Journal of Climate Change
ISSN 2395-7611 Volume 12 (2026) No. 1. pp. 185–197

https://doi.org/10.70917/jcc-2026-011

Article

Imputation techniques for missing rainfall data in the Indian Sundarbans

Priyanka Das [1], Arup Bose [2], Pabitra Banik 3, Aditi Sarkar [4], Mike A Powell [5] and Krishna Chandra Rath [6]

[1] Agricultural and Ecological Research Unit, Indian Statistical Institute, B.T. Road, Kolkata, 700108, West Bengal, India.

[2] Stat-Math Unit, Indian Statistical Institute, B.T. Road, Kolkata, 700108, West Bengal, India.

[3] School of Agriculture and Rural Development, Ramakrishna Mission Vivekanand Educational and Research Institute, Narendrapur, Kolkata 700103, India.

[4] Spatem Geoteck Pvt. Ltd, Kolkata 700084, West Bengal, India.

[5] Dept. of Renewable Resources, University of Alberta, 116 Street and 85 Avenue, Edmonton AB, T6G 2R3, Alberta, Canada.

[6] Department of Geography, Utkal University, Vani Vihar, Bhubaneswar, 751004, India.

* Correspondence author: banikpabitra@gmail.com

Abstract: Climatic station data is crucial for understanding the meteorological characteristics of the Indian Sundarbans, a World Heritage site, where over 70% of the population is engaged in agriculture. However, due to the exiguity of rain gauge stations and insufficient data from the operational stations, quantifying the climate change information and prediction at the ground level is challenging. Moreover, studies on imputing missing rainfall values, particularly in Indian Sundarbans, are limited. The present study experimented various imputation algorithms such as hot deck, k-Nearest Neighbour, Linear Regression and Inverse Distance Weighted. These were applied to the nearest IMD rain gauge stations in the study area. We have also considered the k-NN technique, as applied to IMD gridded data (0.25◦ × 0.25◦). In our analysis, we artificially excluded 6%, 16%, and 25% of the data points at random. The results obtained from the various imputation methods were then compared with the actual observed data, using agreement indices such as R2, MAE, RMSE, and MAPE. The findings reveal that the k-NN applied to IMD gridded data is the most effective approach, achieving an R2 value exceeding 0.9.

Keywords: Indian Sundarbans; Missing Data Imputation; Grid Data; Rain Gauge; IDW; k-nearest neighbours

1. Introduction

The importance of complete and reliable rainfall data in climatic studies cannot be overemphasised. Amin Burhanuddin et al. (2017) have discussed how the availability of high-quality rainfall data will produce more accurate results from an analysis. However, acquiring long-term precipitation data is more difficult than other climatic data types. For example, unlike precipitation, temperature does not fluctuate much in a small region and can be acquired from various sources.

Various factors, such as human error, relocation of measuring stations, faulty instruments, and negligence in measurement, lead to missing values in a dataset. It is one of the most significant issues in many research fields, especially prediction. They impact scientific analysis and eventually lead to inaccurate information in the water resource management sectors (Amin Burhanuddin *et al*., 2017; Lai, 2019). In particular, reliable rainfall data records are crucial in many studies (Aieb *et al*., 2019). The

simplistic approach ignores records containing missing values (Fouad *et al.,* 2021). This approach is inefficient, especially if there is a high proportion of missing values.

Missing values are usually imputed. This is an integral part of climatic studies, especially for rainfall data. The augmented data after imputation would enhance the quality of interpretation for meteorologists, hydrologists, environmentalists, and policymakers.

Usually, rainfall measurements are taken daily. Hence, even though this daily data is aggregated into monthly values for analysis of long-term rainfall data over multiple stations, the missing values primarily occur at the daily level (Meher and Das., 2019). The various imputation methods for rainfall data are sensitive to the magnitude of the data estimates, outliers present, and seasonal variation (Meher *et al*., 2017). The monthly rainfall data are less skewed than the daily data. Moreover, the daily data are sensitive to short-term changes, whereas the monthly data are mainly meant to address the long-term changes in climatological studies. As a result, the imputation technique varies according to the study's intent and the resources available.

Numerous methods have been introduced for estimating and reconstructing missing data (Sattari *et al*., 2016). They can be categorised as empirical, statistical, and functional approaches (Xia *et al*., 1999; Sattari et al., 2016). Most of these methods estimate the missing values using observations from neighbouring stations. For instance, the most basic method is *mean imputation*, where any missing value is imputed by the mean of non-missing values for the attribute (Fouad *et al*., 2021). In a regression approach, the median distribution of regression outputs could be used as the imputed value (Ramos-Calzado *et al*., 2008; Khosravi *et al*., 2014).

One of the most popular and effective methods is the *k-nearest neighbour imputation* (*k*NN). Here, the user determines a positive integer *k* and identifies the *k*-nearest "neighbours" that are "closest" to the missing unit, using some distance measure (usually the Euclidean distance). The average value of these neighbours is employed as the imputed value (Khosravi *et al*., 2014). *k*NN is an efficient method, and its accuracy is better than mean imputation, since the latter averages the values of the entire dataset. However, when the dataset is large, this method is time-consuming since it requires searching within the whole dataset to find the most similar units. Moreover, determining the appropriate value of *k* can be challenging (Batista and Monard, 2003; Rahman and Islam, 2013; Liu et al., 2015; Fouad *et al*., 2021).

Teegavarapu and Chandramouli (2005) applied the *neural network*, *kriging*, and *inverse distance weighting methods* (IDWM) to estimate missing precipitation data. They demonstrated that a better definition of weighting parameters and a surrogate measure for distances could improve the accuracy of the IDWM. The inverse distance weighting method (IDWM) is the most commonly used approach for estimating missing data in hydrology and geographical sciences (Ibid). In the US, especially in operational hydrology literature, IDWM is often referred to as the National Weather Service (NWS) method and is routinely used for the estimation of missing rainfall data (Pratama *et al.,* 2016); Teegavarapu and Chandramouli, 2005). Apart from the distance between the data points, another critical issue relevant to IDWM is the selection of neighborhood observation points to estimate missing data at a point of interest Teegavarapu and Chandramouli (2005).

De Silva *et al*. (2007) worked with missing precipitation data, using methods such as the *Normal Ratio*, IDW, and the *Arithmetic Mean /Local Mean method* in seven agro-climatic zones of Sri Lanka. The study shows that the Inverse Distance method is the most suitable for all three Low-country zones (wet, intermediate, and dry), and the Normal Ratio is the most ideal for the Mid-country and the Upcountry Intermediate zones. The Arithmetic Mean method is more appropriate for the Upcountry Wet zone, while the Aerial Precipitation Ratio method is ideal for the Mid-country Wet zone.

The Sundarbans' low-lying coastal mangrove forests are globally significant, unlike other mangrove forests. This region is home to 0.1% of the global population, making it more populous than approximately 137 countries worldwide (WWF, 2017). The International Union for Conservation of Nature (IUCN) has classified it as a *"World Heritage in Danger."* While the Sundarbans, with their characteristic mangrove forests, are spread over different countries, we focus our attention on the Indian Sundarbans. The local inhabitants primarily depend on climate-sensitive sectors such as agriculture and fishing. The Sundarbans' geographical location and climatic conditions often exacerbate social, economic, and environmental challenges. Therefore, studying long-term rainfall patterns, especially in monitoring the onset, ending, distribution, and intensity of rainfall events, is crucial for understanding the region's vulnerability to the impacts of climate change.

The scarcity of rain gauge observation stations over the study area has resulted in reliance solely on the nearby available stations. Moreover, for the most part, long-term persistent records are missing, even from the existing stations. Consequently, analysing the trend is also challenging. In the deltaic Sundarbans' long-term station records, the consistency and quality of rainfall data are inadequate for analysis. The station data of Sundarbans provided by the IMD contains numerous missing values in

different years over the entire period. Thus, we have also considered the IMD gridded data for imputing missing values. A key contribution of this research lies in the novel application of the k-NN algorithm within a gridded framework. Furthermore, the IDW method was executed on the gridded data, facilitating the derivation of site-specific values of the corresponding coordinates of individual gauge stations.

A comprehensive literature review indicates that no significant studies evaluate various methods for estimating missing precipitation data of the Indian Sundarbans. Moreover, no records or studies have been found using IMD gridded data in the Sundarbans region. Furthermore, there is a notable dearth of comparative research evaluating the performance of gridded datasets against diverse interpolation methods applied to rain gauge networks. Addressing this gap is vital for establishing more robust benchmarks in hydro-climatic modeling and ensuring the precision of regional precipitation, particularly in data-scarce regions where a dense network of suitable rain gauge stations is unavailable. The present work is intended to meet three major objectives: (a) to assess various imputation methods for imputing the missing values; (b) to evaluate the secondary gridded precipitation data set as a source for replacing the missing values of the station data, and (c) to validate an appropriate method statistically.

In Section 2, we have discussed some unique characteristics of the selected study area, the methodologies applied, and the execution process. Section 3 outlines the study's results with comparisons of the performance of the methods using error metrics, followed by the conclusion in Section 4. Our results conclude that the *k*NN applied on IMD gridded data performs exceptionally well when the % of missing data is large, whereas almost all the methods are suitable when the % is small.

2. Material and Methods

2.1. Study Area

The Sundarbans is one of the most extensive contiguous mangrove forests on earth. This tropical forest lies south of the tropic of Cancer, on the delta created by the Ganges, Brahmaputra, and Meghna rivers in the Bay of Bengal. Its area is approximately 10,000 km sq., of which 60% lies in Bangladesh and 40% in India. See Ghosh *et al.,* (2015) for more information. The continuous rise in sea level over time has had a significant adverse effect on the existence of this fragile ecosystem (Pramanik, 2016).

The Indian Sundarbans comprise 106 islands, of which 50% are inhabited. It is divided into the transition zone, the buffer zone, and the core area. The Indian Sundarbans Biosphere Reserve (SBR) is spread over, respectively, 13 and 6 blocks (district administrative units), of the two districts of South 24 Parganas and North 24 Parganas of the state of West Bengal (Sahana et al., 2016).

Based on the bio-geophysical characteristics, the Sundarbans are ecologically categorized into three distinct divisions: the beach or sea face, the swamp forests, and the mature delta. The Muriganga River borders our study area on the west and Harinbhahga and Raimangal Rivers on the east. It is part of the tide-dominated lower deltaic plain.

The deltaic mangrove region enjoys a tropical climate with a dry season between November and May, and a wet monsoonal period from June to October. The annual precipitation is between 1200 to 2200mm. Tropical cyclones result in flooding regularly. The temperature is moderate due to the region's proximity to the Bay of Bengal in the south (Sahana et al., 2016). The maximum and minimum temperatures vary between 26°C to 34°C and 11°C to 25°C respectively. The diurnal tidal activity is a distinct characteristic of a deltaic mangrove region. During a high tide, the water rises to heights between 6 and 10 ft, leading to the complete inundation of plants near the edge of the islands. During a low tide, extensive mud flats are noticeable on both sides of the rivers.

2.2. Research Design

2.2.1. Data

We use two different types of datasets that are available. These are the IMD rain gauge station data and the secondary processed gridded data from different sources.

2.2.1.1 IMD Rain Gauge Station data: There are five IMD rain gauge stations in the Sundarbans: Basirhat, Gosaba, Canning, Diamond Harbour, and Sagar Island. Figure 1 shows their geographical locations. The temporal coverage of these stations varies across the stations. The time points where the values are missing are erratic and also non-homogeneous. Out of the five stations, Basirhat and Gosaba have limited temporal coverage. Hence, we removed them from further consideration and concentrated on the other three (see Table 1).

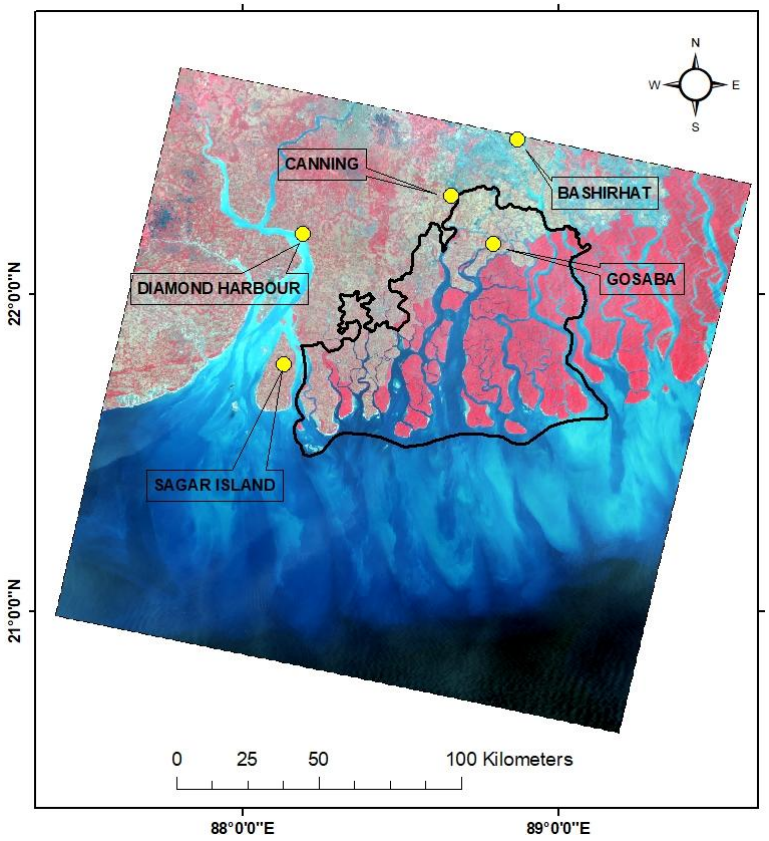


Figure 1. Location of rain gauge stations in the study area. (Source: Indian Meteorological Department.)

Table 1. The details of the rain gauge station available in the study area

Sl.no.	Station	Time-scale
1.	Basirhat (BH)	2011, 2012, 2013, 2016, 2018
2.	Canning (CN)	1982-2018
3.	Diamond Harbour (DH)	1982-2018
4.	Gosaba (GB)	1921- 1956
5.	Sagar Island	1969-2018

2.2.1.2 Gridded data: In the absence of adequate station information, the secondary high-resolution re-analysed gridded database (Meher and Das, 2019) has turned out to be very useful in climate change assessment studies. We used gridded data for some of our imputation methods.

Initially, we considered three different sources of gridded data. CRU (Climate Research Unit), NASA Power data (National Aeronautics and Space Administration), and IMD. As the gridded datasets are prepared from different sources of information, the performance of these datasets must be tested against the observational data (Meher and Das, 2019), especially in the absence of adequate station data. The performance of the gridded CRU and NASA data was assessed earlier to study the precipitation pattern, and a wide range of discrepancies was noticed (Das et al., 2021). On the other hand, the performance of IMD gridded data with 0.25° resolution was found to be superior to the IMD 1° and the CRU 0.5°resolution gridded data for imputation of missing values in the rain gauge data of the rugged Himalayan terrain of Himachal Pradesh and Uttarakhand (Meher and Das, 2019). Hence, we settled on using the IMD gridded data with 0.25° × 0.25° resolution.

2.2.2. Methods of Imputing Missing Values

There are various methods to impute missing rainfall data values. Understandably, due to different situations that may arise, there is no ideal imputation technique (Lo Presti *et al*., 2009; Aieb *et al*., 2019). For example, different imputation techniques are needed for daily and monthly data (see Meher and Das, 2019). Likewise, the percentage of missing values also dictates which imputation methods

may be appropriate. The most significant goal of an imputation method is to maintain data continuity and consistency. In a given context, not all imputation techniques are equally relevant or applicable. Hence, no ideal technique or method exists to estimate the missing values (Lo Presti *et al*., 2009; Aieb *et al*., 2019). However, imputation is generally more reliable when correlations between variables are stronger (see Aieb *et al*., 2019).

To experiment with the different imputation methods, the daily precipitation data for one year was considered for three stations with 100% data availability. We randomly deleted 6%, 16%, and 25% of this data for each station and considered these as missing values. Subsequently, these values were imputed using the methods described below. Then, the data with these imputed values incorporated was compared to the actual data using suitable error statistics.

We now briefly describe the imputation methods that we have considered.

(i) *k*-Nearest Neighbour Imputation (kNN):

To impute any missing value, the *k*NN method uses the *k* observed values/data points of the most similar data points (Aieb *et al*., 2019), where a suitable distance matrix determines similarity. This is one of the easiest and simplest methods to impute and implement. Specific implementations of this include the Hot Deck and the Arithmetic Average methods.

(i.i) Hot deck (HD):

In this implementation, a single nearest neighbour data point is taken (k = 1), and its value is assigned as the missing value. That is, the missing value P_x at location x is replaced by P_i, where i is the selected single nearest neighbour P_i (Rahman *et al*., 2015; Aieb *et al*., 2019), so that,

$$p_x = p_i \tag{1}$$

(i.ii) Arithmetic Average Method (AAM):

The AAM method is usually applied when the long-term average annual rainfall of the stations under consideration is within 10% of the long-term average annual rainfall at the adjoining stations (see Aieb *et al*., 2019). All stations having an average annual rainfall within 10% are considered, and their arithmetic average is the imputed value. Let P_x denote the imputed value of the missing data at location x. Let $P_1, P_2 \ldots, P_m$ be the available values at the other stations according to the 10% criterion. Then

$$P_x = \frac{1}{m}[P_1 + P_2 + P_m] \tag{2}$$

(ii) Linear Regression (LR)

The imputed P_x in this method uses linear regression. Using the historical/past precipitation records, the parameters for the linear regression of x on each neighbouring station were calculated by using

$$b = \frac{\sum_{i=1}^{n} xy - \frac{\sum_{i=1}^{n} x \sum_{i=1}^{n} y}{n}}{\sum_{i=1}^{n} x^2 - \frac{(\sum_{i=1}^{n} x)\, x^2}{n}} \tag{3}$$

$$y = a + bx \tag{4}$$

The highest correlated station (P_i) was used to estimate the missing rainfall data of the target station (P_x).

$$P_X = a + b.\, P_i \tag{5}$$

The estimated parameters a and b, obtained by solving the least squares equations, are given by equations (3) and (4).

(iii) Inverse Distance Weighting Method (IDWM):

In this case, the imputed values are also an average, but unlike the AAM, it is a weighted average. The idea is that the amount of influence that each available value has on the missing value at x should be more for points closest to x than those farther away (Das et al., 2021). In our study for each x, we considered two neighbouring stations and the weights taken to be the inverse of their distances from x. Thus, if r_1 and r_2 are the distances between x and the source stations 1 and 2, then the imputed value at x is obtained by (6).

$$p_x = \left[\frac{(\frac{1}{r1} \times p_1) + (\frac{1}{r2} \times p_2)}{\frac{1}{r1} + \frac{1}{r2}}\right] \tag{6}$$

(iv) IMD Gridded data:

The gridded datasets are very useful in the absence of adequate stationed data. The absence of gauge-stated data in the present study area is not the only problem. The lack of continuity and consistency of available data from the sourced station is another major hindrance. Thus, apart from

relying on the best possible techniques to substitute the data from the available neighbouring stations, identifying an unambiguous and appropriate gridded dataset is also a prerequisite to studying the climate change regime. Thus, this study included and assessed IMD 0.25° gridded data in two ways/methods.

(iv.i) Gridded data interpolation (IMD I):

First, the IMD gridded data on a particular day of a month was interpolated based on the missing day of that month. Subsequently, the corresponding station point on that layer of interpolation was retrieved. Likewise, all the missing values of the year were retrieved and interpreted for each rain gauge station, respectively. The interpolation technique evaluated the 20 nearest grid points to cover all the stations in and around the study area (see Figure 2).

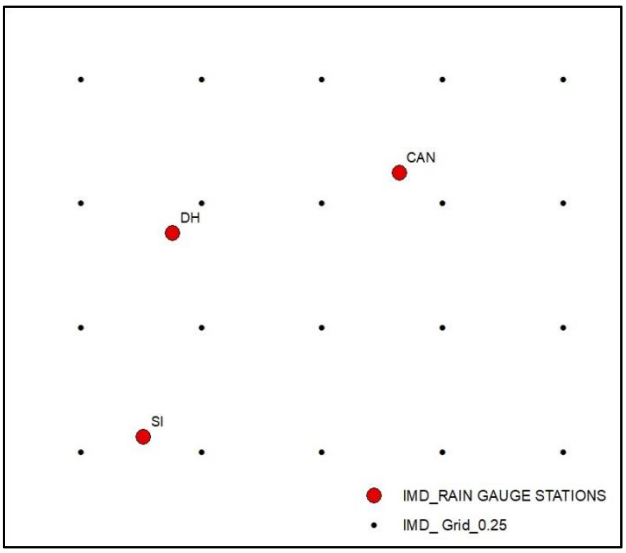


Figure 2. Geographical coordinates of rain gauge stations and the IMD grid points. (Source: Indian Meteorological Department)

(iv.ii) Nearest Neighbour Grid point (IMD NN):

This new method is applied in the present study area to recognise the nearest neighbouring grid points from the target station. Accordingly, the missing values of each station were replicated as it is from the closest grid point (see Figure 2).

2.2.3. Data Analysis Techniques

To quantify and assess the error between the actual and imputed values, we used several widely used measures: Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Root Mean Square Error (RMSE), and the Coefficient of Determination (R^2). The formulae for these measures are given in equations (7) – (10).

While MAE is the average deviation of the predicted values from the corresponding observed values, MAPE is the average of the absolute errors after dividing each error by the actual observation. The RMSE has been used as a standard statistical metric to measure model performance in meteorology, air quality, and climate research studies. For all of these measures, a lower value indicates better accuracy of the method. In the following equations, x denotes actual observations, y denotes estimated/predicted values, i denotes the index of a typical location, and n denotes the total number of data points. The coefficient of determination R^2 is also a popular measure, and a higher value indicates a greater correlation between actual and imputed values.

$$R^2 = \left\{ \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}} \right\}^2 \quad (7)$$

$$MAE = \frac{1}{n}\sum_{t=1}^{n} |x_t - y_t| \quad (8)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (x_t - y_t)^2}{n}} \quad (9)$$

$$\text{MAPE} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{At-Ft}{At}\right| \quad (10)$$

3. Results and Discussion

3.1. Station Data Characteristics

Figure 3 presents the time series of rainfall data (1982-1992) as measured in the three rain gauge stations and the IMD Gridded data. This period was selected in view of the following reasons: first, the beginning year of the full data is distinct for the different stations. Second, there is no missing data in these three rain gauge stations.

The IMD gridded data for this period is similar to the rain gauge station data. The rise and fall of the rainfall patterns are synchronised among the Diamond Harbour and Canning stations. This may be because these stations are more interior in comparison to Sagar Island. The latter, being an isolated location in the Bay of Bengal, exhibited intense rainfall in monsoon months as compared to others, especially in the years 1982, 1986, and 1991.

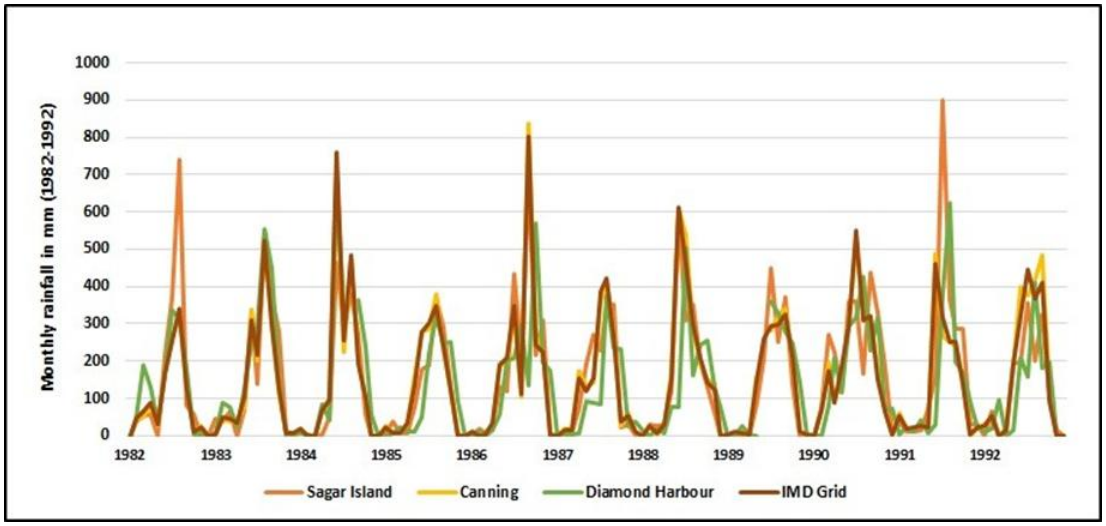


Figure 3. Time series of rainfall data in the gauge stations (SI, CAN, DH) and IMD gridded data (1982-1992).

The variations of the rainfall patterns of the IMD gridded data are consistent with the rain gauge stations during the non-monsoon months (lows in Figure 3) throughout the time series. However, during the monsoon months of 1990, 1991, and 1992, the IMD gridded data (peaks in the graph of Figure 3) were quite dissimilar to the rain gauge station data.

Table 2. Statistical parameters and characteristics of the rain gauge stations

STATIONS	MAX DAILY (mm)	MAX MONTHLY (mm)	MIN (mm)	DAILY AVG (mm)	SD	TOTAL MD DAYS	% MD (1982-2018)
CAN	270.9	838.2	0	153.49	174.206	20	0.50 %
DH	237.2	905.2	0	143.47	167.128	194	1.43%
SI	408.9	900.9	0	161.67	175.99	3466	25.64%

Table 2 provides some descriptive statistics for the rainfall data for 37 years (1982-2018) for CAN, DH, and SI. The average percentage of missing values for each available station varies from 0.5% to 97% (1982-2018).

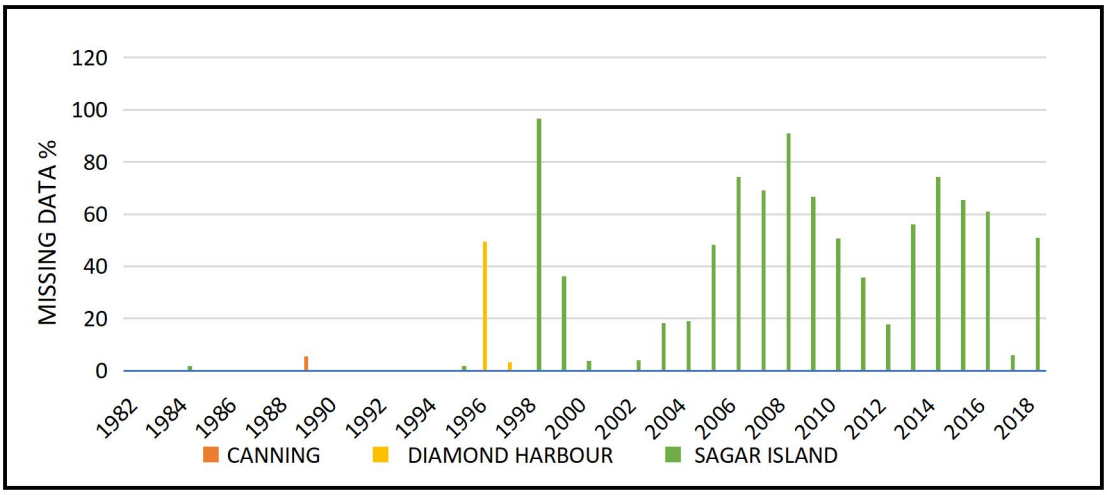


Figure 4. Percentage of missing rainfall data for Canning, Diamond Harbour and Sagar Island.

In this study, the continuity of the data structure is highly variable. In the case of Sagar Island, the daily missing values of a month go up to 96%. Figure 4 exhibits that the SI holds the maximum percent of missing values year-wise from 1982-2018. CAN had only 5.58% of missing values in 1989, whereas DH had 0.27%, 49.45%, and 3.28% missing values for 1995, 1996, and 1997, respectively. SI's maximum daily recorded rainfall is highest at 408.9, compared to 270.9 and 237.2 of CAN and DH, respectively. Kurtosis varies from 0.91 (SI) and 1.05 (CAN) to 2.12 (DH), exhibiting that the outlier characteristics of the distribution are less extreme in Sagar Island. The skewness values are 1.16 (SI), 1.25 (CAN), and 1.46 (DH), implying high asymmetry.

3.2. Analysis and Comparison of the Methods

Table 3 illustrates the prediction results of the Hot Deck, Arithmetic Mean, Linear Regression, Inverse Distance Weighted Method, and IMD gridded methods. The data used by the models in Figure 5 are from the DH, SI, and CAN stations and contain missing values.

Table 3. Comparison of actual and imputed rainfalls.

Missing Data %	Actual/ Obs. Data	Hot deck	AAM	LR	IDWM	IMD Grid interpolation (IMD I)	IMD Grid NN
6% of DH	19.3	2.6	22.4	4.3	15.8	26.8	37
	134.7	187.2	103.4	107.6	131.3	35.0	28
	0	0	0.2	3.3	0.2	1.5	1
16% of SI	2.6	19.3	30.8	27.8	29.7	13.8	5
	47.4	19.8	14.3	28.4	14.8	18.4	27
	60.2	55.2	46.2	73.3	47.0	58.3	66
	34.2	42.2	71.6	56.8	68.9	68.1	53
	0	10	17.8	18.5	13.9	3.4	4
	14.2	35.2	18.6	14.2	26.4	22.5	22
	0	4.1	2.1	13.9	3.1	14.6	15
25% of CAN	1	2.4	6.8	15.9	1.8	2.9	3
	0	14.6	12.6	15.8	11.0	2.4	2
	1.6	18.2	17.4	16.8	13.7	5.5	6
	101	42.2	38.2	20.0	31.7	84.4	84
	17.6	0	0.0	13.9	0.0	7.6	7

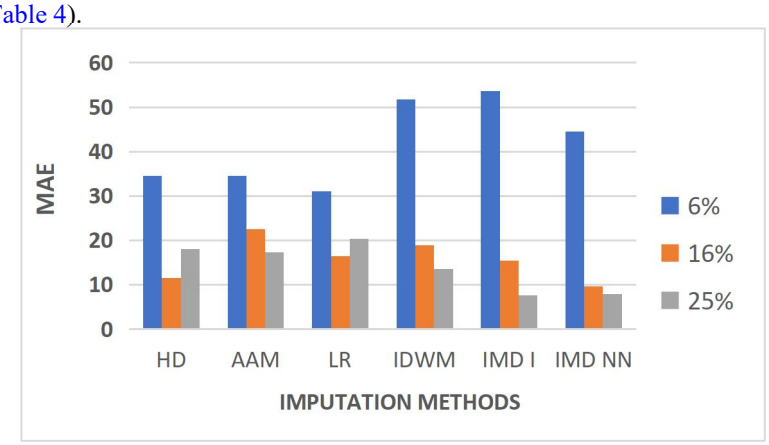
The results show that the AAM and LR methods have the minimum accuracy among all methods under study. The LR method primarily employs precipitation data from the station that demonstrates the strongest correlation with the target station, which may explain its weak performance. The HD shows a good performance when the % of missing values is minuscule. With respect to R^2, RMSE, MAPE and MAE, the IDW shows a good performer for the 25% missingness. The point of view of this study is consistent with the findings of Addi et.al. (2021), in terms of 5%, 10%, 20% and 30% of missing data, IDW performs better than HD in all cases. Rodríguez et al. (2021) explained that IDW

was the best method among other statistical and machine learning methods. A possible explanation is that IDW is the only model that, in addition to considering temporal information, includes spatial information by looking at neighbouring stations to support the imputation process (Ibid). Chen et.al. (2012) evaluated the IDW method to estimate the rainfall distribution in the middle of Taiwan. It explains that the distance and number of stations are crucial in determining the accuracy of the data, and the interpolation accuracy becomes inferior when the number of rainfall stations considered exceeds the optimal value.

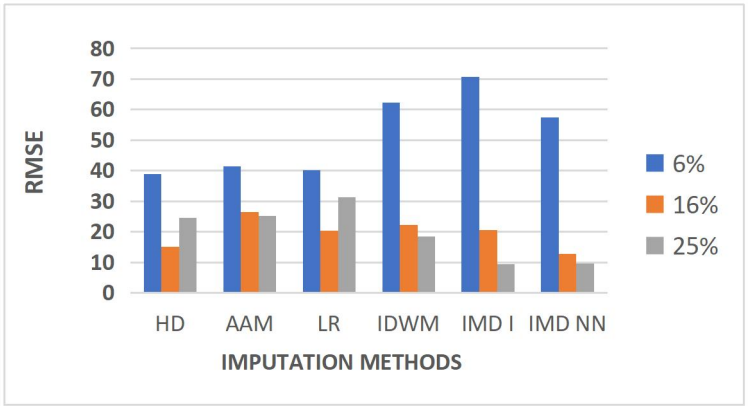
The AAM method is satisfactory if the stations are uniformly distributed over the area and the individual measurements do not vary significantly from the mean (Chow, 1971; Sattari et. al., 2016). The radar graph in Figure 6 between the imputation results and actual rainfall measurements at 6%, 16%, and 25% missing values shows that all methods are acceptable at 6% of missing values.

The distance-based correlation implicitly accounted for the best predictors across all rain gauge stations. *k*NN performs best when there is less than 10%–in that case, there are two to three missing values in a month. In terms of Coefficient of Determination*(R^2),* kNN, AAM, IDWM, and IMDI show the best results for 6% of missing values, with a strong correlation of 0.98, 0.98, and 0.99, respectively. From Figure 6, the LR methods show a good result in terms of MAE (31.06) and MAPE (0.36) in the case of 6% of missing values.

In the case of 16% of missing values, IMDNN methods perform best based on the R^2, MAE, MAPE, and RMSE. It may also be noted that LR and IMD GRID provide good R^2 values of 0.61, 0.61, and 0.76, respectively. The results obtained from the IDWM method are less reliable than the rest of the methods (see Table 4).



(b)



(c)

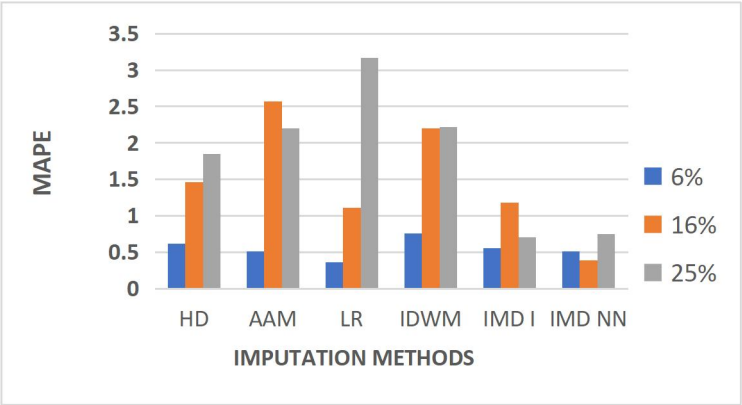


Figure 5. Mean Absolute Error (top), Root Mean Square Error (middle), and Mean Absolute Percentage Error (bottom), using different data imputation methods with 6%, 16%, and 25% of missing values.

Table 4. Performance indicators for rainfall missing data at 6%, 16% and 25%. Coefficient of Determination (R^2), Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE)

Methods	**R^2**			**RMSE**			**MAE**			**MAPE**		
HD	0.82	0.61	0.49	38.8	15.03	24.53	34.5	11.46	18.01	0.62	1.46	2.85
	6%	**16%**	**25%**	**6%**	**16%**	**25 %**	**6%**	**16%**	**25%**	**6%**	**16%**	**25%**
AAM	0.98	0.18	0.57	41.49	26.39	25.08	34.5	22.57	17.35	0.51	2.57	2.25
LR	0.94	0.61	0.36	40.12	20.28	31.24	31.06	16.51	20.36	0.36	1.11	3.17
IDWM	0.98	0.29	0.8	62.23	22.14	18.40	51.75	18.86	13.51	0.76	2.2	2.22
IMD I	0.99	0.45	0.95	70.69	20.6	9.31	53.57	15.5	7.65	0.56	1.18	0.71
IMD NN	0.95	0.76	0.95	57.36	12.72	9.55	44.5	9.68	7.85	0.51	0.39	0.75

In terms of R^2, both IMD I and IMDNN show a high correlation value between observed and predicted values, i.e., 95%, in the case of 25% of missing values. In the case of a large % of missing values, IMD I shows the best results in terms of RMSE, MAE, and MAPE.

This experiment demonstrates that the result derived from the error metrics shows a substantially lower performance of IDWM in terms of R^2, MAE, MAPE, and RMSE. The inverse distance weighting method fails to estimate the predicted values because of the huge distance between the stations and the limited working stations in the study area.

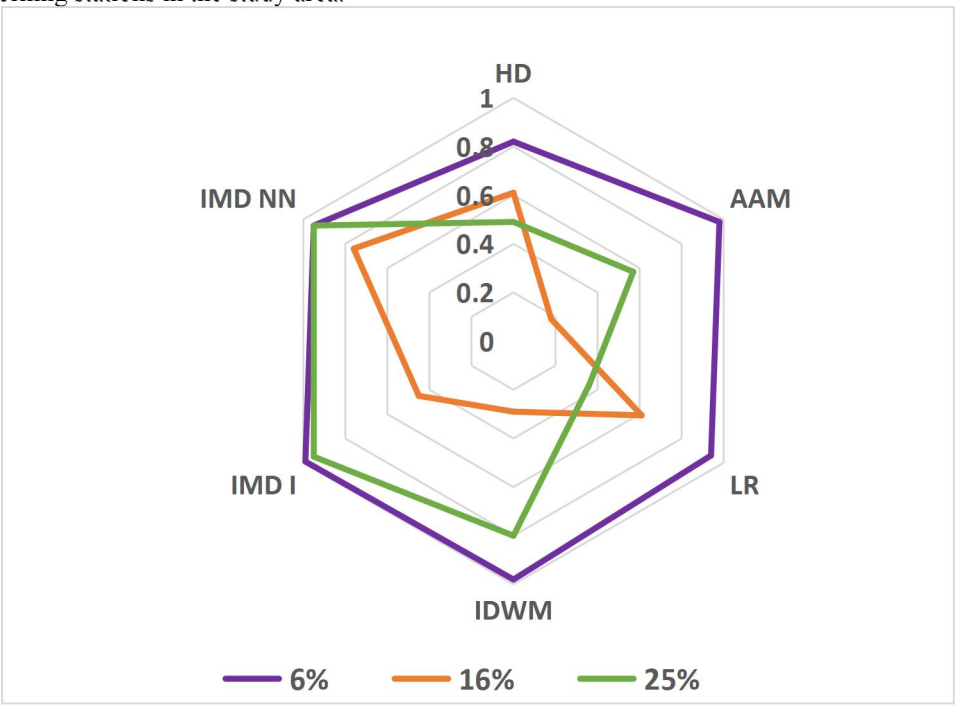


Figure 6. Radar graph of R^2 for different imputation methods with different % of missing values

Overall, using IMD gridded data leads to better performance than other data, simply because the IMD gridded data are generated from the combined influence of the station data. Given the superior performance of the gridded data, it may be used to substitute the missing records in the observational datasets (Meher and Das, 2019).

Imputation of missing rainfall data should be according to well-laid criteria. The distance between the stations plays an important role in providing high imputation accuracy. Especially in tropical countries like India, where rainfall tends to vary on a small spatial scale, closer stations yield better imputation accuracy. Selecting the techniques/methods of imputation is an individual/user's choice. Once the choice is made, its performance is highly affected by the percentage of missing data, the number of stations considered, the distance of the target station from the other stations, and finally, the spatial and temporal resolution of rainfall data of the gauge stations. The most frequently used methods are not necessarily the best ones since each dataset has its special characteristics (Pratama *et al*., 2016). In this study, the percentage of missing data significantly influences the imputation results and error metrics.

Furthermore, analysing the fluctuations and trends in precipitation over an extended period helps assess how these changes may affect agriculture, water resources, infrastructure, and ecosystems. Accurate rainfall data enable planners to estimate water availability, determine sustainable extraction rates, and plan for future needs, which is crucial as climate change alters precipitation patterns (Lettenmaier, 2024). In recent years, there have been growing concerns about the unpredictable and changing patterns of monsoon rainfall. Consequently, it is vital to develop strategies to mitigate potential risks and adapt to evolving climate conditions.

4. Conclusion

Station data frequently suffer from the problem of missing values. Preserving the spatiotemporal continuity of precipitation records necessitates a rigorous, percentage-dependent investigation into imputation methodologies. The efficacy of these techniques is not universal, it remains heavily constrained by the density of the rain gauge network, intrinsic geographical characteristics, regional climatic conditions, and the underlying drainage dynamics. In the case of HD, imputing the missing value is easiest when the percentage of missing values is small. Traditionally, the most popular method is IDWM. It considers both the distance ratio and the values of the nearest stations. However, the current study shows that IDWM is inefficient, as the distance between the stations is huge. The new methods IMDNN and IMDI turned out to be the best when there were 16% and 25% missing values, respectively.

Our study incorporated statistical measures to quantify the errors associated with these imputation methods. All the discussed methods are data sensitive, and each method is important in a particular set of data structures and the percentage of missing values. It is also established that gridded data is the best substitute for the missing station data. The IMD gridded data 0.25°×0.25° can potentially replicate many missing values with the least error. Thus, the proposed method IMDNN is recommended for 16% to 25% of missing values. The performance of the k-nearest neighbour in the IMD gridded data structure directly affects the reliability of classifications and predictions in resource management and environmental forecasting. However, further substantial exploration is needed before it can be applied widely.

Rainfall is a crucial component in the climate change study. Incomplete records and gaps in the modelling assumptions may result in uncertainty in hydrological predictions and inefficient management practices. Therefore, it is a precondition to analyse the imputation techniques and evaluate their implications and limitations. Furthermore, before applying and adapting any imputation method, it needs to be validated for the different conditions of missing values. Otherwise, it may lead to erroneous results and can adversely affect policymakers, farmers, and planning processes dedicated to water resource management.

Acknowledgment
We are grateful to the IMD for providing the rainfall data.

Declarations
Author Contribution PD, AB, and PB contributed to formulating the concept, design, acquisition, and interpretation of the data and writing of the manuscript. AB revised the manuscript critically, and MP revised the manuscript.
Data availability The IMD Grid data were acquired from the IMD's website (https://dsp.imdpune.gov.in/).

Funding
This research has not received any funding from any source.
The researchers have met the cost of the research.

Ethics Declarations
There are no ethical issues in this study.

Disclaimer
The contents and views expressed in this study are the views of the authors and do not necessarily reflect the views of the organisations they belong to

References

Addi, M., Gyasi-Agyei, Y., Obuobie, E., & Amekudzi, L. K.. Evaluation of imputation techniques for infilling missing daily rainfall records on river basins in Ghana. *Hydrological Sciences Journal*, *67*(4), 613–627 (2022). https://doi.org/10.1080/02626667.2022.2030868

Aieb A, Madani K, Scarpa M, Bonaccorso B, and Lefsih K 2019 A new approach for processing climate missing databases applied to daily rainfall data in Soummam watershed, Algeria. *Heliyon*, 5:1–27. https://doi.org/10.1016/j.heliyon.2019.e01247

Amin Burhanuddin S N Z, Deni S and Mohamed Ramli N 2017 Imputation of missing rainfall data using the revised normal ratio method. *Advanced Science Letters*, 23:10981–10985, 11. DOI: 10.1166/asl.2017.10203

Batista G E A P A and Monard M C 2003 An analysis of four missing data treatment methods for supervised learning. *Applied Artificial Intelligence*, 17:519-533. https://doi.org/10.1080/ 713827181

Chen, FW., Liu, CW. Estimation of the spatial rainfall distribution using inverse distance weighting (IDW) in the middle of Taiwan. *Paddy Water Environ* **10**, 209–222 (2012). https://doi.org/10.1007/s10333-012-0319-1

Chow, V. T. (1971). *Hand Book of Applied Hydrology*. McGraw-Hill Book Company.

Das P, Banik P, and Rath K C 2021 Precipitation extremes and anomalies of the Indian Sundarban 1984-2018. *MAUSAM*, 72(4):847-858. DOI: 10.54302/mausam.v72i4.3552

De Silva R, Dayawansa N and Ratnasiri M 2007 A comparison of methods used in estimating missing rainfall data. *Journal of Agricultural Sciences*, 3:101–108, 05. DOI: 10.4038/jas.v3i2.8107

Fouad K M, Ismail M M Azar A T, and M. M. Arafa 2021 Advanced methods for missing values imputation based on similarity learning. *Peer J Computer Science*, 7:1–38. https://doi.org/10.7717/peerj-cs.619

Ghosh A, Schmidt S, Fickert T, and Nusser M 2015 The Indian Sundar- ban mangrove forests: history, utilization, conservation strategies, and local perception. *Diversity*, 7(2):149–169. https://doi.org/10.3390/d7020149

Khosravi G, Nafarzadegan A R, Nohegar A, Fathizad H, and Malekian A 2015 A modified distance-weighted approach for filling annual precipitation gaps: Application to different climates of Iran. *Theoretical and Applied Climatology,* 119(07) 33–42. https://doi.org/10.1007/s00704-014-1091-5

Lai W 2019 A Study on Sequential *k*-Nearest Neighbour SKNN Imputation for Treating Missing Rainfall Data. *International Journal of Advanced Trends in Computer Science and Engineering*, 8:363–368, 06. DOI: 10.30534/ijatcse/2019/05832019

Lettenmaier D 2024 Role of Rainfall in Hydrological Modeling and Water Resource Planning. J Geogr Nat Disasters. 14:319. DOI: 10.35841/2167-0587.24.14.319

Liu Z G, Liu Y, Dezert J and Pan Q 2015 Classification of incomplete data based on belief functions and *k*-nearest neighbours. *Knowledge-Based Systems*, 89:113–125. DOI: 10.1016/j.knosys.2015.06.022

Lo Presti R, Barca E and Passarella G 2009 A methodology for treating missing data applied to daily rainfall data in the Candelaro River Basin (Italy). *Environmental Monitoring and Assessment*, 160:1–22, 01. DOI: 10.1007/s10661-008-0653-3

Meher J, Das L, Akhter J, Benestad R, and Mezghani A 2017 Performance of CMIP3 and CMIP5 GCMS to simulate observed rainfall characteristics over the Western Himalayan Region. *Journal of Climate*, 30(7777–7799), 09. https://doi.org/10.1175/JCLI-D-16-0774.1

Meher J and Das L 2019 Gridded data as a source of missing data replacement in station records. *Journal of Earth System Science*, 128:1–14, 04. https://doi.org/10.1007/s12040-019-1079-8

Pramanik M 2016 Assessment of the Impacts of Sea Level Rise on Mangrove Dynamics in the Indian part of Sundarbans using geospatial techniques. *Journal of Biodiversity, Bioprospecting and Development*, 14:117–127, 01. doi: 10.4172/2376-0214.1000155

Pratama I, Permanasari A E, Ardiyanto I, and Indrayani R. 2016 A review of missing values handling methods on time-series data. *International Conference on Information Technology Systems and Innovation (ICITSI)*, pages 1–6. doi: 10.1109/ICITSI.2016.7858189.

Ramos Calzado P, G´omez Camacho J, Perez-Bernal F and Pita M 2008 A novel approach to precipitation series completion in climatological datasets: Application to Andalusia. *International Journal of Climatology*, 28:1525 – 1534, 09. https://doi.org/10.1002/joc.1657

Rahman M G & Islam M Z 2013 kDMI: A Novel Method for Missing Values Imputation Using Two Levels of Horizontal Partitioning in a Data set. *International Conference on Advanced Data Mining and Applications*. https://doi.org/10.1007/978-3-642-53917-6_23

Rahman S A, Yuxiao H, Claassen J, Heintzman N and Kleinberg S 2015 Combining Fourier and Lagged k-Nearest Neighbor Imputation for Biomedical Time Series Data. *Journal of Biomedical Informatics*, 58:198–207, 10. https://doi.org/10.1016/j.jbi.2015.10.004Get rights and content

Rodríguez, R.; Pastorini, M.; Etcheverry, L.; Chreties, C.; Fossati, M.; Castro, A.; Gorgoglione, A. Water-Quality Data Imputation with a High Percentage of Missing Values: A Machine Learning Approach. Sustainability 2021, 13, 6318. https://doi.org/10.3390/su13116318

Sahana M, Ahmed R and Sajjad H 2016 Analyzing land surface temperature distribution in response to land use/land cover change using split window algorithm and spectral radiance model in Sundarban Biosphere Reserve, India. *Modelling Earth Systems and Environment*, 2:1–11, 05. https://doi.org/10.1007/s40808-016-0135-5

Sattari M T, Rezazadeh-Joudi A and Kusiak A 2016 Assessment of different methods for estimation of missing data in precipitation studies. *Hydrology Research*, 48(4) 032–1044, 09. https://doi.org/10.2166/nh.2016.364

Teegavarapu R and Chandramouli V 2005 Improved weighting methods, Deterministic and Stochastic Data-Driven Models for Estimation of Missing Precipitation Records. *Journal of Hydrology*, 312:191–206, 10. https://doi.org/10.1016/j.jhydrol.2005.02.015

WWF 2017 Sundarban in a Global Perspective: Long-Term Adaptation and Development -Discussion Paper (English).

Xia Y, Fabian P, Stohl A and Winterhalter M 1999 Forest climatology: estimation of missing values for Bavaria, Germany. *Agricultural and Forest Meteorology*, 96:131–144.