Water Resources Management (2024) 38:5945–5972
https://doi.org/10.1007/s11269-024-03936-3

Assessment of Different Methods for Estimation of Missing Rainfall Data

Tuğçe Hırca[1] · **Gökçen Eryılmaz Türkkan**[2]

Received: 21 May 2024 / Accepted: 19 July 2024 / Published online: 31 July 2024

Abstract
Missing data is a common problem encountered in various fields, including clinical research, environmental sciences and hydrology. In order to obtain reliable results from the analysis, the data inventory must be completed. This paper presents a methodology for addressing the missing data problem by examining the missing data structure and missing data techniques. Simulated datasets were created by considering the number of missing data, missing data pattern and missing data mechanism of real datasets containing missing values, which are often overlooked in hydrology. Considering the missing data pattern, the most commonly used methods for missing data analysis in hydrology and other fields were applied to the created simulated datasets. Simple imputation techniques and expectation maximization (EM) were implemented in SPSS software and machine learning techniques such as k-nearest neighbor (kNN), together with the hot-deck were implemented in the Python programming language. In the performance evaluation based on error metrics, it is concluded that the EM method is the most suitable completion method. Homogeneity analyses were performed in the Mathematica programming language to identify possible changes and inconsistencies in the completed rainfall dataset. Homogeneity analyses revealed that most of the completed rainfall datasets are homogeneous at class 1 level, consistent and reliable and do not show systematic changes in time.

Keywords Susurluk basin · Missing rainfall data · Missing data pattern · Missing data mechanism · Expectation–maximization

1 Introduction

One of the most important hydrological factors is rainfall, which is responsible for initiating various hydrological processes within the system and consequently providing data for various different analyses (Wangwongchai et al. 2023). According to the

✉ Tuğçe Hırca
tugcehirca1@gmail.com

Gökçen Eryılmaz Türkkan
gokcen.turkkan@balikesir.edu.tr

[1] Department of Civil Engineering, Faculty of Engineering, Bayburt University, Bayburt, Turkey

[2] Department of Civil Engineering, Faculty of Engineering, Balıkesir University, Balıkesir, Turkey

report prepared by the Center for Research on Disaster Epidemiology (CRED), the most impactful natural disasters in 2022 were droughts and floods influenced by rainfall (CRED 2023). Effective management of these disasters requires optimal water resources planning, which relies on high-quality rainfall data covering significant periods. Data gaps can occur for various reasons, including erroneous manual data entry, equipment errors during data collection or missing data due to defective storage technologies (Gao et al. 2018). Despite data gaps, many hydrological analyses rely on statistical approaches based on full-time series, such as the SPI method and low duration curve.

Understanding and addressing the issue of missing data is critically important for ensuring the validity and reliability of research findings. This study is justified by the need to mitigate the adverse effects that missing data can have on statistical analyses, particularly in hydrological research where precise data is crucial. Because the presence of missing data in any series has several implications: (I) decrease in the power and accuracy of statistical research methods (Roth et al. 1999), (II) the potential for biased estimates of relationships between two or more variables (Pigott 2001), (III) reduced representativeness of samples and (IV) complexities in the analyses used in the study (Kang 2013). Due to these reasons, having a gapless time series is a necessary prerequisite for the statistical and deterministic model approach used in hydrology (Gao et al. 2018). To solve the missing data problem, researchers have focused on two main approaches: deletion and imputation. However, before opting for the deletion of missing data, it is crucial to examine whether the deficiency in the dataset is a structural defect. If the missing data stems from a structural issue, deleting it may introduce bias into the model. Moreover, a significant amount of information may be lost.

Inadequate accounting for missing data, especially for rainfall or flow time series, can lead to a poor basin simulation and due to this fact, ineffective management of water resources might occur. (Gao et al. 2023). As a result, it is necessary to impute missing values with great care. Imputation methods also fall into two main categories: value assignment (Mean, Mode, Median, etc.) or estimation-based imputation. Predictive imputation methods include machine learning techniques (k-Nearest Neighbour, Artificial Neural Networks, Support Vector, Random Forest etc.), multiple imputation methods, and model-based assignment (Maximum Likelihood/EM) methods.

As seen above, although numerous methods exist for missing data imputation in the literature, some prominent ones include the following: *Mean* (Sanusi et al. 2017; Üresin 2021; Zhang and Thorburn 2022). The simplest method is commonly used to fill in missing data in meteorology and climatology. In the use of the arithmetic mean in the missing data, either the normal annual rainfall in the measurements at the surrounding stations should be in the range of 10% of the normal annual rainfall at the target station (Egigu 2020), or arithmetic mean imputation replaces missing values in a variable with the arithmetic mean of the observed values of the same variable (Gao et al. 2023). Another most preferred method is the *Regression Analysis* (Caldera et al. 2016; Mfwango et al. 2018). The whole regression procedure is a two-stage method: In the first step, a regression model is developed using all of the full observations and missing data is then imputed based on that model. One of the most commonly used machine learning techniques for missing data imputation is the *k-nearest neighbour* (kNN) algorithm (Sallaby and Azlan 2021; Sharma and Yuden 2021). The missing observation is estimated using the values of samples (neighbours) that are similar for one or more features. The most preferred model-based assignment method is maximum likelihood-based expectation maximization (EM) (Firat et al. 2012; Malan et al. 2020). Expectation maximization, while filling in missing data, provides accuracy and consistency by measuring how close

the obtained estimates are, compared to the actual data. This case increases the reliability of the analysis results.

In addition to the methods mentioned earlier, recent studies have explored various approaches, leveraging advancing technologies. Owusu et al. (2019) evaluated three satellite rainfall products, TMPA 3B42RT, TMPA 3B42, and CMORPH, against gauged rainfall data using correlation coefficient (r), bias, and percent bias as evaluation methods. They found that TMPA 3B42 performed the best across daily, monthly, annual, and seasonal timescales, while CMORPH consistently overestimated rainfall at gauge locations. Chan Chiu et al. (2021) proposed sine cosine function fitting neural network (SC-FITNET), integrating principal component analysis (PCA) and a sine cosine algorithm, which outperformed other methods in imputing missing rainfall data. Addi et al. (2022) explored statistical imputation techniques for filling missing daily rainfall data, identifying regression, probabilistic principal component analysis (PPCA), and missForest as effective methods, particularly for capturing dry and wet periods and moderate to extreme rainfall events. Nascimento et al. (2022) applied self-organizing maps (SOM) to simulate monthly in flows using satellite-estimated rainfall, while Pinthong et al. (2022) conducted a study to evaluate different techniques for the estimation of missing monthly rainfall data. Their investigation encompassed six machine learning algorithms—Multiple Linear Regression (MLR), M5 model tree (M5), Random Forest (RF), Support Vector Regression (SVR), Multilayer Perceptron (MLP), and Gaussian Processes (GP)—as well as four spatial interpolation methods—Arithmetic Average (AA), Inverse Distance Weighting (IDW), Co-Kriging with Constant (CCW), and Nearest Neighbor (NR). The findings indicated that machine learning approaches exhibited superior performance compared to spatial interpolation methods, attributed to their capability to account for spatial constraints. Among the machine learning algorithms tested, GP demonstrated the highest efficacy in accurately estimating missing rainfall data, underscoring its potential utility in hydrological applications where spatial variability plays a critical role. Sahoo and Ghose (2022) discovered that the feed-forward artificial neural network (FNN), RF, kNN and SOM in completing missing values of rainfall data. The findings highlighted the superior performance of the FNN with error metrics, proving its effectiveness in managing data gaps in complex hydrological systems. Nida et al. (2023) evaluated imputation techniques across weather variables, favoring kNN for rainfall and mean imputation for temperature data. Khampuengson and Wang (2023) introduced full subsequence matching (FSM) as a novel approach for imputing missing values in telemetry water level data, aiming to address issues of incomplete or anomalous data caused by instrument failures. Their study compared FSM against established methods such as Interpolation, kNN, MissForest, and the long short-term memory (LSTM), demonstrating FSM's superior accuracy in imputing missing values, particularly for data exhibiting strong periodic patterns. Wangwongchai et al. (2023) investigated statistical techniques (STs) such as AA, MLR, and nonlinear iterative partial least squares (NIPALS), alongside artificial intelligence-based techniques (AITs) including long-short-term-memory recurrent neural network (LSTM-RNN), M5 model tree, and multilayer perceptron neural networks (MLPNN), for imputing missing daily rainfall data. Their findings highlighted that the M5 model tree (M5-MT) among the AITs and MLR among the STs were particularly effective, with MLR recommended for its accurate performance and straightforward application without requiring extensive prior modeling knowledge. Dariane et al. (2024) investigated various classical and machine learning methods for recovering missing streamflow data. Methods such as linear regression (LR&MLR), artificial neural networks (ANN), SVR, M5 tree, and Adaptive Neuro-Fuzzy Inference Systems (ANFIS) using Subtractive (Sub-ANFIS) and fuzzy C-means (FCM-ANFIS) clustering were compared, with machine learning approaches generally demonstrating superior performance. In the study conducted by Kannegowda et al. (2024),

Kalman Smoothing with structured time series is recommended for small, medium, and large gaps in rainfall data, and Kalman–ARIMA is suggested for very large and mixed gaps. Among multivariate methods, superior performance across varying gap lengths is consistently demonstrated by RF. Kaur et al. (2024) employed multivariate imputation by chained equations and nearest neighbors techniques to handle missing weather data crucial for avalanche forecasting. Their study assessed six key weather variables, demonstrating improved forecasting accuracy and skill scores for artificial neural network-based models following data imputation. Loh et al. (2024) compared kNN, SVR, MR, and ANN techniques for imputing missing fine sediment data, finding ANN to consistently outperform the other methods across different missing data proportions. Apart from these studies, there are other research efforts available in the literature on missing data in these fields that can provide insights for future research on the practical application of hydrological modeling, structural engineering, and theoretical methods: In the study conducted by Tama et al. (2023), rainfall-induced runoff was predicted using a W-flow model. Additionally, Kencanawati et al. (2023) employed the rational method to determine peak discharge derived from surface runoff in their research.

In the field of hydrology, the conventional approach for filling in the missing values generally involves direct regression analyses. However, the development in the machine learning techniques, particularly in recent decades, have introduced alternative methods such as the ANN, the kNN algorithm and ANFIS. When determining the most suitable method among various alternatives, researchers often create a simulated dataset by intentionally deleting some of the data with known values and then estimate these missing values. However, in most hydrology studies, this process only considers the intentional deletion, neglecting the incomplete data structure when forming a simulated dataset. Notably, the actual amount of missing data is often disregarded. Tabachnick and Fidell (2012) argue that incomplete data mechanisms and patterns have a more significant impact on research results than the incomplete data rate.

This study aims to make a significant contribution to the literature by addressing the missing data problem, which is common in hydrology and related fields, by addressing the missing data structure and missing data techniques. Unlike previous studies that often focus on theoretical frameworks or limited case studies, the current approach rigorously applies and compares estimation techniques, including traditional methods such as mean, median, interpolation and innovative machine learning approaches such as k-nearest neighbor and finally EM, a model-based imputation technique using real historical data, to simulated datasets. To the best of the authors' knowledge, this study is one of the innovative studies in hydrology in which the missing data pattern, missing data count, and missing data mechanisms, which are critical evaluation criteria regarding missing data issues, are examined simultaneously and simulated data are created based on this. In this way, it aims to present a methodology regarding the missing data problem in hydrology. Additionally, in this study, the effects of normality assumption and station selection on Expectation Maximization performance were investigated. Furthermore, the study extends beyond imputation accuracy to include comprehensive homogeneity analyses, employing tools like Mathematica to assess the temporal and spatial consistency of completed datasets.

2 Study Area Description and Data Utilized

The Susurluk Basin, located in the western Turkey between 39°-40° north latitude and 27°-30° east longitude, covers approximately 3.11% of Turkey's total surface area, spanning about 24.349 km^2 with a drainage area of 22.399 km^2. The basin is characterized by

its diverse topography, featuring Uludağ, the highest mountain in the Marmara Region, within its bounds. Extending in an east–west direction, this mountain system significantly influences the region. The basin encompasses parts of Balıkesir, Bursa, and Kütahya provinces. Noteworthy water bodies within the Susurluk Basin include Simav Stream, Nilüfer Stream, Mustafa Kemal Paşa Stream and Koca Stream. The Susurluk Basin experiences a transitional climate, exhibiting characteristics of both the Mediterranean and Black Sea climates (SBFMP 2018). With an annual mean rainfall is 688.54 mm and a mean annual flow is 5.43 km^3/year, the basin plays a crucial role in Turkey's water resources. The region's significance is further underscored by the presence of two main freshwater lakes, Manyas Lake (24.400 hectares) and Uluabat Lake (19.900 hectares) both covered by the Ramsar Agreement (Mucan 2022). Given its strategic location and recent developments, including the construction of new water resource structures such as dams, the Susurluk Basin holds considerable economic and social importance for Turkey.

Continuity in the observation series cannot be ensured due to some reasons; such as changes in station locations, opening and closing of observation stations, equipment errors, planned maintenance or updates during the data collection process and lack of crew. Additionally, there are stations in the Susurluk Basin that were operational for a certain period but were later closed. Particularly since 2005, numerous stations have been established; however, the availability of the stations with long-term records is limited. Notably, there are no stations with a sufficient recording history to adequately represent the western part of the Susurluk Basin. To address this limitation, some stations located outside the basin were incorporated into this study. Given that the provinces of Balıkesir, Bursa, and Kütahya encompass a significant portion of the basin, data from all observation stations in these provinces were obtained from the Turkish State Meteorological Service. Subsequently, stations were selected from this extensive group, considering criteria termed adaptation parameters in this study. The selection process aimed to include stations that maximally align with the basin. The adaptation parameters considered including temporality, locationality and similarity.

In the evaluation that concerns temporality, a key consideration was ensuring that selected stations had records from the same starting date until the present day. The determination of the study period's commencement was influenced by the climate reference periods, which represent consecutive 30-year intervals calculated from climate data (Demircan et al. 2013, 2014). Climate modeling studies commonly utilize the data from climate reference periods such as 1961–1990, 1971–2000, and 1981–2010 as climate norms. Consequently, for this study, it was decided that stations with records spanning 1981–2021, including the years 1981–2010, were suitable for inclusion, as this period was deemed to be the representative of the basin and its surroundings within the context of temporality.

In the analysis in corporation with the location criteria, the distance of stations with records from 1981 to 2021 to the basin and whether the basin was located within the Thiessen polygon were influential factors. Using ArcGIS software, Thiessen polygons were delineated for the stations and their impact weights were calculated. It was determined that the Gediz and Gönen stations were located within the Thiessen polygon. Despite the Edremit station is not being situated within the Thiessen polygon of the basin, it was included in the study due to its proximity to west region within the basin. The primary rationale for this selection is to reduce the spatial variations of hydrological and climatic parameters, thereby increasing the reliability and representativeness of the data.

In the assessment that was conducted for similarity, annual rainfall levels in the basin were calculated over the study period. The relationship between stations outside the basin was examined using correlation coefficients, with careful consideration to ensure that stations located outside the basin were at least moderately compatible with the basin. Additionally, the

missing value percentages of stations within the basin played a role to be able to determine which stations outside the basin should be included in the study. Consequently, 13 meteorological stations in the Susurluk Basin and its surroundings, that meet the criteria set under adaptation parameters, were selected for inclusion in the study, as depicted in Fig. 1. This methodology aims to ensure that the selected stations reflect more accurately the hydrological characteristics of the basin and to enhance the reliability of the study results.

Table 1 presents detailed information regarding the locations of the meteorological stations. Table 2 presents general descriptive statistical information of monthly total rainfall data calculated using SPSS software (2013). As seen from this table, the skewness coefficient values vary from 1.06 to 4.90. Notably, the Bursa station exhibits high skewness, indicating that a significant portion of the rainfall is concentrated around lower values, with fewer instances of high rainfall. Moreover, the Uludağ meteorology station registers a monthly total rainfall mean approximately twice of the Bursa meteorology station. Over the study period, the annual total rainfall at Uludağ reaches 2258 mm, compared to 1290.4 mm at the Bursa meteorology station.

3 Methods

3.1 General Considerations of Missing Data

This section examines the percentages of missing data, missing patterns, and various mechanisms of missing data. While these parameters are often overlooked in hydrology missing data imputation studies, they play a crucial role in data analysis and in determining appropriate strategies to be able to handle missing data. Each parameter is discussed below.

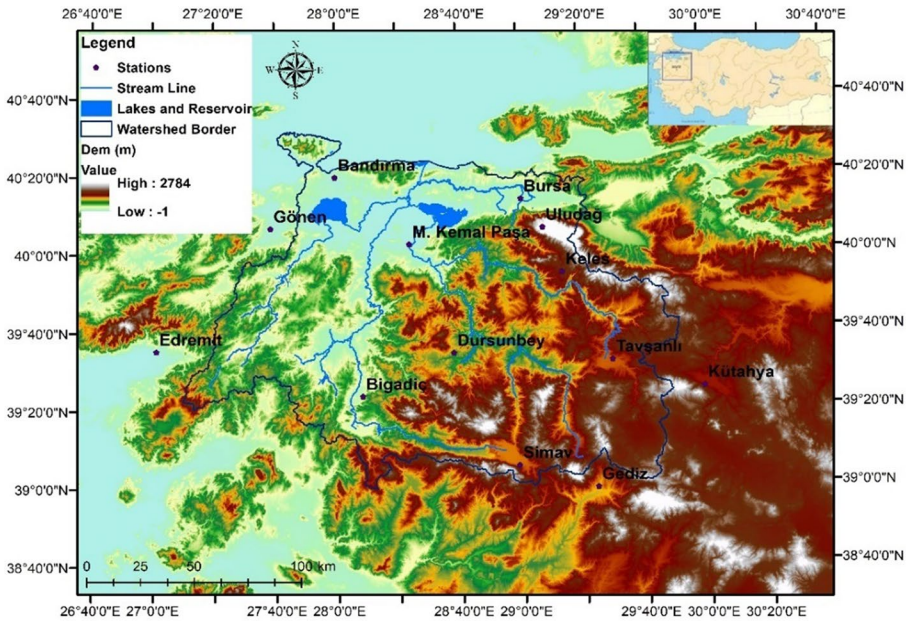

Fig. 1 Study area

Table 1 Location information of the meteorological stations

Province	Station Name	Station Code	Latitude	Longitude	Altitude (m)
Balıkesir	Bandırma	17114	40.3315	27.9965	63
	Bigadiç	17698	39.3953	28.1383	260
	Edremit	17145	39.5895	27.0192	21
	Gönen	17674	40.1135	27.6426	37
Bursa	Bursa	17116	40.2308	29.0133	100
	Dursunbey	17700	39.5778	28.6322	637
	Keles	17695	39.9150	29.2313	1063
	M.Kemal Paşa	17675	40.0425	28.3995	60
	Uludağ	17676	40.1075	29.1290	1877
Kütahya	Gediz	17750	38.9947	29.4003	736
	Kütahya	17155	39.4171	29.9891	969
	Simav	17748	39.1252	28.9919	809
	Tavşanlı	17704	39.5384	29.4941	833

Table 2 General descriptive statistics of monthly total rainfall of the meteorological stations

Province	Station Name	Station Code	Mean (mm)	Standard Deviation	Skewness Coefficient	Variation Coefficient
Balıkesir	Bandırma	17114	58.67	55.52	1.57	0.95
	Bigadiç	17698	43.63	39.56	1.08	0.91
	Edremit	17145	60.44	64.36	1.91	1.06
	Gönen	17674	56.05	51.22	1.43	0.91
Bursa	Bursa	17116	58.02	55.04	4.90	0.95
	Dursunbey	17700	45.56	38.48	1.16	0.84
	Keles	17695	59.52	47.53	1.06	0.80
	M.Kemal Paşa	17675	58.79	49.22	1.14	0.84
	Uludağ	17676	115.83	103.28	2.92	0.89
Kütahya	Gediz	17750	46.02	40.05	1.20	0.87
	Kütahya	17155	45.29	35.45	1.30	0.78
	Simav	17748	59.18	60.14	2.05	1.02
	Tavşanlı	17704	40.55	30.53	1.15	0.75

3.1.1 The Percentage of Missing Data

The percentage of missing data is vital for assessing the representing and reliability of the dataset. Low percentage indicates the more reliable dataset with stronger analysis results, while high percentage requires careful consideration when determining the strategies for handling the missing data. The acceptance of the missing data percentage depends on the research's purpose, sample size, and the mechanism of the missing data. While there is no exact threshold for an acceptable percentage of missing data, some studies have proposed distinct boundary values. For instance, Schafer (1999) suggested that a missing rate of 5% or below has minimal significance. Bennett (2001) proposed that missing data exceeding 10% is likely to introduce bias in statistical analysis. In certain statistical software, such as SPSS, 5% is used as a distinguishing point (Landau and Everitt 2004). When the proportion of the missing data is

below 5% and the missingness is either completely at random (MCAR) or missing at random (MAR), it might be feasible to exclude the missing data or use an appropriate single imputation method. Conversely, in the same scenario (MCAR or MAR missingness), if the proportion of missing data exceeds 5%, sophisticated methodologies for imputing the missing values become necessary. In cases where the missing data is classified as Missing Not At Random (MNAR) and the missingness is attributed to selection bias, corrective techniques like the Heckman adjustment can be employed (Cheema 2014; Osman et al. 2018).

3.1.2 Missing Data Patterns

The concept of missing data patterns involves identifying both missing and observable values within a dataset. It reveals the distribution of missing data and whether these gaps follow a specific pattern. For instance, understanding whether missing values are specific to a particular feature, category, or time period is crucial for comprehending the missing data pattern. While there is no standard list of missing data patterns, the three most common patterns are univariate, monotonic and nonmonotonic, as illustrated in Fig. 2.

- *Univariate:* There is a univariate missing data pattern when only one variable has missing data (Demirtas 2018; Emmanuel et al. 2021).
- *Monotone:* This pattern occurs when the missing data follows a particular order. The presence of a monotone data pattern facilitates the handling of missing values, since the patterns among these missing values may be readily observed (Dong and Peng 2013).
- *Non-Monotone:* This pattern does not follow any particular order or pattern, and the missingness of data occurred randomly or independently. Therefore, the missingness of one variable is not affected by the missingness of other variables (Chen 2022).

3.1.3 Missing Data Mechanisms

In order to learn more about the missing data problem, the cause of the missing data occurrence has been decomposed according to the various missing data mechanisms. Rubin

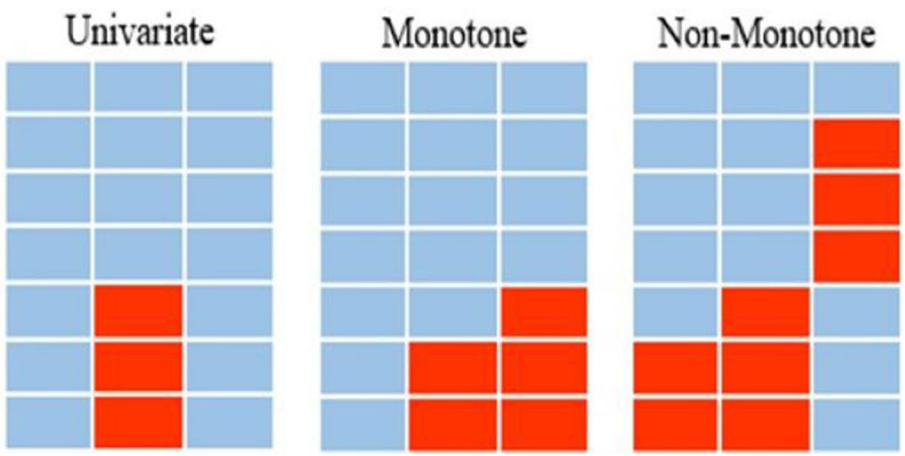

Fig. 2 Missing data patterns (Blue: observed values, red: missing values (Emmanuel et al. 2021))

(1976) appears to have been the first to introduce formally the missing data mechanisms of missing completely at random and missing at random. Rubin identified three mechanisms for missing data: Missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR) (Dong and Peng 2013). It is based on its relationship with the observed or unobserved values in the dataset. For detailed information about missing data mechanisms, the manuscript by Rubin (1976) can be used.

Kalaycıoğlu (2017) symbolized the missing data mechanisms as follows in order to be more easily understood:

In any study, let Y_i represent the dependent variable for each individual i. In this case, the dependent variable Y can be divided into two parts, $Y_{observed}$ and $Y_{missing}$, to indicate the observed and missing values, respectively. Furthermore, in the same study, let the p independent variables observed without missing values be defined by the matrix $X=(X_1, X_2, \ldots, X_k)$ $(k=1,\ldots, p)$. Under these conditions, for each individual i, the missing data index matrix R for the dependent variable Y_i can be defined as follows:

$$R_i = \left\{ \begin{array}{c} 1, \text{If } Y_i \text{ is missing} \\ 0, \text{If } Y_i \text{ is observed} \end{array} \right\}$$

- *Missing completely at random (MCAR):* The probability of the missing data is not associated with any observed or missing value of the dependent variable that contains missing data in the dataset.

$$f(\mathrm{R}|\mathrm{Y}^{\mathrm{observed}}, \mathrm{Y}^{\mathrm{missing}}) = f(R) \tag{1}$$

- *Missing at random (MAR):* The probability of the missing data occurrence in variables with the missing data is only related to the observed values, but from variables with the missing data is independent.

$$f(\mathrm{R}|\mathrm{Y}^{\mathrm{observed}}, \mathrm{Y}^{\mathrm{missing}}) = f(R|\mathrm{Y}^{\mathrm{observed}}) \tag{2}$$

Under this assumption, the probability of the missing data on the dependent variable may also be related to observed or missing data on the independent variables. Namely,

$$f(\mathrm{R}|\mathrm{Y}^{\mathrm{observed}}, \mathrm{Y}^{\mathrm{missing}}) = f(R|\mathrm{Y}^{\mathrm{observed}}, X) \tag{3}$$

- *Missing not at random (MNAR):* The probability of the missing data in the dependent variable, this is related to the missing data $\left(Y^{missing}\right)$, in the variable itself. Under this mechanism, an assumption about why the missing data occurs must be included in the statistical analysis using composite models. However, the inclusion of this assumption, which cannot be verified without prior knowledge of why the data is missing, is possible with more complex statistical models than the other methods. Because of this practical difficulty, the statistical modeling in the presence of non-random missing data has not been widely used in the literature.

The mechanisms for missing data are defined by the probability of missing data occurrence. When this probability is entirely unrelated to other measured variables, it is presumed that the remaining sample is a random subsample (Missing Completely at Random—MCAR). However, if there is a relationship between other measured factors and the likelihood of missing data, it can be inferred that the data is not MCAR. Nevertheless,

MNAR is never definitively ignored because, in practice, the missing data itself is never known. Statistical tests in literature can be employed to determine whether the missing data is entirely random or not. In this study, Little's (1988) MCAR test, one of the most preferred methods, was applied.

3.2 Dealing with the Missing Data in Rainfall Data

Over time, numerous approaches have been developed to estimate missing values in a dataset. This section discusses the missing values approaches used in this study. These approaches can be broadly classified into four categories (Fig. 3).

3.2.1 An Overview of Simple Missing Data Handling Techniques

For decades, dozens of methods have been utilized to address the issue of missing data. In this section, the most commonly used simple methods in the literature are mentioned. The simple imputation strategy involves substituting the missing values for each individual value by utilizing a quantitative or qualitative feature derived from the available non-missing data (García-Laencina et al. 2009). Various approaches, such as mode, mean, or median, are employed in simple imputation to address the missing data by utilizing the existing values. Simple imputation approaches are frequently employed in most research due to their simplicity and their utility as a convenient reference strategy (Jerez et al. 2010). The arithmetic mean approach is used to calculate the incomplete rainfall record when the normal annual rainfall of the neighboring stations is within a range by $\pm 10\%$ of the normal annual rainfall of the target station. However, if this condition is not met, the normal rate method is used for the same purpose or mean imputation can be made by using the values of the station with the missing value. In this study, as stated in Sect. 4.1., since many of the stations have missing values at the same time, each of the station's own values were used in the imputation with the mean. Since the records of the neighboring stations could not be used in the study, both the mean of the series during the study period and the mean of the previous and next two values of the missing value were imputed instead of each missing value in the station with the arithmetic mean. A similar process was applied by calculating the median of the nearby points. One of the simple approaches used to fill in missing data in any time series is spatial interpolation or temporal interpolation methods. In this study, the temporal interpolation technique was used by using the observed values just before and after the missing data.

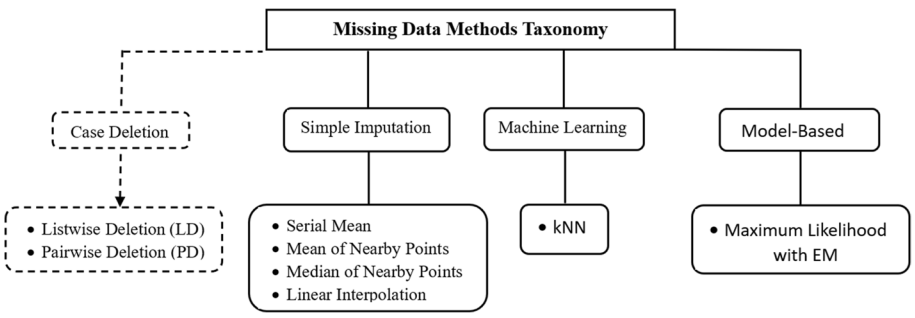

Fig. 3 Taxonomy of the missing data techniques used in this study

3.2.2 k-Nearest Neighbors (kNN) Imputation for Missing Values in Machine Learning

There are several ways available for imputing the missing data, including one of the most often used being hot-deck imputation methods. A deterministic variation of these approaches is the "nearest neighbour" (NN) imputation algorithm (Andridge and Little 2010). The hot-deck imputation approaches include the replacement of the missing values in instances with missing data (recipients) with values obtained from cases (donors) that exhibit similarity to the receiver in terms of observable attributes (Beretta and Santaniello 2016). The major disadvantage of the hot-deck attribution is the difficulty in defining the concept of 'similarity'. Therefore, the hot-deck procedure does not provide a standard path for missing data. However, it is an important technique as it allows missing values to be retrieved from a dataset without the need for additional mathematical or statistical data (Kalton and Kish 1984). Due to its relatively fast and simple algorithm, the hot deck imputation has become very popular among missing data imputation methods (Fadillah and Muchlisoh 2020).

The K-nearest neighbor algorithm (kNN) is one of those algorithms used for classification in *Supervised Learning* based on the distance function created with the parameter k. Several distance measures, including the Minkowski distance, Manhattan distance, Cosine distance, Jaccard distance, and Hamming distance, can be used for kNN imputation; however, the Euclidean distance is reported to be the most efficient and productive (Amirteimoori and Kordrostami 2010; Emmanuel et al. 2021). The Phyton programming language and the Scikit-Learn (Scikit-Learn 2023), Pandas (Pandas 2023) libraries were used for missing imputation with the kNN algorithm in this study. For detailed information about the kNN, the manuscript by Emmanuel et al. (2021) can be used.

3.2.3 Expectation–Maximization (EM) Algorithm

The Expectation–Maximization (EM) algorithm is a commonly employed iterative technique for estimating the maximum likelihood parameters in statistical models (Dempster et al. 1977). Furthermore, it facilitates the process of estimating parameters in probabilistic models that use incomplete data (Dikbas 2017). The missing values are firstly calculated using the estimated model parameters in the application of this method. These completed missing values are then used to recalculate the model parameters and this process is repeated. In the missing data completion, the EM algorithm does not take the cause of the gaps into consideration of the dataset and assumes that they are completely random. One of the most important advantages of the EM method is that the algorithm can be applied even if there are mutually between the missing values in the series and no measured values are neglected. The Gaussian Probability Distribution (normal distribution) of a multivariate data can be represented by the mean and covariance matrix. That means, the mean and covariance matrix are appropriate statistics of the normal distribution. The EM method uses an iterative algorithm and estimates the means, covariance matrix and correlations of quantitative variables with missing values. This method, which is an approach to iterative calculation of maximum likelihood (ML), estimates in various missing data problems. There are two steps in each iteration of the EM algorithm: The first step is the E-step, called the expectation step and the second step is the M-step, called the maximization step. In the E or expectation step, the missing data and the model parameters are estimated with the given observation values. In the M or maximization step, the missing data are assumed to be known and the parameters that will maximize the expected probability function in the E step are determined. This is used in the next step E to determine the distributions of the model parameters. The convergence is achieved as the probability increases with each iteration step with this

algorithm. Do and Batzoglou (2008) offer a comprehensive introduction to the mathematical underpinnings and practical applications of the expectation–maximization (EM) approach. In this study, missing value analysis was conducted using the EM modules in the toolbox of IBM SPSS software (2013).

The effect of the station selection and normality assumption in imputation with Expectation Maximization is also one of the research topics of this study. In this study, the correlation matrix was taken into account for the selection of the station. In the normality assumption, the values obtained from the EM imputation applied to the raw data were compared with the values obtained as a result of the EM imputation to the transformed forms of the data determined to be not normally distributed. In this contex, there are three basic methods that can be used to test the assumption of normality: Descriptive methods; examination of skewness, kurtosis, mean, mode, median values, Graphical methods; examination of histogram, stem-and-leaf graph, box-and-whiskers graph, P-P (probability) and Q-Q (percentage) graphs and statistical methods are Shapiro–Wilk, Kolmogorov–Smirnov, Jarque–Bera etc. In the literature, the fact that skewness and kurtosis are between certain limit values is accepted as an indicator that the data complies with the assumption of normal distribution. These limit values should be between ± 1 according to Hair et al. (2013), ± 1.5 according to Tabachnick and Fidell (2012), and ± 2 according to George and Mallery (2010).

Although skewness and kurtosis values provide researchers with a wider range of evaluations to evaluate the assumption of normality, statistical tests reveal more precise results. Many studies use and compare different tests to validate the normality. In this study, raw and transformed versions of the data were examined with three different approaches: skewness/kurtosis, Shapiro–Wilk (Shapiro and Wilk 1965) and Jarque–Bera (Jarque and Bera 1980) test. The Shapiro–Wilk and Jarque–Bera tests focus on different properties and evaluate different assumptions. While the Shapiro–Wilk test is especially effective in small samples (Pituch and Stevens 2016), the Jarque–Bera test provides a more comprehensive analysis by focusing on features such as skewness and kurtosis. Therefore, the final evaluation of the normality assumption was made with the Jarque–Bera test. However, the changes occurring at Shapiro–Wilk were also followed at every stage. For the Shapiro–Wilk test, SPSS the software normality calculation toolbox and the tseries (tseries 2023) library in R were used to calculate the Jarque–Bera test. The flow chart of the study, prepared to facilitate understanding of which analyzes were carried out at which stage of the study, is presented in Fig. 4.

3.3 Evaluation Criteria for Missing Data Imputation

To evaluate the accuracy of the prediction, error criterion parameters of mean absolute error (MAE), root mean square error (RMSE) and mean biased error (MBE) were used. Root mean square error (RMSE) is a statistical measure that measures the discrepancy between observed and predicted values. The equations of these metrics are shown in the following (Niazkar et al. 2023)

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i^{observed} - x_i^{predicted}\right)^2}{n}} \tag{4}$$

MAE quantifies the mean magnitude of a prediction set's errors as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|x_i^{observed} - x_i^{predicted}\right| \tag{5}$$

MBE is used primarily to estimate the mean bias in the model and to decide whether any steps need to be taken to correct the model bias.

$$MBE = \frac{1}{n}\sum_{i=1}^{n}(x_i^{predicted} - x_i^{observed}) \tag{6}$$

where n denotes the number of data, $x_i^{observed}$ represents the ith observed value, and $x_i^{predicted}$ represents the ith predicted value.

3.4 Homogeneity Test

The utilization of homogeneous series in climate change research holds critical importance. Changes in homogeneous series are attributed to variations in climate and weather patterns (Conrad and Pollak 1950). Several factors, such as the change in the position of the observation station, modifications in the observation format, and structural alterations in the station's surrounding environment, can impact the quality and dependability of long-term climatological time series (Peterson et al. 1998). The presence of discontinuities in non-uniform time series, which are not attributable to environmental variables, introduces uncertainty in accurately determining changes in rainfall when such data are used for climate studies. Therefore, it is imperative to assess the homogeneity of observational data before incorporating it into any research endeavor. In the event that non-homogeneous data is identified, it should be either eliminated or adjusted to achieve homogeneity. Climate scientists have developed and employed numerous approaches to assess the homogeneity of the data under consideration (Klingbjer and Moberg 2003; Ducre-Rubiatille et al. 2003; Tomozeiu et al. 2005; Staudt et al. 2007; Modarres 2008).

In this study, the homogeneity is determined by a two-step approach suggested by Wijngaard et al. (2003). In the first step, the aim is to check the homogeneity of all stations with four tests: (I) Standard Normal Homogeneity Test (SNHT) (Alexandersson 1986), (II) Pettitt's test (1979), (III) Buishand's test (1982), and (IV) Von Neumann's test (1941). The details of these tests are shown in Table 3. The test statistics in the respective table were computed using Mathematica software, and the results were evaluated within a 95% confidence interval.

Homogeneity is checked by testing the null hypothesis (H_0). The H_0 hypothesis shows that there is no change, which implies that the data under investigation is homogeneous. In the second step, the stations are divided into three classes according to the homogeneity results:

- Class 1: Homogeneous (one or zero tests reject the H_0 at the 0.05 significance level)
- Class 2: Doubtful (two tests reject the H_0 at the 0.05 significance level)
- Class 3: Suspect (three or four tests reject the H_0 at the 0.05 significance level)

4 Results and Discussion

4.1 Creation of the Simulated Datasets

The location of the missing data plays a crucial role in dataset integrity. The efficacy of model the performance, particularly in methods like median imputation of nearby points, mean of nearby points or series mean, is directly influenced by the spatial distribution of missing data. A simulated dataset, initially created in the form of a holistic dataset (449*13 data matrix), can lead to inaccurate results, especially during months with seasonal transitions, such as may and august. This is due to the consideration of september

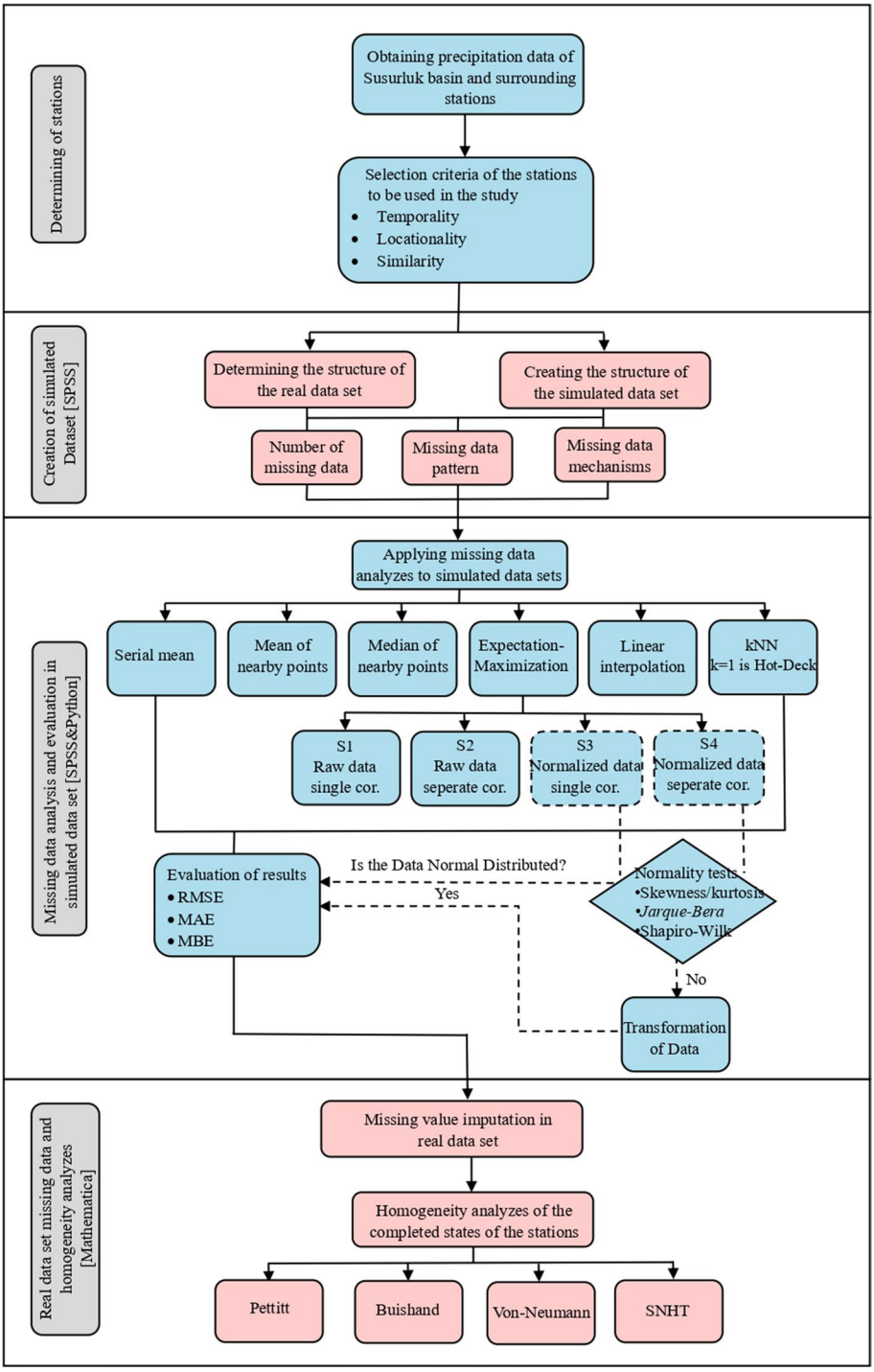

Fig. 4 Flow chart of the study

Table 3 Formulas of the homogeneity tests (Hırca et al. 2022)

Method	Formula	Description
SNHT	$T_k = kZ_1^2 + (n-k)Z_2^2$ k = 1, 2, … , n (7) Where; $Z_1 = \frac{1}{k}\sum_{i=1}^{k}\frac{(Y_i-\overline{Y})}{s}$ *and* $Z_2 = \frac{1}{n-k}\sum_{i=k+1}^{n}\frac{(Y_i-\overline{Y})}{s}$ (8) $T_0 = \max_{1\le k\le n} Tk$ (9)	Y_i = Observation data $\overline{Y}$ = Mean of the series S = Standard deviation of the series T_k = Value dependent on Z_1 and Z_2 T_0 = Test statistics
Pettitt	$X_k = 2\sum_{i=1}^{k} r_i - k(n+1)$ $k = 1, 2, \dots, n$ (10) $X_E = \max_{1\le k\le n} \lvert X_K \rvert$ (11)	r_i = The rank of the data X_k = Test statistics X_E = Critical value
Buishand	$S_0^* = 0$ *and* $S_k^* = \sum_{i=1}^{k}(Y_i - \overline{Y})$ $k = 1, 2, \dots, n$ (12) $R = (maxS_k^* - \min S_k^*)/S$ (13) $Q = R/\sqrt{n}$ (14)	S_0^* *and* S_k^* = Partial sums S = Standard deviation of the series R = Test statistics Q = Critical value
Von Neuman*	$N = \frac{\sum_{i=1}^{n-1}(Y_i - Y_{i+1})^2}{\sum_{i=1}^{n}(Y_i-\overline{Y})^2}$ (15)	N = Test statistics

* Unlike the other tests, the H_0 hypothesis is accepted if the value found in Von-Neumann's test exceeds the critical value

in the missing value imputation for august, resulting in an above-normal rainfall imputation. To address this issue, simulated datasets were generated on a monthly scale, considering the number of missing value, the missing data pattern and the missing data mechanism in the real datasets. The generation of simulated datasets is based on the deletion of rows containing missing values in the real dataset at all stations. Subsequently, in this complete dataset, a simulated dataset was created by replicating the same missing value patterns observed in the real dataset. This process was repeated for each month with missing values. Table 4 presents the mean and standard deviation values for both the real datasets and the simulated datasets. SPSS software was used and Little's MCAR test was applied to determine the missing data mechanisms. The results of Little's MCAR test indicated that the missing value mechanism in the real datasets exhibited a Missing Completely at Random (MCAR) structure. Furthermore, it was confirmed that the simulated datasets shared the same MCAR structure. If missing values in a dataset are MCAR (Missing Completely At Random), it indicates that the probability of a value being missing is unrelated to both observed and unobserved data. This indicates the following:

Randomness The missingness in the dataset is distributed randomly, without any discernible pattern or connection to the observed data or the missing values themselves.

No Bias The absence of data is not causally linked to any particular traits or values within the dataset.

The presence of missing values in an MCAR structure suggests that simple imputation techniques, such as mean or median imputation, can be appropriately utilized in the study. However, for data affected by Missing at Random (MAR) or Not Missing at Random (NMAR) mechanisms, more sophisticated imputation methods such as multiple imputation or predictive modeling techniques may be necessary.

Table 4 Comparison of real and simulated datasets in terms of missing values

Month	Structure of the Real Dataset		Structure of the Simulated Dataset			
	Mean	Std. Dev	Mechanism	Mean	Std. Dev	Mechanism
January	90.33	79.36	MCAR	90.86	79.95	MCAR
May	49.31	38.77	MCAR	50.32	38.72	MCAR
June	36.89	33.98	MCAR	38.26	34.14	MCAR
July	16.45	23.77	MCAR	19.05	26.78	MCAR
August	14.32	22.01	MCAR	16.56	25.04	MCAR
September	30.98	40.34	MCAR	36.73	42.79	MCAR
October	63.82	64.42	MCAR	66.61	65.38	MCAR
November	78.29	57.66	MCAR	79.14	58.03	MCAR
December	98.00	71.38	MCAR	99.85	71.17	MCAR

Figure 5 shows the missing data patterns for july and september for both real and simulated datasets. One of the most important parameters in the missing data studies is the missing data pattern. The procedure generally applied to impute the missing values in hydrology is based on selecting a key station for the station with missing values. In most cases, the missing value at the target station is completed using thes key station data. Therefore, in most imputation approaches (such as regression analysis, normal rate method, and some machine learning methods) estimation of missing rainfall data is possible when data is available at other stations. However, when missing values are found at all stations at the same time, the methods that directly use the data of the key station cannot be preferred. Therefore, missing data patterns should be examined in determining the methods to be used in the study.

In Fig. 5, each row in the dataset represents a different pattern of the missing values and indicates a group of samples with the same pattern of missing values. These patterns or groups of cases are organized depending on the specific variables where the missing values occur. Stations on the x-axis are ranked according to the amount of missing values. When the missing value pattern given as an example is examined, the stations with the highest missing values are on the far right of the graph and the stations with the least missing values (or no missing values) are on the far left. The initial pattern is always one, which contains no missing values. It can be seen that there are 11 different patterns in the missing data pattern for july and 9 different patterns in the missing data pattern for september. For example, while Bursa station could be used to complete Bigadiç station in july, it shows that it cannot be completed in september because there is a missing value in Bursa station in the same year. Therefore, the missing data pattern plays an important role in selecting the key station to be used for imputation.

4.2 Missing Value Imputation in Simulated Datasets

Missing rainfall values in the simulated datasets were estimated monthly by using various imputation methods, including series mean, mean of nearby points, median imputation of nearby points, linear interpolation, Hot-Deck, kNN and EM algorithms. Due to the simultaneous missing values at the stations, the key station-based methods applicable in any month were rendered inapplicable. Consequently, the column-based imputation techniques, which use the station's own records, were preferred in the study. However, EM

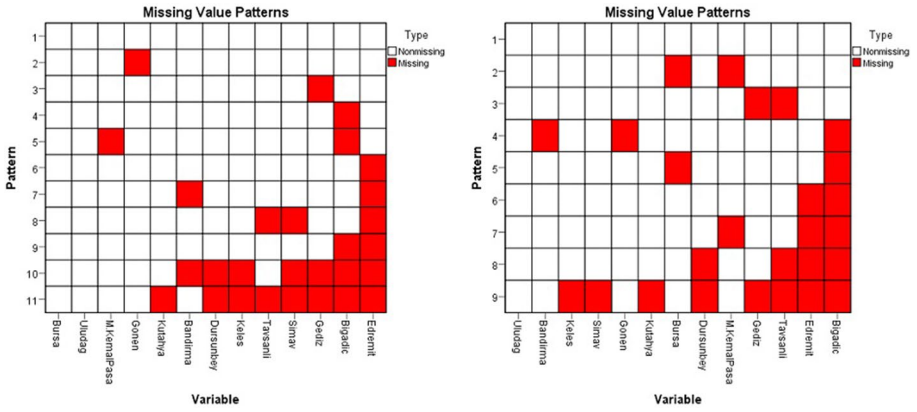

Fig. 5 Missing data pattern for (**a**) july and (**b**) september

imputation, allowing simultaneous completion, was also utilized. Different scenarios were created based on station the selection and normality assumption in EM imputation:

- *Scenario 1:* EM is imputed to the raw data by matching it with the station with which it is most compatible, as indicated by the single correlation matrix (Fig. 6) in the month with the missing value. In this scenario, the same station is used every month for matching.
- *Scenario 2:* EM is imputed to the raw data by matching it with the most compatible station in the month to be completed where the missing value is found Tables 5.
- *Scenario 3:* In the normality test results of the raw data, only the transformed versions of those that are not normally distributed are completed with the station which has the highest correlation in a single correlation matrix (Table 6).
- *Scenario 4:* In the normality test results of the raw data, only the transformed versions that are not normally distributed are completed with the station with the highest correlation in the month in which the missing value is found (Table 7).

When the correlation is examined separately each month, the station with the missing value is matched with the station with the highest correlation according to the monthly correlation analysis results. In a single correlation matrix, simulated rainfall data are listed from January 1981 to December 2021 and the first the normality analyses of the stations are evaluated (Table 6). Then the correlation analysis is performed according to the normality status of the stations. In Scenario 1 and Scenario 3, where a single correlation matrix is used, the stations with the highest correlations are matched in the same way in all months (Fig. 6).

In many studies, the assumption of the normality is often overlooked. However, depending on the result of the normality assumption, the researchers choose between parametric or non-parametric methods. Failing to investigate the normality can lead to erroneous inferences. The statistical tests can be preferred in normality tests because they provide clear results. While different limit values exist in the literature for cases where statistical tests are not preferred, according to Table 5 and Table 7, in most cases where the skewness/kurtosis coefficients are between ± 1, the Jarque–Bera test revealed that the rainfall series follows a normal distribution.

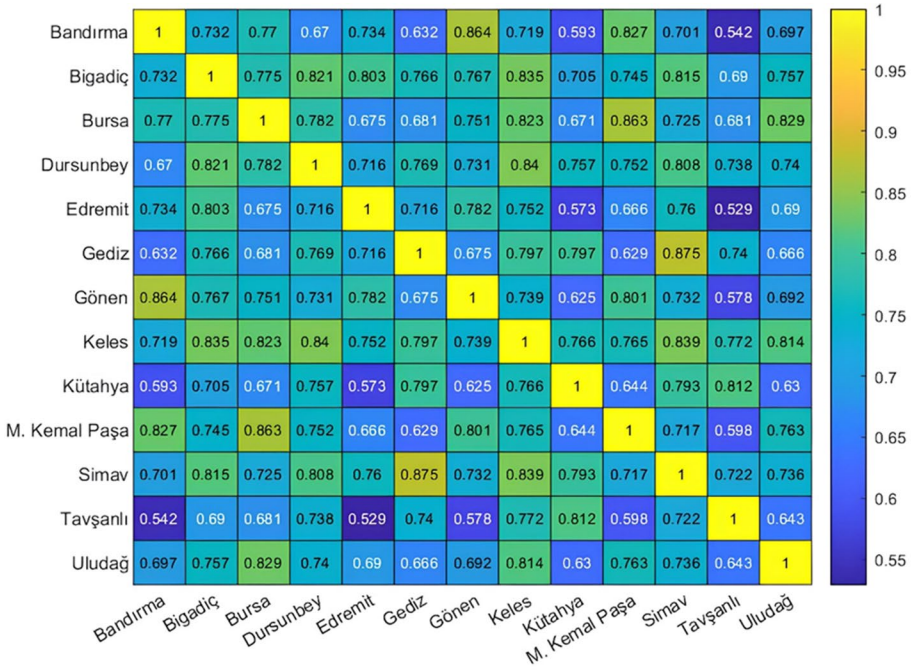

Fig. 6 Spearman's rho correlation analysis of simulated raw data in 449*13 matrix form

Table 5 Normality analysis of simulated raw data (for Scenario 2)

		Bandırma	Bigadiç	Bursa	Dursunbey	Edremit	Gediz	Gönen	Keles	Kütahya	M. Kemal Paşa	Simav	Tavşanlı	Uludağ
January	Skewness	0.803	0.641	0.449	0.834	0.514	0.750	0.527	0.737	0.652	0.912	0.871	1.112	3.913
	Kurtosis	0.706	-0.142	-0.930	0.264	-0.872	0.596	-0.634	-0.100	0.035	1.370	0.352	1.565	20.070
	Shapiro-Wilk	0.054	0.075	0.036	0.023	0.018	0.084	0.064	0.022	0.042	0.045	0.014	0.007	0.000
	Jarque-Bera	0.113	0.274	0.247	0.116	0.221	0.156	0.281	0.179	0.267	0.030	0.094	0.006	0.000
May	Skewness	1.024	1.151	0.763	1.221	1.897	0.740	1.279	0.815	0.561	0.891	1.169	0.591	1.262
	Kurtosis	0.409	1.079	-0.080	2.311	4.415	0.164	0.980	-0.016	-0.321	0.024	1.608	-0.432	1.723
	Shapiro-Wilk	0.002	0.001	0.032	0.006	0.000	0.024	0.000	0.013	0.168	0.006	0.005	0.050	0.003
	Jarque-Bera	0.038	0.010	0.160	0.001	0.000	0.184	0.004	0.126	0.325	0.085	0.004	0.270	0.002
Jun	Skewness	1.281	2.191	1.024	1.317	2.421	1.043	1.529	1.822	0.785	1.599	1.196	1.407	1.561
	Kurtosis	0.959	5.335	1.267	1.975	5.691	0.321	2.798	3.972	0.058	4.221	0.852	1.387	3.167
	Shapiro-Wilk	0.000	0.000	0.024	0.001	0.000	0.000	0.000	0.000	0.016	0.001	0.001	0.000	0.000
	Jarque-Bera	0.008	0.000	0.026	0.001	0.000	0.045	0.000	0.000	0.173	0.000	0.015	0.002	0.000
July	Skewness	0.504	2.658	0.854	1.670	2.298	2.039	0.729	0.787	1.504	2.598	1.857	0.917	2.363
	Kurtosis	-1.009	8.387	0.167	2.503	5.503	4.703	-0.678	-0.391	2.275	8.739	3.646	0.124	6.986
	Shapiro-Wilk	0.007	0.000	0.022	0.000	0.000	0.000	0.002	0.009	0.001	0.000	0.000	0.006	0.000
	Jarque-Bera	0.312	0.000	0.193	0.000	0.000	0.000	0.221	0.233	0.001	0.000	0.000	0.172	0.000
August	Skewness	2.063	2.350	2.069	1.440	1.785	2.427	3.115	3.840	1.558	0.967	2.537	2.093	3.530
	Kurtosis	4.845	6.908	4.828	1.204	2.009	7.624	10.913	17.537	1.662	-0.386	7.652	4.437	15.368
	Shapiro-Wilk	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Jarque-Bera	0.000	0.000	0.000	0.009	0.003	0.000	0.000	0.000	0.003	0.181	0.000	0.000	0.000
September	Skewness	1.912	0.869	0.559	1.235	1.823	1.870	1.954	1.269	2.092	1.067	1.748	0.838	1.427
	Kurtosis	3.390	-0.746	-0.450	0.996	3.001	3.198	3.453	1.410	5.252	0.265	3.177	-0.662	1.972
	Shapiro-Wilk	0.000	0.001	0.053	0.002	0.000	0.000	0.000	0.001	0.000	0.001	0.000	0.001	0.001
	Jarque-Bera	0.000	0.174	0.403	0.025	0.000	0.000	0.000	0.012	0.000	0.077	0.000	0.155	0.002
October	Skewness	1.352	1.562	4.285	0.975	0.972	1.272	2.115	1.947	0.676	1.665	1.400	0.937	1.812
	Kurtosis	2.247	2.331	20.758	0.813	-0.089	1.459	5.933	5.414	-0.218	3.965	2.325	1.252	3.291
	Shapiro-Wilk	0.002	0.000	0.000	0.020	0.001	0.001	0.000	0.000	0.080	0.000	0.001	0.017	0.000
	Jarque-Bera	0.000	0.000	0.000	0.049	0.066	0.004	0.000	0.000	0.240	0.000	0.000	0.038	0.000
November	Skewness	2.413	0.264	0.233	0.544	0.975	0.850	1.667	0.345	1.657	0.551	0.789	0.652	0.584
	Kurtosis	9.470	-0.448	-1.110	0.090	1.880	1.275	4.973	-0.486	3.612	-0.055	0.623	0.141	0.037
	Shapiro-Wilk	0.000	0.689	0.062	0.404	0.035	0.034	0.001	0.477	0.000	0.258	0.107	0.156	0.338
	Jarque-Bera	0.000	0.638	0.296	0.400	0.008	0.049	0.000	0.527	0.000	0.379	0.128	0.270	0.346
December	Skewness	0.725	0.750	0.652	0.840	2.045	0.840	0.869	0.670	1.681	0.261	1.797	1.197	0.848
	Kurtosis	0.329	0.989	0.694	0.251	6.627	0.376	1.005	0.523	4.692	-0.335	4.163	2.842	1.842
	Shapiro-Wilk	0.050	0.051	0.257	0.027	0.000	0.028	0.039	0.289	0.000	0.701	0.000	0.012	0.055
	Jarque-Bera	0.194	0.122	0.226	0.113	0.000	0.110	0.062	0.231	0.000	0.689	0.000	0.000	0.018

Skewness/Kurtosis values exceed ± 1.5 | H_0 is rejected for Shapiro-Wilk | H_0 is rejected for Jarque-Bera

One of the continuous variables may not meet the Pearson correlation normality assumption. In such cases, Spearman's rho correlation is an alternative nonparametric method to determine whether a linear relationship exists between two variables.

Table 6 Normality analysis of simulated raw data in 449*13 matrix form (for Scenario 1 and Scenario 3)

	Bandırma	Bigadiç	Bursa	Dursunbey	Edremit	Gediz	Gönen	Keles	Kütahya	M. Kemal Paşa	Simav	Tavşanlı	Uludağ
Skewness	1.514	1.025	5.119	1.118	1.874	1.145	1.412	1.020	1.285	1.100	1.995	1.092	3.017
Kurtosis	3.324	0.855	55.990	1.420	6.061	1.521	2.773	1.033	3.184	1.457	6.343	1.710	23.218
Shapiro-Wilk	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Jarque-Bera	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Skewness/Kurtosis values exceed ± 1.5 H_0 is rejected for Shapiro-Wilk H_0 is rejected for Jarque-Bera

Therefore, in this study, the Jarque–Bera and Shapiro–Wilk tests were examined comparatively at all stages, where correlation analysis is required. But if the H_0 hypothesis is rejected in Jarque–Bera method, it is accepted that the rainfall series are not normally distributed. Hypotheses created under the assumption of normality;

H_0 The data conforms to normal distribution.

H_1 The data does not comply with normal distribution.

Since the raw datasets are used in Scenario 1 and Scenario 2, the assumption of normality is only used to determine the correlation analysis showing the relationship between stations (Pearson? or Spearman's rho?). Scenario 2 is based on correlation analyses calculated separately for each month. Therefore, according to the normality analysis results in Table 5, the correlation coefficients are calculated for each month and the station with which the station containing the missing value is most compatible is determined.

The normality test results of the raw data prepared within the scope of Scenario 1 and Scenario 2 and are given in Table 5. There are various approaches to investigate the normality of rainfall data. While some studies consider skewness/kurtosis coefficients (Basu et al. 2004; Guo 2022), some studies consider Shapiro–Wilk (Mohammed and Scholz 2023) and others consider Jarque–Bera test (Ünlükara et al. 2010; Ahani et al. 2012; Weslati et al. 2023). Both Shapiro–Wilk and Jarque–Bera tests were performed to evaluate the assumption of the normality. Both tests focus on different characteristics; while the Shapiro–Wilk test is especially effective in small samples (Pituch and Stevens 2016), the Jarque–Bera test provides a more comprehensive analysis by focusing on features such as skewness and kurtosis. In this context, since it was desired to evaluate the skewness and kurtosis properties of the datasets in more detail, the final normality evaluation was carried out according to the Jarque–Bera test.

The procedural steps performed for the normality analysis results in Tables 5, 6 and 7 are explained in detail below;

1. Collect Raw Data: Gather the data for analysis from stations.
2. The normality test is performed using appropriate statistical tests to check if the raw data follows a normal distribution. Common tests include the Jarque–Bera test, Shapiro–Wilk test or skewness/kurtosis coefficients.
3. Based on the results of the normality tests, each dataset is classified as either normally distributed or non-normally distributed.
4. Stations that are not normally distributed are transformed using methods such as square root, logarithmic or cube root.
5. The statistical test is selected based on the normality test results, choosing between Spearman's rho correlation coefficient (for non-normally distributed data) or Pearson's correlation coefficient (for normally distributed data).

Table 7 Normality analysis of transformed versions of only non-normal distributed simulated data (for Scenario 3 and Scenario 4)

		Bandırma	Bigadiç	Bursa	Dursunbey	Edremit	Gediz	Gönen	Keles	Kütahya	M. Kemal Paşa	Simav	Tavşanlı	Uludağ
January	Skewness										0.003		0.116	-0.253*
	Kurtosis										-0.146		0.297	0.889*
	Shapiro-Wilk										0.710		0.630	0.349*
	Jarque-Bera										0.939		0.950	0.603*
May	Skewness	0.174	0.328		0.313	0.585		0.351				0.287		0.333
	Kurtosis	-0.539	-0.542		0.020	0.775		0.041				0.236		0.180
	Shapiro-Wilk	0.737	0.298		0.816	0.155		0.204				0.605		0.922
	Jarque-Bera	0.662	0.520		0.730	0.285		0.675				0.771		0.711
Jun	Skewness	0.347	0.875	0.099	0.310	-0.316*	0.346	0.446	0.809		0.250	0.233	0.673	0.598
	Kurtosis	-0.469	0.740	-0.157	-0.449	0.075*	-0.978	0.105	1.343		0.210	-0.352	-0.140	0.980
	Shapiro-Wilk	0.381	0.036	0.910	0.195	0.614*	0.076	0.409	0.075		0.634	0.929	0.098	0.173
	Jarque-Bera	0.563	0.103	0.910	0.612	0.766*	0.326	0.577	0.071		0.845	0.731	0.261	0.249
July	Skewness		1.014		0.501	1.247	0.655			0.363	0.746	0.598		0.764
	Kurtosis		0.808		-0.378	0.673	-0.069			-0.216	0.392	-0.248		0.843
	Shapiro-Wilk		0.011		0.170	0.001	0.119			0.778	0.037	0.133		0.154
	Jarque-Bera		0.138		0.506	0.071	0.407			0.688	0.296	0.443		0.229
August	Skewness	0.613	0.523	1.004	0.405	1.063	0.734	0.342*	-0.346*	0.474		0.999	0.743	-0.600*
	Kurtosis	-0.067	0.133	0.411	-0.619	-0.139	0.694	-0.899*	0.168*	-0.028		0.929	0.547	0.000*
	Shapiro-Wilk	0.067	0.100	0.017	0.127	0.000	0.256	0.068*	0.311*	0.280		0.034	0.173	0.104*
	Jarque-Bera	0.466	0.640	0.144	0.524	0.156	0.300	0.476*	0.782*	0.607		0.115	0.306	0.455*
September	Skewness	1.060			0.334	0.625	0.753	1.017	0.329	0.756		0.889		0.500
	Kurtosis	0.500			-0.562	0.068	0.527	0.974	-0.445	0.566		0.070		-0.389
	Shapiro-Wilk	0.007			0.516	0.242	0.142	0.018	0.127	0.258		0.016		0.156
	Jarque-Bera	0.070			0.590	0.452	0.268	0.068	0.633	0.251		0.157		0.457
October	Skewness	0.455	0.750	0.760*	0.210		0.544	0.178*	0.058*		0.319	0.303	-0.193	0.770
	Kurtosis	-0.072	0.409	1.905*	-0.247		-0.048	-0.190*	-0.317*		0.966	0.415	0.187	1.125
	Shapiro-Wilk	0.686	0.114	0.121*	0.907		0.307	0.681*	0.896*		0.291	0.457	0.282	0.034
	Jarque-Bera	0.525	0.197	0.030*	0.789		0.418	0.848*	0.860*		0.508	0.740	0.895	0.103
November	Skewness	0.139**				-0.483	-0.467	0.472		0.159				
	Kurtosis	1.875**				0.557	0.446	0.817		0.917				
	Shapiro-Wilk	0.172**				0.261	0.200	0.635		0.471				
	Jarque-Bera	0.147**				0.442	0.486	0.381		0.642				
December	Skewness					0.361				0.188		0.370	-0.226	-0.382
	Kurtosis					1.510				1.126		0.875	0.933	0.811
	Shapiro-Wilk					0.492				0.342		0.395	0.522	0.286
	Jarque-Bera					0.209				0.497		0.475	0.583	0.488

It is the result of square root conversion without any symbol next to it. *It is the result of logarithmic transformation. **It is the result of cube root transformation.

Skewness/Kurtosis values exceed ± 1.5 H_0 is rejected for Shapiro-Wilk H_0 is rejected for Jarque-Bera

Before each stage requiring correlation analysis, normality analyses were performed consistently by adhering to the steps mentioned above. Transformation methods were applied for months and stations that did not show normal distribution in Table 5. Then, these steps were applied for each station. In order to understand the effect of transformation processes on normality, normality tests were performed again and the results are given in Table 7. As seen from the relevant table, the transformation of non-normally distributed data shows improvement according to the skewness/kurtosis, Shapiro–Wilk, and Jarque–Bera test results, indicating that the data either normalize or approach normal distribution.

Under the light of Table 5, 6 and 7 it is possible to make the following comments:

- Shapiro–Wilk is a very sensitive method for evaluating the normality assumption. Even if the skewness and kurtosis values are ± 1, there are datasets that are not normally distributed according to the test.
- The fact that there were 9 missing months in the study and the dataset was divided monthly led to 117 normality tests. Approximately 38% of the raw datasets were determined to be normally distributed. This proves that rainfall data often has a distorted and irregular structure by nature.
- Skewness and kurtosis coefficients provide researchers with broader ranges (e.g., ± 1.5 or ± 2) for assessing normality. Therefore, assuming normality based on these values is generally easier. However, skewness and kurtosis offer only an intuitive perspective on evaluating the normality of a dataset. Hence, using normality tests for a comprehensive assessment leads to more reliable results. In this study, it was determined that the Jarque–Bera test accepts the dataset as normally distributed in most cases where the skewness and kurtosis coefficients are within the ± 1 range. Therefore, in studies where normality is assessed solely based on skewness and kurtosis, considering ± 1 instead of broader thresholds is a more appropriate approach.

- Among the three different normality approachs that are examined in this study, the tests can be ranked from the strongest to the weakest as follows: Shapiro–Wilk, Jarque–Bera, and skewness/kurtosis coefficients.
- The transformation of non-normally distributed data shows improvement according to the skewness/kurtosis, Shapiro–Wilk, and Jarque–Bera test results, indicating that the data either normalize or approach normal distribution.

According to Table 8, which includes the evaluation according to the error metrics, the results according to RMSE and MAE error metrics are close to each other. MBE indicates how much the measurements or predictions deviate from the actual values. If MBE is close to zero, it indicates that the predictions are close to the actual values. Since the error metrics gave similar results and the MBE value of Scenario 2 was -0.19, it was decided to implement this scenario. While median assignment of the nearby points was the least performing method, kNN was determined to be the most effective method after the Expectation Maximization.

Different scenarios have been created to address the necessity of the normality assumption in the expectation maximization process. Based on the created scenarios, the results of the Expectation Maximization can be summarized as follows:

- The use of the Expectation Maximization (EM) algorithm for imputing missing data offers advantages such as flexibility, a robust statistical foundation, an iterative nature, the ability to handle missing data directly, minimizing data loss, preserving data distribution, and widespread availability. These advantages make the EM algorithm an effective and reliable method for missing data analysis.
- In this study, the stations close to each other were not directly matched. For instance, Uludağ and Mustafa Kemal Paşa are stations that are closer to each other (Fig. 1), but they are physically different in terms of topographic, meteorological, and hydrological aspects. The correlation analysis, being a statistical method, does not incorporate physical events, so there is no issue even if the stations are far away. The correlation analysis examines the relationship between two time series with the same units (rainfall). Since the most imputation methods are statistical analyses, the matching stations based solely on their proximity is not a correct approach. In fact, from different months of this study, it was determined that the correlation of the relationship between two stations very close to each other was very weak.
- As a calculation approach, the expectation maximization is not affected by the order. For example, logarithmically matched stations and square roots are calculated based on their order. It is a very useful method as it allows finding missing values at the key station.
- It was determined that EM imputations made after the transformation processes produced biased results.

The findings of the study, as stated by Khalifeloo et al. (2015), suggest that expectation maximization (EM) should be preferred as it offers a fast and iterative approach to missing data imputation.

Table 8 Determining the most appropriate method

Method		Comparison between Actual Values and Estimated Values		
		RMSE*	MAE**	MBE***
Series Average		27.05	17.20	-3.04
Average of Nearby Points		30.07	18.50	-6.14
Nearby Points Median Imputation		31.96	18.58	-9.90
Linear Interpolation		29.22	18.86	-2.27
Hot-Deck		25.92	16.07	-2.15
kNN2		23.31	14.77	-3.61
kNN3		21.61	13.96	-3.94
kNN4		22.71	14.81	-2.88
kNN5		21.73	14.14	-3.93
kNN6		22.13	14.46	-4.32
kNN7		23.07	15.28	-2.77
kNN8		23.60	15.56	-3.15
kNN9		22.78	15.10	-3.71
kNN10		25.26	15.61	-2.56
EM	S1: Raw Data Single Correlation	22.98	14.20	-1.01
	S2: Raw Data Separate Correlation	23.11	14.46	-0.19
	S3: Transformed Data Single Correlation	23.02	13.24	-4.56
	S4: Transformed Data Separate Correlation	23.81	14.10	-4.11

* Root Mean Square Error

** Mean Absolute Error

***Mean Bias Error

4.3 EM Imputation and Homogeneity Analysis to Real Datasets in Scenario 2

After establishing that Scenario 2 was the most suitable method for the simulated rainfall datasets, normality analyses were initially applied to the real data. Following this, correlation analyses (Spearman's rho or Pearson) were conducted for pairwise combinations based on the normality of the stations, and the most compatible station pairs were determined for each month. Finally, the complete rainfall series were obtained by applying EM.

If the rainfall series completed as a result of missing data analyses are to be used in subsequent hydrological, meteorological, climate change, and forecasting studies, they must be hydrologically/statistically reliable. For this reason, Standard Normal Homogeneity Test (SNHT), Pettitt, Buishand, and Von Neumann Ratio homogeneity tests, which are frequently included in the literature, were applied to detect inhomogeneities in the annual total rainfall series. Test statistics for homogeneity tests were calculated in the Mathematica software (2017) and evaluated according to the 95% confidence interval. The findings obtained from the homogeneity analyses are given in Table 9. The study highlights that the Pettitt test is more sensitive in detecting inhomogeneity in series.

According to Table 9, it has been determined that the majority of stations are homogeneous based on the results of homogeneity analyses. This finding provides a solid

Table 9 Homogeneity analyzes*

İstasyonlar	Pettitt	Buishand	SNHT	Von-Neumann	Homojenlik Sınıfı
Bandırma	118	1.197	4.406	1.948	Class 1: Useful
Bigadiç	96	1.026	2.794	1.535	Class 1: Useful
Bursa	96	1.216	2.711	1.845	Class 1: Useful
Dursunbey	126	1.129	3.170	1.765	Class 1: Useful
Edremit	118	1.251	6.137	1.645	Class 1: Useful
Gediz	82	0.783	4.291	2.023	Class 1: Useful
Gönen	**184**	**1.546**	6.461	**1.454**	**Class 3: Suspect**
Keles	72	0.751	6.272	1.848	Class 1: Useful
Kütahya	148	1.277	4.946	1.931	Class 1: Useful
M.Kemal Paşa	138	1.093	3.301	1.957	Class 1: Useful
Simav	**220**	1.209	**10.865**	**1.413**	**Class 3: Suspect**
Tavşanlı	**220**	1.335	7.078	1.824	Class 1: Useful
Uludağ	168	0.988	4.500	1.670	Class 1: Useful

*Critical values for 41 data at 5% significance level; Pettitt = 173.8 = 174, Buishand = 1.532, Von-Neumann = 1.495, SNHT = 8.135

foundation for making accurate predictions and reliably evaluating long-term trends in studies such as climate change research or hydrological modeling.

5 Conclusion

Missing data estimation is important for the sustainable management of water resources, as missing data can make it difficult to determine appropriate policies and strategies. The main purpose of this research is to present a methodology for missing data estimation in hydrology. In this context, simulated datasets were created by considering the number of missing data, missing data pattern and missing data mechanism of real datasets containing missing values, which are often overlooked in hydrology. This paper provides a comparison of simple imputation approaches, machine learning technique and model-based imputation method. For this purpose, a missing data imputation study is carried out for the period 1981–2021. The application of the proposed missing data methodology is given for the monthly total rainfall of the Susurluk Basin. In EM, which is a model-based assignment method, scenarios created on station selection and normality assumption allow comparison of the performance of these selections on the method. EM is determined as the most suitable assignment method, followed by the kNN method. The Jarque–Bera test generally works well for distributions with medium to long tails and test generally indicated that the rainfall series followed a normal distribution when skewness and kurtosis coefficients were within the range of ± 1. Correlation analyses between geographically close stations revealed that proximity alone does not guarantee strong correlation in rainfall patterns, emphasizing the need for a comprehensive statistical approach rather than relying solely on geographical proximity for station matching. In future applied climatological studies, it is recommended to evaluate hybrid methodologies that combine the benefits of various approaches such as statistical techniques (STs) and artificial intelligence-based techniques (AITs) discussed in the introduction, while adhering to the methodology presented in this study, when reliable key stations with no missing data can be selected. These techniques would be even more advantageous if they also account for the critical factor emphasized in this study, namely the missing data pattern.

Author Contribution Both authors contributed to the study conception and design. Data collection, and modeling was performed by T.H. Result analysis and discussion was performed by T.H. and both authors commented on previous versions of the manuscript. Both authors red and approved the final manuscript.

Funding Open access funding provided by the Scientific and Technological Research Council of Türkiye (TÜBİTAK). The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Data Availability The authors have restrictions on sharing data publicly.

Declarations

Ethical Approval The manuscript is conducted in the ethical manner advised by the targeted journal.

Consent to Participate Not applicable.

Consent to Publish The research is scientifcally consented to be published.

Competing Interests The authors declare no conflict of interests.

References

Addi M, Gyasi-Agyei Y, Obuobie E, Amekudzi LK (2022) Evaluation of imputation techniques for infilling missing daily rainfall records on river basins in Ghana. J Des Sci Hydrologiques 67(4):613–627. https://doi.org/10.1080/02626667.2022.2030868

Ahani H, Kherad M, Kousari MR, Zadeh MR, Karampour MA, Ejraee F, Kamali S (2012) An investigation of trends in precipitation volume for the last three decades in different regions of Fars province, Iran. Theor Appl Climatol 109:361–382. https://doi.org/10.1007/s00704-011-0572-z

Alexandersson H (1986) A homogeneity test applied to precipitationdata. J Climatol 6:661–675. https://doi.org/10.1002/joc.3370060607

Amirteimoori A, Kordrostami S (2010) A Euclidean distance-based measure of efficiency in data envelopment analysis. Optimization 59(7):985–996. https://doi.org/10.1080/02331930902878333

Andridge RR, Little RJ (2010) A Review of hot deck ımputation for survey non-response. Int Stat Rev 78:40–64. https://doi.org/10.1111/j.1751-5823.2010.00103.x

Basu GC, Bhattacharjee U, Ghosh R (2004) Statistical analysis of rainfall distribution and trend of rainfall anomalies districtwise during monsoon period over West Bengal. Mausam 55:409–418. https://doi.org/10.54302/mausam.v55i3.1172

Bennett DA (2001) How can I deal with missing data in my study? Aust N Z J Public Health 25(5):464–469

Beretta L, Santaniello A (2016) Nearest neighbor imputation algorithms: a critical evaluation. BMC Med Inform Decis Mak. https://doi.org/10.1186/s12911-016-0318-z

Buishand TA (1982) Some methods for testing the homogeneity of rainfall records. J Hydrol 58:11–27. https://doi.org/10.1016/0022-1694(82)90066-X

Caldera HPGM, Piyathisse VRPC, Nandalal KDW (2016) A comparison of methods of estimating missing daily rainfall data. Engineer: J Inst Eng 49:1–8. https://doi.org/10.4038/engineer.v49i4.7232

Cheema JR (2014) Some general guidelines for choosing missing data han-dling methods in educational research? J Mod Appl Stat Methods 13:53–75

Chan Chiu P, Selamat A, Krejcar O, Kuok 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(7):39–48. https://doi.org/10.9781/ijimai.2021.08.013

Chen YC (2022) Pattern graphs: A graphical approach to nonmonotone missing data. Ann Statist 50(1). https://doi.org/10.1214/21-aos2094

Conrad V, Pollak LW (1950) Methods in Climatology. Harvard University Press, London, England. https://doi.org/10.4159/harvard.9780674187856

CRED (2023) Disasters in numbers. Centre for Research on the Epidemiology of Disasters. https://cred.be/sites/default/files/2022_EMDAT_report.pdf. Accessed 26 June 2023

Dariane AB, Borhan MI (2024) Comparison of classical and machine learning methods in estimation of missing streamflow data. Water Resour Manage 38(4):1453–1478. https://doi.org/10.1007/s11269-023-03730-7

Demircan M, Arabacı H, Bölük E, Akçakaya A, Ekici M (2013) İklim normalleri: üç sıcaklık normalinin ilişkileri ve uzamsal dağılımları. MGM. https://mgm.gov.tr/FILES/iklim/yayinlar/2013/4.pdf. Accessed 20 Aug 2023 (in Turkish)

Demircan M, Demir Ö, Atay H, Eskioğlu O, Tüvan A, Akçakaya A (2014) Climate change projections for Turkey with new scenarios. MGM. https://www.mgm.gov.tr/FILES/iklim/8-ClimateChangeProjectionsForTurkey.pdf. Accessed 20 Aug 2023 (in Turkish)

Demirtas H (2018) Flexible imputation of missing data. J Stat Soft, Book Rev 85(4):1–5. https://doi.org/10.18637/jss.v085.b04

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). J R Stat Soc Ser B (stat Methodol) 39:1–38. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

Dikbas F (2017) Frequency based imputation of precipitation. Stoch Env Res Risk Assess 31(9):2415–2434. https://doi.org/10.1007/s00477-016-1356-x

Do CB, Batzoglou S (2008) What is the expectation maximization algorithm? Nat Biotech 26:897–899

Dong Y, Peng CY (2013) Principled missing data methods for researchers. Springerplus 2(1):222. https://doi.org/10.1186/2193-1801-2-222

Ducre-Rubiatille J, Vincent A, Boulet G (2003) Comparison of tech-niques for detection of discontinuities in temperature series. Int J Climatol 23:1087–1101. https://doi.org/10.1002/joc.924

Egigu ML (2020) Techniques of filling missing values of daily andmonthly rain fall data: a review. SF J Environ Earth Sci 3(1):1036

Emmanuel T, Maupong T, Mpoeleng D, Semong T, Mphago B, Tabona O (2021) A survey on missing data in machine learning. J Big Data 8(1):1–37. https://doi.org/10.1186/s40537-021-00516-9

Fadillah IJ, Muchlisoh S (2020) Perbandingan Metode hot-deck ımputation dan metode KNNI dalam mengatasi missing values. Semnasoffstat 2019:275–285. https://doi.org/10.34123/semnasoffstat.v2019i1.101

Firat M, Dikbas F, Koc AC, Gungor M (2012) Analysis of temperature series: estimation of missing data and homogeneity test. Meteorol Appl 19(4):397–406. https://doi.org/10.1002/met.271

Gao Y, Merz C, Lischeid G, Schneider M (2018) A review on missing hydrological data processing. Environ Earth Sci 77(2):47. https://doi.org/10.1007/s12665-018-7228-6

Gao Y, Semiromi MT, Merz C (2023) Efficacy of statistical algorithms in imputing missing data of streamflow discharge imparted with variegated variances and seasonalities. Environ Earth Sci 82(20):476. https://doi.org/10.1007/s12665-023-11139-z

García-Laencina PJ, Sancho-Gómez J-L, Figueiras-Vidal AR, Verleysen M (2009) K nearest neighbours with mutual information for simultaneous classification and missing data imputation. Neurocomputing 72(7–9):1483–1493. https://doi.org/10.1016/j.neucom.2008.11.026

George D, Mallery M (2010) SPSS for Windows Step by Step: A Simple Guide and Reference, 17.0 update (10a ed.) Boston: Pearson

Guo T (2022) Extreme precipitation strongly impacts the interaction of skewness and kurtosis of annual precipitation distribution on the Qinghai-Tibetan Plateau. Atmosphere (basel) 13:1857. https://doi.org/10.3390/atmos13111857

Hair JF, Black WC, Babin BJ, Anderson RE, Tatham RL (2013) Multivariate data analysis, 8th edn. Edinburgh Gate, Harlow: pearson education limited

Hırca T, Eryılmaz Türkkan G, Niazkar M (2022) Applications of innovative polygonal trend analyses to precipitation series of Eastern Black Sea Basin, Turkey. Theor Appl Climatol 147(1–2):651–667. https://doi.org/10.1007/s00704-021-03837-0

Jarque CM, Bera AK (1980) Efficient tests for normality, homoscedasticity and serial independence of regression residuals. Econ Lett 6(3):255–259. https://doi.org/10.1016/0165-1765(80)90024-5

Jerez JM, Molina I, García-Laencina PJ, Alba E, Ribelles N, Martín M, Franco L (2010) Missing data imputation using statistical and machine learning methods in a real breast cancer problem. Artif Intell Med 50(2):105–115. https://doi.org/10.1016/j.artmed.2010.05.002

Kalaycıoğlu O (2017) An application of sensitivity analysis in the presence of non-random missing data using selection models. J Stat: Stat Actuarial Sci 10(2):76–85 (in Turkish)

Kalton G, Kish L (1984) Some efficient random imputation methods. Commun Statist-Theor Meth 13(16):1919–1939. https://doi.org/10.1080/03610928408828805

Kannegowda N, Udayar Pillai S, Kommireddi CVNK, Fousiya (2024) Comparative assessment of univariate and multivariate imputation models for varying lengths of missing rainfall data in a humid tropical region: a case study of Kozhikode, Kerala, India. Acta Geophys 72(4):2663–2678. https://doi.org/10.1007/s11600-023-01152-y

Kang H (2013) The prevention and handling of the missing data. Korean J Anesthesiol 64:402–406. https://doi.org/10.4097/kjae.2013.64.5.402

Kaur P, Joshi JC, Aggarwal P (2024) Estimation of missing weather variables using different data mining techniques for avalanche forecasting. Nat Haz (dordrecht, Netherlands) 120:5075–5098. https://doi.org/10.1007/s11069-024-06406-6

Kencanawati M, Iranata D, Maulana MA (2023) Hydrologic modeling system HEC-HMS application for direct runoff determination. J Hum Earth Future 4(2):153–165. https://doi.org/10.28991/hef-2023-04-02-02

Khalifeloo MH, Munira M, Heydari M (2015) Application of different statistical methods to recover missing rainfall data in the Klang River catchment. Int J Innov Sci Math 3:2347–9051

Khampuengson T, Wang W (2023) Novel methods for imputing missing values in water level monitoring data. Water Resour Manage 37(2):851–878. https://doi.org/10.1007/s11269-022-03408-6

Klingbjer P, Moberg A (2003) A composite monthly temperaturerecord from Tornedalen in northern Sweden. Int J Climatol 23:1465–1493. https://doi.org/10.1002/joc.946

Landau S, Everitt BS (2004) A Handbook of statistical analyses using SPSS, vol 24. CRC Press, Boca Raton, USA

Little RJA (1988) A test of missing completely at random for multivariate data with missing values. J Am Stat Assoc 83(404):1198–1202. https://doi.org/10.1080/01621459.1988.10478722

Loh WS, Ling L, Chin RJ, Lai SH, Loo KK, Seah CS (2024) A comparative analysis of missing data imputation techniques on sedimentation data. Ain Shams Eng J 15(6):102717. https://doi.org/10.1016/j.asej.2024.102717

Malan L, Smuts CM, Baumgartner J, Ricci C (2020) Missing data imputation via the expectation-maximization algorithm can improve principal component analysis aimed at deriving biomarker profiles and dietary patterns. Nutr Res 75:67–76. https://doi.org/10.1016/j.nutres.2020.01.001

Mathematica (2017) Wolfram research, inc., mathematica, Version 11.2. Champaign, IL. http://wolfram.com

Mfwango LH, Catherine JS, Shija K (2018) Estimation of missing river flow data for hydrologic analysis: the case of Great Ruaha River catchment. Hydrol Curr Res 9(2):299

Modarres R (2008) Regional frequency distribution type of low flowin North of Iran by L-moments. Water Resour Manage 22:823–841. https://doi.org/10.1007/s11269-007-9194-8

Mohammed R, Scholz M (2023) Quality control and homogeneity analysis of precipitation time series in the climatic region of Iraq. Atmosphere (basel) 14:197. https://doi.org/10.3390/atmos14020197

Mucan U (2022) Determination of drought distribution using palmer drought severity ındex: Case study of Susurluk basin. J Global Clim Change 1(2):63–68. https://doi.org/10.56768/10.56768/jytp.1.2.03

Nascimento TVM, Santos CAG, de Farias CAS, da Silva RM (2022) Monthly streamflow modeling based on self-organizing maps and satellite-estimated rainfall data. Water Resour Manage 36(7):2359–2377. https://doi.org/10.1007/s11269-022-03147-8

Niazkar M, Piraei R, Eryılmaz Türkkan G, Hırca T, Gangi F, Afzali SH (2023) Drought analysis using innovative trend analysis and machine learning models for Eastern Black Sea Basin. Theoret Appl Climatol 155:1605–1624. https://doi.org/10.1007/s00704-023-04710-y

Nida H, Kashif M, Khan MI, Ghamkhar M (2023) Comparison of missing data imputation methods using weather data. Pak J Agric Sci 60(2):327–336

Osman MS, Abu-Mahfouz AM, Page PR (2018) A survey on data imputation techniques: Water distribution system as a use case. IEEE Access 6:63279–63291. https://doi.org/10.1109/AccESS.2018.2877269

Owusu C, Adjei KA, Odai SN (2019) Evaluation of satellite rainfall estimates in the pra basin of Ghana. Environ Process 6(1):175–190. https://doi.org/10.1007/s40710-018-0344-1

Pandas (2023) Pandas: a python data analysis library (Version 2.0.3) [Software]. Retrieved from https://pandas.pydata.org. Accessed 07 Sept 2023

Peterson TC et al (1998) Homogeneity adjustments of in situatmospheric climate data: a review. Int J Climatol 18:1493–1517. https://doi.org/10.1002/(SICI)1097-0088(19981115)18:13
Pettitt AN (1979) A non-parametric approach to the change-pointproblem. J Roy Stat Soc: Ser C (appl Stat) 28:126–135. https://doi.org/10.2307/2346729
Pigott TD (2001) A review of methods for missing data. Educ ResEval 7:353–383. https://doi.org/10.1076/edre.7.4.353.8937
Pinthong S, Ditthakit P, Salaeh N, Hasan MA, Son CT, Linh NTT et al (2022) Imputation of missing monthly rainfall data using machine learning and spatial interpolation approaches in Thale Sap Songkhla River Basin. Environmental Science and Pollution Research International, Thailand. https://doi.org/10.1007/s11356-022-23022-8
Pituch KA, Stevens JP (2016) Applied multivariate statistics for the social sciences, 6th edn. Routledge, New York
Roth PL, Switzer FS, Switzer DM (1999) Missing data in multiple item scales: a Monte Carlo analysis of missing data techniques. Organ Res Methods 2:211–232. https://doi.org/10.1177/109442819923001
Rubin DB (1976) Inference and missing data. Biometrika 63(3):581–592
Sahoo A, Ghose DK (2022) Imputation of missing precipitation data using KNN, SOM, RF, and FNN. Soft Comput 26(12):5919–5936. https://doi.org/10.1007/s00500-022-07029-4
Sallaby AF, Azlan A (2021) Analysis of missing value imputation application with K-Nearest Neighbor (K-NN) algorithm in dataset. IJICS (Int J Inform Comp Sci) 5.2:141–144. https://doi.org/10.30865/ijics.v5i2.3185
Sanusi W, Wan Zin WZ, Mulbar U, Danial M, Side S (2017) Comparison of the methods to estimate missing values in monthly precipitation data. Int J Adv Sci Eng Inf Techno /IJASEIT 7(6): 2168–2174. https://doi.org/10.18517/ijaseit.7.6.2637
SBFMP (2018) Susurluk Basin Flood Management Plan. https://www.tarimorman.gov.tr/ (Accessed 14 Sep 2023) (in Turkish)
Schafer JL (1999) Multiple imputation: a primer. Stat Methods Med 8(1):3–15. https://doi.org/10.1191/096228099671525676
Scikit-Learn (2023) Scikit-learn: machine learning in python (Version 1.3.2) [Software]. Retrieved from https://scikit-learn.org
Shapiro SS, Wilk MB (1965) An analysis of variance test for normality (complete samples). Biometrika 52(3–4):591–611. https://doi.org/10.2307/2333709
Sharma V, Yuden K (2021) Imputing missing data in hydrology using machine learning models. Int J Eng Res Technol 10(1):78–82. https://doi.org/10.17577/ijertv10is010011
SPSS (2013) IBM SPSS statistics for windows, Version 22.0. Armonk, NY: IBM Corp
Staudt M, Esteban-parra MJ, Castri-Diez Y (2007) Homogenization of long-term monthly Spainish temperature data. Int J Climatol 27:1809–1823
Tabachnick BG, Fidell LS (2012) Using multivariate statistics. 6. Needham Heights, MA: Allyn & Bacon
Tama DR, Limantara LM, Suhartanto E, Devia YP (2023) The reliability of W-flow run-off-rainfall model in predicting rainfall to the discharge. Civ Eng J 9:1768–78. https://doi.org/10.28991/CEJ-2023-09-07-015
Tomozeiu R, Stefan S, Busuioc A (2005) Winter precipitation variabil-ity and large-scale circulation patterns in Romainia. Theoret ApplClimatol 81:193–201. https://doi.org/10.1007/s00704-004-0082-3
tseries (2023) tseries: time series analysis and computational finance (Version 0.10-55) [Software]. Retrieved from https://CRAN.R-project.org/package=tseries
Üresin U (2021) Correlation based regression imputation (CBRI) method for missing data imputation. Turk J Sci Technol 16(1):39–46
Ünlükara A, Yürekli K, Anlı AS, Örs İ (2010) Evaluation of the drought of Kayseri province based on RDI (reconnaissance) index. Res J Agric Sci 3(1):13–17 (in Turkish)
Von Neumann J (1941) Distribution of the ratio of the mean squaresuc-cessive difference to the variance. Ann Math Stat 12:367–395. https://doi.org/10.1214/aoms/1177731677
Wangwongchai A, Waqas M, Dechpichai P, Hlaing PT, Ahmad S, Humphries UW (2023) Imputation of missing daily rainfall data; A comparison between artificial intelligence and statistical techniques. MethodsX. https://doi.org/10.1016/j.mex.2023.102459
Weslati O, Bouaziz M, Serbaji MM (2023) Precipitation forecasting and monitoring in degraded land: a study case in Zaghouan. Land 12(4):738. https://doi.org/10.3390/land12040738
Wijngaard JB, Klein Tank AMG, Können GP (2003) Homogeneity of 20th century European daily temperature and precipitation series. Int J Climatol 23:679–692. https://doi.org/10.1002/joc.906

Zhang Y, Thorburn PJ (2022) Handling missing data in near real-time environmental monitoring: A system and a review of selected methods. Futur Gener Comput Syst 128:63–72. https://doi.org/10.1016/j.future.2021.09.033