COMPARATIVE ANALYSIS OF MISSING DATA IMPUTATION METHODS FOR FLOOD FEATURES FROM LANGAT RIVER IN SELANGOR, MALAYSIA

Ainaa Hanis Zuhairi[1], Fitri Yakub[2*], Aizul Nahar Harun[3], Mas Omar[4], Muhamad Sharifuddin[5], Amrul Faruq[6], Vijay Sinha[7], Khamarrul Azahari Razak[8], Shahrum Shah Abdullah[9]

[1,2,3,4,5,8,9]Malaysia-Japan International Institute of Technology, Universiti Teknologi Malaysia, 54100 Kuala Lumpur, Malaysia

[6]Department of Electrical Engineering, Universitas Muhammadiyah Malang, 65144 Kota Malang, Indonesia

[7]Department of Computer Science and Engineering, Chitkara University, 174103 Himachal Pradesh, India

Emails: ainaahanis@graduate.utm.my[1], mfitri.kl@utm.my[2*], aizulnahar.kl@utm.my[3], masomar@graduate.utm.my[4], msharifuddin6@graduate.utm.my[5], faruq@umm.ac.id[6], vk.sinha@chitkarauniversity.edu.in[7], khamarrul.kl@utm.my[8],shahrum@utm.my[9]

ABSTRACT

Flooding poses serious risks to lives, infrastructure, and ecosystems, underscoring the need for accurate forecasting. However, missing values in hydrological datasets—often caused by equipment failure or extreme weather—can compromise forecast reliability. This study evaluates five imputation techniques: Last Observation Carried Forward, Next Observation Carried Backward, Linear Interpolation, Spline Interpolation, and K-Nearest Neighbours, to identify the most effective method for reconstructing missing flood-related data. Using temperature, humidity, and water level records from the Langat River, Selangor, Malaysia, each method's performance was assessed via Root Mean Square Error. Results show that Linear Interpolation generally yields the lowest error, while Next Observation Carried Backward performs best when missing data is minimal (1.20%).

Keywords: Flood forecast; Missing Data Imputation; Last Observation Carried Forward; Next Observation Carried Backwards; Linear Interpolation; Cubic Spline Interpolation; K-Nearest Neighbours.

1.0 INTRODUCTION

Flooding poses severe risks to human life, causes livestock losses and landslides, damages buildings and infrastructure, disrupts communication and transportation, and contaminates water supplies [1]. According to Malaysia's disaster service, NADMA, approximately 25,000 people were evacuated across seven Malaysian states in early 2021 due to excessive rains and severe floods. On August 18, 2021, severe floods occurred in Kedah, resulting in six fatalities and mud water reaching 1.5 meters high. The tremendous downpour exceeded a 70-year normal of 278 mm within a few hours. Additionally, heavy rainfall beginning on December 17, 2021, caused floods in eight Malaysian states, displacing 62,999 people. The situation worsened in 2022, with Malaysia facing even more significant flooding. The government had to open 1,200 evacuation centres to accommodate 210,071 displaced people.

The highest number of victims was recorded in December 2022, when monsoon season flooding affected 123,304 people and resulted in 12 fatalities [2]. Given these recurring and severe flood events, accurate flood forecasting has become increasingly important in Malaysia. Effective flood forecasting relies heavily on complete and accurate hydrological data. However, these datasets often contain gaps due to equipment failures, extreme weather conditions, and other unforeseen issues [3]. Missing data in any dataset can significantly compromise the accuracy of the applied forecasting models, making it challenging to predict future events [4]. Having an effective gap-filling method can assist in creating a more accurate forecasting model [5]. The primary purpose of this paper is to examine well-known interpolation methods, such as Last Observation Carried Forward (LOCF), Next Observation Carried Backwards (NOCB), Linear Interpolation, Spline Interpolation, and the machine learning method K-Nearest Neighbour (KNN), sought of the best tool for interpolating missing flood features data.

Classic missing data solutions such as listwise deletion and pairwise deletion discard records with missing values, offering simplicity at the expense of efficiency and unbiasedness when data are not missing completely at random. Single imputation methods, including mean substitution and regression-based imputation, replace missing

entries with predicted or average values. Iterative approaches like the Expectation–Maximization (EM) algorithm obtain maximum-likelihood estimates under MAR by alternating between imputing missing data and updating model parameters. Multiple Imputation (MI), first formalized by Rubin, generates multiple complete datasets via stochastic sampling and combines parameter estimates to reflect imputation uncertainty, emerging as the standard recommended approach for MAR scenarios which are later paired with, if needed techniques to balance dataset [6]. These classic methods provide a foundation for addressing missing data, although the increased adoption of machine learning handling missing values are visible [7]–[9], this is due to its versatility of usage [10], [11].

1.1 Imputation

Missing data is a significant problem in big data analytics that requires attention. This problem is crucial to the big data analysis that has recently gained popularity. Data may be missing for a variety of reasons, including typos, improper data formatting, and equipment malfunctions during data collection. This issue may make it difficult to precisely examine the data. These inadequate data will have an impact on the quality of data analysed and may even result in the construction of an incorrect data model, causing data outputs to differ from the real data. Missing data will cause three main issues. Firstly, most data processing methods are currently unable to handle datasets that contain missing data. Secondly, standard techniques or systems are not designed to handle partial datasets.

The problem of missing records is frequently disregarded in the data analysis process to complete quick tasks and save time, which will produce subpar statistical results. Analyzing datasets with missing records comes in third place. Numerous analysis techniques exhibit high sensitivity to the rate of missing data in the dataset, and a decrease in the effective record count within the dataset may result in a notable reduction or departure in the output of the dataset analysis [12].

In imputation research, missing mechanisms are typically characterized in terms of missing value distribution [4]. In this way, data may be absent under three separate assumptions, missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Under MCAR, the likelihood of missing data value is the same for all data points. MAR happens when the events that cause missingness are entirely random but only within a subset of another observable variable in the dataset. Finally, when none of the preceding two missing mechanisms are at work, but the missingness is directly tied to the actual missing value and/or another variable value, the missing mechanism is referred to be MNAR.

Most solutions for the missing data problem fall into two categories: imputation-based approaches and deletion methods. The deletion algorithms [13] only assess the remaining data; they reject the records that have missing values. The simplest and most often applied data processing techniques are the elimination procedures. But the analysis becomes increasingly erroneous the more data is lost. Conversely, imputation-based techniques [13] aim to fill in missing data by considering potential correlations between the dataset's values.

There are two sorts of imputation-based methods, those that use a data distribution model and those that do not. Without the distribution model, the link between data points may be overlooked, resulting in poor imputation results. To effectively create the model, the distribution model-based technique typically requires domain expertise. When the data is consistent with the distribution model, these approaches can produce satisfactory imputation results. Considering advancements in machine learning research, numerous machine learning-based imputation approaches have been developed to address the missing data issue. These methods employ machine learning techniques to extract rules from input data and estimate the potential value of missing data. This type of technique does not require domain expertise to construct the data distribution model, which can save model development costs while still providing decent imputation results, one of the commonly used approaches include K-nearest neighbour-based imputation [14].

1.2 Existing Well-Known Imputation Methods in Hydrology

In hydrology, a widely used method for imputing missing precipitation data is the arithmetic mean. This technique calculates the average precipitation values from nearby stations and is highly effective in regions where stations are evenly distributed with similar precipitation patterns. However, its simplicity is both an advantage and a limitation. Sattari et al. [15] observed that the arithmetic mean performs best when stations are close and share comparable weather patterns. Conversely, Armanuos et al. [16] demonstrated that more complex methods, such as the normal ratio (NR) or multiple linear regression (MLR), outperform the arithmetic mean in areas with diverse climates or geographic conditions. These findings emphasize the need to tailor the method to the specific characteristics of the study area to maintain accuracy.

The LOCF method is one of the most straightforward approaches for imputing missing data. The approach has been widely used in clinical trials and longitudinal studies due to its simplicity [17][18]. This method is useful for handling gaps in data series, especially when measurements are collected irregularly. However, LOCF assumes that conditions remain unchanged after the last observation, which can introduce bias in dynamic hydrological environments. For instance, if the last recorded rainfall value is carried forward during periods of high variability, it may not accurately represent the current conditions, leading to inaccuracies. Despite its drawbacks, LOCF

remains popular in research due to its ease of implementation and ability to handle data attrition without complex modelling assumptions.

Similar to LOCF, the NOCB method fills in missing values using the next available data point. This method tends to perform better than LOCF when missing data points are near significant changes in the dataset, such as sudden increases in water levels due to heavy rainfall. Jahangiri et al. [19] demonstrated that NOCB is particularly effective when dealing with longitudinal data in cases of intermittent missing data. This suggests that NOCB is better suited for hydrological data, where conditions fluctuate rapidly, compared to LOCF.

Linear interpolation is another widely used technique for imputing missing data in hydrology. This method estimates missing data by drawing a straight line between two known data points. It is most effective when the data follows a relatively smooth trend over time, such as steady changes in river flow or rainfall. However, linear interpolation often fails to account for irregular variability in datasets, which can reduce accuracy in more complex situations, such as during sudden flood events. Niedzielski and Halicki [20] noted that while linear interpolation works well in simple cases, it struggles in environments where conditions change rapidly.

Spline interpolation provides a more advanced approach than linear interpolation by fitting a smooth curve through data points. This method is especially useful for datasets that exhibit non-linear trends, such as fluctuations in river levels during flood events. In hydrological studies, spline interpolation has been found to outperform traditional methods, particularly in cases involving temporal variability. Amirzehni et al. [21] demonstrated that spline interpolation can accurately estimate missing data in situations where linear methods fall short, particularly in semi-arid climates or when working with satellite data that is irregularly collected.

Regression-based approaches are also frequently used to fill gaps in hydrological datasets. However, these methods often underestimate dry days and overestimate non-zero rainfall amounts. Fagandini et al. [22] highlighted that regression models, such as the Food and Agriculture Organization approach, tend to oversimplify spatial variability, resulting in inaccuracies in estimating both wet and dry conditions. Their study showed that geostatistical methods, such as ordinary kriging, outperform regression approaches by incorporating spatial correlations, which makes them more suitable for regions with complex rainfall patterns.

Machine learning techniques, such as KNN, have recently gained popularity for imputing missing data in hydrology. KNN works by identifying the most similar data points, or "neighbours," within the dataset and using their values to estimate the missing data. This method is particularly effective in dealing with datasets that exhibit complex patterns, such as variability in river levels during monsoon seasons. Sahoo and Ghose [23] conducted a study on missing precipitation data in the Cachar watershed and found that KNN performed well in handling moderate levels of missing data. However, the method showed limitations when there were extensive gaps or when neighbouring stations exhibited drastically different precipitation patterns. KNN is most useful when the missing data is non-random, and nearby stations have similar trends.

In summary, the various imputation methods reviewed, including more straightforward approaches such as and NOCB, as well as more advanced techniques like Linear and Spline Interpolation and the machine learning-based KNN, each offer unique advantages depending on the characteristics of the dataset and the extent of missing data. While more straightforward methods such as LOCF and NOCB are easy to implement, they may introduce bias in datasets with high variability. In contrast, more sophisticated approaches like Linear and Spline Interpolation, or KNN, provide more accurate estimations, especially when dealing with complex patterns in hydrological data.

The following section explains the real-world case study using flood-related data from the Langat River in Selangor, Malaysia. The case study will evaluate the performance of each method in handling missing data within this hydrological context in the methodology section, highlighting the practical implications of the comparative analysis presented.

2.0 STUDY AREA AND DATA

The study area that this research focuses on is in Malaysia as per Fig.1, within the state of Selangor, as in Fig.2. As per Fig.3, Bukit Changgang is within Kuala Langat district with latitude of 2.8289° N and longitude of 101.6278° E, located on Peninsular Malaysia's west coast. This district has a population of 307,449 and 2,699 residents of Bukit Changgang [24]. Bukit Changgang is an area prone to flooding. The most recent occurrence affected over 500 victims and 1,007 households in December 2021. The data used in this study was gathered by the Department of Irrigation and Drainage (DID) [25] throughout ten years from 2011 to 2020, however for this comparison analysis, just three years of data from 2013 to 2015 were used as it has complete data, with each feature having 1,095 data points.

3.0 METHODOLOGY

Comparing the findings to genuine missing data is challenging when the actual numbers are missing. As a result, for performance comparison, simulated missing data from a complete dataset is required. To simulate missing data, this flood features dataset was programmed to have values missing entirely at random: Daily mean temperature (˚C) has 27.00% missing values, Daily Mean Relative Humidity (%) with 73.00% and Water Level (m) having the least missing values at 1.20%. For these data sets, simulated missing data are chosen randomly in

conjunction with the mechanism missing completely at random (MCAR). Missing data has been imputed using the recommended imputation methods, and a comparison has been done.

Fig.1: Map of Malaysia.

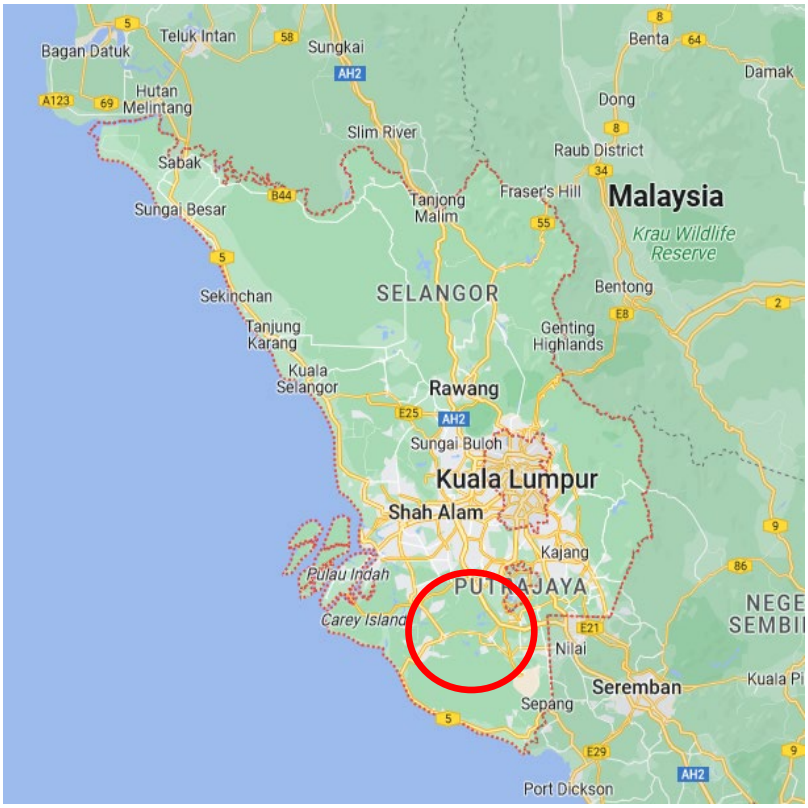

Fig.2: Location of the study area in the state of Selangor.

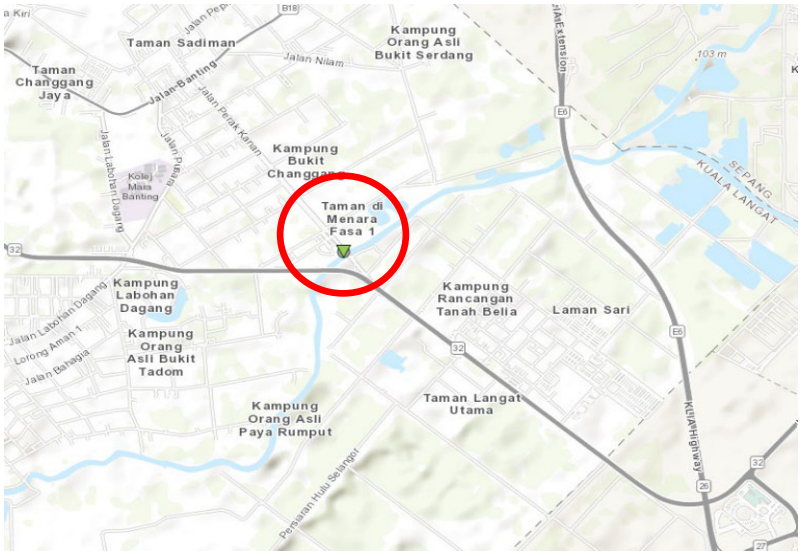

Fig.3: Location of Bukit Changgang Station in Selangor within District of Kuala Langat.

3.1 Last Observation Carried Forward

The LOCF is a standard statistical strategy for analysing longitudinal data that may be missing some follow-up observations. A missing follow-up value is substituted or imputed as the feature's previously observed value in this analysis technique. The last observation is carried forward as per Table 1 [26]. This approach is appropriate for longitudinal, one-dimensional data with measurements taken over time [27][28].

Table 1: LOCF example

Cases	1am	2am	3am	4am
1	1.0	2.1 →	2.1	0.5
2	1.4	1.2	1.6 →	1.6

3.2 Next Observation Carried Backwards

Missing values are imputed by NOCB from the next accessible state, which is carried backwards, as referred to in Table 2. While the NOCB technique is not widely used, it may be found in select publications and regulatory approval documents [29].

Table 2. NOCB example

Cases	1am	2am	3am	4am
1	1.0 ←	1.0	2.2	0.5
2	1.4	1.2 ←	1.2	1.3

3.3 Linear Interpolation

Imputation via interpolation is a technique for estimating unknown data points using two known data points in between. The linear interpolation algorithm used in this study is a Python function that disregards the index and treats all values as evenly spaced [30]. It detects the trend and then uses it in the dataset to fill in the missing values. As per (1), x_1 and y_1 are the first coordinates, x_2 and y_2 are the second coordinates. x is the point at which the interpolation is performed, while y is the interpolated value as shown in Equation (1).

$$y = y_1 + (x - x_1)\frac{(y_2 - y_1)}{(x_2 - x_1)} \quad (1)$$

3.4 Cubic Spline Interpolation

Cubic spline interpolation is a polynomial-based interpolation. It assures that each curve spline's first and second derivatives are continuous [31]. Spline interpolation methods are intended to create a surface based on known errors of surrounding points. An error surface equation is then used to calculate the target position error. As per Equation (2),for $s_i(x)$, s stands for each polynomial segment, x is the input variable being interpolated, the i is the ith polynomial segment. a_i, b_i, c_i and d_i are constant that shapes and position the cubic polynomial while x_{i+1} are the boundaries.

$$s_{i\,(x)} = a_i x^2 + b_i x^3 + c_i x + d_i \; that \; is \; vallid \; for \; x_i \leq x \leq x_{i+1} \qquad (2)$$

3.5 K-Nearest Neighbour

In Nearest neighbour imputation algorithms, every missing value on the records is replaced with a value derived from interconnected cases in the overall records set [32]. In this strategy, as observed in Equations (3) and (4), the k term refers to the number of nearest data points from the nearest neighbour columns with complete values to the missing data point. The weight of the $\mathcal{D}_i$th entry is computed, $\mathcal{D}_i$ is the distance between the ith neighbour and the imputed location. Euclidean Distance is the distance measure used to find the nearest neighbours. After discovering the nearest neighbours, W the weighted mean of the k nearest points will be computed and imputed to the missing point [33][34].

$$D_i = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2} \qquad (3)$$

$$W = \frac{\frac{1}{D_1}}{\sum_{i=1}^{k} \frac{1}{D_1}} \qquad (4)$$

3.6 Evaluation Metrics

The error measurement used to evaluate the performance of imputation algorithms is the Root Mean Square Error (RMSE). These measures quantify the amount of variation between imputations and real values. RMSE is a typical statistical metric for assessing model performance in meteorology, air quality, and climate research investigations. RMSE provides information on short-term efficiency as a benchmark of the difference between forecasted and actual values [35]. The smaller the RMSE, the more accurate the evaluation. Equation (5) explains RMSE, where $(\hat{y}_l)$ is the estimated value, y_i is the actual value of the observation and n is the number of observations [36].

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (\hat{y}_i - y_i)^2 / n} \qquad (5)$$

4.0 RESULTS AND DISCUSSION

This section discusses the results of imputation methods for the simulated datasets. The performance of the different imputation methods on simulated datasets of Langat River is measured with RMSE. Using RMSE, the error of the forecasted values and the actual values in the dataset are shown. The lowest RMSE values are referred to determine the most suitable imputation method. The results of RMSE of different imputation methods on simulated Langat River datasets are shown below.

4.1 Imputation Performance Comparison

The Incomplete datasets have missing values completely at random. The data include temperature, daily mean relative humidity, and water level as features. Figure 6 illustrates 27% missing values for daily mean temperature (˚C), all the incomplete data are visualized using Bar Chart and line graph, Fig.5 for Bar Chart, Fig.6 for temperature, Fig.7 is completed daily mean temperature (˚C) graph using linear interpolation, Fig.8 for daily mean

relative humidity, Fig.9 illustrates completed line graph of daily mean relative humidity (%) using Linear Interpolation while Fig.10 shows water level. Results for each imputation method on the simulated datasets of Langat River are in line graph format as well for easy comparison and illustrates 1.20% missing values for water level (m) in line graph from 1,095 data points. Fig.11 is the completed line graph for water level (m) using NOCB.

The reason for showing Linear Interpolation and NOCB line graph for these three features are due to their good performance. In regard to imputation model performance results based on RMSE, they are illustrated using bar chart in Fig.12, 13 and 14. These figures show that LOCF method consistently performs the worst as per Fig. 12 and 14, with the highest or second highest RMSE value in Fig.11. Although some studies suggest that KNN may outperform other imputation methods for missing flood feature datasets [37], this is not the case in this study. Referring to Fig.12, 13 and 14, linear interpolation has shown to be more consistently effective for these specific datasets. When it comes to small missing value points in this dataset, NOCB outperforms all other models for imputation as per Fig.12.

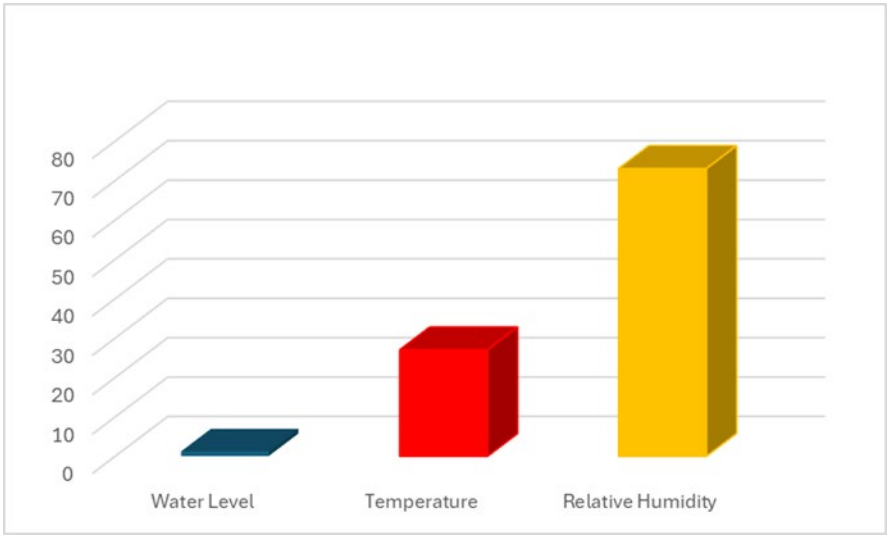

Fig. 5: Data Missing Values Bar Chart.

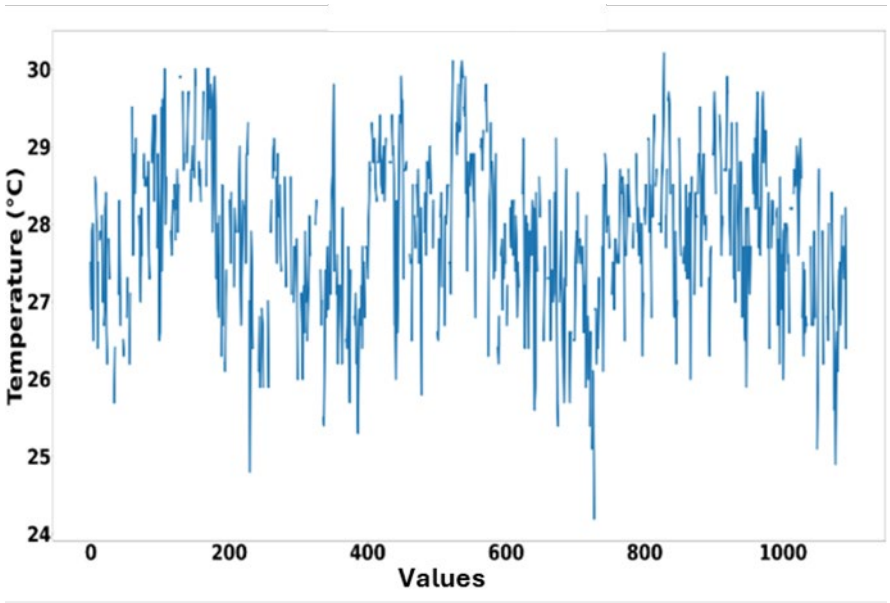

Fig. 6: Temperature Data Line Graph with Missing Values.

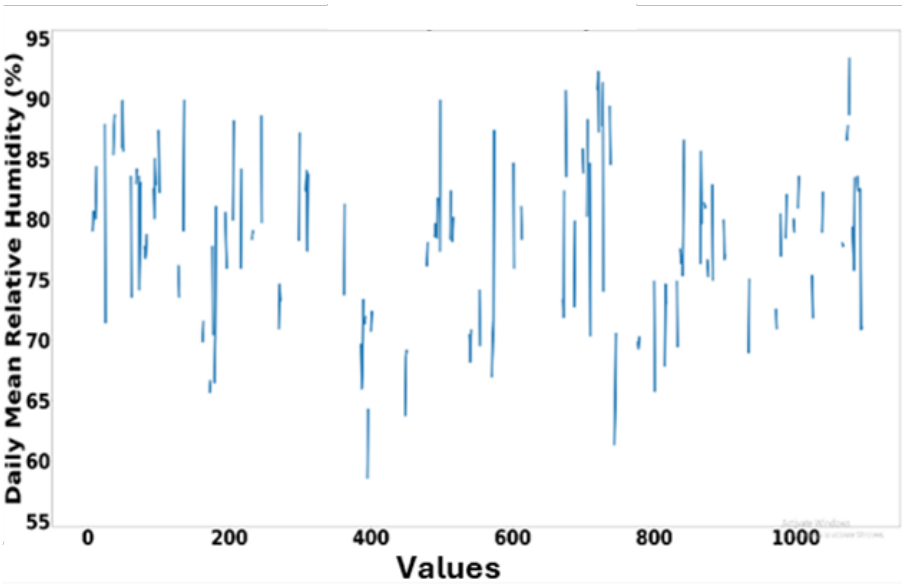

Fig. 7: Temperature Data After Linear Interpolation Line Graph.

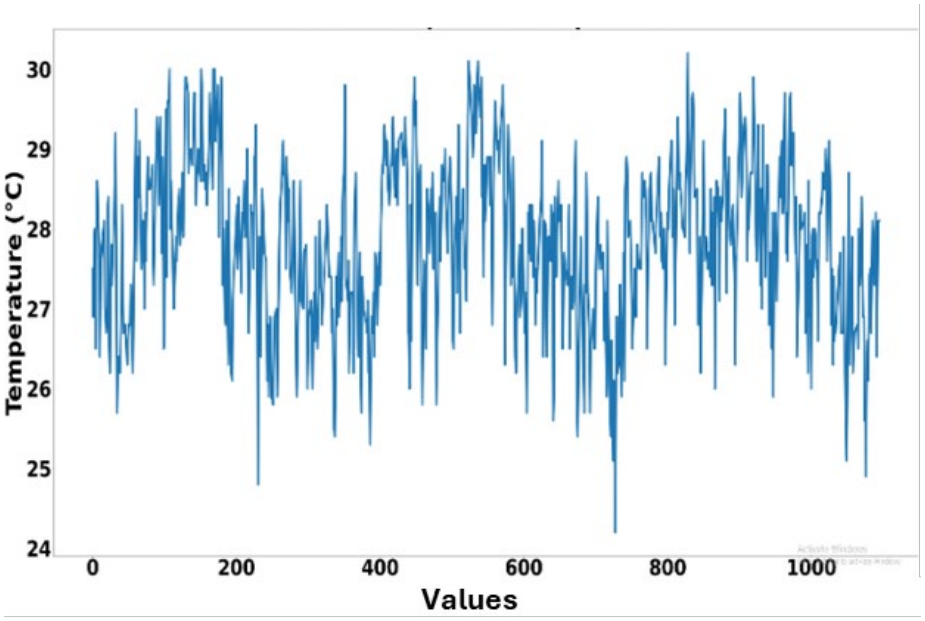

Fig. 8: Relative Humidity Data Line Graph with Missing Values.

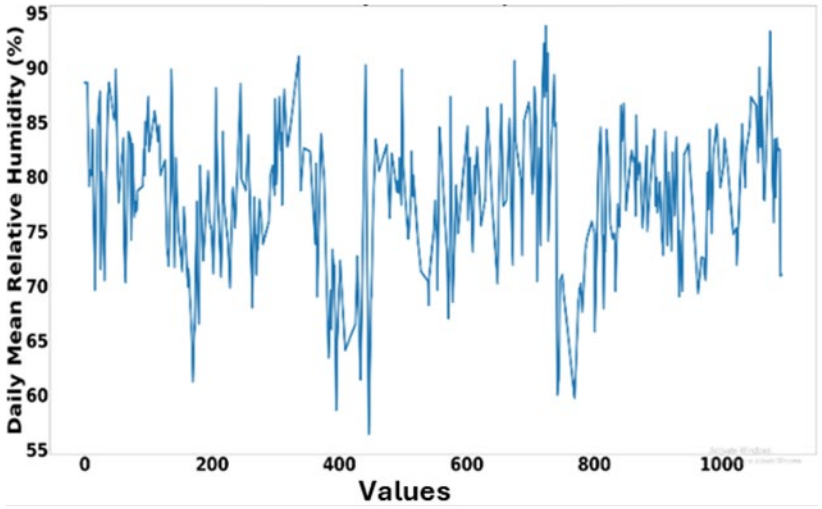

Fig. 9: Relative Humidity Data After Linear Interpolation Line Graph.

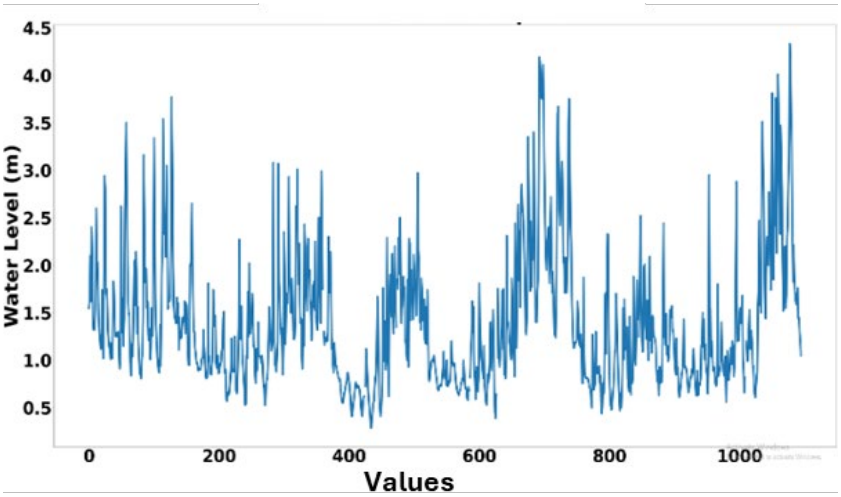

Fig. 10: Water Level Data Line Graph with Missing Values.

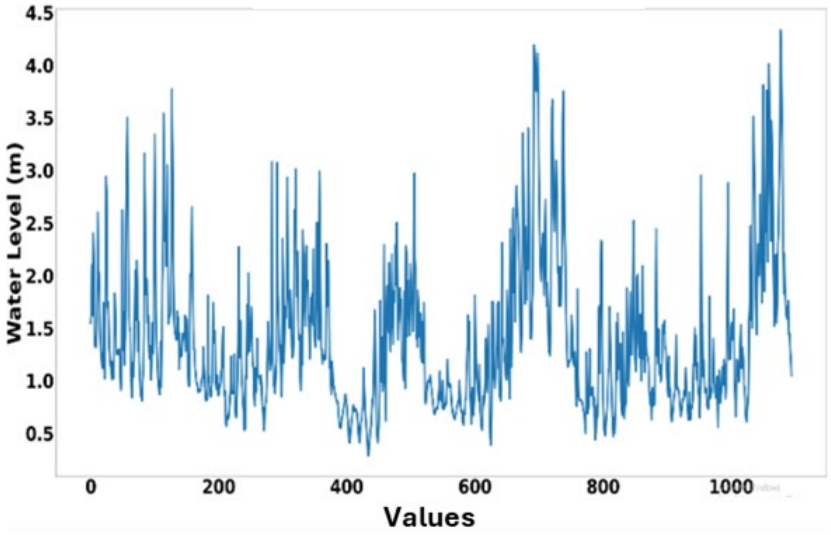

Fig. 11: Water Level Data After NOCB Imputation Line Graph.

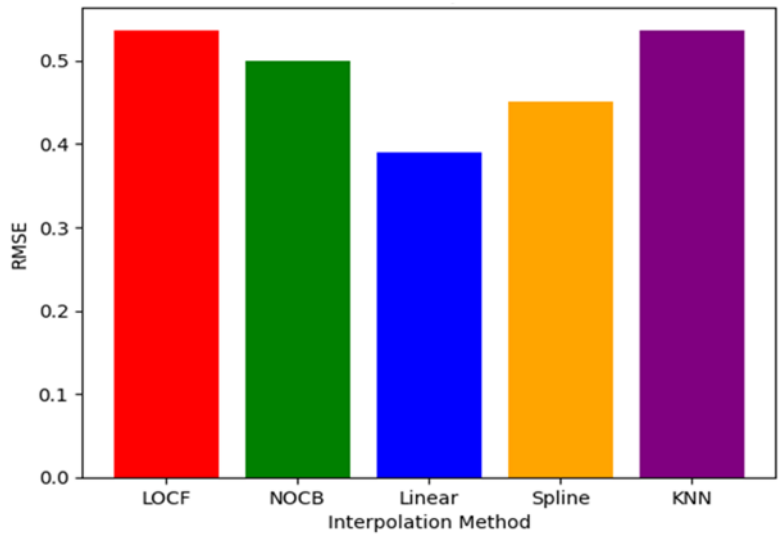

Fig. 12: Langat River simulated daily mean temperature dataset imputation method result.

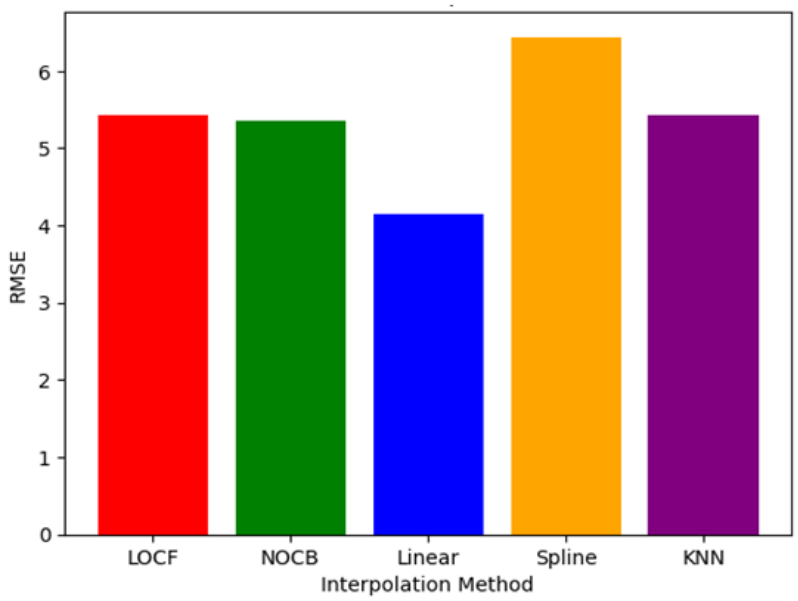

Fig. 13: Langat River simulated daily mean relative humidity dataset imputation method result.

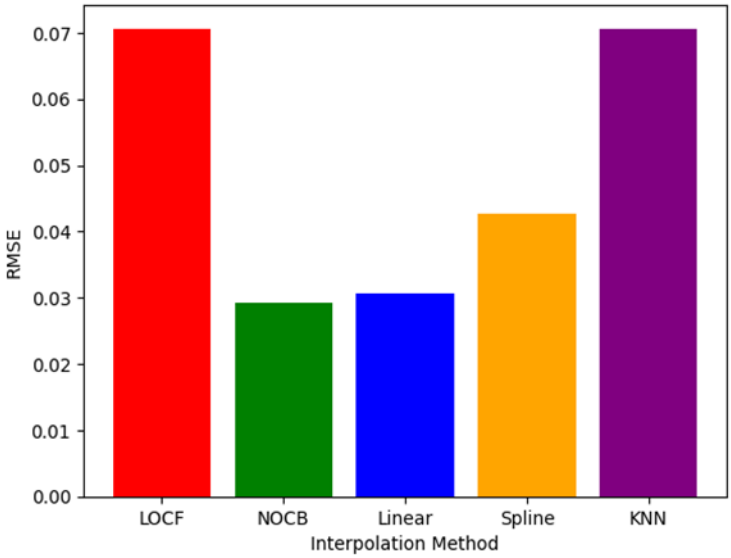

Fig. 14: Langat River simulated water level dataset

5.0 CONCLUSION

The quest for the most efficient approach to imputing missing values for flood characteristics datasets has attracted much attention in numerous research. Missing data is a major concern for academics. Selecting strategies for dealing with missing data is tricky since the same approach might provide greater predicted accuracy rates in certain situations but not in others.

In this study five imputation techniques were utilised, LOCF, NOCB, Linear Interpolation, Cubic Spline Interpolation, and KNN. They are then compared to determine the best strategy for filling in the missing flood characteristics values for the Langat River in Selangor, Malaysia, recreated randomly. The findings demonstrated that Linear Interpolation is the optimum approach for these datasets.

However, NOCB has demonstrated dominance as the approach with the most significant results regarding low percentage missing values. It is also proven that the performance of the imputation approach varies based on the kind of dataset, the condition, and the quantity of missing values. As a result, while determining the optimal approach for data imputation, it is critical to evaluate the dataset after imputation. To summarise, using multiple imputation strategies based on the flood feature datasets is supported, and other research with different methodologies and datasets should be investigated.

5.1 Research Contribution

This study contributes to the field of hydrological data management and flood forecasting by conducting a rigorous comparative analysis of five widely used imputation methods, LOCF, NOCB, Linear Interpolation, Spline Interpolation, and KNN for reconstructing missing flood-related data in the Langat River basin, Selangor, Malaysia.

Unlike previous research, which often focuses on a limited set of methods or generic case studies, this work evaluates the performance of both classic statistical techniques and a prominent machine learning approach within a real-world, flood-prone tropical basin. By systematically assessing imputation accuracy using RMSE across temperature, humidity, and water level records, the study clearly identifies linear interpolation as the most robust method under realistic missing data scenarios, while highlighting the specific effectiveness of NOCB for scenarios with minimal missingness.

The findings provide practical guidance for hydrological data practitioners and researchers working in similar tropical contexts, enabling them to select optimal imputation methods tailored to data conditions and supporting the development of more reliable flood forecasting models. This research thus addresses a critical gap by establishing empirical benchmarks for imputation efficacy in operational flood monitoring, with direct implications for disaster risk management in Malaysia and other data-scarce regions.

6.0 ACKNOWLEDGEMENT

The authors sincerely thanks all those who were involved for their invaluable guidance and insightful feedback throughout the research process. This work was supported by Collaborative Research and External Grant under Japan International Cooperation Agency (JICA) with Cost Center No: S.K130000.0543.4Y359, whose financial assistance made this study possible. We also acknowledge the support of Universiti Technologi Malaysia, Malaysia Japan International Institute of Technology, which provided essential resources and infrastructure. Special appreciation goes to the data providers Malaysian Meteorological Department and Malaysian Department of Irrigation and Drainage for making the datasets available. Finally, we gratefully recognize the constructive comments from anonymous reviewers that helped improve the quality of this manuscript.

REFERENCES

[1] H. A. Rahman, "Climate Change Scenarios In Malaysia: Engaging The Public," *International Journal of Malay-Nusantara Studies*, vol. 1, no. 2. pp. 55–77, Nov. 30, 2018 (Accessed: Sep. 03, 2023). Available: https://journal.unhas.ac.id/index.php/IJoM-NS/article/view/5518.

[2] "Portal Bencana - Laporan." https://portalbencana.nadma.gov.my/ms/laporan (accessed Sep. 03, 2023).

[3] I. F. Kamaruzaman, W. Z. Wan Zin, and N. Mohd Ariff, "A comparison of method for treating missing daily rainfall data in Peninsular Malaysia," *Malaysian J. Fundam. Appl. Sci.*, vol. 13, no. 4–1, pp. 375–380, Dec. 2017, doi: 10.11113/MJFAS.V13N4-1.781.

[4] R. J. A. Little and D. B. Rubin, "Statistical analysis with missing data," *Stat. Anal. with Missing Data*, pp. 1–381, Jan. 2014, doi: 10.1002/9781119013563.

[5] S. van Buuren, "Flexible Imputation of Missing Data, Second Edition," *Flex. Imput. Missing Data, Second Ed.*, Jul. 2018, doi: 10.1201/9780429492259/Flexible-Imputation-Missing-Data-Second-Edition-Stef-Van-Buuren.

[6] A. H. Zuhairi, F. Yakub, M. Omar, M. Sharifuddin, K. A. Razak, and A. Faruq, "Imbalanced Flood Forecast Dataset Resampling Using SMOTE-Tomek Link," *Int. Exch. Innov. Conf. Eng. Sci.*, vol. 10, pp. 845–850, 2024, doi: 10.5109/7323359.

[7] K. Boros and Z. Kmetty, "Identifying missing data handling methods with text mining," *Int. J. Data Sci. Anal.*, pp. 1–13, Jun. 2024, doi: 10.1007/S41060-024-00582-1/Figures/12.

[8] A. Z. Alruhaymi, C. J. Kim, A. Z. Alruhaymi, and C. J. Kim, "Study on the Missing Data Mechanisms and Imputation Methods," *Open J. Stat.*, vol. 11, no. 4, pp. 477–492, Aug. 2021, doi: 10.4236/OJS.2021.114030.

[9] K. Lee, J. Carpenter, R. Little, C. Nguyen, and R. Cornish, "1376 Modern concepts in the handling and reporting of missing data," *Int. J. Epidemiol.*, vol. 50, no. Supplement_1, Sep. 2021, doi: 10.1093/IJE/DYAB168.370.

[10] M. Omar, F. Yakub, S. S. Abdullah, M. S. A. Rahim, A. H. Zuhairi, and N. Govindan, "One-step vs horizon-step training strategies for multi-step traffic flow forecasting with direct particle swarm optimization grid search support vector regression and long short-term memory," *Expert Syst. Appl.*, vol. 252, p. 124154, Oct. 2024, doi: 10.1016/J.ESWA.2024.124154.

[11] M. S. Abd Rahim, F. Yakub, M. Omar, R. Abd Ghani, I. Dhamanti, and S. Sivakumar, "Prediction of Influenza A Cases in Tropical Climate Country using Deep Learning Model," *2nd IEEE Natl. Biomed. Eng. Conf. NBEC 2023*, pp. 188–193, 2023, doi: 10.1109/NBEC58134.2023.10352612.

[12] X. Xu, W. Chong, S. Li, A. Arabo, and J. Xiao, "Missing data imputation based on the evidence Chain," *IEEE Access*, vol. 6, pp. 12983–12992, Feb. 2018, doi: 10.1109/ACCESS.2018.2803755.

[13] R. J. A. Little and D. B. Rubin, "Statistical analysis with missing data," *Stat. Anal. with Missing Data*, pp. 1–381, Jan. 2014, doi: 10.1002/9781119013563.

[14] K. F. Jea, C. W. Hsu, and L. Y. Tang, "A Missing Data Imputation Method with Distance Function," *Proc. - Int. Conf. Mach. Learn. Cybern.*, vol. 2, pp. 450–455, Nov. 2018, doi: 10.1109/ICMLC.2018.8526985.

[15] M. T. Sattari, A. Rezazadeh-Joudi, and A. Kusiak, "Assessment of different methods for estimation of missing data in precipitation studies," *Hydrol. Res.*, vol. 48, no. 4, pp. 1032–1044, 2017, doi: 10.2166/nh.2016.364.

[16] A. M. Armanuos, N. Al-Ansari, and Z. M. Yaseen, "Cross assessment of twenty-one different methods for missing precipitation data estimation," *Atmosphere (Basel).*, vol. 11, no. 4, 2020, doi: 10.3390/ATMOS11040389.

[17] H. Cao, J. Li, and J. P. Fine, "On last observation carried forward and asynchronous longitudinal regression analysis," *Electron. J. Stat.*, vol. 10, no. 1, pp. 1155–1180, 2016, doi: 10.1214/16-EJS1141.

[18] Y. H. Koog, "Opioid Çali Şmalari Nda İLeri Taşi Nmi Ş Son Gözlem Analizine Karşi Temkinli Olunmali Di R," *Turkish J. Rheumatol.*, vol. 28, no. 4, pp. 282–283, 2013, doi: 10.5606/tjr.2013.3735.

[19] M. Jahangiri *et al.*, "A wide range of missing imputation approaches in longitudinal data: a simulation study and real data analysis," *BMC Med. Res. Methodol.*, vol. 23, no. 1, pp. 1–20, 2023, doi: 10.1186/s12874-023-01968-8.

[20] T. Niedzielski and M. Halicki, "Improving Linear Interpolation of Missing Hydrological Data by Applying Integrated Autoregressive Models," *Water Resour. Manag.*, vol. 37, no. 14, pp. 5707–5724, Nov. 2023, doi: 10.1007/S11269-023-03625-7/FIGURES/9.

[21] P. Amirzehni, S. Samadianfard, A. H. Nazemi, and A. A. Sadraddini, "Evaluating capabilities of the spline and cubic spline interpolation functions in reference evapotranspiration estimation implementing satellite image data," *Earth Sci. Informatics*, vol. 16, no. 4, pp. 3779–3795, 2023, doi: 10.1007/s12145-023-01127-z.

[22] C. Fagandini, V. Todaro, M. G. Tanda, J. L. Pereira, L. Azevedo, and A. Zanini, "Missing Rainfall Daily Data: A Comparison Among Gap-Filling Approaches," *Math. Geosci.*, vol. 56, no. 2, pp. 191–217, 2024, doi: 10.1007/s11004-023-10078-6.

[23] A. Sahoo and D. K. Ghose, "Imputation of missing precipitation data using KNN, SOM, RF, and FNN," *Soft Comput.*, vol. 26, no. 12, pp. 5919–5936, Jun. 2022, doi: 10.1007/S00500-022-07029-4/FIGURES/12.

[24] M. Department of Statistics, "Census 2020," 2020. https://www.mycensus.gov.my/index.php/census-product/publication/175-census-2020 (accessed Aug. 29, 2023).

[25] "Department of Irrigation and Drainage." https://www.water.gov.my/index.php/pages/view/583 (accessed Dec. 26, 2023).

[26] F. Mahmud, N. S. Pathan, and M. Quamruzzaman, "Early detection of Sepsis in critical patients using Random Forest Classifier," *2020 IEEE Reg. 10 Symp. TENSYMP 2020*, pp. 130–133, Jun. 2020, doi: 10.1109/TENSYMP50017.2020.9231011.

[27] N. J.Salkind, "Encyclopedia of Research Design," *Encycl. Res. Des.*, May 2010, doi: 10.4135/9781412961288.

[28] J. E. Overall, S. Tonidandel, and R. R. Starbuck, "Last-observation-carried-forward (LOCF) and tests for difference in mean rates of change in controlled repeated measurements designs with dropouts," *Soc. Sci. Res.*, vol. 38, no. 2, pp. 492–503, Jun. 2009, doi: 10.1016/J.SSRESEARCH.2009.01.004.

[29] M. A. Razzaq, I. Cleland, C. Nugent, and S. Lee, "SemImput: Bridging Semantic Imputation with Deep Learning for Complex Human Activity Recognition," *Sensors 2020,* vol. 20, no. 10, p. 2771, May 2020, doi: 10.3390/S20102771.

[30] C. De Mulder, T. Flameling, S. Weijers, Y. Amerlinck, and I. Nopens, "An open software package for data reconciliation and gap filling in preparation of Water and Resource Recovery Facility Modeling," *Environ. Model. Softw.*, vol. 107, pp. 186–198, Sep. 2018, doi: 10.1016/J.ENVSOFT.2018.05.015.

[31] R. Sadikin, I. W. A. Swardiana, and T. Wirahman, "Cubic spline interpolation for large regular 3D grid in cylindrical coordinate: (Invited paper)," *Proc. - 2017 Int. Conf. Comput. Control. Informatics its Appl. Emerg. Trends Comput. Sci. Eng. IC3INA 2017*, vol. 2018-January, pp. 1–6, Jul. 2017, doi: 10.1109/IC3INA.2017.8251730.

[32] L. Beretta and A. Santaniello, "Nearest neighbor imputation algorithms: A critical evaluation," *BMC Med. Inform. Decis. Mak.*, vol. 16, no. 3, pp. 197–208, Jul. 2016, doi: 10.1186/S12911-016-0318-Z/TABLES/5.

[33] W. Y. Lai and K. K. Kuok, "A Study on Bayesian Principal Component Analysis for Addressing Missing Rainfall Data," *Water Resour. Manag.*, vol. 33, no. 8, pp. 2615–2628, Jun. 2019, doi: 10.1007/S11269-019-02209-8/FIGURES/7.

[34] G. E. A. P. A. Batista and M. C. Monard, "A study of k-nearest neighbour as an imputation method," *Front. Artif. Intell. Appl.*, vol. 87, pp. 251–260, 2002, Accessed: Sep. 03, 2023. [Online]. Available: https://www.researchgate.net/publication/220981745_A_Study_of_K-Nearest_Neighbour_as_an_Imputation_Method.

[35] S. M. C. M. 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," *Bull. Electr. Eng. Informatics*, vol. 9, no. 2, pp. 635–643, 2020, doi: 10.11591/eei.v9i2.2090.

[36] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)? -Arguments against avoiding RMSE in the literature," *Geosci. Model Dev.*, vol. 7, no. 3, pp. 1247–1250, 2014, doi: 10.5194/gmd-7-1247-2014.

[37] L. S. Chia, C. W. Yoke, and H. M. Kang, "Comparison of Imputed Statistics Across Rainfall Stations of Four Different Districts in Kelantan, Peninsular Malaysia," *J. Pharm. Negat. Results*, vol. 13, pp. 393–404, Jan. 2023, doi: 10.47750/PNR.2022.13.S10.041.