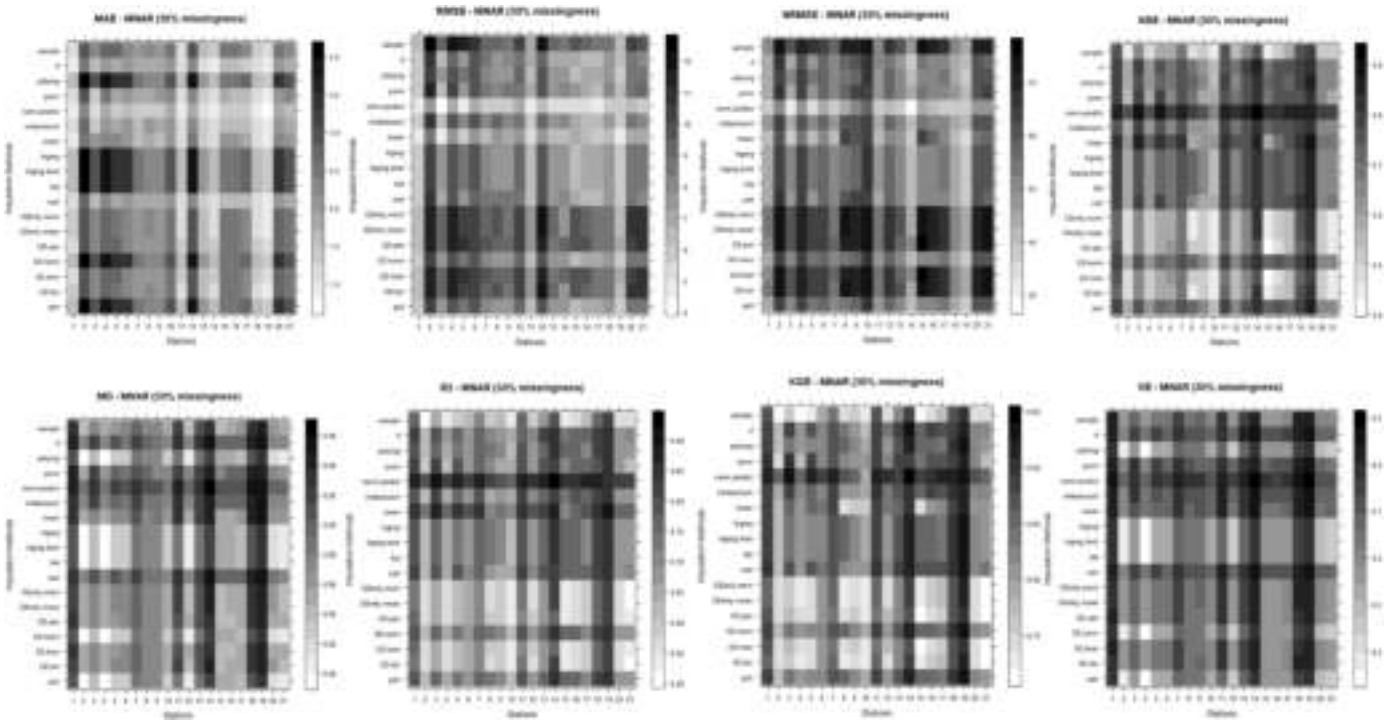

Fig. 4. A heat map of the statistical performance metrics of the imputation methods (MAE, RMSE, NRMSE, NSE, MD, R2, KGE, and VE) based on the example using the imputation of MNAR (30% missingness).

Table 5
Results for the range of the statistical performance metrics (MAE, RMSE, NRMSE, NSE, MD, R2, KGE, and VE) for each of the imputation methods based on MAR, MCAR, and MNAR under 30% missingness level. The alphabet corresponds to the imputation method given in Table 3.

Methods	Type	B	C	D	E	F	G	J	L	M	N
MAE	MAR	1.24–581.75	1.34–4.03	1.06–3.28	1.06–3.28	1.04–3.11	1.04–3.11	0.78–2.25	0.87–2.68	1.04–3.11	1.34–4.03
	MNAR	1.2–2702.71	1.35–3.98	1.05–2.97	1.05–2.97	1.01–3.06	1.01–3.06	0.81–2.17	0.87–2.57	1.01–3.06	1.35–3.98
	MCAR	1.69–332.61	1.65–3.41	1.24–3.41	1.24–3.41	1.3–3.42	1.3–3.42	0.87–2.43	1.06–2.98	1.30–3.42	1.65–4.41
RMSE	MAR	6.25–1160.71	5.67–10.52	6.50–12.18	6.50–12.18	6.30–11.98	6.30–11.98	4.54–7.69	4.75–8.72	6.30–11.98	5.67–10.52
	MNAR	6.52–5409.88	5.76–10.57	6.38–11.68	6.38–11.68	6.40–11.83	6.40–11.83	4.45–7.62	4.77–8.53	6.40–11.83	5.76–10.57
	MCAR	7.43–627.01	6.30–11.20	6.97–12.94	6.97–12.94	7.30–12.69	7.30–12.69	4.87–8.69	5.36–9.45	7.30–12.69	6.30–11.20
NRMSE	MAR	48.1–116.6	38.5–66.0	45.2–74.9	45.2–74.9	42.6–74.8	42.6–74.8	29.0–54.2	33.6–63.8	42.6–74.8	38.5–66.0
	MNAR	46.5–116.4	38.3–66.1	45.0–75.2	45.0–75.2	43.2–74.3	43.2–74.3	29.6–58.3	33.8–64.0	43.2–74.3	38.3–66.1
	MCAR	58.8–118.0	38.7–72.9	47.2–83.0	47.2–83.0	45.6–83.0	45.6–83.0	27.5–63.0	33.1–73.7	45.6–83.0	38.7–72.9
NSE	MAR	−0.36-0.77	0.57–0.85	0.44–0.80	0.44–0.80	0.44–0.82	0.44–0.82	0.71–0.92	0.59–0.89	0.44–0.82	0.57–0.85
	MNAR	−0.36-0.78	0.56–0.85	0.43–0.80	0.43–0.80	0.45–0.81	0.45–0.81	0.66–0.91	0.59–0.89	0.45–0.81	0.56–0.85
	MCAR	−0.39-0.65	0.47–0.85	0.31–0.78	0.31–0.78	0.31–0.79	0.31–0.79	0.6–0.92	0.46–0.89	0.31–0.79	0.47–0.85
MD	MAR	0.59–0.93	0.80–0.93	0.83–0.94	0.83–0.94	0.84–0.94	0.84–0.94	0.88–0.96	0.84–0.95	0.84–0.94	0.80–0.93
	MNAR	0.59–0.93	0.80–0.93	0.84–0.94	0.84–0.94	0.84–0.94	0.84–0.94	0.89–0.96	0.85–0.95	0.84–0.94	0.80–0.93
	MCAR	0.59–0.90	0.77–0.92	0.82–0.93	0.82–0.93	0.81–0.93	0.81–0.93	0.86–0.95	0.82–0.94	0.81–0.93	0.77–0.92
R2	MAR	0.00–0.78	0.62–0.86	0.52–0.81	0.52–0.81	0.52–0.83	0.52–0.83	0.77–0.92	0.71–0.9	0.52–0.83	0.62–0.86
	MNAR	0.00–0.79	0.63–0.86	0.53–0.81	0.53–0.81	0.52–0.83	0.52–0.83	0.75–0.92	0.71–0.9	0.52–0.83	0.63–0.86
	MCAR	0.00–0.69	0.56–0.85	0.44–0.79	0.44–0.79	0.45–0.80	0.45–0.80	0.71–0.93	0.65–0.9	0.45–0.80	0.56–0.85
KGE	MAR	−0.71-0.88	0.77–0.92	0.72–0.90	0.72–0.90	0.72–0.91	0.72–0.91	0.81–0.95	0.74–0.92	0.72–0.91	0.77–0.92
	MNAR	−1.76-0.89	0.78–0.93	0.72–0.90	0.72–0.90	0.72–0.91	0.72–0.91	0.79–0.94	0.75–0.92	0.72–0.91	0.78–0.93
	MCAR	−0.70-0.83	0.74–0.92	0.66–0.89	0.66–0.89	0.67–0.89	0.67–0.89	0.76–0.95	0.69–0.93	0.67–0.89	0.74–0.92
VE	MAR	0.01–0.82	0.37–0.79	0.51–0.84	0.51–0.84	0.53–0.84	0.53–0.84	0.68–0.88	0.6–0.87	0.53–0.84	0.37–0.79
	MNAR	0.00–3.67	0.36–0.79	0.56–0.84	0.56–0.84	0.56–0.84	0.56–0.84	0.69–0.88	0.62–0.87	0.56–0.84	0.36–0.79
	MCAR	0.02–0.74	0.28–0.76	0.46–0.82	0.46–0.82	0.46–0.82	0.46–0.82	0.62–0.88	0.54–0.85	0.46–0.82	0.28–0.76
		O	P	Q	R	S	T	U	V	W	
MAE	MAR	1.34–4.03	1.34–4.03	1.34–4.03	1.34–4.03	0.91–2.79	0.93–2.81	0.92–2.64	1.07–3.20	0.87–2.67	
	MNAR	1.35–3.98	1.35–3.98	1.35–3.98	1.35–3.98	0.86–2.64	0.99–3.05	0.93–2.57	1.09–3.07	0.90–2.64	
	MCAR	1.65–4.41	1.65–4.41	1.65–4.41	1.65–4.41	1.05–2.96	1.16–3.18	1.04–2.80	1.26–3.40	1.10–2.93	
RMSE	MAR	5.67–10.52	5.67–10.52	5.67–10.52	5.67–10.52	6.01–10.75	5.7–10.98	5.94–10.63	6.74–12.50	5.95–10.66	
	MNAR	5.76–10.57	5.76–10.57	5.76–10.57	5.76–10.57	5.67–10.74	5.69–11.39	5.63–10.72	6.37–12.29	6.02–10.71	
	MCAR	6.30–11.20	6.30–11.20	6.30–11.20	6.30–11.20	6.07–11.39	6.59–11.32	6.02–11.25	7.45–12.69	6.31–11.74	
NRMSE	MAR	38.5–66.0	38.5–66.0	38.5–66.0	38.5–66.0	40.2–64.3	38.0–67.0	37.7–65.3	45.4–76.3	38.8–66.9	
	MNAR	38.3–66.1	38.3–66.1	38.3–66.1	38.3–66.1	37.7–65.5	40.5–67.7	37.7–66.1	44.6–75.0	40.3–65.0	
	MCAR	38.7–72.9	38.7–72.9	38.7–72.9	38.7–72.9	37.5–73.5	38.7–73.5	39.2–74.3	47.9–81.6	36.7–72.7	
NSE	MAR	0.57–0.85	0.57–0.85	0.57–0.85	0.57–0.85	0.59–0.84	0.55–0.86	0.57–0.86	0.42–0.79	0.55–0.85	
	MNAR	0.56–0.85	0.56–0.85	0.56–0.85	0.56–0.85	0.57–0.86	0.54–0.84	0.56–0.86	0.44–0.80	0.58–0.84	
	MCAR	0.47–0.85	0.47–0.85	0.47–0.85	0.47–0.85	0.46–0.86	0.46–0.85	0.45–0.85	0.33–0.77	0.47–0.87	
MD	MAR	0.80–0.93	0.80–0.93	0.80–0.93	0.80–0.93	0.86–0.95	0.86–0.95	0.86–0.95	0.83–0.94	0.86–0.95	
	MNAR	0.80–0.93	0.80–0.93	0.80–0.93	0.80–0.93	0.87–0.95	0.85–0.94	0.87–0.95	0.84–0.94	0.87–0.95	
	MCAR	0.77–0.92	0.77–0.92	0.77–0.92	0.77–0.92	0.85–0.95	0.84–0.94	0.85–0.95	0.82–0.93	0.85–0.94	
R2	MAR	0.62–0.86	0.62–0.86	0.62–0.86	0.62–0.86	0.62–0.84	0.6–0.86	0.62–0.86	0.50–0.8	0.61–0.86	
	MNAR	0.63–0.86	0.63–0.86	0.63–0.86	0.63–0.86	0.62–0.86	0.58–0.84	0.61–0.86	0.52–0.81	0.62–0.84	
	MCAR	0.56–0.85	0.56–0.85	0.56–0.85	0.56–0.85	0.54–0.80	0.54–0.86	0.54–0.85	0.47–0.78	0.57–0.87	
KGE	MAR	0.77–0.92	0.77–0.92	0.77–0.92	0.77–0.92	0.79–0.91	0.77–0.93	0.78–0.93	0.71–0.90	0.78–0.92	
	MNAR	0.78–0.93	0.78–0.93	0.78–0.93	0.78–0.93	0.78–0.93	0.75–0.91	0.78–0.93	0.72–0.90	0.79–0.92	
	MCAR	0.74–0.92	0.74–0.92	0.74–0.92	0.74–0.92	0.74–0.93	0.74–0.92	0.73–0.92	0.68–0.88	0.74–0.93	
VE	MAR	0.37–0.79	0.37–0.79	0.37–0.79	0.37–0.79	0.60–0.86	0.59–0.86	0.60–0.86	0.52–0.83	0.60–0.86	
	MNAR	0.36–0.79	0.36–0.79	0.36–0.79	0.36–0.79	0.62–0.87	0.59–0.85	0.62–0.86	0.55–0.83	0.62–0.86	
	MCAR	0.28–0.76	0.28–0.76	0.28–0.76	0.28–0.76	0.54–0.85	0.53–0.84	0.54–0.85	0.47–0.81	0.55–0.85	

challenges for imputation methods. Based on MAE, under MAR and MNAR, methods like *norm. predict*, *rf*, *pmm*, and *midastouch* maintain lower MAE values, ranging between 0.78 and 2.25, 0.92–2.64, 0.93–2.81, 0.87–2.67, and 0.81–2.17, 0.93–2.57, 0.99–3.05, 0.9–2.64, respectively. This indicates their reliability in imputing missing data when it follows a random pattern. Conversely, under MCAR, where data is missing randomly without any systematic pattern, *norm. predict* (0.87–2.43) stands out with consistently lower MAE values, highlighting the advantage of imputing data when it is missing randomly.

Under MAR and MCAR, most imputation methods show consistent RMSE values. The range of RMSE values across methods is relatively narrow, indicating stable performance. Notably, the *norm. predict* method exhibits lower RMSE values of 4.54–7.69, and 4.87–8.69 under MAR and MCAR, respectively, suggesting its effectiveness in imputing data when missingness is random. Under MNAR, *norm. predict* (4.45–7.62) and *mean* (4.77–8.53) demonstrate relatively lower RMSE values. A similar performance was observed under NRMSE with the *norm. predict* method, which exhibits lower NRMSE values of 29–54.2, and 27.5–63 under MAR and MCAR, respectively. Under MNAR, *norm. predict* (29.6–58.3) and *mean* (33.8–64) demonstrate relatively lower NRMSE values.

Under all types of missingness, most imputation methods exhibit a consistent range of values, indicating reliable performance. Notably, the *norm. predict* method consistently shows higher NSE values of 0.71–0.92, 0.6–0.92, and 0.66–0.91, under MAR, MCAR, and MNAR, respectively. Similarly, the *norm. predict* method consistently shows the highest MD values of 0.88–0.96, 0.86–0.95, and 0.89–0.96, under MAR, MCAR, and MNAR, respectively. Most imputation methods show a relatively high range of R2 values, indicating a high degree of explanatory power. The *norm. predict* method consistently shows the highest R2 values of 0.77–0.92, 0.71–0.93, and 0.75–0.92, under MAR, MCAR, and MNAR, respectively. Based on KGE, most imputation methods demonstrate a good level of efficiency. The *norm. predict* method consistently outperforms other methods, showing the highest KGE values of 0.81–0.95, 0.76–0.95, and 0.79–0.94 under MAR, MCAR, and MNAR, respectively suggesting its effectiveness in capturing the statistical characteristics of the observed data in the presence of random missingness. Based on VE, the *norm. predict* method stands out as the most effective method,

consistently achieving higher values of 0.68–0.88, 0.62–0.88, and 0.69–0.88 under MAR, MCAR, and MNAR, respectively. This indicates that this method can replicate the volumetric properties of the observed data even when dealing with randomly missing values.

In comparison to the *norm. predict* method, the *mean*, *rf*, and *cart* methods also consistently display competitive performance across various metrics and types of missingness. These methods exhibit competitive performance in terms of MAE, RMSE, NRMSE, NSE, MD, R2, KGE, and VE, making them one of the higher-performing methods across the board. On the other hand, for MAR and MNAR, in many cases, the *quadratic* method showed lower performance across multiple metrics. The *sample* method also tends to have lower performance, with lower scores in multiple metrics. For MCAR, the *quadratic* and *sample* methods continue to have lower scores in various metrics, whereas the *polyreg* and *logreg. boot* also tend to have lower scores.

For objective evaluation and to cater for the large output from the statistical metrics, the CPI was used as a way to determine a compromise solution or ranking that optimally balances trade-offs between the different performances of the imputation methods under different metrics, different missingness types, and different levels of missingness. By incorporating CPI as a decision-supportive tool, the trade-offs between several statistical performance metrics are taken into account through the use of aggregated ranking. Therefore, CPI enhances the quality of decision in selecting the top-performing imputation methods, which was lacking in previous work that relied on subjective evaluation of each individual statistical metric (Fakhruddin Kamaruzaman et al., 2017; Jahan et al., 2019). Table 6 presents the final rank for each imputation method, obtained by re-aggregation using CPI in the performance study under various missingness types and levels, with *norm. predict* consistently ranking first, *mean* method second, and *quadratic* approach consistently last. The ranking of methods third and below varied based on missingness type and level, but overall, a consistent ranking pattern was observed, indicating that different imputation methods had a stable position regardless of missingness level. Given that *norm. predict* consistently ranked the highest, this method was selected for subsequent analysis.

5.3. Validation of the imputation method

5.3.1. Performance analysis

Figs. 5–8 illustrate a statistical performance analysis comparing *norm. predict* ability to impute missing values under various types (MAR, MNAR, and MCAR) and levels of missingness (5%, 10%, 20%, 30%). The *norm. predict* performance decreased with higher levels of missingness, as evident from comparing the percentage of change between the 10% and 30% imputed datasets to the 5% imputed dataset, where a higher percentage of error was observed with increased missingness.

Under increasing levels of missingness (5%–30%), the MAE showed a slight increase, ranging from 48 to 98%, 49–98%, and 49–99% for MAR, MNAR, and MCAR, respectively, indicating decreased imputation accuracy with larger errors in predictions, possibly due to the method's limited ability to capture underlying patterns or relationships in rainfall data. Similar percentage increases in error were observed for RMSE and NRMSE, ranging from 19 to 83% and 20–85% for RMSE, and 26–88% and 27–89% for NRMSE, respectively, under MAR, MNAR, and MCAR, suggesting higher deviations between imputed and actual values and decreased imputation accuracy with higher errors, indicating poorer performance of the imputation method. The imputation method showed lower increments in error for NSE (2–37%, 1–45%, and 3–55%), MD (1–11%, 1–11%, and 2–14%), R2 (2–24%, 1–32%, and 3–28%), KGE (0–20%, 1–23%, and 2–25%), and VE (3–40%, 3–39%, and 4–52%) under MAR, MNAR, and MCAR, respectively, indicating its effectiveness in reproducing observed data and achieving strong concurrence and similarity between imputed and actual values, as well as a good fit of the regression model. Additionally, in terms of KGE and VE, the imputation method accurately reproduces the imputed data in terms of both mean and variability, as well as the correlation between actual and imputed values, indicating reliable and comparable imputed values to the actual values.

Increased missingness in rainfall data reduces the imputation method's performance due to information loss and reduced sample size, leading to fewer accurate estimates and broader confidence ranges, potentially impacting the reliability and accuracy of the findings. There is a greater loss of information and potential underperformance of the imputation method as the percentage of missing data rises from 10% to 30%, resulting in more missing data points. Additionally, the potential for bias estimation may increase, which reduces the accuracy and reliability of the subsequent analysis. Higher missingness introduces more variability into the time series, leading to increased uncertainty in the calculation of imputed values and decreased reliability and performance of the imputation method.

5.3.2. Probability Density Function (PDF)

Fig. 9 illustrates the performance of imputed stations (upstream, 1538117; middle, 1738131; and downstream, 1933151) compared with actual data based on monthly rainfall PDF under varying levels of

Table 6
Rank of the statistical performance assessment for each imputation method for each type of missingness, namely, MCAR, MAR, and MNAR, under 5%, 10%, 20%, and 30%, respectively, based on CPI. The alphabet corresponds to the imputation method given in Table 3.

	MCAR				MAR				MNAR				Final
Rank	5%	10%	20%	30%	5%	10%	20%	30%	5%	10%	20%	30%	
1	J	J	J	J	J	J	J	J	J	J	J	J	J
2	L	L	L	L	L	L	L	L	L	L	L	L	L
3	T	U	U	S	U	T	U	S	U	U	T	U	U
4	U	W	S	U	S	W	S	U	P	P	U	S	S
5	S	T	W	W	T	U	W	W	C	C	S	W	W
6	W	S	T	T	P	S	T	T	Q	Q	W	T	T
7	M	M	P	P	C	P	P	P	O	O	P	P	D
8	V	D	C	C	Q	C	C	C	N	N	C	C	M
9	D	V	Q	Q	O	Q	Q	Q	R	R	Q	Q	V
10	P	P	O	O	N	O	O	O	S	T	O	O	P
11	C	C	N	N	R	N	N	N	T	S	N	N	C
12	Q	Q	R	R	W	R	R	R	W	W	R	R	Q
13	O	O	D	D	D	V	D	M	D	V	M	M	O
14	N	N	E	E	E	M	E	F	E	D	F	F	N
15	R	R	V	M	M	F	V	G	V	E	G	G	R
16	E	E	M	F	F	G	M	D	M	M	V	D	E
17	F	F	F	G	G	D	F	E	F	F	D	E	F
18	G	G	G	V	V	E	G	V	G	G	E	V	G
19	B	B	B	B	B	B	B	B	B	B	B	B	B

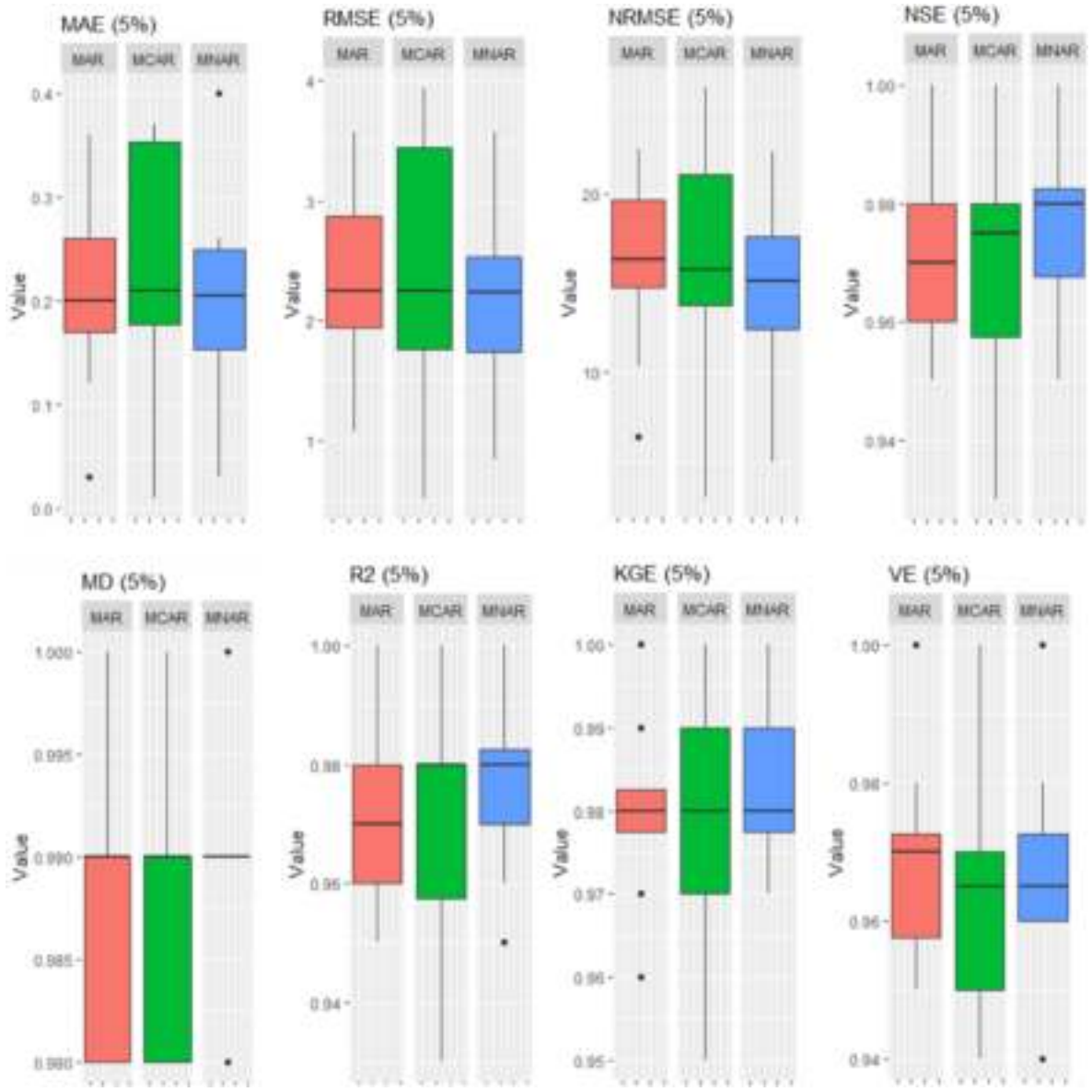

Fig. 5. Results for the statistical performance of the *norm. predict* imputation methods based on MAR, MCAR, and MNAR under 5% missingness based on MAE, RMSE, NRMSE, NSE, MD, R^2, KGE, and VE.

missingness. The PDFs demonstrated that even with up to 30% missingness, the rainfall characteristics and the shape of the distribution were retained, as evidenced by improved symmetry and bell-shaped curves in the imputed datasets compared to the actual data (Kamaruzaman et al., 2017). The similarity in central tendency (median) and data distribution between actual and imputed datasets confirms the successful estimation of missing data and the accurate representation of data variability by the imputation method. The imputed datasets displayed higher kurtosis, indicating increased variability and extreme values in the imputed rainfall compared to the actual data with missing values, highlighting the importance of obtaining a reliable estimate of extreme rainfall events. Additionally, a trend of higher peaks with an increasing level of missingness was observed in most cases, indicating that the imputation method tends to overestimate the probability of high extreme values as the level of missingness increases in the actual data.

6. Conclusions

Numerous studies have focused on finding the best method for filling in missing rainfall data, and in this study, 19 methods in the MICE R package were used to identify the most suitable methods for imputation of daily rainfall in JRB. Across different missingness types and levels (up to 30%), the *norm. predict* method outperformed others, making it the most appropriate choice for this dataset. The *mean*, *rf*, and *cart* methods are another one that seems to work rather well for different types of missingness. These methods are among the best overall since they perform competitively in terms of MAE, RMSE, NRMSE, NSE, MD, R2, KGE, and VE. In particular, the application of the CPI has been instrumental in establishing a well-balanced compromise ranking among imputation methods, effectively addressing the trade-offs inherent in their performance across diverse statistical metrics. The incorporation of CPI as a decision-support tool has enabled an objective and holistic consideration of multiple performance metrics, marking a significant departure from prior research that relied on subjective evaluation based on individual metrics.

Nonetheless, the *norm. predict* method comes with certain limitations that need to be considered, such as its reliance on the assumption of normality, which may not hold in many real-world datasets. Deviations

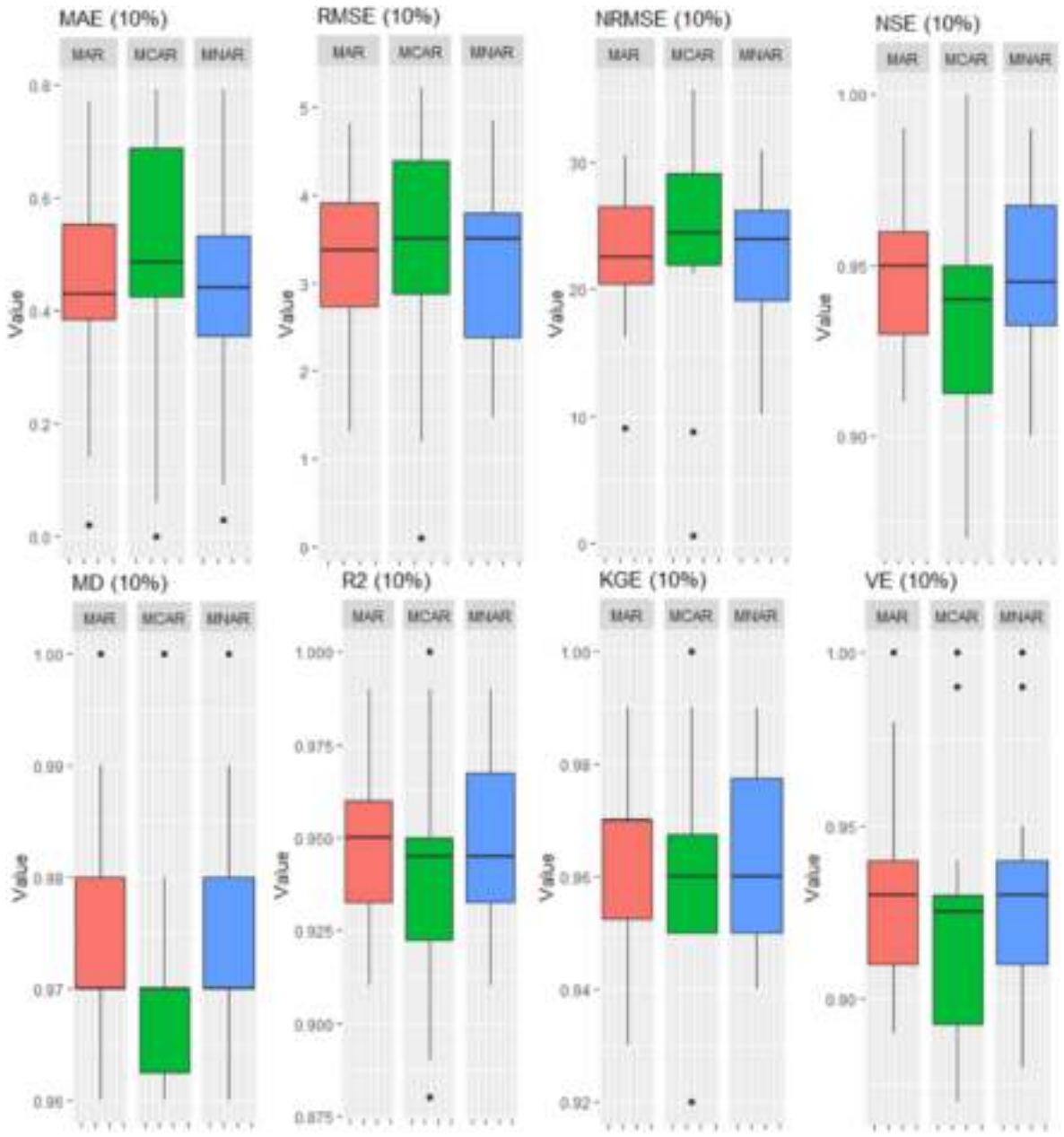

Fig. 6. Results for the statistical performance of the *norm. predict* imputation methods based on MAR, MCAR, and MNAR under 10% missingness based on MAE, RMSE, NRMSE, NSE, MD, R^2, KGE, and VE.

from normality can result in imputed values that do not accurately reflect the underlying data distribution. Additionally, it can be limited by the subjective selection of the neighboring stations, which may introduce bias and affect the uncertainty of imputed values and accuracy. This is due to individual judgement and assumptions in selecting the neighboring stations, which may lead to inconsistent results between studies and hamper the ability to replicate findings. It is also sensitive to the missing data mechanism, performing better in situations where data is MAR or MCAR but less so when data is MNAR. On the other hand, even though the CPI-based ranking methodology has proven to be valuable in this analysis, the choice of performance metrics to be included in the assessment and their relative weights can still introduce subjectivity into the decision-making process. Future research could explore ways to standardize this aspect further. The study also focused specifically on daily rainfall imputation in the JRB region, which may vary in other geographic areas or for different types of climate data. While this study focused on imputation at the daily level, temporal resolutions (e.g., hourly or monthly) and spatial resolutions (e.g., regional or local) may necessitate different imputation strategies. In conclusion, while this research offers valuable insights into missing rainfall data imputation and introduces a robust ranking methodology, it is essential to consider these limitations.

Therefore, it is crucial for researchers to establish objective and standardized criteria for station selection, such as geographic proximity or similarity in climatic conditions, which can minimize the impact of subjective influences. Other approaches for objective selection of the imputation methods under multiple statistical performance metrics can be explore, such as multi-criteria decision making (Dayal et al., 2023), cluster analysis (Zhang et al., 2016) and metric weighting (Chhin and Yoden, 2018). In addition, future comparative assessments between multiple imputation methods and machine learning-based imputation, or other methodologies, should be conducted to examine how each approach performs under high rainfall variability in the tropics. This investigation on missing rainfall data in JRB highlights the importance

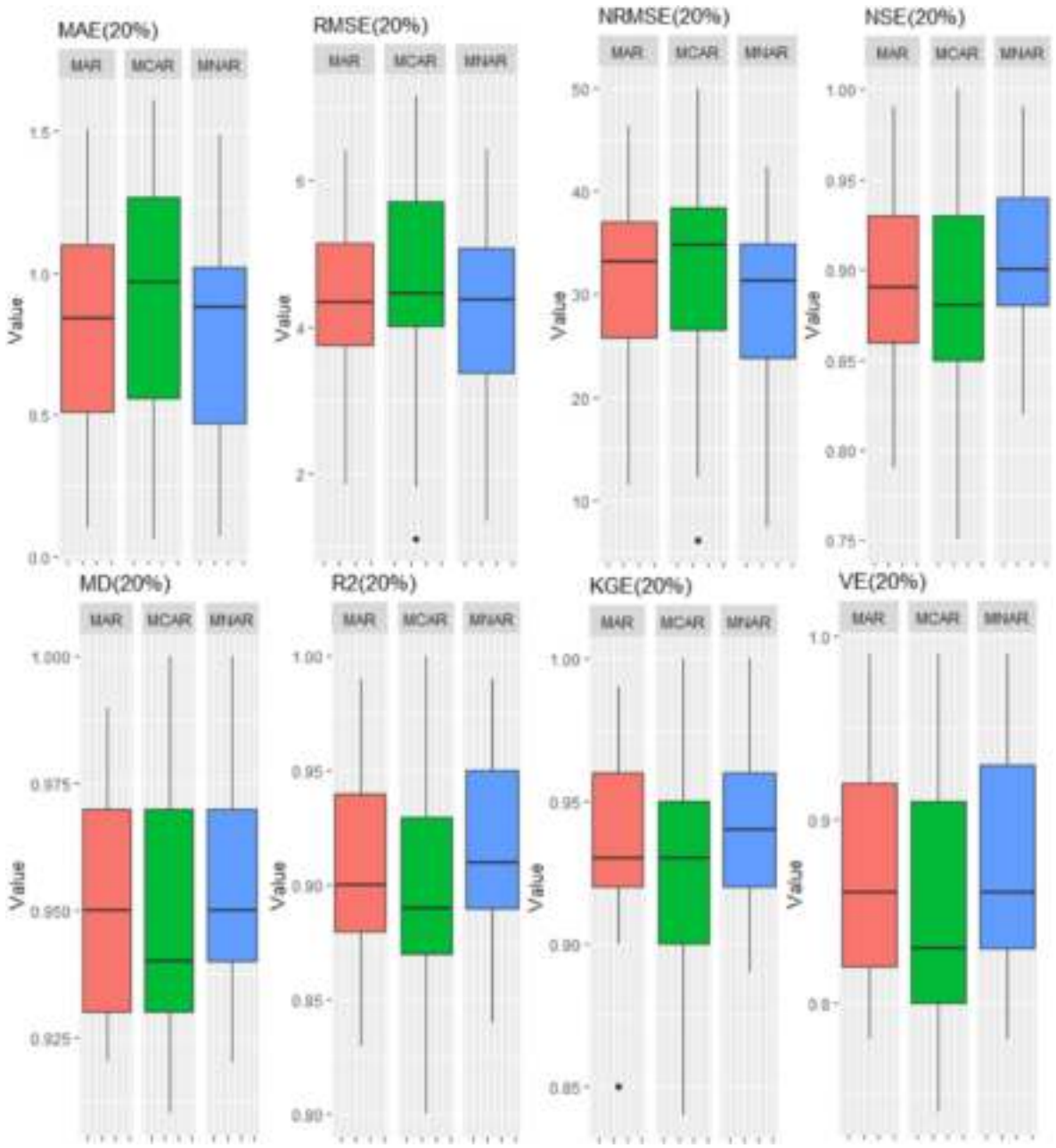

Fig. 7. Results for the statistical performance of the *norm. predict* imputation methods based on MAR, MCAR, and MNAR under 20% missingness based on MAE, RMSE, NRMSE, NSE, MD, R^2, KGE, and VE.

of choosing appropriate imputation methods that align with the area's significant seasonal and spatial rainfall variability. This study highlights the significance of handling missing data to ensure accurate trend analysis, capturing trend direction and amplitude while accounting for high variability patterns absent in actual data, guiding the selection of appropriate imputation techniques for rainfall datasets of similar size, and enhancing rainfall estimation and forecast precision in JRB and other tropical basins in PM and Southeast Asia.

Declaration

Ethical approval

Not applicable.

Consent to participate

Not applicable.

Consent to publish

We, (Zulfaqar Sa'adi Zulkifli Yusop, Nor Eliza Alias, Ming Fai Chow, Mohd Khairul Idlan Muhammad, Muhammad Wafiy Adli Ramli, Zafar Iqbal, Mohammed Sanusi Shiru, Faizal Immaddudin Wira Rohmat, Nur Athirah Mohamad, Mohamad Faizal Ahmad) hereby declare that We participated in the study in the development of the manuscript titled (Evaluating Imputation Methods for Spatiotemporal Rainfall Data Under High Variability in Johor River Basin, Malaysia). We have read the final version and give our consent for the article to be published in the journal of Applied Computing and Geosciences.

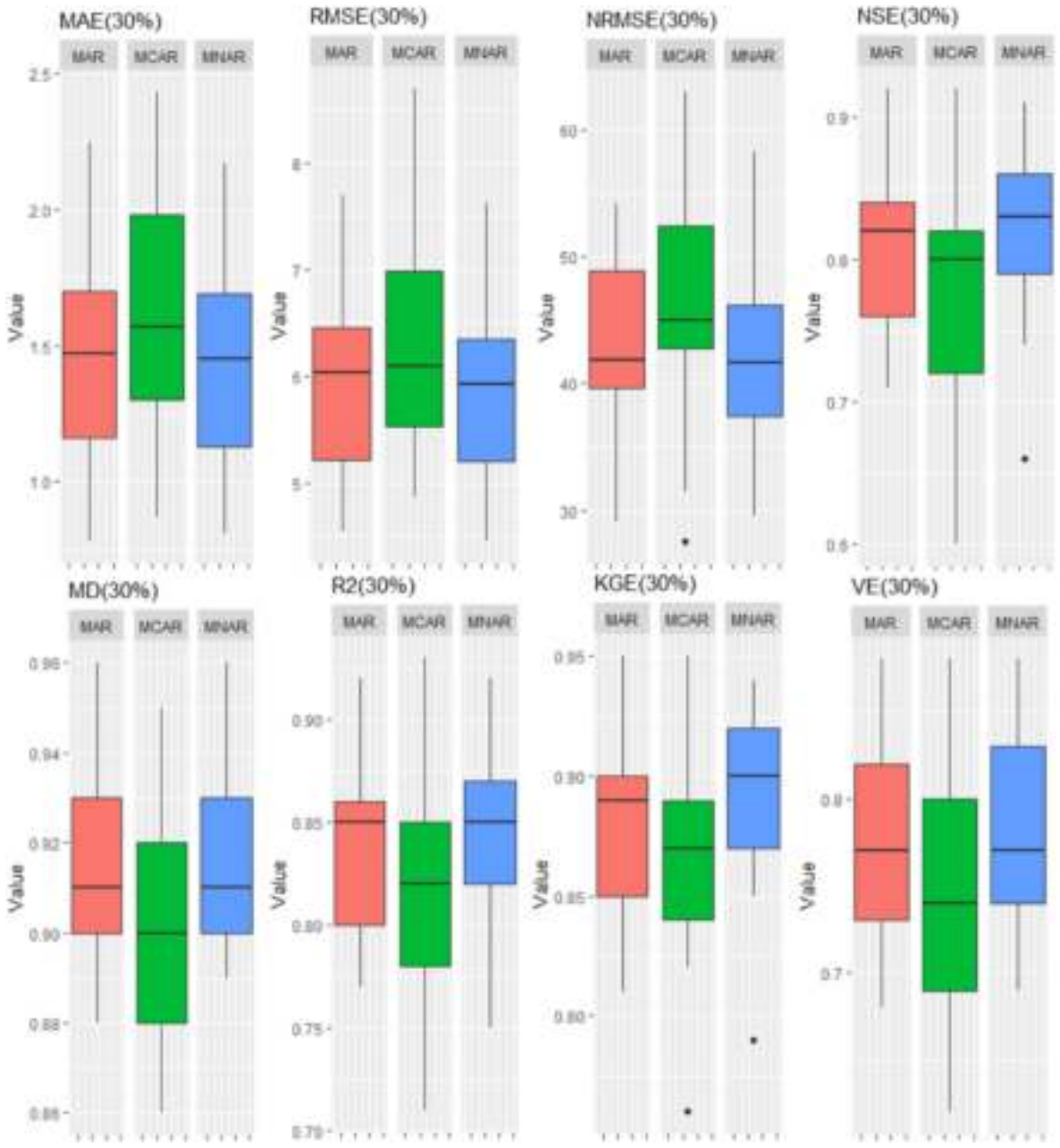

Fig. 8. Results for the statistical performance of the *norm. predict* imputation methods based on MAR, MCAR, and MNAR under 30% missingness based on MAE, RMSE, NRMSE, NSE, MD, R^2, KGE, and VE.

Authors contributions statement

All authors contributed to the study's conception and design. Material preparation, data collection, and analysis were performed by Zulfaqar Sa'adi Zulkifli Yusop, Nor Eliza Alias, Ming Fai Chow, Mohd Khairul Idlan Muhammad, Muhammad Wafiy Adli Ramli, Zafar Iqbal, Mohammed Sanusi Shiru, Faizal Immaddudin Wira Rohmat, Nur Athirah Mohamad, Mohamad Faizal Ahmad. Zulfaqar Sa'adi wrote the first draft of the manuscript. All authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding

This work was supported by the Water Security and Sustainable Development Hub funded by the UK Research and Innovation's Global Challenges Research Fund (GCRF) [grant number: ES/S008179/1].

CRediT authorship contribution statement

Zulfaqar Sa'adi: Writing - review & editing, Writing - original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Zulkifli Yusop:** Writing - review & editing, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization. **Nor Eliza Alias:** Writing - review & editing, Supervision, Resources, Project administration, Investigation, Conceptualization. **Ming Fai Chow:** Writing - review & editing, Methodology, Formal analysis, Data curation, Conceptualization. **Mohd Khairul Idlan Muhammad:** Writing - review & editing, Visualization, Methodology, Formal analysis, Data curation. **Muhammad Wafiy Adli Ramli:** Writing - review & editing, Visualization, Validation, Software, Methodology,

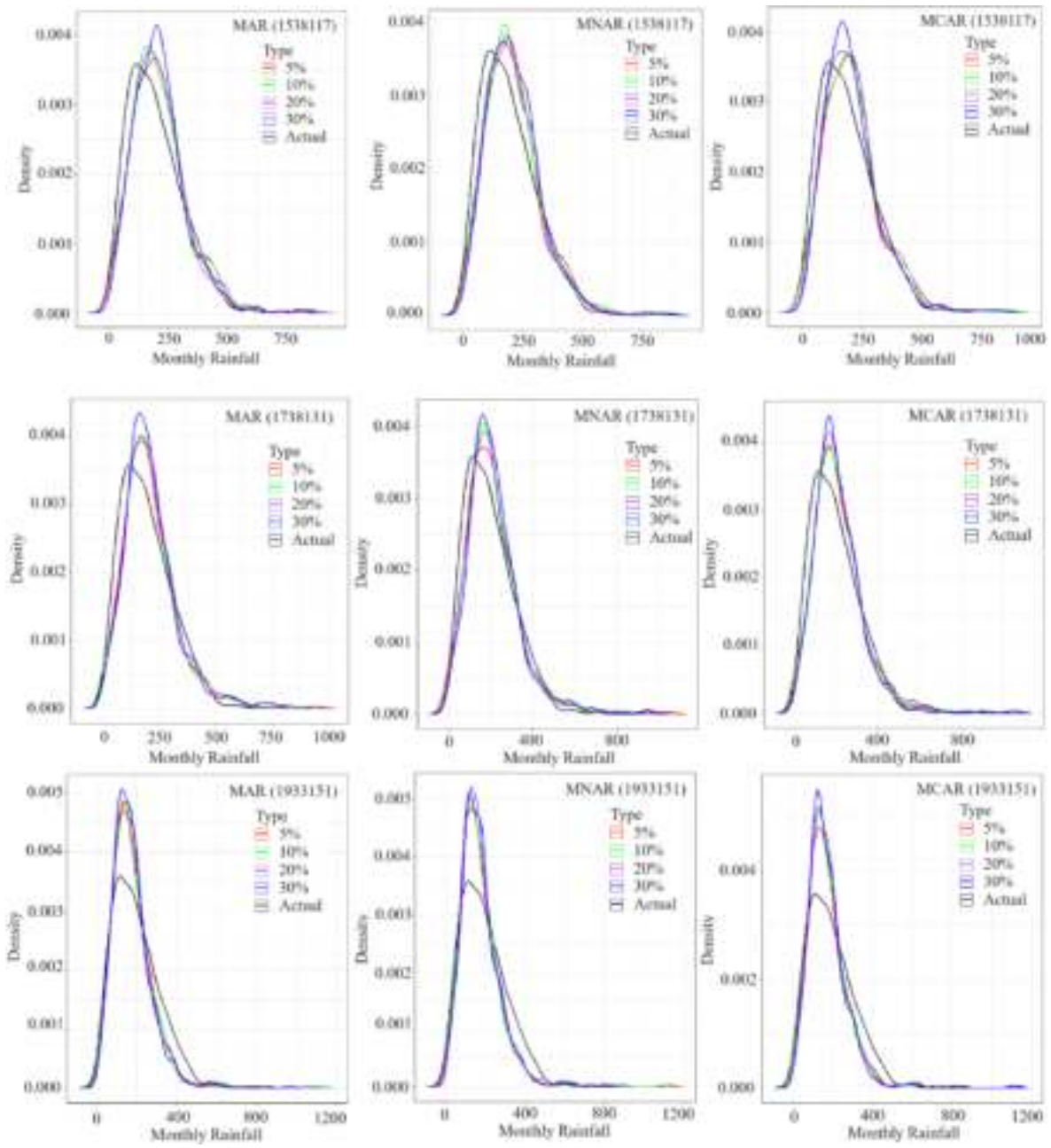

Fig. 9. Comparative assessment based on PDF for monthly rainfall for the selected imputed dataset for St. 1538117, St. 1738131, and St. 1933151 under all missingness.

Formal analysis, Data curation. **Zafar Iqbal:** Writing - review & editing, Software, Methodology, Formal analysis, Data curation. **Mohammed Sanusi Shiru:** Writing - review & editing, Software, Formal analysis, Data curation. **Faizal Immaddudin Wira Rohmat:** Writing - review & editing, Formal analysis, Data curation. **Nur Athirah Mohamad:** Writing - review & editing, Formal analysis, Data curation. **Mohamad Faizal Ahmad:** Writing - review & editing, Formal analysis, Data curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

Addi, M., Gyasi-Agyei, Y., Obuobie, E., Amekudzi, L.K., 2022. Evaluation of imputation techniques for infilling missing daily rainfall records on river basins in Ghana. Hydrol. Sci. J. 67, 613–627. https://doi.org/10.1080/02626667.2022.2030868.

Al-Khwarizmi, P., Tunggal, I., Data, M., Lenyap, H., Terbaik, Y., Abdulraqeb, G., Saeed, A., Chuan, Z.L., Zakaria, R., Syahidah, W.N., Yusoff, W., Mohd, Salleh, Z., 2016. Determination of the best single imputation algorithm for missing rainfall data treatment. J. Qual. Meas. Anal. JQMA 12, 79–87.

Appiah-Badu, N.K.A., Missah, Y.M., Amekudzi, L.K., Ussiph, N., Frimpong, T., Ahene, E., 2022. Rainfall prediction using machine learning algorithms for the various ecological zones of Ghana. IEEE Access 10, 5069–5082. https://doi.org/10.1109/ACCESS.2021.3139312.

Arguez, A., Vose, R.S., 2011. The definition of the standard WMO climate normal: the key to deriving alternative climate normals. Bull. Am. Meteorol. Soc. https://doi.org/10.1175/2010BAMS2955.1.

Balcha, S.K., Hulluka, T.A., Awass, A.A., Bantider, A., Ayele, G.T., 2023. Comparison and selection criterion of missing imputation methods and quality assessment of monthly rainfall in the Central Rift Valley Lakes Basin of Ethiopia. Theor. Appl. Climatol. 154, 483–503. https://doi.org/10.1007/S00704-023-04569-Z/FIGURES/4.

Burhanuddin, S.N.Z.A., Deni, S.M., Shaadan, N., 2021. Controlled Sampling Approach in Improving Multiple Imputation for Missing Seasonal Rainfall Data. https://doi.org/10.21203/rs.3.rs-679692/v1.

Canchala-Nastar, T., Carvajal-Escobar, Y., Alfonso-Morales, W., Cerón, W.L., Caicedo, E., 2019. Estimation of missing data of monthly rainfall in southwestern Colombia using artificial neural networks. Data Brief 26, 104517. https://doi.org/10.1016/J.DIB.2019.104517.

Carvalho, J.R.P. De, Monteiro, J.E.B.A., Nakai, A.M., Assad, E.D., 2017. Model for multiple imputation to estimate daily rainfall data and filling of faults. Rev. Bras. Meteorol. 32, 575–583. https://doi.org/10.1590/0102-7786324006.

Che Ros, F., Tosaka, H., Sidek, L.M., Basri, H., 2016. Homogeneity and trends in long-term rainfall data, Kelantan River Basin, Malaysia. Int. J. River Basin Manag. https://doi.org/10.1080/15715124.2015.1105233.

Chen, L., Xu, J., Wang, G., Shen, Z., 2019. Comparison of the multiple imputation approaches for imputing rainfall data series and their applications to watershed models. J. Hydrol. 572, 449–460. https://doi.org/10.1016/J.JHYDROL.2019.03.025.

Chhin, R., Yoden, S., 2018. Ranking CMIP5 GCMs for model ensemble selection on regional scale: case study of the indochina region. J. Geophys. Res. Atmos. 123, 8949–8974. https://doi.org/10.1029/2017JD028026.

Chiu, P.C., Selamat, A., Krejcar, O., 2019a. Infilling missing rainfall and runoff data for Sarawak, Malaysia using Gaussian mixture model based K-nearest neighbor imputation. Lect. Notes Comput. Sci. 27–38. https://doi.org/10.1007/978-3-030-22999-3_3/COVER, 11606 LNAI.

Chiu, P.C., Selamat, A., Krejcar, O., Kuok, K.K., 2019b. Missing rainfall data estimation using artificial neural network and nearest neighbor imputation. Front. Artif. Intell. Appl. 318, 132–143. https://doi.org/10.3233/FAIA190044.

Chiu, P.C., Selamat, A., Krejcar, O., Kuok, K.K., Herrera-Viedma, E., Fenza, G., 2021. Imputation of rainfall data using the sine cosine function fitting neural network. Int. J. Interact. Multimed. Artif. Intell. 6, 39–48. https://doi.org/10.9781/IJIMAI.2021.08.013.

Chivers, B.D., Wallbank, J., Cole, S.J., Sebek, O., Stanley, S., Fry, M., Leontidis, G., 2020. Imputation of missing sub-hourly precipitation data in a large sensor network: a machine learning approach. J. Hydrol. 588, 125126 https://doi.org/10.1016/J.JHYDROL.2020.125126.

Dayal, D., Pandey, A., Gupta, P.K., Himanshu, S.K., 2023. Multi-criteria evaluation of satellite-based precipitation estimates over agro-climatic zones of India. Atmos. Res. 292, 106879 https://doi.org/10.1016/J.ATMOSRES.2023.106879.

de Carvalho, J.R.P., Almeida Monteiro, J.E.B., Nakai, A.M., Assad, E.D., 2017. Model for multiple imputation to estimate daily rainfall data and filling of faults. Rev. Bras. Meteorol. https://doi.org/10.1590/0102-7786324006.

Dewan, A., Shahid, S., Bhuian, M.H., Hossain, S.M.J., Nashwan, M.S., Chung, E.S., Hassan, Q.K., Asaduzzaman, M., 2022. Developing a high-resolution gridded rainfall product for Bangladesh during 1901–2018. Sci. Data 91 (9), 1–16. https://doi.org/10.1038/s41597-022-01568-z, 2022.

Enders, C.K., 2010. Applied Missing Data Analysis. Guilford Press.

Erdal, H.I., Karakurt, O., 2013. Advancing monthly streamflow prediction accuracy of CART models using ensemble learning paradigms. J. Hydrol. 477, 119–128. https://doi.org/10.1016/J.JHYDROL.2012.11.015.

Fakhruddin Kamaruzaman, I., Zawiah, W., Zin, W., Ariff, M., 2017. A comparison of method for treating missing daily rainfall data in Peninsular Malaysia. Malaysian J. Fundam. Appl. Sci. 13, 375–380. https://doi.org/10.11113/MJFAS.V13N4-1.781.

Farzandi, M., Rezaee-Pazhand, H., 2021. Introduction of MICE method for imputation missing meteorological data and comparison by regression; case study: 130 Years of monthly temperature in mashhad, jask and bushehr. J. Water Sustain. Dev. 8, 31–42. https://doi.org/10.22067/JWSD.V8I3.2104.1038.

Gorshenin, A.K., Martynov, O.P., 2019. Hybrid extreme gradient boosting models to impute the missing data in precipitation records. Inform. i ee Primen. 13, 34–40. https://doi.org/10.14357/19922264190306.

Hamzah, F.B., Hamzah, F.M., Razali, S.F.M., Samad, H., 2021. A comparison of multiple imputation methods for recovering missing data in hydrological studies. Civ. Eng. J. 7, 1608–1619. https://doi.org/10.28991/CEJ-2021-03091747.

Hanaish, I.S., Ibrahim, K., Jemain, A.A., 2013. On the applicability of bartlett lewis model: with reference to missing data. Mat. Malaysian J. Ind. Appl. Math. 29, 53–65. https://doi.org/10.11113/MATEMATIKA.V29.N.359.

Hyndman, Donald, Hyndman, David, 2016. Natural Hazards and Disasters. Cengage Learning.

Jahan, F., Sinha, N.C., Rahman, Md Mahfuzur, Rahman, Md Morshadur, Mondal, M.S.H., Islam, M.A., 2019. Comparison of missing value estimation techniques in rainfall data of Bangladesh. Theor. Appl. Climatol. 136, 1115–1131. https://doi.org/10.1007/S00704-018-2537-Y/FIGURES/2.

Jakhar, Y.K., Mishra, N., Poonia, R., 2018. Predication accuracy analysis of data mining algorithms on meteorological data using R programming. SSRN Electron. J. https://doi.org/10.2139/SSRN.3166223.

Kalteh, A.M., Hjorth, P., 2009. Imputation of missing values in a precipitation–runoff process database. Nord. Hydrol 40, 420–432. https://doi.org/10.2166/NH.2009.001.

Kamaruzaman, I.F., Zawiah, W., Zin, W., Ariff, M., 2017. A comparison of method for treating missing daily rainfall data in Peninsular Malaysia. Malaysian J. Fundam. Appl. Sci. 13, 375–380. https://doi.org/10.11113/MJFAS.V13N4-1.781.

Lai, W.Y., Kuok, K.K., 2019. A study on bayesian principal component analysis for addressing missing rainfall data. Water Resour. Manag. 33, 2615–2628. https://doi.org/10.1007/S11269-019-02209-8/FIGURES/7.

Latrubesse, M., de Farias, K.M.S., Bayer, M., Duarte, L.V., Formiga, K.T.M., Costa, V.A.F., 2022. Comparison of methods for filling daily and monthly rainfall missing data: statistical models or imputation of satellite retrievals?, 2022 Water 14, 3144. https://doi.org/10.3390/W14193144. Page 3144 14.

Martínez, J.L.M., Horta-Rangel, F.A., Segovia-Domínguez, I., Morua, A.R., Hernández, J.H., Martínez, J.L.M., Horta-Rangel, F.A., Segovia-Domínguez, I., Morua, A.R., Hernández, J.H., 2019. Analysis of a new spatial interpolation weighting method to estimate missing data applied to rainfall records. Atmósfera 32, 237–259. https://doi.org/10.20937/ATM.2019.32.03.06.

Milo, E., Ekonomi, L., Margo, L., Donefski, E., 2019a. Seasonal means estimation and missing data in real data time series. Appl. Math. Sci. 13, 25–32. https://doi.org/10.12988/ams.2019.812192.

Milo, E., Ekonomi, L., Margo, L., Donefski, E., 2019b. Seasonal means estimation and missing data in real data time series. Appl. Math. Sci. 13, 25–32. https://doi.org/10.12988/ams.2019.812192.

Miró, J.J., Caselles, V., Estrela, M.J., 2017. Multiple imputation of rainfall missing data in the Iberian Mediterranean context. Atmos. Res. 197, 313–330. https://doi.org/10.1016/J.ATMOSRES.2017.07.016.

Muhammad, M.K.I., Nashwan, M.S., Shahid, S., Ismail, T. bin, Song, Y.H., Chung, E.S., 2019. Evaluation of empirical reference evapotranspiration models using compromise programming: a case study of Peninsular Malaysia. Sustain. Times. https://doi.org/10.3390/su11164267.

Nashwan, M.S., Ismail, T., Ahmed, K., 2019. Non-stationary analysis of extreme rainfall in Peninsular Malaysia. J. Sustain. Sci. Manag. 14, 17–34.

Nor, S.M.C.M., Shaharudin, S.M., Ismail, S., Zainuddin, N.H., Tan, M.L., 2020. A comparative study of different imputation methods for daily rainfall data in east-coast Peninsular Malaysia. Bull. Electr. Eng. Informatics 9, 635–643. https://doi.org/10.11591/EEI.V9I2.2090.

Norazizi, N.A.A., Deni, S.M., 2019a. Comparison of artificial neural network (ANN) and other imputation methods in estimating missing rainfall data at kuantan station. In: Communications in Computer and Information Science. https://doi.org/10.1007/978-981-15-0399-3_24.

Norazizi, N.A.A., Deni, S.M., 2019b. Comparison of artificial neural network (ANN) and other imputation methods in estimating missing rainfall data at kuantan station. Commun. Comput. Inf. Sci. 1100, 298–306. https://doi.org/10.1007/978-981-15-0399-3_24/COVER.

Pak, H.Y., Chuah, C.J., Yong, E.L., Snyder, S.A., 2021. Effects of land use configuration, seasonality and point source on water quality in a tropical watershed: a case study of the Johor River Basin. Sci. Total Environ. 780, 146661 https://doi.org/10.1016/J.SCITOTENV.2021.146661.

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

Pinthong, S., Ditthakit, P., Salaeh, N., Hasan, M.A., Son, C.T., Linh, N.T.T., Islam, S., Yadav, K.K., 2022. Imputation of missing monthly rainfall data using machine learning and spatial interpolation approaches in Thale Sap Songkhla River Basin, Thailand. Environ. Sci. Pollut. Res. 1, 1–17. https://doi.org/10.1007/S11356-022-23022-8/FIGURES/8.

Poyatos, R., Sus, O., Badiella, L., Mencuccini, M., Martínez-Vilalta, J., 2018. Gap-filling a spatially explicit plant trait database: comparing imputation methods and different levels of environmental information. Biogeosciences. https://doi.org/10.5194/bg-15-2601-2018.

Salgado, C.M., Azevedo, C., Proença, H., Vieira, S.M., 2016. Missing Data. In: Secondary Analysis of Electronic Health Records. Springer. Cham. https://doi.org/10.1007/978-3-319-43742-2_13.

Sattari, M.-T., Rezazadeh-Joudi, A., Kusiak, A., 2017. Assessment of different methods for estimation of missing data in precipitation studies. Nord. Hydrol 48, 1032–1044. https://doi.org/10.2166/NH.2016.364.

Saudi, A.S.M., Azid, A., Juahir, H., Toriman, M.E., Amran, M.A., Mustafa, A.D., Azaman, F., Kamarudin, M.K.A., Saudi, M.M., 2015. Flood risk pattern recognition using integrated chemometric method and artificial neural network: a case study in the Johor River Basin. Scopus 74, 165–170. https://doi.org/10.11113/JT.V74.3772.

Shaharudin, S.M., Andayani, S., Binatari, N., Kurniawan, A., Basri, M.A.A., Zainuddin, N.H., 2020. Imputation methods for addressing missing data of monthly rainfall in Yogyakarta, Indonesia. Int. J. Adv. Trends Comput. Sci. Eng. 9, 646–651. https://doi.org/10.30534/ijatcse/2020/9091.42020.

Suhaila, J., Deni, S.M., Jemain, A.A., 2008. Detecting inhomogeneity of rainfall series in Peninsular Malaysia. Asia-Pacific J. Atmos. Sci.

Suhaila, J., Yusop, Z., 2018. Trend analysis and change point detection of annual and seasonal temperature series in Peninsular Malaysia. Meteorol. Atmos. Phys. https://doi.org/10.1007/s00703-017-0537-6.

Sun, F., Roderick, M.L., Farquhar, G.D., 2018. Rainfall statistics, stationarity, and climate change. Proc. Natl. Acad. Sci. U.S.A. 115, 2305–2310. https://doi.org/10.1073/PNAS.1705349115/SUPPL_FILE/PNAS.1705349115.SAPP.PDF.

Tan, M.L., Ficklin, D.L., Ibrahim, A.L., Yusop, Z., 2014. Impacts and uncertainties of climate change on streamflow of the Johor River Basin, Malaysia using a CMIP5 general circulation model ensemble. J. Water Clim. Chang. 5, 676–695. https://doi.org/10.2166/WCC.2014.020.

Tan, M.L., Ibrahim, A.L., Yusop, Z., Duan, Z., Ling, L., 2015. Impacts of land-use and climate variability on hydrological components in the Johor River basin, Malaysia. Hydrol. Sci. J. https://doi.org/10.1080/02626667.2014.967246.

Tasho, E.M., Zeqo, L.M., 2022. Comparison Methods of Estimating Missing Data in Real Data Time Series. https://doi.org/10.1142/S179355712250243615, 10.1142/S1793557122502436.

Tefera, G.W., Dile, Y.T., Ray, R.L., 2023. Evaluating the impact of statistical bias correction on climate change signal and extreme indices in the jemma sub-basin of blue nile basin. Sustain. Times 15, 10513. https://doi.org/10.3390/SU151310513/S1.

Tong, G., Li, F., Allen, A.S., 2020. Missing data. Princ. Pract. Clin. Trials 1–21. https://doi.org/10.1007/978-3-319-52677-5_117-1.

van Buuren, S., Groothuis-Oudshoorn, K., 2011. mice: Multivariate imputation by chained equations in R. J. Stat. Software. https://doi.org/10.18637/jss.v045.i03.

White, I.R., Royston, P., Wood, A.M., 2011. Multiple imputation using chained equations: issues and guidance for practice. Stat. Med. 30, 377–399. https://doi.org/10.1002/SIM.4067.

Wissler, A., Kelly, |, Blevins, E., Buikstra, J.E., 2022. Missing data in bioarchaeology II: a test of ordinal and continuous data imputation. Am. J. Biol. Anthropol. 179, 349–364. https://doi.org/10.1002/AJPA.24614.

Wong, C.L., Venneker, R., Uhlenbrook, S., Jamil, A.B.M., Zhou, Y., 2009. Variability of rainfall in peninsular Malaysia. Hydrol. Earth Syst. Sci. Discuss. https://doi.org/10.5194/hessd-6-5471-2009.

Worku, G., Teferi, E., Bantider, A., Dile, Y.T., 2019. Observed changes in extremes of daily rainfall and temperature in jemma sub-basin, upper blue nile basin, Ethiopia. Theor. Appl. Climatol. 135, 839–854. https://doi.org/10.1007/S00704-018-2412-X/FIGURES/13.

Zhang, Y., Moges, S., Block, P., 2016. Optimal cluster analysis for objective regionalization of seasonal precipitation in regions of high spatial–temporal variability: application to western Ethiopia. J. Clim. 29, 3697–3717. https://doi.org/10.1175/JCLI-D-15-0582.1.

Zvarevashe, W., Krishnannair, S., Sivakumar, V., 2019. Analysis of rainfall and temperature data using ensemble empirical mode decomposition. Data Sci. J. 18, 46. https://doi.org/10.5334/DSJ-2019-046.