Imputing Missing Data in Hydrology using Machine Learning Models

Article *in* International Journal of Engineering and Technical Research · January 2021

DOI: 10.17577/IJERTV10IS010011

2 authors:

Vasker Sharma
Royal University of Bhutan
9 PUBLICATIONS **6** CITATIONS
SEE PROFILE

Kezang Yuden
College of Science and Technology
1 PUBLICATION **3** CITATIONS
SEE PROFILE

Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 01, January-2021

Imputing Missing Data in Hydrology using Machine Learning Models

Vasker Sharma *
Department of Civil Engineering and Surveying,
Jigme Namgyel Engineering College, Dewathang,
Samdrupjongkhar, Royal University of Bhutan

Kezang Yuden
Department of Civil Engineering and Surveying,
Jigme Namgyel Engineering College, Dewathang,
Samdrupjongkhar, Royal University of Bhutan

***Abstract*— Missing data has been a common problem and has been confronted by many researchers in the field of hydrology. Rainfall and Temperature time series data are often found missing and such missingness have huge implication on hydrological modelling, flood frequency analysis, trend analysis and dam operation schemes. Owing to the presence of missing data it hinders the performance analysis of the data and inhibits in concluding the correct inferences from the data. In this study, missing data in the rainfall and temperature has been imputed using kNN model and Tree-based model and subsequently these imputed data have been used as predictors to predict the river flow data using Artificial Neural Network (ANN). Uncertainty from kNN imputation model has been found with bootstrapping techniques, while the tree based and ANN model were assessed by Root Mean Square Error (RMSE) and Mean Absolute Error (MAE).**

Keywords*— *Missing values, hydrology, kNN, ANN, Regression, Decision tree.

I. INTRODUCTION

The long-term hydro-meteorological variables can be utilized for understanding the regional weather and climate and can also be used for the vulnerability assessment of water resource within the region or community of interest [1]. Such data can also be used for planning and managing the water resource at the basin level using different physical based models like hydrological and hydraulic models. However, such variables are often confronted with missing data which makes the analysis difficult or sometimes makes it impossible to analyse. The missingness is ubiquitously introduced owing to defect in the recording sensors, during relocation of the sensors and errors made while noting or observing the data. Due to the presence of the missing rainfall and flow data, it becomes increasingly difficult to calibrate and validate the hydrological model for a basin. Furthermore, robust and complete data is of utmost important for regional flood frequency analysis, hydraulic design and Dam operation schemes. Rainfall and temperature are the key environmental variables that are used to understand the different atmospheric, cryosphere and climatic processes within any region of interest. In absence of complete observed data, many studies [2][3] resort to remote sensing-based data which are model based coupled with station data. However, such data has higher uncertainty and bias introduced when downscaling to station data.

Ignoring the missing data of one variable usually means compromising the observed data of other variables, which results in drastic loss of overall data. Such case is usually confronted when using statistical models like Multiple Linear Regression Model (MLRM) which assumes the linearity among the different observed variables, where corresponding observed value of one variable is dropped for the missingness of other variable which thereby results in the loss of statistical inferences of the data. On the other hand, physical-based models like hydrological models per se can be used to impute the missing river flow data provided that the model is well calibrated and validated using historically observed data. However, finding such long-term historically observed data is not only difficult for an ungauged basin but the data per se tends to have missing data. Therefore, it necessitates to impute the missing data preferably using some contemporary Machine Learning models. Machine Learning models like Artificial Neural Network is basically evolved based on the working method of human brain for classification, identification and recognition. Primarily developed for the medical and neurological study [4], it now find its practice in myriads of disciplines. Although Machine learning concepts were developed as early as 1950s, the applicability has indeed progressed only in recent decades owing to technological advancement in computational power of a computer. Researchers now are able to leverage such technologies to develop complex machine learning models which have a vast usage in different application areas.

However, imputing missing values disparagingly depends on the nature of missing data which is described as follows [5][6][7]:

i. MCAR: Missing Completely at Random, where missingness has no association with any data that is observed or not observed. In such case imputation is advisable. Discarding the missing data will not bias the data however it will lead to loss of sample size especially dealing with multiple variables.
ii. MAR: Missing at Random, where missingness in one variable depends on some other observed variable and discarding missing values may bias the overall data which is not considered ideal for this case. Imputation has to be carried out cautiously.
iii. MNAR: Missing Not at Random, where missingness of the one variable is related to an unobserved value in some other variable relevant to the assessment of interest.

Several researcher [8][9][10] has used machine learning models for estimating flow missing data and have achieved reliable accuracy. [11] compared the Machine Learning and Hydrological models as the imputation model and found that Machine learning performs better in imputing missing data. ANN can be used as an alternative to different environmental and physical-based models used for Rainfall-Runoff

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 01, January-2021

modelling, ground water modelling, water quality modelling and flow modelling and it is indeed found more accurate [12]. Nevertheless, [13] also consider the Normal Ratio Method (NRM) and Inverse Distance Weighted Method (IDM) for imputing missing rainfall data. Although, such method can reconstruct the data at lower frequency (i.e. annual time series), it has to resort to percentage contribution of nearest rainfall station to annual rainfall in order to convert to higher frequency data (i.e. daily or hourly time series).

[14] has used Recurrent Neural Network (RNN) for imputing the missing data of a time series and such model achieved reasonable accuracy providing useful information of the data. Studies like [15][16] has provided wider perspectives of using machine learning models for the analysis of water quality data. Further, [17] showcased the use of ANN and Support Vector Machine (SVM) to predict the nonlinear time series like ground water level and found that SVM performed slightly better than the ANN model, nevertheless both the models well represented the nonlinearity of the data. [18] studied predictability of a flow using kNN and ANN model for different scenarios and advocated that kNN offers better predictability of flow. [19] used regression trees and ANN to reconstruct the missing rainfall data which provided promising streamflow prediction using hydrological models such as Soil Water Assessment Tool (SWAT). [13] and [20] advocated the use of random forest based decision trees to reconstruct the missing values in rainfall data while [21] used sequential imputation considering random forest technique.

In this study, multiple linear regression model, kNN imputation and decision tree-based imputation were used to impute the missing data in rainfall and temperature. MLRM was assessed based on the regression coefficient while uncertainty of kNN imputation were carried out using bootstrapping techniques and decision trees were assessed with Root Mean Square Error. Further, these imputed data were used as predictors to predict the flow in the two gauging station located in the basin using ANN considering back propagation technique. The choice of predictors used for the predicting is solely dependent on the background knowledge of the user and its relationship with the response variable.

II. DATA AND METHODS

The flow data, rainfall and temperature data has been obtained from National Centre of Hydrology and Meteorology (NCHM), Thimphu, Bhutan. Rainfall and temperature data from six meteorological station has been used which fall in the Kholongchu basin located in the eastern region of the Himalayan Kingdom of Bhutan. There are also two flow gauging station as shown in the figure 1.

In the Rainfall and temperature data, there were few missing data and these missing data has been imputed using the approaches stated below. Having the missing data imputed and its efficiency validated, these data has been used as a predictor to predict the flow in the basin.

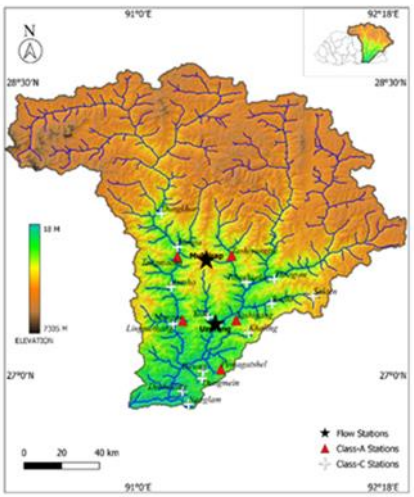

Fig 1 Location of meteorological station

Table 1: Station Details

Station	Abbreviated	Lat	Lon
Radhi	Radhi	27.3647	91.7081
Kanglung	klung	27.2820	91.5185
Tshenkarla	tkarla	27.4754	91.5722
Trashi yangtse	tyang	27.6127	91.4946
Yadi	yadi	27.2961	91.3687
Sherichu	Sherichu	27.2580	91.4165
Uzorong	UZ	27.26	91.41
Muktirap	MR	27.59	91.36

A. Imputation with Multiple Linear Regression Model (MLRM)

In this method, missing values in one station (response variable) was imputed with regressing with the multiple other station (independent variables) where data was complete. Months (a categorical variable) were also used as an independent variable for imputing the missing data. R-package by [22] has been used to impute the missing values. Once the missing data in the station of interest was imputed, it was subsequently treated as independent variable to impute the missing data in remaining stations. Mathematically, MLRM for rainfall were performed in the following way:

$$P\_radhi \sim \hat{\beta}_0 + \hat{\beta}_1.P\_klung + \hat{\beta}_2.P\_tyang + \hat{\beta}_3.Month \quad (1)$$

$$P\_tkarla \sim \hat{\beta}_0 + \hat{\beta}_1.P\_klung + \hat{\beta}_2.P\_tyang + \hat{\beta}_3.P\_radhi + \hat{\beta}_4.Month \quad (2)$$

$$P\_yadi \sim \hat{\beta}_0 + \hat{\beta}_1.P\_klung + \hat{\beta}_2.P\_tyang + \hat{\beta}_3.P\_radhi + \hat{\beta}_4.P\_tkarla + \hat{\beta}_5.Month \quad (3)$$

$$P\_Sherichu \sim \hat{\beta}_0 + \hat{\beta}_1.P\_klung + \hat{\beta}_2.P\_tyang + \hat{\beta}_3.P\_radhi + \hat{\beta}_4.P\_tkarla + \hat{\beta}_5.P\_yadi + \hat{\beta}_6.Month \quad (4)$$

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 01, January-2021

Where initially rainfall in Kanglung(klung) and Trashi Yangtse (tyang) station was complete. Similarly, MLRM for maximum and minimum temperature were also performed

B. Imputation with k-Nearest Neighbours (kNN model)

kNN uses the distance-weighted aggregation techniques where it aggregates the values from the neighbours to obtain the replacement for a missing value. It does so using the weighted mean, where weights are inverted distances from each neighbour. Closer neighbour has more impact on the imputed values. It is often a good practice to sort the variables increasingly by number of missing values before performing kNN imputation. Here, kNN imputation was performed considering 5 (k=5) nearest neighbourhood using R- package developed by [23].

Uncertainty of kNN imputation model with bootstrapping techniques

Whenever analysis or modelling is performed on imputed data, uncertainty from imputation should be adequately accounted. Running a model or performing an analysis on one-time imputed data ignores the fact that imputation estimates the missing values with uncertainty. The solution to this is multiple imputation and one way to implement is by bootstrapping. Bootstrapping is one technique where data are sampled with replacement to get the original data. It is a technique to get the inference of a population data using a sample data. It works with MCAR and MAR data. Here multiple imputation with 1000 boot replicates were generated where each boot replicate represents the regression coefficients calculated as per eqn. 5. Subsequently standard error and bias associated with the replicated and original data were assessed.

$$P\_Sherichu \sim \hat{\beta}_0 + \hat{\beta}_1.P\_klung + \hat{\beta}_2.P\_tyang + \hat{\beta}_3.P\_radhi + \hat{\beta}_4.P\_tkarla + \hat{\beta}_5.P\_yadi + \hat{\beta}_6.Month \quad (5)$$

C. Tree based imputation with random forest

It is based on the non-parametric approach where no assumption is made on the relationship between variables. It can pick up the complex nonlinear patterns and it is often better than the statistical models. Tree based imputation uses random forest behind the hood and builds separate random forest to predict the missing values for each variable one by one. In this study, it utilizes Miss Forest imputation algorithm in R-environment developed by [24], where in the first iteration, missing data is initially imputed with mean of the data and then for each variable containing missing values, it fits a random forest based on the non-missing values and then later predicts the missing values. The iteration continues to repeat until it reaches a stopping criterion or meets the user-specified iteration number.

The algorithm also gives the Out-of-Bag (OOB) error associated with the imputation and hence there is no need to evaluate its efficiency separately. Here the error has been minimized by taking a 1000 decision trees. Increasing the decision tree might improve the imputation model, but it will also require higher computation time, therefore, there is always a speed-accuracy trade-off to be made during computation.

D. Artificial Neural Network (ANN)

In this method, imputed data from the kNN model were considered as an input vectors for the neural network along with the flow data from the two-flow gauging station as an output vector. Here, Neural network was developed using neuralnet R-package developed by [25] to predict the flow in the Uzorong and Muktirap station with the different inputs vectors as shown in the Figure 2. Logistic Activation function with backpropagation option were used while running the ANN model. The stopping criteria for the model simulation was based on an error threshold of 0.01. To have an adequate predictability of the ANN model, the data were first transformed using min-max normalization technique (eqn. 6), by which all the data ranged from 0 to 1. Subsequently the data was randomly split into training and testing data, where each variable in training data had 4458 observation while testing data had 1486 observation.

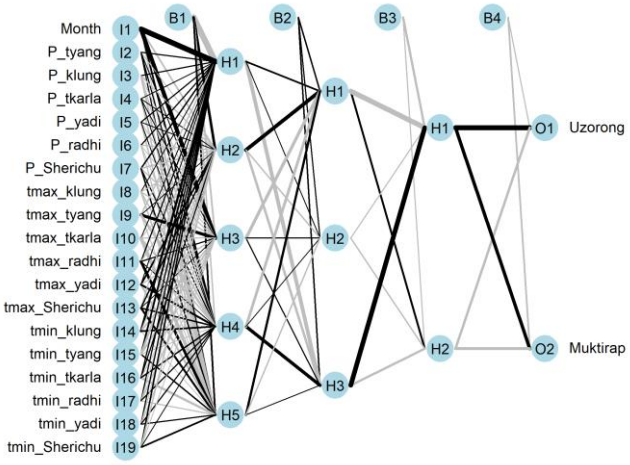

Fig 2: Neural Nework with three hidden layers

$$y = \frac{x - min(x)}{max(x) - min(x)} \quad (6)$$

Where y is the normalized data, x is the original data. Mathematically, neural network is expressed as:

$$y = f\sum_{i=1}^{N} w_i \times x_i + b \quad (7)$$

where y is the output vector and x_i is the input vector in the neural network, N is the number of neurons, w_i is the connection weight between input and output, f is the activation function, and b is the bias term.

Weights and bias are adjusted using the ANN's back-propagation algorithm, where the objective function (also known as loss function) is the error between the network's output and the observed output. The error is minimized using the optimization algorithm known as "Gradient descent" which minimizes the error value by taking steps from an initial guess until it reaches the best value. This make Gradient descent useful, when it is not possible to solve where the derivative of the objective function is equal to zero. The step size is usually calculated by providing the learning rate and is expressed as follows:

$$stepsize = slope \times learningrate \quad (8)$$

Where *slope = slope of objective function.*

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 01, January-2021

III. RESULTS AND DISCUSSION

A. *Imputing with MLRM, kNN and tree based model*

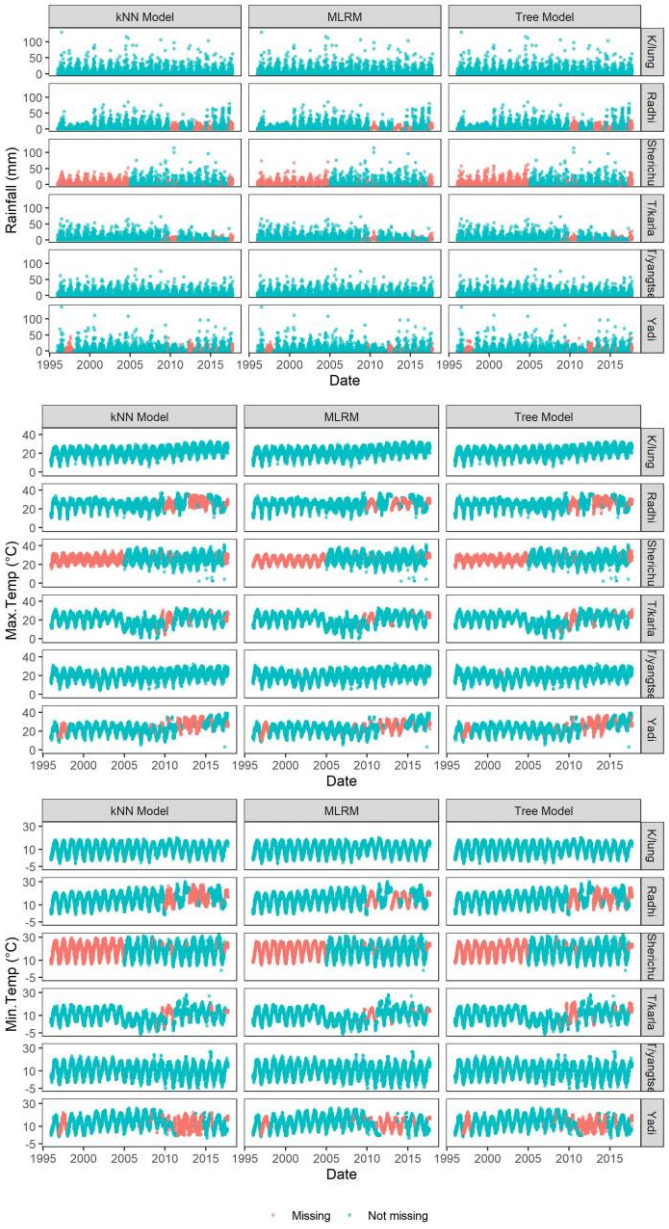

Fig 3: Imputation (a) Rainfall (b) Max. Temperature (c) Min.Temperature

Imputation of the Rainfall and Temperature were carried out using MLRM, kNN and Tree based model. The result of the imputation is as shown in Fig 2. Uncertainty from kNN imputation was performed using bootstrapping technique in which 1000 boot replicates were generated. The boot strap replicates here represent the regression coefficients. The bias and standard error were calculated based on the regression coefficients of replicates and the original imputed data, which are as shown in Table 2. Bias is the difference between the original regression coefficient and the replicate's regression coefficient while Standard error indicates the standard deviation of bootstrap replicates. From the result it is observed that bias and standard error is very minimum and has been accepted for further analysis. Further Root mean square error associated with Decision tree-based model are also shown in Table 3 where Sherichu indicates the maximum RMSE for all meteorological data. This is mainly because there were many missing values in Sherichu at the initial period as compared with the other station, while the RMSE for remaining station fairly remains below 3. Nevertheless, from the error, it can be implied that kNN imputation performed slightly well than that of Tree based model. The decision tree-based model can be further improved by increasing the number of decision tree used in the model however, with the increase in decision tree, the computation time also increases and eventually the user has to make speed- accuracy trade-offs. Finally, the imputation was carried out with MLRM (Fig 3) which clearly shows that the imputed data fits the variability of overall data.

Variable	Original	bias	Std error
Rainfall	0.347	0.0040 1	0.050169 26
Maximum Temperature	0.064	-0.01884	0.020895 82
Minimum Temperature	0.137	-0.00567	0.016366 58

Station	Rainfall	Max. Temp	Min.Temp
T/Yang	0	2.52	1.70
K/lung	0	0	0
T/karla	5.34	2.67	2.55
Yadi	5.2	2.55	2.22
Radhi	6.73	2.32	2.19
Sherichu	5.54	2.99	2.00

B. *Predicting flow using ANN*

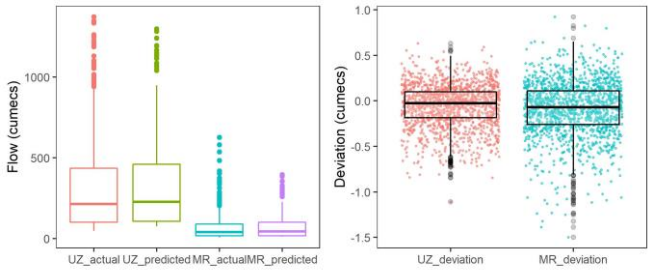

Fig 4: (a) Flow output (b) Flow deviation

Based on the imputed data from kNN model, ANN model was developed to predict the flow at Uzorong and Muktirap. The Fig 4 show the result of the ANN predictability based on the randomly selected testing data. The boxplot (Fig 4a) show the data variability in the actual and predicted data while Fig 4(b) show the deviation of prediction data from the actual data. From the result it is observed that absolute mean deviation is 0.054% and 0.088 % for Uzorong and Muktirap respectively, while the accuracy of the model was 94.45% and 91.11% respectively.

IV. CONCLUSION

Missing values in the hydro-meteorological data has always been found owing to defects in the sensors or maintenance of sensors. The missingness is also introduced due to relocation of station and error in observation which is usually treated as missing. Such missingness in the data inhibits the researchers in the field of hydrology and climate to draw inference from the data or sometimes leads to abstract inferences from the data. In this paper, missing values in six-meteorological station located in Eastern Bhutan has been imputed using several machine

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 01, January-2021

learning models such as kNN and Decision tree model. The data has also been imputed with statistical model using multiple linear regression model. The uncertainty in imputed data from the kNN model was performed with bootstrapping technique and the result showed that the bootstrap replicates follows normal distribution indicating usability of the imputed data. The tree-based model was assessed using the in-built model's OOB error which indicates minimum error associated with the imputation. The data was finally imputed with multiple linear regression model, where data was found to fairly represent the variability of overall data.

Based on the kNN imputed data as an input vector, ANN model was developed to predict the flow at Uzorong and Muktirap flow station. The model was developed considering backpropagation algorithm to calculate the weights and gradient descent optimisation algorithm to minimize the prediction error. The model was trained using the training data and subsequently tested on the testing data. Based on the testing data, the absolute mean deviation for the flow at Uzorong was 0.054% while for Muktirap was 0.088%. Accordingly, the model accuracy was 94.53% and 91.11% for Uzorong and Muktirap respectively. The model accuracy can be further improved by taking more training data which can consider the variability in the overall data.

REFERENCES

[1] K. Choden, J. Wangchuk, D. Yoezer, N. Wangdi, S. Wangchuk, and K. Tenzin, "Climate Change Vulnerability Assessment in Kurichhu Watershed: A case of Gangzur and Kengkhar, Bhutan.," UWICER Press, Lamai Goempa, Bumthang., 2018.

[2] NCHM, "Analysis of Historical Climate and Climate Projection for Bhutan," Royal Government of Bhutan, Thimphu, Bhutan, 2019.

[3] K. Adhikari, Y. Choden, T. Cheki, L. Gurung, T. Denka, and V. Gupta, "Performance evaluation of satellite precipitation estimation with ground monitoring stations over Southern Himalayas in Bhutan," Acta Geophys., 2020.

[4] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," Bull. Math. Biophys., vol. 5, pp. 115–133, 1943.

[5] J. Luengo, S. García, and F. Herrera, "On the choice of the best imputation methods for missing values considering three groups of classification methods," Knowl. Inf. Syst., vol. 32, pp. 77–108, 2012.

[6] M. A. Ben Aissia, F. Chebana, and T. B. M. J. Ouarda, "Multivariate missing data in hydrology – Review and applications," Adv. Water Resour., vol. 110, pp. 299–309, 2017.

[7] M. S. Osman, A. M. Abu-Mahfouz, and P. R. Page, "A Survey on Data Imputation Techniques: Water Distribution System as a Use Case," IEEE Access, vol. 6, pp. 63279–63291, 2018.

[8] M. . Mispan, N. F. A. Rahman, M. F. Ali, K. Khalid, M. H. A. Bakar, and S. H. Haron, "MISSING RIVER DISCHARGE DATA IMPUTATION APPROACH USING ARTIFICIAL NEURAL NETWORK," ARPN J. Eng. Appl. Sci., vol. 10, no. 22, pp. 10480–10485, 2015.

[9] T. R. Petty and P. Dhingra, "Streamflow Hydrology Estimate Using Machine Learning (SHEM)," J. Am. Water Resurces Assoc., vol. 54, no. 1, pp. 55–68, 2018.

[10] F. B. Hamzah, F. Mohdhamzah, S. F. Razali, O. Jaafar, and N. Abduljamil, "Imputation methods for recovering streamflow observation : A methodological review," Cogent Environ. Sci., vol. 6, no. 1, pp. 1–21, 2020.

[11] M. Kim, S. Baek, M. Ligaray, J. Pyo, M. Park, and K. H. Cho, "Comparative Studies of Different Imputation Methods for Recovering Streamflow Observation," Water, vol. 7, pp. 6847–6860, 2015.

[12] R. Tanty and T. S. Deshmukh, "Application of Artificial Neural Network in Hydrology- A Review," Int. J. Eng. Res. Technol., vol. 4, no. 06, pp. 184–188, 2015.

[13] M. T. Sattari, A. Rezazadeh-joudi, and A. Kusiak, "Assessment of different methods for estimation of missing data in precipitation studies," Hydrol. Res., vol. 48, no. 4, pp. 1032–1044, 2017.

[14] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, "Recurrent Neural Networks for Multivariate Time Series with Missing Values," Sci. Rep., vol. 8, pp. 1–12, 2018.

[15] A. Shkurin and E. Sarkola, "WATER QUALITY ANALYSIS USING MACHINE LEARNING," MAMK University of Applied Sciences, 2016.

[16] R. H. Ngouna, R. Ratolojanahary, K. Medjaher, F. Dauriac, M. Sebilo, and J. Junca-Bourié, "A data-driven method for detecting and diagnosing causes of water quality contamination in a dataset with a high rate of missing values," Eng. Appl. Artif. Intell., vol. 95, pp. 1–13, 2020.

[17] H. Yoon, S. Jun, Y. Hyun, G. Bae, and K. Lee, "A comparative study of artificial neural networks and support vector machines for predicting groundwater levels in a coastal aquifer," J. Hydrol., vol. 396, pp. 128–138, 2011.

[18] A. Ahani, M. Shourian, and P. Rahimi Rad, "Performance Assessment of the Linear, Nonlinear and Nonparametric Data Driven Models in River Flow Forecasting," Water Resour. Manag., vol. 32, pp. 383–399, 2018.

[19] J.-W. Kim and Y. A. Pachepsky, "Reconstructing missing daily precipitation data using regression trees and artificial neural networks for SWAT streamflow simulation," J. Hydrol., vol. 394, pp. 305–314, 2010.

[20] M. T. Sattari, K. Falsafian, A. Irvem, S. Shahab, and S. Qasem, "Potential of kernel and tree-based machine- learning models for estimating missing data of rainfall," Eng. Appl. Comput. Fluid Mech., vol. 14, no. 1, pp. 1078–1094, 2020.

[21] U. Mital, D. Dwivedi, J. B. Brown, and B. Faybishenko, "Sequential Imputation of Missing Spatio-Temporal Precipitation Data Using Random Forests," Front. Water, vol. 2, no. 20, pp. 1–14, 2020.

[22] M. Van der Loo, "Package ' simputation ' Simple Imputation. R package version 0.2.4.," 2020.

[23] A. Kowarik and M. Templ, "Imputation with the R package VIM," J. Stat. Softw., vol. 74, no. 7, pp. 1–16, 2016.

[24] D. J. Stekhoven and P. Bühlmann, "Missforest-Non-parametric missing value imputation for mixed-type data," Bioinformatics, pp. 1–13, 2011.

[25] S. Fritsch, F. Guenther, M. N. Wright, M. Suling, and S. M. Mueller, "Package 'neuralnet' Training of Neural Networks R package version 1.44.2.,"2019.