Advancing Climate Analytics with Enhanced Data Imputation

Team Name: **FIAS**

Savannah Flanagan
srf8585@nyu.edu

Amy Cho
sc8088@nyu.edu

Frederic Dai
qd2044@nyu.edu

Iris Wu
yw5653@nyu.edu

Abstract

Climate change mitigation and prevention efforts are plagued by missing data which greatly reduces the capabilities of important analyses. This phenomenon systematically and disproportionately affects developing nations, as they often lack the infrastructure and support to gather the appropriate amount of data. The International Monetary Fund's 'Annual Surface Temperature Change' dataset is one of the most commonly used climate datasets in the world and is unfortunately plagued by this issue of missing data. In order to encourage more accurate analyses using this database, we explore a variety of machine learning (ML) and non-ML-based imputation methods to attempt to impute this missing data, also investigating how the inclusion of related climate variables can improve imputation accuracy. Our analysis reflected that ML-based methods outperformed non-ML-based methods on the simple temperature change dataset, but those models often performed less accurately with the inclusion of additional climate features. Most models were able to achieve a high imputation accuracy, with the iterative random-forest-based model MissForest achieving the best performance. We successfully applied this optimal model to the IMF dataset to impute its missing values, creating a robust, non-missing dataset for use by future researchers. In future research, we suggest enhancing the MissForest model by optimizing its parameters for better accuracy and developing hybrid models that combine MissForest with time series analysis techniques. These improvements aim to refine predictions for climate datasets with temporal patterns, increasing the model's utility in climate science research.

1. Introduction

Null Data, in other words, missing values, in a dataset indicates the absence of information in certain data fields, which can lead to significant biases in statistical analysis and potentially influence the reliability of resulting conclusions.

In the context of climate change prevention, the integrity and fullness of data are crucial. Accurate and comprehensive datasets enable researchers and policymakers to develop effective strategies for both mitigating and adapting to climate change. The fullness of data enhances the validity of the results of the research and supports evidence-based to shape critical climate policies.

In particular, the climate data organized by the International Monetary Fund (IMF) is one of the most utilized datasets for climate change analyses. The IMF is a global organization that works to achieve sustainable growth and prosperity for all of its 190 member countries [1]. The data organized by the IMF provides a detailed and expansive view of climate-related indicators, allowing for nuanced analysis and informed conclusions across diverse geographic regions and temporal scales.

Unfortunately, even the data gathered by global corporations like the IMF contains missing values. To avoid the issues of containing missing data in the research and risking the decrease of the statistical evidence of the analysis, researchers often delete rows containing missing data [2]. However, this approach can skew climate change mitigation and prevention strategies. Moreover, this strategy will disproportionately affect developing nations that contain more missing values compared to others due to systematic issues in data collection and infrastructure challenges. While some normal and non-machine learning imputation methods do exist to address these missing values, these methods are often implemented without a specific understanding of the material, leading to a potential decrease in accuracy.

To combat these issues, we decided to explore the effectiveness of various machine learning methods and sets of features at imputing missing mean temperature changes from the IMF dataset to fill in missing values as accurately as possible for further analysis.

Our approach starts by intentionally deleting some known values from the dataset to create a synthetically missing dataset on which to test our imputation methods. With this dataset, we were able to test the imputation accuracy of various imputation models across two different feature sets. After comparing these accuracy results from each method, we applied the most accurate model to impute missing data into the actual IMF dataset, creating a dataset free of missing values for future analyses.

2. Related Work

SimpleImputer The 'SimpleImputer' method is an advanced strategy in the sci-kit learn library [3], which aims to provide more nuanced imputations by replacing missing values using a specified constant or utilizing statistical measures like the mean or median of each column containing missing data. While effective for basic application processes, this method might not capture the full complexity and concept needed in diverse datasets such as those involved in climate analysis.

MissingIndicator The 'MissingIndicator' method is an advanced strategy in the sci-kit learn library [3], which aims to help handle missing data by providing a binary representation indicating the presence of missing values in a dataset. This method can be especially useful in datasets where the pattern of missing data itself is informative and can be used as an input feature in predictive modeling, furthermore helps uncover hidden patterns in missing data that were simply imputed or ignored. While effective for datasets where the pattern of missing data is informative, it may not be suitable for accurately imputing missing data, particularly in scenarios like climate data from developing countries. Since in some cases, the absence of data requires more sophisticated approaches to predict or understand the underlying data patterns.

MultipleImputation Multiple Imputation (MI) method developed by Donald Rubin [4] creates multiple sets of plausible values for missing data, reflecting uncertainty about the right value to impute. By generating several complete datasets with these multiple sets, each of which is analyzed using standard statistical methods such as Linear Regression, Logistic Regression, ANOVA (Analysis of Variance), Survival Analysis, and GLM (Generalized Linear Models). The results of these analyses are then pooled to create final estimates and inferences. This method allows the imputation process to account for the uncertainty inherent in predicting missing values. While effective for accounting for the uncertainty inherent in predicting missing values, this method might not be suitable for our research due to its requirement of the missing data. Multiple Imputation requires that the missing data mechanism be properly classified as MCAR (Missing Completely at Random), MAR (Missing at Random), or NMAR (Not Missing at Random), which may not always be possible or accurate in our data context related to climate change. Additionally, the assumption that imputed values are unbiased estimates might not hold if the underlying assumptions about the data and missing mechanisms are violated, potentially leading to erroneous conclusions in our specific research context.

3. Data and Preprocessing

The IMF's climate change data provides indicators related to climate change such as carbon dioxide atmospheric concentrations, trends in global warming, rising sea levels, rising temperatures, and frequency of natural disasters to monitoring climate change and its impacts on populations.

We decided to address null values in the 'Annual Surface Temperature Change' dataset since the annual surface temperature change data is critical in monitoring and analyzing global climate change. To impute missing values effectively, we decided to use the correlated data from the other datasets as explanatory variables. As the correlated data, we decided to use the 'World Monthly Atmospheric Carbon Dioxide Concentrations', 'Forest and Carbon', 'Land Cover and Land Cover Altering Indicator', and 'Climate-related Disasters Frequency' datasets.

We decided to extract data only from the Years 1992 - 2020 from the datasets. Excluding data before 1992 is due to the establishment of the United Nations Framework Convention on Climate Change (UNFCCC) [5]. UNFCC is a United Nations' (UN) international environmental treaty adopted in 1992 to combat climate change by stabilizing greenhouse gas concentrations in the atmosphere at a level that would prevent dangerous anthropogenic interference with the climate system. As the first agreement to specifically focus on climate change as a global issue, the treaty was signed in 1992 and entered into force on March 21, 1994. This policy also influenced the availability of some datasets, which only started in 1992, aligning with the treaty's adoption. Excluding data after 2020 is due to the COVID-19 pandemic. According to the International Monetary Fund [6], the COVID-19 pandemic's unique economic and social impacts altered public perceptions, which could change the further performance of climate data analysis. As more comprehensive post-pandemic datasets become available, there will be more opportunities to input pandemic and post-pandemic data to enhance the robustness of future studies. Focusing on data from and after these crucial years was to ensure consistency and relevance in the context of current climate frameworks and agreements.

The 'Annual Surface Temperature Change' dataset, 'IMF_surface_Temperature.csv', presents the mean surface temperature change during the period 1961-2021, using temperatures between 1951 and 1980 as a baseline. We extracted data containing only the temperature data from Year 1992 - 2020 and the corresponding country as columns.

The 'World Monthly Atmospheric Carbon Dioxide Concentrations' dataset, 'IMF_Atmospheric_CO2.csv', presents the concentration of carbon dioxide in the atmosphere, on a monthly and yearly basis, dating back to 1958. Averaged the calculated atmospheric carbon dioxide concentration data that was calculated monthly by Year. We extracted data containing only the averaged calculated atmospheric carbon dioxide concentration data from the Years 1992 - 2020 and the corresponding country as columns.

The 'Forest and Carbon' dataset, 'IMF_Forest_and_Carbon.csv', presents the calculated value of carbon stocks in the forest(million tonnes), forest area(1000 HA), index of carbon stocks in the forest(Index), index of forest extent(index), and land area(1000 HA) of different countries. We extracted data containing only the rows with "Forest Area" as an Indicator from the Years 1992 - 2020 and the corresponding country as columns.

The 'Land Cover and Land Cover Altering Indicator', 'IMF_Land_Cover.csv', presents the changes in land cover over time, grouping land cover into those types that have climate influencing, climate regulating, and climate neutral impacts. We extracted data containing only the rows with the "Climate Altering Land Cover Index" as an Indicator from the Years 1992 - 2020 and the corresponding country as columns.

The 'Climate-related Disasters Frequency', 'IMF_Climate-Related_Disasters..csv', presents the relationship between climate and natural disasters, showing the trend in climate-related disasters over time. We extracted data containing only the rows with the "Climate-related disasters frequency, Number of Disasters: TOTAL", for overall values, from the Years 1992 - 2020 and the corresponding country as columns. Also imputed 0s for all missing values because the dataset was only marked when and how frequently disasters occurred each year, leaving other files blank.

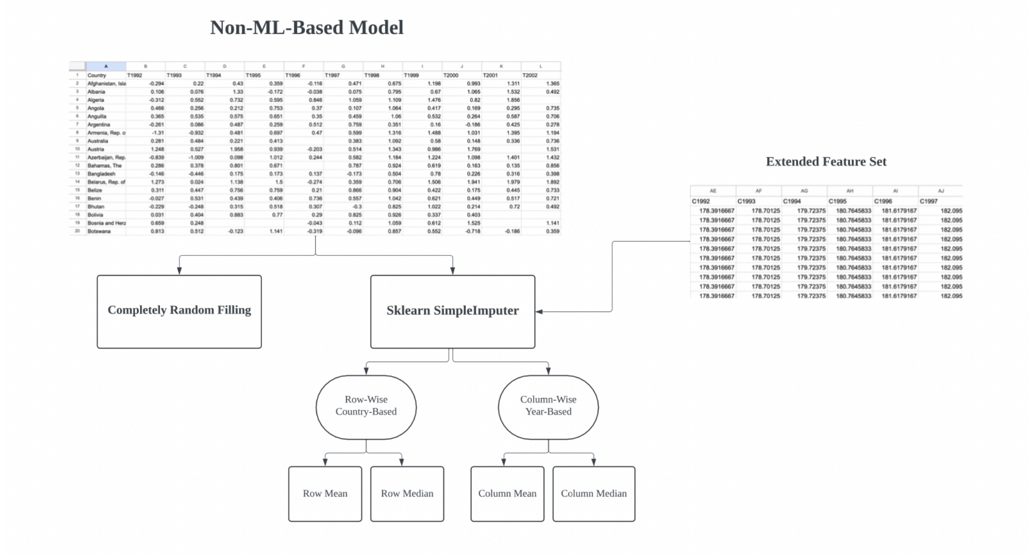
To investigate how the inclusion of additional features improves the accuracy of imputation, we decided to explore two different feature sets. The Temperature Feature Set only consists of the annual surface temperature change data. The Extended Feature Set includes this data along with all the other features we extracted from the additional IMF climate datasets, namely atmospheric carbon dioxide, forest area, land cover, and total natural disasters. We obtained the dataset for this feature set by merging all the IMF datasets mentioned previously, with each column being a certain climate variable for a certain year between 1992 and 2020. Each row contains data for a specific country, with only those included in the original temperature dataset included in the final merged dataset. We removed rows with missing values in this merged dataset to obtain the final dataset from which we could make our dataset for testing the accuracy of various imputation methods. Regardless of the feature set being explored, only the annual surface temperature change across various years was attempted to be imputed, as this is our variable of interest for this study.

In order to impute the most accurate data possible into the annual surface temperature change dataset, we needed to get a sense of each model's accuracy in imputing this data, and for each feature set. To test imputation accuracy, we created synthetic missing datasets by randomly deleting values in the annual surface temperature change columns of both the Temperature and Extended Feature Set databases. Missing values were artificially included in the same proportion as they were in the original dataset to echo the actual scenario as much as possible, meaning that about 5% of values were deleted. We then used these synthetic datasets as inputs to imputation methods so that their missing values could be imputed, and compared these imputed values to the actual values in the dataset. We used both root mean squared error (RMSE) and mean absolute error (MAE) to measure the accuracy of imputation for both feature sets. We chose both these metrics so that we might get a sense of the methods' accuracies both with and without the large influence of outliers so that we might see if outliers are commonly present in some methods and get a sense of how the models perform overall with their influence being less sizable. Both of these metrics can be interpreted in the same units as the original dataset (i.e. temperature change in degrees Celsius).

4. Methodology

Our methodology is structured into two principal sections: the utilization of non-machine learning-based models and machine learning-based models to impute missing data.

Non-Machine Learning-Based Models In the non-machine learning-based component of our methodology, we employed two distinct imputation techniques: Completely Random and the Sklearn SimpleImputer. Completely Random serves as a straightforward approach, exclusively utilizing the Temperature Feature Set to populate missing entries with randomly generated values in the range of the temperature data. On the other hand, the Sklearn SimpleImputer offers a more nuanced method. This technique facilitates imputation along both country-based (row-wise) and year-based (column-wise) axes, allowing for the calculation of either the mean or median for specific rows or columns of data. Additionally, for the SimpleImputer method, we extended our experiments to include the combined dataset, which integrates the Temperature Feature Set with additional climate-related variables. Image 1 provides a specific visualization of the non-machine learning-based model.



	A	B	C	D	E	F	G	H	I	J	K	L
1	Country	T1992	T1993	T1994	T1995	T1996	T1997	T1998	T1999	T2000	T2001	T2002
2	Afghanistan, Isla	-0.294	0.22	0.43	0.359	-0.116	0.471	0.675	1.198	0.993	1.311	1.365
3	Albania	0.106	0.076	1.33	-0.172	-0.038	0.075	0.795	0.67	1.065	1.532	0.492
4	Algeria	-0.312	0.552	0.732	0.595	0.846	1.059	1.109	1.476	0.82	1.856	
5	Angola	0.466	0.256	0.212	0.753	0.37	0.107	1.064	0.417	0.169	0.295	0.735
6	Anguilla	0.365	0.535	0.575	0.651	0.35	0.459	1.06	0.532	0.264	0.587	0.706
7	Argentina	-0.261	0.086	0.487	0.259	0.512	0.759	0.351	0.16	-0.186	0.425	0.278
8	Armenia, Rep. o	-1.31	-0.932	0.481	0.697	0.47	0.599	1.316	1.488	1.031	1.395	1.194
9	Australia	0.281	0.484	0.221	0.413		0.383	1.092	0.58	0.148	0.336	0.736
10	Austria	1.248	0.527	1.958	0.939	-0.203	0.514	1.343	0.986	1.769		1.531
11	Azerbaijan, Rep.	-0.839	-1.009	0.098	1.012	0.244	0.582	1.184	1.224	1.098	1.401	1.432
12	Bahamas, The	0.286	0.378	0.801	0.671		0.787	0.924	0.619	0.163	0.135	0.856
13	Bangladesh	-0.146	-0.446	0.175	0.173	0.137	-0.173	0.504	0.78	0.226	0.316	0.398
14	Belarus, Rep. of	1.273	0.024	1.138	1.5	-0.274	0.359	0.706	1.506	1.941	1.979	1.892
15	Belize	0.311	0.447	0.756	0.759	0.21	0.866	0.904	0.422	0.175	0.445	0.733
16	Benin	-0.027	0.531	0.439	0.406	0.736	0.557	1.042	0.621	0.449	0.517	0.721
17	Bhutan	-0.229	-0.248	0.315	0.518	0.307	-0.3	0.825	1.022	0.214	0.72	0.492
18	Bolivia	0.031	0.404	0.883	0.77	0.29	0.825	0.926	0.337	0.403		
19	Bosnia and Herz	0.659	0.248			-0.043	0.112	1.059	0.612	1.525		1.141
20	Botswana	0.813	0.512	-0.123	1.141	-0.319	-0.096	0.857	0.552	-0.718	-0.186	0.359

AE	AF	AG	AH	AI	AJ
C1992	C1993	C1994	C1995	C1996	C1997
178.3916667	178.70125	179.72375	180.7645833	181.6179167	182.095
178.3916667	178.70125	179.72375	180.7645833	181.6179167	182.095
178.3916667	178.70125	179.72375	180.7645833	181.6179167	182.095
178.3916667	178.70125	179.72375	180.7645833	181.6179167	182.095
178.3916667	178.70125	179.72375	180.7645833	181.6179167	182.095
178.3916667	178.70125	179.72375	180.7645833	181.6179167	182.095
178.3916667	178.70125	179.72375	180.7645833	181.6179167	182.095
178.3916667	178.70125	179.72375	180.7645833	181.6179167	182.095
178.3916667	178.70125	179.72375	180.7645833	181.6179167	182.095
178.3916667	178.70125	179.72375	180.7645833	181.6179167	182.095

Image 1:Structure of Non-Machine Learning Based Model

Machine Learning-Based Models Our machine learning model is broadly categorized into four groups of imputation methods: ensemble-based, time series-based, iterative, and statistical. These methods were implemented using various existing packages in Python, including MissForest, FancyImpute, Impyute, and SKlearn. Assume default arguments were used unless otherwise stated. Please see their documentation, cited in the references section, for further details.

Ensemble Methods:

MissForest (MissForest) In our implementation of the MissForest method, we start by filling in missing values with the mean of the column. The process then continues iteratively, building a Random Forest model for each variable with missing data. This model treats all other variables as predictors, with rows where the variable is missing serving as the test set and those where it is present serving as the training set. After each round of imputation, the newly estimated values are updated in the dataset. This dataset, with updated values, is used in subsequent iterations to again estimate missing values. The process repeats until the changes in imputed values between iterations are negligible, signaling that convergence has been achieved [7].

Random Forest (Sklearn) Unlike the direct implementation of the MissForest Method, here we build a Random Forest Regressor for each country (each represented as a row of data) and split each country's data into training, validation, and testing subsets. Initially, we train the regressor using temperature change data exclusively from the past three years: delta temperature at year t-1, t-2, and t-3. This model is then used to predict the temperature change for the subsequent year, t. After training, we input test data into the model to generate predictions for previously omitted values.

Next, we enhance the training process by incorporating an extended feature set alongside the original temperature features. This extended set includes additional climate-related information, such as carbon dioxide concentrations and disaster frequency, providing a more comprehensive dataset for modeling. The enhanced Random Forest model is then retrained with these features and subsequently tested to estimate the temperature change at year t once again.

This procedure is repeated for each of the 165 countries in our dataset, both with and without the extended feature set. We evaluate the performance of the models through metrics such as Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) for each country. To synthesize these individual assessments into a collective measure of performance, we compute the average RMSE and MAE across all countries. This aggregated result allows us to compare the efficacy of our modified approach against other methods in predicting temperature changes. Image 2 displays a flowchart illustrating the structure of a Random Forest Regressor, detailing its operational framework.

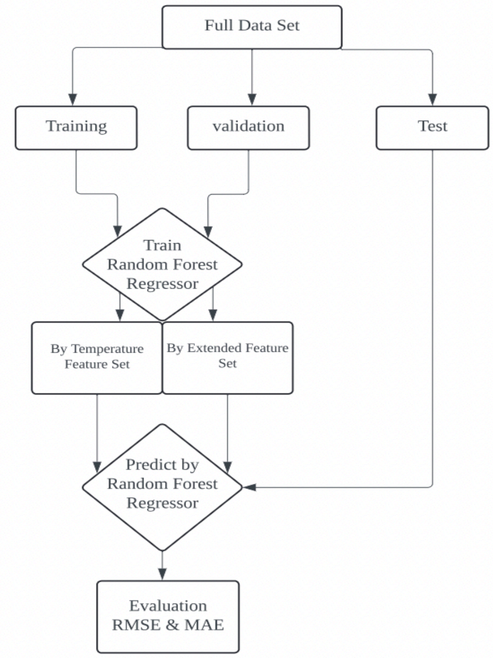

Full Data Set
Training
validation
Test
Train Random Forest Regressor
By Temperature Feature Set
By Extended Feature Set
Predict by Random Forest Regressor
Evaluation RMSE & MAE

Image 2: Structure of Random Forest Regressor

Time Series-based Methods:

Last Observation Carried Forward (Impyute) The Last Observation Carried Forward method addresses missing values in a dataset by utilizing the last observed value prior to the missing one. This method operates on the assumption that the most accurate prediction for a missing value is the value that was last recorded. In practice, when a missing value is encountered in a series, LOCF simply substitutes it with the most recent non-missing value found earlier in the sequence. This approach is straightforward and commonly used in time series data where it is reasonable to assume that conditions have not changed significantly between consecutive data points [8].

Moving Window (Impyute) In our application of the Moving Window technique to address missing data, we utilize the mean as the chosen statistic. We begin by determining the size of the window and then calculate the mean for the observations contained within this window as it traverses the dataset. This average, derived from the observed values in the window, is subsequently used to fill in missing entries. This method proves beneficial in datasets where there is a serial correlation, as the mean of nearby data points offers a dependable estimate for the missing values [8].

Iterative Methods:

IterativeImputer (SKlearn) The IterativeImputer technique is grounded in the principles of multivariate imputation by chained equations (MICE) and is designed to handle missing data by considering each feature with missing entries as a dependent variable within a regression framework, utilizing all other available features as predictors. Initially, the process involves filling all missing values with initial estimates, typically using the mean or median of each feature. Subsequently, a specific feature column is chosen to act as the dependent variable. A regression model is then fitted based on the observed values of this feature, with the remaining features serving as independent variables. This model is used to predict the missing values for the chosen feature. This procedure is iteratively applied to each feature in the dataset, systematically addressing all missing data [3].

IterativeSVD (FancyImpute) IterativeSVD is an imputation technique based on the principle of matrix completion through Singular Value Decomposition (SVD), which breaks down a matrix into three separate matrices to expose its fundamental geometric and algebraic properties. The process initiates with an initial estimation of the missing values, typically using simple imputation methods like the mean or median. Subsequently, a low-rank approximation of the data matrix is computed using SVD. This approximation is utilized to fill in the missing values. The procedure is inherently iterative; with each cycle, the SVD is recalculated based on the newly updated matrix. This iterative recalibration continues, progressively refining the estimates of the missing values until the adjustments between iterations diminish below a predetermined threshold or until a maximum number of iterations is achieved. This method efficiently leverages the underlying structure of the data to estimate missing values accurately [9].

Statistical Methods:

Expectation Maximization (Impyute) The Expectation Maximization (EM) algorithm is a statistical method designed to estimate parameters in models dealing with incomplete datasets. It operates through two main phases: the Expectation step (E-step) and the Maximization step (M-step). In the E-step, the algorithm estimates the missing values based on the current parameter estimates, computing the expected value of the likelihood function as if these parameter estimates were the actual true parameters. Following this, the M-step involves updating the parameters by maximizing the likelihood function, using the data estimated during the E-step. This maximization typically entails optimizing the parameters to enhance the fit to the data, encompassing both the originally observed and the newly estimated values. Through these iterative steps, EM effectively refines the parameter estimates to accurately handle and interpret incomplete data [8].

KNNImputer (SKlearn) The KNNImputer is based on the assumption that similar data points are located in close proximity within the feature space, allowing for the estimation of missing values by leveraging information from the nearest neighbors. When a missing value is encountered, the imputer determines the k-nearest neighbors using a distance metric, typically Euclidean, focusing on the available non-missing features. Once these neighbors are identified, the missing values are imputed by calculating the mean of these neighbors' values. This method effectively utilizes the inherent relationships and patterns within the data, providing a robust mechanism for handling missing data by drawing on the similarities among data points [3].

5. Results and Evaluation

Table 1 in the appendix shows the results of all our imputation methods on both feature sets. The results of completely random imputation with the Extended Feature Set were not included, as completely random imputation was performed only to serve as a baseline. Including extended features into the range of random imputation would only provide results skewed by the distribution of the other features, providing a metric that is not useful for any comparisons nor very effective.

Of the non-ML-based methods, row-wise imputation of mean and median with the Extended Feature Set was extremely inaccurate. This is due to the fact that the extended features span a much wider range of values than the temperature features, making any row-wise metrics heavily inflated. We could have standardized these additional features to fix this issue, but we decided to keep them in their original units so that our RMSE and MAE might be in the original units of the data, making our results more easily interpretable. Since these results are so skewed and not representative of meaningful trends, they have been excluded from all further visualizations and analyses.

Figure 1 shows each model's RMSE and MAE on the Temperature Feature Set, colored by their respective model type. These results are about as expected; all models did better than the completely random imputation baseline, and all ML-based models performed better on both metrics compared to non-ML-based methods. Our RF model, denoted with a triangle on the figure, performed moderately well, surpassing all non-ML-based methods but not as successful as the other ML-based methods, likely because of their additional complexity. The three best-performing models for this feature set were MissForest, IterativeSVD, and Moving Window.

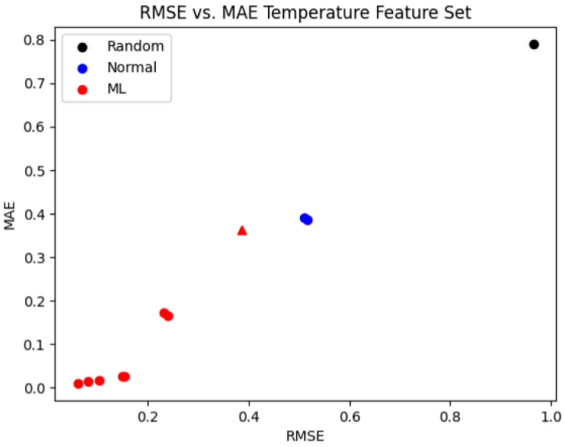


Figure 1: Performance of models on the Temperature Feature Set in terms of RMSE and MAE

Figure 2 reflects the same data as Figure 1 but for the Extended Feature Set, and does not show the same expected or consistent trends. We see some non-ML-based models surpassing ML-based models in terms of both RMSE and MAE, and one model even has a higher RMSE than the random baseline.

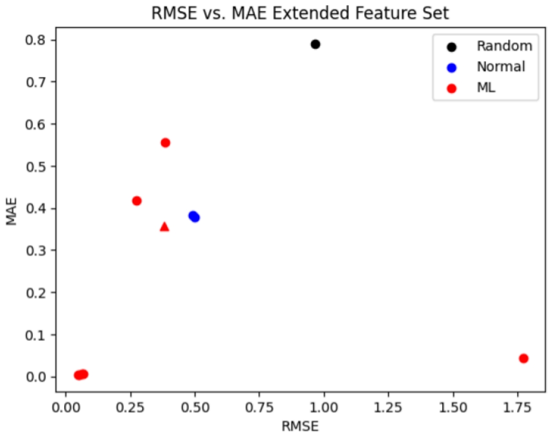


Figure 2: Performance of models on the Extended Feature Set in terms of RMSE and MAE

This outlying RMSE value of 1.772 degrees Celsius was a result of the Moving Window model. While this model had a high RMSE, it had an MAE of only 0.044, which is comparable to that of similar models; this disparity indicates the skewed RMSE was a result of a few high outliers, as RMSE is less robust to outlying values. As mentioned previously, this model works by aggregating the values of a window of cells around a missing data point to impute this value. Missing values on the outskirts of the temperature feature columns likely had moving windows that included data from the other climate features. As these features are on different scales, this likely inflated the predicted imputed value, leading to the inflated RMSE. So, while this model was quite successful on the Temperature Feature Set, it is impractical to deploy on the Extended Feature Set, or, more generally, to be used with features that have different distributions.

Two other ML-based models performed more poorly than several non-ML-based models on both RMSE and MAE: KNNImputer and IterativeImputer. The reason behind their low performance is less intuitive. We surmise that these models needed to be more complex to learn the complex relationships between the extended features and temperature, as they are the least complex of the ML-based models. Our RF model, a bit more complex being an ensemble method, performed a bit better than these models, though worse than the more complex ML-based models, EM, LOCF, IterativeSVD, and MissForest, which were the best-performing for this feature set.

Figure 3 shows the RMSE graphed against the MAE for all models to compare the relative performance of the models between the feature sets. Surprisingly, no feature set universally has better performance when being used to test models than the other. The results are quite mixed. In general, models trained on the Extended Feature Set have the best performance in terms of the smallest RMSE and MAE, but not by a large margin by any means.

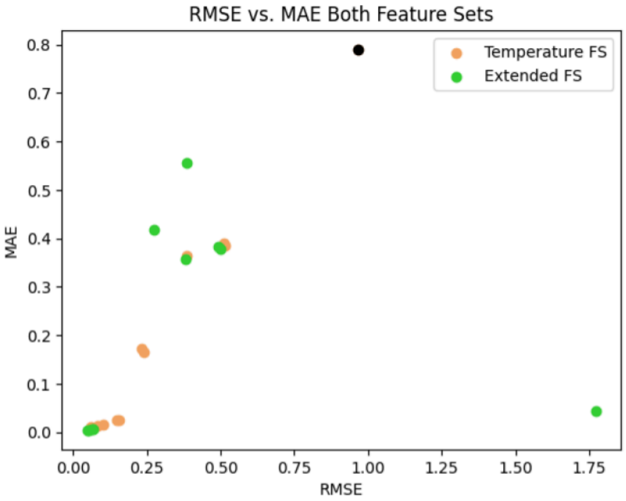


Figure 3: Performance of models on both feature sets

From the results of the previous graph, it is unclear whether the Extended Feature Set had a significantly positive effect on increasing the accuracy of imputations. To get a better understanding of this debate, we decided to break down the influence of the Temperature versus the Extended Feature Set by the model. Here, we averaged the RMSE and MAE of each model for each feature set to get one metric to represent the accuracy of each model. We then obtained the difference in each model's performance on the Extended and Temperature Feature Sets and graphed them in Figure 4.

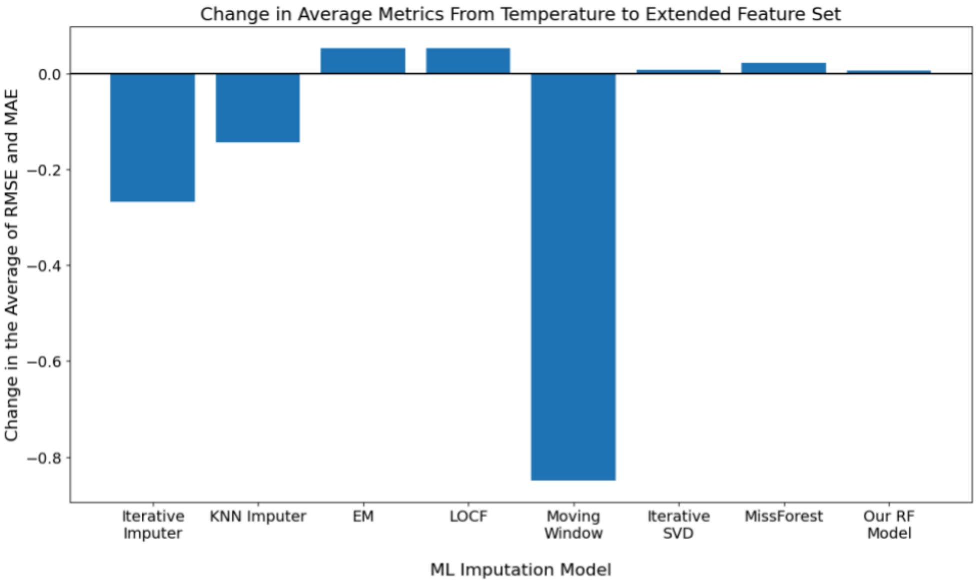


Figure 4: Performance comparison between feature sets for each model

In this graph, negative changes indicate that the model performed worse when trained on the Extended Feature Set than when trained on the Temperature Feature Set. Moving Window is one model that certainly follows this trend; the same is true for IterativeImputer and KNNImputer, as we saw previously. The models that do perform better on the Extended Feature Set, our RF model, EM, LOCF, IterativeSVD, and MissForest only showed a small amount of improvement, with the maximum change in the average of RMSE and MAE being a reduction of only 0.053. These results indicate that the Extended Feature Set did not provide significant improvement in the accuracy of imputation. This means that, for future imputations of the same kind, dealing with the logistical challenges of gathering extended features for imputation is largely unnecessary; simple temperature data would serve quite well, as long as the right model is selected for imputation.

With this being said, since we have the data from the Extended Feature Set readily available, we will certainly include it in our models if it aids in their performance. We can identify the best model for imputation as the MissForest model trained on the Extended Feature Set, which had the lowest RMSE and MAE out of all the models. We will thus use this model to impute the missing data in the actual IMF dataset; the link to the resulting CSV is provided in the appendix under Attachment 1.

6. Conclusion and Future Work

Our methodology involved two main strategies: using traditional non-machine learning methods and advanced machine learning models to fill in missing data. Initially, we worked with a basic Temperature Feature Set, focusing solely on annual temperature changes. Later, we enhanced our dataset by adding more complex climate-related variables such as land cover changes, disaster frequencies, and carbon metrics, to create what we called the Extended Feature Set. This allowed us to compare traditional data-filling methods against more sophisticated machine-learning approaches.

In our experiments with non-machine learning methods, such as the Sklearn SimpleImputer and Completely Random imputation, we found that while these methods were straightforward, they were not as effective in managing the complexities introduced by the Extended Feature Set. The results were often skewed, especially with row-wise imputation, due to the diverse range of values from different data types. On the machine learning side, we explored several methods, including ensemble techniques like Random Forest and sophisticated algorithms like MissForest and IterativeSVD. MissForest proved to be particularly effective. It works by making initial guesses for missing values and iteratively improving these guesses using a Random Forest approach. The iterative process continues until the changes in imputed values stabilize, a method that has shown excellent results in terms of accuracy and reliability.

Our findings demonstrate that while non-ML methods are simpler and sometimes adequate, ML methods, especially MissForest, provide a significant improvement in accuracy. They are particularly adept at handling datasets that incorporate a variety of features and data complexities. Despite the challenges of integrating and processing an extended set of features, MissForest's ability to handle these complexities makes it a valuable tool for future climate research and policy development. Thus, moving forward, we plan to use the MissForest model, applied to the Extended Feature Set, to impute missing data in the IMF dataset. This approach not only addresses the technical challenges effectively but also enhances the inclusivity and comprehensiveness of climate data analysis.

We then have some specific areas for further development and research, building on the findings from the current study of imputing missing climate data using the MissForest model. Two main avenues are proposed for enhancing the model's effectiveness and applicability to broader climate data challenges.

Optimization of MissForest Parameters: The performance of the MissForest model can be significantly influenced by its parameter settings. Future research should focus on an exhaustive parameter-tuning process to determine the optimal configuration for climate datasets. This includes adjusting the number of trees in the forest, the depth of each tree, and the method of bootstrap sampling. Each of these parameters can affect the model's ability to capture the underlying patterns in the data, and finding the right balance can improve both accuracy and computational efficiency. Moreover, incorporating feature weighting within the algorithm could help identify which variables most significantly impact the model's predictions, providing valuable insights for climate science research.

Development of Hybrid Models: Integrating MissForest with other predictive techniques could lead to the creation of hybrid models that capitalize on the strengths of multiple approaches. Specifically, the combination of MissForest with models designed for time series analysis could enhance imputation accuracy for data with inherent temporal structures, such as historical weather patterns or seasonal variations in environmental indicators. This hybrid approach would be particularly advantageous for datasets where predictions depend heavily on understanding temporal dynamics, offering a more nuanced and precise tool for climate data analysis. Such models could better accommodate the complexities of climate change indicators, ensuring more robust predictions and supporting more informed decision-making in climate research.

By pursuing these specific areas of future work, the project can extend the utility of the MissForest model and contribute more broadly to the field of climate data analysis, ultimately supporting efforts to combat and understand global climate change.

Acknowledgments

We thank Professor Grace Lindsay and Teaching Assistant Kartik Jindgar from New York University for their mentorship and invaluable assistance throughout this project and the course
(DS-UA 301 Advanced Topics in Data Science: Machine Learning for Climate Change).

References

1. International Monetary Fund (IMF). “About the IMF.” *International Monetary Fund (IMF).*
2. Tracyrenee. “The Importance of Deleting Columns with Null Values When Making Predictions.” *Medium*, 12 Mar. 2022.
3. Scikit-Learn. “6.4. Imputation of Missing Values.” *Scikit*, scikit-learn.org/stable/modules/impute.html#:~:text=Missing%20values%20can%20be%20imputed,for%20different%20missing%20values%20encodings. Accessed 30 Apr. 2024.
4. Rubin, Donald B. "Multiple Imputation After 18+ Years." *Journal of the American Statistical Association*, vol. 91, no. 434, 1996, pp. 473-489. Taylor & Francis, Ltd.
5. United Nations Framework Convention on Climate Change. Text of the UNFCCC. 1992.
6. Mohammad, A., Pugacheva, E. (2022). Impact of COVID-19 on Attitudes to Climate Change and Support for Climate Policies. IMF Working Paper WP/22/23. International Monetary Fund.
7. “MissForest.” *PyPI*, pypi.org/project/MissForest/.
8. *Impyute*, impyute.readthedocs.io/en/master/index.html.
9. “Fancyimpute.” *PyPI*, pypi.org/project/fancyimpute/.

Appendix

Table 1: Results of Imputation Model Tests on Synthetic Datasets

	Temperature Feature Set		Extended Feature Set	
Method	RMSE	MAE	RMSE	MAE
Non-ML-Based Methods				
Completely Random	0.965	0.790	N/A	N/A
Column Mean (sk-learn SimpleImputer)	0.510	0.390	0.491	0.382
Column Median (sk-learn SimpleImputer)	0.517	0.386	0.500	0.377
Row Mean (sk-learn SimpleImputer)	0.526	0.382	15129.675	4296.106
Row Median (sk-learn SimpleImputer)	0.529	0.383	95.536	91.259
ML-Based Methods				
IterativeImputer (sk-learn)	0.232	0.173	0.385	0.556
KNNImputer (sk-learn)	0.240	0.165	0.275	0.419
Expectation Maximization (impyute)	0.149	0.026	0.064	0.005
Last Observation Carried Forward (impyute)	0.154	0.026	0.069	0.005
Moving Window (impyute)	0.103	0.016	1.772	0.044
IterativeSVD (fancyimpute)	0.061	0.010	0.052	0.004
MissForest	0.082	0.014	**0.048**	**0.004**
Our RF Model	0.387	0.363	0.382	0.357

Attachment 1: Imputed IMF Temperature Dataset

For access to the final imputed IMF 'Annual Surface Temperature Change' dataset, please use this link. If you encounter any issues, please contact any member of this team.