Journal of Agrometeorology

ISSN : 0972-1665 (print), 2583-2980 (online)
Vol. No. 26 (3) : 318-323 (September - 2024)
https://doi.org/10.54386/jam.v26i3.2657
https://journal.agrimetassociation.org/index.php/jam

Research Paper

Univariate and multivariate imputation methods evaluation for reconstructing climate time series data: A case study of Mosul station-Iraq

KHALID QARAGHULI[1,2], M. F. MURSHED[1*], M. AZLIN M. SAID[1], ALI MOKHTAR[3] and IMAN ROUSTA[4]

[1]*School of Civil Engineering, Engineering Campus, Universiti Sains Malaysia, Malaysia*
[2]*Al-Mussaib Technical Institute, Al-Furat Al-Awsat Technical University, Babil, Iraq*
[3]*Department of Agricultural Engineering, Faculty of Agriculture, Cairo University, Egypt*
[4]*Department of Geography, Yazd University, Yazd, Iran*
**Corresponding Author Email: cefaredmurshed@usm.my*

ABSTRACT

Comprehensive climate time series data is indispensable for monitoring the impacts of climate change. However, observational datasets often suffer from data gaps within their time series, necessitating imputation to ensure dataset integrity for further analysis. This study evaluated six univariate and multivariate imputation methods to infill missing values. These methods were applied to complete the subsets of time series data for precipitation, temperature, and relative humidity from Mosul station spanning 1980–2013. Artificial gaps of 5%, 10%, 20%, and 30% missing observations were introduced under scenarios of missing completely at random (MCAR) missing at random (MAR), and missing not at random (MNAR). Evaluation metrics including RMSE and Kling-Gupta Efficiency were utilized for performance evaluation. Results revealed that seasonal decomposition was the most effective univariate imputation method across all variables. For the multivariate imputation, kNN demonstrated superior performance in infilling the precipitation missing data under MCAR, while norm.predict exhibited optimal performance in the temperature missing data under all missing scenarios. Moreover, missForest was identified as the most suitable method for infilling missing relative humidity data. This study's methodology offers insights into selecting appropriate imputation methods for other climate stations, thereby enhancing the accuracy of the climate change effects analysis.

Keywords: Missing data, Data reconstruction, Climate change, MNAR, MAR, MCAR

Recent severe catastrophes worldwide, like floods, droughts, and wildfires, were primarily driven by climate change. Understanding and addressing this issue necessitates robust databases that facilitate the analysis of climate signals, continual monitoring of their progression, and the developing of more precise predictive models for future changes (Diouf *et al.,* 2022). Climate time series exhibit complex spatio-temporal characteristics and frequently encounter the challenge of incomplete or missing values (Saubhagya *et al.,* 2024). Many analytical methods necessitate the replacement of missing values with estimations to facilitate subsequent analysis. This process, known as imputation, is a fundamental aspect of statistical data pre-processing.

The missing data mechanism falls into three categories and has a significant impact on the validity and efficacy of the imputation procedures (Afrifa-Yamoah *et al.,* 2020). The data 'missingness' may occur in one of three types, (i) Missing Completely At Random (MCAR), that is when the probability of a variable having missing data is not related to the values of that variable, or to the other variables that have been measured. (ii) Missing At Random (MAR), which is when the probability of missing data is related to the values of other measures in our dataset but not the missing values themselves. (iii) Missing Not At Random (MNAR), or specifically when the probability of a missing value is dependent on the variable being analyzed (Diouf *et al.,* 2022).

In general, two approaches have been devised to address the issue of time series missing data. Typically, these approaches

Article info - DOI: *https://doi.org/10.54386/jam.v26i3.2657*
Received: 8 July 2024; Accepted: 16 July 2024 ; Published online : 01 September 2024

encompass both conventional statistical approaches and machine learning techniques (Saubhagya *et al.,* 2024; Sridhara *et al.,* 2024). The statistical interpolation approach utilizes statistical knowledge to fill in data, including mean, median, mode, etc. This strategy works best with linear assumptions and minimal missing data (Ou *et al.,* 2024). On the other hand, the machine learning imputation methods are advancing rapidly, primarily focusing on the development of predictive approaches to address missing data through the use of unsupervised or supervised learning (Emmanuel *et al.,* 2021). However, these approaches fall under two categories, (i) single imputation which concerns replacing a missing value with a single plausible value. (ii) Multiple imputation involves generating several possible values for each missing data item and then combining the results. This approach minimizes the error that may occur due to imputation (Diouf *et al.,* 2022).

The meteorological data retrieved from the IMOS database for all stations in Nineveh, Iraq display a significant number of missing values, leading to biased findings in climate studies reliant on this dataset. Consequently, addressing this deficiency within the temporal continuity of meteorological observations becomes imperative. This necessity has spurred our exploration of appropriate methodologies aimed at effectively reconstructing the missing data. Therefore, this study evaluates the univariate and multivariate imputation approaches to fill in rainfall, temperature, and relative humidity time series data from Mosul station to identify the optimal imputation method. The R packages include methods such as missForest (based on the Random Forest technique), MICE, and Visualisation and Imputation of Missing Values (VIM) have been chosen to conduct missing data imputation. A cross-validation methodology has been suggested to evaluate the estimations produced by the imputation methods. This involves generating incomplete time series data under various mechanisms, MAR, MCAR, and MNAR.

MATERIAL AND METHODS

Study area

Nineveh governorate, Iraq is situated in the northwestern region of the country, with coordinates ranging from 41° 30'–44° 30' longitude and 35° 00'–37° 00' latitude, as depicted in Fig. 1. The region serves as the focal point of Iraq's agricultural economy, supplying substantial amounts of wheat and barley crops essential to meet the population's needs.

In Iraq, particularly within Nineveh province, the adverse impact of wars and the presence of terrorist organizations over the past decade have led to operational disruptions in several weather stations. As a consequence, acquiring accurate and comprehensive weather data from all stations has become challenging. For that, it is mandatory to find an appropriate approach to fill gaps in gauge records for drought analysis models.

Data and sources

Mosul station is the main meteorological ground station in the governorate and holds long-term monthly time series data, specifically for three variables: precipitation, temperature, and relative humidity. The period covered by the data is from 1980 to 2022. The station's data records are complete (with no missing values) from 1980 to 2013. However, from 2014 to 2022, the records have missing temperature and relative humidity values, with 9% and 10% respectively. Therefore, reconstructing these data is essential for further environmental analysis. The data records are obtained from the Iraqi Meteorological Organization and Seismology. Fig. 2 shows the missingness percentage in the time series data.

Data amputation

The assessment of the imputation algorithm performance often entails the creation of missing values. To generate artificial gaps in datasets, we assumed three distinct mechanisms: MCAR, MAR, and MNAR. The percentage of missingness was simulated at rates of 5%, 10%, 20%, and 30%. Two different scenarios were applied to generate missing data: univariate and multivariate. The former considered individual variables and created gaps in precipitation, temperature, and relative humidity data separately, while the latter considered multiple variables and created gaps in the aforementioned variables as a composite.

Univariate imputation methods

Two methods of single imputation, namely seasonal decomposition and Kalman smoothing, have been employed for imputing univariate precipitation, temperature, and relative humidity

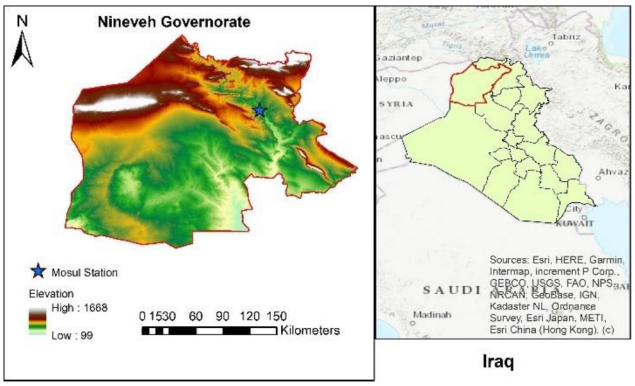

Fig. 1: Location of the study area

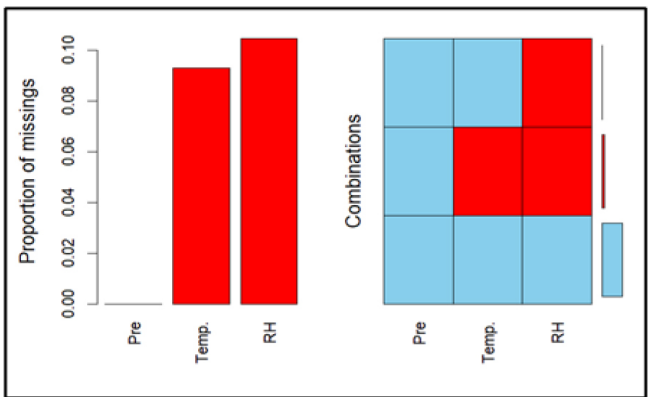

Fig. 2: Data missingness percentage

missing data. These approaches were selected for their effectiveness in producing the best results with more complex and longer time series datasets (Umar and Gray, 2023). The methods were executed using the R package imputeTS version 3.3.

Seasonal decomposition

Trend extraction from time series data is a fundamental task in many engineering and meteorological domains. Numerous studies have been conducted on trend extraction techniques, which can be broadly classified into two categories: smoothing-based and non-smoothing-based techniques. The seasonal-trend decomposition based on loess is one of the well-known methods utilizing the smoothing approach. It de-seasonalizes the time series by removing the seasonal component, applies imputation to the de-seasonalized series, and then returns the seasonal component.

Kalman smoothing

Kalman filter is a statistical algorithm that allows for specific calculations to be performed on a model represented in the state space form (Chaudhry *et al.,* 2019). The Kalman smoothing approach is applied to either the state space representation of an ARIMA model obtained by auto.arima or a basic structural model obtained by StructTS (Moritz & Bartz-Beielstein, 2017). In this study, Kalman filters were employed to fit ARIMA models to forecast missing values by leveraging trends observed in previously observed data.

Multivariate imputation methods

This study evaluated four imputation techniques to fill in missing values. These imputation approaches were: kNN, missForest, norm.predict, and random forest (rf). The former two are single imputation and the latter two are multiple imputation methods.

k-nearest neighbors (k-NN)

The k-nearest neighbor technique relies on donor observations. It involves aggregating the values of the *k* closest neighbors to derive an imputed value. The method of aggregation varies depending on the nature of the variable. The distance between two observations is computed as the weighted mean of each variable's contributions, with weights reflecting the variable's significance (Kowarik and Templ, 2016). This study utilized the VIM package in R software to implement the kNN method.

missForest

The missForest algorithm is an enhanced version of the random forests approach, which employs an iterative imputation scheme, where it initially trains a random forest (RF) using observed values. It then predicts the missing values and continues this process iteratively (Vidal-Paz *et al.*, 2023). In this study, the missForest package in R software was used to implement this function.

Multivariate imputation by chained equations (MICE)

Multivariate imputation by chained equations (MICE) is an imputation method based on fully conditional specification. It utilizes separate models to impute incomplete attributes. MICE is capable of handling missing values across datasets with continuous, binary, and categorical attributes, employing a distinct model for each attribute (Li et al., 2024). This study utilizes the MICE R package to impute multivariate missing data, employing two mice functions: norm.predict and random forest (rf). The first method imputes the "optimal value" based on the linear regression model, commonly referred to as regression imputation, while the second method fills in the univariate missing data using random forests which stand out as a prevalent technique for reconstructing the missing values (Van Buuren and Groothuis-Oudshoorn, 2011).

Evaluation metrics

In this study, to evaluate different imputation methods under different scenarios, two evaluation metrics namely: Root Mean Square Error (RMSE) and Kling-Gupta efficiency (KGE) were utilized. RMSE is recognized as a predominant and extensively employed performance measure in imputation research (Jadhav *et al.,* 2019). KGE serves as a rigorous validation measure and is extensively applied in distributed hydrological modeling. KGE ranges from -∞ to 1, with a score of 1 representing an ideal model fit.

RESULTS AND DISCUSSION

Univariate approach

The assessment of two individual imputation methods, namely seasonal decomposition and Kalman smoothing was conducted in this study. As shown in Table 1, the analysis showed that the two methods perform differently depending on the variable considered. In the context of precipitation missing value imputation, the RMSE analysis highlights a discrepancy in the performance of the Kalman smoothing method, which generates negative values across all missing mechanisms, contrary to the nature of precipitation data where negative values are not feasible. As a result, this method is disregarded. On the other hand, the seasonal decomposition method yields imputed values with RMSE ranging from 7.194 to 16.736, 8.095 to 21.594, and 9.511 to 25.48 for MAR, MCAR, and MNAR mechanisms, respectively. Notably, the method demonstrates a pronounced level of efficacy particularly in scenarios where data exhibit random missingness (MAR), with subsequent performance comparable to instances of missing data characterized by a complete absence of discernible pattern (MCAR). In contrast, MNAR exhibits the least favorable performance, as the mechanism underlying data absence is contingent upon unobserved data, thus presenting heightened complexities for imputation methodologies. It is worth mentioning that at lower missingness levels (particularly at 5% and 10%) the performance significantly improves, as depicted in Fig. 3.

The seasonal decomposition method demonstrates superior performance in temperature imputation, showing RMSE values ranging from 0.514 to 2.462, 0.466 to 1.65, and 0.505 to 2.387 for MAR, MCAR, and MNAR mechanisms, respectively. Remarkably, the most consistent and optimal result occurs under MCAR for all levels of missingness. This signifies their reliability in imputing missing data when it conforms to a random pattern. Aligning well with the typical occurrence of temperature missingness under this mechanism. In contrast, Kalman smoothing imputation RMSE was slightly higher from seasonal decomposition under 5% and 10%, and the percentage of differences increased when the missingness

Table 1: RMSE for seasonal decomposition and Kalman smoothing methods

Missingness (%)	Method	MAR			MCAR			MNAR		
		Pre	Temp	RH	Pre	Temp	RH	Pre	Temp	RH
5	seadec	7.194	0.514	1.387	8.095	0.466	1.661	9.511	0.505	1.585
	KS_ARIMA	7.684	0.532	4.700	8.062	0.494	4.847	10.007	1.094	5.054
10	seadec	10.079	0.818	2.450	11.884	0.751	1.616	14.802	1.452	2.294
	KS_ARIMA	10.809	0.934	9.111	12.32	0.859	8.011	19.923*	3.268	7.563
20	seadec	12.788	1.175	2.817	14.945	1.000	3.571	21.009	1.268	5.096
	KS_ARIMA	17.034*	1.140	16.933	14.727	1.617	13.84	21.196*	1.646	9.039
30	seadec	16.736	2.462	6.301	21.594	1.650	4.774	25.480	2.387	6.490
	KS_ARIMA	33.076*	5.009	26.652	21.444*	2.973	6.180	25.749*	3.802	14.846

Note. asterisk (*) indicates negative values

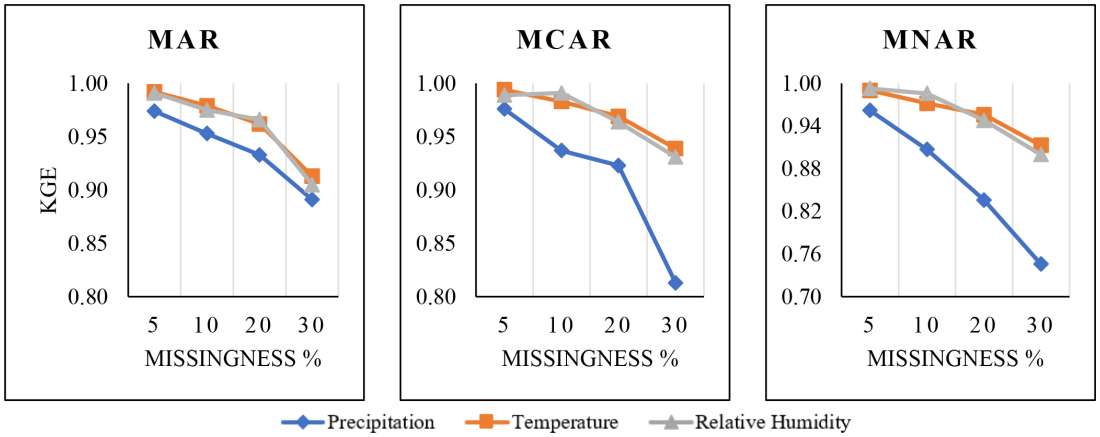

Fig. 3: KGE for precipitation, temperature, and relative humidity for seasonal decomposition method under all missingness types and levels

increased.

In the case of relative humidity imputation, the mean RMSE values for the seasonal decomposition method range between 1.387 and 6.301 for MAR, 1.661 and 4.774 for MCAR, and 1.585 and 6.49 for MNAR. Notably, the minimum and maximum values for the MAR mechanism are lower compared to those for MCAR and MNAR. Additionally, the mean RMSE tends to increase as the level of missingness increases for any missing data scenario, indicating that this method performs more effectively when dealing with fewer missing values as opposed to higher rates of missingness. In contrast, Kalman smoothing performs poorly under all missingness rates.

Multivariate approach

The effectiveness of four methods, namely missForest, kNN, norm.predict, and random forest (rf), encompassing two single imputation methods (missForest and kNN) and two multiple imputation methods (norm.predict and rf), were examined. The findings presented in Table 2 reveal significant insights into the efficacy of various methods for precipitation imputation under different missing data mechanisms. Indeed, employing the kNN method in scenarios governed by MCAR yields superior results across all levels of missingness as evidenced by the minimum RMSE value of 2.13. Contrary to initial hypotheses suggesting that lower missingness levels might correspond to reduced variability and potentially improved accuracy, our analysis demonstrates that kNN exhibits enhanced performance under MCAR at a 10% missingness level compared to a 5% level in precipitation imputation. In scenarios characterized by MAR, both missForest and kNN exhibit favorable performance across all rates. Particularly noteworthy is the superiority of kNN over missForest at 10% and 20% missingness levels, while missForest outperforms kNN at 5% and 30% missingness rates under MAR.

Conversely, under the MCAR mechanism, kNN outperforms missForest at 5% and 20% missingness rates. It is imperative to highlight that the imputation of negative values by the norm.predict method, which contradicts the inherent nature of precipitation, has been disregarded in our analysis. In the context of filling in missing temperature values, norm.predict emerged as the top performer under MAR across all levels of missingness, except for the 5% scenario where kNN yielded the best RMSE of 0.289. missForest ranked as second-best, followed by kNN. Under MCAR, norm.predict once again exhibited superior performance, followed by missForest and kNN, except at 10% missingness where kNN surpassed missForest. Conversely, under MNAR, norm.predict proved to be the best method at 5% and 10% missingness, with missForest following suit. Conversely, at 10% and 30%, missForest surpassed norm.predict. Notably, random forest performed the least effectively across all scenarios.

Regarding the relative humidity imputation, the

Table 2: RMSE for missForest, kNN, norm.predict, and random forest imputation methods under all missingness

Missingness (%)	Method	MAR			MCAR			MNAR		
		Pre	Temp	RH	Pre	Temp	RH	Pre	Temp	RH
5	missForest	4.822	0.333	0.246	6.266	0.219	0.594	6.191	0.348	0.608
	kNN	5.391	0.289	0.136	6.089	0.396	0.827	3.662	1.405	1.930
	norm.predict	3.989	0.335	0.648	4.866	0.216	0.712	5.375	0.272	0.960
	random forest	5.734	0.405	0.490	6.423	0.315	0.833	6.736	0.427	0.794
10	missForest	4.676	0.885	0.810	3.382	0.406	0.823	13.528	0.515	1.083
	kNN	3.352	0.863	0.901	2.132	0.370	0.859	14.605	0.588	1.318
	norm.predict	3.545	0.860	1.252	1.775	0.452	0.943	12.922	0.624	0.998
	random forest	5.618	1.089	1.454	1.734	0.617	1.086	6.736	0.427	0.794
20	missForest	9.932	0.682	1.654	7.609	0.911	1.285	9.528	0.613	1.926
	kNN	9.197	0.941	1.827	7.264	1.082	1.570	9.128	0.66	2.087
	norm.predict	7.743	0.658	1.855	5.374	0.921	1.508	8.235	0.584	1.907
	random forest	12.253	1.060	2.600	9.939	1.140	2.244	14.835	0.808	2.350
30	missForest	10.735	1.282	1.794	4.875	1.168	1.617	6.191	0.348	0.608
	kNN	11.025	1.324	1.815	3.662	1.405	1.930	12.502	1.337	2.493
	norm.predict	11.577	1.141	1.492	3.484	1.057	1.749	10.019	0.945	2.220
	random forest	13.106	1.357	2.479	5.444	1.926	2.559	17.483	1.570	3.133

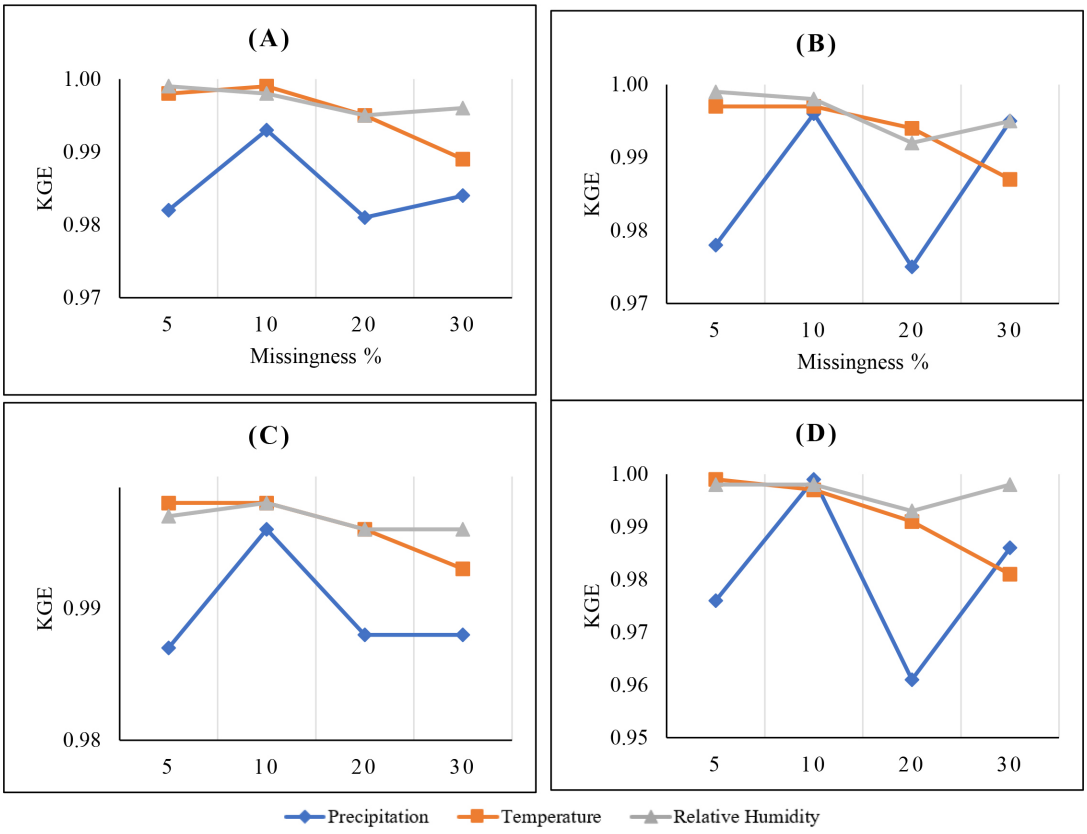

Fig. 4: KGE for precipitation, temperature, and relative humidity for (A) missForest, (B) kNN, (C) norm.predict, and (D) rf methods under MCAR missingness

missForest method demonstrated superior performance across all scenarios, except in instances of MAR where kNN and norm.predict were optimal at 5% and 30% missingness rates. Similarly, under MNAR, kNN and norm.predict emerged as the top performers at 10% and 20% missingness rates. It is noteworthy that random forest consistently exhibited the least effective performance across all scenarios. Fig. 4 shows the KGE results for the multivariate methods MCAR.

CONCLUSION

In summary, the presence of missing values in meteorological datasets represents a persistent challenge with significant implications for subsequent analyses, particularly in the context of extreme weather event forecasting and management, notably in developing countries. Inadequate handling of missing data prior to analysis can lead to erroneous conclusions and recommendations, compromising the integrity of the decision-

making processes. By assessing performance metrics like RMSE and KGE, it was determined that from the analysis of univariate imputation, the seasonal decomposition method exhibited favorable performance in accurately estimating missing monthly precipitation, temperature, and relative humidity data under all scenarios. Therefore, the seasonal decomposition method is recommended for imputation in univariate meteorological variables. From the analysis of the multivariate imputation, our findings indicate that among the imputation methods evaluated, the kNN method exhibited the highest efficacy for imputing missing precipitation values, closely followed by missForest. For missing temperature values, norm. predict emerged as the most effective method, with kNN performing admirably as well. In the case of the missing relative humidity data, the missForest method consistently demonstrated superior performance across all scenarios, followed by norm.predict.

It is noteworthy that this study specifically concentrates on imputing monthly data. It is plausible that the efficacy of these methods may differ in other geographical areas or with diverse types of climate data. Additionally, while our investigation primarily focuses on monthly imputation, it is conceivable that different temporal resolutions (e.g., daily) and spatial resolutions might necessitate the adoption of alternative imputation strategies.

ACKNOWLEDGEMENT

The authors thank the School of Civil Engineering, University Sains Malaysia (USM) for providing the necessary facilities for the research work.

Funding: This research did not receive any financial support.

Conflict of interest: The authors report no conflicts of interest.

Data availability: Monthly meteorological datasets for this research were obtained from the Iraqi Meteorological Organization and Seismology (IMOS).

Authors contribution: **K. Qaraghuli**: Conceptualization, Methodology, Visualization; Writing-original draft; **M. F. Murshed**: Supervision, Methodology, Reviewing and editing; **M. Azlin M. Said**: Supervision, Reviewing and editing; **A. Mokhtar**: Conceptualization, Reviewing; **I. Rousta:** Methodology, manuscript review.

Disclaimer: The contents, opinions, and views expressed in the research article are the views of the authors and do not necessarily reflect the views of the organization they belong to.

Publisher's Note: The periodical remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

REFERENCES

Afrifa-Yamoah, E., Mueller, U. A., Taylor, S. M. and Fisher, A. J. (2020). Missing data imputation of high-resolution temporal climate time series data. *Meteorol. Applications*, 27(1): 1-18. https://doi.org/10.1002/met.1873

Chaudhry, A., Li, W., Basri, A. and Patenaude, F. (2019). A Method for Improving Imputation and Prediction Accuracy of Highly Seasonal Univariate Data with Large Periods of Missingness. *Wireless Commun. Mobile Comput.,* 2019. https://doi.org/10.1155/2019/4039758

Diouf, S., Dème, E. H. and Dème, A. (2022). Imputation methods for missing values: the case of Senegalese meteorological data. *African J. Applied Statist.,* 9(1): 1245-1278. https://doi.org/10.16929/ajas/2022.1245.267

Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B. and Tabona, O. (2021). A survey on missing data in machine learning. *J. Big Data,* 8(1):140. https://doi.org/10.1186/s40537-021-00516-9

Jadhav, A., Pramod, D. and Ramanathan, K. (2019). Comparison of Performance of Data Imputation Methods for Numeric Dataset. *Applied Artificial Intel.,* 33(10): 913-933. https://doi.org/10.1080/08839514.2019.1637138

Kowarik, A. and Templ, M. (2016). Imputation with the R package VIM. *J. Statist. Software*, 74(7). https://doi.org/10.18637/jss.v074.i07

Li, J., Guo, S., Ma, R., He, J., Zhang, X., Rui, D., Ding, Y., Li, Y., Jian, L., Cheng, J. and Guo, H. (2024). Comparison of the effects of imputation methods for missing data in predictive modelling of cohort study datasets. *BMC Medical Res. Methodol.,* 24(1): 1-9. https://doi.org/10.1186/s12874-024-02173-x

Moritz, S., and Bartz-Beielstein, T. (2017). imputeTS: Time series missing value imputation in R. *R Journal*, 9(1), 207-218. https://doi.org/10.32614/rj-2017-009

Ou, H., Yao, Y. and He, Y. (2024). Missing Data Imputation Method Combining Random Forest and Generative Adversarial Imputation Network. *Sensors*, 24(4): 1112. https://doi.org/10.3390/s24041112

Saubhagya, S., Tilakaratne, C., Lakraj, P. and Mammadov, M. (2024). A Novel Hybrid Spatiotemporal Missing Value Imputation Approach for Rainfall Data: An Application to the Ratnapura Area, Sri Lanka. *Applied Sci.,* 14(3): 999. https://doi.org/10.3390/app14030999

Sridhara, S., Soumya B. R. and Kashyap, G. R. (2024). Multistage sugarcane yield prediction using machine learning algorithms. *J. Agrometeorol.,* 26(1): 37-44. https://doi.org/10.54386/jam.v26i1.2411

Umar, N. and Gray, A. (2023). Comparing Single and Multiple Imputation Approaches for Missing Values in Univariate and Multivariate Water Level Data. *Water (Switzerland)*, 15(8): 1-21. https://doi.org/10.3390/w15081519

Van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *J. Statist. Software*, 45(3): 1-67. https://doi.org/10.18637/jss.v045.i03

Vidal-Paz, J., Rodríguez-Gómez, B. A. and Orosa, J. A. (2023). A Comparison of Different Methods for Rainfall Imputation: A Galician Case Study. *Applied Sci.,* 13(22): 12260. https://doi.org/10.3390/app132212260