ILKOM Jurnal Ilmiah Vol. 18, No. 1, April 2026, pp. 97-108
Accredited 2nd by KEMDIKTISAINTEK No. 10/C/C3/DT.05.00/2025; E-ISSN 2548-7779 | P-ISSN 2087-1716

Research Article Open Access (CC–BY-SA)

Evaluating the Effectiveness of TBaWI for Imputation of Missing Rainfall Data

Lukman Syafie [a,1]; Narendra Awangga [a,2,*]; Yulita Salim [a,3]
[a] *Universitas Muslim Indonesia, Jl. Urip Sumoharjo Km.5, Makassar, 90231, Indonesia*
[1] *lukman.syafie@umi.ac.id;* [2] *13020220022@student.umi.ac.id;* [3]*yulita.salim@umi.ac.id*
* *Corresponding author*

Article history: Received December 24, 2025; Revised February 11, 2026; Accepted March 28, 2026; Available online April 20, 2026

Abstract

Daily rainfall data plays an important role in hydrological and climatological analysis, especially in tropical regions characterised by high rainfall variability and sharp seasonal changes. However, observational data often has gaps, which can reduce model accuracy and obscure relevant climatological signals. This study addresses these issues by applying the Trend-Based Window Imputation (TBaWI) method, an adaptive imputation approach that considers local temporal trends and seasonal dynamics in estimating missing rainfall values. This method was tested using CHIRPS data for the Makassar region for the period 2014–2023 with synthetic data loss scenarios of 10%, 15%, 20%, and 25%. The results show that TBaWI consistently provides a lower Mean Absolute Error (MAE) value, namely 6.14–7.65 mm, compared to linear interpolation, which produces 6.46–7.75 mm. The SMAPE value of TBaWI is also lower, for example 33.16% in the 15% data loss scenario, compared to interpolation at 35.06%. In addition, this method showed an improvement in the ability to identify dry days through the Zero Hit Rate (ZHR), which reached 60.08% in the 20% data loss scenario, higher than the interpolation of 58.32%, while the Rainy Hit Rate (RHR) remained in a stable range of 79–88%. These findings indicate that TBaWI is more effective in maintaining climatological consistency and numerical accuracy of tropical rainfall data. Further research is expected to integrate spatial aspects and optimise machine learning-based parameters to improve the generalisation of the method under various climatic conditions.

Keywords: Rainfall Imputation, Tbawi, Adaptive Window, CHIRPS, Tropical Rainfall, Missing Data

Introduction

Rainfall data plays a vital role in various fields, such as climate analysis, hydrological modeling, agricultural planning, as well as disaster risk mitigation [1], [2], [3], [4]. In tropical regions such as Indonesia, seasonal rainfall patterns that tend to be extreme make daily rainfall data an important element in the data-driven decision-making process [3], [5], [6]. As happened to the observation system in Makassar City, especially the Paotere Maritime Station which represents the coastal area and the BMKG Hasanuddin Station (Maros) which reflects the land area. These two stations play a strategic role in rainfall monitoring in western South Sulawesi, but often experience data gaps [7] . This kind of data loss not only creates bias in basic statistical analysis and precipitation trend tracking, but also has a direct impact on the accuracy of hydrological modeling as well as the effectiveness of early warning systems. In the context of increasingly real climate change, the availability of complete rainfall data is increasingly crucial. An increase in the frequency and intensity of extreme rainfall in various tropical regions, such as the southern part of Java Island, indicates a shift in the re-period of extreme events [8]. The lack of data in these conditions has the potential to obscure important climatological signals and reduce the accuracy of analyzing climate change and evaluating the risk of hydrometeorological disasters. Therefore, an imputation method is needed that is not only numerically accurate, but also capable of representing climatological patterns consistently, especially in tropical regions that have a high level of rainfall variability.

Efforts to overcome the problem of data loss are generally carried out using simple methods such as mean, median, Linear Interpolation or Naive Bayes [9], [10], [11]. This approach is popular for its ease of implementation, but it has limitations in representing the discrete, seasonal, and highly volatile characteristics of tropical rainfall [12], [13], [14]. In some cases, very low or zero rainfall values also pose a challenge in estimation, as they can lead to zero probability in the classification model, so smoothing techniques are needed to maintain the accuracy of the estimate [15]. The interpolative method also often produces over-smoothed values, thus failing to capture the sharp transition between

http://dx.doi.org/10.33096/ilkom.v18i1.3273.97-108

dry days and high-intensity rainy days [16], [17], [18], [19]. In addition, model-based approaches such as machine learning and artificial neural networks often require large training data and high computing resources, which are not always available at the observation site [20], [21], [22], [23], [24]. Nevertheless, some lightweight learning approaches such as Gaussian Naïve Bayes or simple neural networks can be applied to small datasets efficiently without the need for special hardware. Although a number of studies have examined statistical model-based imputation techniques as well as adaptive interpolation, studies that explicitly consider the seasonal dynamics typical of the tropics and maintain climatological balance in the missing data are still limited [25],[26].

To answer this gap, this study introduces Trend-Based Window Imputation (TBaWI), a Trend-Based imputation method with seasonally adaptive windows for daily rainfall in tropical settings that leverages local trend direction and strength to impute missing values, improving accuracy while preserving wet–dry occurrence patterns. The method is validated on CHIRPS data under multiple missingness levels using MAE, SMAPE, ZHR, RHR, and Pearson correlation, and is benchmarked against linear interpolation.

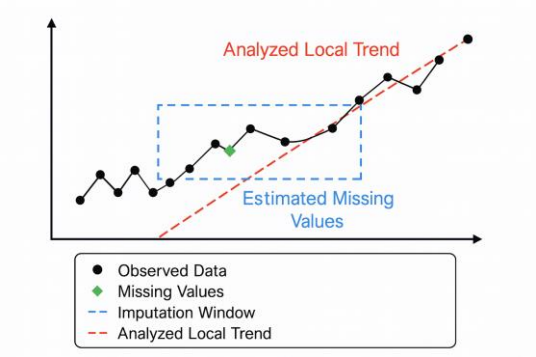

Figure 1. Illustration of the TBaWI Method

Figure 1 illustrates the basic mechanism of TBaWI in estimating lost rainfall values by utilising local trends around the data loss points. Black dots indicate actual observation data, while green dots mark lost values. The blue area represents the imputation window used to analyse data change patterns around the missing points. The red line shows the analysed local trend calculated based on valid data within the window. The missing values are then estimated following the direction and slope of the local trend so that the imputation results remain consistent with the dynamics of rainfall changes around the missing point.

This method combines three main components, namely (1) Kalman Filter to reduce noise and maintain signal stability, (2) Gaussian Masked Smoothing to maintain temporal continuity, and (3) adaptive window based on local variance to global variance ratio to dynamically adjust the estimated size. The purpose of this study is to evaluate the effectiveness of TBaWI in imputing tropical daily rainfall data compared to conventional interpolative methods.

Method

This research began with the collection of daily rainfall data from the Climate Hazards Group InfraRed Precipitation with Station (CHIRPS) satellite. Synthetic data loss scenarios of 10%, 15%, 20%, and 25% were then integrated with the original dataset for comparison at the evaluation stage. The combined dataset then underwent comprehensive pre-processing, which included outlier detection using Modified Z-score, adaptive filtering through Kalman Filter, and temporal smoothing using Gaussian Smoothing. The imputation stage was performed using the TBaWI method, which estimates missing values based on temporal dynamics and seasonal context“The performance of this method was evaluated using MAE, modified SMAPE, ZHR, RHR, and Pearson correlation (r). A summary of the complete analytical workflow from data acquisition, scenario construction, pre-processing, imputation, to evaluation is systematically described in **Figure 2**.

.

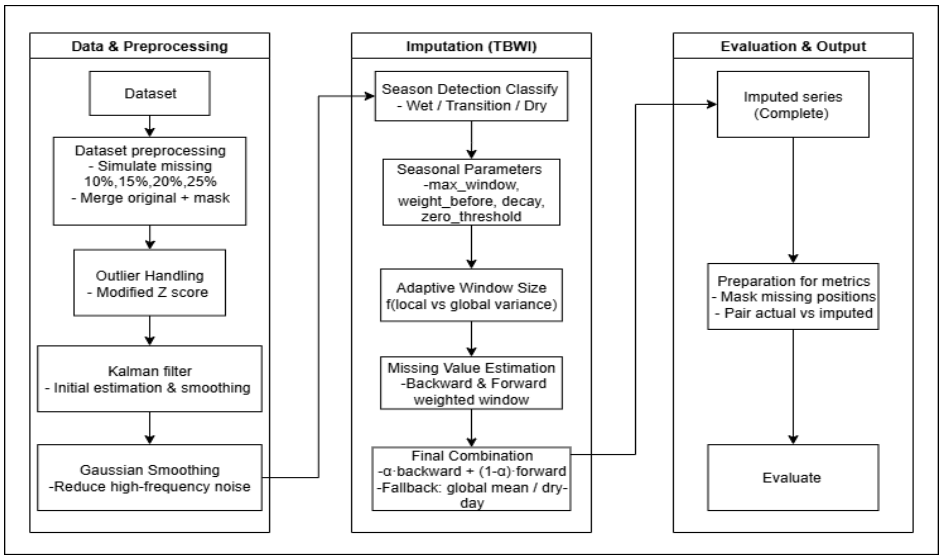

Figure 2. Research design

In this paper, rainfall at day t is denoted by x_t (mm). The observed time series containing missing values is y_t, where y_t may be NaN at missing positions. The imputed estimate is denoted by x̂_t. A day is considered "dry" if rainfall is below r_0, where r_0 = 0.1 mm. Seasonal parameters include maximum window size w_max, minimum window size w_min, seasonal backward weight w_b (weight_before), exponential decay factor δ (weight_decay), and seasonal zero tolerance θ_0 (zero_threshold).

A. Data Collections

The dataset used in this study comes from the CHIRPS satellite at the coordinates of Makassar City (119.41° E, 5.13° LS) with a spatial resolution of 0.05° and a daily temporal resolution. The data period used covers the years 2014 to 2023, then a simulation of synthetic data loss is carried out with four scenarios of the proportion of data lost, namely 10%, 15%, 20%, and 25%. Each missing scenario is randomly generated to represent the state of data loss due to sensor interference or observation errors. The dataset consists of three main columns, namely Date as the time of recording, RR_Original as the actual rainfall value, and Missing as a column containing values that are randomly deleted according to the data loss scenario. The selection of CHIRPS is based on the completeness of the data and its good validation of surface observations in the territory of Indonesia [27], [28].

B. Data Pre-processing

The pre-processing stage is designed to (i) suppress abnormal extreme values (outliers) and (ii) produce a more stable series to support the subsequent imputation process. The output of this stage is denoted by z_t, which is used as the main input for estimating the imputed values $\hat{x}_t$.

To distinguish observed and missing positions, an observation indicator is defined as:

$$m_t = \begin{cases} 1, & y_t \text{ is observed} \\ 0, & y_t \text{ is missing (NaN)} \end{cases} \quad (1)$$

a. Outlier Treatment using Modified Z-score

Outliers are identified using the modified Z-score based on the median and the MAD (Median Absolute Deviation). The median of observed values is:

$$\tilde{y} = \text{median}\{\, y_t \mid m_t = 1 \,\} \quad (2)$$

and the MAD is computed as:

$$MAD = \text{median}\{ \quad \mid y_t - \tilde{y} \mid \quad \mid m_t = 1 \,\} \quad (3)$$

The modified Z-score is defined by:

$$Z_t^* = 0.6745 \frac{| y_t - \tilde{y} |}{MAD} \quad (4)$$

Following the implementation, a data point is flagged as an outlier when:

$$Z_t^* > 3.5$$

When an outlier is detected, its value is replaced by a local median computed within a ± 3-day neighborhood (ignoring NaN values in the window):

$$y_t^{(1)} = median\,\{y_r \mid \tau \in [t-3, t+3], m_t = 1\} \quad (5)$$

For non-outlier positions, values remain unchanged ($y_t^{(1)} = y_t$).

b. Adaptive 1D Kalman Filtering (Optional, Missing-only Update)

To obtain a smoother and more robust series against random fluctuations, a one-dimensional Kalman filter is applied using a random-walk model with measurement matrix $H = 1$. The noise parameters are set adaptively from the data variance:

$$\sigma^2 = \mathrm{Var}\left\{y_t^{(1)}\right\} \quad (6)$$

$$Q = \mathrm{clip}\left(\frac{\sigma^2}{100},\ 0.05,\ 0.30\right), \qquad R = \mathrm{clip}\left(\frac{\sigma^2}{10},\ 0.50,\ 2.00\right) \quad (7)$$

Let $k_t$denote the Kalman-filtered output at time t. Consistent with the proposed design, Kalman filtering is used only to fill missing locations in the original series, while observed values are preserved:

$$y_t^{(2)} = m_t\, y_t^{(1)} + (1 - m_t)\, k_t \quad (8)$$

c. Masked Gaussian Smoothing (Missing-only Update)

After the Kalman step (when enabled), Gaussian smoothing is employed to produce a more gradual day-to-day pattern. To prevent missing entries from biasing the smoothing operation, a masked Gaussian smoothing scheme is used with:

$$\sigma_g = 0.3$$

Let G_{σ_g}be the 1D Gaussian kernel and $*$denote convolution. The masked-smoothed value $g_t$is computed as:

$$g_t = \frac{(G_{\sigma_g} * (y^{(2)} \cdot m))_t}{\left(G_{\sigma_g} * m\right)_t} \quad (9)$$

Similar to the Kalman step, Gaussian smoothing is applied only at the originally missing indices, ensuring that observed values remain intact. The final pre-processed series is defined by:

$$z_t = m_t\, y_t^{(2)} + (1 - m_t)\, g_t \quad (10)$$

The resulting series $z_t$is subsequently used to determine the adaptive window and to estimate the imputed rainfall $\hat{x}_t$.

C. Imputation with TBaWI

After the preprocessing stage, missing values in the observed rainfall series $y_t$are imputed using the TBaWI approach. Unlike conventional interpolation with fixed neighborhoods, TBaWI adapts its estimation behavior by considering (i) seasonal characteristics and (ii) local variability around the missing index. Let $z_t$denote the pre-processed rainfall series obtained from the previous stage, and let $\hat{x}_t$denote the imputed estimate at time t. A day is considered dry when rainfall is below r_0, where $r_0 = 0.1$mm.

a. Season Detection

A key component of TBaWI is the seasonal-aware configuration. The seasonal category is determined automatically from the month of observation using a month-dependent rainfall probability $p(m)$. The season category $s(t)$is assigned by the following rule:

- Heavy wet season if $p(m) \geq 0.70$
- Light wet season if $0.45 \leq p(m) < 0.70$
- Transitional dry if $0.30 \leq p(m) < 0.45$
- Heavy dry season if $p(m) < 0.30$

The classification of the seasons can be seen in **Table 1**.

Table 1. Month-based rainfall probability and season category

Month	Probability of Rain	Season Category
Jan	0.85	Heavy Wet
Feb	0.80	Heavy Wet
Mar	0.70	Heavy Wet
Apr	0.55	Light Wet
May	0.35	Transitional Dry
Jun	0.20	Heavy Dry
Jul	0.10	Heavy Dry
Aug	0.10	Heavy Dry
Sep	0.25	Heavy Dry
Oct	0.45	Light Wet
Nov	0.65	Light Wet
Dec	0.80	Heavy Wet

The assigned season $s(t)$controls the configuration of imputation parameters, including window size and weighting behavior. This design enables the imputation process to better follow typical tropical rainfall dynamics.

b. Season Configuration

Once the season category is determined, TBaWI assigns a season-specific parameter set. The regulated parameters include:

- maximum window size $w_{\max}$,
- backward-side weight w_b(weight assigned to the "before" side),
- exponential decay factor δ,
- zero tolerance threshold θ_0(used in dry-pattern checking).

Each season has its own configuration as summarized in **Table 2**.

Table 2. Seasonal parameter configuration

Season	$w_{\max}$	w_b	δ	θ_0
Heavy Wet	5	0.60	0.80	0.10
Light Wet	4	0.55	0.85	0.20
Trans. Dry	4	0.50	0.90	0.30
Heavy Dry	3	0.45	0.95	0.40

A minimum window size $w_{\min}$ is enforced for all seasons (in this study, $w_{\min} = 2$).

c. Adaptive Window Size Estimation

The adaptive window size ($W_{adaptive}$) in the TBaWI method is determined based on the comparison between local variance and global variance of rainfall data. Local variance (σ^2_{local}) is calculated from valid data

around the missing point by considering the forward (after) and backward (before) directions separately. Meanwhile, the global variance (σ^2_{global}) is calculated from the entire time series of pre-processed rainfall data so that it reflects the general stability of rainfall patterns during the observation period.

The adaptive window size in TBaWI is determined by comparing local variance around the missing point with the global variance of the entire pre-processed series. The variance of a sample is computed as:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(u_i - \bar{u})^2 \tag{11}$$

where $u_i$are sample observations and $\bar{u}$is the sample mean.

For a missing index t, local variability is measured separately on the backward and forward sides (within the seasonal maximum window). Let $\sigma^2_{\text{bwd}}(t)$and $\sigma^2_{\text{fwd}}(t)$denote the local variances computed from valid values of $z_t$in the backward and forward neighborhoods, respectively. To accommodate asymmetry, a directional combination weight $w_r$is defined as:

$$w_r = \begin{cases} 0.7, & \sigma^2_{\text{fwd}}(t) > 2\,\sigma^2_{\text{bwd}}(t) \\ 0.3, & \sigma^2_{\text{bwd}}(t) > 2\,\sigma^2_{\text{fwd}}(t) \\ 0.5, & \text{otherwise} \end{cases} \tag{12}$$

The combined local variance is then computed by:

$$\sigma^2_{\text{local}}(t) = w_r\,\sigma^2_{\text{bwd}}(t) + (1 - w_r)\,\sigma^2_{\text{fwd}}(t) \tag{13}$$

Let σ^2_{global}be the variance of the full pre-processed series $\{z_t\}$. The variance ratio is bounded to prevent extreme shrinking:

$$\rho(t) = min\left(0.95, \frac{\sigma^2_{\text{local}}(t)}{\sigma^2_{\text{global}}}\right) \tag{14}$$

Finally, the adaptive window size is computed as:

$$W_{adaptive} = \max\{W_{min}, min\{W_{max}, W_{max}(1 - \rho(t)^{0.7})\}\} \tag{15}$$

where $w_{\max}$is selected from **Table 2** based on the season $s(t)$.

d. Missing Value Estimation Strategy

After determining $w(t)$, the missing rainfall is estimated using valid values around t. The window is split into backward and forward parts using the seasonal weight w_b:

$$w^-(t) = \lfloor w(t)\,w_b \rfloor, w^+(t) = w(t) - w^-(t) \tag{16}$$

Let $n_b$and $n_a$denote the numbers of valid (non-NaN) samples available on the backward and forward sides, respectively.

e. Dry-pattern (zero-pattern) check

To avoid generating false rainfall during dry conditions, TBaWI computes the proportions of near-zero values (below r_0) on both sides:

$$Z_{\text{before}}(t) = \frac{1}{n_b}\sum_{i=1}^{n_b}\mathbf{1}[z_{t-i} < r_0], Z_{\text{after}}(t) = \frac{1}{n_a}\sum_{j=1}^{n_a}\mathbf{1}[z_{t+j} < r_0] \tag{17}$$

If both proportions exceed the seasonal tolerance threshold θ_0, the missing value is set to zero:

$$Z_{before}(t) > \theta_0 \wedge Z_{after}(t) > \theta_0) \Rightarrow \hat{x}\_t = 0 \tag{18}$$

f. Exponential weighting on both sides

If the dry-pattern condition is not met, estimation is performed using exponential decay weighting with factor δ. The backward-side estimate is:

$$\hat{x}_{\text{before}}(t) = \frac{\sum_{i=1}^{n_b} \delta^{\,n_b-i+1}\, z_{t-i}}{\sum_{i=1}^{n_b} \delta^{\,n_b-i+1}} \tag{19}$$

and the forward-side estimate is:

$$\hat{x}_{\text{after}}(t) = \frac{\sum_{j=1}^{n_a} \delta^{\,j-1}\, z_{t+j}}{\sum_{j=1}^{n_a} \delta^{\,j-1}} \tag{20}$$

The final imputed value is obtained by combining both sides using w_b:

$$\hat{x}_t = w_b\, \hat{x}_{\text{before}}(t) + (1 - w_b)\, \hat{x}_{\text{after}}(t) \tag{21}$$

If only one side provides valid samples, the estimate uses that side only. If both sides have no valid samples, $\hat{x}_t$falls back to the mean of $\{z_t\}$(or 0if undefined). Finally, the estimate is constrained to be non-negative:

$$\hat{x}_t = \max(0, \hat{x}_t) \tag{22}$$

D. Evaluation of imputation results

To evaluate the performance of TBaWI in reconstructing missing rainfall values, a quantitative assessment is conducted by comparing the imputed estimates $\hat{x}_i$against the corresponding ground-truth values $x_i$only at the indices that are simulated as missing. The evaluation emphasizes not only numerical accuracy but also the preservation of tropical rainfall characteristics, particularly the correct identification of dry and rainy days using the dry-day threshold r_0(with $r_0 = 0.1$mm).

Let n be the number of evaluated missing points :

a. Modified Symmetric Mean Absolute Percentage Error (modified SMAPE)

A stabilized and modified SMAPE is used to avoid undefined behavior near zero rainfall. With stabilizer $\epsilon = 0.5$, the metric is defined as :

$$SMAPE_{\varepsilon} = \left(\frac{100}{n}\right) \sum_{i=1}^{n} \begin{cases} 0, & if\ (x_i < \varepsilon) \wedge (\hat{x}_i < \varepsilon); \\ \dfrac{|x_i - \hat{x}_i|}{|x_i| + |\hat{x}_i| + \varepsilon}, & otherwise \end{cases} \tag{23}$$

with:

- x_i: actual rainfall value (mm),
- $\hat{x}_i$: imputation result value (mm),
- n: the amount of data evaluated,
- $\epsilon = 0.5$

Smaller SMAPE values indicate more accurate estimates.

b. Mean Absolute Error (MAE)

MAE is used to measure the average of the absolute difference between the actual value and the imputation result [29], [30]. This metric provides an intuitive numerical error measure in the same unit as rainfall data.

$$MAE = \frac{1}{n} \cdot \sum_{i=1}^{n} |\, x_i - \hat{x}_i \,| \tag{24}$$

The smaller the MAE value, the better the method performance in minimizing estimation errors.

c. Zero Hit Rate (ZHR)

The ZHR measures the method's ability to recognize days without rain (rainfall value < 0.1 mm). This metric is calculated as the proportion of the number of dry days that are successfully correctly predicted against the actual total dry days.

$$ZHR = \frac{TP_{dry}}{N_{dry}} \cdot 100\% \tag{25}$$

with:

- TP_{dry}: the number of actual dry days that are also predicted to be dry,
- N_{dry}: total actual dry days (value $x_i < 0.1$).

The higher the ZHR value, the better the model's ability to maintain a representation of a dry day.

d. Rainy Hit Rate (RHR)

The RHR measures the method's ability to recognize rainy days (rainfall value ≥ 0.1 mm). It is calculated as the percentage of actual rainy days that are predicted to rain.

$$RHR = \frac{TP_{rain}}{N_{rain}} \cdot 100\% \tag{26}$$

with:

- TP_{rain}: the number of actual rainy days that are also predicted to rain,
- N_{rain}: Total Actual Rainy Day (value $x_i \geq 0.1$).

ZHR and RHR are jointly used to assess the model's ability to maintain a balance between dry and wet events.

e. Pearson Correlation (r)

Pearson correlation is used to assess the strength and direction of the linear relationship between the imputation result data and the actual data.

$$r = \frac{\mathrm{Cov}(x, \hat{x})}{\sigma_x \cdot \sigma_{\hat{x}}} \tag{27}$$

with:

- $\mathrm{Cov}(x, \hat{x})$: the covariance between the actual value and the imputation result,
- $\sigma_x, \sigma_{\hat{x}}$: the standard deviation of each variable.

Values closer to 1 indicate stronger agreement in temporal variation.

This evaluation is conducted only at the indices that are simulated as missing, so the reported metrics directly reflect the method's capability to reconstruct values that are unavailable in the observed series. The metrics are used jointly to provide a comprehensive assessment: MAE measures absolute error, modified SMAPE captures proportional error under near-zero rainfall conditions, ZHR and RHR evaluate the preservation of dry–wet event patterns that are critical in tropical hydrological analysis, and Pearson's correlation indicates whether temporal rainfall variability is consistently maintained after imputation.

Results and Discussion

A. Results

The performance evaluation of TBaWI was conducted using CHIRPS daily rainfall data for the 2014–2023 period under four synthetic missingness levels (10%, 15%, 20%, and 25%). The imputation results were benchmarked against linear interpolation as a baseline, and the comparative performance is summarized in **Table 3**.

Table 3. Average Results of TBaWI Evaluation and Linear Interpolation (2014–2023)

Missing (%)	Method	MAE	SMAPE	ZHR (%)	RHR (%)	r
10	TBaWI	7.65	33.59	66.29	84.88	0.52
	Linear Interpolation	7.75	34.32	63.03	86.36	0.52
15	TBaWI	6.14	33.16	66.61	79.89	0.56
	Linear Interpolation	6.46	35.06	62.37	84.25	0.57
20	TBaWI	6.31	34.93	60.08	88.06	0.56
	Linear Interpolation	6.50	35.80	58.32	91.55	0.58
25	TBaWI	6.23	34.99	59.51	86.10	0.51
	Linear Interpolation	6.67	35.25	62.07	84.94	0.50

Overall, TBaWI consistently achieves lower MAE and SMAPE than linear interpolation across all missingness scenarios. In terms of MAE, TBaWI improves over linear interpolation by approximately 1.29% (10%), 4.95% (15%), 2.92% (20%), and 6.60% (25%). Similarly, SMAPE is reduced by about 2.13% (10%), 5.42% (15%), 2.43% (20%), and 0.74% (25%), indicating more stable magnitude estimates under the highly skewed and intermittent nature of tropical rainfall.

For occurrence-based metrics, TBaWI yields higher ZHR than linear interpolation at 10–20% missingness (e.g., 66.61% vs 62.37% at 15%), suggesting better preservation of dry-day (zero rainfall) structure in these scenarios. However, at 25% missingness, ZHR is lower for TBaWI (59.51%) than for linear interpolation (62.07%), indicating that dry-day preservation becomes more challenging under heavier data loss. In contrast, linear interpolation shows higher RHR at 10–20% missingness, whereas at 25% missingness TBaWI becomes slightly higher (86.10% vs 84.94%). Pearson correlation values remain comparable for both methods (approximately 0.50–0.58), showing similar agreement in temporal variability.

B. Discussion

The consistent improvements in MAE and SMAPE can be attributed to TBaWI's adaptive window mechanism and seasonal weighting, which prioritize locally relevant temporal information and reduce the influence of less representative points during periods with different rainfall regimes. This design is particularly important for tropical rainfall time series, where variability is sharp, episodic, and highly seasonal.

The occurrence metrics (ZHR and RHR) reveal a meaningful trade-off between preserving dry-day structure and rainy-day occurrence, depending on the missingness level. At 10–20% missingness, the higher ZHR obtained by TBaWI indicates that the zero-pattern check helps reduce false non-zero ("light-rain") imputations on truly dry days, which is hydrologically relevant because dry days typically dominate tropical rainfall records. Meanwhile, linear interpolation tends to produce small non-zero values between surrounding points, which explains its higher RHR at 10–20% missingness—interpolation more readily classifies days as rainy even when the true occurrence is uncertain.

At 25% missingness, TBaWI maintains stronger magnitude accuracy (lower MAE and SMAPE) and slightly higher RHR, but its ZHR drops below the baseline. This suggests that under heavier data loss, the occurrence decision becomes less reliable and the balance between preserving dry-day zeros (ZHR) and non-zero events (RHR) can shift. Nevertheless, the correlation values remain similar between methods, indicating that both approaches capture comparable overall temporal co-variation, while TBaWI provides more consistent gains in magnitude accuracy and competitive preservation of the wet–dry event structure.

Conclusion

This study demonstrates that TBaWI imputes tropical daily rainfall data more accurately and consistently than linear interpolation, while better preserving wet–dry occurrence patterns. The combination of seasonally adaptive windows, dry-pattern checks, and seasonal weighting enables TBaWI to produce stable estimates that remain consistent with the climatological characteristics of the study area. A key limitation is that TBaWI relies on empirical parameter settings, is univariate in nature, and is primarily evaluated under synthetic missingness (random masking), which may not fully represent real-world missing-data mechanisms. Future work should incorporate spatial information and automatic parameter optimisation to improve generalisability across diverse climatic conditions. Practically, TBaWI is well-suited for reconstructing daily rainfall time series in tropical regions where preserving dry–wet event structure is critical for hydrological analysis and downstream modelling.

Acknowledgement

The author would like to thank the Climate Hazards Group for the availability of the CHIRPS dataset which is the main source of rainfall data [31]. Thanks, were also expressed to the team for providing technical and conceptual input during the algorithm development process and evaluation of the TBaWI method.

References:

[1] M. Addi, Y. Gyasi-Agyei, E. Obuobie, And L. K. Amekudzi, "Evaluation Of Imputation Techniques For Infilling Missing Daily Rainfall Records On River Basins In Ghana," *Hydrological Sciences Journal*, Vol. 67, No. 4, Pp. 613–627, Mar. 2022, Doi: 10.1080/02626667.2022.2030868.

[2] R. L. Prasetio, A. Zakaria, A. Ashruri, And S. Sumiharni, "Analisis Data Curah Hujan Yang Hilang Dengan Menggunakan Metode Normal Ratio, Inversed Square Distance, Rata - Rata Aljabar Dan Metode Modifikasi," *Jurnal Rekayasa Sipil Dan Desain*, Vol. 9, No. 3, Pp. 559–570, Nov. 2021, Doi: 10.23960/Jrsdd.V9i3.1992.

[3] A. Wangwongchai, M. Waqas, P. Dechpichai, P. T. Hlaing, S. Ahmad, And U. W. Humphries, "Imputation Of Missing Daily Rainfall Data; A Comparison Between Artificial Intelligence And Statistical Techniques," *MethodsX*, Vol. 11, P. 102459, Dec. 2023, Doi: 10.1016/J.Mex.2023.102459.

[4] V. D. Banda, R. B. Dzwairo, S. K. Singh, And T. Kanyerere, "Hydrological Modelling And Climate Adaptation Under Changing Climate: A Review With A Focus In Sub-Saharan Africa," *Water (Basel).*, Vol. 14, No. 24, P. 4031, Dec. 2022, Doi: 10.3390/W14244031.

[5] F. R. Muhammad, S. W. Lubis, And S. Setiawan, "Impacts Of The Madden–Julian Oscillation On Precipitation Extremes In Indonesia," *International Journal Of Climatology*, Vol. 41, No. 3, Pp. 1970–1984, Mar. 2021, Doi: 10.1002/Joc.6941.

[6] A. Mulsandi, Y. Koesmaryono, R. Hidayat, A. Faqih, And A. Sopaheluwakan, "On The Interannual Variability Of Indonesian Monsoon Rainfall (Imr): A Literature Review Of The Role Of Its External Forcing," *Jurnal Meteorologi Dan Geofisika*, Vol. 24, No. 2, Pp. 115–127, Mar. 2024, Doi: 10.31172/Jmg.V24i2.1049.

[7] H. Darwis And F. Umar, "Precipitation Missing Data Prediction Using Recommendation System," *International Journal Of Scientific & Technology Research*, Vol. 8, No. 11, 2019, [Online]. Available: Www.Ijstr.Org

[8] Giarno *Et Al.*, "Changing Of Return Periods Of Extreme Rainfall Using Satellite Observation In Java Island," *E3s Web Of Conferences*, Vol. 464, P. 11005, Dec. 2023, Doi: 10.1051/E3sconf/202346411005.

[9] L. Syafie, F. Umar, A. Mude, H. Darwis, Herman, And Harlinda, "Missing Data Handling Using The Naive Bayes Logarithm (Nbl) Formula," In *2018 2nd East Indonesia Conference On Computer And Information Technology (Eiconcit)*, Ieee, Nov. 2018, Pp. 318–321. Doi: 10.1109/Eiconcit.2018.8878538.

[10] R. Rodríguez *Et Al.*, "Water-Quality Data Imputation With A High Percentage Of Missing Values: A Machine Learning Approach," *Sustainability*, Vol. 13, No. 11, P. 6318, Jun. 2021, Doi: 10.3390/Su13116318.

[11] S. M. Shaharudin, "Imputation Methods For Addressing Missing Data Of Monthly Rainfall In Yogyakarta, Indonesia," *International Journal Of Advanced Trends In Computer Science And Engineering*, Vol. 9, No. 1.4, Pp. 646–651, Sep. 2020, Doi: 10.30534/Ijatcse/2020/9091.42020.

[12] L. V. Duarte, K. T. M. Formiga, And V. A. F. Costa, "Comparison Of Methods For Filling Daily And Monthly Rainfall Missing Data: Statistical Models Or Imputation Of Satellite Retrievals?," *Water (Basel).*, Vol. 14, No. 19, P. 3144, Oct. 2022, Doi: 10.3390/W14193144.

[13] Z. Sa'adi *Et Al.*, "Evaluating Imputation Methods For Rainfall Data Under High Variability In Johor River Basin, Malaysia," *Applied Computing And Geosciences*, Vol. 20, P. 100145, Dec. 2023, Doi: 10.1016/J.Acags.2023.100145.

[14] S. Mariana Che Mat Nor, S. M. Shaharudin, S. Ismail, N. H. Zainuddin, And M. L. Tan, "A Comparative Study Of Different Imputation Methods For Daily Rainfall Data In East-Coast Peninsular Malaysia," *Bulletin Of Electrical Engineering And Informatics*, Vol. 9, No. 2, Apr. 2020, Doi: 10.11591/Eei.V9i2.2090.

[15] L. Syafie, D. Indra, And Yuhandry, "Modifikasi Rumus Bayesian Network Untuk Klasifikasi Dokumen," Oct. 2015.

[16] G. V. Nguyen, X.-H. Le, L. Nguyen Van, D. T. T. May, S. Jung, And G. Lee, "Machine Learning Approaches For Reconstructing Gridded Precipitation Based On Multiple Source Products," *J. Hydrol. Reg. Stud.*, Vol. 48, P. 101475, Aug. 2023, Doi: 10.1016/J.Ejrh.2023.101475.

[17] B. Dey *Et Al.*, "Monitoring Groundwater Potential Dynamics Of North-Eastern Bengal Basin In Bangladesh Using Ahp-Machine Learning Approaches," *Ecol. Indic.*, Vol. 154, P. 110886, Oct. 2023, Doi: 10.1016/J.Ecolind.2023.110886.

[18] O. Mudele, F. M. Bayer, L. F. R. Zanandrez, A. E. Eiras, And P. Gamba, "Modeling The Temporal Population Distribution Of Ae. Mosquito Using Big Earth Observation Data," *Ieee Access*, Vol. 8, Pp. 14182–14194, 2020, Doi: 10.1109/Access.2020.2966080.

[19] L. E. Alejo-Sanchez *Et Al.*, "Missing Data Imputation Of Climate Time Series: A Review," *MethodsX*, Vol. 15, P. 103455, Dec. 2025, Doi: 10.1016/J.Mex.2025.103455.

[20] C. Li, X. Ren, And G. Zhao, "Machine-Learning-Based Imputation Method For Filling Missing Values In Ground Meteorological Observation Data," *Algorithms*, Vol. 16, No. 9, P. 422, Sep. 2023, Doi: 10.3390/A16090422.

[21] M. Alabadla *Et Al.*, "Systematic Review Of Using Machine Learning In Imputing Missing Values," *Ieee Access*, Vol. 10, Pp. 44483–44502, 2022, Doi: 10.1109/Access.2022.3160841.

[22] B. O'sullivan And G. Kelly, "Efficient Likelihood And Machine-Learning Models For Spatiotemporal Rainfall Estimation And Imputation," *International Journal Of Climatology*, Aug. 2025, Doi: 10.1002/Joc.70072.

[23] M. Toure *Et Al.*, "Machine Learning Approaches For Imputing Missing Meteorological Data In Senegal," *Applied Computing And Geosciences*, Vol. 27, P. 100281, Sep. 2025, Doi: 10.1016/J.Acags.2025.100281.

[24] Herman *Et Al.*, "Comparison Of Artificial Neural Network And Gaussian Naïve Bayes In Recognition Of Hand-Writing Number," In *2018 2nd East Indonesia Conference On Computer And Information Technology (Eiconcit)*, Ieee, Nov. 2018, Pp. 276–279. Doi: 10.1109/Eiconcit.2018.8878651.

[25] S. Pinthong *Et Al.*, "Imputation Of Missing Monthly Rainfall Data Using Machine Learning And Spatial Interpolation Approaches In Thale Sap Songkhla River Basin, Thailand," *Environmental Science And Pollution Research*, Vol. 31, No. 41, Pp. 54044–54060, Sep. 2022, Doi: 10.1007/S11356-022-23022-8.

[26] G. R. Arathy Nair, S. Adarsh, A. El-Shafie, And A. N. Ahmed, "Enhancing Hydrological Data Completeness: A Performance Evaluation Of Various Machine Learning Techniques Using Probabilistic Fusion Imputer With Neural Networks For Streamflow Data Reconstruction," *J. Hydrol. (Amst).*, Vol. 639, P. 131583, Aug. 2024, Doi: 10.1016/J.Jhydrol.2024.131583.

[27] J. Suryanto, Amprin, And Anisum, "Validasi Curah Hujan Harian Chirps Precipitation Satellite Product Di Provinsi Kalimantan Barat," *Jurnal Ilmiah Rekayasa Pertanian Dan Biosistem*, Vol. 11, No. 1, Pp. 73–88, Mar. 2023, Doi: 10.29303/Jrpb.V11i1.442.

[28] M. Shahid *Et Al.*, "Assessing The Potential And Hydrological Usefulness Of The Chirps Precipitation Dataset Over A Complex Topography In Pakistan," *Hydrological Sciences Journal*, Vol. 66, No. 11, Pp. 1664–1684, Aug. 2021, Doi: 10.1080/02626667.2021.1957476.

[29] K. Seu, M.-S. Kang, And H. Lee, "An Intelligent Missing Data Imputation Techniques: A Review," *Joiv : International Journal On Informatics Visualization*, Vol. 6, No. 1–2, P. 278, May 2022, Doi: 10.30630/Joiv.6.1-2.935.

[30] M. Mastika, T. Siswantining, And A. Bustamam, "Comparison Of Missing Value Imputation Using Mean, Bayesian Knn, And Non-Bayesian Knn On Tep Gene Expression Data," *Media Statistika*, Vol. 18, No. 1, Pp. 61–72, Oct. 2025, Doi: 10.14710/Medstat.18.1.61-72.

[31] Climate Hazards Center (CHC), UC Santa Barbara, "CHIRPS: Rainfall Estimates from Rain Gauge and Satellite Observations," [Online]. Available: https://www.chc.ucsb.edu/data/chirps. Accessed: 12 Nov. 2025.