

Article

Imputation of Multi-Dimensional High-Frequency Climate Data to Predict Air and Surface Temperatures in Kuwait

Shehroz S. Khan *  and Rami Al-Hajj * 

College of Engineering and Technology, American University of the Middle East, Egaila 54200, Kuwait

* Correspondence: shehroz.khan@aum.edu.kw (S.S.K.); rami.alhajj@aum.edu.kw (R.A.-H.); Tel.: +965-2225-1400 (S.S.K. & R.-A.H.)

Abstract

Missing values may arise in climate data collection due to sensor malfunction, transmission errors, device calibration and operational issues. This problem can be more catastrophic in the case of multi-dimensional and high-frequency climate data sets, where some or all climate readings could be missing at multiple timestamps. These missing data in high-frequency climate modeling could lead to inaccurate prediction models, which in turn affect overall assessments, planning, and climate-related measures and policy. In this paper, we evaluate the performance of three imputation techniques based on the mean, k-nearest neighbor, time-based interpolation and a new temporal cross-year climate imputation approach using a random forest, long short-term memory (LSTM) model and contextual embedding-based Transformer regression methods. We discussed our findings on four years of multi-output, high-frequency and multi-dimensional climate data collected in Kuwait. Using a leave-one-year-out cross-validation approach, our results show that all imputation methods perform better than no imputation, with LSTM and time-based interpolation emerging as the best combination. Imputing climate data based on previous years' timestamps did not yield good results, highlighting the variability of climate data across years.

Keywords: climate modeling; missing data; imputation; deep learning

1. Introduction

High-frequency climate data is an essential resource to capture complex intraday weather patterns, seasonality, and temporal auto-correlation. These data are crucial for monitoring short-duration extreme events, understanding land-atmosphere interactions, and supporting operational forecasting and early warning systems. The multi-dimensionality in these data sets must account for both temporal and spatial or inter-variable dependencies (e.g., temperature, pressure, and humidity). However, predicting multiple outputs (e.g., air and surface temperatures) in high-frequency multi-dimensional climate data is a challenging task [1]. Missing data at fine temporal resolutions can significantly degrade downstream tasks such as temperature prediction, heat stress assessment, and climate risk management [2]. Missing data normally occurs in climate research due to several reasons, such as sensor malfunction, depleted batteries, power failures [3,4], transmission and data logging failures [5], harsh weather conditions, incorrect installations [6], and operational issues [3]. Arid and hyper-arid regions present particularly challenging conditions for high-frequency climate data collection. Kuwait serves as a representative case study due to



Academic Editor: Jérôme Gensel

Received: 17 December 2025

Revised: 5 February 2026

Accepted: 18 February 2026

Published: 25 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

its extreme temperatures, frequent dust events, and strong seasonal variability [1], making it a suitable test bed for evaluating imputation strategies under harsh climatic conditions.

Beyond station-based sensor networks, missing data is also a well-recognized challenge in reanalysis products [7], urban climate sensing [8], and remote sensing applications [9], where data gaps arise due to retrieval failures, cloud cover, sensor degradation, or spatial–temporal inconsistencies. In these contexts, gap-filling and reconstruction techniques are critical for maintaining data continuity and enabling downstream modeling, the monitoring of climate extremes, and urban heat assessment. While the present study focuses on high-frequency ground-based climate observations, the methodological emphasis on robust imputation and its downstream predictive impact is complementary to challenges encountered in these related domains.

Unlike low-frequency climate data, high-frequency and extreme weather environmental data sets are more likely to have missing values due to sensor outages, communication failures, and harsh operating conditions [10,11]. Therefore, maintaining temporal structure and prediction utility requires the effective imputation of such data. Missing data in climate and meteorological time series is a well-recognized challenge, particularly because incomplete records can bias downstream analyses and compromise prediction or forecasting performance. A recent systematic review of imputation strategies for climate time series highlights a diverse set of approaches, ranging from simple statistical methods to advanced machine learning (ML) and deep learning (DL) techniques [12]. Below, we present a summary of these approaches that can be broadly categorized as follows:

- **Conventional Statistical Methods:** Early work in this domain predominantly utilized statistical techniques, such as mean imputation or seasonal averaging, simple and multiple linear regression, and interpolation. These methods are typically computationally fast and easy to interpret; however, they may fail to capture complex temporal dynamics in climate data, especially at fine temporal resolutions [12,13].
- **Time Series Models:** For high-resolution climate time series data, models such as structural time series with Kalman smoothing and autoregressive integrated moving average models have been examined for missing value reconstruction [14]. In a notable case study on hourly climate variables, such time series approaches performed competitively, demonstrating the plausibility of imputed values under moderate missingness, but showed variable efficacy across locations and variables [14].
- **Machine Learning Techniques:** More recently, machine learning methods have been applied to meteorological data imputation, including K-nearest neighbor (KNN), random forest (RF), missForest, MICE, Gradient Boosting [12], and have been shown to achieve good accuracy for standard meteorological variables (e.g., temperature and humidity) under moderate missing rates, with tree-based learners often outperforming other classic ML alternatives [15]. These techniques, however, usually work with daily or hourly data and may not scale well with high-frequency and multi-output settings without significant adaptation.
- **Deep Neural Networks:** Advances in deep learning for missing data imputation have also been explored in climate studies. Some of the common deep learning methods used in this application, including graphical neural networks [16], generative adversarial networks [17], bi-long short-term memory models (LSTM) [18], Transformers [19], recurrent neural networks [20], deep spatio-temporal models [21], multi-layered perceptron [22] and convolution neural networks [23]. These methods have shown competitive or sometimes better results; however, they incur high computational costs [12].

Demirhan and Renwick [24] examined 36 imputation techniques' accuracy for solar irradiance series using a real-world Australian climate data under 16 experimental

conditions. Their objective was to estimate missing values in minutes, hours, daily, and weekly solar irradiance series. They showed that for high-frequency climate series data, simple techniques such as interpolation can work well. More sophisticated techniques, e.g., the Kalman filter, can provide benefits in a broader range of contexts, although time and resource complexity could be a deterrent. If the task involves predicting multiple outputs (e.g., different values of air and surface temperatures), then the quality of imputed missing data becomes more critical to avoid deterioration in predictive performance.

Despite this breadth of research, two fundamental gaps remain insufficiently addressed in the existing literature:

- Limited focus on high-frequency, multi-output climate data: Much of the prior work has focused on univariate series (e.g., temperature and precipitation) or lower sampling frequencies, leaving a gap in understandings of how these methods perform for densely sampled, multi-variable data sets.
- Insufficient treatment of cross-year dependencies: Standard imputation methods often do not exploit recurrent seasonal and diurnal patterns across successive years, which can be an important source of information for high-frequency data.

The existing literature has not examined the effect of different imputation strategies on inter-annual cyclic patterns in multi-output temperature prediction models. This lack of integrated evaluation obscures the practical implications of imputation choices for high-frequency climate analytics and limits the reliability of predictive systems used in extreme climate monitoring and forecasting. This research is motivated by the need for high-fidelity climate data in Kuwait, where extreme heat and dust storms frequently cause simultaneous sensor failures across multiple dimensions.

In this paper, we consider 4-years' worth of high-frequency multi-output climate data collected at the American University of the Middle East, Kuwait. The data was collected between the years 2021–2024 and contains numerous climate variables collected every 5 s, including air density, wind direction, dew point, heat index, pressure, relative humidity, solar radiation, air and surface temperature, and related variables (see Table 1). This data contains two types of missing values: (i) the entire record is missing, i.e., all sensor values are missing for a given timestamp, or (ii) some sensor values are missing for a given timestamp. The details of the data and missing values are presented in the Section 2.1. The hypothesis is that imputing missing values in high-frequency climate data sets improves the predictive accuracy of multi-output regression models for air and surface temperature compared to models trained on non-imputed data. To validate the hypothesis, we explore three traditional missing imputation algorithms, including mean, knn and time-based interpolation. Additionally, we also develop a baseline temporal cross-year climate imputation algorithm (TCYC) that leverages climate variables' values from previous years to impute the current year's missing values for a given timestamp. We tested three multi-output regression approaches to predict six climate variables, including RF, the long short-term memory model (LSTM) and a contextual embedding-based Transformer (FTTransformer) [25]. We followed a leave-one-year-out cross-validation strategy to study the generalization of imputation techniques across different years. The main contributions are summarized below:

- A comparative evaluation of imputation techniques for high-frequency, multi-dimensional climate data in an arid region: Kuwait.
- An analysis of downstream predictive impact, showing how imputation choices influence air and surface temperature prediction accuracy.
- Introduction of TCYC as a cyclic-aware baseline imputation approach for multi-variate climate variables.

Our results show that the imputation of missing climate data helps in building better predictive models in comparison to no imputation. The optimal combination of regression and imputation techniques was found to be the LSTM with time-based interpolation. The results also show that utilizing data from previous years to impute a given year may not yield better results than other traditional imputation methods, yet it performs better than no imputation.

Table 1. Different weather parameters captured in the AUMET data set; source [1].

#	Parameter	Unit	Description
1	Air_Density	Kg/m ³	Density of the air
2	Air_Temp_Avg	°C	Average temperature of the air at three meters above the surface over 5 min
3	Air_Temp_Min	°C	Minimum temperature of the air at three meters above the surface over 5 min
4	Air_Temp_Max	°C	Maximum temperature of the air at three meters above the surface over 5 min
5	Corr_Wind_Dir_Avg	degree	Average corrected wind direction if the wind sensor is moving over 5 min
6	Corr_Wind_Dir_Min	degree	Minimum corrected wind direction if the wind sensor is moving over 5 min
7	Corr_Wind_Dir_Max	degree	Average corrected wind direction if the wind sensor is moving over 5 min
8	DPoint_Avg	°C	Average dew point temperature over 5 min. Dew point is the temperature at which air becomes saturated with moisture (water vapor).
9	DPoint_Min	°C	Minimum dew point temperature over 5 min.
10	DPoint_Max	°C	Maximum dew point temperature over 5 min.
11	HeatIndex_C	°C	represents how hot it feels to the human body when relative humidity is combined with the air temperature. It is also known as the feels-like temperature in hot conditions.
12	Pressure_Avg	hPa	Average air pressure
13	Pressure_Min	hPa	Minimum air pressure
14	Pressure_Max	hPa	Maximum air pressure
15	Relative_Wind_Dir_Avg	angleo	Average relative wind direction.
16	Relative_Wind_Dir_Min	angleo	Minimum relative wind direction
17	Relative_Wind_Dir_Max	angleo	Maximum relative wind direction
18	Relative_Wind_Speed_Avg	m/s	Average wind speed
19	Relative_Wind_Speed_Min	m/s	Minimum wind speed
20	Relative_Wind_Speed_Max	m/s	Maximum wind speed
21	RH_Avg	pct	Average relative humidity
22	RH_Min	pct	Minimum relative humidity
23	RH_Max	pct	Maximum relative humidity
24	Solar_Radiation_Avg	w/m ²	Average amount of solar radiation in 5 min
25	Solar_Radiation_Min	w/m ²	Minimum amount of solar radiation in 5 min
26	Solar_Radiation_Max	w/m ²	Maximum amount of solar radiation in 5 min
27	Sunshine_Hr	H	The cumulative sunshine duration (in hours) during which direct solar irradiance is >120 W/m ² recorded over a given day, and it is reported at 5-minute intervals
28	Surface_Temp_Avg	°C	Average air temperature at 10 cm above the surface over 5 min
29	Surface_Temp_Min	°C	Minimum air temperature at 10 cm above the surface over 5 min
30	Surface_Temp_Max	°C	Maximum air temperature at 10 cm above the surface over 5 min
31	WetBulbTemp	°C	The Wet Bulb Temperature. It is the temperature a parcel of air would have if it were cooled to saturation.
32	Windchill	-	999
33	Timestamp	UTC+3, DD/MM/YYYY H:M:S	Time of sensor reading.

2. Materials and Methods

2.1. AUMET Climate Data

The AUMET climate data set is collected from a weather monitoring system deployed at the research and innovation center of the American University of the Middle East (AUM), Kuwait, latitude (north) 29° 9′ 53.28″ N, and longitude (east) 48° 5′ 53.61″ E. This system is situated on a 25-meter high elevated rooftop above ground level. The system is solar-powered by a solar panel and contains a data logger and a cellular antenna for remote access. Figures 1 and 2 show the general structure of the AUMET weather management system along with its geographical location.

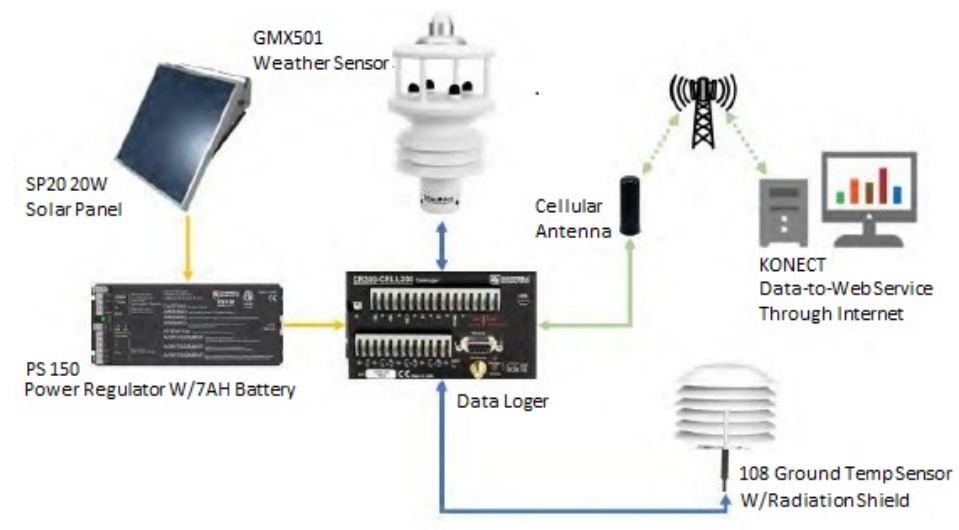


Figure 1. AUMET weather monitoring system diagram.



Figure 2. Location of AUMET weather monitoring system.

The AUMET data set contains high-frequency weather and solar radiation data collected at 5-minute time intervals since 2019. This data contains a total of 33 climate parameters, as shown in Table 1. More details on the data set and the system are presented in the works of Khan and Al-Hajj [1]. The data set contains no duplicate values; however, each year contains several missing values, mostly due to one or more sensor malfunctions. The degree of missingness depends on the level of outage of the sensor(s), as discussed below:

- (1) In the year 2021, the solar panel malfunctioned and there are no data between March 29, 14:15:00, and May 09, 11:25:00. Additionally, it also has some missing data at the beginning of the year from midnight (00:00:00) to 00:03:00, March 29, between 08:05:00 and 08:10:00, and 08:20:00 and 13:40:00, and between May 09, 18:20:00, and May 10, 05:55:00. Therefore, 12,047 records have missing data, leading to 11.46% missing data in the year 2021.
- (2) In the year 2022, there are missing data at various time intervals from January 01 03:00:00 to 03:00:00 (one record missing), July 21, 07:30:00 to 08:10:00, August 18, 08:05:00 to 09:05:00, 09:10:00 to 09:15:00 (two rows with NaN readings), and 09:30:00 to August 18 09:30:00 (one record missing), as well as November 27, 09:45:00 to 12:15:00. This leads to 55 missing records, which is 0.05% missing data. The rows with NaN values are also considered missing values in this case.
- (3) In the year 2023, only one record is missing on January 01 at 03:00:00, which is almost equal to 0% of the total data.
- (4) In the year 2024, the data are missing from January 01, 03:00:00 to 06:00:00, and March 10, 00:09:00 to 00:09:15 (four rows with NaN readings). This leads to 37 records missing, equaling 0.04% of missing data. The rows with NaN values are also considered missing values in this case.

Tables 2 and 3 show a detailed analysis of the missing data for each year. It is expected that each year will contain 105120 records, with the year 2024 being a leap year to contain 105408 records (corresponding to the number of timestamps spaced at 5-min time intervals in that year). The year 2021 has the most missing values due to a malfunction with the solar panels for a long duration, whereas the year 2023 only contains one missing record. Overall, the AUMET climate data set is robustly collected over four years, except for the year 2021.

Khan and Al-Hajj [1] use the high-frequency AUMET climate data set to predict six multiple outputs, including air and surface temperatures, and their minimum and maximum values within a 5-minute interval. The authors removed the missing values from their analysis. They employed several multi-output regression models in a leave-one-year-out cross-validation framework and found that the LSTM and FTTransformer performed the best on this data set. The RF regression also worked the best among the traditional machine learning methods. However, due to omitting the missing values, the sequential relationship between various climate variables could be broken and lead to sub-optimal prediction results.

In this paper, we used a similar methodology as Khan and Al-Hajj [1]. Out of 33 climate variables, windchill was excluded due to a constant value. Six temperature variables were identified as dependent variables for the multi-output prediction, including air and surface temperatures, and their maximum and minimum values. The timestamp parameter is recorded in UTC+3, in the DD/MM/YYYY H:M:S format, and is converted to extract the hour, minute, day of the week, and month corresponding to a given timestamp. The day of the week and month are numerically encoded from 1 to 7 and 1 to 12; thus, the overall number of features in the final data set are 30. RF, LSTM and FTTransformer regression are trained on these 30 features to output six temperature variables. The regression methods are trained on data with missing values (missing values were omitted) and after imputing the missing values, as described below. Both LSTM and FTTransformer were trained for 20 epochs with the same architecture as presented by Khan and Al-Hajj [1].

Table 2. Missing data time intervals in years 2021, 2022, 2023 and 2024. The date format is year–month–date. The missing data duration format is x days hours:minutes:seconds of missing data.

Start Time	End Time	Duration
2021-01-01 00:00:00	2021-01-01 03:00:00	0 days 03:00:00
2021-03-29 08:05:00	2021-03-29 08:10:00	0 days 00:10:00
2021-03-29 08:20:00	2021-03-29 13:40:00	0 days 05:25:00
2021-03-29 13:50:00	2021-03-29 14:05:00	0 days 00:20:00
2021-03-29 14:15:00	2021-05-09 11:25:00	40 days 21:15:00
2021-05-09 18:20:00	2021-05-10 07:55:00	0 days 13:40:00
2022-01-01 03:00:00	2022-01-01 03:00:00	0 days 00:05:00
2022-07-21 07:30:00	2022-07-21 08:10:00	0 days 00:45:00
2022-08-18 08:05:00	2022-08-18 09:05:00	0 days 01:05:00
2022-08-18 09:30:00	2022-08-18 09:30:00	0 days 00:05:00
2022-11-27 09:45:00	2022-11-27 12:15:00	0 days 02:35:00
2023-01-01 03:00:00	2023-01-01 03:00:00	0 days 00:05:00
2024-01-01 03:00:00	2024-01-01 06:00:00	0 days 03:05:00

Table 3. Missing data analysis for years, 2021, 2022, 2023, and 2024.

Year	Total Expected Timestamps (Freq = 5 min)	Number of Missing Timestamps	Percentage of Missing Timestamps (%)
2021	105,120	12,047	11.46
2022	105,120	55	0.05
2023	105,120	1	0.00
2024	105,408	37	0.04

2.2. High-Frequency Climate Data Imputation

As discussed in Section 1, traditional imputation approaches tend to perform equivalently to complex deep learning approaches and can be computationally faster. Considering the large amount of data and limited resources, we explore the imputation approaches described below. The imputation was performed on each year’s data separately to avoid data leakage during leave-one-year cross-validation. This will further preserve year-specific idiosyncrasies present in the data and will not corrupt other years by imputing all the data together.

- **Mean Imputation:** In this imputation method, a missing feature value is replaced by its mean value over a complete year. This approach works effectively with small number of missing values; however, when the missingness ratio is high (e.g., year 2021), this approach might be problematic and show limitations. The reason is that all the large number of missing rows will have the same value for each feature; thus, they may not be adding much value to the regression algorithm. In general, mean imputation may not work well with time series data as it disregards the temporal aspect. However, in our feature extraction pipeline, we extract additional four features for each date and timestamp (which are already known). Therefore, even if the imputed climate data values for many missing rows are identical, these four feature values will be different.
- **Knn Imputation:** This imputation method leverages the relationship between variables by estimating missing values based on the values of the k-nearest neighbor. We chose

a default value of $k = 5$. This approach can be a good estimator for data with fewer missing values and with similarity to other data points. This approach may fail, especially when a large chunk of data is missing (e.g., year 2021). In such a case, the nearest neighbor estimation could be inaccurate as the information about the neighbors may also be missing. One point to note is that the missing chunk of data could be a transition between seasons (e.g., cold to hot); thus, nearest neighbor imputation may not be ideal in those cases. Moreover, the value of k is not known in advance and estimating it is not straightforward. For small values of k , the algorithm is, in general, sensitive to noise and subject to a risk of over-fitting. For large data sets, the computation time for this approach can be high since the distance between all pairs of data points needs to be computed.

- **Interpolation Imputation:** In this method, interpolation estimates missing values by creating data points within the range of existing discrete points, often assuming a linear or non-linear relationship between consecutive observations. This method is particularly suited for ordered or time series data. Using the Python 3.12.3 interpolation implementation [26], we employ time-based interpolation with the parameter 'method' set to 'time' as it works for daily and higher resolution data to interpolate a given length of interval; the parameter 'limit_direction' is set to 'both' to fill consecutive NaNs values in both forward and backward directions. This approach may work well for ordered time series data by assuming linear relationships in short intervals.

2.3. Temporal Cross-Year Climate Imputation Algorithm (TCYC)

We now present the TCYC imputation algorithm specific to multi-year, high-frequency time series climate data. It is developed to be used in a leave-one-year-out cross-validation framework, where three years of data are used for training and one for testing, but it can be extended to any number of years. The method leverages cross-year cyclical temporal similarity using the same timestamps from different years to fill missing or NaN values. For each training fold, data from three years are merged into one data set. Missing or incomplete timestamps are imputed by referencing the identical time positions in previous years. The algorithm follows a hierarchy of four rules:

- (i) If a timestamp's data is missing, compute the average of the same timestamp values from the previous two years.
- (ii) If only one of the two previous years has a valid record for that timestamp, use that record directly.
- (iii) If both previous years are missing, search backward or forward (within a window) for the nearest valid rows, and take their average to impute the missing timestamp.
- (iv) For leap years (e.g., 2024), if the data for February 29 is missing in prior years, use the same time from February 28 or March 1 as a fallback for imputation.

If none of the fallbacks are met, then TCYC can leave some data with NaN instead of imputing with unrealistic values. TCYC can broadly preserve seasonal and diurnal consistency across years and avoids global averaging by using temporally aligned data. It minimizes data loss and ensures all timestamps remain intact, while preventing data leakage by restricting imputation to training years only. However, various fallback can lead to imputation mistakes, e.g., if the timestamps for all the previous three years are missing and for leap years. This approach cannot learn feature correlation. The biggest challenge would be imputing prolonged missing intervals that may result in averaged data that loses temporal variability by imputing uniform values. Additionally, TCYC assumes that the same timestamps across different years follow similar distribution or values. However, this may not be the case as yearly weather could differ and imputed values could introduce unwarranted variations in the data set.

Each regression method is trained on mean, knn, time-based interpolation and TCYC-imputed versions of the original data set, and also with no imputation. The no imputation method refers to the original data with missing values. In this situation, the missing data (or rows) are removed and the regression methods are trained. The no imputation method serves as a baseline to compare against the other imputation methods. All the regression methods are trained with default parameters in a leave-one-year-out cross-validation regimen, where the data of one year is tested and the remaining three years' data is used for training. In our case, four folds are formed: when the year 2024 is tested, the training data comprises 2021, 2022 and 2024, and so on for other testing years. The test set in each fold is not imputed to avoid data leakage and fair comparison. The performance metrics reported are the R2 score, Mean Squared Error (MSE) and Mean Absolute Error (MAE) across all the six output variables. Figure 3 shows a detailed diagrammatic representation of different phases involving data processing, cross-validation, imputation methods, model training and testing, and the evaluation and analysis performed on the AUMET data set.

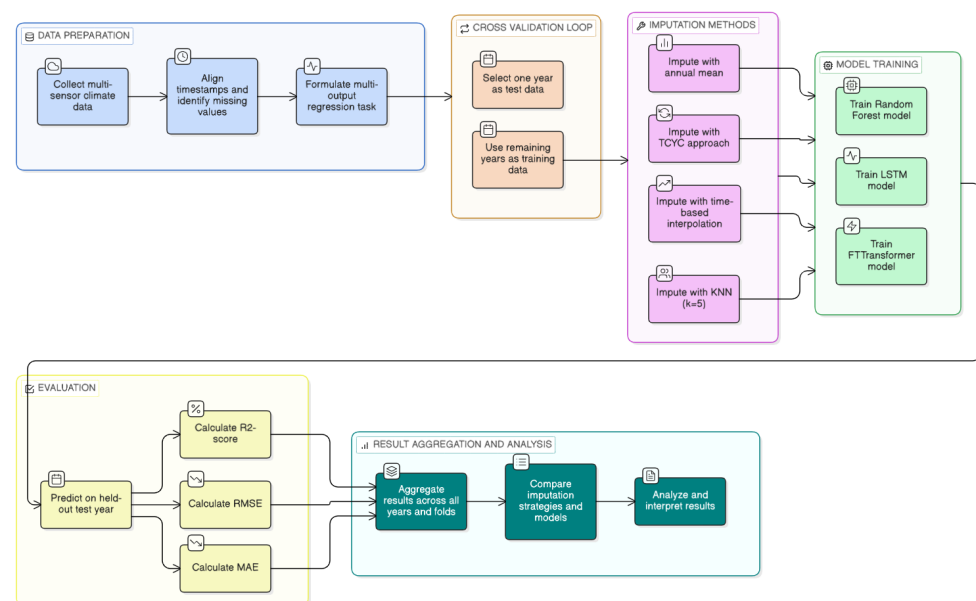


Figure 3. Diagram showing different phases of data processing, cross-validation, imputation, training, and evaluation on the AUMET data set. The diagram is generated using eraser.io.

3. Results

We present the results across folds and within folds to highlight the global and local performance of various regression methods under different imputation approaches.

3.1. Across Folds

Tables 4–6 show the average values of the R2, MSE and MAE metrics for each imputation method, along those with no imputation, across all folds. It can be observed that, with reference to no imputation, both the LSTM model and FTTransformer improve performance across all the metrics when imputation is performed, whereas RF performance remains the same or slightly worse after using imputation. LSTM, with time-based interpolation, shows major improvement in MSE and MAE and minor improvement in R2 scores among all combinations of regression and imputation methods. For RF, LSTM and FTTransformer, the best imputation methods (for R2, MSE and MAE metrics) turn out to be no imputation, time-based interpolation and mean imputation. Alternatively, LSTM and FTTransformer emerged as the top performing regression methods for no imputation and all other imputation methods.

Table 4. R2-score across all folds (higher is better). ‘no’ means original missing data without imputation.

Regressor	Imputation				
	knn	Time-Based	Mean	TCYC	no
LSTM	0.9988	0.9988	0.9987	0.9987	0.9987
RF	0.9982	0.9983	0.9982	0.9983	0.9983
FTTransformer	0.9983	0.9986	0.9987	0.9982	0.9981

Table 5. MSE across all folds (lower is better). ‘no’ means original missing data without imputation.

Regressor	Imputation				
	knn	Time-Based	Mean	TCYC	no
LSTM	0.1237	0.1178	0.1296	0.1276	0.1268
RF	0.1751	0.1751	0.1756	0.1751	0.1749
FTTransformer	0.1701	0.1415	0.1334	0.1787	0.1826

Table 6. MAE across all folds (lower is better). ‘no’ means original missing data without imputation.

Regressor	Imputation				
	knn	Time-Based	Mean	TCYC	no
LSTM	0.2215	0.2187	0.2283	0.2210	0.2258
RF	0.2698	0.2697	0.2703	0.2696	0.2695
FTTransformer	0.2897	0.2566	0.2542	0.3038	0.3119

3.2. Within Folds

Table 7 shows the results for each fold after applying no imputation versus all examined imputation methods for each regression method. The results are sorted by R2 score. Despite the global results obtained, individual folds can yield specific results corresponding to the data they are trained and tested on. Table 8 shows the regressor and imputation method combinations for each fold corresponding to the best R2 score, MSE and MAE. It can be observed that LSTM + time-based showed the best performance for two folds (folds 2 and 3), whereas for fold 3 FTTransformer+knn was the best for all metrics, and for fold 1 FTTransformer+mean was the best combination for the R2 score and MSE and LSTM+knn was the best combination for MAE. This shows that the best-performing imputation and regression depends on the data being trained and tested. As an example, the training data for fold 4 comprises years 2021, 2022 and 2024, which have the most missing values. In this case, LSTM did not fare better in comparison to FTTransformer.

Table 9 and Figure 4 show the performance of all the regressors across all the imputation methods and folds, with LSTM performing the best in all metrics, followed by FTTransformer and lastly RF (except for MAE). Table 10 and Figure 5 show the performance of all the imputation methods across all the regressors and folds. The time-based interpolation and mean imputation performed the best across all metrics, followed by knn and TCYC. Every imputation method outperformed the absence of imputation. The TCYC algorithm did not surpass the other traditional imputation methods, primarily because previous years’ data may not be directly representative of a given year. This is an important finding that highlights variable yearly climate conditions, and could have direct consequences on developing future forecasting models.

Table 7. All regression–imputation combinations per fold (sorted by R2 score).

Fold	Regressor	Imputation	R2	MSE	MAE
1	FTTransformer	mean	0.9989	0.1096	0.2183
1	LSTM	no	0.9989	0.1118	0.2195
1	LSTM	time-based	0.9989	0.1174	0.2211
1	LSTM	mean	0.9988	0.1268	0.2336
1	LSTM	knn	0.9987	0.1281	0.2179
1	LSTM	TCYC	0.9987	0.1331	0.2305
1	FTTransformer	TCYC	0.9987	0.1344	0.2337
1	FTTransformer	time-based	0.9982	0.1862	0.3003
1	RF	TCYC	0.9981	0.1935	0.2823
1	RF	knn	0.9981	0.1935	0.2827
1	RF	time-based	0.9981	0.1941	0.2830
1	RF	no	0.9981	0.1942	0.2830
1	RF	mean	0.9981	0.1944	0.2838
1	FTTransformer	no	0.9980	0.2088	0.3247
1	FTTransformer	knn	0.9979	0.2183	0.3508
2	LSTM	time-based	0.9990	0.1115	0.2138
2	LSTM	knn	0.9989	0.1182	0.2264
2	FTTransformer	time-based	0.9989	0.1189	0.2460
2	LSTM	mean	0.9988	0.1236	0.2252
2	LSTM	TCYC	0.9988	0.1260	0.2211
2	FTTransformer	no	0.9988	0.1299	0.2251
2	FTTransformer	mean	0.9988	0.1332	0.2533
2	LSTM	no	0.9988	0.1334	0.2392
2	FTTransformer	TCYC	0.9987	0.1383	0.2874
2	RF	no	0.9986	0.1535	0.2560
2	RF	time-based	0.9986	0.1539	0.2564
2	RF	knn	0.9986	0.1541	0.2563
2	RF	mean	0.9986	0.1542	0.2564
2	RF	TCYC	0.9986	0.1546	0.2568
2	FTTransformer	knn	0.9983	0.1775	0.2807
3	LSTM	time-based	0.9988	0.1047	0.2073
3	LSTM	TCYC	0.9988	0.1085	0.2100
3	LSTM	knn	0.9988	0.1109	0.2144
3	LSTM	no	0.9986	0.1227	0.2220
3	FTTransformer	mean	0.9985	0.1318	0.2296
3	FTTransformer	time-based	0.9985	0.1329	0.2387
3	LSTM	mean	0.9985	0.1364	0.2362

Table 7. *Continued.*

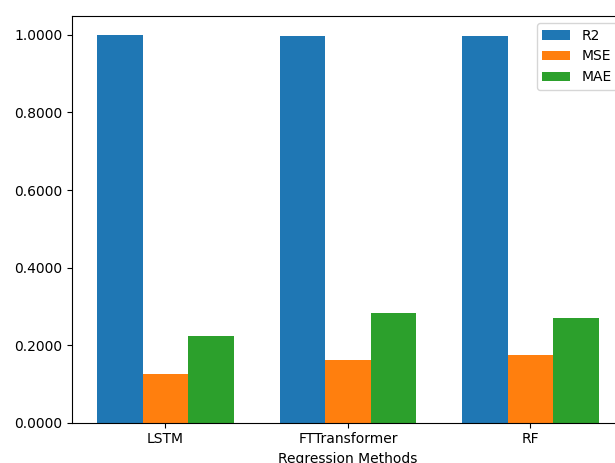
Fold	Regressor	Imputation	R2	MSE	MAE
3	RF	TCYC	0.9982	0.1646	0.2654
3	RF	no	0.9982	0.1647	0.2653
3	RF	time-based	0.9982	0.1652	0.2656
3	RF	knn	0.9982	0.1656	0.2660
3	RF	mean	0.9981	0.1661	0.2661
3	FTTransformer	knn	0.9980	0.1794	0.3155
3	FTTransformer	TCYC	0.9976	0.2182	0.3232
3	FTTransformer	no	0.9975	0.2235	0.3686
4	FTTransformer	knn	0.9990	0.1052	0.2118
4	FTTransformer	time-based	0.9987	0.1281	0.2416
4	LSTM	mean	0.9987	0.1317	0.2180
4	LSTM	time-based	0.9987	0.1378	0.2327
4	LSTM	knn	0.9986	0.1378	0.2275
4	LSTM	no	0.9986	0.1394	0.2225
4	LSTM	TCYC	0.9986	0.1428	0.2222
4	FTTransformer	mean	0.9984	0.1592	0.3155
4	FTTransformer	no	0.9984	0.1682	0.3291
4	RF	time-based	0.9982	0.1870	0.2739
4	RF	knn	0.9982	0.1873	0.2741
4	RF	no	0.9982	0.1874	0.2737
4	RF	TCYC	0.9982	0.1878	0.2738
4	RF	mean	0.9982	0.1878	0.2747
4	FTTransformer	TCYC	0.9978	0.2239	0.3710

Table 8. Regression and imputation method combinations for each best-performing metric for each fold.

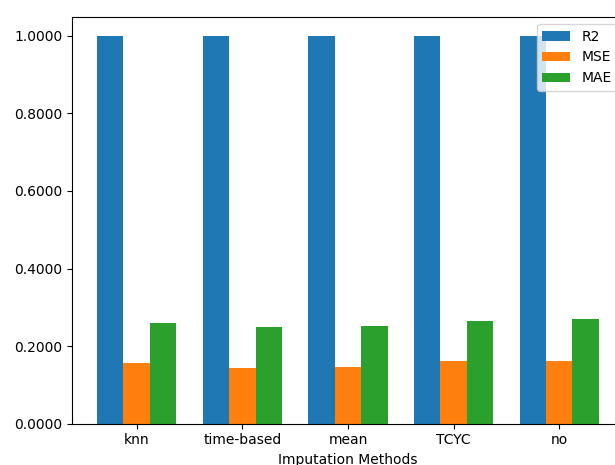
	Best Metric	Regressor + Imputation	Metric
Fold 1	R2	FTTransformer + mean	0.9989
	MSE	FTTransformer + mean	0.1096
	MAE	LSTM + knn	0.2179
Fold 2	R2	LSTM + time-based	0.9990
	MSE	LSTM + time-based	0.1115
	MAE	LSTM + time-based	0.2138
Fold 3	R2	LSTM + time-based	0.9988
	MSE	LSTM + time-based	0.1047
	MAE	LSTM + time-based	0.2073
Fold 4	R2	FTTransformer + knn	0.9990
	MSE	FTTransformer + knn	0.1052
	MAE	FTTransformer + knn	0.2118

Table 9. Average performance by regression methods (across all imputation methods and folds).

Imputation Methods	Performance Metrics		
	R2	MSE	MAE
LSTM	0.9987	0.1251	0.2230
FTTransformer	0.9984	0.1613	0.2832
RF	0.9982	0.1752	0.2698

**Figure 4.** Graphical representation of average performance by regression methods (across all imputation methods and folds).**Table 10.** Average performance by imputation methods (across all regression methods and folds).

Imputation Methods	Performance Metrics		
	R2	MSE	MAE
knn	0.9984	0.1563	0.2603
time-based	0.9986	0.1448	0.2484
mean	0.9985	0.1462	0.2509
TCYC	0.9984	0.1605	0.2648
no	0.9984	0.1615	0.2691

**Figure 5.** Graphical representation of average performance by imputation methods (across all regression methods and folds).

3.3. Discussion

Although this study is grounded in climate data from Kuwait, the proposed imputation and evaluation framework is not location specific. The methods evaluated in this work are applicable to high-frequency, multi-dimensional environmental data sets commonly encountered in other urban climate monitoring, systems, meteorological sensor networks, and environmental observation systems. In particular, regions experiencing harsh climates, sparse sensor coverage, or frequent data gaps may benefit from systematic imputation strategies to improve downstream predictive modeling. The reliable imputation of missing high-frequency climate data has practical implications for heat stress monitoring, urban micro-climate analysis, and early warning systems, where incomplete sensor data can compromise timely and accurate decision making.

The experimental results demonstrate that imputing missing values generally improves the prediction performance of multi-output regression models for both air and surface temperatures. This supports the hypothesis that handling missingness explicitly is critical when working with high-frequency, multi-dimensional climate data sets. However, the results also indicate the variation in performance using different imputation techniques, highlighting the complexity of climate data dynamics. Previous studies have emphasized the importance of data completeness in climate modeling, often focusing on low-frequency or single-variable data sets. In contrast, this study extends earlier work by evaluating imputation strategies under high-frequency, multi-variate and multi-output settings, which are increasingly common with modern sensor deployments. The observed improvements in downstream prediction align with prior findings that emphasize the role of preprocessing in climate analytics. However, the absence of a universally superior imputation method contrasts with studies conducted on lower-dimensional or less temporally dense data sets, indicating that high-frequency, multi-variate climate data pose additional challenges that warrant further investigation.

The imputation methods used in this work could have limitations due to the choice of parameters. In the case of knn imputation, $k = 5$ was used. Larger neighborhood sizes or alternative distance metrics, such as temporal or seasonally weighted similarities, may yield improved performance, albeit at the cost of potentially high computational complexity. These choices, however, introduce additional hyperparameters that require careful tuning in high-frequency and multi-dimensional climate data settings. Averaging the same timestamp of previous years in TCYC and the annual mean imputation provides a lower-bound reference for this problem. While seasonally stratified baselines (e.g., monthly or weekly averages) may better preserve the underlying distribution in climate data, they introduce additional modeling choices related to temporal resolution and smoothing. For interpolation imputation, other complex methods can be used (e.g., polynomial and splines); however, their impact on over-fitting needs to be assessed.

This work contributes to the United Nations' Sustainable Development Goals, particularly SDG 13 (Climate Action), by supporting more robust climate data analytics that can improve climate monitoring and forecasting under real-world data constraints. Additionally, the findings are relevant to SDG 11 (Sustainable Cities and Communities), as accurate urban temperature prediction is critical for managing heat-related risks and supporting climate-resilient urban planning.

4. Conclusions and Future Work

High-frequency climate data sets are increasingly used in climate studies to support fine-grained environmental monitoring, urban heat assessment, and predictive modeling; however, their utility is often compromised by missing data caused by sensor faults, communication failures, or maintenance interruptions. This study demonstrated that such

missingness degrades the performance of multi-output temperature prediction models and that appropriate imputation is necessary to enhance downstream prediction task.

Using four years of high-frequency, multi-dimensional climate data from Kuwait, we evaluated several imputation algorithms under a multi-output regression setting. Across all experiments, imputing missing values consistently improved predictive performance compared to using incomplete data, confirming the central hypothesis that restoring temporal continuity is critical for reliable climate prediction. Among the evaluated methods, time-based interpolation combined with LSTM models yielded the best results, highlighting the effectiveness of simple, temporally aware imputation when paired with sequence-based learners. In contrast, the limited improvements observed for RF suggest that not all predictive models benefit equally from imputation, particularly when temporal dependencies are not explicitly modeled.

The time-based interpolation performed well in this study; future work could explicitly analyze its performance as a function of missing gap duration (e.g., short versus prolonged sensor outages). Such stratified analyses would help to assess the robustness of interpolation methods under extended data loss scenarios common in harsh climate environments. The TCYC did not outperform other imputation methods; however, its value lies in providing an interpretable, cycle-aware reference that highlights the benefits and limitations of simple temporal assumptions in high-frequency climate data. This finding underscores the need for caution when transferring historical information across years in a changing climate scenario. While TCYC was formulated as a simple cyclic-aware baseline, it does not currently account for local temporal trends surrounding missing periods. Future extensions could integrate trend-aware mechanisms, such as incorporating moving averages or weighted windows over adjacent days to correct inter-annual differences. However, determining appropriate window sizes and weighting strategies introduces additional hyperparameters that may vary across climate regimes and sampling frequencies. Exploring adaptive or data-driven approaches for trend correction while preserving interpretability represents an important direction for extending TCYC in the context of climate change and increasing inter-annual variability. Future work needs to also explore the adaptive selection of k and time series-aware distance measures to further enhance knn imputation for climate data. Season-aware mean imputation and non-linear imputations should also be investigated to understand their impact on improving high-frequency climate modeling problems.

The findings from this work further emphasize the importance of designing data pipelines that explicitly address missingness in high-frequency environmental data, as well as the value of combining imputation strategies with models that can exploit temporal structures. From an applied perspective, improved data continuity can enhance downstream applications, such as heat stress monitoring, early warning systems, and urban climate analytics, contributing indirectly to climate resilience and sustainability goals. This also implies developing robust sensors that can withstand adverse weather conditions to attenuate missingness in future climate data. Future work could explore the translation of various imputation strategies to spatially distributed or hybrid sensing environments, including reanalysis, urban and remote sensing data sets.

Furthermore, future work involves the development of ameliorated forecasting models, i.e., accurately predicting future air and surface temperatures (or other climate parameters) for both shorter and longer horizons. Imputing missing data for previous years could improve the overall predictability of future climate variables. Advanced imputation techniques based on recent DL algorithm, especially large language models [27], need further exploration. Another potential future direction is model-based imputation, where a deep

learning model can be trained on some parts of the data and used to predict future climate variables values, thus helping in imputing gaps in climate data.

Author Contributions: Conceptualization, S.S.K. and R.A.-H.; methodology, S.S.K.; software, S.S.K.; validation, S.S.K. and R.A.-H.; resources, R.A.-H.; data curation, R.A.-H.; writing—original draft preparation, S.S.K.; and writing—review and editing, S.S.K. and R.A.-H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data that support the findings of this study are available from the American University of the Middle East, Kuwait, but restrictions apply to the availability of these data, which were used under a license for the current study and so are not publicly available. The data are, however, available from the authors upon reasonable request and with permission of the American University of the Middle East, Kuwait.

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

References

1. Khan, S.S.; Al-Hajj, R. Multi-output deep learning for high-frequency prediction of air and surface temperature in Kuwait. *Sci. Rep.* **2025**, *15*, 39027.
2. Zwiers, F.W.; Alexander, L.V.; Hegerl, G.C.; Knutson, T.R.; Kossin, J.P.; Naveau, P.; Nicholls, N.; Schär, C.; Seneviratne, S.I.; Zhang, X. Climate extremes: Challenges in estimating and understanding recent changes in the frequency and intensity of extreme climate and weather events. In *Climate Science for Serving Society: Research, Modeling and Prediction Priorities*; Springer: Dordrecht, The Netherlands, 2013; pp. 339–389.
3. Mateo, E.I.; Mark, B.G.; Hellström, R.Å.; Baraer, M.; McKenzie, J.M.; Condom, T.; Rapre, A.C.; Gonzales, G.; Gómez, J.Q.; Encarnación, R.C.C. High temporal resolution hydrometeorological data collected in the tropical Cordillera Blanca, Peru (2004–2020). *Earth Syst. Sci. Data* **2022**, *14*, 2865–2882.
4. Žliobaitė, I.; Hollmén, J.; Junninen, H. Regression models tolerant to massively missing data: A case study in solar-radiation nowcasting. *Atmos. Meas. Tech.* **2014**, *7*, 4387–4399.
5. Hayawi, K.; Shahriar, S.; Hacid, H. Climate data imputation and quality improvement using satellite data. *J. Data Sci. Intell. Syst.* **2025**, *3*, 87–97.
6. Zech, C.; Schöne, T.; Illigner, J.; Stolarczuk, N.; Queißer, T.; Köppl, M.; Thoss, H.; Zubovich, A.; Sharshebaev, A.; Zakhidov, K.; et al. Hydrometeorological data from a remotely operated multi-parameter station network in central Asia. *Earth Syst. Sci. Data* **2021**, *13*, 1289–1306.
7. Varotsos, K.V.; Katavoutas, G.; Giannakopoulos, C. On the use of reanalysis data to reconstruct missing observed daily temperatures in Europe over a lengthy period of time. *Sustainability* **2023**, *15*, 7081.
8. Wang, P.; Zhang, H.; Cheng, S.; Zhang, T.; Lu, F. Quickly forecasting the future state of urban sensors by the missing-data-tolerant deep learning approach. *Sustain. Cities Soc.* **2025**, *118*, 106044.
9. Shen, H.; Li, X.; Cheng, Q.; Zeng, C.; Yang, G.; Li, H.; Zhang, L. Missing information reconstruction of remote sensing data: A technical review. *IEEE Geosci. Remote Sens. Mag.* **2015**, *3*, 61–85.
10. Liu, Q.; Wang, Y.; Zhao, F.; Zheng, C.; Xie, J. A Review of the Research Progress of Sensor Monitoring Technology in Harsh Engineering Environments. *Sensors* **2025**, *25*, 6308.
11. Dong, X.; Li, D.; Chen, A.; Zheng, D. Humidity Sensing in Extreme Environments: Mechanisms, Materials, Challenges, and Future Directions. *Chemosensors* **2026**, *14*, 20.
12. Alejo-Sanchez, L.E.; Márquez-Grajales, A.; Salas-Martínez, F.; Franco-Arcega, A.; López-Morales, V.; Acevedo-Sandoval, O.A.; González-Ramírez, C.A.; Villegas-Vega, R. Missing data imputation of climate time series: A review. *MethodsX* **2025**, *15*, 103455.
13. Magyari-Sáska, Z.; Haidu, I.; Magyari-Sáska, A. Experimental Comparative Study on Self-Imputation Methods and Their Quality Assessment for Monthly River Flow Data with Gaps: Case Study to Mures River. *Appl. Sci.* **2025**, *15*, 1242.
14. Afrifa-Yamoah, E.; Mueller, U.A.; Taylor, S.M.; Fisher, A.J. Missing data imputation of high-resolution temporal climate time series data. *Meteorol. Appl.* **2020**, *27*, e1873.
15. Li, C.; Ren, X.; Zhao, G. Machine-learning-based imputation method for filling missing values in ground meteorological observation data. *Algorithms* **2023**, *16*, 422.
16. Dimitri, G.M.; Cappelli, I.; Scarselli, F.; Fort, A.; Gori, M. Graph Neural Networks for Missing Data Imputation in Time Series from Meteorological Sensors. In *Proceedings of the 2024 IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering (MetroXRINE)*; IEEE: New York, NY, USA, 2024; pp. 1242–1247.

17. Viana, D.; Teixeira, R.; Soares, T.; Baptista, J.; Pinto, T. Generative Adversarial Networks for Synthetic Meteorological Data Generation. In *Proceedings of the EPIA Conference on Artificial Intelligence*; Springer: Cham, Switzerland, 2024; pp. 197–206.
18. Kulanuwat, L.; Chantrapornchai, C.; Maleewong, M.; Wongchaisuwat, P.; Wimala, S.; Sarinnapakorn, K.; Boonya-Aroonnet, S. Anomaly detection using a sliding window technique and data imputation with machine learning for hydrological time series. *Water* **2021**, *13*, 1862.
19. Wang, L.; Zhang, Y. Filling GRACE data gap using an innovative transformer-based deep learning approach. *Remote Sens. Environ.* **2024**, *315*, 114465.
20. Shan, S.; Xie, X.; Fan, T.; Xiao, Y.; Ding, Z.; Zhang, K.; Wei, H. A deep-learning based solar irradiance forecast using missing data. *IET Renew. Power Gener.* **2022**, *16*, 1462–1473.
21. Xu, T.; Zhang, Y.; Zhang, C.; Abodoukayimu, A.; Ma, D. Deep learning tool: Reconstruction of long missing climate data based on spatio-temporal multilayer perceptron. *Theor. Appl. Climatol.* **2024**, *155*, 5835–5847.
22. Park, J.; Müller, J.; Arora, B.; Faybishenko, B.; Pastorello, G.; Varadharajan, C.; Sahu, R.; Agarwal, D. Long-term missing value imputation for time series data using deep neural networks. *Neural Comput. Appl.* **2023**, *35*, 9071–9091.
23. Boujoudar, M.; El Ydrissi, M.; Abraim, M.; Bouarfa, I.; El Alani, O.; Ghennioui, H.; Bennouna, E.G. Comparing machine learning algorithms for imputation of missing time series in meteorological data. *Neural Comput. Appl.* **2025**, *37*, 23773–23787.
24. Demirhan, H.; Renwick, Z. Missing value imputation for short to mid-term horizontal solar irradiance data. *Appl. Energy* **2018**, *225*, 998–1012.
25. Gorishniy, Y.; Rubachev, I.; Khrulkov, V.; Babenko, A. Revisiting deep learning models for tabular data. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 18932–18943.
26. SciPy. Interpolation. 2026. Available online: <https://docs.scipy.org/doc/scipy/tutorial/interpolate.html> (accessed on 3 February 2026).
27. Franch, G.; Tomasi, E.; Wanjari, R.; Poli, V.; Cardinali, C.; Alberoni, P.P.; Cristoforetti, M. GPTCast: A weather language model for precipitation nowcasting. *Geosci. Model Dev.* **2025**, *18*, 5351–5371.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.