Reconstructing Missing Data in Hydro-Meteorology using Machine Learning: A Case in Himalayan Kingdom of Bhutan

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

Sanjana Pokhrel
Department of Civil Engineering
College of Science and Technology,
Royal University of Bhutan Phuntsholing, Bhutan

Ugyen Choden
Druk Green Power Corporation
Thimphu, Bhutan

Kirtan Adhikari
Department of Civil Engineering
College of Science and Technology,
Royal University of Bhutan Phuntsholing, Bhutan

***Abstract*—Missingness in the hydro-meteorological data is ubiquitous and it is inevitable owing to the frequent breakdown and maintenance of the meteorological sensor from time to time. It is well understood that 30 years of continuous weather data is necessary for the analysis of any significant trend and pattern. However, the missingness in the data creates a huge gap. Such missingness in the data inhibits researchers from concluding the correct statistical inference and thereby makes the whole data futile. Furthermore, the non-random nature of missingness possess a challenge for the researchers in selecting imputing models. This study presents the Machine Learning imputation techniques using Random Forest (RF). Available data from the different meteorological stations in Bhutan were collected and missing data were imputed using random forest employing the miss Forest technique. The imputation error in meteorological variables was assessed with Out-of-Bag (OOB) error. After missing data in the meteorological variable was imputed, they were considered as an independent variable with the flow as a response variable. A random forest model was created to predict the missing flow given the meteorological variables. The model was assessed using RMSE, R^2 and MAE.**

Keywords—Missing data; Imputation; Hydro-meteorology, Random forest, Machine Learning

I. INTRODUCTION

Hydro-meteorological variables viz. rainfall, temperature, and river discharge play an instrumental role in quantifying national and international water resource balance via a hydrological model. It also plays a significant role in augmenting regional and global climate models. Furthermore, flood forecasting models particularly depend on the real-time monitoring of these variables. Changes in the hydro-climatic variable concerning frequency and intensity are going to severely affect the environment and society, particularly affecting the climate-sensitive sections like agriculture, hydropower, and forest management. Rainfall and flow in Bhutan play a fundamental role in the sustenance of agriculture and hydropower and the knowledge of the hydro-met variables allows for proper decision-making and better preparedness. As the Himalayas ecosystem is highly susceptible to climate change and acts as a pivotal landmark, the impacts are likely to be observed first and therefore it is necessary to analyze and impute the missing values in the hydro-meteorological data to facilitate better analysis of these variables [1]

However, the development of hydrological, climatic, and flood forecasting model disparagingly depends on the quality of the data being observed. A long-term hydro-meteorological observation would tremendously help in developing such models however such observation is usually confronted with missing data. The missing data in hydro-meteorology is not uncommon and dealing with such data is a challenge for the hydrologist and meteorologist. Such missingness is prevalently existent owing to the breakdown and maintenance of meteorological and gauging sensors. Sometimes the missingness has to be intentionally created when the observed data is erratic (outlier) and does not follow the overall variability of data, for instance, recording of the temperature of 60 degrees during winter in a cold region is impossible based on the climate regime of the station. Such outlier is usually created in the data owing to errors in noting or observing. However, outlier such as flood extreme as a result of extreme rainfall has to be removed cautiously. With such missing data introduced in the variable, the long-term observation becomes futile as it cannot infer any meaningful statistical inferences. It will also be difficult to calibrate and validate hydrological and climate models. Due to the presence of missing data in such variables many researchers [2], [3] prefer to use data from global repositories such as Climate Research Unit (CRU), Aphrodite, and Tropical Rainfall Mission (TRMM), and ECMWF Climate Reanalysis. However, such data tend to have higher uncertainty because they are downscaled from Global Circulation models (GCM) to the regional level. Such data may not represent the actual regional climate or climate phenomena.

Deleting the missing data can be one of the first options to consider but can be only done when the missing data is relatively low (< 2%). In the case of many missing data, deletion of values from one variable may lead to ignoring the prominently observed values of some other dependent variables which result in the loss of useful data.

Imputing missing values disparagingly depends on the nature of missing data which is described as follows [4]–[6]:

i. MCAR: Missing Completely at Random, where missingness has no association with any data that is observed or not observed. In such a case imputation is advisable. Discarding the missing data will not bias the data however it will lead to a 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.

Machine Learning (ML) models have a humongous potential for imputing the missing data based on the observation of other observed values. Simple statistical models such as the Multiple Linear Regression Model (MLRM) also seem to be a viable option for imputing the data, as they can reproduce the overall variability of the original data however, it requires at least one response or independent variable which is free from missing data which is seldom the case[7]. ML models which were quintessentially developed for medical and neurological study [8], now find their practice in myriads of disciplines. However, the researcher across myriad disciplines could not leverage such techniques due to the limitation in the computation capacity of the computer then. With technological and computing advancements in recent decades, the application of ML models has increased manyfold in myriad disciplines. Machine Learning is a data-driven model, which learns even from complex and non-linear data patterns to aid in prediction, classification, and regression analysis.

Machine Learning model like kNN and Random Forest was well tested by [7] for imputing the missing data in a meteorological variable and found that they are a reliable method for imputation. Similarly, [9] also found satisfactory results from the ML model while imputing flow data. Although [10] have highlighted the challenges of the ML model in hydrology, several researchers [11]–[17] have advocated the use of ML models for estimating the missing flow data considering their reliable accuracy. ML models such as artificial neural network (ANN) can be used alternative to physical-based hydrological models to estimate the flow data and provides enormous applicability in hydrological studies [18]. Further [19] have also used deep learning models employing Long Short-Term Memory (LSTM) for time series prediction. Apart from hydro-met variables, the imputing can also be done for other environmental variables such as water quality variables[20] and groundwater level [21], and have found a reliable imputed value. Multiple imputations using integrated chain equation (MICE) has also been found better by employing the random forest as the imputing technique [22].

Since the hydro-meteorological variables are only indicators of understanding the regional weather and climatic phenomena, handling its missing data with the latest techniques allows for gathering a greater number of useful observations which could later be used to analyze the climate phenomena. It is well understood that a minimum of 30 years of data are required to analyze any climatic phenomena [23], therefore in this study attempt has been made to impute the missing data in hydro-meteorological variables using random forest and kNN and subsequently predict the missing flow data with ANN using the imputed meteorological variables as the dependent variables. Such a study has the potential to create the complete hydro-meteorological variable for the region without any missing values which could be used for various weather monitoring and climatological, hydrological, flood forecasting, agricultural and environmental research.

II. DATA AND METHODS

Daily rainfall and temperature data from Class A and Class C meteorological stations along with flow data was acquired from National Centre of Hydrology and Meteorology (NCHM), Thimphu, Bhutan. There are 20 class A station, 56 Class C station and 16 flow gauging station considered in this study. The location of these station is as shown in Fig 1.

The data from 1996–2020 has been considered in this study. Fig 2 shows the number of station and its associated range of percentage missing for both the rainfall and temperature data. There are 76 Class A and Class C rainfall station while there are only 70 Class A and Class C temperature station.

Out of 76 stations considered in the analysis, 46 station have missing data ranging from 0 to 20 percent while only 2 stations have 80 to 100 percent missing data.

Class A and Class C meteorological data was imputed using random forest and subsequently the missing flow data was predicted using ANN model where imputed rainfall and temperature data from Class A and Class C were used as independent variable.

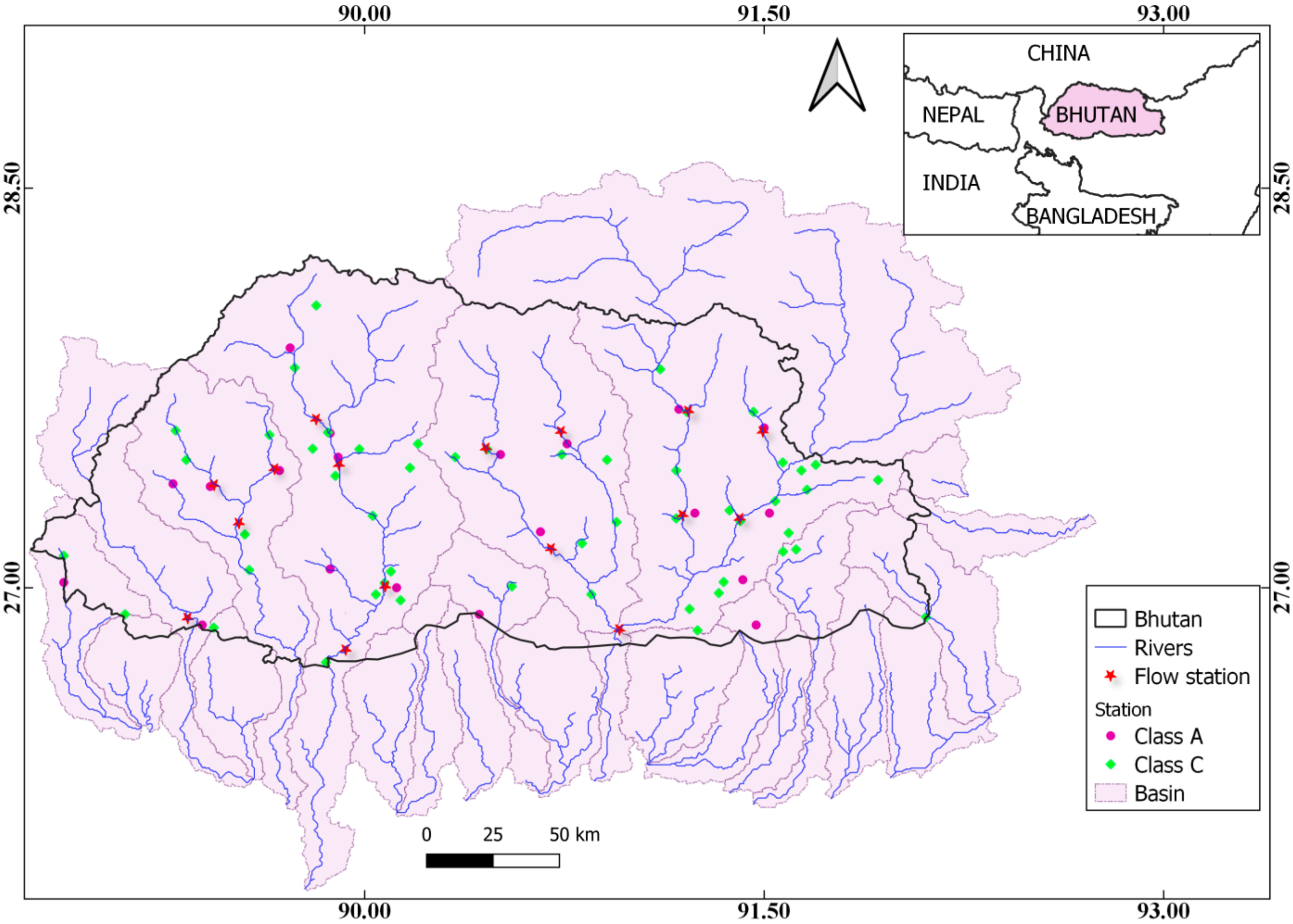

Fig 1: Location of Meteorological station

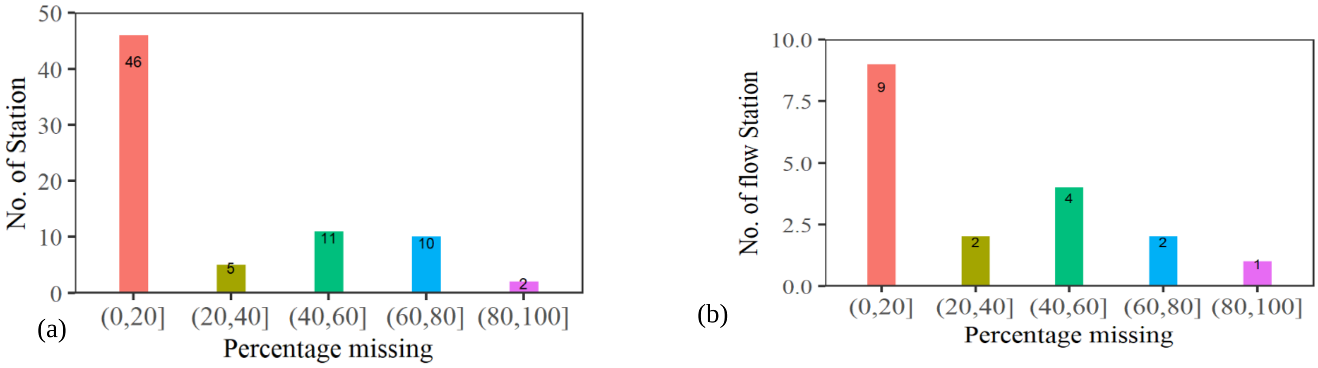

Fig 2 Histogram showing the number of station and percentage missing in (a) meteorological station

A. Imputation with random forest

Random forest (RF) is based on the non-parametric approach where no assumption is made on the relationship between variables. It can pick up complex nonlinear patterns and it is often better than the statistical models. Tree-based imputation uses a random forest behind the hood and builds separate random forests to predict the missing values for each variable one by one. RF considers two types of missing data, (1) missing data in the original dataset used to create the random forest and (2) missing data in the new sample that is to be categorized.

The general idea of dealing with the missing data in this context is to make an initial guess and gradually refine the guess until the error associated with the guessing is minimized. In this process, the missing values are firstly initialized with the median value. Now, these median imputed guesses are refined by first determining which observations are similar to the ones with the imputed data. This is determined by building a random forest and running the data for each tree (in this case there are 500 trees developed). From the first tree, the observations (including the median imputed) that fall on the same leaf node are considered to be similar. The similarity is tracked using a numeric value in the proximity matrix. Then these proximity values are divided by the number of decision trees considered. Now the proximity values of the imputed data to make a better guess of the missing values.

In this study, it utilizes the Miss Forest imputation algorithm in R- environment developed by [24], wherein the first iteration, missing data is initially imputed with the 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. Interestingly, the algorithm also calculates the Out-of-Bag (OOB) error associated with the imputation and hence there is no need to evaluate its efficiency separately. The OOB are those samples that were left out from the bootstrap data that were randomly used in each tree. Ultimately, the accuracy of the random forest is measured by the proportion of OOB samples that were correctly classified by the random forest. The proportion of OOB samples that are incorrectly classified is known as OOB error. Here the error has been minimized by taking 500 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.

B. Imputing flow data with a random forest model

As opposed to using the Miss Forest technique to impute the missing values in meteorological variables, this method uses a regular random forest technique to predict the flow values based on a set of independent variables. The imputed meteorological variables have been used as the independent variable in predicting the missing flow values. Firstly, the data was normalized using a min-max normalization technique (Eqn 1) where all the data ranged from 0 to 1. Subsequently, the data was organized into training and testing data with 75 % of randomly observed data as training and the remaining 25 % of randomly observed data as the testing data.

$$y = \frac{x - \min(x)}{\max(x) - \min(x)} \qquad (1)$$

Where y is the normalized data, and x is the original data. The training and testing data are selected in such a way that there are no missing values. For a particular flow gauging station, the independent variable is selected considering their presence in the same basin as the flow gauging station because it is well understood that flow is more influenced by rainfall and temperature data located within the basin.

The random forest model is built based on the training data using the caret package in R which is developed by [25]. Further, the model has been enhanced by taking 10-fold cross-validation with parameter, mtry (no of predictors) ranging from one to fifty. With such assessments, the parameter corresponding least error is identified. Such cross validation makes the model better than single split train and test data. Once the model was developed, the model was tested on the testing data and the errors were assessed. After assessing the errors from the testing data, the model was applied to predict the missing flow data

III. RESULTS AND DISCUSSION

The missing data on rainfall and temperature have been imputed using the Miss Forest technique which employs a random forest technique to impute the missing data. To improve the accuracy of the model, 500 decision tree was considered for each imputation. From the imputation model, it was observed that the model efficiently imputes the missing data with little to no variation in before and after imputation. This is evident in Fig 3 and Fig 4 which show the imputed and observed data in a time series. The imputation error was assessed using OOB error and it is observed that for temperature the error ranged from 0 – 5 while for the rainfall data the error ranged from 0 – 50. Such difference in error between temperature and rainfall data is particularly because of a large variation in rainfall data. The rainfall is usually highly variable and further, it contains many outliers which occur as a result of occasional intense rainfall. Owing to such phenomena, the imputation error for rainfall is relatively higher than that of temperature. Nevertheless, such imputation error can be reduced using a larger decision tree it will which consumes higher processing time, however, there are always speed-accuracy trade-offs to be made while running such machine learning techniques.

The imputation error was also assessed with different proportions of missingness in the data, and it is observed that the proportion of missing data did not have a serious impact on the imputation as is evident from Fig 4.

After imputation of meteorological variables and assessing the associated errors, a separate random forest model was created to predict the missing data in the flow data. The meteorological variable was considered as an independent variable while the flow was treated as the response variable. The errors were assessed using Root Mean Square error (RMSE) and Mean Absolute Error (MAE) while accuracy was assessed using R-square (R^2). The errors were assessed for both the training and testing data which is as shown in Table 1. For both the training and testing data, the RMSE ranged from 0.017–0.083, R-square ranged from 0.5–0.82 and MAE ranged from 0.01–0.06. It is interesting to observe that for both the training and testing data, the model performed quite well as there is no overfitting of the model as metrics for both the training and testing data were found to be similar. Using this model, the missing data for flow was predicted and imputation can be observed in Fig 6. Therefore, the random forest can be a highly sought technique to predict the missing data.

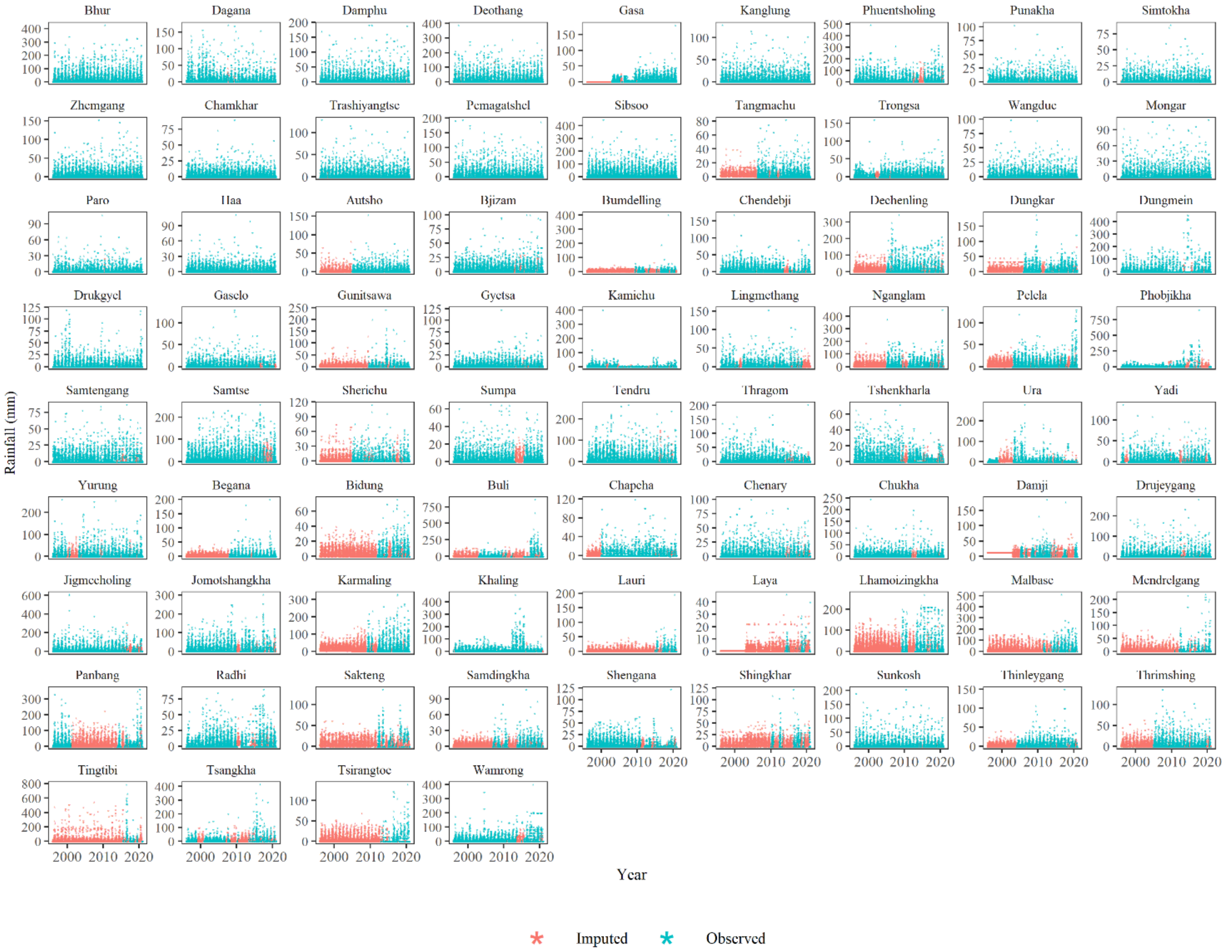

Fig 3 Class A and Class C imputed rainfall

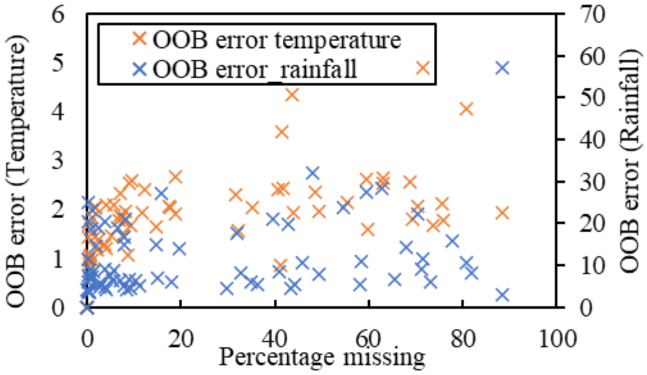

Fig 4 OOB error for a range of proportion of missingness in meteorological data

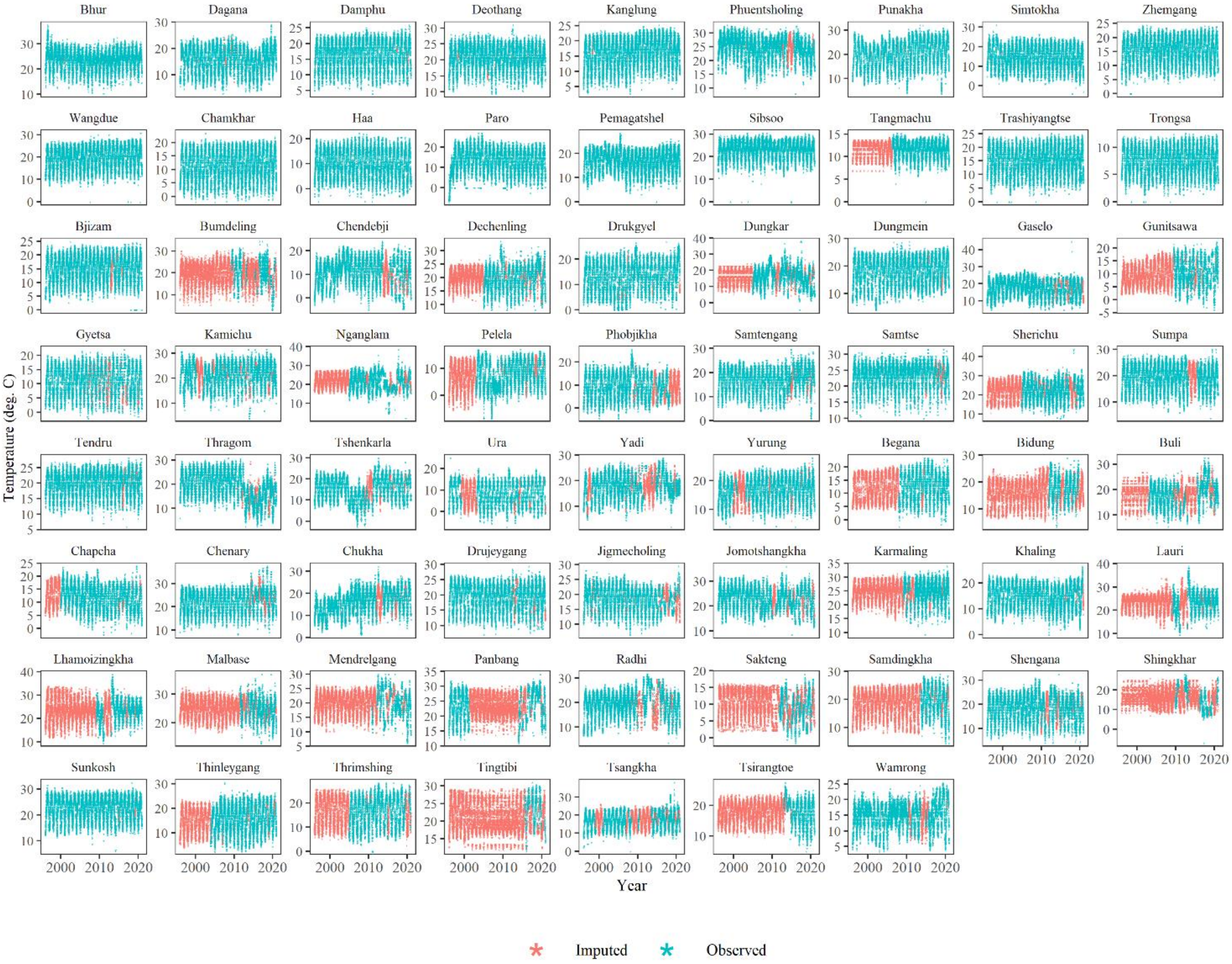

Fig 5 Class A and Class C imputed temperature

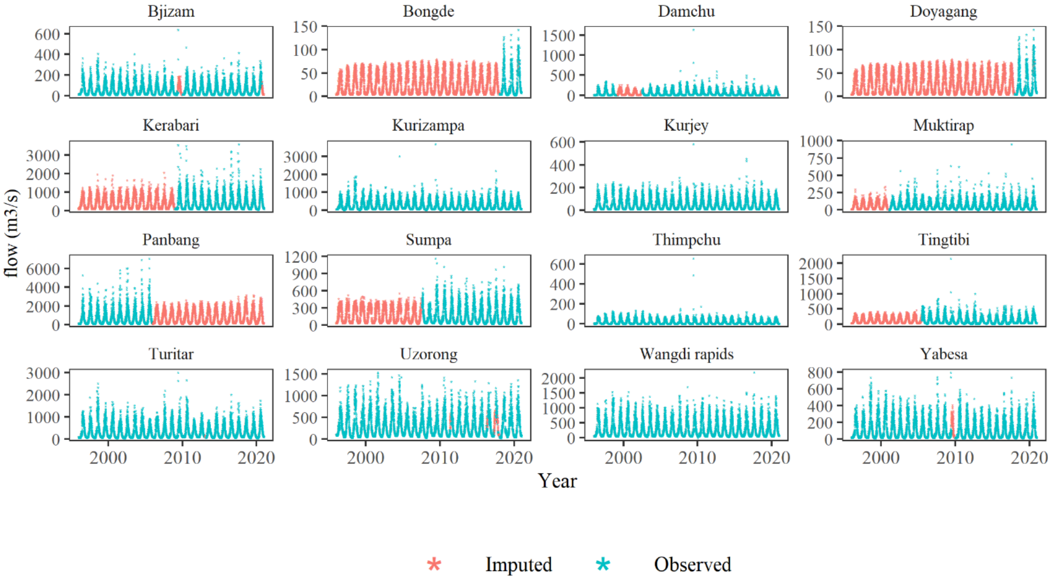

Fig 4 Imputation of flow data

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 08, August-2022

Table 1 Model metrics for training and testing data

Flow Station	mtry	RMSE		R-square		MAE	
		Training data	Testing data	Training data	Testing data	Training data	Testing data
Doyagang	3	0.041	0.049	0.570	0.500	0.024	0.025
Bongde	15	0.095	0.107	0.765	0.750	0.050	0.060
Damchu	14	0.024	0.022	0.630	0.656	0.013	0.01
Thimpchu	6	0.017	0.029	0.731	0.535	0.011	0.012
Kerabari	19	0.055	0.052	0.796	0.807	0.032	0.030
Turitar	20	0.051	0.055	0.790	0.785	0.031	0.032
Wangdi Rapids	12	0.057	0.057	0.805	0.800	0.034	0.033
Yabesa	17	0.066	0.067	0.797	0.802	0.039	0.038
Kurjey	38	0.037	0.042	0.816	0.784	0.022	0.023
Panbang	9	0.067	0.061	0.691	0.713	0.035	0.032
Bjizam	34	0.043	0.041	0.800	0.819	0.024	0.023
Tingtibi	32	0.031	0.029	0.735	0.738	0.017	0.017
Sumpa	4	0.0834	0.083	0.667	0.653	0.051	0.051
Kurizampa	3	0.044	0.045	0.646	0.641	0.026	0.026
Muktirap	20	0.034	0.041	0.734	0.692	0.017	0.018
Uzorong	27	0.072	0.074	0.820	0.816	0.043	0.044

IV. CONCLUSION

Missingness in the hydro-meteorological variable is ubiquitous and it poses a humongous challenge for the researchers in the field of hydrology, hydraulics, and environment to handle and deal with such data. Although there are many techniques such as linear and mean models to impute the missing data, such techniques may not represent the original variability of the data. This study attempts to recreate the missing data in hydro-meteorological variables monitored in Bhutan using the random forest technique. Missing data in 76 rainfall stations and 70 temperature stations were imputed using the miss forest technique. Further, the imputed meteorological variable was used as the independent variable for predicting the flow data using random forest. The error in the imputation of the meteorological variable was minimized by taking 500 decision trees. It was observed that the imputation error for rainfall data was relatively higher than that of temperature data. The imputation error (OOB error) for rainfall ranged from 0 to 50 while for temperature the error ranged from 0 to 5. Such a difference in error between rainfall and temperature is basically because of variance in the data. The rainfall data usually tend to have higher variance due to the presence of outliers which is basically due to intense rainfall events while the temperature data do not have as many outliers as compared with rainfall data. Therefore, if data has a higher variance, the OOB error is also quite higher. Nevertheless, miss forests are a better imputation model than other prominently used models such as linear models and mean models. It is one of the techniques to recreate the past missing data which can be used for statistical analysis. After the meteorological variable was imputed, a separate random forest model was created to predict the flow data using imputed meteorological data as the independent variable. The data then was normalized using the min-max normalization technique and subsequently, the normalized data were divided into training (75%) and testing (25%) data. The error was assessed for both the training and testing data and it was found that the model performed well because there was no overfitting of the data and the metrics for both the training and testing data were found similar. The RMSE ranged from 0.017–0.083, R-square ranged from 0.5–0.82 and MAE ranged from 0.01–0.06. Since the errors are relatively less and the model was built on a meteorological variable as the independent variable, the model result can be acceptable for predicting the flow data, because the flow is largely influenced by the meteorological variable. The model can also be used for understanding the future flow regime of the river given the different meteorological and weather scenarios from climate data. To further improve the predictive rate of the model, different machine learning such as Artificial Neural networks and Deep Learning can be employed although such techniques consume higher computation time.

REFERENCES

[1] V. Sharma and K. Adhikari, “Rainfall and rainy days trend and ENSO phenomena in Himalayan Kingdom of Bhutan,” *Acta Geophys.*, Jun. 2022, doi: 10.1007/S11600-022-00839-Y.

[2] NCHM, “Analysis of Historical Climate and Climate Projection for Bhutan,” National Center for Hydrology and Meteorology Royal Government of Bhutan PO Box: 2017 Thimphu, 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, doi: 10.1007/s11600-020-00434-z.

[4] 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, doi: 10.1007/s10115-011-0424-2.

[5] 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, doi: 10.1016/j.advwatres.2017.10.002.

[6] 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, doi: 10.1109/ACCESS.2018.2877269.

[7] V. Sharma and K. Yuden, “Imputing Missing Data in Hydrology using Machine Learning Models,” *Int. J. Eng. Res. Technol.*, vol. 10, no. 01, pp. 78–82, Jan. 2021, doi: 10.17577/IJERTV10IS010011.

[8] 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.

[9] P. Arriagada, B. Karelovic, and O. Link, “Automatic gap-filling of daily streamflow time series in data-scarce regions using a machine

learning algorithm," *J. Hydrol.*, vol. 598, p. 126454, 2021, doi: https://doi.org/10.1016/j.jhydrol.2021.126454.

[10] A. Gharib and E. G. R. Davies, "A workflow to address pitfalls and challenges in applying machine learning models to hydrology," *Adv. Water Resour.*, vol. 152, p. 103920, 2021, doi: https://doi.org/10.1016/j.advwatres.2021.103920.

[11] 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.

[12] 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, doi: 10.1111/1752-1688.12555.

[13] 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, doi: 10.1080/23311843.2020.1745133.

[14] F. B. Hamzah, F. M. Hamzah, S. Fatin, M. Razali, and H. Samad, "A Comparison of Multiple Imputation Methods for Recovering Missing Data in Hydrological Studies," *Civ. Eng. J.*, vol. 7, no. 09, pp. 1608–1619, 2021.

[15] D. Heras and C. Matovelle, "Machine-learning methods for hydrological imputation data: analysis of the goodness of fit of the model in hydrographic systems of the Pacific-Ecuador," *Rev. Ambient. Água*, vol. 16, 2021.

[16] M. Zounemat-Kermani, O. Batelaan, M. Fadaee, and R. Hinkelmann, "Ensemble machine learning paradigms in hydrology: A review," *J. Hydrol.*, vol. 598, p. 126266, 2021, doi: https://doi.org/10.1016/j.jhydrol.2021.126266.

[17] H. Mosaffa, M. Sadeghi, I. Mallakpour, M. Naghdyzadegan Jahromi, and H. R. Pourghasemi, "Chapter 43 - Application of machine learning algorithms in hydrology," in *Computers in Earth and Environmental Sciences*, H. R. Pourghasemi, Ed. Elsevier, 2022, pp. 585–591. doi: https://doi.org/10.1016/B978-0-323-89861-4.00027-0.

[18] E. Rozos, P. Dimitriadis, and V. Bellos, "Machine Learning in Assessing the Performance of Hydrological Models," *Hydrology*, vol. 9, no. 1, 2022, doi: 10.3390/hydrology9010005.

[19] C. Shen and K. Lawson, "Applications of Deep Learning in Hydrology," in *Deep Learning for the Earth Sciences: A Comprehensive Approach to Remote Sensing, Climate Science, and Geosciences*, John Wiley & Sons Ltd, 2021, pp. 285–297. doi: 10.1002/9781119646181.ch19.

[20] R. Rodriguez Núñez, M. Pastorini, L. Etcheverry, C. Chreties, M. Fossati, A. Castro, and A. Gorgoglione, "Water-quality data imputation with a high percentage of missing values: A machine learning approach," *Sustainability*, vol. 13, no. 11, pp. 11–7, 2021, doi: 10.3390/su13116318.

[21] L. Kulanuwat, C. Chantrapornchai, M. Maleewong, P. Wongchaisuwat, S. Wimala, K. Sarinnapakorn, and S. Boonya-aroonnet, "Anomaly Detection Using a Sliding Window Technique and Data Imputation with Machine Learning for Hydrological Time Series," *Water*, vol. 13, no. 13, 2021, doi: 10.3390/w13131862.

[22] X. Jing, J. Luo, J. Wang, G. Zuo, and N. Wei, "A Multi-imputation Method to Deal With Hydro-Meteorological Missing Values by Integrating Chain Equations and Random Forest," *Water Resour. Manag.*, vol. 36, no. 4, pp. 1159–1173, 2022, doi: 10.1007/s11269-021-03037-5.

[23] WMO, *Guide to Climatological Practices*, 2018 editi., no. WMO-No. 100. Geneva, Switzerland: World Meteorological Organization, 2018.

[24] D. J. Stekhoven and P. Bühlmann, "Missforest-Non-parametric missing value imputation for mixed-type data," *Bioinformatics*, pp. 1–13, 2011, doi: 10.1093/bioinformatics/btr597.

[25] M. Kuhn, "caret: Classification and Regression Training." 2021. [Online]. Available: https://cran.r-project.org/package=caret