*J.Natn.Sci.Foundation Sri Lanka 2026 54 (2): 259 - 271*
DOI: http://dx.doi.org/10.4038/jnsfsr.v54i2.12534

**RESEARCH ARTICLE**

## Statistics

# Spatiotemporal techniques for imputing sequential missing values in rainfall data collected from Ratnapura area: A comparative study

**HRNC Heendeniya and CD Tilakaratne***
*Department of Statistics, Faculty of Science, University of Colombo, Colombo.*

Submitted: 06 March 2025; Revised: 04 September 2025; Accepted: 22 February 2026

**Abstract:** Modeling and forecasting rainfall is vital for issuing early flood warnings, but missing data, particularly sequential gaps, make this task challenging. Even though there are several techniques proposed in the literature aiming at the estimation of sequential missing values, selecting the most suitable technique is challenging. It is worth considering both temporal patterns and spatial patterns in such imputations. This study evaluates four widely used spatiotemporal techniques, K-nearest neighbors (KNN), MissForest, denoising autoencoder, and spatiotemporal kriging, to identify the best approach for imputing sequential missing values of rainfall at Ratnapura. Analysis was carried out using daily rainfall data of Rathnapura and six neighboring gauging stations, collected from 2015 to 2019. The accuracy of each estimated series was evaluated against the actual data using root mean square error (RMSE) and mean absolute error (MAE). The results show that the KNN is the most suitable technique under both wet and dry seasons, showing lower RMSE and MAE values. MissForest also performed well, surpassing denoising autoencoder and spatiotemporal kriging, though it generally fell slightly short of KNN. Among the methods compared, spatiotemporal kriging demonstrated the lowest accuracy in both wet and dry seasons due to a lack of reference stations and other meteorological variables. However, none of the techniques were successful in accurately imputing sudden extreme rainfall events, highlighting an important limitation. Forecasted rainfall at the Ratnapura station using both the imputed data and the actual data revealed that the forecasts obtained using the imputed data were not significantly different from those obtained using the actual data.

**Keywords:** Spatiotemporal data, sequential missing value imputation, rainfall data, machine learning, deep learning

## INTRODUCTION

Modeling and forecasting rainfall in a particular area is essential for issuing early warnings of floods in that area. However, in a developing country like Sri Lanka, modeling rainfall is very complicated as these series consist of missing values, especially when data are missing continuously in time for considerably long periods as a result of natural events or technical failures. If data of a series are missing continuously in time for a period, such missing values can be identified as sequential missing values.

Even though a few techniques proposed in literature aiming the estimation of sequential missing values (Feng et al., 2014; Mital et al., 2020; Saubhagya et al., 2024), selecting the most suitable technique is challenging, as the performance of the techniques will depend on the nature of the missing data series as well as the characteristics of the aiding data, which are used to estimate the missing values. In estimating sequential missing values of rainfall data, it is worth considering not only their temporal patterns but also their spatial patterns as well (Li et al., 2020; Mital et al., 2020; Saubhagya

* Corresponding author (cdt@stat.cmb.ac.lk; https://orcid.org/0000-0003-0330-845X)

et al., 2024). However, studies which evaluate advanced spatiotemporal techniques for imputing missing values, especially a sequence of missing values in rainfall data, are extremely rare.

In the Sri Lankan context, one study (Silva et al., 2007) focuses on three conventional methods, namely the arithmetic mean method, the normal ratio method, and the inverse distance method, and they introduced a new method called the aerial precipitation ratio method. however, according to the RMSE and mean absolute percentage error (MAPE) accuracy of these methods was very low. In another study, Saubhagya et al. (2024) developed a new hybrid method which combines spatial kriging and artificial neural networks. However, it does not examine how the seasonal variation of data affects missing value imputation and also does not compare the novel technique developed with widely used spatiotemporal missing value imputation techniques. In the global context, (Mital et al., 2020) developed a novel sequential imputation algorithm with the aim of imputing missing values in spatiotemporal daily rainfall measures. They used a sequential imputation algorithm, which was incorporated within a spatial interpolation method based on random forests. However, they did not compare this algorithm with other available techniques.

The main objective of this study is to compare widely used spatiotemporal missing value imputation techniques with the intention of identifying the most suitable missing value imputation technique to estimate the sequential missing values of the rainfall at Rathnapura gauging station. Rathnapura was selected as it is the most vulnerable area for floods in Sri Lanka (Thevakaran et al., 2019; Weerakoon & Dharmapriya, 2021). Moreover, it is important to mention that there is a research gap in identifying the most suitable spatiotemporal missing value technique to estimate sequential missing values of rainfall data.

Apart from finding a more accurate spatiotemporal missing data imputation technique, this study examines how different imputation techniques perform during different seasons and investigates the sensitivity of the selected spatiotemporal imputation techniques to different levels of missingness. Furthermore, this study will recommend the imputation technique that better estimates sequential missing data in rainfall records in the Ratnapura area. Finally, it investigates the pertinence of the imputed missing values in developing the forecasting models.

The rest of the paper is organized as follows: the next section describes the materials and methods used in achieving the objectives of this study. The theories behind the selected widely used spatiotemporal techniques for imputing sequential missing values are also briefly presented in this section. The third section presents and interprets the results obtained. The final section presents the conclusion derived and proposes directions for future research.

## MATERIALS AND METHODS

Four widely used sequential missing value imputation techniques, namely, K-nearest neighbors (KNN) (Zhu and Cheng, 2016), Missforest (Zhang et al., 2021), denoising autoencoder (Wardana, Gardner and Fahmy, 2022) and spatiotemporal kriging (Bae et al., 2018; Yang et al., 2018) were identified from a comprehensive literature review. Sequential missing value imputation for Rathnapura gauging station was carried out using daily rainfall data of Rathnapura and six neighbouring gauging stations (Moralioya, Detanagala, Keragala, Elston, Balangoda, and Ilubbuluwa) from 2015 to 2019. This period was selected as complete rainfall data records (i.e. without missing values) were available for all selected stations. Initially, a preliminary analysis was carried out using the selected dataset as spatial analysis and temporal analysis. As the Spearman correlation coefficient has the ability to measure linear as well as non-linear correlation between variables, it was applied to identify relationships between rainfall records in Rathnapura and other neighbouring stations (i.e. spatial dependencies).

Rainfall records are deleted during selected periods of different lengths to artificially create sequential missing values. Then, sequential missing values were estimated using the four selected imputation techniques. The accuracy of each estimated series of missing values was evaluated against the actual data using RMSE and MAE. Sequential missing values were imputed with a focus on both wet and dry seasons. For each season, missing periods of different durations (one week, two weeks, and one month) were examined. The selection of missing periods for wet and dry seasons was carried out according to Priyakumara and Ranasinghe (2023), considering the rainfall values (whether it is high (wet season) or low (dry season)) in the dataset. They mentioned that the primary source of rainfall for the Ratnapura area is the southwest monsoon, typically occurring from May to September. According to this statement, September was selected to represent the wet season, and January was selected to represent the dry season. Therefore, missing values were generated for both September (representing the wet season) and January (representing the dry season) in the years 2015, 2017, and 2019. This was done for varying durations, including one week, two weeks, and one month, to evaluate how the imputation techniques performed under different lengths of missing data. To implement KNN, Missforest, and spatiotemporal kriging, daily data from 2015 to 2019 were used without considering the train-test split since these techniques only need datasets with missing values. However, the denoising autoencoder was trained using data from 1994 to December 2014 since it needs a large number of data points to avoid overfitting, while the test sample included data from January 2015 to December 2019. The model accepts input as a collection of 7×7 matrices, where each matrix consists of 7 columns representing rainfall data from different stations, including Ratnapura. Each matrix also contains 7 rows, where the first 6 rows correspond to the previous 6 days of observed data, and the final row represents the current day. This setup allows the autoencoder to impute sequential missing values during both the wet and dry seasons in the test set.

The following sub-sections briefly explain the theory behind four selected techniques relevant to estimating sequential missing value imputation of rainfall at Rathnapura gauging station.

**K nearest neighbors**

The KNN algorithm used in this study applies a weighted mean from a sample of data at the target location. This requires the presence of data at the multi-site station network considered. Let us define a dataset where the rows represent days ($\boldsymbol{r_i}$; $i = 1, \ldots, i_0, \ldots, I$) and the columns represent stations ($\boldsymbol{c_s}$; $s = 1, \ldots, s_0, \ldots, S$). Each entry in the dataset corresponds to the value of a variable $X_{i,s}$, observed for a specific station ($s$) on a particular day ($i$). Consider a row vector ($\boldsymbol{r_{i0}}$) containing a missing value in the target station ($s_0$). This row vector is compared with other row vectors ($\boldsymbol{r_i}$) by computing a distance measure ($d_i|\boldsymbol{r_{i0}}$) which measures dissimilarity between the day with a missing value ($\boldsymbol{r_{i0}}$) and a particular day ($\boldsymbol{r_i}$). For this purpose, the commonly used Euclidean norm is adopted with a weight (defined by equation (2)) (Juna et al., 2022).

$$d_i|\boldsymbol{r_{i0}} = \sqrt{weigth \times sqaured\ distance\ \ from\ persent\ coordinates} \quad ...(01)$$

$$sqared\ distance\ \ from\ present\ coordinates = \sum_{s=1}^{S} (x_{i,s} - x_{i0,s})^2 \quad ...(02)$$

$$weight = \frac{total\ number\ of\ stations\ without\ the\ target\ station}{present\ coordinates} \quad ...(03)$$

Then, data at the target station ($\boldsymbol{c_{s0}}$) are ranked according to the ascending order of $d_i$ ($\boldsymbol{c_{s0}}|d_i$).

$$\boldsymbol{c_{s0}}|d_i \rightarrow x_{s0,k} = \{x_{s0,1}, \ldots, x_{s0,k}, \ldots, x_{s0,I}: d_k \leq d_{k+1}\}$$

Then, the target datum is estimated as the weighted mean:

$$\hat{x}_{i0,s0} = \sum_{k=1}^{K} d_k^{-1} x_{s0,k} \quad ...(04)$$

where K is the parameter indicating the number of nearest neighbours considered for the estimation.

**Missforest**

The Missforest imputation (S. Zhang et al., 2021) procedure starts by initializing an estimate for the missing values in the dataset. This initial estimate is commonly obtained through a simple imputation method such as mean imputation. Additionally, the variables in the dataset are sorted based on their respective counts of missing values, arranging them from lowest to highest. Each variable undergoes an imputation process in which missing values are imputed by employing a random forest. This involves training the random forest model with observed responses and predictors for the given variable. Subsequently, the trained random forest is applied to predict the missing values by utilizing the corresponding set of predictors for the missing entries. This procedure is repeated until the difference (normalized mean square error (NMSE)) (defined by equation (5)) between the imputed data ($X_{new}^{imp}$), and the previous iteration output data ($X_{old}^{imp}$) do not increase.

$$NMSE = \frac{\sum_{j \in N} \left(X_{new}^{imp} - X_{old}^{imp}\right)^2}{\sum_{j \in N} (X_{new}^{imp})^2} \quad ...(05)$$

where X = Matrix of data points, N = Number of data points.

**Spatiotemporal kriging**

In this study, a spatiotemporal process $\{Z(\boldsymbol{s}, t): (\boldsymbol{s}, t) \in D \subseteq \mathbb{R}^d \times \mathbb{R}\}$ was assumed, where $\boldsymbol{s} \in \mathbb{R}^d$, d=2 represented the two dimensions in space (latitude and longitude), and $t \in \mathbb{R}$ represented only one dimension, namely time. For this spatiotemporal process, spatiotemporal kriging can be performed mainly in two ways.

- Directly fit the variogram to the original data and perform kriging to impute missing values (Yang et al., 2018)
- Initially, fit a multiple linear regression model to the data and fit the variogram to the residuals of the multiple linear regression model (de Medeiros et al., 2019)

In the first approach, using original rainfall data ($Z(\boldsymbol{s},t)$), the sample variogram $\gamma_{st}(h_s, h_t)$ was calculated. Here, $h_s$ refers to a predefined spatial lag, representing a fixed distance interval between pairs of rainfall stations, while $h_t$ refers to a predefined temporal lag, representing a fixed time interval between observations. Then the theoretical model was selected that was best suited for the sample variogram based on the smallest mean square error (MSE) between the sample variogram and fitted theoretical variogram (this is achieved by selecting a variogram model (such as spherical, exponential, or Gaussian) and adjusting its parameters (range, sill, and nugget) to closely match the empirical or sample variogram, which is calculated from the observed data). After that, use the least-squares approach or maximum likelihood estimation (MLE) to fit the theoretical variogram model to the sample variogram. Then the fitted variogram can be used to impute the missing rainfall values.

In the second approach, the spatiotemporal variation $Z(\boldsymbol{s},t)$ was established by the decomposition of the deterministic *m (s, t)* and stochastic residual $\varepsilon(\boldsymbol{s},t)$ as follows:

$$Z(\boldsymbol{s},t) = m(\boldsymbol{s},t) + \varepsilon(\boldsymbol{s},t) \qquad \text{...(06)}$$

Where $m(\boldsymbol{s},t)$ is a deterministic part of the $Z(\boldsymbol{s},t)$ which can be estimated as follows (See equation (7)). Here $\varepsilon(\boldsymbol{s},t)$ is the residual term.

$$\widehat{m}(\boldsymbol{s},t) = \sum_{k=0}^{k} \widehat{\beta_k}\, X_k(\boldsymbol{s},t)\ ;$$

$$\widehat{\beta_k} = \textit{Estimated Coefficients of a multiple linear regression model} \qquad \text{...(07)}$$

Where $X_k(\boldsymbol{s},t)$ represents the observed predictors at the location $\boldsymbol{s}$ and time $t$ in the dataset.

Usually, $m(\boldsymbol{s},t)$ component is determined by an ordinary least square regression model. After that for the residual part, the same procedure was applied as in approach one.

**Denoising autoencoder**

Denoising autoencoders commonly use convolutional layers to build the neural network. It consists of two main parts: an encoder and a decoder (as shown in Figure 1)

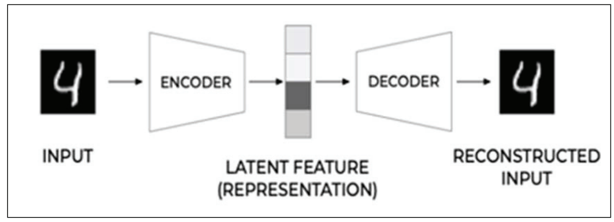


**Figure 1:** General structure of an autoencoder (Source: Michelucci, 2022, (http://arxiv.org/abs/2201.03898))

In general, the encoder can be written as a function $g$ that will depend on some parameters as in equation (8):

$$h_i = g(x_i) \qquad \text{...(08)}$$

Where $h_i \in \mathbb{R}^q$ (latent feature representation) is the output of the encoder $(i.e.\ g: \mathbb{R}^n \to \mathbb{R}^q; q < n,$ $\textit{where } n \textit{ is the original dimension of the data})$. The decoder (and the output of the network can be indicated as $\tilde{x}_i$ ) can be written then as a second generic function $f$ of the latent features as in equation (9):

$$\tilde{x}_i = f(h_i) = f(g(x_i)) \qquad \text{...(09)}$$

here $\tilde{x}_i \in \mathbb{R}^n$. Training and autoencoder simply means finding the functions $g(.)\ and\ f(.)$ that satisfy,

$$\arg\min_{f,g} < [\Delta(x_i, f(g(x_i)))] > \qquad \text{...(10)}$$

where $\Delta$ indicates a measure of how the input and the output of the Autoencoder differ (basically the loss function will penalize the difference between input and output) and $<.>$ indicates the average over all observations (Baldi, 2012).

**Further verification of imputed values**

Finally, to clarify whether the imputed dataset gives accurate forecasting results, LSTM (long-short-term memory) models were fitted using actual data and imputed data under different scenarios, and the forecasted results obtained from two types of models were compared by means of MAE. LSTM was selected as the past studies (Zhang et al., 2020) confirm its suitability in forecasting time series data for short as well as long horizons.

Three LSTM models were trained using three overlapping time periods for the training and test sets with the aim of obtaining generalized results. The training and test sets for each model were as follows:

- **Model 1:** The training set covered from January 2015 to December 2016, and the test set included a three-month period from January 2017 to March 2017.
- **Model 2:** The training set spanned from January 2016 to December 2017, with the test set covering January 2018 to March 2018.
- **Model 3:** The training set ranged from January 2017 to December 2018, while the test set included January 2019 to March 2019.

Forecasting accuracy was evaluated for models trained with actual data and imputed data, considering different random seeds and various lengths of missing data periods (one month, two weeks, and one week). These missing periods were selected because, in the Ratnapura area, sequential missing values in rainfall typically occur for a maximum duration of about one month. Therefore, the chosen scenarios reflect realistic conditions, covering up to one month of continuous missing data.

a) For ten random seeds, generate five one-month missing periods.
b) For ten random seeds, generate five two-week missing periods.
c) For ten random seeds, generate ten weekly missing periods.

## RESULTS AND DISCUSSION

### Results of the preliminary analysis

As the relationships between daily rainfall at the target station, Ratnapura, and its neighboring stations are not linear (for example, see Figure 2), Spearman's correlation coefficient was used to measure the correlations. Figure 3 shows the calculated correlation measures of these stations. It shows positive correlations between the target station, Ratnapura, and all reference stations, especially strong correlations (>0.6) with Keragala, Moralioya, and Elston. Therefore, it is worth incorporating the rainfall readings of these reference stations in imputing missing rainfall values at Ratnapura station.

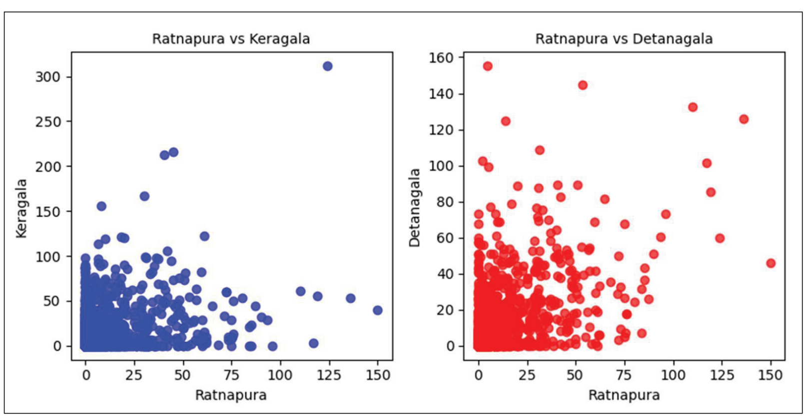


**Figure 2:** Scatter plots of daily rainfall at Ratnapura versus Keragala and Detanagala (2015–2019)

Figure 4 shows the distribution of daily rainfall at Rathnapura station by month. The plot shows high occurrences of rainfall from May to October. Additionally, an extreme rainfall event, which was very high compared to the other values, occurred in May. Conversely, during January to April and November to December, rainfall is lower compared to the other months. The details observed in the graph align with findings from previous studies (for example, Priyakumara & Ranasinghe, 2023). These studies have pointed out that the primary source of rainfall for the Ratnapura area is the southwest monsoon, typically occurring from May to September. Figure 4 confirms this matter, as it clearly illustrates high rainfall events during these months, consistent with the influence of the southwest monsoon.

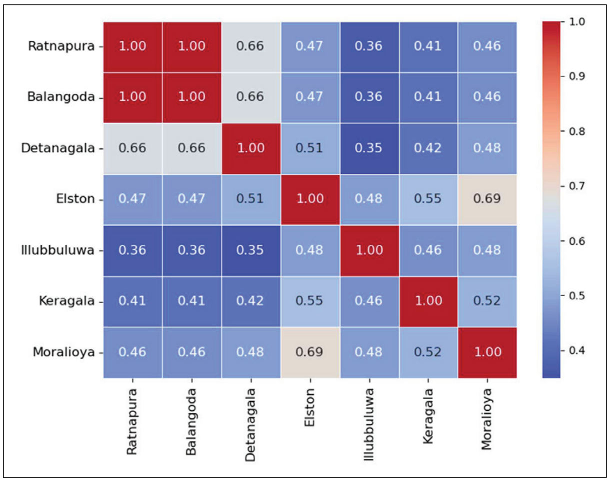


**Figure 3:** The heatmap of the correlation between rainfall gauging stations

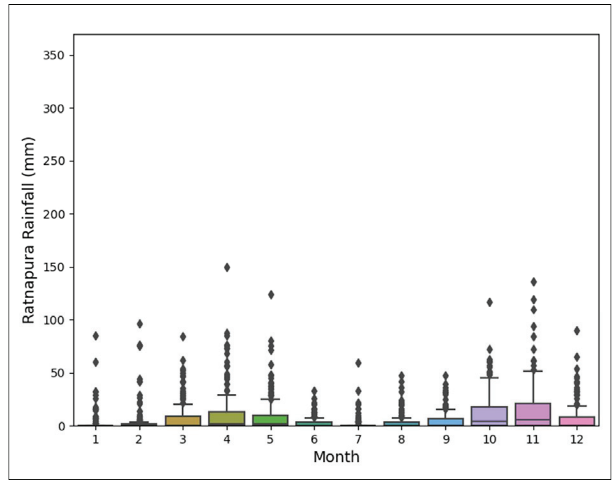


**Figure 4:** Distribution of daily rainfall at Rathnapura station by months (from 2015 to 2019)

Figures 5 and 6 present the distributions of daily rainfall at reference stations by month. It indicates that other stations exhibit similar rainfall patterns as Ratnapura station, with minor deviations in some months. Since this section mainly focuses on identifying the wet and dry seasons in the Ratnapura area, the variation of rainfall in other stations is considered less significant. In here, the wet season, characterized by high rainfall events, occurs from May to September, while the dry season, with lower rainfall, takes place from January to April and from October to December.

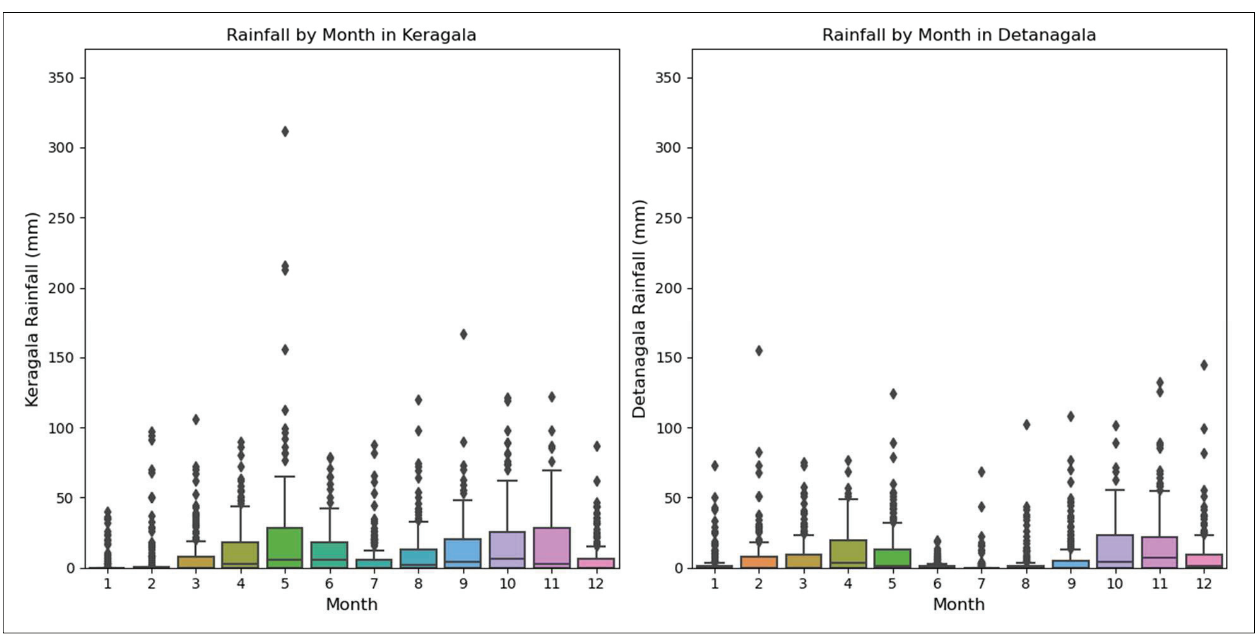


**Figure 5:** Distribution of daily rainfall of Keragala and Detanagala stations by months (from 2015 to 2019)

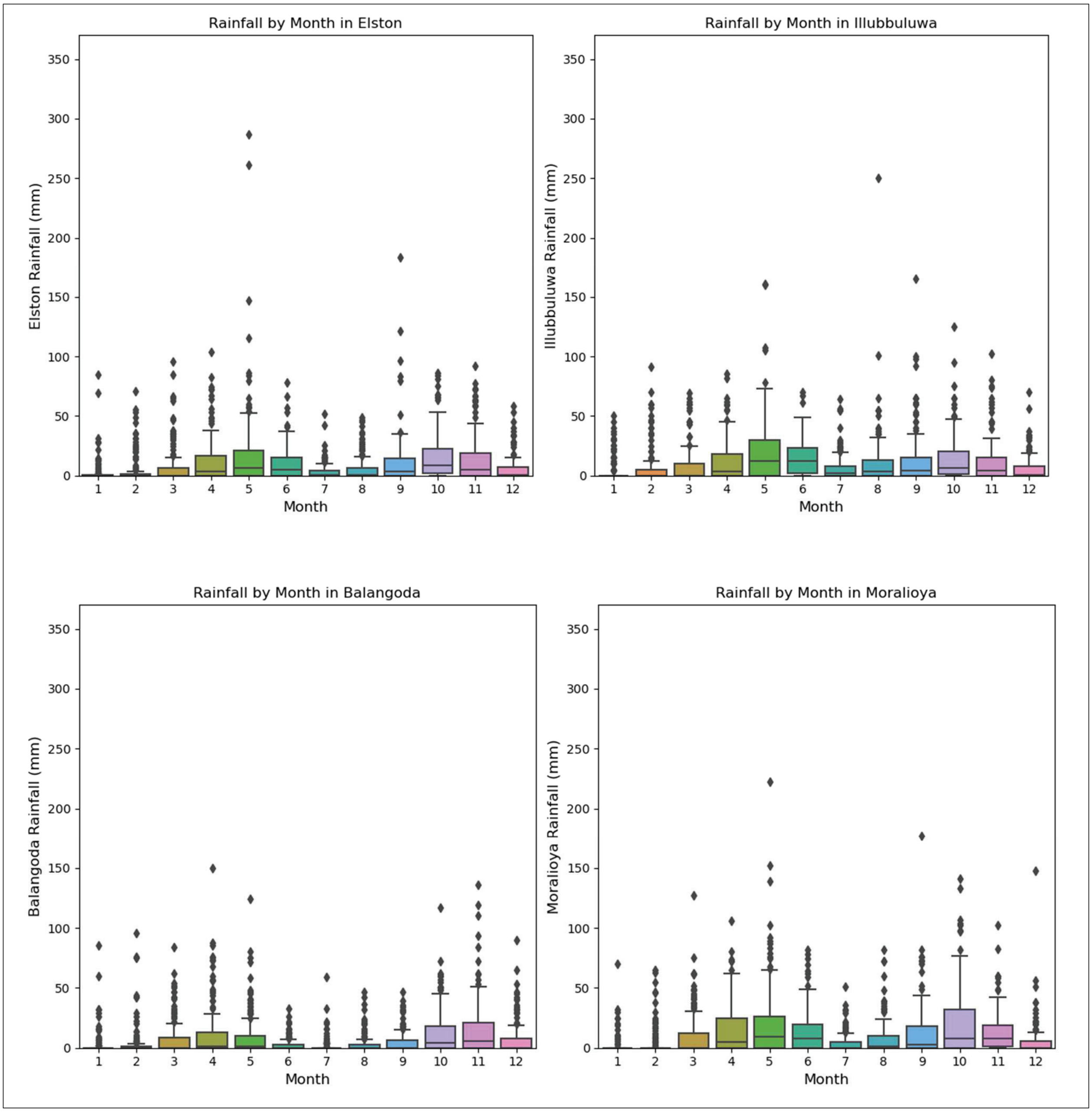


**Figure 6:** Distribution of daily rainfall of Elston, Illubbuluwa, Balangoda, and Moralioya stations by months (from 2015 to 2019)

**Evaluation of underlying assumptions for imputation methods**

For spatiotemporal kriging, the validity of underlying assumptions, including second-order stationarity and spatial isotropy (Gräler et al., 2016) was assessed prior to application. To assess spatial isotropy, directional spatial semi-variograms were computed (see Figure 7). The semi-variograms differed across directions, indicating the presence of anisotropy (Oliveira et al., 2022), which was subsequently incorporated into the fitted theoretical variogram models. In addition, the temporal stationarity of each station's time series was evaluated using the augmented Dickey-Fuller (ADF) test. All series returned p-values close to 0, indicating that all the time series were stationary.

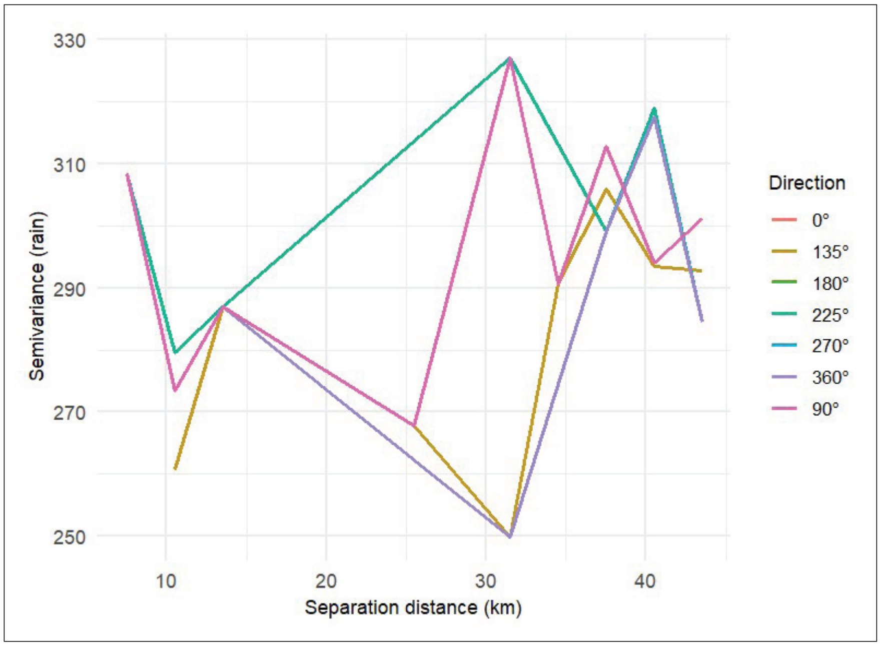


**Figure 7:** Directional spatial semi-variograms

In contrast, MissForest (Stekhoven & Bühlmann, 2012), KNN, and denoising autoencoder methods do not require any prior assumptions regarding the data distribution or correlation structure. Consequently, these nonparametric and machine learning–based approaches were applied directly, and the evaluation of imputation accuracy was performed without preliminary assumption checking.

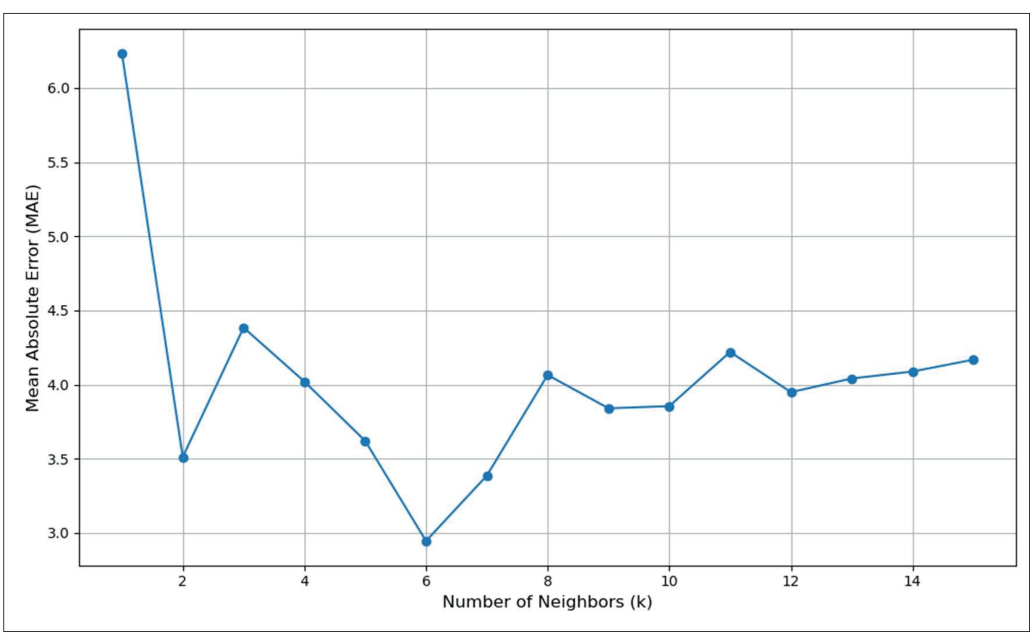


**Figure 8:** Plot of K values vs MAE values in KNN imputation

**Comparison of performance of the four missing value imputation techniques**

Table 1 presents the MAE and RMSE values obtained to compare the accuracy of imputed missing values for the wet and dry seasons in 2019. For KNN and MissForest, no train-test split was used, as these lazy learning (instance-based) algorithms automatically identify and iteratively impute missing values directly on the dataset without producing a reusable model. Therefore, these techniques were implemented using data from 2015 to 2019, the period in which sequential missing values are included. Furthermore, when performing KNN, the optimal K value was selected using the MAE vs K value plot (see Figure 8). In this study, the range of K values considered spans from 1 to 15. The criterion for selecting the optimal K value is based on identifying the K value that yields the minimum MAE for the imputation.

For spatiotemporal kriging, to fit the variogram, rainfall values from 2015 to 2019 of reference stations were used, as it needs to have complete rainfall records of other stations to impute Ratnapura station's (unknown location according to the kriging theory) rainfall missing values in the period of 2015 to 2019. However, for the denoising autoencoder, training was conducted using data from 1994 to 2014, with the test set used for imputation spanning 2015 to 2019, as discussed in Section 2. The reason for using a full dataset to implement this method is that denoising autoencoders consist of convolutional layers with more parameters (weights) to tune during the model-training procedure. Therefore, this model needs to have sufficient data to overcome the overfitting problem when training.

Table 1 shows that the MAE and RMSE values of the KNN algorithm (in the wet season) are the lowest. Therefore, it is evident that for sequential missing value imputation for the wet season, KNN outperformed the other three techniques. For the dry season, KNN outperformed all the other techniques when the missing period is one week and one month. However, there is a conflict between KNN and Missforest imputation results for the two-week missing value period. In this case, Missforest performed better than KNN according to both the RMSE and MAE values. This is because both methods imputed a low rainfall event (2019-Jan-11) as a high rainfall event and an extreme rainfall event (2019-Jan-14) as a low rainfall event (See Figure 9). Referring to Table 1, it can be mentioned that both KNN and Misforest produce better imputation results in dry periods than wet periods. This may be due to the lower variability of rainfall in dry seasons.

**Table 1:** Sequential missing values imputation results for four methods in 2019

| Missing Rates | KNN | | Missforest | | Spatiotemporal Kriging | | Denoising Autoencoder | |
|---|---|---|---|---|---|---|---|---|
| | MAE (mm) | RMSE (mm) | MAE (mm) | RMSE (mm) | MAE (mm) | RMSE (mm) | MAE (mm) | RMSE (mm) |
| Wet: Week | 1.85 | 2.99 | 4.26 | 5.72 | 2.90 | 4.94 | 1.99 | 2.51 |
| Dry: Week | 0.22 | 0.31 | 0.35 | 0.38 | 5.86 | 6.60 | 1.08 | 1.09 |
| Wet: Two Weeks | 4.60 | 7.88 | 6.78 | 10.88 | 7.87 | 12.89 | 6.87 | 12.12 |
| Dry: Two Weeks | 3.21 | 6.48 | 2.54 | 5.37 | 6.25 | 7.32 | 3.71 | 5.55 |
| Wet: Month | 6.21 | 9.13 | 7.82 | 10.66 | 11.47 | 19.39 | 9.13 | 13.86 |
| Dry: Month | 1.87 | 4.52 | 2.24 | 4.50 | 7.91 | 8.81 | 2.69 | 4.13 |

MAE and RMSE represent mean absolute error and Root mean square error, respectively

To verify the performance of the four techniques, the experiments were repeated for another two years. RMSE and MAE were used to evaluate performance. For this purpose, 2015 and 2017 were selected. Tables 2 and 3 compare MAE and RMSE values obtained from the four techniques for 2015 and 2017, respectively.

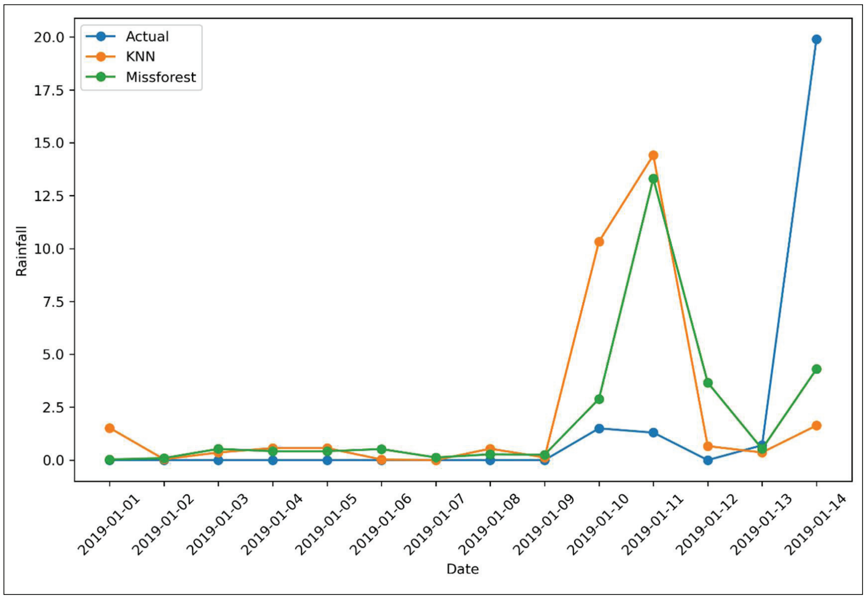


**Figure 9:** Actual vs imputed rainfall values with KNN and Missforest (1st to 14th January 2019)

According to Tables 2 and 3, in most of the missing scenarios, KNN outperformed the other three techniques in the aforementioned two years, as well as in 2019. Another noticeable point is that the MAE and RMSE values of all the techniques showed higher values in the wet season of 2017. The reason for this is that there is an extreme rainfall (120mm) compared to other observations in that period.

Although spatiotemporal kriging was expected to outperform MissForest and KNN, as reported in previous research (Yang et al., 2018), it exhibited the lowest performance in this study, with higher MAE and RMSE values across all missing data scenarios (Tables 1–3) in both wet and dry seasons. As per (de Medeiros et al., 2019), this can be due to the lack of capturing spatiotemporal rainfall variability of the fitted variogram, and also due to the small number of locations. As a solution for the lack of capturing spatiotemporal rainfall variability, fitting a deterministic trend component to the rainfall data to capture some of the spatiotemporal variability that cannot be captured by the spatiotemporal variogram was suggested by several researchers (de Medeiros et al., 2019; Hu et al., 2017; van Zoest et al., 2020). However, fitting a deterministic component $m(\boldsymbol{s}, t)$, as defined in sub-section 2.3 and Equation (6), was not appropriate,

**Table 2:** Sequential missing values imputation results for four methods in 2015

| Missing rates | KNN | | Missforest | | Spatiotemporal kriging | | Denoising autoencoder | |
|---|---|---|---|---|---|---|---|---|
| | MAE (mm) | RMSE (mm) | MAE (mm) | RMSE (mm) | MAE (mm) | RMSE (mm) | MAE (mm) | RMSE (mm) |
| Wet: Week | 6.12 | 10.65 | 5.27 | 7.05 | 8.52 | 13.66 | 8.54 | 11.56 |
| Dry: Week | 2.07 | 5.08 | 2.75 | 4.89 | 5.39 | 6.08 | 2.25 | 3.45 |
| Wet: Two weeks | 5.11 | 7.40 | 5.10 | 6.37 | 7.80 | 11.30 | 8.33 | 10.68 |
| Dry: Two weeks | 1.04 | 3.59 | 1.49 | 3.52 | 6.55 | 7.02 | 1.55 | 2.30 |
| Wet: Month | 7.27 | 10.26 | 7.65 | 10.24 | 10.48 | 15.93 | 8.54 | 16.60 |
| Dry: Month | 2.40 | 4.62 | 3.09 | 5.67 | 10.05 | 12.24 | 2.97 | 6.20 |

as most of the coefficients in the deterministic regression model were not significant, the model explained little of the variability ($R^2 = 0.2$), and the residuals violated assumptions such as normality.

**Table 3:** Sequential missing values imputation results for four methods in 2017

| Missing rates | KNN | | Missforest | | Spatiotemporal kriging | | Denoising autoencoder | |
|---|---|---|---|---|---|---|---|---|
| | MAE (mm) | RMSE (mm) | MAE (mm) | RMSE (mm) | MAE (mm) | RMSE (mm) | MAE (mm) | RMSE (mm) |
| Wet: Week | 30.84 | 48.17 | 34.73 | 51.31 | 41.89 | 67.18 | 35.23 | 56.14 |
| Dry: Week | 0.71 | 0.36 | 0.93 | 1.28 | 5.98 | 6.66 | 0.821 | 0.91 |
| Wet: Two weeks | 18.32 | 35.49 | 22.87 | 38.37 | 27.80 | 49.35 | 24.83 | 42.82 |
| Dry: Two weeks | 0.23 | 0.80 | 0.99 | 1.06 | 8.25 | 9.37 | 0.974 | 1.06 |
| Wet: Month | 12.84 | 25.32 | 13.24 | 26.04 | 22.15 | 36.96 | 15.86 | 30.98 |
| Dry: Month | 1.76 | 3.89 | 3.71 | 6.65 | 10.08 | 11.81 | 1.36 | 2.24 |

In general, KNN performs better than the other three imputation techniques considered in this study.

## Forecasting results of the LSTM model

A similar procedure to the study done by Ahn et al. (2021), which was conducted to compare missing data imputation methods in time series forecasting, was followed. Here, forecasted values from LSTM models trained on actual and KNN-imputed datasets were compared to examine the appropriateness of using imputed values for rainfall forecasting. The three LSTM models were fitted, and forecasting was carried out according to the procedure described in sub-section 2.5. For all the missing data scenarios tested in this study, a single set of parameters was found to be optimal from the search space considered. Table 4 summarizes the optimal configuration, which was applied consistently across all scenarios and the search space of each parameter where these optimal values were obtained.

**Table 4:** Optimal parameters for LSTM models and search space

| Parameter | Optimal value | Search space |
|---|---|---|
| Look-back period | 4 days | 2, 3, 4, 5, 6 |
| Number of LSTM layers | 1 | 1, 2, 3 |
| LSTM units per layer | 32 | 16, 32, 64, 128 |
| Batch size | 32 | 16, 32, 64, 128 |
| Learning rate | 0.001 | 0.0001, 0.001, 0.005, 0.01 |
| Dropout rate | 0.2 | 0.1, 0.2, 0.3, 0.4 |
| Epochs | 100 | 50, 100, 150, 200 |

Table 5 compares the MAE values calculated from forecasted values obtained from the two types of LSTM models (fitted with actual and imputed data). Only the test data sets are used for calculating MAE. Results show that MAE values obtained from the LSTM models trained with imputed data do not deviate much from those obtained from the counterpart models. This suggests that the KNN-imputed data are closely aligned with the actual dataset, consistent with findings from previous studies (Ahn et al., 2021), which reported KNN as one of the most accurate imputation techniques for time series forecasting.

**Table 5:** MAE values (for test data) of the LSTM Models trained with imputed and actual data

| Train/Test period | Missing period length | Without missing values | Seed | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | 1234 | 2222 | 2789 | 3456 | 4821 | 5093 | 6178 | 7342 | 8569 | 9027 |
| 2015-2016/ 2017 Jan-March | One week | 6.67 | 6.59 | 6.50 | 6.66 | 6.55 | 6.72 | 6.76 | 6.66 | 6.62 | 6.66 | 6.60 |
| | Two weeks | | 6.60 | 6.72 | 6.60 | 6.49 | 6.53 | 6.56 | 6.63 | 6.51 | 6.64 | 6.59 |
| | One month | | 6.62 | 6.55 | 6.61 | 6.55 | 6.28 | 6.61 | 6.34 | 6.40 | 6.57 | 6.28 |
| 2016-2017/ 2018 Jan-March | One week | 6.84 | 6.83 | 6.88 | 6.50 | 6.54 | 6.61 | 6.61 | 6.54 | 6.61 | 6.61 | 6.62 |
| | Two weeks | | 6.90 | 6.74 | 6.49 | 6.55 | 6.60 | 6.63 | 6.51 | 6.59 | 6.61 | 6.56 |
| | One month | | 6.82 | 6.52 | 6.42 | 6.49 | 6.53 | 6.65 | 6.45 | 6.58 | 6.63 | 6.54 |
| 2017-2018/ 2019 Jan-March | One week | 11.00 | 11.04 | 10.76 | 10.99 | 10.88 | 10.87 | 10.99 | 10.79 | 10.98 | 10.99 | 11.01 |
| | Two weeks | | 11.17 | 10.95 | 11.01 | 10.96 | 10.95 | 10.97 | 11.01 | 11.02 | 11.01 | 10.99 |
| | One month | | 10.91 | 11.01 | 11.10 | 10.98 | 11.13 | 11.10 | 10.99 | 10.97 | 10.95 | 10.93 |

## CONCLUSIONS

This study investigated the performance of different spatiotemporal imputation techniques for sequential missing values in rainfall data collected from the Ratnapura gauging station, with particular attention to how imputation accuracy varies across wet and dry seasons and to the sensitivity of methods to varying durations. According to the imputation results, it can be concluded that KNN outperformed the other three techniques in both wet and dry seasons in imputing sequential missing values of rainfall series collected from Ratnapura gauging station. The Missforest algorithm showed better performance compared to denoising autoencoder and spatiotemporal kriging during both wet and dry seasons, closely matching the results of the KNN algorithm, but worse than those of KNN in general. However, spatiotemporal kriging showed weaker performance in both wet and dry seasons compared to KNN, MissForest, and denoising autoencoder. Due to its limitation in capturing rainfall variability through the deterministic component, future studies are recommended to include additional meteorological variables when modeling spatiotemporal variability to improve imputation accuracy. Even though these techniques showed reasonable accuracy, they were not successful in imputing sudden extreme rainfall events accurately. Therefore, it is vital to further improve the methods applied to this study or develop novel techniques to impute such sudden extreme rainfall events accurately. Another important observation is that as the duration of missing values increases, the accuracy of the imputation methods decreases consistently across all techniques.

The LSTM forecasting results for both the imputed and actual datasets show very similar MAE values. This indicates that the imputation process, particularly using the KNN technique, has a minimal impact on the forecasting accuracy. Therefore, KNN can be recommended as a reliable method to impute sequential missing values of rainfall data at the Rathnapura gauging station, and such imputed values can be used in building flood forecasting models for Rathnapura.

**Acknowledgment**

Rainfall data purchased from the Meteorology Department of Sri Lanka for another research project (a PhD research) funded by the University of Colombo collaborative research grant No. AP/3/2/2019/CG/30, were used for this study, with the permission of the grant holders. The second author is the principal investigator of this grant.

## REFERENCES

Ahn, H., Sun, K., & Kim, K. P. (2021). Comparison of missing data imputation methods in time series forecasting. *Computers, Materials and Continua*, *70*(1), 767–779. https://doi.org/10.32604/cmc.2022.019369

Bae, B., Kim, H., Lim, H., Liu, Y., Han, L. D., & Freeze, P. B. (2018). Missing data imputation for traffic flow speed using spatio-temporal cokriging. *Transportation Research Part C: Emerging Technologies*, *88*(February), 124–139. https://doi.org/10.1016/j.trc.2018.01.015

Baldi, P. (2012). Autoencoders, Unsupervised Learning, and Deep Architectures. *ICML Unsupervised and Transfer Learning*, 37–50. https://doi.org/10.1561/2200000006 de Medeiros, E. S., de Lima, R. R., de Olinda, R. A., & os Santos, C. A. C. (2019a). Modeling spatiotemporal rainfall variability in Paraíba, Brazil. *Water (Switzerland)*, *11*(9). https://doi.org/10.3390/w11091843

Feng, L., Nowak, G., O'Neill, T. J., & Welsh, A. H. (2014). CUTOFF: A spatio-temporal imputation method. *Journal of Hydrology*, *519*(PD), 3591–3605. https://doi.org/10.1016/j.jhydrol.2014.11.012

Gräler, B., Pebesma, E., & Heuvelink, G. (2016). Spatio-temporal interpolation using gstat. *R Journal*, *8*(1), 204–218. https://doi.org/10.32614/rj-2016-014

Hu, D., Shu, H., Hu, H., & Xu, J. (2017). Spatiotemporal regression Kriging to predict precipitation using time-series MODIS data. *Cluster Computing*, *20*(1), 347–357. https://doi.org/10.1007/s10586-016-0708-0

Juna, A., Umer, M., Sadiq, S., Karamti, H., Eshmawi, A. A., Mohamed, A., & Ashraf, I. (2022). Water Quality Prediction Using KNN Imputer and Multilayer Perceptron. *Water (Switzerland)*, *14*(17), 1–19. https://doi.org/10.3390/w14172592

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

Li, H., Li, M., Lin, X., He, F., & Wang, Y. (2020). A spatiotemporal approach for traffic data imputation with complicated missing patterns. *Transportation Research Part C: Emerging Technologies*, *119*(July), 102730. https://doi.org/10.1016/j.trc.2020.102730

Mital, U., Dwivedi, D., Brown, J. B., Faybishenko, B., Painter, S. L., & Steefel, C. I. (2020). Sequential Imputation of Missing Spatio-Temporal Precipitation Data Using Random Forests. *Frontiers in Water*, *2*(August), 1–15. https://doi.org/10.3389/frwa.2020.00020

Oliveira, A. L. G., Lima, J. P., Brasco, T. L., & Amaral, L. R. (2022). The importance of modeling the effects of trend and anisotropy on soil fertility maps. *Computers and Electronics in Agriculture*, *196*. https://doi.org/10.1016/j.compag.2022.106877

Priyakumara, U. L. A. S., & Ranasinghe, E. M. S. (2023). Rainfall Variability in Ratnapura District: A Comparative Study of the Two 30-year Standard Periods of 1961-1990 and 1991-2020. *Sri Lanka Journal of Social Sciences and Humanities*, *3*(2), 165–177. https://doi.org/10.4038/sljssh.v3i2.109

De Silva, R. P., Dayawansa, N. D. K., & Ratnasiri, M. D. (2007). A comparison of methods used in estimating missing rainfall data. *Journal of Agricultural Sciences* – Sri Lanka, 3(2), 101–108. https://doi.org/10.4038/jas.v3i2.8107

Stekhoven, D. J., & Bühlmann, P. (2012). Missforest-Non-parametric missing value imputation for mixed-type data. *Bioinformatics*, *28*(1), 112–118. https://doi.org/10.1093/bioinformatics/btr597

Thevakaran, A., Suppiah, R., & Sonnadara, U. (2019). Trends in extreme rainfall events in Sri Lanka,1961-2010. *Journal of the National Science Foundation of Sri Lanka*, *47*(3), 285–295. https://doi.org/10.4038/jnsfsr.v47i3.9280

Van Zoest, V., Osei, F. B., Hoek, G., & Stein, A. (2020). Spatio-temporal regression kriging for modelling urban NO2 concentrations. *International Journal of Geographical Information Science*, *34*(5), 851–865. https://doi.org/10.1080/13658816.2019.1667501

Wardana, I. N. K., Gardner, J. W., & Fahmy, S. A. (2022). Estimation of missing air pollutant data using a spatiotemporal convolutional autoencoder. *Neural Computing and Applications*, *34*(18), 16129–16154. https://doi.org/10.1007/s00521-022-07224-2

Weerakoon, K. G. P. K., & Dharmapriya, J. A. D. U. (2021). Factors Affecting the Effectiveness of Flood Resilience Plan for Urban Sustainability of Ratnapura Municipal Council Area. *International Journal of Real Estate Studies*, *15*(2), 32–42. https://doi.org/10.11113/intrest.v15n2.127

Yang, H., Yang, J., Han, L. D., Liu, X., Pu, L., Chin, S. M., & Hwang, H. L. (2018a). A Kriging based spatiotemporal approach for traffic volume data imputation. *PLoS ONE*, *13*(4), 1–11. https://doi.org/10.1371/journal.pone.0195957

Zhang, C. J., Zeng, J., Wang, H. Y., Ma, L. M., & Chu, H. (2020). Correction model for rainfall forecasts using the LSTM with multiple meteorological factors. *Meteorological Applications*, *27*(1), 1–15. https://doi.org/10.1002/met.1852

Zhang, S., Gong, L., Zeng, Q., Li, W., Xiao, F., & Lei, J. (2021). Imputation of GPS coordinate time series using missforest. *Remote Sensing*, *13*(12), 1–17. https://doi.org/10.3390/rs13122312

Zhu, M., & Cheng, X. (2016). Iterative KNN imputation based on GRA for missing values in TPLMS. *Proceedings of 2015 4th International Conference on Computer Science and Network Technology, ICCSNT 2015*, *Iccsnt*, 94–99. https://doi.org/10.1109/ICCSNT.2015.7490714