International Journal of Mathematics and Statistics (IJMS)

DOI: 10.5281/zenodo.22985985 | ISSN: 3141-6438 | Volume: 2 | Issue: 1 | 2026

A Comparative Imputations Statistics Approaches With Application to Nigerian Rainfall Data

Taiwo Wale Ayanniyi[1*], Efosa Michael Ogbeide[2]

[1,2]Department of Mathematics and Statistics, Ambrose Alli University, Ekpoma, Edo State, Nigeria

Corresponding Author: ayanniyitaiwo7@gmail.com

Abstract

Missing values in rainfall datasets reduce the reliability of statistical inference and undermine decision-making in agriculture, climate studies, and environmental management. This study evaluated the performance of five imputation techniques; Expectation-Maximization (EM), Multiple Imputation (MI), Regression Imputation (RI), Bootstrap Expectation-Maximization (BEM), and Random Forest (RF) using the 2019 Nigerian rainfall dataset obtained from the National Bureau of Statistics. The methods were compared using Raw Bias, Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Variance to assess estimation accuracy, predictive performance, and stability. The results revealed that MI produced the lowest bias (-0.003), making it the most suitable method for minimizing systematic estimation error. In contrast, RF achieved the highest predictive accuracy, recording the lowest MSE (0.9530) and RMSE (0.9762). Although BEM exhibited the lowest variance (0.9844), indicating greater stability, it was associated with relatively high bias, limiting its overall effectiveness. The findings demonstrate that no single method is universally optimal; rather, the choice of imputation technique should be guided by the primary analytical objective. RF is recommended for applications requiring high predictive accuracy, whereas MI is preferable when unbiased parameter estimation is essential. The study provides empirical evidence to support the adoption of robust imputation techniques by agencies such as the Nigerian Meteorological Agency (NiMet), thereby improving the quality of national climate databases and strengthening evidence-based agricultural and environmental decision-making.

Keywords: Expectation-Maximization, Multiple Imputation, Regression Imputation, Bootstrap Expectation-Maximization, and Random Forest.

1. Introduction

In the contemporary digital economy, data has become a critical asset for decisions across healthcare, finance, marketing, and environmental science. However, modern datasets are frequently compromised by missing observations across several digital forms. While rapid digitization yields vast volumes, extracting meaningful insights requires addressing these incomplete records. Resolving this challenge is essential for informed strategic planning and gaining competitive advantages [5, 19].

Data mining represents a foundational analytical process focused on extracting latent patterns, non-trivial relationships, and systematic knowledge from complex, large-scale repositories to convert raw data into actionable intelligence [5, 18]. Core paradigms such as classification, clustering, and association rule mining empower modern organizations to decipher intricate consumer behaviours, flag transactional anomalies, optimize operational workflows, and execute high-precision predictive modelling [2, 19].

Despite its analytical power, a critical barrier to effective knowledge discovery in data mining remains the ubiquitous presence of incomplete records. Missing values represent a systemic vulnerability arising from heterogeneous sources, such as human errors during manual logging, sensor or equipment malfunctions, participant non-responses in empirical surveys, and synchronization mismatches during multi-source data integration. If not rigorously addressed during the data pre-processing phase, these information gaps inevitably diminish statistical power, introduce severe selection bias, and generate systematically misleading downstream conclusions. Consequently, deploying advanced mitigation strategies is essential to maintain model validity and guarantee the reliability of empirical findings [1, 15].

In academic literature, climate and weather-related observational streams serve as critical assets for environmental monitoring, precision agricultural planning, water resource management, and national disaster preparedness framework designs. Within agrarian developing economies such as Nigeria, localized precipitation records are uniquely paramount; rainfall functions as the foundational source of water sustaining sub-Saharan agricultural productivity, directly dictating national food security, regional hydrological modelling, and public climate adaptation initiatives. However, the institutional utility of these meteorological repositories is frequently compromised by missing values, which typically occur due to transmission dropout, station power failures, or sensor malfunctions within ground-based networks. Left unmitigated, these fragmented datasets inject systemic uncertainty, introduce significant diagnostic noise, and heavily degrade the accuracy of downstream predictive climate modelling and seasonal forecasting architectures. Consequently, robust missing data resolution has become a foundational prerequisite for securing reliable, data-driven climate resilience and early warning systems across vulnerable topographies

To mitigate the bias and statistical power loss associated with observational missingness, researchers rely on a spectrum of statistical and machine learning imputation paradigms designed to preserve both the structural dependencies and intrinsic variance of incomplete datasets. Within the climate sciences, these methodologies are routinely employed to address data gaps arising from instrument failure, transmission dropouts, or irregular sampling regimes. Key analytical classes include: Expectation–Maximization (EM) Algorithm: A likelihood-based iterative procedure that computes maximum likelihood parameter estimates in the presence of missing data [14]. Multiple Imputation (MI): A stochastic framework that generate plausible completed datasets to explicitly account for and propagate missing-data uncertainty across downstream statistical inferences [4].

Regression-Based Imputation: A deterministic or stochastic approach that leverages parameter estimation across observed predictor variables to model and fill missing conditional observations (Başakın et al., 2023).Bootstrap Expectation–Maximization (BEM): An integrative approach combining non-parametric bootstrap resampling with the EM algorithm to stabilize parameter estimation and refine confidence interval construction under non-standard likelihoods, small sample sizes, or complex missingness patterns [14].

Random Forest (RF) Imputation: A non-parametric machine learning strategy (e.g., MissForest, MissARF) capable of capturing complex non-linear relationships, multi-way interactions, and high-dimensional space dependencies without requiring explicit distributional assumptions [9].

Applying appropriate missing-data imputation strategies to rainfall datasets is critical to ensuring temporal completeness, enhancing the fidelity of regional climate analyses, and supporting data-driven decision-making across environmental, agricultural, and hydrometeorological sectors [6] Evaluating alternative imputation paradigms on empirical observational data, specifically Nigerian precipitation series characterized by localized spatial variability and temporal dropouts provides empirical insight into their comparative estimation efficiency, structural accuracy, and operational suitability for climate research [4, 16].

This study evaluates the performance of parametric, non-parametric, and machine learning imputation techniques using the 2019 Nigerian Bureau of Statistics (NBS) rainfall dataset. Specifically, the paper compares five methodologies: Expectation–Maximization (EM), Multiple Imputation (MI), deterministic and stochastic Regression Imputation, Bootstrap EM (BEM), and Random Forest (RF). The objective is to determine the most accurate and efficient approach for modeling missing climate data by analyzing four key performance metrics: Variance, Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Raw Bias.

2. Methodology

Data mining extracts patterns across diverse fields but requires high-quality data, as missing values distort results and necessitate robust imputation methods [3, 6, 9]. Advanced machine learning, deep learning, and Explainable AI (XAI) capture non-linear interactions without strict assumptions, supporting applications like healthcare diagnostics [3, 8, 9, 12, 14, 17]. Based on MCAR, MAR, and MNAR taxonomies, improper handling of missingness reduces sample sizes and introduces systematic bias (Casanellas-Bernat et al., 2025; Farhangfar et al., 2008]. While vital for environmental planning, rain gauge records regularly suffer from spatial and temporal gaps (Casanellas-Bernat et al., 2025]. in Nigeria, this is exacerbated by constrained monitoring infrastructure [16].To reconstruct missing precipitation records, various statistical and machine learning frameworks are deployed and evaluated using performance metrics like RMSE, MAE, Bias, MSE, and variance [6, 10, 11]:

2.1. Expectation Maximization (Em) Method

The Expectation-Maximization (EM) algorithm is a two-step iterative optimization framework designed to compute maximum likelihood parameter estimates and impute missing data simultaneously. Widely used for incomplete datasets or hidden-variable models, the algorithm repeats its core phases until statistical convergence is reached. The primary operational step is structured as follows:

Expectation (E) Step: Evaluates the conditional expectation of the missing data given the observed data and current parameter estimates. In practice, the algorithm calculates the most likely values to ”fill in” the missing observation gaps based on the active model parameters.

Maximization (M) Step:

Following the E-step's missing data estimation, the M-step updates model parameters via maximum likelihood estimation, treating the dataset as complete. The EM algorithm alternates between these steps until convergence. While effective for MCAR and MAR data, it can be computationally intensive on large datasets. Additionally, because it ignores imputation uncertainty, it may underestimate standard errors and compromise inferential validity when missingness is substantial. Let the rain fall data vector be represented as:

$Y = \ (Y_{obs}\ ,\ Y_{mis})$ (1)

Where:

Y_{obs} is the observed rainfall values

Y_{mis} is the missing rainfall values

2.1.1. Likelihood Function

Let $\theta = (\mu, \Sigma)$ (2)

denote the unknown parameter vector. The complete data likelihood function is defined as:

$L\ (\theta|Y) = p\,(Y_{obs}\ ,\quad Y_{mis}|\theta)$ (3)

The corresponding log-likelihood is

$l\,(\theta|Y) = \log p\,(Y_{obs},\ Y_{mis}|\theta)$ (4)

2.1.2. Expectation Step (E-Step)

In the E-Step, the expected value of thr complete data log-likelihood is computed conditional on the observed data and current parameter estimates $\theta^{(t)}$

$Q\left(\theta \,\middle|\, \theta^{(t)}\right) = E\left[l\,(\theta|Y)\ \mid Y_{obs},\quad \theta^{(t)}\ \right]$ (5)

That is,

$Q\left(\theta \,\middle|\, \theta^{(t)}\right) = E\left[\log f\,(Y_{obs}, Y_{mis} \mid \theta) \mid Y_{obs}, \theta^{(t)}\right]$ (6)

2.1.3. Maximization Step (M- step)

In the M-Step, updated parameter estimates are obtained by maximizing the expected log-likelihood:

$\theta^{(t+1)} = \text{argmax}_{\theta}\ Q\left(\theta \,\middle|\, \theta^{(t)}\right)$ (7)

2.1.4. Iterative Algorithm

The EM algorithm alternates between the E-step and M-step until convergence:

$\theta^{(0)} \rightarrow\ \theta^{(1)} \rightarrow\ \theta^{(2)} \rightarrow \cdots \rightarrow\ \theta^{(k)}$

$\|\theta^{(k+1)} - \theta^{(k)}\| < \varepsilon$ (8)

where ε is a predetermined convergence tolerance.

2.2. Multiple Imputation (Mi)

Multiple Imputation (MI) is a simulation-based approach that accounts for missing-data uncertainty by replacing each absent value with several plausible values to generate a distinct, completed datasets. The framework executes across three continuous phases, beginning with imputation, where missing entries are filled using independent draws from predictive distributions. Next, during the analysis phase, each completed dataset is evaluated separately using standard statistical workflows. Finally, the pooling phase combines these individual parameter estimates and standard errors using Rubin's rules to produce a single, unified inference that explicitly propagates the structural uncertainty of the missingness process.

The pooling procedure is as follows:

Let $\hat{\theta} = \frac{1}{m}\ \sum_{i=1}^{m} \hat{\theta}_i$ (9)

The total variance is computed as:

$T\ = \overline{U} + \left(1 + \frac{1}{m}\right)\ B$ (10)

where:

$\overline{U} = \frac{1}{m}\ \sum_{i=1}^{m} U_i,\ B =\ \frac{1}{m-1}\ \sum_{i=1}^{m} \left(\hat{\theta}_i - \overline{\theta}\right)^2$ (11)

2.3. Regression Imputation

Regression imputation estimates missing values by fitting a regression model where the incomplete variable serves as the dependent variable and the completely observed variables act as predictors. The process executes sequentially by identifying the target variables with missing entries and selecting the appropriate auxiliary predictors. A regression model is then fitted exclusively on the complete data observations. Finally, the resulting model equations are used to predict the missing values, substituting these calculated estimates directly back into the dataset to reconstruct a complete profile.

This model specification is as follows:

$Y_i =\ \ \beta_0 +\ \beta_1\ X_1 + \ldots +\ \beta_k\ X_k + \varepsilon$ (12)

Y_i = missing rainfall

X_1, X_2 = predictor variables

$\beta_0, \beta_{1,}, \beta_k$ = regression coefficient

ε = error term

2.4. Bootstrap Expectation maximization

The Bootstrap Expectation–Maximization (BEM) framework combines bootstrap resampling with the EM algorithm to stabilize parameter estimation and refine confidence intervals. The process operates by drawing multiple bootstrap resamples with replacement from the incomplete dataset, executing the EM algorithm independently on each resample, and pooling the resulting estimates into a unified empirical sampling distribution. The formal model specification is defined as:

$\hat{\theta}^{(b)} = \text{EM}\left(X^{(b)}\right)$ (13)

$\hat{\theta} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^{(b)}$ (14)

where X is missing values

$X^{(b)} = b^{th}$ bootstrap sample (15)

B is number of bootstraps resamples, θ is parameter estimated, $\hat{\theta}^{(b)}$ is EM estimate from bootstrap sample b and $\hat{\theta}$ is final combined estimate.

2.5. Random Forest Imputation (MissForest)

Random Forest Imputation (RFI) is an ensemble learning method that handles non-linear relationships and high-dimensional interactions without strict distributional assumptions. The algorithm trains parallel regression trees on bootstrapped data samples, predicting missing values sequentially across variables until the completed dataset reaches structural convergence. The formal model specification for a target variable Y predicted by auxiliary features X across a forest ensemble is defined as:

$\widehat{Y} = \frac{1}{T}\sum_{t=1}^{T} f_t(X)$ (17)

where T = number of trees in the forest, X = dataset with predictors and missing values.

$f_t(X)$ = prediction from tree t, $\widehat{Y}$ = final imputed rainfall value.

To evaluate and compare the performance of the competing imputation techniques, this study employs the following explicit error and performance metrics:

Raw Bias (RB). $= \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)$ (18)

Variance $= \frac{\sum_{i=1}^{n}(Y_i-\overline{Y})^2}{n-1}$ (19)

Mean Square Error (MSE) MSE $= \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2$ (20)

Root Mean Square (RMSE) $= \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2}$ (21)

3. Results and Analysis

In this section, we apply the five selected approaches: Expectation–Maximization (EM), Multiple Imputation (MI), Regression Imputation, Bootstrap Expectation–Maximization (BEM), and Random Forest Imputation (RFI), to reconstruct the incomplete precipitation records. The analysis utilizes the 2019 Nigerian rainfall dataset, which contains missing observational values across nine (9) distinct states.

Table 3.1: Reported in Nigeria bureau abstract 2019

State	2014	2015	2016	2017	2018
Abia	2161	2268	1980	2220	2439
Adamawa	873	930	718	917	858
Akwa-Ibom	2411	2232	1912	2281	2487
Anambra	1545	1790	2273	1920	1935
Bauchi	1054	965	1621	1072	1111
Bayelsa	2444	2543		2595	2625
Benue	1189	1046	1402	1092	1210
Borno	669	540	588	597	562
Cross-River	2986	2761	2522	2744	2599
Delta	1859	1681	1766	1890	1871
Ebonyi	1230	1349		1949	1963
Edo	2226	2096	2123	2131	
Ekiti	2309	1490		1471	1409
Enugu	1839	1610	1757	1667	1834
Gombe	787	879	857	913	314
Imo	2386	2117	2738	2196	2345
Jigawa	670	1051		771	758
Kaduna	1189	1118	1268	1221	1309
Kano	981	931		926	1000
Katsina	537	562	474	576	559
Kebbi	668	738	1196	772	758
Kogi	1169	1053	1632	1108	1229
Kwara	1191	1128	1352	1308	1287
Lagos	1722	1401	1392	1362	1520
Nasarawa	1120	1179	1566	1248	1228
Niger	1185	1157	1422	1179	1196
Ogun	1500	1353	1466	1295	1626
Ondo	1424	1430	1310	1434	1383
Osun	1341	1336	1278	1333	1309
Oyo	1406	1190	1702	1243	1328
Plateau	1244	1270	1237	1214	1268
Rivers	2478	2212	2602	2227	2251
Sokoto	617	669	603	659	626
Taraba			1569	1134	1411
Yobe	667		367	773	759
Zamfara	884	921	1006	924	891
FCT	1474	1464	1445	1389	1481

Source: National Bureau of Statistics 2019

3.1. Regression Approach

The estimated values of regression approach are shown in Table 3.2.

Table 3.2: Result from the application of R.I approach where missing observation exist

State	2014	2015	2016	2017	2018
Abia	2161	2268	1980	2220	2439
Adamawa	873	930	718	917	858
Akwa-Ibom	2411	2232	1912	2281	2487
Anambra	1545	1790	2273	1920	1935
Bauchi	1054	965	1621	1072	1111
Bayelsa	2444	2543	**2503**	2595	2625
Benue	1189	1046	1402	1092	1210
Borno	669	540	588	597	562
Cross-River	2986	2761	2522	2744	2599
Delta	1859	1681	1766	1890	1871
Ebonyi	1230	1349	**2723**	1949	1963
Edo	2226	2096	2123	2131	**2195**
Ekiti	2309	1490	**1719**	1471	1409
Enugu	1839	1610	1757	1667	1834
Gombe	787	879	857	913	314
Imo	2386	2117	2738	2196	2345
Jigawa	670	1051	**309**	771	758
Kaduna	1189	1118	1268	1221	1309
Kano	981	931	**1040**	926	1000
Katsina	537	562	474	576	559
Kebbi	668	738	1196	772	758
Kogi	1169	1053	1632	1108	1229
Kwara	1191	1128	1352	1308	1287
Lagos	1722	1401	1392	1362	1520
Nasarawa	1120	1179	1566	1248	1228
Niger	1185	1157	1422	1179	1196
Ogun	1500	1353	1466	1295	1626
Ondo	1424	1430	1310	1434	1383
Osun	1341	1336	1278	1333	1309
Oyo	1406	1190	1702	1243	1328
Plateau	1244	1270	1237	1214	1268
Rivers	2478	2212	2602	2227	2251
Sokoto	617	669	603	659	626
Taraba	**1201**	**1070**	1569	1134	1411
Yobe	667	**861**	367	773	759
Zamfara	884	921	1006	924	891
FCT	1474	1464	1445	1389	1481

Table 3.2 presents the results of the regression imputation method applied to the Nigerian

2019 rainfall dataset. The method uses observed variables to predict missing rainfall values based on a fitted linear regression model, providing single deterministic estimates for the missing observations. Bold values indicate the imputed values.

3.2. Bootstrap-EM Approach

The estimated values of Bootstrap-EM Approach are shown in Table 3.3

Table 3.3: Result from the application of bootstrap-EM approach where missing observation exist

State	2014	2015	2016	2017	2018
Abia	2161	2268	1980	2220	2439
Adamawa	873	930	718	917	858
Akwa-Ibom	2411	2232	1912	2281	2487
Anambra	1545	1790	2273	1920	1935
Bauchi	1054	965	1621	1072	1111
Bayelsa	2444	2543	**2445**	2595	2625
Benue	1189	1046	1402	1092	1210
Borno	669	540	588	597	562
Cross-River	2986	2761	2522	2744	2599
Delta	1859	1681	1766	1890	1871
Ebonyi	1230	1349	**2570**	1949	1963
Edo	2226	2096	2123	2131	**2286**
Ekiti	2309	1490	**1674**	1471	1409
Enugu	1839	1610	1757	1667	1834
Gombe	787	879	857	913	314
Imo	2386	2117	2738	2196	2345
Jigawa	670	1051	**492**	771	758
Kaduna	1189	1118	1268	1221	1309
Kano	981	931	**1105**	926	1000
Katsina	537	562	474	576	559
Kebbi	668	738	1196	772	758
Kogi	1169	1053	1632	1108	1229
Kwara	1191	1128	1352	1308	1287
Lagos	1722	1401	1392	1362	1520
Nasarawa	1120	1179	1566	1248	1228
Niger	1185	1157	1422	1179	1196
Ogun	1500	1353	1466	1295	1626
Ondo	1424	1430	1310	1434	1383
Osun	1341	1336	1278	1333	1309
Oyo	1406	1190	1702	1243	1328
Plateau	1244	1270	1237	1214	1268
Rivers	2478	2212	2602	2227	2251
Sokoto	617	669	603	659	626
Taraba	**1273**	**1143**	1569	1134	1411
Yobe	667	**824**	367	773	759
Zamfara	884	921	1006	924	891

State	2014	2015	2016	2017	2018
FCT	1474	1464	1445	1389	1481

Table 3.3 presents the results of the Bootstrap EM method applied to the Nigerian 2019 rainfall dataset. The method combines the EM algorithm with bootstrap resampling to improve the estimation of variability by generating a distribution of parameter estimates across repeated samples. Bold values indicate the imputed Bootstrap EM values.

3.3. Multiple Imputation

The estimated values of multiple imputation approach are shown in Table 4.4

Table 3.4: Result from the application of MI approach where missing observation exist

State	2014	2015	2016	2017	2018
Abia	2161	2268	1980	2220	2439
Adamawa	873	930	718	917	858
Akwa-Ibom	2411	2232	1912	2281	2487
Anambra	1545	1790	2273	1920	1935
Bauchi	1054	965	1621	1072	1111
Bayelsa	2444	2543	**1980**	2595	2625
Benue	1189	1046	1402	1092	1210
Borno	669	540	588	597	562
Cross-River	2986	2761	2522	2744	2599
Delta	1859	1681	1766	1890	1871
Ebonyi	1230	1349	**1912**	1949	1963
Edo	2226	2096	2123	2131	**2345**
Ekiti	2309	1490	**2123**	1471	1409
Enugu	1839	1610	1757	1667	1834
Gombe	787	879	857	913	314
Imo	2386	2117	2738	2196	2345
Jigawa	670	1051	**603**	771	758
Kaduna	1189	1118	1268	1221	1309
Kano	981	931	**367**	926	1000
Katsina	537	562	474	576	559
Kebbi	668	738	1196	772	758
Kogi	1169	1053	1632	1108	1229
Kwara	1191	1128	1352	1308	1287
Lagos	1722	1401	1392	1362	1520
Nasarawa	1120	1179	1566	1248	1228
Niger	1185	1157	1422	1179	1196
Ogun	1500	1353	1466	1295	1626
Ondo	1424	1430	1310	1434	1383
Osun	1341	1336	1278	1333	1309
Oyo	1406	1190	1702	1243	1328
Plateau	1244	1270	1237	1214	1268
Rivers	2478	2212	2602	2227	2251

State	2014	2015	2016	2017	2018
Sokoto	617	669	603	659	626
Taraba	**1185**	**1179**	1569	1134	1411
Yobe	667	**738**	367	773	759
Zamfara	884	921	1006	924	891
FCT	1474	1464	1445	1389	1481

Table 3.4 presents the results of the Multiple Imputation (MI) technique applied to the dataset. The method generates multiple complete datasets by replacing missing values with plausible estimates, analyses them independently, and combines the results using Rubin's rules. Bold values indicate the rainfall observations imputed using the MI approach.

3.4. Expectation Maximization Approach

The estimated values of expectation maximization approach are shown in Table 3.5

Table 3.5: result from the application of EM approach where missing observation exist

State	2014	2015	2016	2017	2018
Abia	2161	2268	1980	2220	2439
Adamawa	873	930	718	917	858
Akwa-Ibom	2411	2232	1912	2281	2487
Anambra	1545	1790	2273	1920	1935
Bauchi	1054	965	1621	1072	1111
Bayelsa	2444	2543	**2767**	2595	2625
Benue	1189	1046	1402	1092	1210
Borno	669	540	588	597	562
Cross-River	2986	2761	2522	2744	2599
Delta	1859	1681	1766	1890	1871
Ebonyi	1230	1349	**3139**	1949	1963
Edo	2226	2096	2123	2131	**2213**
Ekiti	2309	1490	**1169**	1471	1409
Enugu	1839	1610	1757	1667	1834
Gombe	787	879	857	913	314
Imo	2386	2117	2738	2196	2345
Jigawa	670	1051	**180**	771	758
Kaduna	1189	1118	1268	1221	1309
Kano	981	931	**1307**	926	1000
Katsina	537	562	474	576	559
Kebbi	668	738	1196	772	758
Kogi	1169	1053	1632	1108	1229
Kwara	1191	1128	1352	1308	1287
Lagos	1722	1401	1392	1362	1520
Nasarawa	1120	1179	1566	1248	1228
Niger	1185	1157	1422	1179	1196
Ogun	1500	1353	1466	1295	1626

State	2014	2015	2016	2017	2018
Ondo	1424	1430	1310	1434	1383
Osun	1341	1336	1278	1333	1309
Oyo	1406	1190	1702	1243	1328
Plateau	1244	1270	1237	1214	1268
Rivers	2478	2212	2602	2227	2251
Sokoto	617	669	603	659	626
Taraba	**687**	**1005**	1569	1134	1411
Yobe	667	**829**	367	773	759
Zamfara	884	921	1006	924	891
FCT	1474	1464	1445	1389	1481

Table 3.5 presents the results of the Expectation-Maximization (EM) technique applied to the dataset. The method iteratively estimates missing rainfall values and updates model parameters until convergence. Bold values indicate the imputed rainfall observations obtained using the EM approach.

3.5. Random Forest

The estimated values of random forest approach are shown in Table 3.6

Table 3.6: Result from the application of R.F approach where missing observation exist

State	2014	2015	2016	2017	2018
Abia	2161	2268	1980	2220	2439
Adamawa	873	930	718	917	858
Akwa-Ibom	2411	2232	1912	2281	2487
Anambra	1545	1790	2273	1920	1935
Bauchi	1054	965	1621	1072	1111
Bayelsa	2444	2543	**2272**	2595	2625
Benue	1189	1046	1402	1092	1210
Borno	669	540	588	597	562
Cross-River	2986	2761	2522	2744	2599
Delta	1859	1681	1766	1890	1871
Ebonyi	1230	1349	**1899**	1949	1963
Edo	2226	2096	2123	2131	**2267**
Ekiti	2309	1490	**1590**	1471	1409
Enugu	1839	1610	1757	1667	1834
Gombe	787	879	857	913	314
Imo	2386	2117	2738	2196	2345
Jigawa	670	1051	**946**	771	758
Kaduna	1189	1118	1268	1221	1309
Kano	981	931	**990**	926	1000
Katsina	537	562	474	576	559
Kebbi	668	738	1196	772	758
Kogi	1169	1053	1632	1108	1229
Kwara	1191	1128	1352	1308	1287

State	2014	2015	2016	2017	2018
Lagos	1722	1401	1392	1362	1520
Nasarawa	1120	1179	1566	1248	1228
Niger	1185	1157	1422	1179	1196
Ogun	1500	1353	1466	1295	1626
Ondo	1424	1430	1310	1434	1383
Osun	1341	1336	1278	1333	1309
Oyo	1406	1190	1702	1243	1328
Plateau	1244	1270	1237	1214	1268
Rivers	2478	2212	2602	2227	2251
Sokoto	617	669	603	659	626
Taraba	**1375**	**1266**	1569	1134	1411
Yobe	667	**791**	367	773	759
Zamfara	884	921	1006	924	891
FCT	1474	1464	1445	1389	1481

Table 3.6 presents the results of the Random Forest imputation method applied to the rainfall dataset. The method uses an ensemble of decision trees to capture nonlinear and complex relationships in the observed data and predict missing values. Bold values indicate the imputed rainfall observations obtained using the Random Forest approach.

3.6. Evaluation Criteria of EM, RI, MI, BT-EM, and RF Approach for the Rainfall Data

Table 3.7: Error estimate using variance, mean squared error, root mean squared error and bias

Method	Raw Bias	MSE	RMSE	Variance
EM	0.0049	1.0699	1.0344	1.0757
MI	-0.003	1.0113	1.0056	1.0168
EMB	1.0986	0.9840	0.9920	0.9844
REG	0.6769	0.9950	0.9975	0.9453
RF	0.4417	0.9530	0.9762	0.9524

The imputation methods were evaluated using Raw Bias, Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Variance**.** Multiple Imputation (MI) and Expectation-Maximization (EM) recorded the lowest raw bias, indicating the least systematic error, while Bootstrap EM showed the highest bias. In terms of predictive accuracy, Random Forest (RF) achieved the lowest MSE (0.9530) and RMSE (0.9762), making it the most accurate method, whereas EM recorded the highest error values. Regarding stability, Bootstrap EM had the lowest variance, indicating the most consistent estimates, while EM was the least stable. Overall, MI was the least biased, RF was the most accurate, and Bootstrap EM was the most stable, suggesting that the choice of method depends on the objective of the analysis.

4. Conclusion

This study evaluated the empirical performance of five statistical and machine learning imputation techniques: Expectation-Maximization (EM), Multiple Imputation (MI), Regression Imputation (RI), Bootstrap Expectation-Maximization (BEM), and Random Forest (RF) using the 2019 National Bureau of Statistics (NBS) Nigerian rainfall dataset across nine states. The comparative analysis demonstrated that no single imputation architecture is universally superior, as performance varies depending on the targeted statistical priority. Multiple Imputation (MI) proved to be the most effective framework for eliminating systematic estimation error, yielding the lowest raw bias (-0.003). For applications prioritizing high predictive precision and minimized error bounds, the non-parametric Random Forest (RF) algorithm outperformed traditional statistical paradigms by achieving the lowest Mean Squared Error (0.9530) and Root Mean Squared Error (0.9762). Although the Bootstrap EM (BEM) framework offered the highest algorithmic stability with a minimum variance of 0.9844, its elevated raw bias limits its utility for general climate applications. These empirical insights provide a clear data-driven roadmap for agencies like the Nigerian Meteorological Agency (NiMet) to transition away from ad-hoc data removal, thereby strengthening the predictive validity of national climate databases, hydrological modeling, and evidence-based agricultural planning across sub-Saharan ecosystems.

References

[1] Afkanpour, M., Hosseinzadeh, E., & Tabesh, H. (2024). Identify the most appropriate imputation method for handling missing values in clinical structured datasets: A systematic review. *BMC Medical Research Methodology*, *24*(1), Article 188.

[2] Alayo, B. (2024). Missing data: Causes, types, and handling techniques. LinkedIn Pulse.

[3] Aman, A., & Chhillar, R. S. (2020). Data mining techniques in healthcare for disease prediction: A review. *International Journal of Computer Applications*, *175*(23), 28–35.

[4] Başakın, E. E., Ekmekcioğlu, Ö., & Özger, M. (2023). Providing a comprehensive understanding of missing data imputation processes in evapotranspiration-related research: A systematic literature review. *Hydrological Sciences Journal*, *68*(15), 2089–2104.

[5] Bordoloi, R., Réda, C., Trautmann, O., Bej, S., & Wolkenhauer, O. (2025). Handling missing data in downstream tasks with distribution-preserving guarantees. *Journal of Neuroinflammation*, *22*(1), 45–58.

[6] Casanellas-Bernat, A., Julia, C., & Camprodon, J. (2025). Performance of missing data imputation methods in daily precipitation time series across different climate zones. *Journal of Hydrology*, *642*, 131850.

[7] Farhangfar, A., Kurgan, L., & Dy, J. (2008). Impact of imputation of missing values on classification error for discrete data. *Pattern Recognition*, *41*(12), 3692–3705.

[8] García-González, L., Fernández-López, A., & Rodríguez-Díaz, A. (2023). Explainable

artificial intelligence in tabular data processing and imputation: A systematic survey. *Information Fusion*, *98*, 101842.

[9] Golchian, P., Kapar, J., Watson, D. S., & Wright, M. N. (2025). Missing value imputation with adversarial random forests – MissARF. arXiv.

[10] Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J., & Kolehmainen, M. (2004). Methods for imputation of missing values in air quality data sets. *Atmospheric Environment*, *38*(18), 2895–2907.

[11] Kim, J. W., & Pachepsky, Y. A. (2010). Reconstructing missing daily precipitation data using regression trees and time series analysis. *Journal of Hydrology*, *394*(3–4), 305–312.

[12] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

[13] Little, R. J., & Rubin, D. B. (2019). Statistical analysis with missing data (3rd ed.). John Wiley & Sons.

[14] Malan, L., Smuts, C. M., Baumgartner, J., & Ricci, C. (2020). Missing data imputation via the expectation-maximization algorithm can improve principal component analysis aimed at deriving biomarker profiles and dietary patterns. *Nutrition Research*, *75*, 67–76.

[15] Moradi, M., & Mazoochi, M. (2025). A method to information retrieval from missing data using data mining techniques and genetic algorithm. *Journal of Statistical Technology and Information Management*, *13*(2), 2092–2105.

[16] Ogbeide, E. M., Shuaibu, M., & Siloko, U. I. (2023). Theory and application of two (2) iterative imputation approaches to Nigeria annual rainfall data reported. *African Journal of Mathematical Sciences and Statistics*, *6*(3), 1–14.

[17] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135–1144). ACM.

[18] Siddiqui, F. H. (2024). Handling missing data in big data analytics (Technical Report No. BD-2024-03). Scribd Materials. scribd.com

[19] Yu, Q., Umpierrez, G., Davis, G., Dong, X., & Peng, L. (2026). Accurate prediction tool for time in range using statistical learning on inpatient continuous glucose monitoring with missing data. *Diabetes*, *75*(Suppl. 1), 1932