Ain Shams Engineering Journal 15 (2024) 102717

Contents lists available at ScienceDirect

Ain Shams Engineering Journal

journal homepage: www.sciencedirect.com

A comparative analysis of missing data imputation techniques on sedimentation data

Wing Son Loh [a], Lloyd Ling [b], Ren Jie Chin [b,*], Sai Hin Lai [c,d], Kar Kuan Loo [a], Choon Sen Seah [e]

[a] Department of Mathematical and Actuarial Sciences, Lee Kong Chian Faculty of Engineering and Science, Universiti Tunku Abdul Rahman, 43000 Kajang, Malaysia
[b] Department of Civil Engineering, Lee Kong Chian Faculty of Engineering and Science, Universiti Tunku Abdul Rahman, 43000 Kajang, Malaysia
[c] Department of Civil Engineering, Faculty of Engineering, Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia
[d] UNIMAS Water Centre (UWC), Faculty of Engineering, Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia
[e] Faculty of Computing, Universiti Teknologi Malaysia, 81310 Johor Bahru, Malaysia

ARTICLE INFO

Keywords:
Missing data
Imputation techniques
artificial neural network (ANN)
Fine sediment
Sedimentation

ABSTRACT

Sediment data pertains to various hydrological variables with complex sediment hydrodynamics such as sedimentation rates which are often incompletely presented. Thus, the availability of sedimentation data is of utmost necessity for data accessibility. A comparative analysis on the missing fine sediment data imputation performance was made based on four different techniques, namely the k-Nearest Neighbourhood (k-NN), Support Vector Regression (SVR), Multiple Regression (MR), and Artificial Neural Network (ANN), under the single imputation (SI) and multiple imputation (MI) regimes. Across different missing data proportions (10%-50%), the ANN demonstrated optimal results with consistent performance metrics recorded over both SI and MI regimes. For the highest missing data proportion (50%), the ANN presented the best imputation performance with a reported root mean squared error (RMSE) 0.000882, mean absolute error (MAE) 0.000595, coefficient of determination (R^2) 71%, and Kling-Gupta Efficiency (KGE) 72%. The imputation performance ranking is as follows: ANN, SVR, MR, and k-NN.

1. Introduction

1.1. Background and problem Statement

The transport mechanism of sediment particles constitutes a critical aspect of the hydrological cycle, influencing the sustainability of the aquatic ecosystems, balance of water quality and quantity, maintaining the aquatic habitat conditions, and the overall ecosystem preservation. Throughout the recent years, the intensified anthropogenic activities stemming from urbanisation, timber extraction, and agriculture have introduced heavy sediment loads into the locations of dams, rivers and oceans, carrying detrimental impacts to both the environment as well as the economy [1,2]. The motion of fine sediment particles during the settling process in water bodies wields substantial influence towards siltation rates [3]. Additionally, it is common that real data derived from the hydrological studies typically encounters data incompleteness issue such as instrumental failures or budget constraints [4]. Thus, the pivotal role of missing data imputation techniques must not be trivialized in the context of sedimentation data. In fact, the existence of missing data presents an obstacle in deciphering the complex sediment hydrodynamics such as sedimentation rate of fine sediments in water [5]. Furthermore, lacking of a complete series of data and / or using inaccurate data values for analysis would produce misleading results and eventually lead to invalid research studies and decisions being made. In order to properly handle missing data without sacrificing the data reliability and validity, appropriate imputation techniques must be considered. In this regard, different types of imputation techniques were analyzed and compared in this study, ranged from basic methods, to complex and algorithm based modeling techniques. The missing data imputation process was carried out on the missing sedimentation database based on four stipulated missing proportion, 10 %, 20 %, 30 %, 40 %, and 50 %. The proportion of missing data is a dominant factor in the studies of missing data imputation as the availability of the complete oservations from the data set reduces [6]. Past literatures had suggested that a common range between 10 % and 50 % of missing proportion was adapted in missing data related studies [6,]. Based on the rule of thumb, the underlying assertion regarding the missing proportion in this study is that the missing data imputation procedure is not cost effective and is

* Corresponding author.
E-mail address: chinrj@utar.edu.my (R.J. Chin).

https://doi.org/10.1016/j.asej.2024.102717
Received 30 September 2023; Received in revised form 13 February 2024; Accepted 18 February 2024
Available online 11 March 2024

considered to be insignificant whenever the missing proportion is below 5 % [7]. On the contrary, excessive missing data has extremely high potential of introducing bias to the analysis as a result of an imbalanced data set [8]. As a consequence of a biased analysis which was extracted from analyzing the remaining available data from a largely incomplete data set, biased estimated parameters with high error fluctuation will be produced. Besides, the characterization of the data depends heavily on the completeness of the data set and thus there will be a high likelihood that the data that were missing carried significant properties and influential information from the original complete data set. Such results hold utterly deficient statistical power which would hinder the computed statistical analyses [8,9].

1.2. Missing data mechanisms

Over the past decades, it had been widely recognized that issues invited by the presence of missing data is a pervasive concern within a multitude of hydrological databases. Such examples encompasses of missing observations from precipitation data [10], riverflow data [7], rainfall and runoff data [10], water quality index data [8,12], and sediment load data [5]. There were handful of factors that contributes to the presence of missing sedimentation data. For instance the discrepancy in calibration readings [10], ramification of defective sensor components and failure of in-situ measuring instruments [13,14], occurrence of unexpected catastrophic disasters like landslides and flash floods due to excessive downpour of stormwater [15], and error-prone manual data entry processes [16].

While it is vital to develop a reliable and technically sound approach to impute the missing sedimentation data, the missing mechanisms must be understood to ensure the imputation techniques appropriately address the underlying association between the studied variables as well as the probability of the observed data that is missing [17,18]. Generally, the types of missing data mechanisms can be broken down into three principal categories, which are the missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) [19].

First and foremost, the MCAR mechanism suggests the scenario of an almost zero or absolute absence of a relationship between the dependency of the variables observed and the likelihood of the unobserved data being missing [15]. In other words, missing data classified under the MCAR mechanism assumes that the original data value is fully independent of the missingness, which is completely random. Cases such as missing recorded data due to the inappropriate use of measuring tools, impaired laboratory equipments, non-responsive data transmissions and overlooked value caused by human related errors are clear examples from the MCAR mechanism [20]. MAR instead interprets the missingness to be related to the observable complete data values, but is unrelated to the unobserved missing data values. Hence, it can be said that MAR claims that non-available missing data as a result of disregarded records follow a random stochastic manner which is predictable from the data pattern discoverable from the observed data [21]. Last but not least, the MNAR missing mechanism states that the missingness of the unobserved data is directly associated with the other missing unobserved data values. This means that the likelihood of the data point being missing with the observed data supplied, has full dependence of the remaining unobserved missing value, and completely independent of the observed complete data set. In this regard, the MNAR mechanism is known to be the most challenging missing mechanism to address [20,22].

By defining a general set of data matrix, $\boldsymbol{D}$ that consists both the observable and missing data variables denoted by $\overrightarrow{D}_O$ and $\overrightarrow{D}_M$ respectively, the interconnected relationship between the different variables based on the data missingness could be visualized in Fig. 1, where Q represents the cause of the missingness that is unrelated to the $\overrightarrow{D}_M$, and R represents the resulting missingness.

More specifically, the likelihood of the sample observation, $\underline{\theta}$, associated with the missing data patterns that are described by the three distinct missing mechanisms can be expressed in accordance with the mathematical equations for MCAR (Eq (1), MAR (Eq (2), and MNAR (Eq (3) [19].

$$\Pr\left(\underline{\theta} \in \left(\overrightarrow{D}_O, \overrightarrow{D}_M\right)|\boldsymbol{D}\right) = \Pr\left(\underline{\theta} \in \left(\overrightarrow{D}_O, \overrightarrow{D}_M\right)\right)\forall \boldsymbol{D}_{ij} \tag{1}$$

where $\underline{\theta}$ is independent of $\overrightarrow{D}_O$ and $\overrightarrow{D}_M$.

$$\Pr\left(\underline{\theta} \in \left(\overrightarrow{D}_O, \overrightarrow{D}_M\right)|\boldsymbol{D}\right) = \Pr\left(\underline{\theta} \in \left(\overrightarrow{D}_O, \overrightarrow{D}_M\right)\right)\forall \boldsymbol{D}_{ij} \tag{2}$$

where $\underline{\theta}$ is independent of $\overrightarrow{D}_M$.

$$\Pr\left(\underline{\theta} \in \left(\overrightarrow{D}_O, \overrightarrow{D}_M\right)|\boldsymbol{D}\right) = \Pr\left(\underline{\theta} \in \left(\overrightarrow{D}_O, \overrightarrow{D}_M\right)\right)\forall \boldsymbol{D}_{ij} \tag{3}$$

where $\underline{\theta}$ is dependent of $\overrightarrow{D}_M$.

1.3. Missing data imputation techniques

In the past decade, there were a large number of studies carried out to perform missing data imputation accross various fields such as applications in financial data [23], biological gene expressions [24], educational production functions [25], ground electromagnetism from the magnetic data acquisition system [20], drill cutting settling rate predictotion [26], and more. Nevertheless, missing data imputation is also actively being researched in the context of missing hydrological databases as mentioned previously. The nature of imputation techniques could be generally grouped into two variations, namely the theoretical based imputation technique, and the empirical based (i.e. function modelling) imputation technique [10,27]. In most cases, the theoretical based approach requires fundamental theories derived from the domain knowledge of the specific field. Such approaches are usually supported by a list of theoretical assumptions which are required to be satisfied.

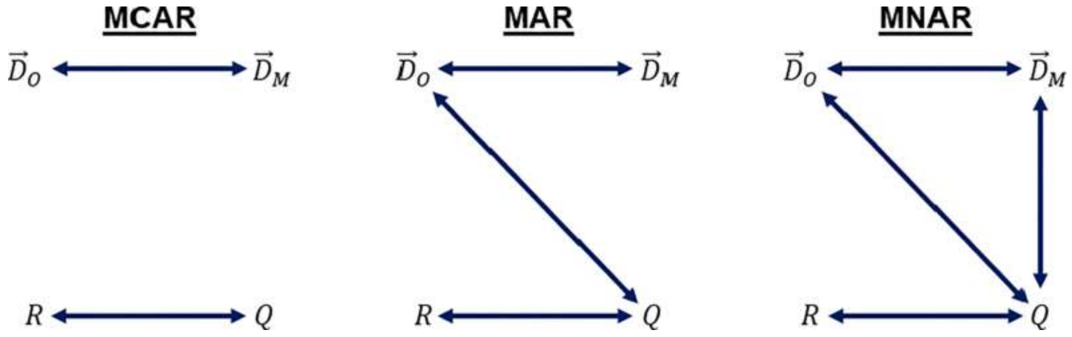


Fig. 1. Missing mechanism relationship illustration.

Hence, to solve the missing data problem via the traditional theory based methods, researchers have to clearly identify the underlying distribution of the data set, ensuring the aspects of data characteristics and data quality aligned with all of the distribution assumptions.

On the contrary, empirical based approaches gain information through inferences that is drawn from the real-world data [28]. Empirical methods are prevalent due to the flexibility in the modelling technique and computational algorithms applied. The most straightforward and well known technique of all is the simple arithmetic average (SAA), which imputes the missing data point values according to the computed arithmetic average of the variable of interest [10,29]. Although the SAA technique is simple enough, the imputed missing data values were undesirable in almost all cases, especially when the variation within the data variables were large, or/and when outliers existed. Similarly, the simple median (SM) imputation replaces the unobserved missing data with the computed median of the variable of interest. Median presents a robust statistical measure compared to mean, and studies have also shown that the SM imputation technique yielded slightly improved results compared to the SAA technique in the presence of data outliers [6,29].

Besides SAA and SM, hot deck (HD) is another common imputation technique that was commonly used [20,30]. Under the HD technique, missing data values were substituted with the most similar observed data point based on the closest neighbour which is assumed to best resemble the missing observation itself. Some studies have applied a similar approach known as the k-nearest neihbourhood (k-NN) imputation technique [29]. Overall, the aforementioned approaches were not developed based on the idea of function fitting but rather relied on the closeness between the missing data and the *k* observed neighbouring point(s). Therefore, such mechanisms have high potential of introducing an inductive bias towards the imputation techniques [31].

Furthermore, regression models are one of the alternatives of the missing data imputation techniques. Multiple regression (MR) is an extended version of the basic linear regression model, enabling not just an interpretation of a bivariate linear relationship, but from a multivariate aspect, with a richer analysis on the correlation of the investigated variables [15]. Subsequently, a multivariate relationship is established between the explanatory variables and the response variable under the MR imputation technique, and the combinations between the variables with the corresponding best weight coefficients is finalized for the missing data imputation. Based on several studies on the missing hydrological data imputation (e.g. wind, temperature and rainfall), it had been discovered that the MR model outperformed the SAA and k-NN techniques respecively by producing a higher estimation accuracy [32,33].

In addition to the MR imputation technique, the support vector regression (SVR) pressented a non-traditional approach, extended from the support vector machine (SVM) model. Particularly, SVR performs regression for estimations in contrast to the SVM which performs classification for target data labels. Briefly, the SVR creates a hyperplane and performs estimation based on the location relative to the decision boundary lines with a specific margin. The strength of SVR lies in its robustness against data variations, with no underyling distributional assumptions [26].

On top of the regression models, the artificial neural network (ANN) model garnered immense professional acclaim and popularity especially in the recent years revolutionized by artificial intelligence (AI). An ANN is a general model under the subset of the machine learning category where it mimics the human brain, applying the same mechanism to facilitate machines to learn from instances (i.e. training), and generate a prediction output [34]. There is an abundance of types of ANN models, each with different layout structures as well as training mechanisms, serving for different prediction purposes. Despite of the various available types of ANN models, the basic layout structure of a typical ANN as illustrated in Fig. 2 follows the standard design by having a single input layer, followed by one or more hidden layers, and a single output layer at the terminal part [3]. Nodes representing input data in the input layers are called input nodes whereas the final node that provides the output value is called the output node. Similarly, nodes found within the hidden layer(s) are known as hidden nodes.

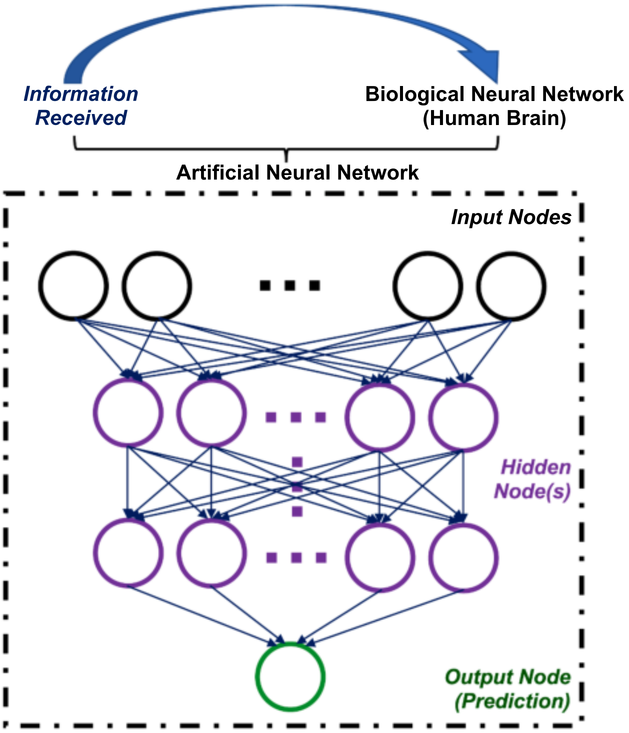


Fig. 2. General layout structure of an ANN model mimicking information received by biological neural network in human brain.

Each layer contains information transmitters known as the nodes which is similar to the biological neurons in the human brain, feeding input signals to the subsequent connected nodes through a learning mechanism, stimulated by the input instances. In addition, for a typical ANN model, adjacent nodes in the same layer are not interconnected. Furthermore, in the ANN model training process, the nodes processes the input data and computes a temporary net value, Ω to be fed to each of the subsequent nodes in the next layer. Iteratively in a feed-forward manner, the training continues layer by layer within each of the nodes from the consecutive layers, until the net output value is finally passed to the terminal node where a prediction output is generated [35]. Each node is associated with two vital components known as the weight, **W** and bias, **B,** and is assigned with a pre-defined transfer function. As the name suggested, the weight term informs a particular node on the weightage of receiving the input data **X**, to calculate the net output value that are generated by each of the previous nodes as shown in (Eq. (4)) [29].

$$\Omega = \mathbf{W} \bullet \mathbf{X} + \mathbf{B} \tag{4}$$

where Ω is the net value received at each node; **W** and **B** represents the weights and biases matrices respectively.

The bias term enables the ANN model to offset the null weighted sum, preventing the nodes from ignoring a particular node during the training process. The transfer function must meet the monotonicity criteria, and must be differentiable at all points. It processes the weighted sum and computes the output value at each node based on the pre-defined function [36]. Table 1 shows the commonly used examples of transfer function, including the linear function, logistic (i.e. sigmoid) function, hyperbolic tangent function (i.e. tanh), and the rectified linear unit (ReLU) function. Although the linear function is a choice for the transfer function, it is not implemented as most study requires ANN

Table 1
Commonly used examples of transfer functions [29,36].

Type of Transfer Function	Mathematical Formula
Linear	$\phi(\Omega) = \Omega$
Logistic	$\phi(\Omega) = \dfrac{1}{1+e^{-\Omega}}$
Hyperbolic Tangent	$\phi(\Omega) = \dfrac{2}{1+e^{-2\Omega}} - 1$
Rectified Linear Unit	$\phi(\Omega) = \begin{cases} 0, \Omega < 0 \\ \Omega, \Omega \geq 0 \end{cases}$

models to address the non-linear complexity among variables during the training process.

Notably, the ANN models, also known as universal approximators due to the non-parametric modelling mechanism are possible to approximate any form of functions [36,37]. Together with the stellar performance across the estimation performance in numerous hydrological applications [6], the ANN model is an excellent candidate with great potential of producing highly accurate results.

Despite of the dozens of missing data imputation techniques discussed, the case deletion method may be a solution, subjected to the percentage of missing data and the condition of missing mechanism. Based on the literatures, the deletion method includes listwise deletion and pairwise deletion [38,39]. Given the missing proportion is 5 % or less, the listwise deletion technique can be used to discard the entire observation if one or more variables is found missing. Pairwise deletion on the other hand was more preferred over listwise deletion as it attempts to preserve the size of the available data set. In pairwise deletion, a statistical analysis, usually the correlation between the pair of missing data variables is assessed, and the pair of variables will be deleted together if the correlation is insignificant. Else, other imputation methods will be used to replace the missing data points [40]. However, the use of the deletion techniques are only suitable in the context of the MCAR mechanism.

A summary on the performances of the existing estimation approach from different studied was presented in Table 2. However, the results from the studies cannot be compared directly due to different application type and context. In particular, the commonly used error metrics such as the mean absolute error (MAE) and the root mean squared error (RMSE) from the studies possessed distinct unit measurements, which leaded to the highly disparate in their values. Hence, the error metrics were not able to directly differentiate model performances across different studies. However, a smaller error value is always preferred in all cases of model estimations. Fortunately, the unitless coefficient of determination, R^2 is a sensible indicator to compare model performances. In general, it could be examined that the ANN technique outperformed other modelling techniques since a higher R^2 was discovered across the studies. The ANN exhibited superior performance regardless of the type of the study which consisted missing data, or has no missing data at all (i.e. 0 % missing percentage).

Nevertheless, possessing a complete and valid data set in the context of sedimentation studies is highly significant as the hydrodynamic properties of sediment particles plays an important role in governing the sediment transporation mechanism, affecting siltation rates. The primary objective of this study is to compare the efficacy of applying some of the most renowned missing data imputation techniques on the missing sedimentation data set. Particularly, a comparative analysis on the missing data imputation techniques was performed by implementing the k-NN, SVR, MR as well as the popular ANN.

2. Methodology

2.1. Missing sedimentation data set and workflow

The original complete sedimentation data set was collected from the particle image velocimetry (PIV) experiment. The PIV machine captured instantaneous images of each of the fine seedling particles in the water with laminar flow from the sedimentation basin and then the particle positions as well as other hydrokinetic properties such as the particle sedimentation rate were computed via the Dantec Dynamics PIV software [3,46,47]. The sedimentation data set comprised of 6240 complete observations of the captured sedimentation rate of fine particles associated with four hydraulic parameters which were the fine seedling particle sizes within 5 μm to 50 μm; the inlet depth of water flow ranged between 6 cm and 10.5 cm, and the horizontal as well as vertical position (in pixels) of the particles that was extracted from the captured images under the PIV experiment [46]. Table 3 presents the descriptive statistics of the studied variables in the fine sediment data set.

Based on the MAR mechanism, the missing sedimentation rate data points were simulated at a stipulated proportion of 10 %, 20 %, 30 %, 40 %, and 50 %. The missing data points were imputed by the selected imputation techniques according to two distinct imputational procedures, called the single imputation (SI) and multiple imputation (MI). All the selected missing data imputation techniques were implemented under both procedures. At the end of the procedures, each of the imputation techniques were evaluated based on the three performance indicators which includes the mean absolute error (MAE), root-mean squared error (RMSE), the coefficient of determination ($R^{2.}$), and the Kling-Gupta Efficiency (KGE) [48]. The best imputation technique was selected at the final stage of the comparative analysis. The general workflow of this study is illustrated in Fig. 3.

2.2. Data pre-processing

The sedimentation data set were pre-processed before used. Firstly,

Table 2
Performance of the common estimation techniques in different application.

Source	Context of Study	Maximum Missing Percentage	Techniques Applied	MAE	RMSE	R^2
[41]	Total Suspended Solid	0 %	ANN	NIL	0.475—0.993	0.623–0.920
[42]	Local Scour Depth DownStream	0 %	ANN	0.130	NIL	0.740
			MR	0.200	NIL	0.670
[43]	Suspended Sediment Concentration	0 %	ANN	28.06–29.89	50.00–61.74	0.470–0.630
			MR	31.55–64.93	105.88–210.46	0.560–0.625
[44]	Suspended Sediment Load	21 %	ANN	820–1614	1646–4731	0.65–0.97
			MR	3285–3246	5444–7535	0.59–0.75
[45]	Streamflow	30 %	k-NN	NIL	0.313	0.644
[22]	Air Quality	40 %	k-NN	0.852	1.067	NIL
[26]	Drill Cuttings Settling Velocity	40 %	ANN	0.065	0.090	0.612
[18]	Water Quality	45 %	k-NN	NIL	0.411	0.4823
[20]	Ground Electromagnetism	80 %	SAA	9.432	10.604	NIL
			Hot-Deck	7.93	10.130	NIL
			k-NN	2.600	7.282	NIL
			SVR	0.510	0.560	NIL
			ANN	0.100	< 0.100	NIL

Table 3
Descriptive statistics of the studied variables in the fine sediment data set.

Variables	Unit	Mean	Median	Minimum	Maximum	Standard Deviation
Fine particle size	μm	25.8974	20.0000	5.0000	50.0000	18.3509
Inlet depth	cm	8.7747	9.0000	6.0000	10.5000	1.6007
x- particle position	cm	13.8357	12.0000	0.0000	42.0000	9.3779
y- particle position	cm	8.7840	10.0000	0.0000	20.0000	5.6787
Sedimentation rate	m/s	0.0017	0.0009	8.4900 x 10^{-8}	0.0844	0.0026

the outlier observations which could be caused by the variability or discrepancies in the data values recorded from the experimental measurements were removed. This is because the presence of outlier is more susceptible to errors, which would highly affect the quality of the data set as well as the missing data imputation results. After the outlier observations were removed, each data variable, z was rescaled to obtain a unitless value, z_{norm} via the min–max scaler equation as defined in (Eq (5) [49].

$$z_{norm} = \frac{z - z_{min}}{z_{max} - z_{min}} \tag{5}$$

The purpose of applying the min–max data normalization process was to improve the variable interpretability as each variable was projected onto a value bounded between 0 and 1, avoiding the dominance of variables due to larger unit values, while the data set distribution properties and their characteristics were preserved. In addition, the normalized data set allowed the computational task to be simplified [49].

2.3. Single imputation (SI)

The SI imputational procedure generates a single estimated value for each of the missing data points. In this study, the imputation techniques of k-NN, SVR, MR, and ANN were selected for the comparative analysis. The computational steps of each of the imputation technique to obtain the imputation value $\widehat{\theta}$ were provided below:

2.3.1. k-Nearest Neighbourhood (k-NN)

Under the k-Nearest Neighbourhood technique, each of the missing data value was imputed based on the simple arithmetic average according to the k-nearest data reference points towards the input data learned (Eq (6) [11]. The Euclidean distance metrics was used as the reference measure for the k nearest data.

$$\widehat{\theta}_i = \widehat{\theta} = \frac{\sum_{i=1}^{k} y_i}{k} \tag{6}$$

where y is the observable data values based on the variable of interest.

2.3.2. Support vector regression (SVR)

Under the SVR technique, each of the missing data value was imputed based on the modelled hyperplane that best fits the missing data points. Using the applied kernel function Φ, the input data variables x_i were transformed into a high dimensional feature space for which the hyperplane minimizes the margin distance between the hyperplane and the nearest data points. The regression function was computed based on (Eq (7) [50].

$$f(x_i) = \underline{w}^T \Phi(x_i) + \mathrm{b} \tag{7}$$

The SVR fits a hyperplane which will be located geometrically at the middle of the boundary lines and only input data points within the decision boundary lines of minimum error rate around the hyperplane were considered. The minimization process was performed based on (Eq (8) with a band width margin $\pm\varepsilon$, and slack factors $\xi_i^{\ +}$, $\xi_i^{\ -}$ [50].

$$\min_{\underline{w},b,\xi,\xi^*} \frac{1}{2}\underline{w}^T \underline{w} + \frac{\lambda}{2} \sum_{i=1}^{N} (\xi_i^{\ +} + \xi_i^{\ -}) \tag{8}$$

The constraints involved in the minimization process are defined in (Eq (9).

$$-(\varepsilon + \xi_i^{\ -}) \le y_i - f(x_i) \le \varepsilon + \xi_i^{\ +} \tag{9}$$

where $\xi^-, \xi^+ \ge 0$. Both factors equals to zero if the data points fall within the boundaries.

The applied non-linear Gaussian kernel function in the SVR is defined in Eq. (10).

$$\Phi(x_i) = e^{-\frac{\|x - x_i\|^2}{2\sigma}} \tag{10}$$

where σ is the variance hyperparameter of the Gaussian kernel function.

2.3.3. Multiple regression (MR)

Under the MR technique, each respective missing data value was imputed based on the mathematical regression formula as shown in Eq (11) [10,44].

$$\widehat{\theta}_k = \beta_0 + \beta_1 y_{k1} + \beta_2 y_{k2} + \beta_3 y_{k3} + \beta_4 y_{k4} + \epsilon \tag{11}$$

where y_k is the observable data values of each of the non-missing variable data points, β_0 is the constant (intercept) term, β_1, β_2, β_3 and β_4 are the slope coefficients, also known as the regression weights for each of the variables, and ϵ is the residual (random error) term.

2.3.4. Artificial neural network (ANN)

The ANN employed in this study combined both the feed-forward as well as the back-propagation learning rule in the model training phase. Through the trial and error approach, the number of nodes were explored for a single hidden layer. The ANN model was trained with the available observations first through the feed-forward rule from the input layer, to the hidden layer, then the terminal node in the final output layer. The moment the feed-forward learning completed, the output value will be compared with the true data observations. The sum of squared error was calculated, and this was where the back-propagation rule commence. The weight parameters of the previous nodes were adjusted in a backwards passing fashion based on the computed sum of squared error [49]. For the ANN model in this study, the logistic transfer function was employed.

2.4. Multiple imputation (MI)

As the name suggested, the MI imputational procedure generates multiple estimated values (i.e. point estimates) from the fitted models to each of the imputed data set containing the missing data points [28]. In this particular study, the bootstrap resampling was integrated with the MI procedure [51] was carried out with 100 sets of sample data that were obtained under the bootstrap resampling method. The integrated boostrapping with the MI approach is a straightforward application of the standard percentile-based bootstrap confidence interval for the estimator [17]. The conceptual framework of the bootstrap resampling method was illustrated in Fig. 4, to collect a total of 10 bootstrap

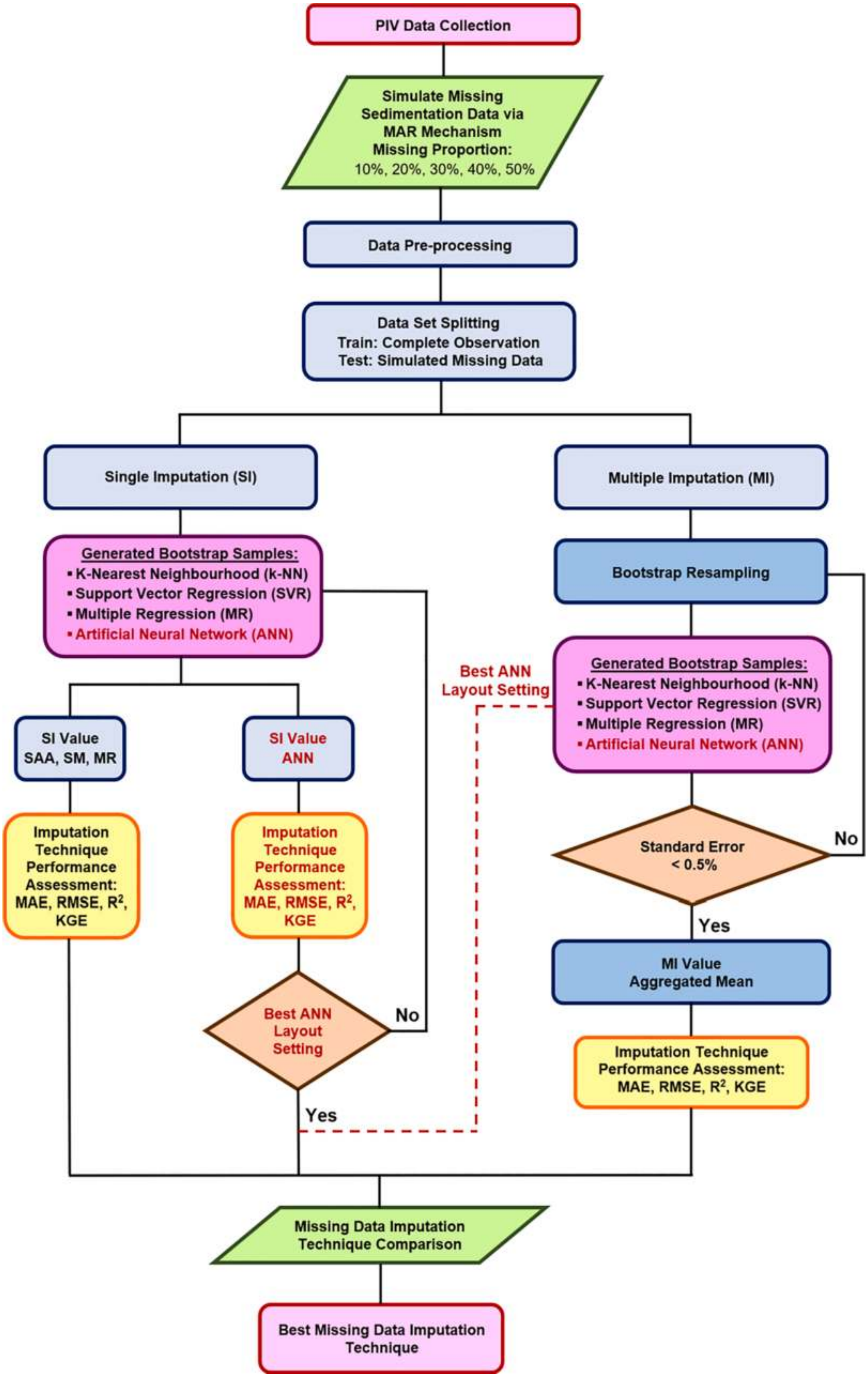

PIV Data Collection
Simulate Missing Sedimentation Data via MAR Mechanism Missing Proportion: 10%, 20%, 30%, 40%, 50%
Data Pre-processing
Data Set Splitting Train: Complete Observation Test: Simulated Missing Data
Single Imputation (SI)
Multiple Imputation (MI)
Generated Bootstrap Samples:
▪ K-Nearest Neighbourhood (k-NN)
▪ Support Vector Regression (SVR)
▪ Multiple Regression (MR)
▪ Artificial Neural Network (ANN)
Bootstrap Resampling
Best ANN Layout Setting
Generated Bootstrap Samples:
▪ K-Nearest Neighbourhood (k-NN)
▪ Support Vector Regression (SVR)
▪ Multiple Regression (MR)
▪ Artificial Neural Network (ANN)
SI Value SAA, SM, MR
SI Value ANN
Imputation Technique Performance Assessment: MAE, RMSE, R², KGE
Imputation Technique Performance Assessment: MAE, RMSE, R², KGE
Standard Error < 0.5%
No
Yes
MI Value Aggregated Mean
Best ANN Layout Setting
No
Yes
Imputation Technique Performance Assessment: MAE, RMSE, R², KGE
Missing Data Imputation Technique Comparison
Best Missing Data Imputation Technique

Fig. 3. General workflow based on the comparative analysis for the missing data imputation techniques.

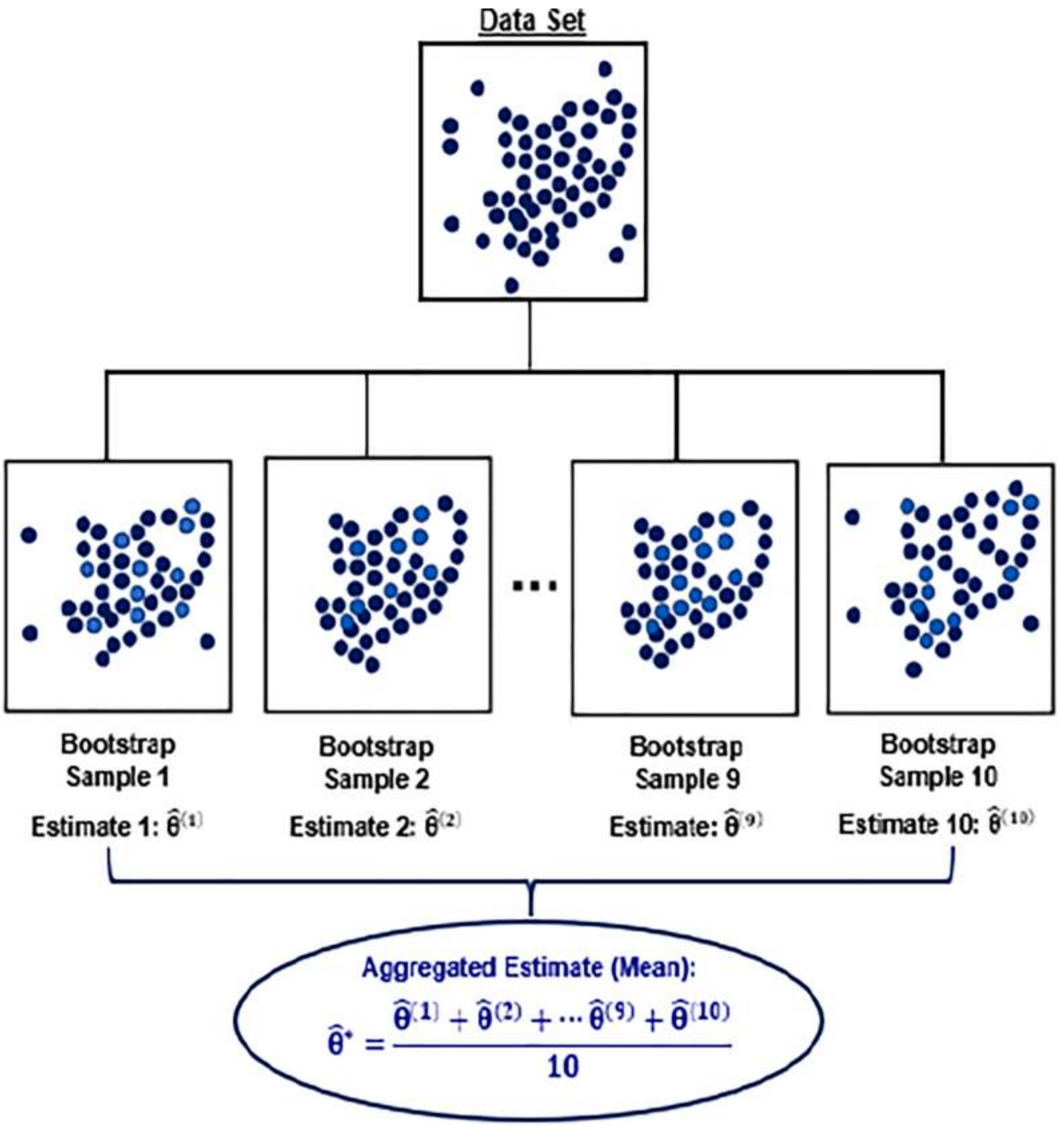

Data Set
Bootstrap Sample 1
Estimate 1: $\hat{\theta}^{(1)}$
Bootstrap Sample 2
Estimate 2: $\hat{\theta}^{(2)}$
...
Bootstrap Sample 9
Estimate: $\hat{\theta}^{(9)}$
Bootstrap Sample 10
Estimate 10: $\hat{\theta}^{(10)}$
Aggregated Estimate (Mean):
$\hat{\theta}^{*} = \frac{\hat{\theta}^{(1)} + \hat{\theta}^{(2)} + \cdots \hat{\theta}^{(9)} + \hat{\theta}^{(10)}}{10}$

Fig. 4. Illustration of the bootstrap resampling method.

samples.

Applying the bootstrap method, the same size of data samples were obtained through sampling with replacement. Hence, it is definitely possible to have repeated data (indicated by the light blue sample points in Fig. 4) being resampled in the formation of more robust bootstrap samples. After producing 10 sets of bootstrap samples, the missing data imputation techniques were applied. As a result, 10 sets of imputed values of the missing data points will be generated. Then, the imputed results from the 10 sets of bootstrap samples were aggregated by computing their means. To ensure the reliability of the boostrap samples, the imputed values must not have a standard error greater than 0.5 %. Once the reported standard error exceeds 0.5 %, the bootstrap method will be reinitiated to produce another 10 sets of boostrap sample, and the previous sets will be discarded. In this study, the best network layout setting of the ANN model was determined under the SI procedure under the trial and error method separately for the 10 %, 20 %, 30 %, 40 %, and 50 % missing proportion. The best ANN layout settings were implemented in the MI procedure based on the corresponding data missing proportion.

2.5. Performance metrics

The performance metrics, namely the mean absolute error (MAE) in (Eq (12), the root-mean squared error (RMSE) in (Eq (13), the coefficient of determination (R^2) in (Eq (14), and the Kling Gupta Efficiency (KGE) in (Eq (15) were applied to evaluate the performance of each of the missing data imputation techniques in estimating the missing values of the sedimentation rates [48].

$$MAE = \frac{\sum_{i=1}^{N} |y_i - \bar{y}_i|}{M} \tag{12}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (\bar{y}_i - y_i)^2}{M}} \tag{13}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (\hat{y}_i - y_i)^2}{\sum_{i=1}^{N} (\bar{y}_i - y_i)^2} \tag{14}$$

$$KGE = 1 - \sqrt{(1 - R)^2 + (1 - \alpha)^2 + (1 - \beta)^2} \tag{15}$$

where the ratio between estimated and observed mean (bias ratio),

$$\alpha = \frac{\sum_{i=1}^{N} \widehat{y}_i}{\sum_{i=1}^{N} y_i}$$

the ratio between estimated and observed standard deviation (variability ratio),

$$\beta = \frac{N\sum_{i=1}^{N}\left(\widehat{y}_i^{\,2}\right) - \left(\sum_{i=1}^{N}\widehat{y}_i\right)^2}{N\sum_{i=1}^{N}\left(y_i^{2}\right) - \left(\sum_{i=1}^{N}y_i\right)^2}$$

The MAE and RMSE are error metrics, for which the lower the error value based on the imputed results, the more favourable the imputation technique. R^2 on the contrary quantifies the amount of variability that the imputation techniques addressed by comparing the imputed value and the observed value. The value of R^2 usually lies between 0 and 1, signifying the strength of the model in capturing the variability of the predicted data. More specifically, the estimations could be considered as perfect if the value of R^2 is exactly one, whilst a zero R^2 value suggests the equivalence of the estimated error variation and the deviation between the mean and data points. In other words, the zero R^2 value signifies that the used model for imputation has an equal performance compared to the case of estimation using the mean. Furthermore, the R^2 if found to fall below zero (negative), it suggests that the model had a poorer imputation performance as compared to when using the mean as estimator [48]. The mathematical interpretation of this consequence can be inspected from the method of computing the R^2, for which a poor prediction assoaciates with a very large difference between the predicted and observed data resulting in the large numerator, dominating the denominator which takes the total squared difference between the individual data and their mean. On the whole, the imputation techniques should account for the unvertainty about the missing observations and so, higher R^2 value is always preferred. On top of that, the KGE has been prominently applied to assess the mode calibration and evaluation performances, where it is able to confront the inadequacy of the R^2 metric in terms of the variability and bias. Extended beyond the correlation index term, R, the constitutive components of the KGE are collectively considered to be more comprehensive than other indicators although they cannot be directly compared due to the distinct dependence of the coefficients. Similarly, the KGE of value 1 indicates a perfect match between imputations and observations. Moreover, the zero value in R suggests the particular imputation model has equivalent explanatory power as the mean estimator. On the contrary, negative KGE values indicates that the model suffers from poor imputation results. Also, while most studies suggested that a positive KGE reflects that the simulated output results were better than when the mean was used as the estimator, there were several cases which allude that the prediction results are less satisfactory when the value of KGE is 0.5 or less [48,52,53]. It should be noted that the mentioned results from the different studies were analyzed based on the estimation using a fully complete data set.

3. Results and discussion

3.1. Single imputation (SI) results

The optimum number of hidden nodes which allowed for the best ANN layout setting was searched through a trial and error approach. Table 4 below provides the statistical performance metrics for each of the combination based on the layout setting of the ANN model with the 10 % stipulated missing proportion. It is clear that the layout setting of 4–13-1 produced the lowest MAE and RMSE, as well as the highest R^2 and KGE values.

Table 4
Missing data imputation results for 10% missing proportion (ANN) under the SI regime.

Number of Hidden Nodes	MAE	RMSE	R^2	KGE
1	0.000651	0.000839	0.631794	0.655818
2	0.000582	0.000761	0.695583	0.708435
3	0.000570	0.000779	0.681679	0.653111
4	0.000511	0.000699	0.744196	0.765093
5	0.000563	0.000751	0.714241	0.782966
6	0.000514	0.000696	0.745174	0.754854
7	0.000500	0.000682	0.756698	0.789241
8	0.000527	0.000711	0.733548	0.735433
9	0.000515	0.000677	0.759363	0.779460
10	0.000492	0.000674	0.761004	0.767331
11	0.000503	0.000675	0.761476	0.788914
12	0.000513	0.000696	0.745903	0.766781
13	0.000510	0.000676	0.761318	0.804456
14	0.000512	0.000689	0.753664	0.804014
15	0.000520	0.000697	0.746321	0.788073

Table 5
Missing data imputation results for 10% missing proportion under the SI regime.

Missing Data Imputation Technique	MAE	RMSE	R^2	KGE
k-NN	0.001025	0.001409	0.000000	−0.396031
SVR	0.000454	0.000649	0.782048	0.660030
MR	0.001017	0.001335	0.157921	0.187814
ANN (4–13-1)	0.000510	0.000676	0.761318	0.804456

After the best ANN model layout setting was selected, the imputed results of the other three imputation techniques under the SI procedure when the missing proportion was set as 10 % were compared as shown in Table 5. Remarkably, the SVR technique was able to provide the imputation results of the minimum MAE and RMSE, with a significantly high R^2 value of 0.782048 as well as moderate KGE value of 0.660030. On the other hand, the 4–13-1 layout of the ANN exhibited a comparable imputation results with similar error metrics and slightly lower R^2 value of 0.761318, but a remarkably high KGE value of 0.804456. With the reported zero R^2 and negative KGE values, the k-NN technique showed unsatisfactory performance. The MR technique offered better imputed results when compared to the k-NN due to the positive R^2 and KGE indicator as well as the lower error metrics. However, the relatively low KGE value obtained by the MR technique implied its imputation performance were lacking. In brief, both ANN and SVR in this case presented excellent results. The distinct difference in the KGE and R^2 indicated that ANN had an overall better performance in the missing data imputation. This could be explained by the nature of KGE indicator which considered multiple aspects of the model performance beyond considering solely on the correlation measure among the imputed data. The greatly discounted KGE from the R^2 indicator suggested that SVR had performed well in addressing the correlation but sacrificed on the bias and variability component.

Table 6
Missing data imputation results for 20% missing proportion (ANN) under the SI regime.

Number of Hidden Nodes	MAE	RMSE	R^2	KGE
1	0.000674	0.000925	0.566330	0.586693
2	0.000617	0.000864	0.622859	0.663409
3	0.000602	0.000856	0.631051	0.670494
4	0.000540	0.000795	0.000795	0.725605
5	0.000566	0.000812	0.673774	0.740200
6	0.000557	0.000810	0.669106	0.723263
7	0.000523	0.000771	0.701544	0.745460
8	0.000539	0.000787	0.686504	0.727239
9	0.000520	0.000734	0.726823	0.766121
10	0.000575	0.000819	0.660236	0.693826
11	0.000520	0.000749	0.714415	0.738909
12	0.000524	0.000759	0.707430	0.737259
13	0.000486	0.000685	0.761331	0.788649
14	0.000489	0.000716	0.739128	0.763032
15	0.000549	0.000762	0.707044	0.754183

The imputation results for 20 % missing proportion from the 15 combinations of the ANN model layout settings were summarized in Table 6. The ANN model with 13 hidden nodes was seen to outperform others with the least MAE value of 0.000486, and RMSE value of 0.000685, and also with the largest R^2 of 0.761331 as well as KGE value of 0.788649.

Table 7 shows the imputed results under the SI procedure when the missing proportion was set as 20 %. Similarly, the selected best ANN model with the network layout setting of 4–13-1 was seen to provide better imputation results with the smallest RMSE among the imputation techniques, with a substantially high R^2 value of 0.761331 and KGE of 0.788649. Moreover, the calculated error metrics from each of the imputation techniques had increased on the overall as compared to the missing proportion of 10 %. This is due to more unknown data values introduced to the sedimentation data set in the model fitting or training phase, affecting the quality of imputation results. Notably, the performance of the ANN has not declined drastically. Comparatively, the SVR had a slightly poorer performance than ANN in both cases of 10 % and 20 % missing proportion, as indicated by the lower R^2 and KGE values. The k-NN techniques yielded the largest MAE and RMSE, associated with the lowest R^2 and KGE values, indicating the poorest performance across all of the imputation techniques.

Similarly, the imputation results of the 30 % missing proportion from the 15 combinations of the ANN model layout settings were summarized in Table 8. By careful inspection, it could be discovered that the 4–13-1 and 4–14-1 settings for the ANN model achieved a similarly optimum results as compared to the other layout settings. Considering the trade-off by the lower KGE value but with similar values in the error metrics, the 4–13-1 ANN layout had marginally better imputed results compared to the 4–14-1 setting.

After the ANN model with the optimum layout setting has been chosen, the results were further compared with the other imputation techniues based on the imputed results under the SI procedure when the missing proportion was set as 30 %, provided in Table 9 below. Despite of the increased missing proportion, the ANN model with a 4–14-1 layout setting managed to obtain the R^2 value of 0.743228 and KGE of 0.779721, which signified that the ANN model have successfully explained the missing data correlation by approximately 74 % as well as the variability and bias component based on the overall imputed missing data. The best ANN model results were also supported by the lowest RMSE of 0.0.000755. The SVR was able to secure the lowest MAE but possessed a lower R^2 and a greatly discounted KGE value. Besides, the MR technique had shown poor performance whilst the k-NN technique failed to provide reasonable imputation results based on all of the performance metrics. In brief, the results implied that the ANN model outperformed other imputation techniques.

Similarly, the imputation results for the 40 % missing proportion from the 15 combinations of the ANN model layout settings were summarized in Table 10. Evidently, the 4–14-1 ANN had outperformed other layout settings. The lowest MAE and RMSE as well as highest R^2 and KGE were reported.

In spite of a suffered greater observation loss due to the higher proportion (40 %) of missing data, the 4–14-1 ANN managed to maintain a similar R^2 and KGE of approximately 0.754960 and 0.763172 respectively as shown in Table 11. The remarkable imputation results by ANN was more significant compared to the SVR technique that has an R^2 of 0.704345 and KGE of only 0.563332. Similarly, the MR technique yielded a below average performance with a positively low values in the R^2 and KGE values. The MR has similar error metrics values with the k-NN technique but it has produced an unsignificant value in the R^2 of 0.009222, and an undesirable KGE value of − 0.346660, suggesting its uncomparable imputational performance.

Last but not least, the imputation results for the 50 % missing proportion from the 15 combinations of the ANN model layout settings were summarized in Table 12. By careful inspection, it was found that the 4–14-1 and 4–15-1 layout settings for the ANN model shared disputable

Table 7
Missing data imputation results for 20% missing proportion under the SI regime.

Missing Data Imputation Technique	MAE	RMSE	R^2	KGE
k-NN	0.001036	0.001453	0.040873	−0.266534
SVR	0.000469	0.000744	0.718325	0.648367
MR	0.001027	0.001354	0.157044	0.211024
ANN (4–13-1)	0.000486	0.000685	0.761331	0.788649

Table 8
Missing data imputation results for 30% missing proportion (ANN) under the SI regime.

Number of Hidden Nodes	MAE	RMSE	R^2	KGE
1	0.000678	0.000919	0.615703	0.594434
2	0.000629	0.000864	0.661436	0.659542
3	0.000629	0.000859	0.664972	0.619862
4	0.000594	0.000835	0.684209	0.689598
5	0.000575	0.000819	0.696266	0.709078
6	0.000587	0.000817	0.697809	0.702790
7	0.000577	0.000800	0.709408	0.721948
8	0.000556	0.000788	0.718347	0.731169
9	0.000532	0.000745	0.747788	0.750139
10	0.000541	0.000789	0.717855	0.722174
11	0.000544	0.000772	0.729387	0.731315
12	0.000545	0.000772	0.729766	0.740193
13	0.000551	0.000755	0.743228	0.779721
14	0.000507	0.000716	0.766998	0.759619
15	0.000554	0.000753	0.743254	0.758095

Table 9
Missing data imputation results for 30% missing proportion under the SI regime.

Missing Data Imputation Technique	MAE	RMSE	R^2	KGE
k-NN	0.001012	0.001480	0.004297	−0.360414
SVR	0.000494	0.000770	0.733351	0.631397
MR	0.001055	0.001397	0.178645	0.206801
ANN (4–14-1)	0.000551	0.000755	0.743228	0.779721

Table 10
Missing data imputation results for 40% missing proportion (ANN) under the SI regime.

Number of Hidden Nodes	MAE	RMSE	R^2	KGE
1	0.000709	0.001038	0.593693	0.559312
2	0.000697	0.001023	0.605138	0.557732
3	0.000661	0.000972	0.644193	0.612560
4	0.000626	0.000953	0.659425	0.676419
5	0.000574	0.000904	0.692583	0.698257
6	0.000576	0.000899	0.695631	0.695575
7	0.000595	0.000889	0.701798	0.679867
8	0.000562	0.000871	0.713924	0.698408
9	0.000556	0.000835	0.736975	0.735783
10	0.000566	0.000874	0.712259	0.703029
11	0.000557	0.000848	0.728627	0.721976
12	0.000635	0.000921	0.681241	0.687482
13	0.000595	0.000882	0.707401	0.718822
14	0.000544	0.000807	0.754960	0.763172
15	0.000576	0.000876	0.711190	0.710878

Table 11
Missing data imputation results for 40% missing proportion under the SI regime.

Missing Data Imputation Technique	MAE	RMSE	R^2	KGE
k-NN	0.001058	0.001623	0.009222	−0.346660
SVR	0.000524	0.000893	0.704345	0.563332
MR	0.001112	0.001537	0.174784	0.180469
ANN (4–14-1)	0.000544	0.000807	0.754960	0.763172

Table 12
Missing data imputation results for 50% missing proportion (ANN) under the SI regime.

Number of Hidden Nodes	MAE	RMSE	R^2	KGE
1	0.000733	0.001240	0.520683	0.453899
2	0.000721	0.001230	0.528905	0.452593
3	0.000677	0.001163	0.579600	0.503165
4	0.000626	0.001136	0.598677	0.547326
5	0.000609	0.001114	0.612906	0.561521
6	0.000619	0.001129	0.603464	0.550496
7	0.000604	0.001078	0.637705	0.585148
8	0.000605	0.001091	0.628832	0.561462
9	0.000574	0.001036	0.665606	0.614285
10	0.000602	0.001065	0.646686	0.580699
11	0.000598	0.001056	0.652570	0.584719
12	0.000609	0.001053	0.654744	0.624656
13	0.000595	0.001064	0.647025	0.585714
14	0.000581	0.001066	0.648297	0.665640
15	0.000563	0.001015	0.678511	0.628115

imputation results. However, the 4-14-1 ANN attained the highest KGE value of 0.665640, with the correspondingly sifnificant R^2 value of 0.648297, slightly lower than the R^2 value when the layout setting employs 15 hidden nodes (4-15-1). Additionally, there were no significant difference between the evaluated error metrics.

The performance measures for the different missing data imputation techniques when the missing data proportion was set at 50 % were tabulated in Table 13 below. Based on the performance masures, the ANN model had exceptionally outperformed compared to other imputation techniques although a great deal of data were missing. The reported values of R^2 and KGE were approximately 65 % or more. The remarkable imputation results by ANN showed distinct significance compared to the SVR technique that has an addressed R^2 of 0.617921 and KGE of only 0.457485. Similarly, the MR technique yielded a below average performance with a positively low values in the R^2 and KGE values. Also, the MR has similar error metrics values with the k-NN technique but it has an unsignificant value in the R^2 of 0.047152, and an undesirable KGE value of − 0.271489, restating its uncomparable imputational performance.

The overall imputation results by each of the imputation techniques were illustrated in Fig. 5, Fig. 6, Fig. 7, and F8 respectively, with each of the sub-plots presenting the scatterplot of the imputed against observed missing data values. Fig. 5 depicted the ANN model with superior imputation performance, followed by the SVR technique depicted in Fig. 6, MR shown in Fig. 7, and finally k-NN in Fig. 8 that has the worst performance. Data points which coincide with the red linear line represents a perfect match between the imputed and observed missing data values.

For lower missing proportion such as 10 % and 20 %, ANN and SVR were compatible in terms of the imputation results. This was examined between Figs. 5 and 6, where the point estimates generated by both imputation techniques followed closely along the linear line which indicated a good match between the imputed and observed missing data points. However, as the imputation performance of the ANN dominated the SVR as the missing proportion increased to 30 %, 40 %, and 50 %. Specifically when the 50 % missing proportion scatterplots for the ANN and SVR were compared, majority of the data points with larger observed values were located far below from the red linear line. On the contrary, the data points were located closer and spreaded more uniformly, balancing along the red linear line. These results suggested that the ANN was not only able to capture the missing data correlation and variability, but also with a relatively low bias. The SVR imputation technique however produced results with high bias (i.e. underestimate) despite of addressing the correlation and variability.

Table 13
Missing data imputation results for 50% missing proportion under the SI regime.

Missing Data Imputation Technique	MAE	RMSE	R^2	KGE
k-NN	0.001087	0.001777	0.047152	−0.271489
SVR	0.000551	0.001118	0.617921	0.457485
MR	0.001160	0.001680	0.167987	0.139492
ANN (4–14-1)	0.000581	0.001066	0.648297	0.665640

Furthermore, the illustrated scatterplots for the MR imputation technique in Fig. 7 suggested that the imputed missing data were mostly inaccurate as majority of the data points were located far away from the red linear line. The pattern of imputation were consistent across all missing proportions. For smaller values of missing data observations, the MR technique presented a uniform imputation along the red linear line resembling some unbiasness although they were all far aways from the actual data points on the red linear line. Apparently, the major source of bias contribution existed on the larger values of the missing data observations. Specifically, all of the imputed data points were lower and greatly distanced from the red linear line, signifying low accuracy with high bias in the MR imputation technique. Thus, the MR ecxhibited a low estimation ability with a high tendency to perform underestimation based on the imputation results.

Nevertheless, the k-NN technique has not provided plausible imputation results across all missing data proportions as depicted in Fig. 8. The thick horizontal trend formed by the imputed data points showed that the imputed values were very similar, regardless of the magnitude of the observed missing data value. The k-NN imputation performance was seen to be slightly better than the SAA and SM method where an identical estimated value is assigned to all missing data points. In the case of the SAA and SM method, a linear horizontal line represents the imputed missing data points. The k-NN has some variation in the imputations but failed to capture the correlation and variability of the missing data points. Besides, the k-NN has showed biasness where most of the imputed data points lied below the red linear line. Hence, the k-NN imputation technique has underperformed and tends to underestimate based on the imputation results.

3.2. Multiple imputation (MI) results

The bootstrap resampling method followed by the MI procedure was implemented. In the bootstrap resampling phase, 100 bootstrap samples were generated and the respective imputation techniques were applied to produce the point estimates based on the bootstrap samples. To ensure the results validity, the standard error of the point estimates from the respective imputation techniques must be consistent, without exceeding a standard error of 0.5 %. Specifically, the layout setting for the ANNs in the MI regime were derived from the best selected layout setting from the ANN under the SI regime. Consequently, the same network setting was applied to the MI imputation regime corresponding to each of the stipulated missing proportion.

The overall results including both SI and MI regime across the five missing proportions for all imputation techniques were reported in Table 14. The error metrics (MAE and RMSE) yielded from the MI regime were all slightly above the error metrics from the SI regime. On the other hand, the R^2 values were mostly lower for the MI regime except for the k-NN where a higher R^2 across all missing percentage was observed in the MI imputation results. In addition, the ANN technique showed a minor increase in the R^2 value when the missing proportion was 10 %. Similarly, the KGE values were reported to be lower for majority of the cases of missing proportions as well as the types of imputation technique. On top of that, it was evident that there were substantial differences in the KGE value for the SVR and MR techniques when the missing proportion was at 50 %. Also, it was inspected that the higher the stipulated missing proportion, the poorer the missing data imputation performances. This was reflected by the increased error metrics and declined R^2 and KGE values as the missing proportion increased.

Figs. 9–12 depicted the summarized performance measures

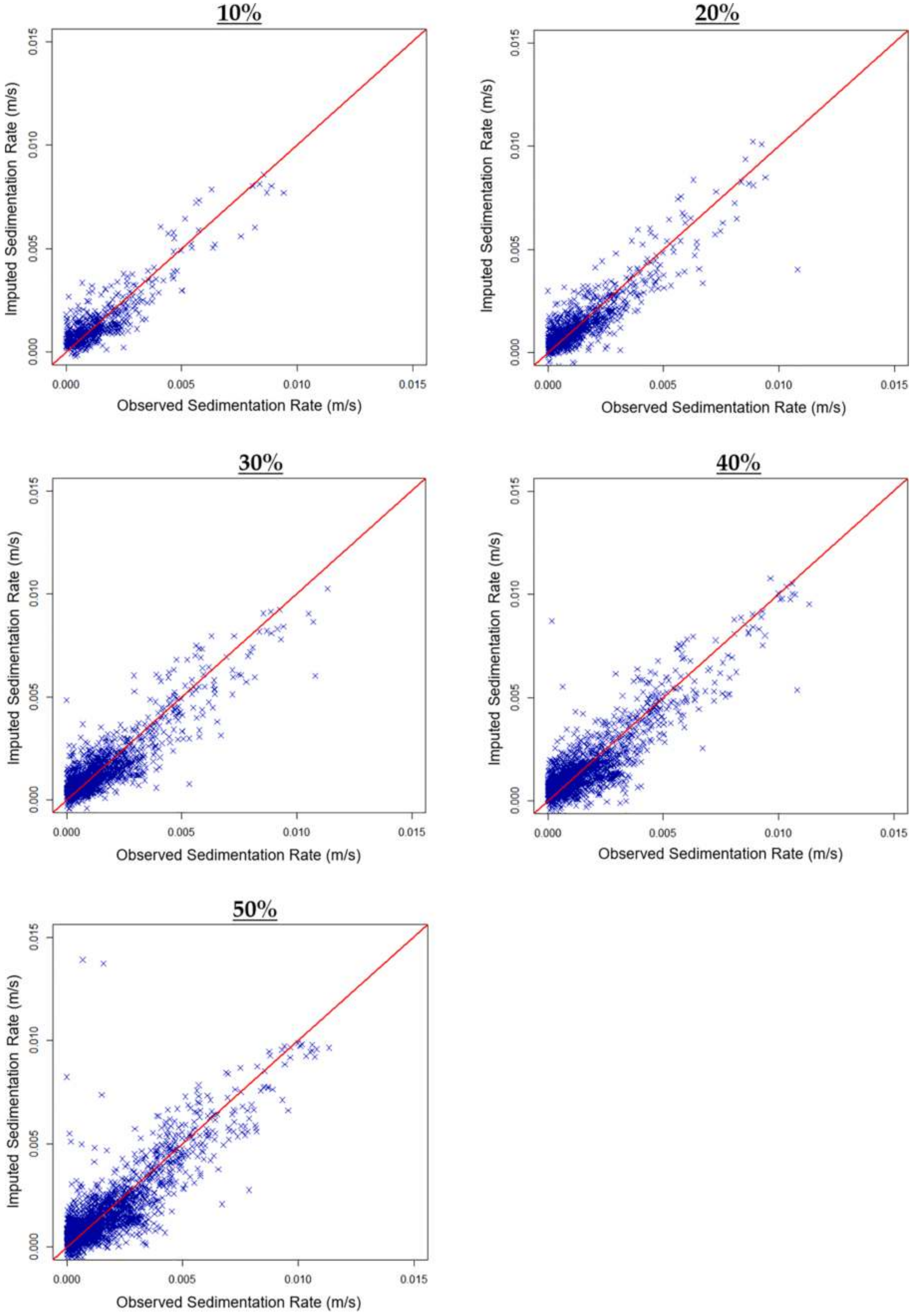


Fig. 5. Scatterplots of imputed against observed missing data values of the best ANN under different missing proportion.

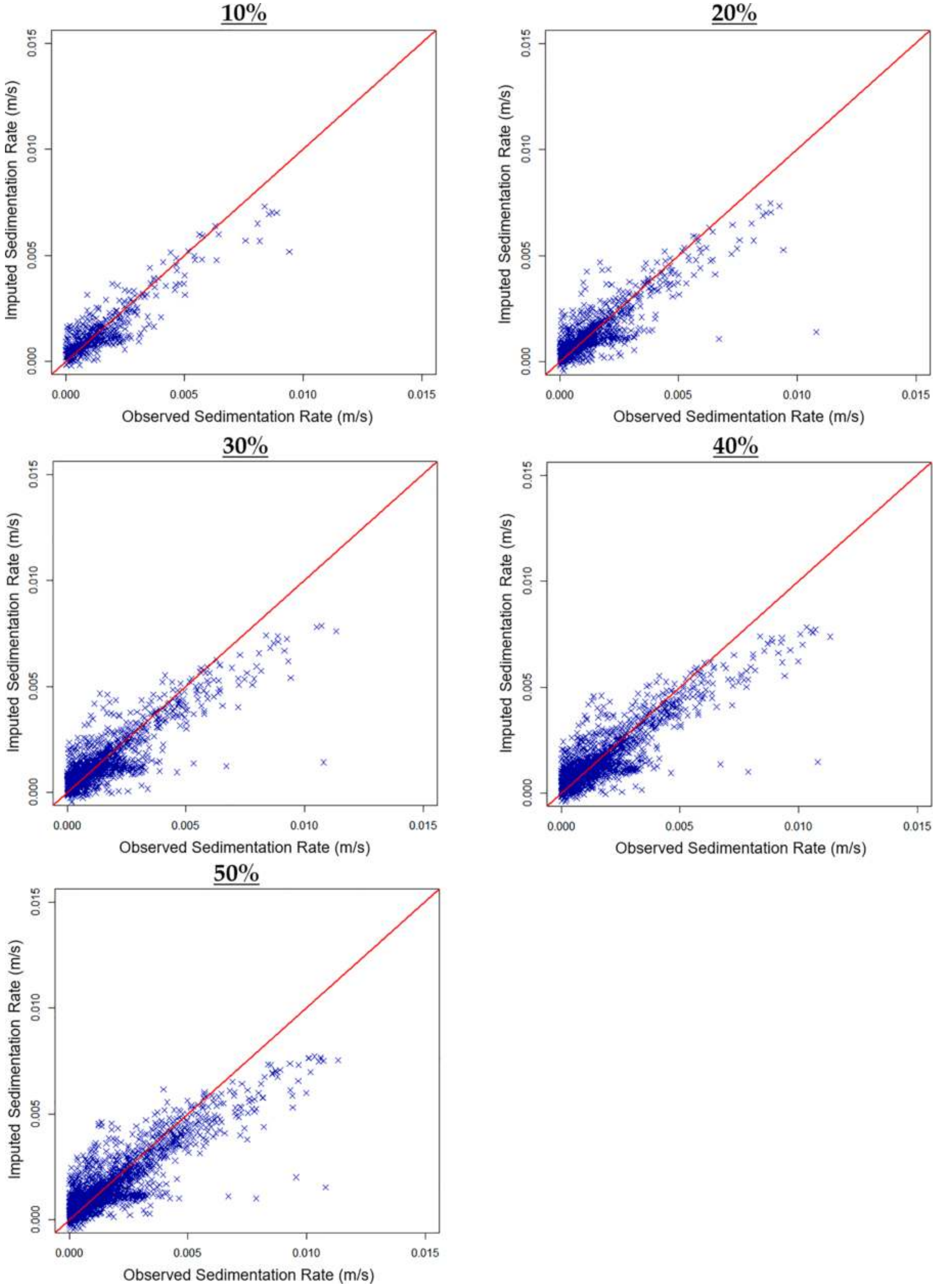
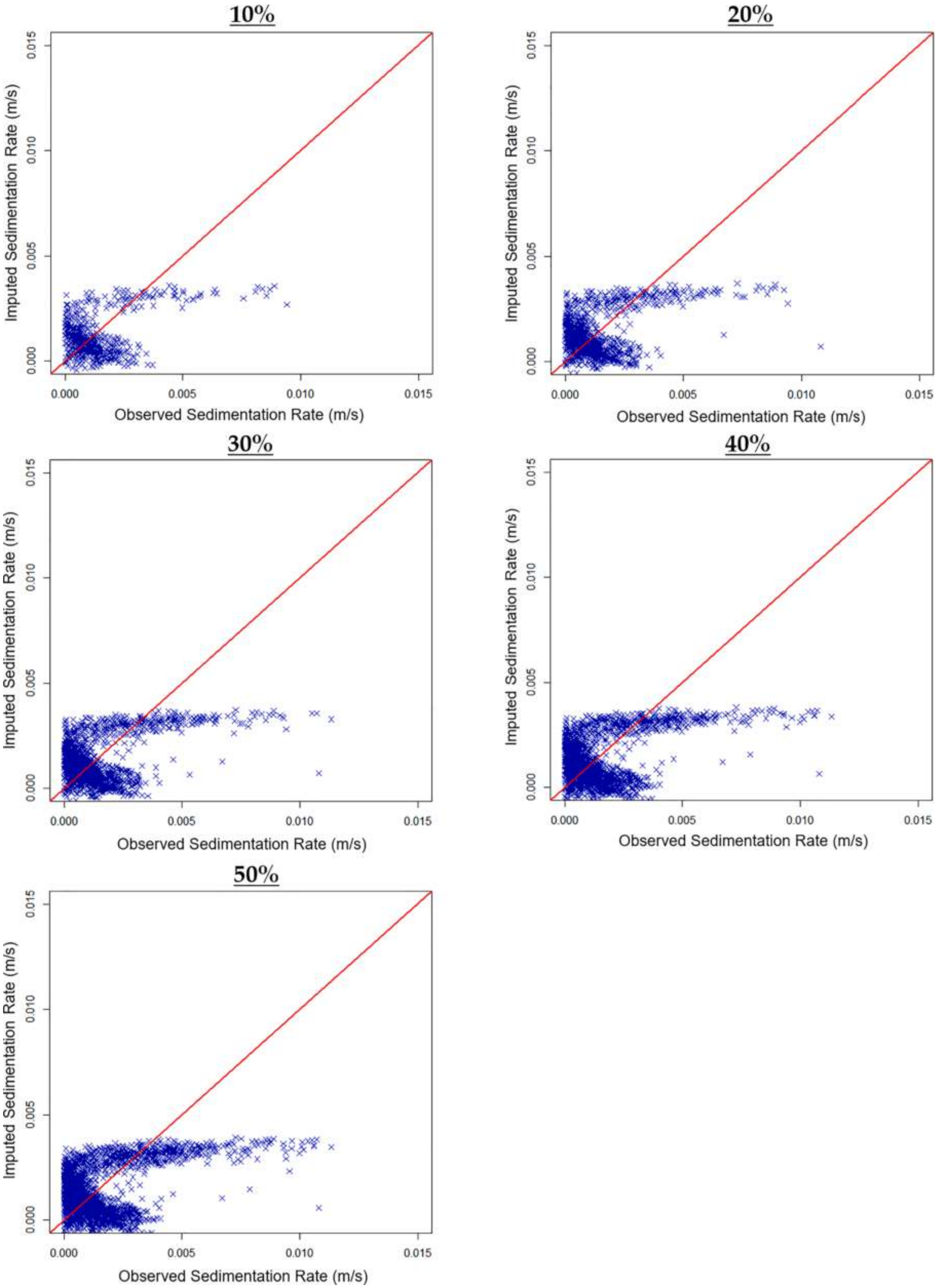


Fig. 6. Scatterplots of imputed against observed missing data values of SVR under different missing proportion.



10%

20%

30%

40%

50%

Fig. 7. Scatterplots of imputed against observed missing data values of MR under different missing proportion.

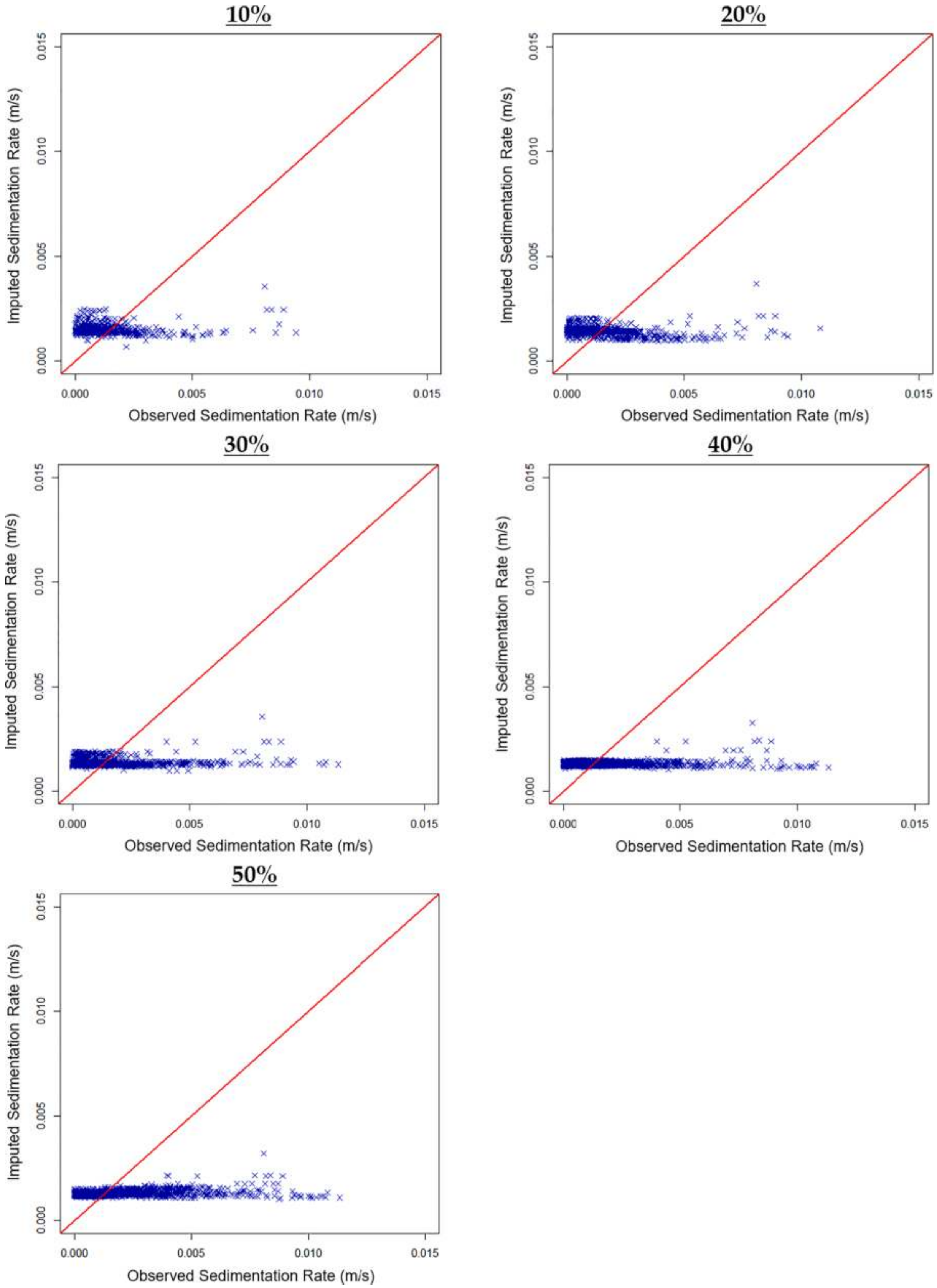


Fig. 8. Scatterplots of imputed against observed missing data values of k-NN under different missing proportion.

Table 14
Missing data imputation results for 10% missing proportion under the MI regime.

Missing Data Imputation Technique – Missing Percentage	MAE		RMSE		R^2		KGE	
	SI	MI	SI	MI	SI	MI	SI	MI
k-NN (10 %)	0.001025	0.001167	0.001409	0.001515	0.000001	0.013691	−0.396031	−0.380688
SVR (10 %)	0.000454	0.000546	0.000649	0.000661	0.782048	0.776147	0.660030	0.572782
MR (10 %)	0.001017	0.001092	0.001335	0.001413	0.157921	0.157919	0.187814	0.198678
ANN(10 %)	0.000500	0.000607	0.000682	0.000836	0.756698	0.763547	0.789241	0.728016
k-NN (20 %)	0.001036	0.001104	0.001453	0.001491	0.040873	0.044134	−0.266534	−0.295501
SVR (20 %)	0.000469	0.000516	0.000744	0.000792	0.718325	0.712380	0.648367	0.649496
MR (20 %)	0.001027	0.001075	0.001354	0.001402	0.157044	0.156415	0.211024	0.229867
ANN (20 %)	0.000523	0.000538	0.000771	0.000780	0.701544	0.707102	0.745460	0.759609
k-NN (30 %)	0.001012	0.001123	0.001480	0.001543	0.004297	0.020080	−0.360414	−0.357956
SVR (30 %)	0.000494	0.000592	0.000770	0.000880	0.733351	0.726479	0.631397	0.555964
MR (30 %)	0.001055	0.001148	0.001397	0.001493	0.178645	0.178511	0.206801	0.199345
ANN (30 %)	0.000551	0.000558	0.000755	0.000823	0.743228	0.716493	0.779721	0.747770
k-NN (40 %)	0.001058	0.001123	0.001623	0.001647	0.009222	0.022097	−0.346660	−0.343940
SVR (40 %)	0.000524	0.000588	0.000893	0.000958	0.704345	0.694568	0.563332	0.557826
MR (40 %)	0.001112	0.001182	0.001537	0.001592	0.174784	0.173684	0.180469	0.197149
ANN (40 %)	0.000595	0.000617	0.000882	0.000987	0.707401	0.679630	0.718822	0.681925
k-NN (50 %)	0.001087	0.001253	0.001777	0.001861	0.047152	0.052643	−0.271489	−0.333830
SVR (50 %)	0.000551	0.000775	0.001118	0.001349	0.617921	0.610742	0.457485	0.255491
MR (50 %)	0.001160	0.001372	0.001680	0.001893	0.167987	0.166821	0.139492	0.045922
ANN (50 %)	0.000595	0.000677	0.001064	0.001168	0.647025	0.637247	0.585714	0.577463

respectively under the results comparison between the SI and MI regimes across all missing proportions. The MR imputation technique (in yellow) has the most undesirable results with the highest value across all missing proportions by comparing and contrasting the MAE metrics between the SI and MI regimes (Fig. 9). The k-NN (in red) has slightly lower MAE on the overall when compared to the MR. On the contrary, the ANN technique (in green) showed highly satisfactory results with significantly lower MAE compared to the MR and k-NN techniques. The MAE yieled from the ANN was comparable with the SVR technique (in blue) with minor differences. In brief, the lower MAE exhibited by the SVR and ANN techniques indicated smaller absolute differences between the imputed and observed missing data values.

Similarly, the comparison of SI and MI results based on the RMSE metric in Fig. 10 suggested that the ANN and SVR were the most desirable missing data imputation techniques as both techniques yielded the lowest RMSE across all missing proportions for both SI and MI regimes. The relatively smaller RMSE indicated that the average magnitude of the differences between imputed values and observed values have smaller deviations, implying a higher accuracy and better overall imputation performance. For the ANN technique, the overall RMSE values were seen to be greater for a lower missing proportion (10 % and 20 %) but smaller for higher missing proportion (30 %, 40 %, and 50 %) when compared to the SVR. In addition, there was a significant difference between the RMSE values of ANN and SVR in the MI regime. When the missing proportion was set at 10 %, the RMSE of the SVR was comparatively lower than the ANN. On the contrary, the RMSE of the SVR was comparatively higher than the ANN when the missing proportion was set as 50 %. The phenomenon suggested that the ANN was abel to provide more accurate imputation results although the missing proportion was high whereas the higher accuracy exhibited by the SVR technique was constrained by the missing data proportion.

The compared coefficient of determination, R^2 values as illustrated in Fig. 11 has again confirmed the distinctive ability of the ANN and SVR imputation techniques for missing data across all missing proportions. There were minor discrepancies between the values for ANN and SVR except for the case where the stipulated missing proportion was 50 %.

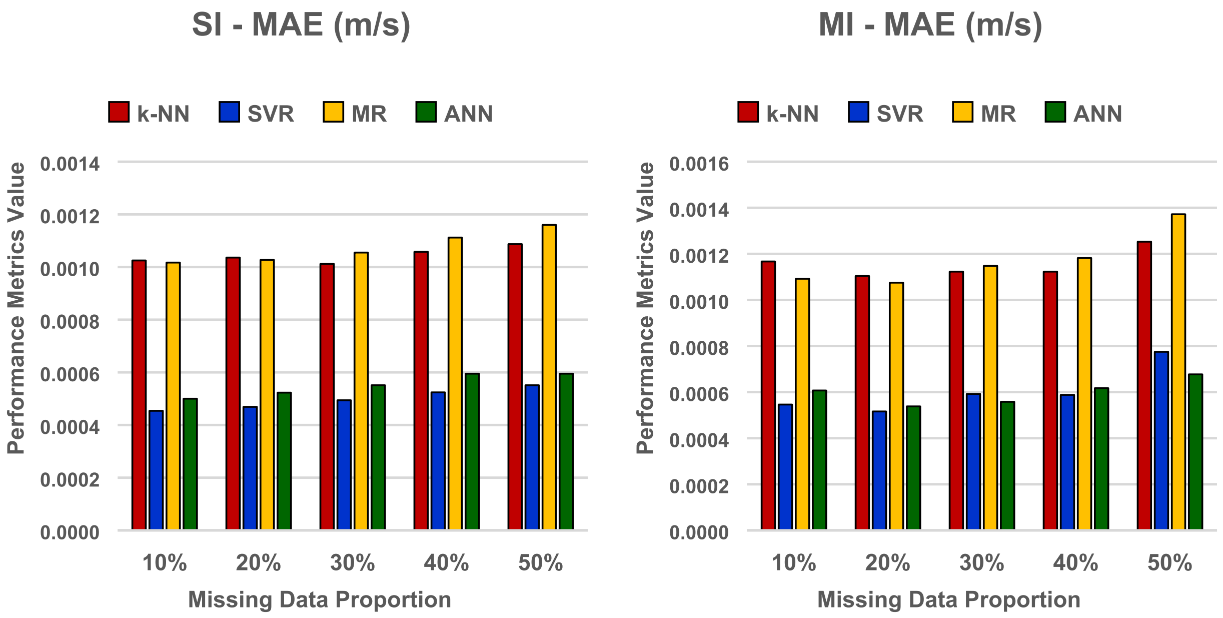


Fig. 9. Results comparison based on the MAE metric.

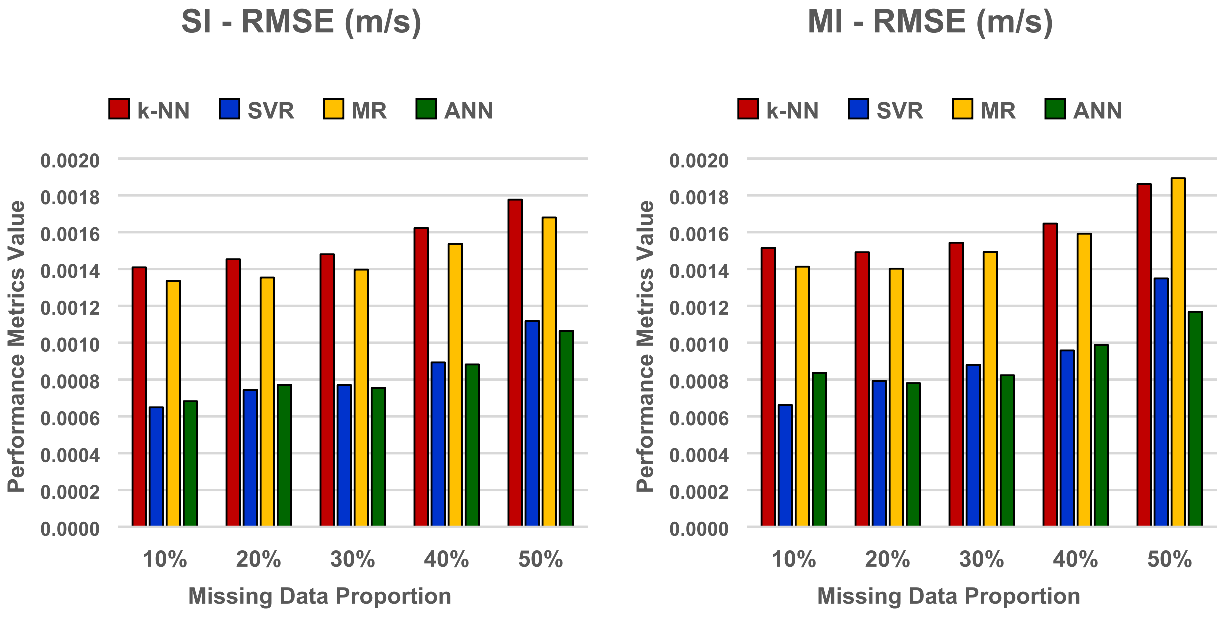


Fig. 10. Results comparison based on the RMSE metric.

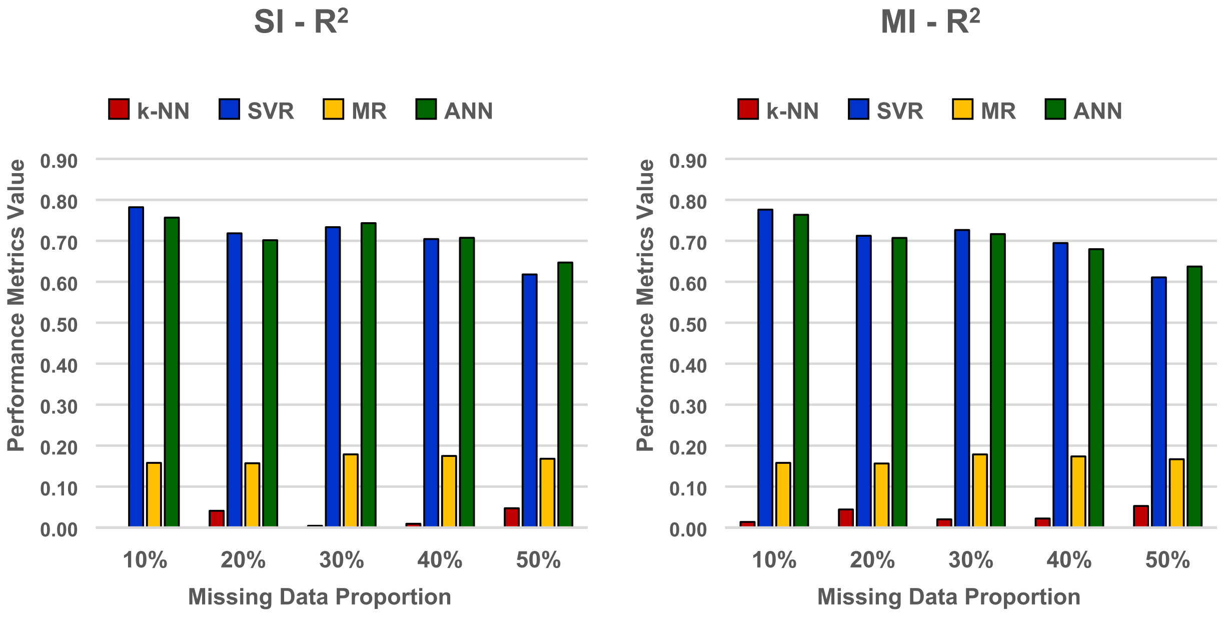


Fig. 11. Results comparison based on the R^2 metric.

The larger value of R^2 for the imputation results of ANN under both the SI and MI regimes indicated that the ANN was more capable in imputing missing data of larger missing proportions compared to the SVR technique. In other words, both the ANN and SVR had shown good performance in capturing larger proportion of the observed data variation with high accuracy. The imputation performance was relatively lower for the MR technique as it was able to attain small values of R^2. Lastly, the k-NN technique yielded the worst imputation performance with the lowest R^2 reported across all missing proportions for both the SI and MI regimes. It was expected to have a decreasing trend for the R^2 as the missing proportion increased. The random fluctuation in the R^2 for the k-NN and MR imputation techniques further suggested that their imputation performances were not consistent with the missing proportion. In short, the ANN and SVR were evidently reliable imputation techniques based on the distinguishable results of R^2.

Based on the illustration in Fig. 12, the KGE value was the highest for the ANN imputation technique, followed by the SVR, MR, and the k-NN techniques. Evidently, the KGE for the k-NN was the least across all missing proportions for both the SI and MI regimes. Such imputation results were undesirable as the yielded KGE were all negative values. The overall KGE exhibited a declining trend as the missing proportion increased except for the SVR and k-NN techniques. The fluctuation in the KGE values across the increased missing proportion suggested that the SVR and k-NN were not able to provide consistent imputation performances. With the highest KGE values and all above 0.5, the ANN exhibited remarkable imputation performance across all missing proportions as well as in both SI and MI regimes. Although the SVR managed to produce satisfactory KGE values in most cases under the SI regime, the results under the MI regime revealed that the SVR was less reliable due to the discounted KGE values and a huge declination in the value from 40 % to 50 % missing proportion. The lowest KGE value was between 0.2 and 0.3 which was lower than 0.5. The MR showed

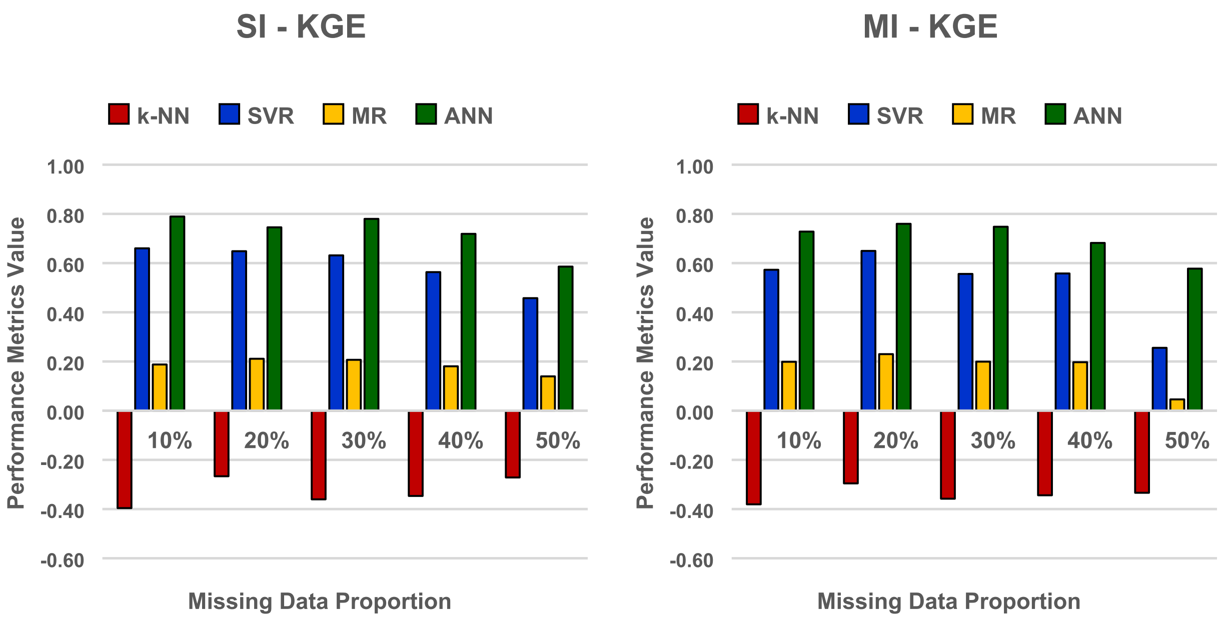


Fig. 12. Results comparison based on the KGE metric.

positively small and consistent values of approximately 0.2 except for the case of 50 % missing proportion under the MI regime. The results suggested that MR has shown some, but highly limited imputation performance, with low consistency.

In essence, the different performance measures aided in understanding how well the imputation techniques estimated the missing data values from different perspectives. Firstly, the ANN and SVR imputation techniques were able to secure low values of MAE and RMSE. The low error metrics signified that the imputed values by the ANN and SVR closely aligned with the actual observed data values, showing high accuracy (low error). However, the MR and k-NN techniques provided comparatively larger MAE and RMSE values, showing lower imputation accuracy (higher error). Notably, the selected best ANN models from the SI regime, which the identical layout setting were applied to the MI regime were capable in producing highly satisfactory imputation results with significant values in both R^2 and KGE values, all at least 0.5 in all cases of missing proportions.

Although when the missing proportion was large, the ANN imputation technique was exceptionally able to provide promising imputation results of low error metrics as well as KGE and R2 values of at least 0.5. The concurrently high values of R^2 and KGE signified the distinct ability of the ANN in capturing the variability within the missing data during imputation. In particular, the high KGE could be interpreted by the successful imputation results that leveraged the three performance aspects, which are the correlation, variability bias, and mean bias that are considered by the KGE when the imputation results were examined.

In general, the KGE provided a more comprehensive assessment towards the imputation performance as compared to the R^2 although the maximum value of both performance measures are 1 (indicating a perfect score). Hence, it is not surprising that an imputation technique would have higher R^2 but lower KGE. From this perspective, the ANN outperformed the SVR technique because the produced R^2 results were considered to be close to the value of KGE for the ANN but not for the SVR. Based on the analyzed imputation results, the SVR tended to produced higher R^2 values but traded with a greatly discounted KGE value compared to the R^2 values. This indicated that the SVR has limited ability in addressing the variability bias and mean bias when imputing the missing data values for the fine sediment data.

Collectively, most of the imputation results under the MI regime in this study yielded a slightly lower performance than the SI regime based on the imputation performance measures. This was due to the process of the MI regime which involved in generating multiple sets of imputed data. Subsequently, randomness or variability were introduced into the MI processes to account for the missing data uncertainty for better generalization of the results. The inherent variability within the imputed data resulted in the larger discrepancies between the imputed and observed data values which ultimately leaded to the slight lower imputation performances. On the contrary, imputations under the SI regime has slightly better performances as only one imputed data value was generated. Consequently, this leaded to the potentiality of the imputation process to fail in capturing the underlying missing data uncertainties. The exhibited lower apparent error was traded with a cos of underestimated variability associated with the imputed missing data values.

Nevertheless, the differences between the performance measures were relatively small for the imputed results of the ANN across the different missing data proportions. This reflects that the ANN imputation technique is reliable as it produced consistent imputation performance. Conversely, the SVR technique exhibited less reliable performance when it was compared between the SI and MI regimes. Specifically, there was a notable fall in the KGE value in the SVR imputation results when the missing proportion was specified at 50 %.

Moreover, the MR imputation technique showed insignificant imputation results. The highly evaluated error metrics but low R^2 and KGE values across all different missing proportions as well as the SI and MI regimes confirmed that the MR was not capable of capable of capturing the complex dynamics of fine sediment within the dataset, thus producing inaccurate imputation results. Likewise, the k-NN has also failed to provide plausible imputation results as it exhibited the poorest performance with the largest error metrics, very low R^2 values, and undesirably negative KGE values. Similaly, it is highly possible that the k-NN failed in producing sensible imputation results due to the sophisticated underyling dynamics within the sedimentation data. Since the k-NN imputation mechanism was based on similarity between the targeted missing data and its surrounding k most similar instances, it would be the case that there were no obvious groups of similar observations for the k-NN to successfully impute the missing data accurately.

To rank the overall performances based on the four imputation techniques implemented in this study, it was evident that the ANN had the best imputation results, followed by the SVR, the MR, and then the k-NN. Firstly, for the missing proportion of 10 %, the ANN under the SI regime with layout setting 4–13-1 showed the overall best imputation performance with the lowest error metrics of 0.000500 in the MAE and 0.000682 in the RMSE, and highest KGE of about 79 %. However, the highest R^2 was resulted from the SVR technique with a value of approximately 78 %. Secondly, for the missing proportion of 20 %, the 4–13-1 ANN has the best imputation results. The SI and MI regime both have negligible differences in their error metrics. Both had the lowest value compared to other imputation techniques with an approximate value of 0.0005 in the MAE and 0.00077 in the RMSE. The ANN under the MI regime returned the best KGE value of 76 %. However, the best R^2 value was held by the SVR technique under the SI regime with a value of approximately 72 %. Thirdly, the best imputation performance again belonged to the ANN under the SI regime and with layout setting of 4–14-1 for the 30 % missing proportion. This was associated by the lowest RMSE of 0.000755, and highest R^2 of 74 % as well as KGE of 78 %. The lowest MAE was held by the SVR results of 0.000494. For the 40 % missing proportion, the ANN under the SI regime and with layout setting of 4–14-1 exhibited the best performance. It has the lowest RMSE of 0.000882, and highest R^2 of 71 % as well as KGE of 72 %. However, the lowest MAE was attained by the SVR with a value of 0.000524. Ultimately, the best performance for the 50 % missing proportion belonged to the 4–14-1 ANN under the SI regime, with the lowest RMSE of 0.001064, and highest R^2 of 65 % as well as KGE of 59 %. In brief, the lowest error metrics in this study provided the MAE ranged between 0.0005 and 0.0006, while the lowest RMSE ranged between 0.00068 and 0.00011. The maximum R^2 ranged between 65 % and 78 % whereas the maximum KGE ranged between 59 % and 79 %. The outstanding results were comparable to the existing studies that were aforementioned [18,26,41,42,43,44].

4. Conclusions

Overall, the best imputation performances were exhibited by the ANN technique under both SI and MI regimes, for all missing proportions. The corresponding imputation results were consistent between two regimes, signifying that the ANN imputation technique was not only highly accurate, but highly reliable with good generalization from the imputed results. The main goal in this study is to perform a comparative analysis on the different imputation techniques. As a results, the ANN with such promising results are very appealing especially when the missing proportion was specified at high values such as 40 % and 50 %. Despite the loss of a huge portion of data, the ANN was able to retain a highly satisfactory imputation performance. This was confirmed by the significant performance metrics of R^2 and KGE which were at least 59 %. On top of that, the SVR imputation technique had also shown satisfactory results with some limitations especially in addressing the imputation bias. Moreover, it was evident the coefficient of determination, R^2 alone could not fully decipher the imputation performance. As such, it should be always studied along with other performances metrics such as the KGE which was utilized in this study.

Based on the best superior performance of the ANN imputation technique, there are future improvements that could be made to enhance the imputation performance. In particular, the best ANN layout setting which may not the be the global optimum network design in terms of the number of hidden layers and number of nodes within each hidden layers. Although the trial and error approach was used to search for the best ANN layout setting, the number of hidden layers was restricted to a single layer, where the maximum exploration of the optimum number of hidden nodes were capped at a maximum number of 15 number. However, an increased node number increases the overall network complexity which then increases the computational complexity and the corresponding computational time. Therefore, it would be recommended that further enhancement on the imputation technique could be implemented by extending the current work of methodology. For example, the incorporation of hybrid models, ensembled learning algorithms, as well as the integration of metaheuristic optimization algorithms such as the particle swarm optimization (PSO) to boost the performance of the existing ANN model.

Nevertheless, the MI regime was preferred although it may provide an increased error metrics. The MI acknowledges and quantifies the uncertainty associated with the missing data values. It generates multiple sets of imputed values, incorporating variability to address different possible values. The MI regime was able to impose the genuine uncertainty in a comprehensive approach, which cannot be achieved by the SI regime. Noneless, this study have yielded insights of the superiror imputation performances of the ANN, which is one of the machine learning models commonly applied in missing data imputation. The study suggested that the ANN could provide highly reliable estimation results but researchers were highly recommended to implement hybrid techniques which potentially works better than single developed models.

CRediT authorship contribution statement

Wing Son Loh: Methodology, Formal analysis, Writing – original draft. **Lloyd Ling:** Conceptualization, Writing – original draft, Writing – review & editing, Supervision, Funding acquisition. **Ren Jie Chin:** Conceptualization, Supervision, Funding acquisition. **Sai Hin Lai:** Conceptualization, Methodology. **Kar Kuan Loo:** Formal analysis. **Choon Sen Seah:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research was supported by the Universiti Tunku Abdul Rahman Research Fund (IPSR/RMC/UTARRF/2023-C1/*L*01) and was partly supported by Malaysia Toray Science Foundation (4417/0005). The authors would like to thank the Lee Kong Chian Faculty of Engineering and Science, Universiti Tunku Abdul Rahman, and the Faculty of Engineering, University Malaya for all the technical support provided for the study.

References

[1] Academy of Sciences Malaysia. Erosion and Sedimentation. *ASM Position Paper* **2017**.

[2] Gupta LK, Pandey M, Raj PA, Shukla AK. Fine sediment intrusion and its consequences for river ecosystems: a review. J Hazard Toxic Radioact Waste 2023; 27:1. https://doi.org/10.1061/(ASCE)HZ.2153-5515.0000729.

[3] Loh WS, Chin RJ, Ling L, Lai SH, Soo EZX. Application of machine learning model for the prediction of settling velocity of fine sediments. Mathematics 2021;9:3141. https://doi.org/10.3390/math9233141.

[4] Ouyang Y. A gap-filling tool: predicting daily sediment loads based on sparse measurements. Hydrology 2022;9:181. https://doi.org/10.3390/hydrology9100181.

[5] Mitchell PJ, Spence MA, Aldrige J, Kotilainen AT, Diesing MA. Sedimentation rates in the baltic sea: a machine learning approach. Cont Shelf Res 2021;214:104325. https://doi.org/10.1016/j.csr.2020.104325.

[6] Michel TB, Wanderson PP, Isamara MS, Antonio SFM, José ATR. Methodological approaches for imputing missing data into monthly flows series. Rev Ambien Água 2022;17:2. https://doi.org/10.4136/ambi-agua.2795.

[7] Jakobsen JC, Gluud C, Winkel P, Lange T, Wetterslev J. The thresholds for statistical and clinical significance - a five-step procedure for evaluation of intervention effects in randomised clinical trials. BMC Med Res Methodol 2014;14:34.

[8] Kermorvant C, Liquet B, Litt G, Jones JB, Mengersen K, Peterson EE, et al. Reconstructing missing and anomalous data collected from high-frequency in-situ sensors in fresh waters. Int J Environ Res Public Health 2021;18(23):12803. https://doi.org/10.3390/ijerph182312803.

[9] Helsel D.R.; Hirsch M.R.; Ryberg K.R.; Archfield S.A.; Gilroy E.J. Statistical Methods in Water Resources Techniques and Methods 4-A3. https://doi.org/10.3133/tm4A3.

[10] Sattari, M.T.; Joudi, A.R.; Kusiak, A. 2016. Assessment of Different Methods for Estimation of Missing Data in Precipitation Studies. *Hydrology Res*. **2017**, 48(4), 1032–1044. https://doi.org/10.2166/nh.2016.364.

[11] Chiu, P.C.; Selamat; A., Krejcar, O. Infilling Missing Rainfall and Runoff Data for Sarawak, Malaysia Using Gaussian Mixture Model Based K-Nearest Neighbor Imputation. *IEA/AIE, Lecture Notes in Computer Science* **2019**, 11606, 27-38. https://doi.org/10.1007/978-3-030-22999-3_3.

[12] Rodríguez R, Pastorini M, Etcheverry L, Chreties C, Fossati M, Castro A, et al. Water-quality data imputation with a high percentage of missing values: a machine learning aproach. Sustainability 2021;13(11):6318. https://doi.org/10.3390/su13116318.

[13] Ben Aissia MA, Chebana F, Ouarda TBMJ. Multivariate missing data in hydrology – review and applications. Adv Water Resour 2017;110:299–309.

[14] Chivers BD, Wallbank J, Cole SJ, Sebek O, Stanley S, Fry M, et al. Imputation of missing sub-hourly precipitation data in a large sensor network: a machine learning approach. J Hydrology 2020;588:12156. https://doi.org/10.1016/j.jhydrol.2020.125126.

[15] Gao L, Zheng Y, Wang Y, Xia J, Chen X, Li B, et al. Reconstruction of missing data in weather radar image sequences using deep neuron networks. Appl Sci 2021;11(4):1491.

[16] Kashani MH, Dinpashoh Y. Evaluation of efficiency of different estimation methods for missing climatological data. Stoch Env Res Risk A 2012;26(1):59–71.

[17] Bartlett JW, Hughes RA. Bootstrap inference for multiple imputation under uncongeniality and misspecification. Stat Methods Med Res 2020;29(12):3533–46.

[18] Norzanah MS, Zalhan MZ, Ismail MN, Termizi AB. Comparative analysis of missing data imputation methods for continuous variables in water consumption data. Int J Adv Trends in Comp Sci & Eng 2019;8(1.6):471–8. https://doi.org/10.30534/ijatcse/2019/6981.62019.

[19] Little RJA. A Test of missing completely at random for multivariate data with missing values. J Am Stat Assoc 1988;83(404):1198–202.

[20] Muhammad AH, Nur DKA, Nooritawati MT, Zatul IAL, Mohamad HJ, Yoshikawa A. Missing data imputation of MAGDAS-9's ground electromagnetism with supervised machine learning and conventional statistical analysis models. Alex Eng J 2022;61(404):937–47. https://doi.org/10.1016/j.aej.2021.04.096.

[21] Garciarena U, Santana R. An extensive analysis of the interaction between missing data types, imputation methods, and supervised classifiers. Expert Syst Appl 2017;89(15):52–65. https://doi.org/10.1016/j.eswa.2017.07.026.

[22] Alsaber AR, Pan J, Al-Hurban A. Handling complex missing data using random forest approach for an air quality monitoring dataset: a case study of kuwait environmental data (2012 to 2018). Int J Environ Res Public Health 2021;18:1333. https://doi.org/10.3390/ijerph18031333.

[23] Svetlana, B.; Sven, L.; Martin L.; Markus, P. Missing Financial Data (May 11, 2022). Available at SSRN: https://ssrn.com/abstract=4106794.

[24] Fadilah, B.; Zuraini, A.S.; Saedudin, R.R.D.; Shahree, K.; Seah, C.S. Research On Missing Data Imputation Methods On Gene Expression. *Academia of Information Computing Research, Excelligent Academia*. **2020**, 1(1), 37-45.

[25] Elasra A. Multiple imputation of missing data in educational production functions. Computation 2022;10:49. https://doi.org/10.3390/computation10040049.

[26] Agwu OE, Akpabio JU, Dosunmu A. Artificial neural network model for predicting drill cuttings settling velocity. Petroleum 2020;6(4):340–52. https://doi.org/10.1016/j.petlm.2019.12.003.

[27] Xia J, Chen YD. Water problems and opportunities in hydrological Sciences in China. Hydrol Sci J 2001;46:907–21.

[28] Jared SM. Multiple imputation: a review of practical and theoretical findings. Stat Sci 2018;33(2):142–59.

[29] Yang, R. Analyses of Approaches to Deal with Missing Data in Water Quality Data Set. *Advances in Economics, Business and Management Research*. **2022**, Proceedings of the 2022 7th International Conference on Social Sciences and Economic Development.

[30] Hunt LA. Missing Data Imputation and Its Effect on the Accuracy of Classification. in: Data Sci., Springer; 2017. p. 3–14.

[31] Qi X, Guo H, Wang W. A reliable KNN filling approach for incomplete interval-valued data. Eng Appl Artif Intel 2021;100:104175.

[32] Afrifa-Yamoah E, Mueller UA, Taylor SM, Fisher AJ. Missing data imputation of high-resolution temporal climate time series data. Meteorol Appl 2020;27(1):1–18.

[33] Borges PdA, Franke J, da Anunciação YMT, Weiss H, Bernhofer C. Comparison of spatial interpolation methods for the estimation of precipitation distribution in distrito federal, Brazil. Theor Appl Climatol 2016;123(1-2):335–48.

[34] Sarker IH. AI-based modeling: techniques, applications and research issues towards automation, intelligent and smart systems. SN Comput Sci 2022;3(2). https://doi.org/10.1007/s42979-022-01043-x.

[35] Tersita CN, Yesid C, Wilfredo A, Wimlar MC, Eduardo C. Estimation of missing data of monthly rainfall in southwestern columbia using artificial neural networks. Data in Brief 2019;26:104517. https://doi.org/10.1016/j.dib.2019.104517.

[36] Emanuel RHK, Docherty PD, Lunt H, Möller K. The effect of activation functions on accuracy, convergence speed, and misclassification confidence in CNN text classification: a comprehensive exploration. J Supercomput 2024;80(1):292–312.

[37] Chin RJ, Lai SH, Loh WS, Ling L, Soo EZX. Assessment of inverse distance weighting and local polynomial interpolation for annual rainfall: a case study in peninsular malaysia. Eng Proc 2023;38:61. https://doi.org/10.3390/engproc2023038061.

[38] Badari F, Shah ZA, Saedudin RDR, Kasim S, Seah CS. Research on missing data imputation methods on gene expression. Acad Inform Comput Res 2020.

[39] Peugh JL, Enders CK. Missing data in educational research: a review of reporting practices and suggestions for improvement. Rev Educ Res 2004;74(4):525–56. https://doi.org/10.3102/00346543074004525.

[40] Little RJ, Rubin DB. Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. Annu Rev Public Health 2000;21:121–45. https://doi.org/10.1146/annurev.publhealth.21.1.121. PMID: 10884949.

[41] Balahaha HZS, Wong JK, Balahaha FZS, Chow MF, Yusuf E, Ali NA, et al. Investigating the reliability of machine learning algorithms as a sustainable tool for total suspended solid prediction. Ain Shams Eng J 2021;12(2):1607–22. https://doi.org/10.1016/j.asej.2021.01.007.

[42] Yasser AMM. Modeling of local scour depth downstream hydraulic structures in trapezoidal channel using GEP and ANNs. Ain Shams Eng J 2013;4(4):717–22. https://doi.org/10.1016/j.asej.2013.04.005.

[43] Van CP, Le H, Chin LV. Estimation of the Daily Flow in River Basins using the Data-driven Model and Traditional Approaches: An Application in the Hieu River Basin. Practice and Technology: Vietnam; 2022.

[44] Ulke A, Tayfur G, Ozkul S. Predicting suspended sediment loads and missing data for gediz River, Turkey. J Hydrol Eng 2009;14(9):954–65.

[45] Hamzah FB, Mohd Hamzah F, Mohd Razali SF, Samad H. A Comparison of multiple imputation methods for recovering missing data in hydrological studies. Civ Eng J 2021;7(9):1608–19.

[46] Kashani, M.M.; Lai, S.H.; Ibrahim, S.; Meriam, N.; Sulaiman, N. A Study on Hydrodynamic Behavior of Fine Sediment in Retention Structure Using Particle Image Velocimetry. *Water Environ. Res.* **2016**, 88.

[47] Czernek K, Ochowiak M, Janecki D, Zawilski T, Dudek L, Witczak S, et al. Sedimentation tanks for treating rainwater: CFD Simulations and PIV experiments. Energies 2021;14:7852.

[48] Wouter JMK, Jim EF, Ross AW. Technical Note: Inherent Benchmark or Not? Comparing Nash Sutcliffe and Kling-Gupta Efficiency Scores. Hydrol Earth Syst Sci 2019;327. https://doi.org/10.5194/hess-2019-327.

[49] Aksu G, Guzeller CO, Eser T. The effect of normalization method used in different sample sizes on the success of artificial neural network model. Int J of Assess Tools in Edu 2019;6:170–92.

[50] Rushd S, Hafsa N, Al-Faiad M, Arifuzzaman M. Modelling the settling velocity of a sphere in newtonian and non-newtonian fluids with machine-learning algorithms. Symmetry 2021;13:71.

[51] Shao J, Sitter RR. Bootstrap for imputed survey data. J Am Stat Assoc 1996;91(435):1278–88.

[52] Rogelis MC, Werner M, Obregón N, Wright N. Hydrological model assessment for flood early warning in a tropical high mountain basin. Hydrol Earth Syst Sci Discuss 2016;1–36.

[53] Piazza AD, Conti FL, Noto LV, Viola F, Loggia GL. Comparative analysis of different techniques for spatial interpolation of rainfall data to create a serially complete monthly time series of precipitation for Sicily, Italy. Int J App Earth Observ Geoinform Discuss 2011;12:396–408.

Wing Son Loh is currently pursuing the part-time Doctor of Philosophy (Science) programme at the Lee Kong Chian Faculty of Engineering and Science (LKC FES) at Universiti Tunku Abdul Rahman (UTAR). He holds a Master's degree in Mathematics and an Honours degree in Actuarial Science. Additionally, he is also serving as a lecturer at the Department of Mathematical and Actuarial Sciences (DMAS) for the undergraduate programmes under the LKC FES. Prior to joining UTAR as a lecturer, he worked as a full-time research assistant. He has successfully secured a project funding from the UTAR Research Fund (UTARRF) 2023 Cycle 1 and have published a journal paper as well as several conference proceedings.

Lloyd Ling is an Associate Professor and the Deputy Dean of Research and Development and Postgraduate Programmes at the Lee Kong Chian Faculty of Engineering and Science (LKC FES) at Universiti Tunku Abdul Rahman (UTAR). He earned a Master's degree in Engineering and two MBA degrees in Finance and International Business. He also prolonged his Post-MBA studies with Stanford University in Advanced Project Management and Negotiation. Before returning to Malaysia, he held several managerial positions in S&P 100 corporate and served on the corporate contingency planning committee, He also started up a catering services company in the United States. Ir. Dr. Ling Lloyd is also an Associate Fellow of the ASEAN Academy of Engineering and Technology (AAET) and serves on the panels of the Engineering Technology Accreditation Council (ETAC) and the Engineering Accreditation Council (EAC) in Malaysia. Additionally, he is a member of the Earthquake Technical Committee of the Department of Standards Malaysia under the Ministry of International Trade and Industry (MITI). He is a registered professional engineer in Malaysia, a member of the Institution of Engineers, Malaysia (IEM), serving the Urban Engineering Development Special Interest Group, and a certified trainer with a Train-The-Trainer (TTT) certification. Ir. Dr. Ling Lloyd has been an active principal researcher and research team member on national and industrial research grants in Malaysia. He was the speech champion and commencement speaker at California State University, Northridge in the USA for two consecutive years (1997 and 1998). Five of his students have won four final year project poster competition championships, one bronze medal at UTAR, and two IEEE final year project (FYP) national championships in Malaysia between 2016 and 2022.

Ren Jie Chin received his BEng. and Ph.D. degrees in Environmental Engineering from University of Malaya, Malaysia, in 2015 and 2019, respectively. He is currently an assistant professor in the Department of Civil Engineering, Lee Kong Chian Faculty of Engineering and Science, Universiti Tunku Abdul Rahman, Malaysia. He is a registered Professional Technologist (Malaysia Board of Technologists). He has been the leader or been part of the team for about several national and international research projects. His research direction focused on flood, drought, and water resources management in the context of climate change, which involved computational simulation, development of decision support system, artificial intelligent, and optimization models. He has published a total number of 24 research papers in reputed ISI journals and conference proceedings.

Sai Hin Lai is a Professor at the Department of Civil Engineering, Faculty of Engineering, Universiti Malaysis Sarawak. He is a registered Professional Engineer (PEng, Malaysia), and Chartered Engineer (CEng, UK). He serves as a Fellow of Asean Academy of Engineering; Technology (FAAET) and Institution of Engineering and Technology(FIET). He has been the leader or been part of the team for about 40 national and international research projects. His research is focused on flood, drought, and water resources management in the context of climate change, which involved computational simulation, development of decision support system, artificial intelligent, and optimization models. He has published more than 80 research papers in reputed SCI journals.

Kar Kuan Loo is a student at University Tunku Abdul Rahman, pursuing a Bachelor of Science (Honours) in Applied Mathematics with Computing. Her research interests lie at the intersection of applied mathematics and machine learning, where she explores innovative solutions to real-world challenges. Loo is committed to academic excellence and aspires to contribute significantly to these fields.

Choon Sen Seah obtained his Doctorate in Information Technology from Universiti Tun Hussein Onn Malaysia (UTHM) and has a keen interest in technology. Prior to joining Universiti Tunku Abdul Rahman (UTAR), he ran his own tech company. Currently, he serves as an Assistant Professor at UTAR. His research/technical interest and experience encompass Data Science, Digital Entrepreneurship, Financial Technology, Precision Farming & Information System. In terms of research output, he has secured around RM550 thousand worth of research grants & consultation projects as principal investigator. Seah Choon Sen has achieved significant accomplishments in his field, including publishing over 20 indexed articles and books, receiving multiple awards, and supervised more than 10 teams in winning awards in international innovation competitions. He is an accredited trainer with HRD Corp, an ecosystem builder with MaGIC, a Meta Certified Community Manager, an alumni of Microsoft Learn Student Ambassador, and the Vice President of Huawei Malaysia Seeds for the Future Alumni. In his spare time, he acted as a mentor for the startup community both on and off campus.