Natural Hazards (2024) 120:5075–5098
https://doi.org/10.1007/s11069-024-06406-6

ORIGINAL PAPER

Estimation of missing weather variables using different data mining techniques for avalanche forecasting

Prabhjot Kaur[1] · Jagdish Chandra Joshi[2] · Preeti Aggarwal[1]

Received: 10 January 2023 / Accepted: 3 January 2024 / Published online: 2 February 2024

Abstract
The availability of continuous weather data is essential in many applications such as the study of hydrology, glaciology, and modelling of extreme catastrophic events such as landslides, heavy precipitation, cloud burst and snow avalanches. Weather data are collected either manually or automatically, and due to variety of reasons, it becomes difficult to maintain continuous records of these data. In the present study, different data mining techniques like multivariate imputation by chained equations and nearest neighbour have been used to address the missing data problem for avalanche forecasting over the Himalayas. Six weather variables, maximum temperature, minimum temperature, wind speed, pressure, fresh snow and relative humidity used in all avalanche and weather forecasting models, have been made available from 1996 to 2019. Missing data are generated randomly to create 10, 15, 20 and 30% in order to study the algorithms. Scatter plots, root-mean-square error and coefficient of determination (R^2) of the generated missing data have been computed. Case analysis of imputed major snow events is done from 2017 to 2019, demonstrating proficient imputation. The performance of artificial neural network-based avalanche forecasting models has been compared with and without data imputation. Results of the study are promising as HSS and accuracy for avalanche forecasting models accelerates to 0.36 from 0.31 and 0.74 from 0.71, respectively, for Station-1 and HSS to 0.3 from 0.24 and accuracy to 0.72 from 0.68 for Station-2 after missing data imputation.

Keywords Data imputation · Multivariate Imputation by Chained Equations method (MICE) · Nearest neighbour (NN)

1 Introduction

Most meteorological studies involve analysis of field data, which comprises of inevitable gaps in recorded climate data especially at high elevations (Kanda et al. 2018). Existence of gaps in the records of data acquisition systems are often attributed to various reasons such as absence of the observer, instrumental failures and communication line breakdown.

✉ Prabhjot Kaur
pavibti@gmail.com

1 University Institute of Engineering and Technology, Panjab University, Chandigarh, India

2 Armament Research and Development Establishment, Pune, India

(Kashani and Dinpashoh 2012). Although some research can be carried out with incomplete data series, yet the significance of complete data series cannot be overlooked (Firat et al. 2012). A well-known approach to deal with the missing data problems is complete case analysis (CCA), which omits subjects with missing values from the analysis. It is a simple solution to ignore the observation with missing values and no significant problem occurs when there are very few observations with missing values. However, deleting a large number of observations with missing values causes a significant loss of information (Zhang 2015). In some cases, where missing ratio is high, CCA leads to inefficient analysis due to information loss causing biased inferences about the parameters of interest (Sterne et al. 2009). It also decreases the statistical power and efficiency of the data (Kwak and Kim 2017). And quality of data plays very critical role for model building in machine learning. For modellers who work on numerical weather forecast, complete historical series of meteorological data are important for the initialization, training and verification of the models (Carvalho et al. 2017). Therefore, estimation of missing data is the first and significant phase in most climatological, environmental and hydrological studies and any procedure which is effective to deal with this problem plays a vital role in such studies (Tabony 1983). Also in the field of geosciences, one of the most pressing concerns is the ongoing issue of climate change. To address this challenge, it is crucial to develop effective models aimed at minimizing loss of life and property. Complete weather data play a vital role in this effort, as does the use of accurate techniques for data imputation. Not only for meteorological variable, imputation has also gain importance in various other fields like medical data, etc. as addresses in research works by Orczyk and Porwik (2013), Chhabra et al. (2017), Javadi et al. (2021), KA et al. (2022), etc.

Different researchers have used different data-driven models listed in Table 1 for missing data estimation of meteorological variable worldwide. Models performance varies from one geographical location to another. Costa et al. (2021) highlight the potential of the MICE technique to fill gaps in daily data from time series of meteorological variables. According to past studies, using multiple imputations instead of single imputations for missing data estimation takes into account the statistical uncertainty involved in the process. The chained equations method is also highly adaptable and can handle various types of variables (continuous or binary) and complexities such as bounds or skipped patterns. Alruhaymi and Kim (2021) has stated MICE is the best approach to dealing with missing at random (MAR) missingness. Other than missing data estimations in weather datasets Khan and Hoque (2020) developed SICE (Single Center Imputation from Multiple Chained Equation) extended version of MICE in which they used mean value for numerical data and mode value for categorical data set instead of basic MICE techniques on three open datasets. Khan and Hoque stated that SICE had 20% higher F-measure for binary data imputation and 11% less error for numeric data imputations than MICE with same execution time. Kim and Pachepsky (2010) stated that better accuracy was accomplished with the combined regression tree and ANN rather than using them independently. Kim et al. (2019) stated *k*-nearest neighbour (*k*NN) provided the most appropriate missing data imputation for weather data used in PV forecasting in Korea. The *k*NN imputation is based on machine learning, which has been extensively used for classification, regression, and imputation (Batista and Monard 2002). Also in *k*NN, new data can be added seamlessly. Inverse distance weighing (IDW) is also used a lot in literature which works on the same principle as *k*NN. Other methods like regularized EM algorithm (Schneider 2001), the Fourier fit, the EM-Markov chain Monte Carlo (Yozgatligil et al. 2013) and the Bayesian network (Lara-Estrada et al. 2018) not listed in Table 1 are also used by the researchers for imputing missing observations on daily and monthly precipitation, temperature and humidity data.

Table 1 Literature Survey on different data-driven models for weather data imputation

Authors	Area	Variables	Techniques
Costa et al. (2021)	Brazil	Precipitation, temperature, relative humidity, atmospheric pressure, wind speed and insolation	Multivariate imputation by chain equations (MICE)
Afrifa-Yamoah et al. (2020)	Australia	Temperature, humidity, wind speed	Autoregressive integrated moving average (ARIMA) model with Kalman smoothing and multiple linear regression
Kim et al. (2019)	Korea	All weather variables	Linear interpolation (LI), mode imputation (MI), *k*-nearest neighbours (*k*NN), and multivariate imputation by chain equations (MICE)
Kajewska-Szkudlarek and Stańczyk (2018)	Wroclaw, Poland	Average, maximum and minimum air temperature, relative air humidity	Artificial Neural Networks: MLP (Multi-Layer Perceptron) and RBF (Radial Basis Function), which differ in terms of modelling the input–output relation and SVR (Support Vector Regression)
Kanda et al. (2018)	Karakorum range of Indian Himalayas	Temperature	Simple arithmetic averaging, inverse distance weighing (IDW) interpolation, normal ratio method (NR), single best estimator (SBE), multiple regression using the least absolute deviation criterion (MLAD), closest station method (CSM) and UK traditional method (UKT)
Sattari et al. (2017)	Iran	Precipitation	Arithmetic averaging (AA), inverse distance interpolation, linear regression (LR), multiple imputations (MI), multiple linear regression analysis (MLR), non-linear iterative partial least squares (NIPALS) algorithm, normal ratio (NR), single best estimator (SIB), UK traditional (UK) and M5 model tree
Aprianti and Mukhlash (2015)	Indonesia	Rainfall	Rough set algorithm
Che Ghani et al. (2014)	Malaysia	Rainfall	Gene expression programming (GEP)
Choge and Regulwar (2013)	India	Precipitation	Artificial Neural Networks

Table 1 (continued)

Authors	Area	Variables	Techniques
Kashani and Dinpashoh (2012)	Iran	Precipitation	Normal Ratio, Inverse Distance, UK traditional method, multiple linear regression, Multiple Imputations, and three modern data mining methods including Multi-layer perceptron, Support vector regression and *k*-Nearest neighbours
Kim and Pachepsky (2010)	USA	Precipitation	Regression tree and ANN
Dastorani et al. (2010)	Iran		Normal ratio method, correlation method, an artificial neural network (ANN), and an adaptive neuro-fuzzy inference system (ANFIS)
Teegavarapu et al. (2009)	Eastern part of Kentucky	Precipitation	Genetic algorithm and distance weighting method
Kotsiantis et al. (2006)	Greece	Temperature	Model Trees and Rules
Tabony (1983)	UK	Temperature	Principal component analysis (PCA)

Ongoing climate change and complex interactions between snow and meteorological features are resulting in frequent avalanches in the snow bound region of the Indian Himalayas leading to massive destruction of property and life. Model building in machine learning for avalanche forecasting demands good quality data which is improbable during peak winters or in the case of any extreme event in snow bound areas of the Indian Himalayas due to harsh weather conditions, avalanches and topographical influences. Hardly any efforts are made for missing data estimation in this region. As literature suggests MICE and *k*NN being one of the powerful tools in missing data estimations not only for meteorological variables but in other application such as medical studies to incorporate better knowledge in the model estimation for weather variables. Hence, the objective of this study is to assess the effectiveness of MICE and *k*NN estimation techniques in analysing meteorological data from snowy mountainous regions in the Indian Himalayas. Additionally, this study aims to compare the performance of machine learning models with and without data imputation over the study area. The imputation methods are evaluated with the help of RMSE, standard deviations, scatter plots, coefficient of determination, Taylor diagram and avalanche prediction model using probability of detection (POD), Heidke skill score (HSS), false alarm rate (FAR), bias and accuracy.

2 Study area and data

The Indian Himalayan Region classified into Karakoram Range, Great Himalayan Range and Pir Panjal Range stretches across a length of 2500 km and width of 250–300 km and receives moderate to heavy snowfall during winter (Nov–Apr) due to western disturbance. Indian Himalayas experiences wide diversity in climatic and precipitation patterns (Sharma 2000). The Defence Geo-Informatics Research Establishment (DGRE), India, has an observational network of manual observatories all over the Indian Himalayas collecting weather/meteorological data (temperature, wind speed, pressure, etc.) daily at 08:30 and 17:30 Indian Standard Time (IST) (0300 and 1200 UTC/GMT). Station-1, an observatory of DGRE in J&K, India, is situated at an altitude of 2650 m in Pir Panjal Mountain ranges of Lower Himalayan. Station-2 an observatory of DGRE in Higher Himalayas or Great Himalaya Range, situated at an altitude of 3300 m in Ladakh, India. Figure 1 depicts the study area and the location of the stations in the Indian Himalaya. Table 2 elaborates on the meteorological and geographical differences in the study areas. Though avalanche occurrences are more in Station 1 but type of avalanche and intensity of avalanches are hazardous for Station 2.

In this study, meteorological data of Station-1 in Pir Panjal range and Station-2 in Great Himalayan range are analysed from December, 1992–March, 2019 having 6544 and 6545 tuples, respectively. Gaps in the meteorological variables vary from nil to more than 50%. The climatic variables analysed in this study are used as a principal source of inference in building models to impute missing data. Complete case analysis (CCA) is done by case wise deletion of observations that has a missing value for any variable and only complete observations are analysed. Data at the target stations were assumed to be missing for the purpose of estimation for testing various methods. Therefore, 10% (2017–2019), 15 and 20% of the CCA data were considered missing randomly for testing methods, and the remaining 90, 85 and 80% of the data were used to develop simulation network for imputation. Data variables and tuples corrupted more than 50% are omitted from the database as they can lead to a biased result in data imputation (Madley-Dowd et al. 2019).

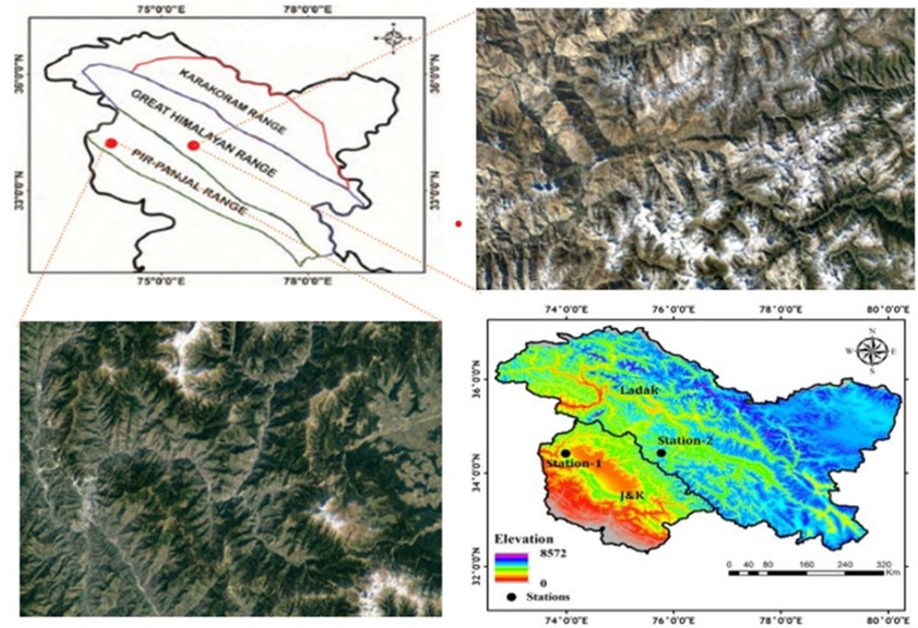


Fig. 1 Study area

Table 2 Meteorological and geographical study of Station 1 and Station 2 (Sharma 2000)

Factors	Lower Himalayan Zone (Station 1)	Upper Himalayan Zone (Station 2)
Terrain/geographical factor		
Altitude	3200–4100 m (76%)	5000–5600 m (100%)
Slope	30–38 (64%)	28–32 (67%)
Ground	Tall grassy cover	Rocky, scree and glacial
Meteorological factors		
Snowfall in a storm	20–80 cm (56%)	10–20 cm (51%)
Average total yearly snowfall	15–18 m	7–8 m
Temperature (°C)		
Highest max	20.2	9.0
Mean max	6.8	− 8.1
Mean min	− 1.6	− 27.7
Lowest min	− 12	− 41

3 Methodology

3.1 Types of missing mechanism

Mechanism of missing data is related to three terms: Missing at Random (MAR), Missing Completely at Random (MCAR) and Missing Not at Random (MNAR). Adopting generic notation, where Y_{com} as complete data and partition in (Y_{obs}, Y_{mis}), where Y_{obs} and Y_{mis} are the observed and missing parts, respectively. Rubin (1976) defined missing data to be MAR if the distribution of missingness does not depend on Y_{mis}. In other words, MAR allows the probabilities of missingness to depend on observed data but not on missing data. In MAR, there exists systematic relationship exists between one or more measured variables and the probability of missing data. This represents an important practical problem for missing data analysis because maximum likelihood estimation and multiple imputation assume an MAR mechanism. Whereas, there is no practical MAR mechanism to confirm that the probability of missing data on Y_{mis} is solely a function of other measured variables (Enders 2010). However, Gómez-Carracedo et al. (2014) stated if data are lost because of a system shutdown, faults in power supply, etc. but not because of the values themselves it can be accepted that missing data have a MAR structure. A special case of MAR, is missing completely at random (MCAR), occurs when the distribution does not depend on Y_{obs} either. The probability of missing data on a variable Y_{mis} is unrelated to other measured variables and is unrelated to the values of Y_{mis} itself. MCAR is a special condition of MAR as it is more restrictive condition than MAR because it assumes that missingness is completely unrelated to the data. When probability of Y_{mis} depends on Y_{mis} and Y_{obs}, the missing data are said to be missing not at random (MNAR). Like the MAR mechanism, there is no way to verify that data are MNAR. MAR is called ignorable nonresponse whereas MNAR is called non-ignorable nonresponse (Alruhaymi and Kim 2021).

3.1.1 Consequences of MCAR, MAR, and MNAR

The main consequence of MCAR is loss of statistical power. The good thing about MCAR is that analyses yield unbiased parameter estimates (i.e., estimates that are close to population values). MAR (i.e., when the cause of missingness is taken into account) also yields unbiased parameter estimates. The reason MNAR is considered a problem is that it produces biased parameter estimates (Alruhaymi and Kim 2021; Enders 2010).

3.2 Different data imputation techniques

3.2.1 Simple imputation

Single imputation techniques generate a specific value for a missing real value in a dataset. This technique has less computational cost. There are many single imputation methods proposed by the researchers. The imputation can be obtained by measures such as mean, median, mode of the available values of that variable. Other approaches, such as machine learning-based techniques like ANN, KNN, SVM are also used in single imputation (Khan and Hoque 2020). But filling all the missing values using only single imputation may not correctly address the uncertainty of the dataset and likely to produce bias imputation (Khan and Hoque 2020).

3.2.2 Multiple imputation

Single imputation of values obtained by the regression models fails to proper variability and there exists uncertainty of the imputed records that is not communicated to the analysis stage, which can be achieved by multiple imputation (Pickles 2005). Multiple imputation methods introduced by Rubin (1987) in which multiple values were simulated for the imputation of a single missing value using different simulation models. Multiple imputation methods are complex in nature, but they do not suffer from biasness like single imputation. In multiple imputation, each missing data is replaced with m values obtained from m iterations (where $m > 1$ and m normally lies between 3 and 10). By imputing multiple times, multiple imputation accounts for the uncertainty and range of values that the true value could have taken. Multiple imputation reduces bias, improve validity, increases precision and results in robust statistics. One of the popular approach is Multivariate Imputation by Chained Equations (MICE). MICE algorithm, proposed by V. S. Buuren and K. Groothuis-Oudshoorn, is widely used for multiple imputation. MICE is simulated using Predictive Mean Matching, Multiple Random Forest Regression Imputation, Multiple Bayesian Regression Imputation, Multiple Linear Regression using Non-Bayesian Imputation, Multiple Classification and Regression Tree (CART), Multiple Linear Regression with Bootstrap Imputation, etc. Markov chain Monte Carlo (MCMC) is another method for multiple imputation.

In the study, data imputation is done with help of nearest neighbour (simple imputation) where complete past data are analysed and multiple imputation by MICE where different imputed dataset are simulated and best among all is selected for imputation. The framework of proposed research is conducted in several steps as illustrated in Fig. 2 and imputation models used are discussed below. Performance analysis of the imputed and non-imputed datasets for avalanche forecasting is done with the artificial neural neuron network using sklearn in python.

3.3 *k*NN: *k*-nearest neighbour

*k*NN is a simple imputation technique with efficient statistical methods and machine learning technique having applications in different scenarios such as regression, classification or imputation. *k*NN is considered lazy, instance-based learning algorithm and among top 10 data mining algorithms (Wu et al. 2008). *k*NN as imputer can easily handle and predict both quantitative features and qualitative features. The major drawback of *k*NN as imputer is when the algorithm searches through all the dataset making it very critical for large databases but a robust procedure at the same time for missing data estimation. To apply *k*NN for missing data imputation, one of the important step is to select an appropriate distance metric. Uniform or inverse distance weighing are commonly used in KNN for simulations. Equations 1 and 2 elaborated both the techniques in detail.

$$\text{Feature set } = \{x_1, x_2, x_3 \ldots x_n\}$$

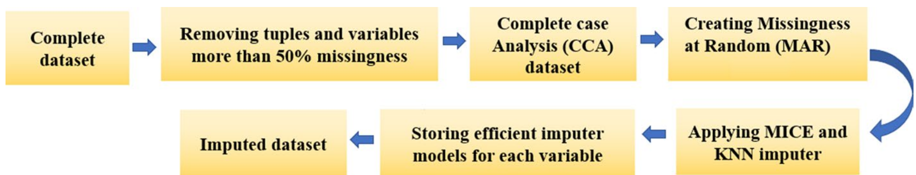


Fig. 2 Steps for data imputation

$$\text{Eucledian distances } (D_i) = \left(\left(x_1 - x_1'\right)^2 + \left(x_2 - x_2'\right)^2 + \ldots (x_n - x_n')^2 \right)^{0.5}$$

$$\text{Inverse distance weighing } (W_i) = \frac{1}{D_i}$$

$$X_{\text{predict}} = (V_1 + V_2 + V_3 + \cdots V_n)/K \tag{1}$$

Or

$$X_{\text{predict}} = \left((V_1 * W_1) + (V_2 * W_2) + \cdots (V_n * W_n)\right)/(W_1 + W_2 + \cdots W_n) \tag{2}$$

where V is the target value of the nearest neighbours and K or 10 nearest neighbours are considered for data imputation in the present simulation (Pozdnoukhov et al. 2008). Not only data records but variables can also carry weightage based on their significance for the targeted imputed variables. Various distance metrics such as Euclidean and grey can be used to fetch the nearest neighbours. Several studies on missing data imputation in different disciplines have been conducted using *k*NN. Kim et al. (2019) stated *k*NN performed best among other data imputation for weather variables used in Photovoltaic system forecasting over Korea. García-Laencina et al. (2009) proposed feature-weighted distance metric based on mutual information (MI) using *k*NN on two incomplete open datasets, Voting and Hepatitis, are from the UCI repository. Other studies on *k*NN for imputation include Batista and Monard (2002), Troyanskaya et al. (2001), Kim et al. (2019), Brás and Menezes (2007), Zhang (2011), Huang and Lee (2004), Zhang (2012), Tlamelo et al. (2021) and Choudhury and Kosorok (2020).

In the present study, *k*NN Imputer of Sklearn library in python is used for imputation of meteorological variable. Inverse Euclidean distance weighing metric with 10 nearest neighbours are used in the proposed work. The algorithm self-organises if more than one feature of the data tuple is missing and computes distance accordingly.

3.4 MICE: multivariate imputation by chained equation

Multivariate imputation by chained equations (MICE) was introduced by Van Buuren (1999), where he created imputed datasets based on a set of imputation models, one model for each variable with missing values. MICE is also known as "fully conditional specification" or "sequential regression multiple imputation". It specifies the multivariate imputation model on a variable-by-variable basis by a set of conditional densities obtained by different regression models, one for each incomplete variable.

MICE is a robust and informative method to deal with missing data. It imputes missing data in a dataset through an iterative series of predictive models. In each iteration, each specified variable in the dataset is imputed using the other variables in the dataset. These iterations continue until convergence is met. MICE methods are heavily reliant on the assumption of missing values being MAR which means that the probability that a value is missing depends only on observed values and not on unobserved values (Schafer and Graham 2002). MICE provides multiple values corresponding to one missing value by creating a series of regression (or other suitable) models, depending on its 'method' parameter. In MICE, each missing variable is treated as a dependent variable, and left out in the record is treated as an independent variable. Figure 3 explains the general principal on which MICE operates.

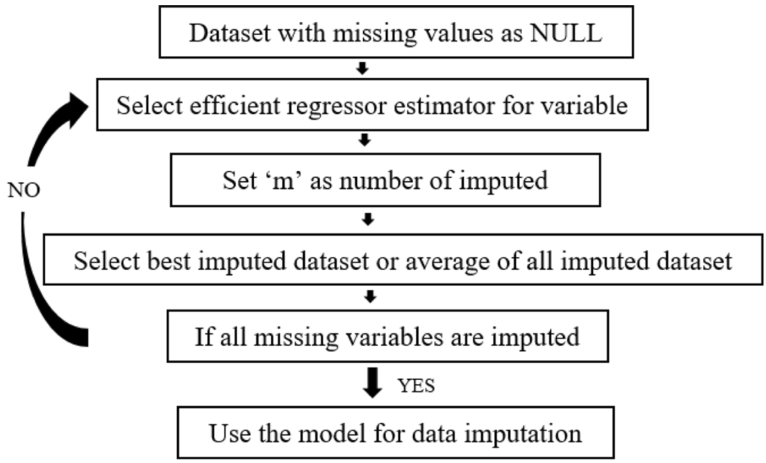


Fig. 3 Working of MICE

MICE is used in different discipline for data imputes includes research of Khan and Hoque (2020), Chhabra et al. (2017), Javadi et al. (2021), Wesonga (2015), Carvalho et al. (2017), Azur et al. (2011), Buuren and Groothuis-Oudshoorn (2011), Kim et al. (2019); Norazizi and Deni (2019), Alruhaymi and Kim (2021), Costa et al. (2021) and many more.

In the present study, IterativeImputer of Sklearn library in python is used for imputation of meteorological variables over Station-1 and Station-2. Iterative imputer of Sklearn in python is a replica of MICE package in R. IterativeImputer models each feature with missing values as a function of other features, and uses that estimate for imputation. In Sklearn, IterativeImputer is provides with four inbuilt estimators namely Bayesian Ridge, KNeighboursRegressor, ExtraTreesRegressor and DecisionTreeRegressor for MICE implementation. Each estimator works as follows (James et al. 2013; Hackeling 2017; Tlamelo et al. 2021; Alruhaymi and Kim 2021).

3.4.1 Bayesian ridge (MICE-BR)

Type of Bayesian regression which estimates a probabilistic model of the regression problem using ridge regression. It uses *L2* regularization for finding a maximum a posteriori estimation under a Gaussian prior over the coefficients w with precision λ^{-1}. Model is trained to find the best suited lambda to the simulation.

3.4.2 Decision tree (MICE-DT)

They are a non-parametric supervised learning method used for classification and regression. They predict the missing values by learning simple decision rules (if–then–else decision rules) inferred from the data features. Functions used to measure the quality of a split are "MSE" (used in data imputation) for the mean squared error, "friedman_MSE" (mean squared error with Friedman's improvement score for potential splits), "MAE" for the mean absolute error and "poisson" which uses reduction in Poisson deviance to find splits.

3.4.3 Extra tree regressor (MICE-ETR)

It implements a meta estimator that fits a number of randomized decision trees (extra-trees or ensemble of decision trees) on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. Extra tree regressor has MSE and MAE as its supporting criteria for splits.

3.4.4 *k* neighbours regressor (MICE-*k*NN)

Implements learning based on the k (= 10 used in the study) nearest neighbours, where k is an integer value specified by the user. k Neighbours Regressor is different from *k*NN imputation, which learns from samples with missing values by using a distance metric that accounts for missing values, rather than imputing them. Other applicability like distance metrics are similar to *k*NN Imputer.

3.5 Artificial neural network (ANN)

A single hidden layer multi-layer perceptron ANN with single output node for prediction of avalanche occurrence has been developed for both the stations using imputed and non-imputed weather data to predict avalanche occurrences. ANN is implemented using Sklearn MLPclassifier library with stochastic gradient as a technique to optimize weights and biases for the network. The methodology of development of ANN has been discussed in detail by Joshi et al. (2020). They have used ANN for simulation of snowpack parameters and prediction of avalanche hazard using Class-II data. In the present study, the ANN has been parameterized to deliver avalanche predictions in terms of occurrence and non-occurrence of avalanches. The ANN parameterized for avalanche forecasting has single hidden layer architecture with 14 input neurons 5 hidden neurons and 1 output neurons that correspond to avalanche occurrence by using predict_proba function of MLPclassifier which defines avalanche day based on the inputs. Weights and bias are initialized by the MLPclassifiers. The network is trained with a learning rate 0.001 and momentum 0.01. The network used 2,00,000 epochs for training with sigmoid as activation function and fault tolerance to 10^{-6}. ANN is specifically used to see the improvement in avalanche forecasting by improving data quality after data imputation for both the stations. ANN has been used worldwide (Joshi et al. 2020; Kaur et al. 2022; Dekanová et al. 2018; Schirmer et al. 2009; Singh and Ganju 2008 etc.) by different research in avalanche predictions.

4 Results and discussions

The purpose of the study is to develop an efficient data imputation technique for snow meteorological data for Station-1 and Station-2. Data imputation is carried on meteorological variable includes relative humidity, maximum temperature, minimum temperature, wind speed, fresh snow and pressure.

Performance measures used to define the suitability of imputation models are root-mean-square error, standard deviation, Coefficient of Determination, Taylor diagram and scatter plots. As discussed in methodology *k*NN imputer (*k*NN) and Iterative Imputer (MICE) are used for data imputation. MICE further has four different estimators for imputation in sklearn. Therefore, study of *k*NN and MICE for imputation is

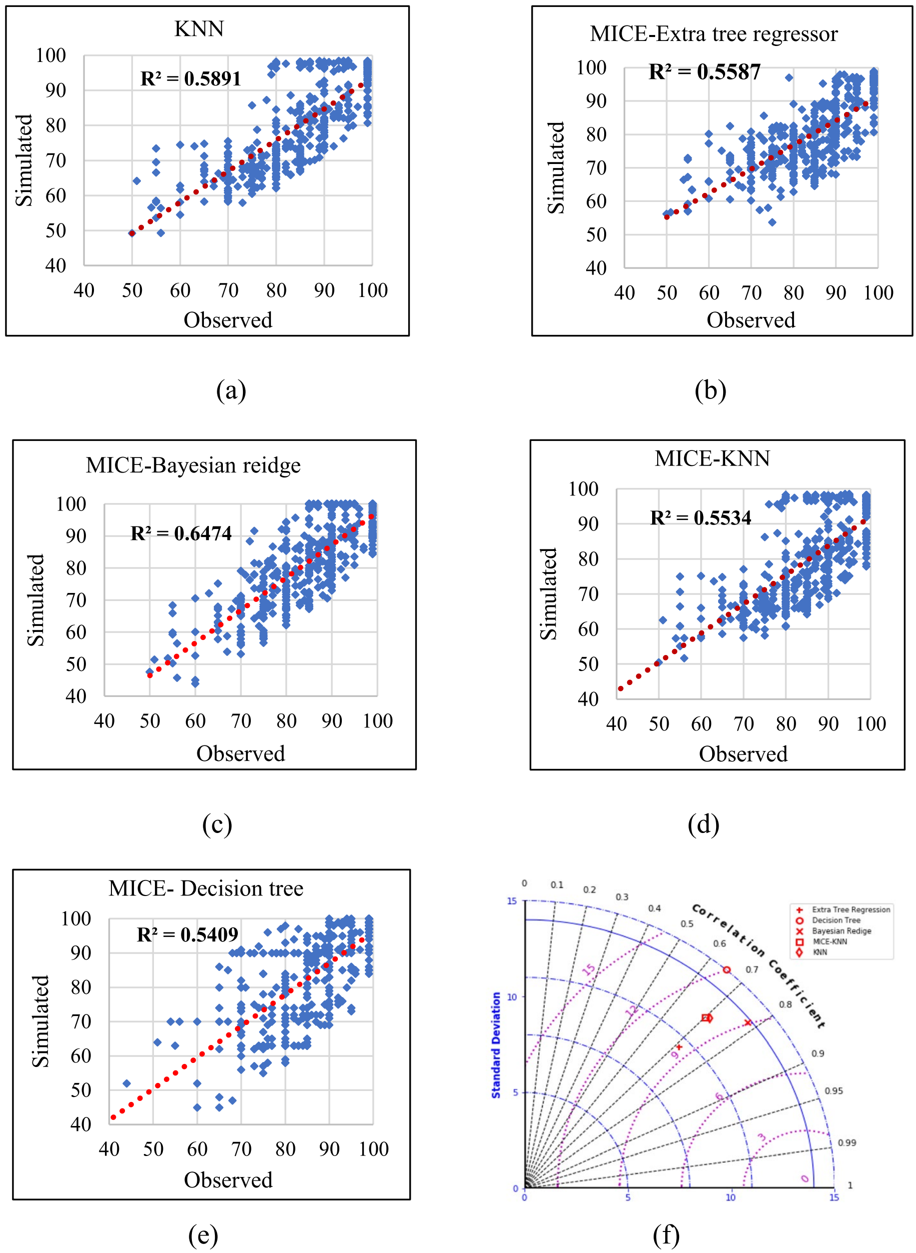


Fig. 4 Relative humidity imputation over Station-1 using **a** *k*NN Imputer, **b** Iterative Imputer Extra tree regressor, **c** Iterative Imputer Bayesian Ridge, **d** Iterative Imputer *k* nearest neighbour, **e** Iterative Imputer decision tree, **f** Taylor diagram for RH

done in two phases, first using relative humidity (RH) with 10% missingness as shown in Fig. 4a–f to study all the imputation approaches. From scatter plot in Fig. 4a–e, Taylor diagram Fig. 4f and Table 3, $RH_{Station1}$ estimation sequence MICE-BR, MICE-ETR,

Table 3 Standard deviation and RMSE of relative humidity using *k*NNImputer and iterative imputer for Station-1

Technique	Standard deviation (%)	RMSE (%)
*k*NN	20	9.4
BR	20	8.6
*k*NN-MICE	20	9.9
ETR	20	8.7
DT	20	10.2

*k*NN, MICE-*k*NN, MICE-DT (RMSE 8.6, 8.7, 9.4, 9.9 and 10.2, respectively) with standard deviation of 20. RMSE is preferred above all the performance measures as the result produced is in same units are more informative than relative performances. Although all the models imputed relative humidity with satisfactory RMSE as compared to standard deviation as shown in Table 3, MICE-BR holds highest correlation (> 0.8) with observations,and has the standard deviation of 8.6 and R^2 value 0.64. Figure 4f provides a summary of the relative skills with which models simulate the pattern of relative humidity. Decision tree had a low pattern correlation (< 0.7), R^2 0.54 and RMSE of 10.2. Although *k*NN and MICE-*k*NN have almost same correlation and deviation but different RMSE, MICE-ETR simulates with second best results because of its robustness to noise and inadequate features. Since MICE-BR has best performance, it was used as a MICE estimator in further imputation of missing variable. Moreover, MICE-BR uses poor distributions that allow to incorporate external knowledge into model which helps in efficient estimation. Difference between the best and worst model simulated RMSE is 1.6. MICE-DT was unable to efficiently impute the data because of its inadequate ability towards regression estimations and highly sensitive to small changes in data resulting in large changes in the tree structures. For evaluating current data with past similar prevailing condition KNN imputer was used to carry further imputation of the meteorological variable. In a study by Afrifa-Yamoah et al. (2020), relative humidity is imputed over four different locations of Australia using three different techniques whose RMSE varies from 3.5 to 13.05. In the proposed study of imputation humidity, RMSE is stated 9.4 (*k*NN) and 8.6 (MICE-Bayesian ridge) comparable to the humidity imputed over Australia by Afrifa-Yamoah et al. (2020). In Costa et al. (2021), MICE imputation of RH on daily scale showed correlations from 0.5 to 0.8 and a RMSE from 6.7 to 14.6%, similarly present study techniques KNN and MICE-BR showed correlation between 0.7 to 0.9 and RMSE from 8.5 to 9.5 better than the former study.

Table 4 Comparative analysis of *k*NN and MICE with variation in missingness

	Minimum temperature				Maximum temperature				Fresh Snow			
Station-1												
Missingness	10%	15%	20%	30%	10%	15%	20%	30%	10%	15%	20%	30%
R^2 (MICE)	0.81	0.78	0.77	0.77	0.9	0.89	0.89	0.88	0.88	0.86	0.85	0.84
R^2 (*k*NN)	0.63	0.62	0.62	0.61	0.59	0.58	0.57	0.55	0.23	0.21	0.21	0.18
Station-2												
Missingness	10%	15%	20%	30%	10%	15%	20%	30%	10%	15%	20%	30%
R^2 (MICE)	0.82	0.81	0.8	0.78	0.92	0.89	0.88	0.85	0.96	0.96	0.96	0.94
R^2 (*k*NN)	0.83	0.81	0.80	0.78	0.67	0.66	0.66	0.66	0.82	0.72	0.69	0.7

Second phase to test the algorithm, CCA is applied with random missingness of 10, 15, 20, and 30% for temperature (minimum and maximum) and fresh snow for Station-1 and Station-2. Table 4 illustrates the increase in coefficient of determination (R^2) as missingness decreased for both the stations, stating difficulty for the algorithms to efficiently impute the values and if missingness is more than 50% the variable cannot be estimates or it will provide biased imputations as training data are corrupted/biased (Aprianti and Mukhlash 2015).

Figure 5a-l represents comparative study through Coefficient of Determination and scatter plots of maximum temperature, minimum temperature, wind speed, pressure and fresh snow missing data imputation (with 10% missingness) through *k*NN and MICE-BR over Station-1. Table 5 represents RMSE corresponding to the standard deviation of the variables for both the techniques. MICE-BR and *k*NN performance varies from variable to variable. For Maximum temperature $_{\text{Station1}}$: MICE-BR, *k*NN; Minimum temperature $_{\text{Station1}}$: MICE-BR; *k*NN, Relative humidity $_{\text{Station1}}$: MICE-BR; *k*NN, Snowfall $_{\text{Station1}}$: MICE-BR; *k*NN, Wind Speed $_{\text{Station1}}$: *k*NN; MICE-BR, Pressure $_{\text{Station1}}$: *k*NN; MICE-BR. Overall MICE-BR demonstrated better results for Station-1 corresponding to temperatures, relative humidity, fresh snow. For fresh snow instead of best imputed model of MICE average of all the model SICE is used in estimation as suggested by Khan and Hoque (2020) in order to incorporate all the abilities like poor distribution from MICE-BR, sensitive to data from MICE-DT, robust to noise and irrelevant features from MICE-ETR and past knowledge from MICE-*k*NN. Coefficient of Determination (R^2) is greater than 0.6 for most of the variables imputed and reaches to max 0.9 for minimum temperature imputation. Similarly Fig. 6a-l represents comparative study through scatter plots of maximum temperature, minimum temperature, wind speed, pressure and fresh snow missing data imputation through *k*NN and MICE-BR over Station-2 station. Table 6 illustrates the RMSE and standard deviation of the snow meteorological data of Station-2. Figure 6 and Table 6 demonstrate snow meteorological data imputation stating MICE-BR efficiently handling missing data for Station-2 station except for wind speed and pressure. Sequence of model performance for Station-2 is as follow: Maximum temperature$_{\text{Station2:}}$ MICE-BR; *k*NN, Minimum temperature $_{\text{Station2}}$: MICE-BR; *k*NN, Snowfall $_{\text{Station2}}$: MICE-BR; *k*NN, Wind Speed $_{\text{Station2}}$: *k*NN; MICE-BR, Pressure$_{\text{Station2}}$: MICE-BR; *k*NN. Relative Humidity in Station-2 had missingness more than 75%, hence imputation of humidity was omitted for Station-2 as it can lead to biasness in the data. Based on the results shown in Table 6, coefficient of determination and scatter plots imputation of missing data has been done efficiently. For both Station 1 and Station 2 MICE imputed with better results than *k*NN but efficiency of MICE for Station 2 is more than Station 1. *k*NN for Indian Himalayas was not capable in precisely identifying anisotropies that are present in non-homogeneous regions such as mountains (Tung 1983). Distance alone, however, cannot affect the positive autocorrelation in climatological data; becoming major limitation of *k*NN in estimating meteorological variables in snow bound areas of Indian Himalayas.

Kanda et al. (2018) proposed data imputation methods include simple arithmetic average, inverse distance weighing, normal ratio method, single best estimator, multiple regression using the least absolute deviation criterion, UK traditional method and closest station method for maximum temperature and minimum temperature and precipitation for Karakorum range of Himalayas on different locations. RMSE stated by Kanda et al. (2018) for maximum temperature vary from 1.1 to 3.9 °C and for minimum temperature vary from 1.07 to 3.5 °C. Another study by Afrifa-Yamoah et al. (2020) over Australia used structural time series model autoregressive integrated moving average (ARIMA) model with Kalman smoothing and multiple linear regression for temperature, humidity and wind

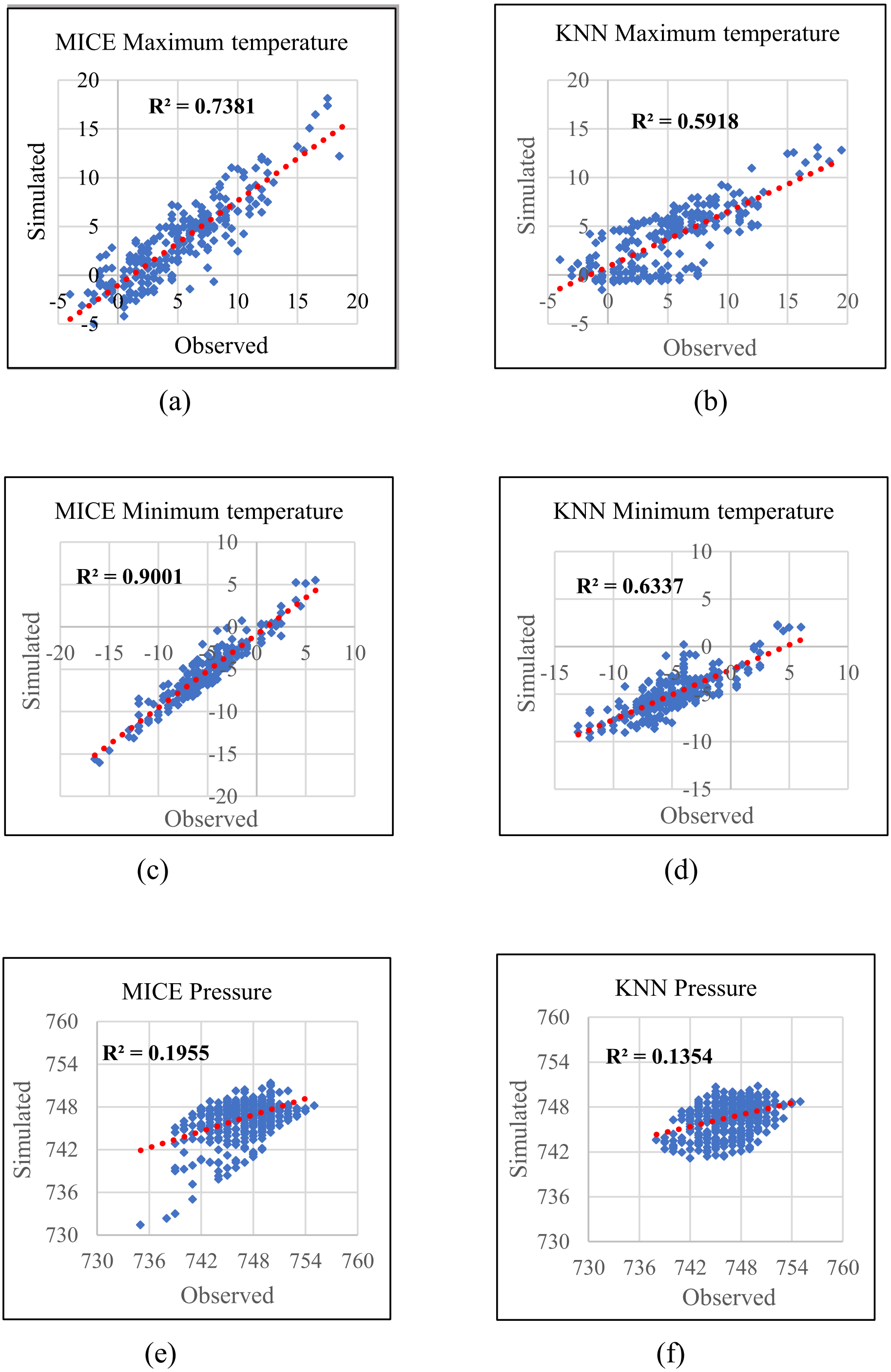


Fig. 5 Comparative study of Mice and *k*NN over snow meteorological data over Station-1

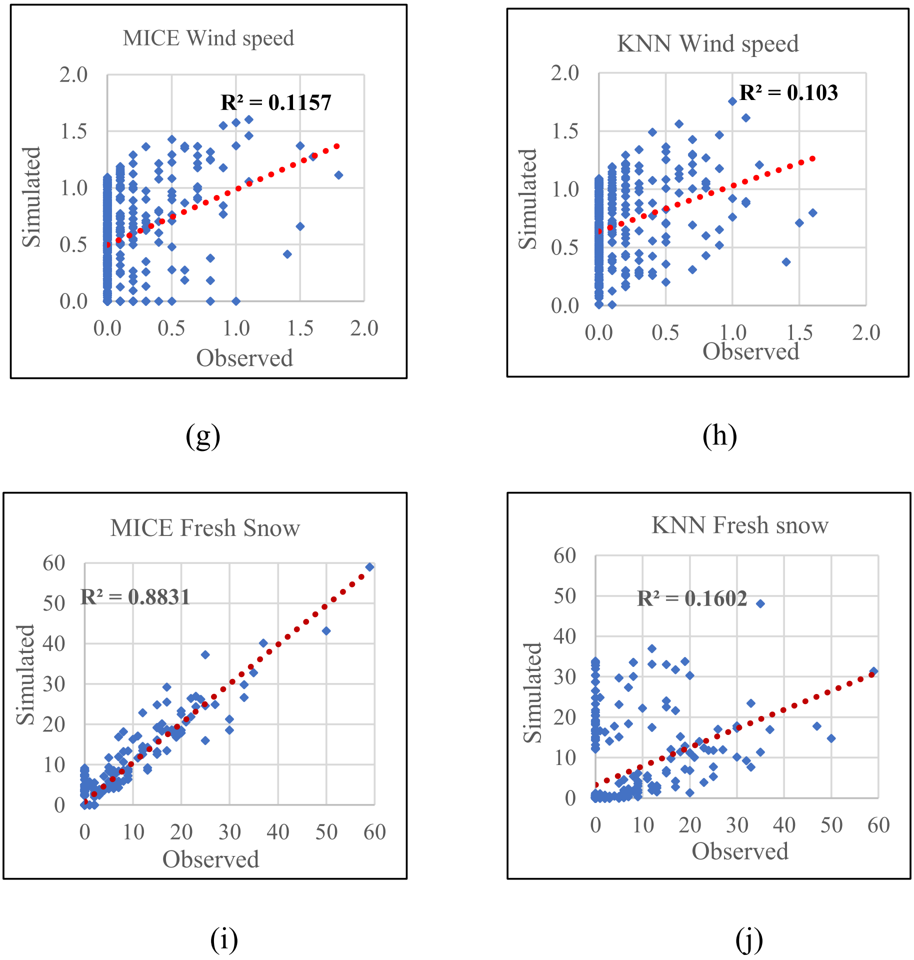


Fig. 5 (continued)

Table 5 Standard Deviation, RMSE using MICE and KNN over snow meteorological variable of Station-1

Variables	Standard deviation	RMSE *k*NN	RMSE MICE
Max Temperature	4.1 °C	3.1 °C	2.4 °C
Minimum Temperature	3.4 °C	2 °C	1.2 °C
Wind Speed	1.4 km/h	0.6 km/h	0.78 km/h
Relative Humidity	20%	9.4%	8.6%
Pressure	3.4 hPa	2.9 hPa	3 hPa
Fresh Snow	9 cm	8 cm	2.8 cm

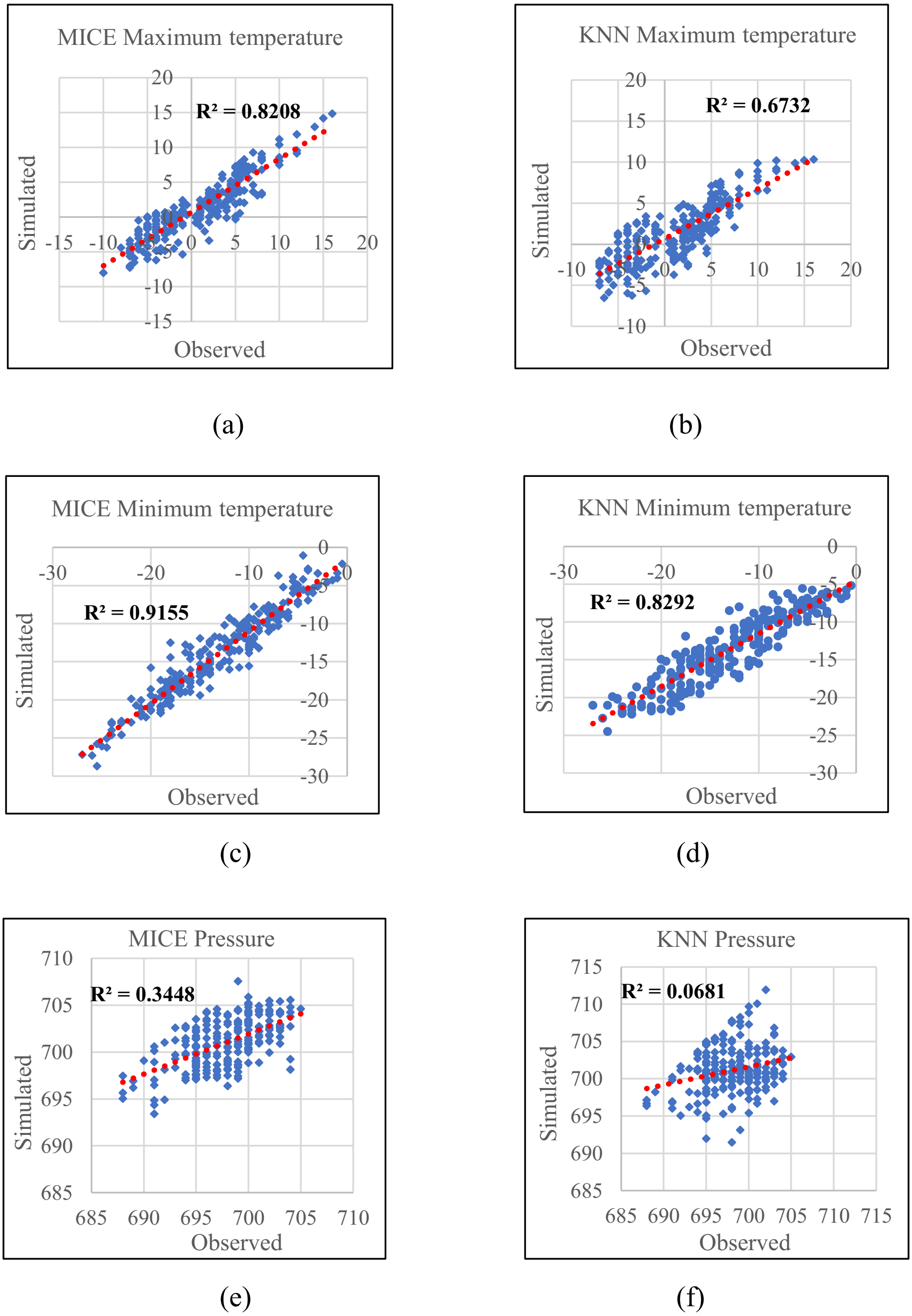
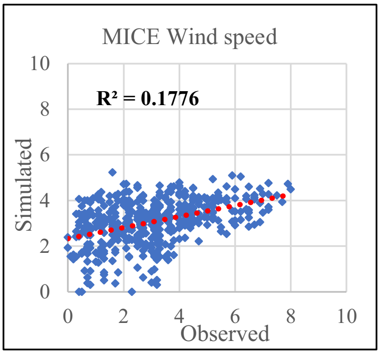
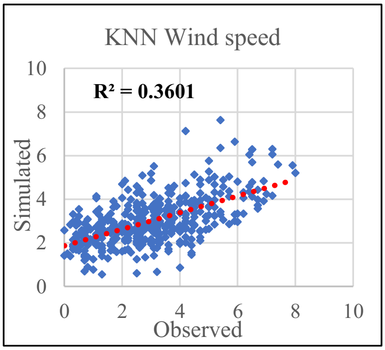


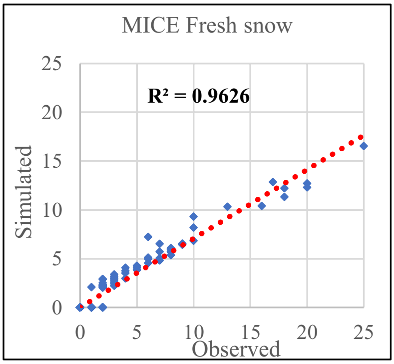
Fig. 6 Comparative study of Mice and *k*NN over snow meteorological data over Station-2



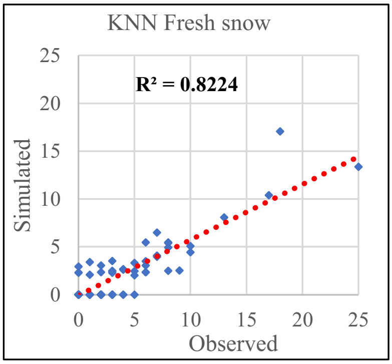
(g)



(h)



(i)



(j)

Fig. 6 (continued)

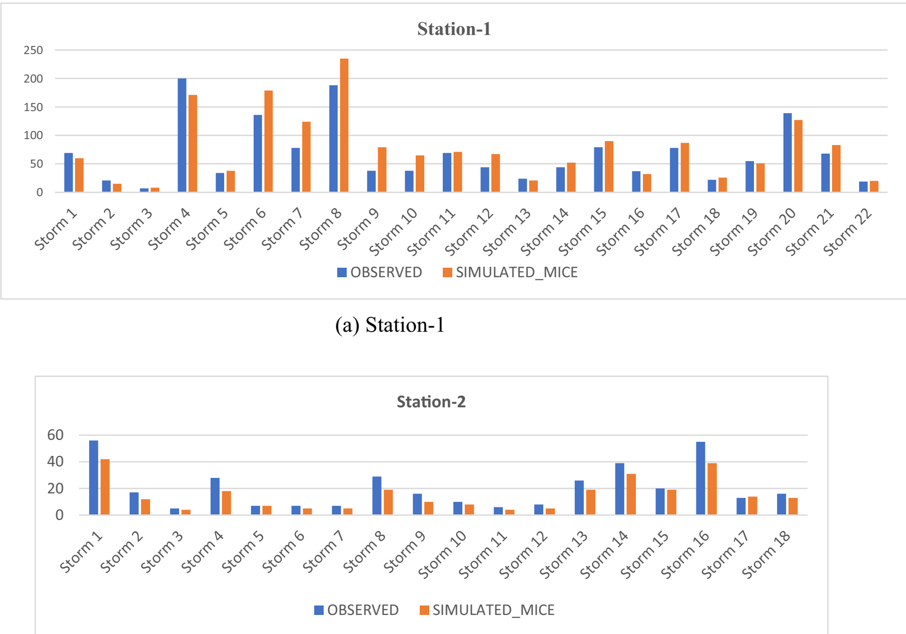
Table 6 Standard Deviation, RMSE using MICE and KNN over snow meteorological variable of Station-2

Variables	Standard deviation	RMSE KNN	RMSE MICE
Max Temperature	5.8 °C	2.8 °C	2.1 °C
Minimum Temperature	6.8 °C	2.6 °C	1.9 °C
Wind Speed	2.5 km/h	1.3 km/h	1.6 km/h
Relative Humidity	–	–	–
Pressure	7.7 hPa	4.9 hPa	4.2 hPa
Fresh Snow	3.6 cm	1.2 cm	1 cm

speed imputation. RMSE for temperature vary from 0.8 to 1.4 °C and wind speed from 1.9 to 2.8 km/h. Kotsiantis et al. (2006) proposed different methods for filling missing temperature in weather data banks. The least RMSE of 2.2 °C is achieved on using three years data. The proposed imputation techniques over meteorological variable for Station-1 and Station-2 in the study produces RMSE of 2.4 and 2.1 °C for maximum temperature and 1.2

and 1.9 °C for minimum temperature for Station-1 and Station-2. The RMSEs attained in the proposed study are comparable to research studies carried in past. Costa et al. (2021) imputed (MICE) temperature ranging RMSE ranged from 0.9 to 1.9 °C whereas the MICE-BR in current study imputed temperature 1.2 to 2.4 °C in comparable range. For atmospheric pressure Costa et al. (2021) RMSE ranges from 1 to 5 hPa whereas in the study from 2.9 to 4.2 hPa. However, for the wind speed, the proposed study had an RMSE of 0.6 and 1.3 km/h better than Afrifa-Yamoah et al. (2020) and comparable to Costa et al. (2021) (0.8–1.9 m/s) proposed technique but according to scatter plots and R^2 wind speed and pressure are need improvement in estimation for both stations. The main reason for deprived estimation in wind speed and pressure is the curvature and topology of mountain surfaces, as well as their presence, can impact the vertical movement of heat and moisture. This can have an influence on cloud formation and precipitation in the surrounding area, as mountains can act as barriers to large-scale atmospheric flows causing difficulties for MICE and *k*NN to learn the trend and imputing wind and pressure data. Precipitation study on Karakorum Himalayas by Kanda et al. (2018) stated RMSE between 2.1 and 3.3 cm when it is missing at random. Purposed data imputation imputed fresh snow at an RMSE of 2.8 cm for station-1 and 1 cm for station-2 by MICE is outperforming the study proposed by Kanda et al. (2018) and Costa et al. (2021) (RMSE from 4 to 12 mm). RMSEs of all the variables are less than the standard deviation of the data in the database.

The performance of MICE imputed fresh snow has further been evaluated by comparing imputed and observed total snowfall during major snowfall during 2017–2019 (MAR-10%, test data). The observed and MICE imputed storm snow during 2017–2019 for station-1 and station-2 as shown in Fig. 7 represents the imputation has reproduced snowfall with reasonable accuracy. However, it has over-predicted heavy snowfall events for station-1



(a) Station-1

(b) Station-2

Fig. 7 Major snow events 2017–2019 observed and predicted by MICE. **a** Station-1 and **b** Station-2 such as storm 6, 7 and 8. The overall performance of MICE imputation model for snowfall has been found considerably good during the validation period for both stations. Hence, snowfall and temperature data imputed by MICE can be used for various applications, including implementation of avalanche forecasting models in regions where observed weather data are missing.

The high efficiency of advanced methods such as artificial neural networks has been reported by Joshi et al. (2020), Teegavarapu and Chandramouli (2005), Ustoorikar and Deo (2008) and Kashani and Dinpashoh (2012). Therefore, an ANN-based avalanche forecasting model has been developed for station-1 and station-2 using MLP classifies of sklearn to validate data imputation. Hyperparameters such as type of activation function, threshold, momentum, learning rate and iteration are kept same for both networks (i.e. ANN model without data imputation and ANN model with data imputation). In case of Station-1 POD incremented to 0.71 from 0.67, HSS to 0.36 from 0.31 accuracy to 0.74 from 0.71 after missing value imputation of the variables having miss percentage less than 50%. False alarms and bias decrement to 0.59 from 0.62 and 1.73 from 1.76 as stated in Fig. 8a. In case of Station-2, HSS incremented to 0.3 from 0.24 accuracy to 0.72 from 0.68 after missing value imputation of the variables having missingness less than 50%. False alarms and bias decrement to 0.57 from 0.63 and 1.32 from 1.53 as stated in Fig. 8b. Though POD remains same to 0.56, overall performance of

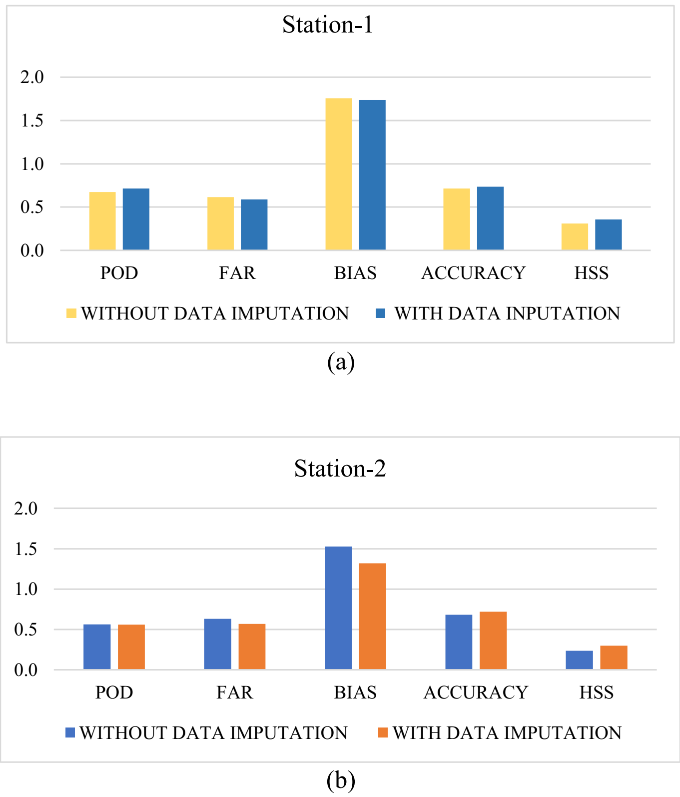


Fig. 8 Performance measures of avalanche prediction models with and without data imputation

avalanche forecasting model for Station-2 after missing data imputation is improved. HSS greater than 0.25 for the forecast of a natural random process such as avalanche is considered better than the random forecast (Joshi et al. 2020) which improved with data imputation for both the stations. ANN though have complex structure but at the same time is immune to noise and outliers. Difference in geographical and meteorological features, missingness ratio in the study areas has affected ANN performance but at the same time there an evident improvement in the forecast in both the areas. Hence, Fig. 8 states the merits of missing data imputation in significantly enhancing avalanche forecasting for both the station.

The prominence of estimating missing climate data cannot be overlooked in regions such as mountains and forests where data are affected by topography and microclimates of the region (Kashani and Dinpashoh 2012). Based upon the results obtained, enhancement in performance of avalanche forecasting model, their comparative analysis, RMSE and scatter plots, it is inferred that the MICE is suitable for estimating missing values of temperature, relative humidity and fresh snowfall over Indian Himalayas.

5 Conclusions and future scopes

Snow Meteorological datasets are subjected to suffer a common drawback, missing or incomplete data resulting in atrocious training of the avalanche prediction model increasing the risk in avalanche prone areas. In the proposed research, snow meteorological data imputation technique is designed for two different locations of Indian Himalayas Station-1 in lower Himalayas and Station-2 in Greater Himalayas using *k*-nearest neighbour (*k*NN Imputer) and multivariate imputation by chained equation (MICE). The methods studied have demonstrated their suitability in imputing missing data in maximum temperature, minimum temperature, humidity and fresh snow on daily basis. A comparative study was carried out between *k*NN Imputer (*k*NN) and Iterative Imputer (MICE) on the locations where the latter has accurately estimated missing data. The methods' performance was assessed using various measures such as root-mean-square error, coefficient of determination, standard deviation, scatter plots, Taylor diagram, and performance metrics like POD, HSS, accuracy, FAR, and bias for avalanche forecasting model (ANN). Additionally, major snow events during the testing period were compared for evaluation purposes. The RMSE of all the imputed weather variables has been found significantly smaller than their standard deviations. RMSE's of the variables were found equivalent to the other studies conducted worldwide for imputation of temperature, wind, fresh snow, pressure and humidity (Kanda et al. 2018; Afrifa-Yamoah et al. 2020; Kotsiantis et al. 2006, Costa et al. 2021). Overall accuracy and HSS of both the station incremented to 0.74 from 0.71 and 0.72 from 0.68, 0.36 from 0.31 and 0.3 from 0.24 for station-1 and Station-2, respectively. It is imperative to consider utilizing multiple imputation models as a flexible technique for accommodating various variables when filling in missing data gaps in snow meteorology. Research has proven its efficiency, making it a viable option for experts in designing avalanche forecasting models.

The study can further be enhanced using mean imputation from different imputation model. In MICE, instead of best data imputation, mean imputation of the imputed variable can be considered. Artificial neural networks and support vector machines have shown their applicability in imputation; therefore, these and other machine learning methods can be considered. Moreover, KNN and MICE can be more exhaustively explored in the python libraries to achieve more accuracy.

Acknowledgements The authors acknowledge Director SASE for his approval to initiate this work under DRDO-funded project—Him Sandesh. The technical staff of SASE is also acknowledged for their valuable contribution in the collection of manual field data.

Author contributions All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by PK and JCJ. The first draft of the manuscript was written by PK and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Declaration

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

References

Afrifa-Yamoah E, Mueller UA, Taylor SM, Fisher AJ (2020) Missing data imputation of high-resolution temporal climate time series data. Meteorol Appl 27(1):e1873
Alruhaymi AZ, Kim CJ (2021) Study on the missing data mechanisms and imputation methods. Open J Stat 11(4):477–492
Aprianti W, Mukhlash I (2015) Handling missing value on meteorological data classification with rough set based algorithm. Global J Pure Appl Math 11(3):1147–1155
Azur MJ, Stuart EA, Frangakis C, Leaf PJ (2011) Multiple imputation by chained equations: What is it and how does it work? Int J Methods Psychiatr Res 20(1):40–49
Batista GE, Monard MC (2002) A study of K-nearest neighbour as an imputation method. HIS 87:48
Brás LP, Menezes JC (2007) Improving cluster-based missing value estimation of DNA microarray data. Biomol Eng 24(2):273–282
Che Ghani NZ, Abu Hasan Z, Tze Liang L (2014) Estimation of missing rainfall data using GEP: case study of raja river, Alor Setar, Kedah. Adv Artif Intell 2014:6
Chhabra G, Vashisht V, Ranjan J (2017) A comparison of multiple imputation methods for data with missing values. Indian J Sci Technol 10(19):1–7
Choge HK, Regulwar DG (2013) Artificial neural network method for estimation of missing data. Int J Adv Technol Civ Eng 2(1):1–4
Choudhury A, Kosorok MR (2020) Missing data imputation for classification problems. arXiv preprint arXiv:2002.10709.
Costa RL, Barros Gomes H, Cavalcante Pinto DD, da Rocha Júnior RL, dos Santos Silva FD, Barros Gomes H, Luís Herdies D (2021) Gap filling and quality control applied to meteorological variables measured in the northeast region of Brazil. Atmosphere 12(10):1278
Dastorani MT, Moghadamnia A, Piri J, Rico-Ramirez M (2010) Application of ANN and ANFIS models for reconstructing missing flow data. Environ Monit Assess 166(1):421–434
de Carvalho JRP, Almeida Monteiro JEB, Nakai AM, Assad ED (2017) Model for multiple imputation to estimate daily rainfall data and filling of faults. Revista Brasileira De Meteorologia 32:575–583
Dekanová M, Duchoň F, Dekan M, Kyzek F & Biskupič M (2018) Avalanche forecasting using neural network. In: 2018 ELEKTRO, IEEE, pp 1–5
Enders CK (2010) Applied missing data analysis. Guilford press, New York
Firat M, Dikbas F, Koc AC, Gungor M (2012) Analysis of temperature series: estimation of missing data and homogeneity test. Meteorol Appl 19(4):397–406
García-Laencina PJ, Sancho-Gómez JL, Figueiras-Vidal AR, Verleysen M (2009) K nearest neighbours with mutual information for simultaneous classification and missing data imputation. Neurocomputing 72(7–9):1483–1493

Gómez-Carracedo MP, Andrade JM, López-Mahía P, Muniategui S, Prada D (2014) A practical comparison of single and multiple imputation methods to handle complex missing data in air quality datasets. Chemom Intell Lab Syst 134:23–33

Hackeling G (2017) Mastering machine learning with scikit-learn. Packt Publishing Ltd., Mumbai

Huang CC, Lee HM (2004) A grey-based nearest neighbor approach for missing attribute value prediction. Appl Intell 20(3):239–252

James G, Witten D, Hastie T, Tibshirani R (2013) An introduction to statistical learning. Springer, New York, p 18

Javadi S, Bahrampour A, Saber MM, Garrusi B, Baneshi MR (2021) Evaluation of four multiple imputation methods for handling missing binary outcome data in the presence of an interaction between a dummy and a continuous variable. J Probab Stat 2021:1–14

Joshi JC, Kaur P, Kumar B, Singh A, Satyawali PK (2020) HIM-STRAT: a neural network-based model for snow cover simulation and avalanche hazard prediction over North-West Himalaya. Nat Hazards 103(1):1239–1260

KA ND, Tahir NM, Abd Latiff ZI, Jusoh MH, Akimasa Y (2022) Missing data imputation of MAGDAS-9's ground electromagnetism with supervised machine learning and conventional statistical analysis models. Alex Eng J 61(1):937–947

Kajewska-Szkudlarek J, Stańczyk J (2018) Filling missing meteorological data with Computational Intelligence methods. In: ITM web of conferences, vol 23, EDP Sciences, Les Ulis, p 00015

Kanda N, Negi HS, Rishi MS, Shekhar MS (2018) Performance of various techniques in estimating missing climatological data over snowbound mountainous areas of Karakoram Himalaya. Meteorol Appl 25(3):337–349

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

Kaur P, Joshi JC, Aggarwal P (2022) A multi-model decision support system (MM-DSS) for avalanche hazard prediction over North-West Himalaya. Nat Hazards 110(1):563–585

Khan SI, Hoque ASML (2020) SICE: an improved missing data imputation technique. J Big Data 7(1):1–21

Kim JW, Pachepsky YA (2010) Reconstructing missing daily precipitation data using regression trees and artificial neural networks for SWAT streamflow simulation. J Hydrol 394(3–4):305–314

Kim T, Ko W, Kim J (2019) Analysis and impact evaluation of missing data imputation in day-ahead PV generation forecasting. Appl Sci 9(1):204

Kotsiantis S, Kanellopoulos D, Pintelas P (2006) Handling imbalanced datasets: a review. GESTS Int Trans Comput Sci Eng 30(1):25–36

Kwak SK, Kim JH (2017) Statistical data preparation: management of missing values and outliers. Korean J Anesthesiol 70(4):407–411

Lara-Estrada L, Rasche L, Sucar E, Schneider UA (2018) Inferring missing climate data for agricultural planning using Bayesian network. Land 7(4):1–13

Madley-Dowd P, Hughes R, Tilling K, Heron J (2019) The proportion of missing data should not be used to guide decisions on multiple imputation. J Clin Epidemiol 110:63–73

Norazizi NAA, Deni SM (2019) Comparison of artificial neural network (ANN) and other imputation methods in estimating missing rainfall data at Kuantan station. In: Soft computing in data science: 5th international conference, SCDS 2019, Iizuka, Japan, Springer, Singapore, pp 298–306

Orczyk T, Porwik P (2013) Influence of missing data imputation method on the classification accuracy of the medical data. J Med Inform Technol 22 pp. 111–116

Pickles A (2005) Missing data: problems and solutions: problems and solutions. In: Encyclopedia of social measurement. Academic Press, Ltd

Pozdnoukhov A, Purves RS, Kanevski M (2008) Applying machine learning methods to avalanche forecasting. Ann Glaciol 49:107–113

Rubin DB (1976) Inference and missing data. Biometrika 63(3):581–592

Rubin DB (1987) Multiple imputation for nonresponse in surveys. John Wiley & Sons Inc., New York. https://doi.org/10.1002/9780470316696

Sattari MT, Rezazadeh-Joudi A, Kusiak A (2017) Assessment of different methods for estimation of missing data in precipitation studies. Hydrol Res 48(4):1032–1044

Schafer JL, Graham JW (2002) Missing data: our view of the state of the art. Psychol Methods 7(2):147

Schirmer M, Lehning M, Schweizer J (2009) Statistical forecasting of regional avalanche danger using simulated snow-cover data. J Glaciol 55(193):761–768

Schneider T (2001) Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values. J Clim 14(5):853–871

Sharma SS (2000) An overview of snow and avalanche research in Indian Himalaya. In: Proceedings of the international snow science workshop, pp 558–565

Singh A, Ganju A (2008) Artificial neural networks for snow avalanche forecasting in Indian Himalaya. In: Proceedings of 12th international conference of international association for computer methods and advances in geomechanics, IACMAG, vol 16
Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, Carpenter JR (2009) Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. Bmj. https://doi.org/10.1136/bmj.b2393
Tabony RC (1983) The estimation of missing climatological data. J Climatol 3(3):297–314
Teegavarapu RS (2009) Estimation of missing precipitation records integrating surface interpolation techniques and spatio-temporal association rules. J Hydroinf 11(2):133–146
Teegavarapu RS, Chandramouli V (2005) Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records. J Hydrol 312(1–4):191–206
Tlamelo E, Thabiso M, Dimane M, Thabo S, Banyatsang M, Oteng T (2021) A survey on missing data in machine learning. J Big Data 8(1):1–37
Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB (2001) Missing value estimation methods for DNA microarrays. Bioinformatics 17:520–525
Tung YK (1983) Point rainfall estimation for a mountainous region. J Hydraul Eng 109(10):1386–1393
Ustoorikar K, Deo MC (2008) Filling up gaps in wave data with genetic programming. Mar Struct 21(2–3):177–195
Van Buuren S, Groothuis-Oudshoorn K (2011) Mice: multivariate imputation by chained equations in R. J Stat Softw 45:1–67
Van Buuren S, Boshuizen HC, Knook DL (1999) Multiple imputation of missing blood pressure covariates in survival analysis. Stat Med 18(6):681–694
Wesonga R (2015) On multivariate imputation and forecasting of decadal wind speed missing data. SpringerPlus. https://doi.org/10.1186/s40064-014-0774-9
Wu X, Kumar V, Quinlan JR, Ghosh J, Yang Q, Motoda H, Steinberg D (2008) Top 10 algorithms in data mining. Knowl Inform Syst 14(1):1–37
Yozgatligil C, Aslan S, Iyigun C, Batmaz I (2013) Comparison of missing value imputation methods in time series: the case of Turkish meteorological data. Theoret Appl Climatol 112:143–167
Zhang S (2011) Shell-neighbor method and its application in missing data imputation. Appl Intell 35:123–133
Zhang S (2012) Nearest neighbor selection for iteratively kNN imputation. J Syst Softw 85(11):2541–2552
Zhang Z (2015) Missing values in big data research: some basic skills. Ann Transl Med 3(21):323

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.