algorithms

MDPI

Article

Machine-Learning-Based Imputation Method for Filling Missing Values in Ground Meteorological Observation Data

Cong Li [1,2,*], Xupeng Ren [1] and Guohui Zhao [2]

[1] School of Computer and Communication, LanZhou University of Technology, LanZhou 730050, China; renxp@lut.edu.cn
[2] National Cryosphere Desert Date Center, Northwest Institute of Eco-Environment and Resources, Chinese Academy of Sciences, Lanzhou 730000, China; zhgh@lzb.ac.cn
* Correspondence: shuimulicong@lut.edu.cn

Abstract: Ground meteorological observation data (GMOD) are the core of research on earth-related disciplines and an important reference for societal production and life. Unfortunately, due to operational issues or equipment failures, missing values may occur in GMOD. Hence, the imputation of missing data is a prevalent issue during the pre-processing of GMOD. Although a large number of machine-learning methods have been applied to the field of meteorological missing value imputation and have achieved good results, they are usually aimed at specific meteorological elements, and few studies discuss imputation when multiple elements are randomly missing in the dataset. This paper designed a machine-learning-based multidimensional meteorological data imputation framework (MMDIF), which can use the predictions of machine-learning methods to impute the GMOD with random missing values in multiple attributes, and tested the effectiveness of 20 machine-learning methods on imputing missing values within 124 meteorological stations across six different climatic regions based on the MMDIF. The results show that MMDIF-RF was the most effective missing value imputation method; it is better than other methods for imputing 11 types of hourly meteorological elements. Although this paper applied MMDIF to the imputation of missing values in meteorological data, the method can also provide guidance for dataset reconstruction in other industries.

Keywords: meteorological data; missing value imputation; machine learning; reconstruction

Citation: Li, C.; Ren, X.; Zhao, G. Machine-Learning-Based Imputation Method for Filling Missing Values in Ground Meteorological Observation Data. *Algorithms* **2023**, *16*, 422. https://doi.org/10.3390/a16090422

Academic Editors: Danilo Pelusi and Asit Kumar Das

Received: 31 July 2023
Revised: 30 August 2023
Accepted: 31 August 2023
Published: 2 September 2023

1. Introduction

Ground meteorological observation data (GMOD) include multiple elements such as temperature, humidity, wind speed, wind direction, air pressure, precipitation, cloud cover, visibility, etc. These elements reflect meteorological conditions and changes on the earth's surface and are the basis of research on earth-related disciplines. For example, GMOD are inevitably used in the process of weather forecasting [1,2] or climate change analysis [3–5]. Additionally, they are an important reference for societal production and life and can be used for disaster warning [6–8], agriculture [9,10], forestry [11–13], tourism [14,15], marine fisheries [16,17], water conservancy [18,19], transportation [20,21], and other fields. Effective use of GMOD may help stakeholders, including the government, to avoid environmental problems and damages, providing a wide range of social benefits.

To meet different service requirements and functions, weather stations can use various sensors to monitor different environmental factors and transmit data within minutes or even seconds. However, there are inevitable problems such as operational errors, sensor failures, network transmission failures, and storage failures that can cause missing values within the meteorological observation dataset. Consequently, adopting rational techniques to handle missing data during the data preprocessing phase is essential for ensuring the integrity and dependability of the data.

The most direct method for missing value processing is deletion [22], which means deleting records with missing values or outliers. The advantage of this method is that it is

simple to operate, but it may cause information loss, lead to data deviation [23], destroy the continuity of time series, and cause bias or even errors in the analysis results [24].

Imputing missing values is an effective way to make data complete and to avoid analysis errors or the inability to perform analysis due to missing values. Of course, imputing missing values also has certain limitations and risks, such as introducing some biases or errors, resulting in inaccurate or unreliable data analysis results. Therefore, when choosing the method of imputing missing values, it is necessary to choose the appropriate method according to the type, distribution, missing mechanism, and other factors of the data. The simplest imputation methods include the mean imputation [25] (replacing missing values with the average of non-missing values in the attribute), mode imputation [26] (replacing missing values with the most frequent value in the attribute), and median imputation [27] (replacing missing values with the middle value in the sorted data). However, these methods tend to reduce the overall variance of the imputed dataset and cause a large amount of data homogenization. Another common and simple imputation method is hot-deck imputation, which replaces missing data with values from existing data that meet a set of rules. The advantage of this method is that it uses the relationship between data to estimate missing values, but the disadvantage is that the rule setting is more subjective. The above two types of imputation methods are often used as benchmark methods for comparison with other imputation methods [28,29].

Another frequently used technique for estimating missing values is multiple imputation. It has a better imputation effect than traditional methods such as mean imputation and median imputation [30,31], but the imputation process may fail. Cattram [32] summarized the causes of imputation failure, including perfect prediction and collinearity.

The imputation method based on machine learning can adapt to any missing patterns and has good robustness and small deviation. It also interpolates complex high-dimensional data better than mean, hot-deck, and multiple imputation methods [33]. This method learns the rules in the data through relevant algorithms and uses them to interpolate the data. In the field of meteorological missing value imputation, this method has three basic forms of application.

The first form is to use machine-learning algorithms directly. Some researchers have confirmed the validity of some machine-learning imputation methods. For instance, random forest (RF) [34] and multi-layer perceptron neural networks (MLP) [35] have the ability to accurately reconstruct meteorological variables, particularly temperature. In the research by Taewon [36], two statistical imputation methods (linear and spline) and three machine-learning methods (multivariate linear regression (MLR), RF, and MLP) interpolated the missing values of seven meteorological variables, respectively. The findings indicate that machine-learning imputation methods are more applicable than statistical imputation methods for handling missing data in both the short and long term. Moreover, MLP exhibited the highest level of accuracy across all experiments.

The second form is based on the improvement of the existing algorithm. Bo [37] designed an auto-encoder architecture with a convolution layer to reconstruct missing wind speed data. Compared with the six traditional data reconstruction methods, the designed network had the lowest reconstruction error.

The third version combines various machine-learning techniques. For example, Jinhua [38] used long short-term memory networks (LSTM) and support vector regression (SVR) to interpolate missing wind pressure data. Samal's model [39] is composed of temporal convolutional networks (TCN) and convolutional neural networks (CNN), which are used to estimate the missing value of PM2.5. The experimental results indicated that the combined model has a higher imputation accuracy than the single model. Multi-model combination can also be used to interpolate some meteorological parameters that are difficult to recover, such as precipitation. Precipitation data are a special challenge for the recovery of missing data, because these data are random and highly unbalanced in the duration of rainfall and non-rainfall. Benedict [40] proposed an imputation model of rainfall that was classified and then predicted. Gradient tree boosting, K-nearest neighbors (KNN), RF,

support vector machines (SVM), and neural networks were tested via classification and prediction algorithms, respectively. The test results showed that a combination of RF and neural networks is superior to surface fitting technology in recovering missing precipitation data at a 30 min resolution.

By reviewing the existing literature, it can be seen that the imputation method based on machine learning has advantages in the field of reconstruction of ground weather station missing values, but there are also some problems. First, much of the literature tends to focus on the imputation of a certain meteorological element, but when multiple parameters (such as temperature, humidity, wind speed, etc.) in the dataset have random missing values at the same time, using the existing data in the dataset to fill these missing values is a problem that is rarely discussed in the literature. Second, there are many machine-learning methods, but their verification in the field of meteorological data imputation is insufficient. Most machine-learning method types are verified in [39], but only eight methods are verified overall. Third, in addition to Joseph's extensive tests on imputation algorithms of 134 weather stations [34], a limited number of studies have assessed the performance of machine-learning meteorological imputation methods in multiple locations or examined the effectiveness of different algorithms in various climatic regions. Finally, some comparisons are unfair, such as comparing a combined model with a single model, comparing a fine-tuned model with an untuned model, and running the algorithm with carefully selected and processed datasets. Therefore, within the existing literature, it is hard to find an answer to the question of: "Which machine learning model can most accurately impute missing meteorological data?" or "Under what conditions is method x suitable?"

In this study, a multidimensional meteorological data imputation framework (MMDIF) that uses machine-learning predictions to reconstruct missing meteorological data was designed. Based on MMDIF, the imputation performance of 20 typical machine-learning methods for the missing values of 124 pieces of ground meteorological observation data in six climatic regions was verified. Each method was automatically individually tuned to guarantee optimal performance.

2. Data

The data used in this paper were sourced from the National Water and Climate Center (NRCS), which is affiliated with the U.S. Department of Agriculture. The agency has 226 automatic monitoring stations set up across the United States to monitor hourly changes in soil, atmosphere, precipitation, wind, and other elements in real time, and it has built the largest database of soil, water resource, and climate data in the United States. This study collected hourly meteorological data of 124 stations from 2010 to 2019, including 11 meteorological elements, namely, air temperature (TOBS), maximum air temperature (TMAX), minimum air temperature (TMIN), maximum wind speed (WSPDX), average wind speed (WSPDV), solar radiation (SRADV), dew point temperature (DPTP), air pressure (PVPV), relative humidity (RHUM), minimum humidity (RHUMN), and maximum humidity (RHUMX). According to the Köppen climate classification index, we divided the collected station data into six types of climate types, with each type represented by three letters. The main climate condition is represented by the first letter, which can be one of the following: (A) equatorial climate, (B) arid climate, (C) warm climate, (D) snow climate, or (E) polar climate. The second and third letters indicate precipitation and temperature conditions, respectively. For example, BWh indicates that the area is located in a dry and rainless desert climate. To obtain further information on the climate classification, please refer to [41]. Figure 1 illustrates the distribution of the 124 stations studied in the experiment.

This paper performed outlier detection on the experimental data and deleted the data items that contained outliers to avoid the influence of outliers on some machine-learning methods that rely on certain assumptions and calculations, such as linear regression, logistic regression, support vector machine, etc.

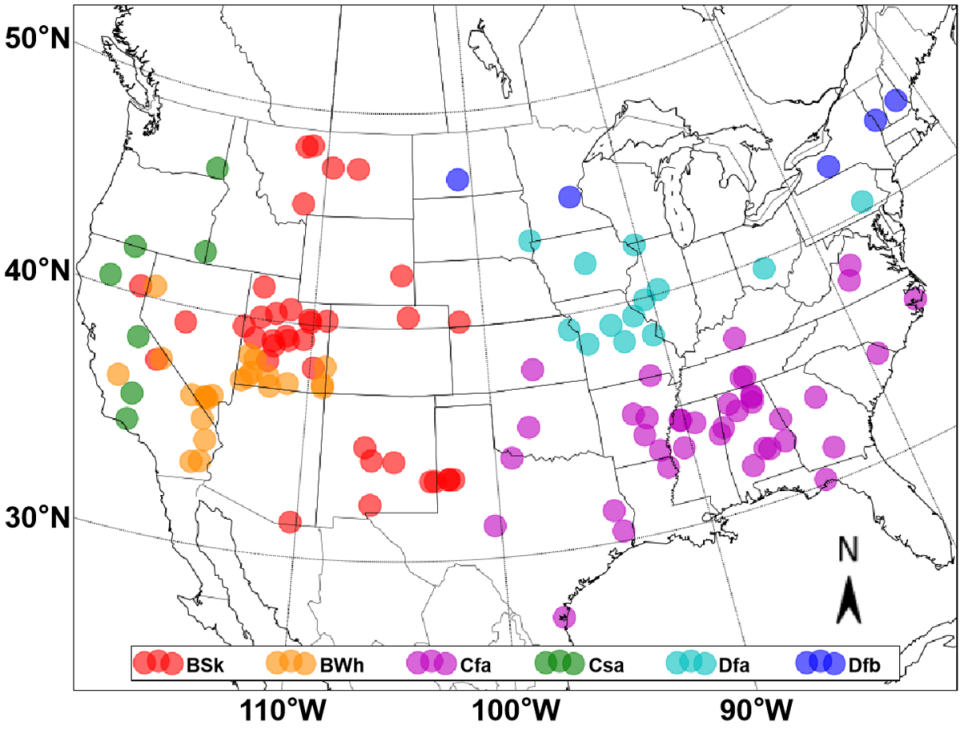

Figure 1. Map of climatic zoning site.

When imputation is applied directly to the original data, the extent of discrepancy between the imputation result and the true value cannot be assessed due to the unavailability of the true value for the missing data. This paper selected the non-missing records of all stations, randomly deleted some units according to the proportion, recorded the deleted part as the true value, and then imputed data according to the proposed framework, so that the difference between the imputed value and the true value could be evaluated. The missing value rate in this study was set at 10%, 20%, 40%, 60%, and 80%.

3. Methodology

This paper outlines the designed multidimensional meteorological data imputation framework (MMDIF) based on machine learning, which uses the predicted values of machine-learning models as the imputed values of missing data. The non-missing data in the feature to be imputed are used as the label data for training the model, and the features in the dataset that have a correlation with the feature to be imputed are used as the input data for the model. It is assumed that the dataset contains N observation values (such as temperature, humidity, wind speed, etc.), which we usually call the features of the data. Each feature has a certain percentage of randomly missing values. The MMDIF designed in this paper is as follows:

Step 1: Choose the feature with the least missing values in the dataset as the feature to be filled in, mark it as data″, and mark the remaining features as data′.
Step 2: Divide data′ by day into layers, and replace the missing cells in each layer with the mean of the recorded cells in that layer as temporary imputation values.
Step 3: Take the non-missing data in data″ as the label of machine learning and record the line number z where the missing value is located in data″.
Step 4: Calculate the correlation between each feature and label in data′ except for row z, and keep the features with a correlation greater than 0.2 as the machine-learning trainset (a correlation coefficient below 0.2 means that there is a very weak correlation between the two variables [42]). The correlation analysis uses the spearman correlation coefficient,

which can reflect the degree of correlation between the two variables, x and y, based on Equation (1), where n represents the sample size, and $\overline{x}$ and $\overline{y}$ represent the sample mean:

$$r = \frac{\sum_i^n (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i^n (x_i - \overline{x})^2 \sum_i^n (y_i - \overline{y})^2}} \quad (1)$$

Step 5: Select the features consistent with trainset in data", and select the z line as the testset.
Step 6: Train the machine learning model using the label and trainset created in Steps 3 and 4.
Step 7: Input the testset into the model trained in Step 6, and impute the output values of the model to the corresponding positions in data.
Step 8: Check the data. If there are missing values, return to Step 1 and continue the imputation program; otherwise, exit the program.

The MMDIF imputation process is illustrated in Figure 2. The machine-learning module depicted in the figure employs a variety of prediction mechanisms, including 20 mainstream machine-learning methods. These methods encompass linear, tree-based, instance-based, kernel-based, probability-based, and neural network-based methods. Table 1 provides a brief overview of these methods. The neural network methods are implemented using the Keras and Tensorflow modules, and the remaining methods are implemented using the scikit-learn modules. The implementation process requires setting hyperparameters for each model, and the choice of model hyperparameters greatly affects the performance of the model. In order for each model to achieve the best possible performance, in this paper, the most advantageous hyperparameter combination for the model was sought out via a grid search of the variable space [43]. Because the weight matrix initialization greatly influences the performance of neural network-based methods, we repeated these algorithms ten times on the same data and computed the mean of the results. This allowed us to compare them fairly with other machine-learning methods.

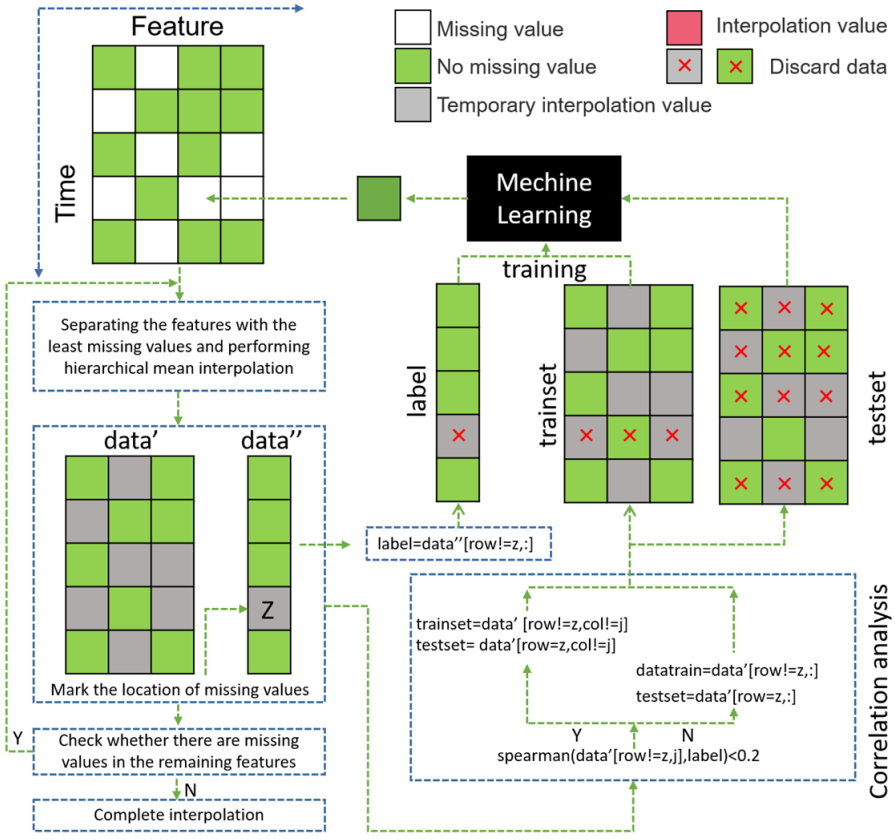

Figure 2. MMDIF flowchart.

Table 1. Typical machine-learning regression method.

Method	Short Description
Linear regression	Multivariate linear regression (MLR), ridge regression (Ridge), lasso regression (Lasso), and ElasticNet regression (ENet). These methods are used to establish the relationship between independent and dependent variables by fitting the data through minimizing the sum of squared residuals. The difference between these four methods is that they improve the generalization ability of the model by adding different types of regularization terms.
Probability-based	Bayesian ridge regression (BR) and automatic relevance determination regression (ARD) are based on Bayesian linear regression. ARD is a linear regression model that is solved using Bayesian inference in statistics, assuming that the prior distribution of the regression coefficients is an elliptical Gaussian distribution parallel to the coordinate axis, whereas BR assumes that the prior distribution of the regression coefficients is a spherical normal distribution.
Instance-based	K-nearest neighbor (KNN). For a given test sample, based on the distance metric, find the K-closest training samples in the training set, and then make predictions based on the information of these K "neighbors".
Tree-based	Decision tree regression (DTR), random forest (RF), adaptive boosting algorithm (AdaBoost), gradient boosting decision tree (GBDT), extremely randomized tree (ERT), bootstrap aggregating (Bagging). DTR is a simple regression algorithm, whereas Bagging, RF, AdaBoost, GBDT, and ERT are ensemble learning algorithms based on decision trees. They improve prediction performance by using different methods to construct and combine decision trees.
Kernel-based	Support vector regression (SVR) uses kernel functions to transform data into higher-dimensional space to simulate nonlinearity.
Neural nework-based	Perceptron, multilayer perceptron neural networks (MLP), recurrent neural networks (RNN), long short-term memory networks (LSTM), bidirectional LSTM networks (BiLSTM), and temporal convolutional networks (TCN) are based on the perceptron, but with differing structures and functions. For example, MLP is a feedforward neural network, and RNN, LSTM, and BiLSTM can all handle sequence data. On the other hand, TCN uses convolution to handle sequence data.

4. Results

This section discusses the imputation performance of machine-learning models for meteorological datasets with missing values from four aspects. First, the imputation accuracy of each model was evaluated under different missing value rates. Second, the imputation performance of each model was contrasted across various climate types. Third, the best missing value imputation method for each observation value was discussed. Fourth, since a long model training time affects real-time tasks, the training time of the imputation models was also discussed. Finally, to validate the efficacy of the MMDIF introduced in this research, we conducted a comparison with algorithms featured in other research papers.

4.1. Model Performance at Different Observations

Since each meteorological station produces observation data containing multiple meteorological elements, and the units of these meteorological elements are not consistent, this section used two dimensionless indicators, the determination coefficient (R^2) and the symmetric mean absolute percentage error (SMAPE), to evaluate the imputation ability of different machine-learning methods for meteorological datasets with missing values. The closer the R^2 is to 1, the more accurately the algorithm fits the data, and the closer the SMAPE is to 0, the smaller and more precise the prediction error of the algorithm is. Their definitions are shown in:

$$R^2 = 1 - \frac{\sum_{l=1}^{m}(y_l - \widehat{y_l})^2}{\sum_{l=1}^{m}(y_l - \overline{y})^2} \tag{2}$$

$$SMAPE = \frac{100\%}{m}\sum_{l=1}^{m}\frac{|\widehat{y_l} - y_l|}{(|\widehat{y_l}| + |y_l|)/2} \tag{3}$$

where m represents the sample size, y_l represents the true value, $\widehat{y_l}$ represents the model output value, and $\overline{y}$ represents the mean value of attribute y.

Table 2 displays the average results of 20 machine-learning methods used to impute missing values in 124 meteorological observation datasets under different missing value rates within the MMDIF. According to the results recorded in the table, the data missing rate is the primary factor that impacts the accuracy of imputation. Regardless of whether it is based on the R^2 or SMAPE, all methods will achieve the best performance of the method at a low missing rate (10%). The reason for this is that when the missing value rate is small, the missing items have a limited influence on the overall distribution of the data samples. As the missing rate increases gradually, the original distribution characteristics of the data are altered by the growing number of missing values, which leads to a decline in the imputation method's performance with the higher missing rate. Among all methods, the RF is the most reliable imputation method. Based on the SMAPE, the RF performs better than other algorithms in all imputation experiments with all missing value rates presented in this paper. Meanwhile, based on the R^2, the RF performs better than other methods when the missing rate is 10%, 20%, and 40%, and the RF's R^2 is slightly lower than the TCN only when the missing rate is 60% and 80%.

Table 2. The imputation accuracy of various algorithms under different missing rates.

Method	Missing Rate (R^2 | SMAPE)				
	10%	**20%**	**40%**	**60%**	**80%**
AdaBoost	0.84 | 9.17	0.81 | 9.90	0.73 | 11.51	0.65 | 13.07	0.57 | 14.54
Perceptron	0.90 | 6.14	0.87 | 6.95	0.82 | 8.46	0.75 | 9.97	0.68 | 11.57
ARD	0.87 | 7.49	0.84 | 8.33	0.78 | 9.75	0.71 | 11.21	0.63 | 12.89
Bagging	0.91 | 5.05	0.89 | 5.57	0.84 | 6.60	0.79 | 7.90	0.72 | 9.48
BR	0.87 | 7.48	0.84 | 8.33	0.78 | 9.75	0.71 | 11.22	0.63 | 12.89
BiLSTM	0.91 | 5.42	0.89 | 6.17	0.84 | 7.52	0.77 | 9.10	0.70 | 10.80
MLP	0.90 | 6.16	0.87 | 6.97	0.81 | 8.46	0.75 | 9.98	0.68 | 11.59
DTR	0.85 | 7.80	0.81 | 8.62	0.74 | 9.99	0.65 | 11.48	0.53 | 13.02
ENet	0.87 | 7.62	0.84 | 8.46	0.78 | 9.82	0.71 | 11.26	0.63 | 12.83
ERT	0.84 | 5.98	0.80 | 6.70	0.73 | 8.11	0.64 | 9.75	0.51 | 11.56
GBDT	0.90 | 5.97	0.87 | 6.70	0.82 | 8.07	0.77 | 9.43	0.70 | 10.85
KNN	0.89 | 6.71	0.86 | 7.53	0.80 | 8.82	0.75 | 10.08	0.67 | 11.42
Lasso	0.87 | 7.62	0.84 | 8.46	0.78 | 9.82	0.71 | 11.26	0.63 | 12.83
MLR	0.88 | 7.46	0.84 | 8.32	0.78 | 9.74	0.71 | 11.21	0.63 | 12.89
LSTM	0.90 | 5.80	0.88 | 6.47	0.82 | 7.92	0.76 | 9.55	0.68 | 11.21
RF	0.92 | 4.90	0.90 | 5.39	0.86 | 6.36	0.81 | 7.59	0.74 | 9.13
Ridge	0.87 | 7.47	0.84 | 8.33	0.78 | 9.75	0.71 | 11.23	0.63 | 12.90
RNN	0.90 | 6.31	0.87 | 7.14	0.81 | 8.62	0.75 | 10.16	0.67 | 11.76
SVR	0.90 | 6.67	0.88 | 7.17	0.83 | 8.36	0.77 | 9.76	0.69 | 11.32
TCN	0.86 | 7.51	0.86 | 7.56	0.84 | 8.13	0.82 | 9.10	0.76 | 10.23

Figure 3 shows the imputation results of the MMDIF combined with RF on all sites. Each scatter point in the figure represents the imputation result of MMDIF-RF on a dataset. The distribution of scatter points in the figure reflects the robustness of the algorithm on different datasets, and the more densely the data points are distributed in the vertical direction, the more stable the algorithm is. RF performed very well at a low data missing rate of 10%, with the R^2 above 0.9 for 116 stations (93.5% of the total) and the SMAPE below 5% for 91 stations (73.4% of the total). As the data missing rate increases to 20%, the R^2 remains above 0.9 for 85 stations (68.5% of the total), and the SMAPE stays below 5% for 65 stations (52.4% of the total). However, at a high data missing rate of 60%, RF's R^2 drops significantly to around 0.8 for most stations (97 out of 124), and the SMAPE rises to an average of 7.59%, with a large variation among stations. At an extremely high data missing rate of 80%, both the R^2 and SMAPE become sparse and unstable, indicating poor data imputation results.

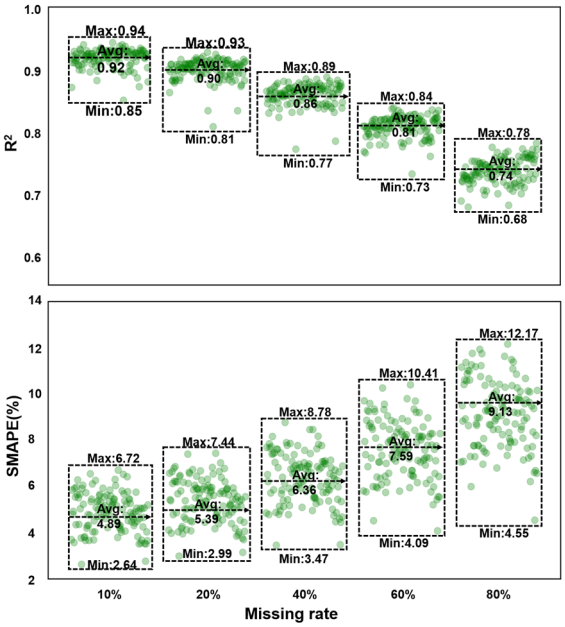

Figure 3. Distribution of the imputation results of RF.

Neural network-based techniques fall into the second category. Of these methods, BiLSTM performs well. This may be related to its special structure, which can capture the time series characteristics in the data, and this may enable it to cope with missing data [44].

The results of the six linear regression methods (MLR, Ridge, Lasso, Enet, BR, and ARD) are very close on R^2 and SMAPE, and there is no obvious difference in their performance. They have a weak ability to impute missing data, because these algorithms are all based on linear assumption models, whereas GMOD have nonlinear or complex patterns, which causes the models to not fit or predict the data well. We noticed that ERT's R^2 was the lowest at all missing rates, suggesting that ERT is not suitable for imputing missing values in GMOD. This may lead to a decrease in the correlation and distribution between the imputed data and the original data.

As the missing rate increases, the prediction accuracy of all algorithms decreases, but the degree of decrease varies. This paper used Equations (4) and (5) to calculate the decrease in the imputation accuracy of each algorithm when the missing rate rose from 10% to 80%, where $R^2_{10\%}$ and $R^2_{80\%}$, $SMAPE_{10\%}$ and $SMAPE_{80\%}$ represent the R^2 and SMAPE obtained using one of the methods when the missing rate is 10% and 80%, respectively. The results are shown in Figure 4.

$$R^2 \ decline \ range = \left(R^2_{10\%} - R^2_{80\%}\right) / R^2_{10\%} \tag{4}$$

$$SMAPE \ decline \ range = (SMAPE_{80\%} - SMAPE_{10\%}) / SMAPE_{10\%} \tag{5}$$

In general, there was a certain positive correlation between the decrease in R^2 and the decrease in SMAPE, that is, the larger the decrease in R^2, the larger the decrease in SMAPE. However, Figure 4 shows that neural network-based algorithms (such as Perceptron, BiLSTM, MLP, LSTM, RNN, etc.) have smaller R^2 decreases and larger SMAPE decreases than other types of algorithms, such as linear regression, probability-based algorithms, and tree-based methods. Such a difference suggests that neural network algorithms perform better on the R^2 metric but worse on the SMAPE metric. A possible explanation for this could be that neural network algorithms can capture the complex features and nonlinear relationships in the data more effectively [45], which improves the R^2 value, but they also tend to overfit the noise or outliers in the data, which lowers the SMAPE value.

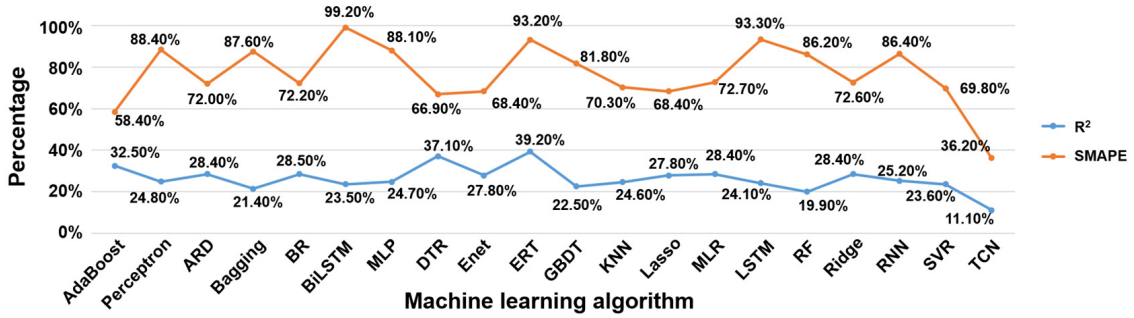

Figure 4. The degree of decline in imputation accuracy of each algorithm when the missing rate increases from 10% to 80%.

As shown previously, RF is less accurate than TCN, except for at 60% and 80% missing rates, but more accurate at 10%, 20% and 40%. However, Figure 4 reveals that the R^2 drops by 19.9%, and the SMAPE rises by 86.2% for RF, whereas TCN only sees a 11.1% R^2 decline and a 36.2% SMAPE increase as the missing rate grows from 10% to 80%. This means TCN is more robust for missing rate changes, which is crucial in practice, as we cannot control or predict a missing rate. A sensitive algorithm would require frequent parameter tuning or replacement, increasing our work and cost. An insensitive algorithm can reliably handle any missing data without affecting the imputation quality.

In addition to RF, five other tree-based methods (DTR, AdaBoost, GBDT, ERT, Bagging) saw the R^2 drop by 21.4%–39.2% and the SMAPE rise by 58.4%–93.2%, with ERT exhibiting the worst performance for both metrics, showing its low stability and accuracy for missing data imputation. The kernel-based (SVR) and instance-based (KNN) methods have similar imputation accuracy declines as the missing rate increases. Their R^2 falls by around 24% and their SMAPE climbs by around 70% when the missing rate goes from 10% to 80%. They interpolated well with increasingly low missing data but performed poorly on increasingly high missing data.

4.2. Model Performance under Different Climate Zones

Figures 5 and 6 show the best imputation methods for each climate type and the number of times they achieved the best imputation results based on R^2 and SMAPE criteria, respectively. The figures indicate that RF emerged as the most dependable imputation method. It achieved higher R^2 than others in most sites across all climate zones for 10–40% missing rates. For instance, in the BSk zone with 37 sites, RF beat the other 19 methods at 33 (89.1% of BSk), 35 (94.6% of BSk), and 35 (94.6% of BSk) sites for 10%, 20%, and 40% missing rates, respectively. However, TCN became more competitive with higher missing rates. In the BSk zone, TCN outperformed others at 29 (78.4%) sites for an 80% missing rate, whereas RF only achieved this at 7 (18.9%) sites. Using the SMAPE as the criterion, RF dominated other methods across all climate types and missing rates. BiLSTM also performed well with missing rates below 60% but struggled with higher ones. Other algorithms such as Bagging, LSTM, GBDT, KNN, DTR, SVR, and ERT excelled in some cases but lacked consistency overall.

The average performance of each algorithm in each climate zone was calculated, and the results are given in Figure 7. In the Cfa, Dfa, and Dfb climate zones, the R^2 of all algorithms was generally higher than those in the BSk, BWh, and Csa climate zones, and according to the SMAPE indicator, the algorithms performed better in the Dfb, Dfa, and Csa climate zones. Considering both indicators, all algorithms performed poorly in the BWh and BSk climate zones. When comparing various climate types, it can be observed that the BSk (semi-arid steppe climate) and BWh (desert climate) experience greater levels of evaporation than precipitation, with the precipitation being mainly concentrated in either summer or winter. Therefore, the distribution of observation data related to precipitation, such as the relative humidity (RHUM), minimum humidity (RHUMN), maximum humidity

(RHUMX), and dew point temperature (DPTP), is not uniform across the dataset. As a result, the data used to train the model and the data to be interpolated will exhibit a significant difference in data distribution. This should be the main reason for the poor performance of the imputation model under the BWh and BSk climate conditions.

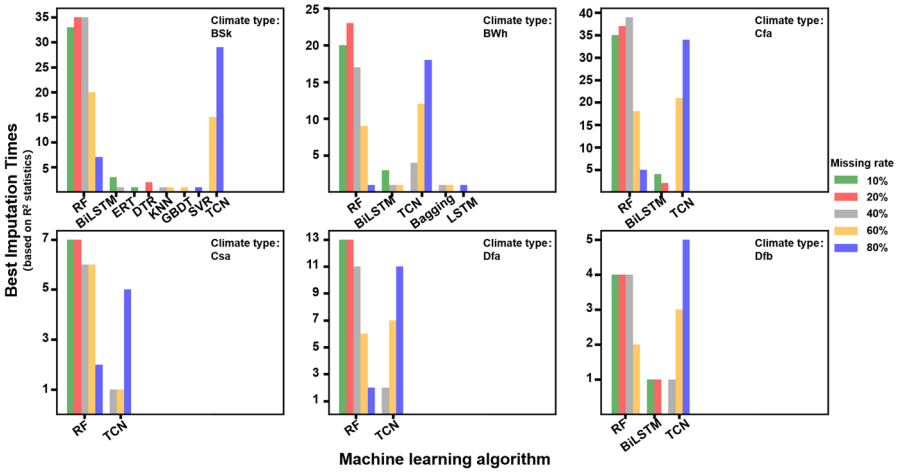

Figure 5. Best imputation methods for each climate category and their statistics (based on R^2).

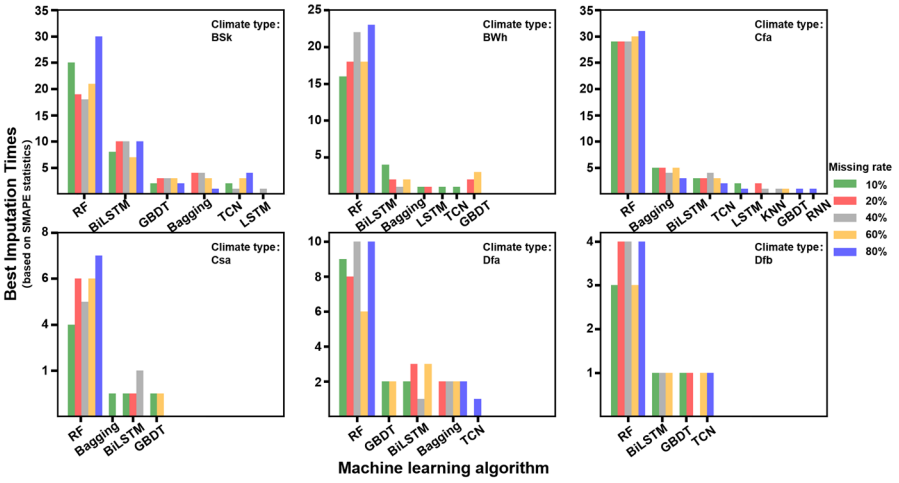

Figure 6. Best imputation methods for each climate category and their statistics (based on SMAPE).

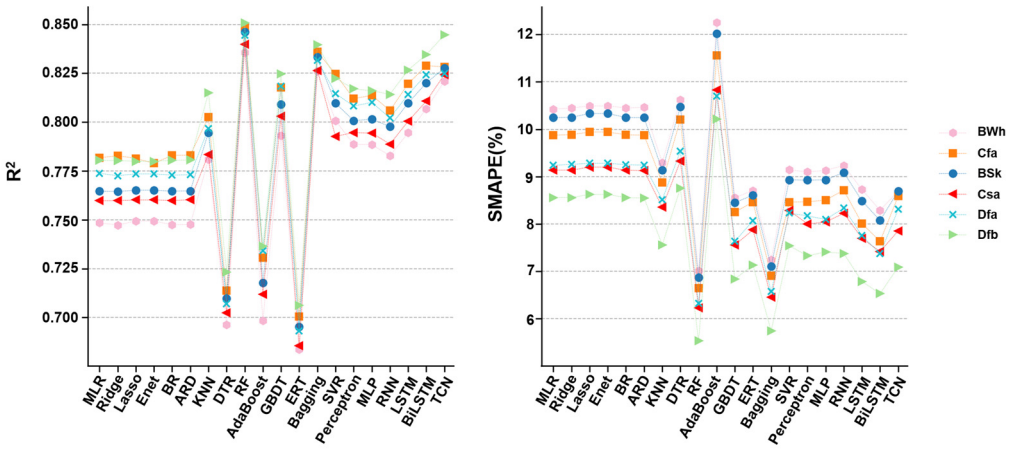

Figure 7. Performance of each algorithm in different climate types.

4.3. Model Performance for Each Observation Element

Table 3 displays the optimal and least effective imputation techniques for 11 meteorological observation elements. The first line of each cell in the table records the best method, and the second line records the worst method. The RMSE, MAE, R^2, and SMAPE were the average values of the imputation results of the imputation algorithm at five missing value rates. The definitions of the RMSE and MAE are shown in Equations (6) and (7), respectively, and the definitions of the R^2 and SMAPE are given in Section 4.1.

$$RMSE = \sqrt{\frac{1}{m}\sum_{l=1}^{m}(y_l - \hat{y}_l)^2} \tag{6}$$

$$MAE = \frac{1}{m}\sum_{l=1}^{m}|(y_l - \hat{y}_l)| \tag{7}$$

According to the records in Table 3, the RF is undoubtedly the most ideal method for imputing missing values. For the 11 meteorological observations presented in this paper, the imputation results of the RF could maintain the highest R^2 and the lowest RMSE, MAE, and SMAPE. AdaBoost, ERT, Enet, BR, ARD, SVR, and TCN had the worst records, especially AdaBoost, which is not ideal for imputing missing values of the temperature (TOBS, TMIN, and TMAX) and humidity (RHUM, RHUMN, and RHUMX) data. The 20 imputation methods used in this paper performed poorly on the WSPDX, WSPDV, and SRADV, which can be observed very clearly with the R^2 and SMAPE, two dimensionless indicators. The R^2 of RF and TCN, which they perform the best, only reached 0.48–0.66 for these three observations. In the future, it is worth paying attention to the imputation of missing values for wind speed and solar radiation.

Table 3. The best and worst imputation methods for each meteorological observation element.

Elements	RMSE		MAE		R^2		SMAPE	
TOBS	RF	1.61 °C	RF	0.93 °C	RF	0.96	RF	5.51
	AdaBoost	3.51 °C	AdaBoost	2.64 °C	Enet	0.78	AdaBoost	11.71
TMIN	RF	1.73 °C	RF	1.04 °C	RF	0.96	RF	6.62
	AdaBoost	3.50 °C	AdaBoost	2.62 °C	AdaBoost	0.87	AdaBoost	12.38
TMAX	RF	1.64 °C	RF	0.89 °C	RF	0.96	RF	5.25
	AdaBoost	3.10 °C	AdaBoost	2.26 °C	AdaBoost	0.89	AdaBoost	10.74
WSPDX	RF	2.95 Mph	RF	2.09 Mph	TCN	0.66	RF	6.47
	ERT	4.10 Mph	AdaBoost	3.01 Mph	ERT	0.29	AdaBoost	11.5
WSPDV	RF	1.88 Mph	RF	1.32 Mph	TCN	0.64	RF	7.21
	ERT	2.62 Mph	ERT	1.83 Mph	ERT	0.28	AdaBoost	12.86
SRADV	RF	194.76 W/m^2	RF	155.79 W/m^2	RF	0.48	RF	15.53
	ERT	278.71 W/m^2	ERT	213.15 W/m^2	ERT	0.07	ERT	17.97
DPTP	RF	2.09 °C	RF	0.75 °C	RF	0.94	RF	12.47
	SVR	7.13 °C	SVR	5.75 °C	AdaBoost	0.75	AdaBoost	19.14
PVPV	RF	0.25 KPa	RF	0.09 KPa	RF	0.93	RF	4.38
	TCN	0.66 KPa	SVR	0.46 KPa	Enet	0.50	BR	6.43
RHUM	RF	5.03%	RF	2.99%	RF	0.93	RF	3.73
	AdaBoost	8.61%	AdaBoost	6.41%	Enet	0.82	AdaBoost	8.46
RHUMN	RF	5.66%	RF	2.89%	RF	0.92	RF	2.68
	AdaBoost	10.18%	AdaBoost	7.21%	Enet	0.75	AdaBoost	9.01
RHUMX	RF	5.34%	RF	3.27%	RF	0.92	RF	2.75
	AdaBoost	9.45%	AdaBoost	7.24%	Enet	0.76	AdaBoost	8.72

Figure 8 shows the imputation results for each observation variable at the 2008 meteorological observation site. Each subplot draws the best and worst imputation results for one observation variable. As mentioned in Section 4.1, the data missing rate is the main factor affecting the data imputation results. When the data missing rate reaches 80%, there is a significant difference between the imputation results and the original data. When the

missing rate is less than 60%, the RF imputation results for temperature (TBOS, TMIN, TMAX, and DPTP), humidity (RHUM, RHUMN, and RHUMX), and air pressure (PVPV) are highly consistent with the original data.

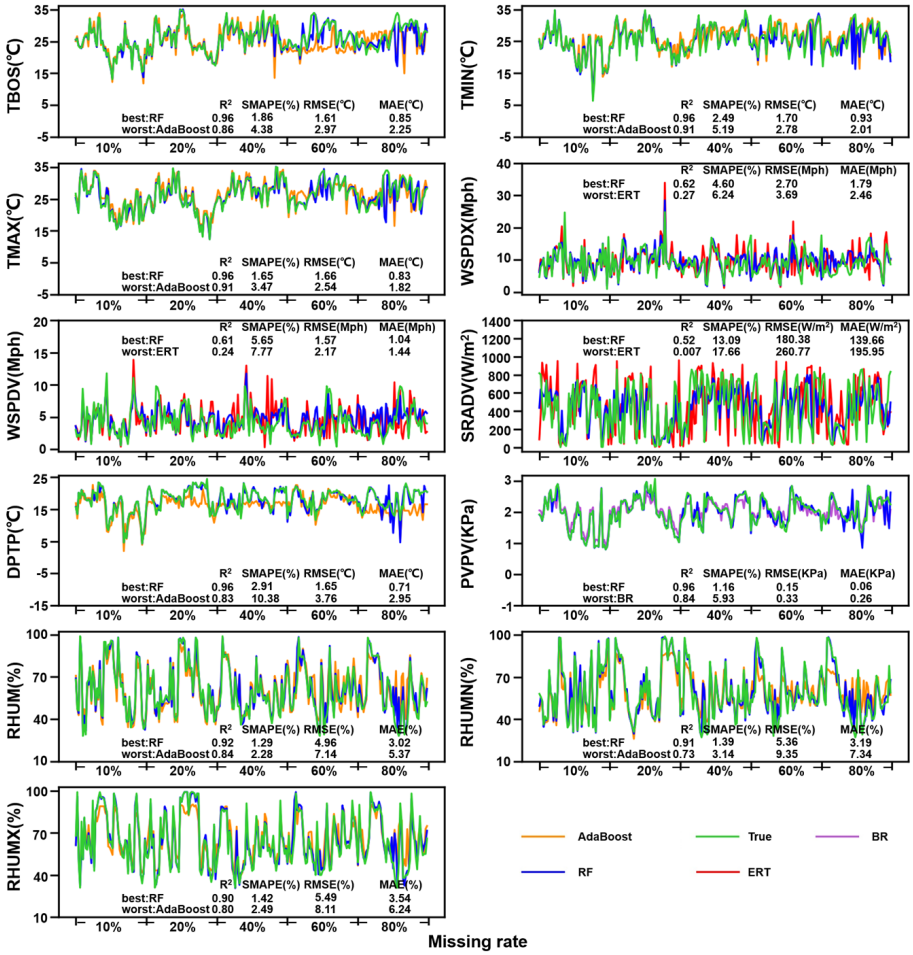

Figure 8. The best and worst imputation results at the 2008 meteorological observation site.

4.4. Training Duration of the Model

Table 4 records the training duration of all methods. The hardware environment for model training is Intel i7-8700 (CPU), 64 G (RAM), 1050 Ti (GPU). The six linear regression methods outlined in this paper had similar imputation effects, but their training time varied greatly, with MLR being the fastest (less than 1 s) and ARD being the slowest (more than 5 min on average).

In both the tree-based categories, the RF was the slowest, taking an average training time of nearly 1 min. However, it had the best imputation effect among all algorithms. A method that is faster than the RF and has a similar imputation effect was Bagging, taking 8 s as an average training time. Its R^2 is 1% lower than that of RF, and its SMAPE is 3% higher than that of the RF.

As seen from previous analysis, BiLSTM, LSTM, and TCN are three neural network imputation methods with a better performance. However, due to their structure, LSTM and BiLSTM can only read and parse one record at a time, and they have to process each record sequentially. Hence, LSTM and BiLSTM take a long time to train. In our experiment, the longest training time of LSTM was 4680.28 s, and that of BiLSTM was 6061.58 s, which is too slow for data imputation tasks at the hourly level. TCN is a novel structure that applies convolutional networks to time series learning tasks, solving the problem of LSTM and BiLSTM being hard to parallelize and significantly reducing the training time.

Table 4. The training time of each algorithm.

Method	Duration(s)					
	Missing Rate 10%	Missing Rate 20%	Missing Rate 40%	Missing Rate 60%	Missing Rate 80%	Average Duration
MLR	0.395	0.396	0.421	0.405	0.331	0.389
Ridge	167.428	153.703	141.661	117.065	90.136	133.999
Lasso	153.925	146.388	143.651	127.361	106.294	135.524
Enet	153.351	147.633	144.169	127.264	107.077	135.899
BR	273.696	265.083	261.950	230.349	193.824	244.980
ARD	382.077	380.746	404.102	377.901	322.452	373.456
KNN	1.844	2.508	3.846	4.373	4.377	3.390
DTR	0.969	0.923	0.884	0.771	0.636	0.837
RF	70.141	67.158	64.987	55.351	45.508	60.629
AdaBoost	2.090	2.730	3.867	4.155	3.460	3.260
GBDT	21.040	19.391	18.618	16.150	12.622	17.564
ERT	0.576	0.574	0.614	0.556	0.471	0.558
Bagging	8.836	8.597	8.632	7.599	6.277	7.988
SVR	43.255	44.390	46.883	38.067	27.519	40.023
Perceptron	127.659	116.002	107.112	93.163	78.459	104.479
MLP	145.523	131.159	119.722	104.000	89.925	118.066
RNN	432.557	382.919	349.538	281.223	231.763	335.600
LSTM	1235.162	1195.449	1025.659	781.857	609.432	969.512
BiLSTM	1856.816	1683.721	1472.877	1139.001	909.946	1412.472
TCN	446.149	423.100	394.256	345.127	299.772	381.681

4.5. Comparison with Other Studies

Joseph [34] introduced an imputation method based on RF in his paper, which is used to impute the missing 15 min and daily maximum/minimum temperature observations of 134 sites in Washington State over 8.5 years, with a data missing rate of only 3.3%. We compared the imputation results with Joseph's best imputation results when the missing rate was 10%, as shown in Table 5. In the case of a higher missing rate, the MMDIF-RF method used in this paper was more accurate than Joseph's method for imputing the TBOS, TMAX, and TMIN.

Table 5. Comparison of the results of the two studies (°C).

Observation Elements	Joseph's Research		MMDIF-RF	
	RMSE	MAE	RMSE	MAE
TBOS	0.63	0.43	0.61	0.35
TMAX	0.72	0.53	0.68	0.34
TMIN	0.92	0.70	0.77	0.48

5. Conclusions

Complete and accurate GMOD are crucial for social and economic development and people's daily lives, but data can be lost due to various issues during collection, transmission, and storage. As a result, numerous techniques have been developed for recovering lost data. Machine-learning methods are among the most widely used methods for interpolating meteorological missing values, as they offer high accuracy, simple operation, and other benefits. However, when utilizing machine-learning methods to interpolate meteorological data, the intrinsic relationship among the observation variables has been overlooked by many studies, and only a small number of studies have thoroughly assessed the performance of different machine-learning methods across a wide range of sites.

In this paper, we propose a machine-learning-based multidimensional meteorological data imputation framework (MMDIF), which can use machine-learning predictions to reconstruct GMOD with random missing values across multiple attributes. We used this

framework to compare the missing value imputation performance of 20 machine-learning methods. The data we used were sourced from 10 years of observation data of 124 stations in six climate zones of the continental U.S., including 11 variables. Based on four evaluation indicators for assessing the predictive performance of models, R^2, SMAPE, RMSE, and MAE, we find that RF, BiLSTM, and TCN had the best imputation effects for missing data. They were more accurate than the other regression methods for most stations, especially the RF. The RF performed better than the other methods in interpolating 11 meteorological variables under the six climate types in this paper, especially for TBOS, TMIN, TMAX, DPTP, PVPV, RHUM, RHUMN, and RHUMX. For these eight variables, RF achieves a R^2 above 0.9 and a SMAPE below 9% (R^2 and SMAPE are the average values of the imputation results at five missing rates). The literature [46] has shown that RF uses random sampling, random splitting, and voting mechanisms in modeling and prediction, which can effectively avoid overfitting or underfitting of the model, improve the accuracy and robustness of the model, and have a better performance than deep learning in dealing with missing data imputation. However, in this study, it was found that none of the 20 imputation methods performed well for wind speed and solar radiation variables (WSPDX, WSPDV, and SRADV). This indicates that the imputation framework proposed in this paper has certain limitations when handling these specific variables. Some of the limitations include modifying the framework structure, incorporating new data (such as the sampling time of the data as an input variable), and handling data separately for different seasons. These needed to be further explored in future research. Although we have tested the MMDIF-RF with GMOD in this study, this method can also be a reference for missing data imputation within other domains.

Author Contributions: Conceptualization, C.L.; methodology, C.L.; software, C.L.; original draft preparation, C.L.; review and editing of manuscript, C.L. and G.Z.; visualization, C.L. and X.R.; data curation, G.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key R&D Program of China (project No. 2022YFF0711700); the project of the School of Computer and Communication, Lanzhou University of Technology, on the research of fine temperature prediction models based on deep learning (project No. H1814cc012); and Light of West China Program of Chinese Academy of Sciences (project No. E2297801).

Data Availability Statement: The data that support the findings of this study are openly available in National Water and Climate Center of the US Department of Agriculture (https://www.nrcs.usda.gov/resources, accessed on 29 August 2023).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Fathi, M.; Haghi Kashani, M.; Jameii, S.M.; Mahdipour, E. Big Data Analytics in Weather Forecasting: A Systematic Review. *Arch. Comput. Methods Eng.* **2021**, *5*, 1247–1275. [CrossRef]
2. Zhou, C.; Li, H.; Yu, C.; Xia, J.; Zhang, P. A station-data-based model residual machine learning method for fine-grained meteorological grid prediction. *Appl. Math. Mech.* **2022**, *43*, 155–166. [CrossRef]
3. Magistrali, I.C.; Delgado, R.C.; dos Santos, G.L.; Pereira, M.G.; de Oliveira, E.C.; Neves, L.D.O.; de Souza, L.P.; Teodoro, P.E.; Junior, C.A.S. Performance of CCCma and GFDL climate models using remote sensing and surface data for the state of Rio de Janeiro-Brazil. *Remote Sens. Appl. Soc. Environ.* **2021**, *21*, 100446. [CrossRef]
4. Sebestyén, V.; Czvetkó, T.; Abonyi, J. The Applicability of Big Data in Climate Change Research: The Importance of System of Systems Thinking. *Front. Environ. Sci.* **2021**, *9*, 70. [CrossRef]
5. Ding, X.; Zhao, Y.; Fan, Y.; Li, Y.; Ge, J. Machine learning-assisted mapping of city-scale air temperature: Using sparse meteorological data for urban climate modeling and adaptation. *Build. Environ.* **2023**, *234*, 110211. [CrossRef]
6. Khan, S.; Kirschbaum, D.; Stanley, T. Investigating the potential of a global precipitation forecast to inform landslide prediction. *Weather. Clim. Extrem.* **2021**, *33*, 100364. [CrossRef]
7. Freitas, A.A.D.; Oda, P.S.S.; Teixeira, D.L.S.; Silva, P.D.N.; Mattos, E.V.; Bastos, I.R.P.; Nery, T.D.; Meetodiev, D.; Santos, A.P.P.d.; Gonçalves, W.A. Meteorological conditions and social impacts associated with natural disaster landslides in the Baixada Santista region from March 2nd–3rd, 2020. *Urban Clim.* **2022**, *42*, 101110. [CrossRef]

8. Zhang, Y.; Wu, Y.; Zhang, F.; Yao, X.; Liu, A.; Tang, L.; Mo, J. Application of power grid wind monitoring data in transmission line accident warning and handling affected by typhoon. *Energy Rep.* **2022**, *8*, 315–323. [CrossRef]
9. Wang, F.; Lai, H.; Li, Y.; Feng, K.; Zhang, Z.; Tian, Q.; Zhu, X.; Yang, H. Dynamic variation of meteorological drought and its relationships with agricultural drought across China. *Agric. Water Manag.* **2021**, *261*, 107301. [CrossRef]
10. Iniyan, S.; Varma, V.A.; Naidu, C.T. Crop yield prediction using machine learning techniques. *Adv. Eng. Softw.* **2023**, *175*, 103326. [CrossRef]
11. Fraccaroli, C.; Govigli, V.M.; Briers, S.; Cerezo, N.P.; Jimenez, J.P.; Romero, M.; Lindner, M.; de Arano, I.M. Climate data for the European forestry sector: From end-user needs to opportunities for climate resilience. *Clim. Serv.* **2021**, *23*, 100247. [CrossRef]
12. Ghafarian, F.; Wieland, R.; Lüttschwager, D.; Nendel, C. Application of extreme gradient boosting and Shapley Additive explanations to predict temperature regimes inside forests from standard open-field meteorological data. *Environ. Model. Softw.* **2022**, *156*, 105466. [CrossRef]
13. Kern, A.; Marjanović, H.; Csóka, G.; Móricz, N.; Pernek, M.; Hirka, A.; Matošević, D.; Paulin, M.; Kovač, G. Detecting the oak lace bug infestation in oak forests using MODIS and meteorological data. *Agric. For. Meteorol.* **2021**, *306*, 108436. [CrossRef]
14. Barnet, A.F.; Ciurana, A.B.; Pozo, J.X.O.; Russo, A.; Coscarelli, R.; Antronico, L.; De Pascale, F.; Saladié, Ò.; Anton-Clavé, S.; Aguilar, E. Climate services for tourism: An applied methodology for user engagement and co-creation in European destinations. *Clim. Serv.* **2021**, *23*, 100249. [CrossRef]
15. Wang, L.; Zhou, X.; Lu, M.; Cui, Z. Impacts of haze weather on tourist arrivals and destination preference: Analysis based on Baidu Index of 73 scenic spots in Beijing, China. *J. Clean. Prod.* **2020**, *273*, 122887. [CrossRef]
16. Cerim, H.; Özdemir, N.; Cremona, F.; Öğlü, B. Effect of changing in weather conditions on Eastern Mediterranean coastal lagoon fishery. *Reg. Stud. Mar. Sci.* **2021**, *48*, 102006. [CrossRef]
17. Amon, D.J.; Palacios-Abrantes, J.; Drazen, J.C.; Lily, H.; Nathan, N.; van der Grient, J.M.A.; McCauley, D. Climate change to drive increasing overlap between Pacific tuna fisheries and emerging deep-sea mining industry. *NPJ Ocean Sustain.* **2023**, *2*, 9. [CrossRef]
18. Jia, D.; Zhou, Y.; He, X.; Xu, N.; Yang, Z.; Song, M. Vertical and horizontal displacements of a reservoir slope due to slope aging effect, rainfall, and reservoir water. *Geod. Geodyn.* **2021**, *16*, 266–278. [CrossRef]
19. Liu, Y.; Shan, F. Global analysis of the correlation and propagation among meteorological, agricultural, surface water, and groundwater droughts. *J. Environ. Manag.* **2023**, *333*, 117460. [CrossRef]
20. Joshua, S.; Yi, L. Effects of extraordinary snowfall on traffic safety. *Accid. Anal. Prev.* **2015**, *81*, 194–203. [CrossRef]
21. Lu, H.P.; Chen, M.Y.; Kuang, W.B. The impacts of abnormal weather and natural disasters on transport and strategies for enhancing ability for disaster prevention and mitigation. *Transp. Policy* **2019**, *98*, 2–9. [CrossRef]
22. Newman, D.A. Missing Data: Five Practical Guidelines. *Organ. Res. Methods* **2014**, *17*, 372–411. [CrossRef]
23. Lokupitiya, R.S.; Lokupitiya, E.; Paustian, K. Comparison of missing value imputation methods for crop yield data. *Environmetrics* **2006**, *17*, 339–349. [CrossRef]
24. Schafer, J.L.; Graham, J.W. Missing data: Our view of the state of the art. *Psychol. Methods* **2002**, *7*, 147–177. [CrossRef]
25. Felix, I.K.; Rian, C.R.; Leon, T. Local mean imputation for handling missing value to provide more accurate facies classification. *Procedia Comput. Sci.* **2023**, *216*, 301–309. [CrossRef]
26. Xu, X.; Xia, L.; Zhang, Q.; Wu, S.; Wu, M.; Liu, H. The ability of different imputation methods for missing values in mental measurement questionnaires. *BMC Med. Res. Methodol.* **2020**, *20*, 42. [CrossRef]
27. Berkelmans, G.F.; Read, S.H.; Gudbjörnsdottir, S.; Wild, S.H.; Franzen, S.; Van Der Graaf, Y.; Eliasson, B.; Visseren, F.L.J.; Paynter, N.P.; Dorresteijn, J.A.N. Population median imputation was noninferior to complex approaches for imputing missing values in cardiovascular prediction models in clinical practice. *J. Clin. Epidemiol.* **2022**, *145*, 70–80. [CrossRef] [PubMed]
28. Vazifehdan, M.; Moattar, M.H.; Jalali, M. A Hybrid Bayesian Network and Tensor Factorization Approach for Missing Value Imputation to Improve Breast Cancer Recurrence Prediction. *J. King Saud. Univ. Comput. Inf. Sci.* **2019**, *31*, 175–184. [CrossRef]
29. Schmitt, P.; Mandel, J.; Guedj, M. A comparison of six methods for missing dataimputation. *J. Biom. Biostat.* **2015**, *6*, 1. [CrossRef]
30. Madan, L.Y.; Basav, R. Handling missing values: A study of popular imputation packages in R. *Knowl.-Based Syst.* **2018**, *160*, 104–118. [CrossRef]
31. Gordana, I.; Tome, E.; Barbara, K.S. Evaluating missing value imputation methods for food composition databases. *Food Chem. Toxicol.* **2020**, *141*, 111368. [CrossRef]
32. Cattram, D.N.; John, B.C.; Katherine, J.J. Practical strategies for handling breakdown of multiple imputation procedures. *Emerg. Themes Epidemiol.* **2021**, *18*, 5. [CrossRef]
33. Jerez, J.M.; Molina, I.; García-Laencina, P.J.; Alba, E.; Ribelles, N.; Martín, M.; Franco, L. Missing Data Imputation Using Statistical and Machine Learning Methods in a Real Breast Cancer Proble. *Artif. Intell. Med.* **2010**, *50*, 105–115. [CrossRef] [PubMed]
34. Joseph, P.B.Z.; David, J.B. Machine learning imputation of missing Mesonet temperature observations. *Comput. Electron. Agric.* **2022**, *192*, 106580. [CrossRef]
35. Franco, B.M.; Hernández-Callejo, L.; Navas-Gracia, L.M. Virtual weather stations for meteorological data estimations. *Neural Comput. Appl.* **2020**, *32*, 12801–12812. [CrossRef]
36. Taewon, M.; Seojung, H.; Jung, E.S. Interpolation of greenhouse environment data using multilayer perceptron. *Comput. Electron. Agric.* **2019**, *166*, 105023. [CrossRef]

37. Jing, B.; Pei, Y.; An, J. Missing wind speed data reconstruction with improved context encoder network. *Energy Rep.* **2022**, *8*, 3386–3394. [CrossRef]
38. Li, J.; Desen Zhu, D.; Li, C. Comparative analysis of BPNN, SVR, LSTM, Random Forest, and LSTM-SVR for conditional simulation of non-Gaussian measured fluctuating wind pressures. *Mech. Syst. Signal Process.* **2022**, *178*, 109285. [CrossRef]
39. Samal, K.; Babu, K.S.; Das, S.K. Multi-directional temporal convolutional artificial neural network for PM2.5 forecasting with missing values: A deep learning approach. *Urban Clim.* **2021**, *36*, 100800. [CrossRef]
40. Benedict, D.C.; John, W.; Georgios, L. Imputation of missing sub-hourly precipitation data in a large sensor network: A machine learning approach. *J. Hydrol.* **2020**, *588*, 125126. [CrossRef]
41. Kottek, M.; Grieser, J.; Beck, C.; Rudolf, B.; Rubel, F. World Map of the Köppen-Geiger climate classification updated. *Meteorol. Z.* **2006**, *15*, 259–263. [CrossRef] [PubMed]
42. Harry, K. Measures of Association: How to Choose? *J. Diagn. Med. Sonogr.* **2008**, *24*, 155–162. [CrossRef]
43. Yagli, G.M.; Yang, D.; Srinivasan, D. Automatic hourly solar forecasting using machine learning models. *Renew. Sustain. Energy Rev.* **2019**, *105*, 487–498. [CrossRef]
44. Ying, H.; Deng, C.; Xu, Z.; Huang, H.; Deng, W.; Yang, Q. Short-term prediction of wind power based on phase space reconstruction and BiLSTM. *Energy Rep.* **2023**, *9*, 474–482. [CrossRef]
45. Sarker, I.H. Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. *SN Comput. Sci.* **2021**, *2*, 420. [CrossRef]
46. Sun, Y.; Li, J.; Xu, Y.; Zhang, T.; Wang, X. Deep learning versus conventional methods for missing data imputation:A review and comparative study. *Expert Syst. Appl.* **2023**, *227*, 120201. [CrossRef]