Received: 17 March 2024 | Revised: 21 May 2024 | Accepted: 21 June 2024 | Published online: 26 June 2024

Journal of Data Science and Intelligent Systems
2025, Vol. 3(2) 87–97
DOI: 10.47852/bonviewJDSIS42022857

BON VIEW PUBLISHING

RESEARCH ARTICLE

Climate Data Imputation and Quality Improvement Using Satellite Data

Kadhim Hayawi[1], Sakib Shahriar[1,*] and Hakim Hacid[2]

[1]College of Technological Innovation, Zayed University, United Arab Emirates

[2]Technology Innovation Institute, United Arab Emirates

Abstract: Combating climate change has emerged as a global concern recently, and meteorological data remain an important measure for analyzing and predicting climate trends. However, ground weather stations and sensors can be impacted by faults due to accidents and unreliability, often resulting in, for example, missing data and lowering the overall quality of the data. This paper explores the impact of using satellite data as an input feature for machine learning algorithms. In particular, temperature, pressure, wind speed, and global horizontal radiation data are imputed using various machine learning algorithms to overcome potential data quality issues resulting from the ground stations. The results from two experiments highlight that the performance of the algorithms significantly increases by using satellite data as input features. For instance, the incorporation of satellite data improved the R^2 values for temperature prediction using Random Forest and XGBoost to 0.86 and 0.84, respectively, demonstrating a notable enhancement compared to models without satellite data. The paper discusses several implications of these findings and outlines future research directions to further enhance the predictive accuracy of meteorological data imputation using satellite inputs.

Keywords: meteorological data imputation, machine learning, weather data cleaning, solar irradiance forecasting, renewable energy, climate change

1. Introduction

In response to the growing challenges posed by climate change, countries around the world are increasingly investing in renewable energy and enhancing the quality of meteorological data. Recognizing the critical role of reliable weather and solar data in transitioning to sustainable energy sources, 159 countries signed a declaration at the 2023 United Nations climate change conference [1]. This declaration emphasizes the importance of ensuring robust climate control, advancing renewable energy technologies, and developing sustainable energy systems. Meteorological data or weather data have been widely used to study various impacts of climate change, including economic impacts [2, 3] and hydrological impacts [4]. Besides climate change, weather data are also used for weather forecasting [5] and urban resource management and planning [6]. Transitioning to renewable energy has been proposed by experts as an important tool to mitigate climate change [7], and solar energy is among the most utilized source of renewable energy [8]. In-depth planning and assessment of solar resources are required to integrate solar energy into the grid for reliable services. Countries in the solar belt region (between latitudes of 40°N and 40°S) receive a large amount of solar radiation, making them optimum for large-scale solar energy production. To fully leverage the benefits of meteorological and solar data, it is necessary to ensure the reliability and the quality of the captured data.

The reliability of ground weather stations and solar data can be, in practice, compromised in various forms. For instance, the sensors may not be detecting the observations accurately or may not be functional due to a lack of power [9]. Moreover, faulty data can be generated in data storage, transmission, and processing. Extreme weather events, including high heat and storms, can also decrease the reliability of ground sensors [10]. In addition to unreliable data, some of these extreme events may also cause the ground stations to not be operational for a certain period of time, resulting in missing time-series data. Therefore, both inaccurate data and missing data must be handled to ensure the reliability of ground weather data. The process of replacing missing data with substituted values is known as data imputation [11]. In essence, rectifying inaccurate data can also be considered a data imputation problem. This is because inaccurate values will first be removed from the dataset, resulting in missing values which will then be imputed.

Meteorological data and solar data are time-series data, that is, the measurements are associated with a timestamp specifying when the measurement is recorded. There are various approaches to time-series data imputation, for example, ARIMA [12]. Researchers have also utilized statistical approaches, including Monte Carlo Markov chains for imputing meteorological data [13]. Other approaches include mathematical modeling [14] and simulation [15]. Recently, machine learning-based approaches have gained popularity for time-series data imputation. To this end, researchers have utilized supervised learning [16], unsupervised learning [17], and deep learning [18]. These algorithms are data-driven and rely on input features, including historical data and other meteorological

[*]**Corresponding author:** Sakib Shahriar, College of Technological Innovation, Zayed University, United Arab Emirates. Email: sakib.shahriar@zu.ac.ae

measures [16], for model training. In this work, we explore the impact of using satellite weather data as an input feature to several machine learning models. To the best of our knowledge, existing works have not investigated the performance of machine learning imputation for meteorological data using satellite data as input features.

Generally, ground weather stations provide more accurate measurements than satellite data. Mendelsohn et al. [19] investigated the accuracy of satellite data in comparison to ground stations for various locations. The study concluded that satellite measurements for phenomenon such as precipitation may not be as accurate as ground stations, but measurements such as temperature are promising for satellite data. It can be safe to assume that there would be a correlation between the ground and satellite data, irrespective of the level of accuracy. Consequently, machine learning algorithms can learn the trend from satellite data to better predict and impute the missing ground station data. In this work, we will formulate two training approaches for machine learning algorithms. In the first approach, the machine learning algorithms will not have any satellite information as input features. In the second approach, we will add the corresponding satellite data for the measurement to be imputed. For instance, the temperature from the satellite will be an input feature when imputing ground temperature data. Following are the main contributions of this paper:

1) It proposes a novel approach for meteorological and solar data imputation using satellite measurements as input features.
2) It provides a comparison of several machine learning algorithms in imputing various meteorological data, including temperature and solar irradiation.
3) It discusses the implications of using satellite data for imputation and outlines future research directions.

The rest of the paper is organized as follows. Section 2 provides a concise literature review covering the existing works in weather and solar data imputation and forecasting. Section 3 describes the methodology, including data collection and experimental setup. Section 4 summarizes the results and Section 5 discusses the implications and future work. Finally, Section 6 concludes the paper.

2. Literature Review

Chivers et al. [20] proposed a machine learning approach for imputing precipitation data of UK weather stations. In addition to predicting the absolute values (regression), a classification model for predicting rain or no rain was also developed. XGBoost and random forest (RF) provided the best performance for classification, whereas neural network was more effective for regression. In Boomgard-Zagrodnik and Brown [21], the authors focused on the imputation of air temperature data using machine learning for weather stations in Washington state. A RF model trained with historical temperature and humidity features was effective for imputation, with a reported mean absolute error (MAE) of 0.43 °C. Doreswamy et al. [16] utilized various machine learning algorithms, including RF and support vector machines, to impute missing weather data. The data from National Climatic Data Center (NCDC), containing weather data from different sources around the globe, were used for their experiment. In terms of R^2, linear regression performed the best for imputing mean dew point and minimum temperature with R^2 values of 0.896 and 0.919, respectively. However, the models were not effective in predicting other weather parameters, such as wind speed and mean visibility. Similarly, Gad et al. [22] employed a deep learning-based approach to impute missing values in the NCDC dataset. The model was trained using data from neighboring weather stations that do not have missing values. The proposed convolutional neural network model obtained the best performance using a stochastic gradient descent optimizer with root mean squared error (RMSE) scores of 0.123 and 0.046 for dew point and minimum temperature, respectively. Kiani and Saleem [23] proposed a K-Nearest Neighbor (K-NN) algorithm for imputing missing surface temperature data. For their experiment, the authors use temperature data from 38 weather stations located in Pakistan. The proposed hybrid approach first identified a cluster of K years to obtain the nearest temperature trend to the missing value. The missing temperature was then imputed by taking the average value for the same date from the K years identified in the previous step.

The process of weather and solar forecasting is relevant to imputation. This is because a forecasting model is able to predict trends based on historical data. The same model can therefore be used to predict any missing data. Recently with the advances of Neural Networks, there has been increasing interest in applying deep networks to solar radiation forecasting. To predict the average daily solar radiation in Kuwait, Bou-Rabee et al. [24] used artificial neural networks (ANN) and achieved a 94.75% efficiency. Similarly, Hussain and AlAlili [25] proposed a hybrid modeling with ANN to estimate solar radiation in the UAE. Utilizing parameters such as temperature, humidity, wind speed, and sunshine duration, the study obtained a minimum RMSE of 2.78%. Kazem et al. [26] deployed machine learning models, including support vector machines, to predict the output current of photovoltaic systems. The input parameters included solar radiation and ambient temperature, and the reported mean squared error was 2.6%. Under clear-sky conditions, the solar output can be simply calculated based on the solar panel's position with respect to the sun, but the challenge arises with the presence of clouds and dust [27]. Most numerical methods that model solar radiation do not consider information about cloud presence. Therefore, Tuohy et al. [27] point out that for short-term forecasting, numerical weather prediction techniques are not as effective as sky imaging and satellite data. To emphasize the importance of renewable energy forecasting, particularly solar and wind, Notton et al. [28] present an in-depth review. Some of the main benefits of forecasting include a reduction in integration costs, a decrease in average annual costs, a decrease in reserve shortfalls, and an increase of percentage reduction in curtailments of PV systems. A review of data-driven approaches for weather forecasting presented by Fathi et al. [29] and Chantry et al. [30] discuss the challenges in the context of machine learning-based weather forecasting. Table 1 presents a comparison of the existing works in weather and solar data forecasting.

As shown in Table 1, existing works have effectively used machine learning algorithms for weather and solar data imputation and forecasting. However, the implications of using satellite data as input features for training the models were not studied by existing works to the best of our knowledge. Hence, the proposed work aims to investigate the effectiveness of satellite data as input features for training machine learning models for meteorological data imputation.

3. Methodology

3.1. Data collection

In this paper, two independent meteorological data sources were used. The first step was obtaining a reliable ground weather station dataset containing various meteorological and solar measurements. To this end, the Canadian Weather Energy and Engineering

Table 1
Comparison of related works

Study	Variable, data	Method	Key metrics	Key findings
Chivers et al. [20]	Precipitation, UK weather stations	XGBoost, Random Forest, Neural Networks	$R^2 = 0.35$ and $F1$ score = 0.92	Effective in predicting rain occurrence and precipitation.
Boomgard-Zagrodnik and Brown [21]	Air temperature, Weather stations in Washington state	Random Forest	MAE of 0.43 °C.	Effective in imputing air temperature.
Doreswamy et al. [16]	Various weather, NCDC global weather data	Random Forest, SVM, Linear Regression	$R^2 = 0.896$, 0.919	Effective for mean dew point and temperature; less effective for wind speed and visibility.
Gad et al. [22]	Various weather, NCDC global weather data	Convolutional Neural Network	RMSE = 0.123, 0.046	Effective in imputing dew point and temperature using neighboring station.
Kiani and Saleem [23]	Surface temperature, Stations in Pakistan	K-Nearest Neighbors	RMSE = 2.2	Utilized temperature trends over years to impute missing surface temperature.
Bou-Rabee et al. [24]	Solar radiation data in Kuwait	Artificial Neural Networks	94.75% efficiency	High efficiency in predicting daily solar radiation.
Hussain and AlAlili [25]	Solar radiation data in the UAE	Hybrid ANN model	Minimum RMSE = 2.78%	Effective in estimating solar radiation using various climatic parameters.
Kazem et al. [26]	Photovoltaic current, Research Lab in Sohar University, Oman.	Support Vector Machines	MSE = 2.6%	Effective in predicting photovoltaic system output using solar radiation and temperature.
Tuohy et al. [27]	Solar radiation forecasting	Review of Numerical methods, sky imaging, satellite	N/A	Discusses challenges in solar radiation forecasting under varying sky conditions.
Notton et al. [28]	Renewable energy forecasting	Review	N/A	Highlights the benefits of forecasting in reducing costs and operational challenges in PV systems.
Fathi et al. [29]	Weather forecasting	Review of data-driven approaches	N/A	Highlighted various data-driven techniques for weather forecasting.
Chantry et al. [30]	Weather forecasting	Review	N/A	Discusses challenges in applying machine learning to weather forecasting.

Datasets (CWEEDS) [31] containing data from 492 Canadian locations with at least ten years of data between 1998 and 2014 were utilized. Specifically, the historical data for Toronto International Airport for the year 2014 were obtained for the proposed experiment. The dataset contained hourly measurements of various weather variables. For simplification, four measurements were included in this study, namely temperature, pressure, wind speed, and global horizontal radiation.

To obtain the satellite data, NASA's Modern-Era Retrospective analysis for Research and Applications, Version 2 (MERRA-2) [32], was used. MERRA-2 provides various satellite meteorological measurements for any given coordinates. Therefore, the coordinates for the Toronto International Airport weather station (Latitude 43.68, Longitude −79.63) were used to obtain the corresponding satellite measurements. The temperature was recorded at 2 meters above ground, and the wind speed was recorded at 10 meters above the ground. The pressure was recorded at ground level, and radiation was the surface incoming shortwave irradiation.

3.2. Data processing

The CWEEDS data are only provided in WY3 (.wy3) data format, which is not compatible with most applications. Consequently, the methodology proposed by Siu and Liao [33] was employed to convert the data files into EnergyPlus Weather File (.epw) format, which was then converted to CSV format using the open-source EnergyPlus software[1]. Information such as leap year observed and the number of holidays were not needed for the experiment and were removed. The dataset did not contain any missing values.

The satellite data were timestamped using universal time. Consequently, the pytz library in Python [34] was used to convert the data into the local (Toronto) time zone. Pandas library [35] was used to concatenate the date and time columns and merge the two datasets. The satellite temperature was recorded in Kelvin and was converted to degree Celsius to match the ground station data. Moreover, the pressure measured in satellite data was in hectopascal, which was converted to pascals to match the ground station unit. In addition to the temperature and pressure, the wind speed was measured in m/s and global horizontal radiation was measured in Wh/m^2.

To obtain temporal features, hour, day, and month features were extracted from the timestamp columns. These features are however cyclic ordinal features. For instance, the value of 12 representing December is close to the value of 1 representing January. This information is not provided to the models explicitly. Therefore,

[1] https://energyplus.net/downloads

trigonometric transformation was utilized [36]. The final dataset contained eleven attributes: three temporal features and eight weather measures (four from each of the two weather sources).

3.3. Algorithms

Machine learning algorithms can be used to train computer systems to learn from data without explicit programming. In this work, the algorithms will be trained to model a specific weather phenomenon. Given that we have the measured values, this is a supervised learning problem. Since all the values to be predicted (weather and solar measurements) are numerical, the problem is specifically a supervised regression problem. Four popular supervised machine learning algorithms were experimented with to provide a comprehensive assessment.

K-NN [37] is a lazy learning algorithm that does not require a dedicated training phase. To make a prediction for a given data point, a distance measure such as Euclidean or Manhattan is employed to identify its K-NNs from the dataset. In the context of regression tasks, the average value of these k neighbors is calculated and used as the prediction for the target data point. For our experiments, we specifically chose a value of k equal to 4, allowing us to balance the trade-off between noise reduction and the accuracy of the predictions.

RF is an ensemble learning technique that combines predictions of multiple decision trees. Each decision tree in the ensemble functions akin to a flow chart, systematically breaking down complex decisions into a series of simpler decisions based on split points derived from the input features [38]. This process of decision-making in trees is based on selecting the best split at each node to maximize the homogeneity of the resultant subgroups. The RF algorithm improves on the decision tree model by creating an ensemble of trees, each trained on a random subset of the data and features. This randomness helps in reducing the model's susceptibility to overfitting on the training data. The aggregation, or "voting," across these multiple decision trees is typically performed by taking the average value of the predictions for regression tasks or the majority vote for classification tasks [39].

A gradient boosting algorithm such as XGBoost [40] also uses multiple decision trees to make predictions. However, unlike RF, which embodies a bagging approach to ensemble learning where decision trees are constructed in parallel, boosting algorithms such as XGBoost deploy trees in a sequential manner. This sequential construction is pivotal because each new tree in the sequence specifically addresses the errors committed by the preceding trees. Consequently, subsequent trees in the sequence incrementally improve the model's performance by focusing on the harder-to-predict instances that earlier trees struggled with [38]. This methodical focus on correcting previous errors allows gradient boosting algorithms to often achieve higher accuracy than bagging techniques, albeit potentially at the cost of increased computational complexity.

The final model we experimented with was a multilayer perceptron (MLP), a sophisticated type of ANN that falls under the broader category of deep learning algorithms. An MLP is composed of multiple layers: it starts with an input layer, followed by one or more hidden layers, and concludes with an output layer. Each layer is made up of neurons that use non-linear activation functions, allowing the network to capture complex patterns and relationships in the data [38]. In our specific application, the MLP regressor was configured with 5 hidden layers, enabling it to perform non-linear regression with enhanced depth and complexity. For simplicity and clarity in further discussions, we will refer to the MLP as ANN.

3.4. Experimental setup

The research framework in this work is displayed in Figure 1. The data collection and pre-processing steps were explained in the previous sections. To compare the effectiveness of using satellite data for imputation, two different machine learning experiments were conducted.

The first experiment trains independent models to predict the different weather attributes with and without satellite data as input features. In this context, K-fold cross-validation was used to measure the performance across the dataset. The value of K was set to 10, implying that the algorithms are repeatedly trained ten times with a fraction of 1/10 training examples left out for testing. This approach provides a more general evaluation of the dataset without bias in selecting a test subset. After training all the algorithms, it is possible to compare any difference in performance as a result of using satellite data as input features.

In the second experiment, the objective is to conduct a more in-depth analysis of the performance. Consequently, three weeks across the dataset were identified as test cases, including one week each from February, June, and October. The models were trained using the same algorithms in the first experiment. However, the training process did not include these three weeks of test cases. After training, the algorithms predicted all weather attributes for these three weeks. The difference in performance using satellite data was then analyzed for these three weeks in terms of evaluation metrics and visualization.

For each model building, the hourly, monthly, and daily values were used as inputs to the model. In addition, the experiments were repeated using satellite data to observe performance improvements (if any) for the specific weather variable. For instance, when imputing temperature values, the machine learning models were only trained using the temporal variables (six features for the hour, month, and day obtained by trigonometric transformation). After this, the variable of interest from the satellite was added for each experiment, bringing the total feature to seven. For example, we used the satellite temperature along with the six temporal features for temperature imputation in the second set of experiments.

3.5. Evaluation

To quantify the performance of each experiment, three popular regression metrics were used. The metrics are MAE, RMSE, and coefficient of determination (R^2). The metrics are defined in Equations (1)–(3).

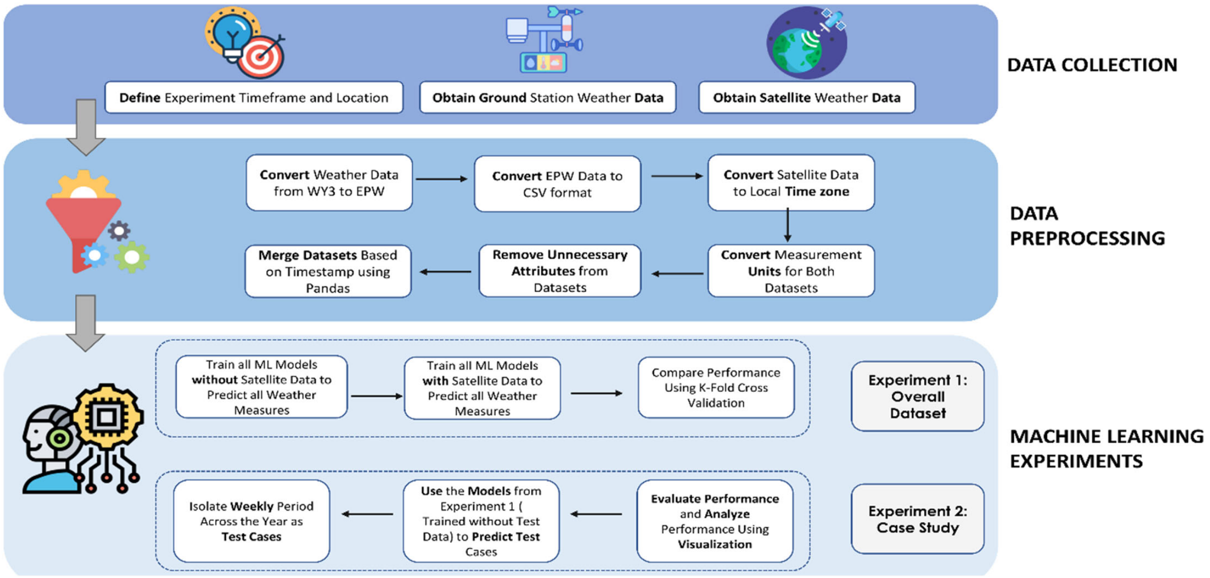
$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{Y}_i\right| \tag{1}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{Y}_i\right)^2}{n}} \tag{2}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{Y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \mu\right)^2} \tag{3}$$

where n represents all values in the dataset, Y_i represents the real value, $\hat{Y}$ represents the predicted value, and μ represents the mean of the real values. A lower value of MAE and RMSE indicates lower error rate, and hence better performance. On the other hand, a higher value for R^2 indicates better model performance.

Figure 1
Proposed research framework


DATA COLLECTION
Define Experiment Timeframe and Location
Obtain Ground Station Weather Data
Obtain Satellite Weather Data
DATA PREPROCESSING
Convert Weather Data from WY3 to EPW
Convert EPW Data to CSV format
Convert Satellite Data to Local Time zone
Convert Measurement Units for Both Datasets
Remove Unnecessary Attributes from Datasets
Merge Datasets Based on Timestamp using Pandas
MACHINE LEARNING EXPERIMENTS
Train all ML Models without Satellite Data to Predict all Weather Measures
Train all ML Models with Satellite Data to Predict all Weather Measures
Compare Performance Using K-Fold Cross Validation
Experiment 1: Overall Dataset
Isolate Weekly Period Across the Year as Test Cases
Use the Models from Experiment 1 (Trained without Test Data) to Predict Test Cases
Evaluate Performance and Analyze Performance Using Visualization
Experiment 2: Case Study

4. Results

In this section, the results from the experiments are presented. First, the results using 10-fold cross-validation are presented to study the overall impacts of using satellite data as input features. Next, three case scenarios are developed to analyze the results in terms of visualization.

4.1. Overall results

In the first part of the experiment, all four machine learning models were trained without any satellite information. Next, the experiments were repeated with each model trained with an additional feature, representing the satellite data for the specific weather phenomenon that was being predicted. The average results across the 10-fold cross-validation for both experiments in predicting temperature are summarized in Table 2.

The results indicate that across all models, using satellite information provides better performance for predicting temperature values. The best performance was obtained using RF and XGBoost with R^2 values of 0.86 and 0.84, respectively. Next, the performance for predicting pressure is presented in Table 3.

In terms of predicting atmospheric pressure, the models perform worse than predicting temperature. Overall, the results indicate that across all models, using satellite information provides better performance. The best performance was obtained using RF and XGBoost with R^2 values of 0.44 and 0.45, respectively. Next, the performance for predicting wind speed is presented in Table 4.

The models perform worse in predicting wind speed than both temperature and pressure. Overall, using satellite information provides better performance. The best performance was obtained using MLP and XGBoost with R^2 values of 0.33 and 0.32, respectively. The performance for predicting global horizontal radiation is compared in Table 5.

The models perform reasonably well for predicting global horizontal radiation. Three of the four models performed best with satellite as input features with the best performance coming from RF and XGBoost with R^2 scores of 0.80 and 0.79, respectively. The only model that performed better without satellite information is K-NN.

Table 2
Imputation of temperature results

	MAE	RMSE	R^2
K-NN	6.10	8.60	0.21
K-NN (With Satellite)	5.43	8.55	0.24
RF	5.56	7.42	0.35
RF (With Satellite)	2.53	3.48	0.86
XGBoost	4.64	5.84	0.64
XGBoost (With Satellite)	2.78	3.75	0.84
MLP	6.15	7.75	0.36
MLP (With Satellite)	4.51	6.01	0.62

Table 3
Imputation of pressure results

	MAE	RMSE	R^2
K-NN	613.9	771.6	−0.13
K-NN (With Satellite)	466.2	688.0	0.11
RF	651.9	820.0	−0.28
RF (With Satellite)	328.1	536.2	0.44
XGBoost	590.6	730.4	0.04
XGBoost (With Satellite)	340.3	537.8	0.45
MLP	386.3	699.5	0.06
MLP (With Satellite)	99172.0	99175.0	−19015.5

4.2. Case study scenarios

In this experiment, we isolated three weeks from the dataset to be used as the test set. All four algorithms were trained using the entire dataset except for these three weeks. The first test case scenario was between February 7 and February 13, 2014. Table 6

Table 4
Imputation of wind speed results

	MAE	RMSE	R^2
K-NN	2.07	2.70	−0.14
K-NN (With Satellite)	1.80	2.37	0.11
RF	2.18	2.82	−0.23
RF (With Satellite)	1.65	2.16	0.26
XGBoost	1.99	2.54	−0.00
XGBoost (With Satellite)	1.58	2.07	0.32
MLP	1.92	2.48	0.05
MLP (With Satellite)	1.57	2.06	0.33

Table 5
Imputation of global radiation results

	MAE	RMSE	R^2
K-NN	65.3	123.3	0.70
K-NN (With Satellite)	75.4	137.2	0.63
RF	63.2	123.2	0.69
RF (With Satellite)	52.6	100.5	0.80
XGBoost	70.0	111.3	0.75
XGBoost (With Satellite)	61.5	102.5	0.79
MLP	96.8	145.4	0.57
MLP (With Satellite)	74.8	132.2	0.65

Table 6
Test case 1 R^2 comparison

	Temperature	Pressure	Wind	Radiation
K-NN	0.82	−1.12	−0.27	0.67
K-NN (With Satellite)	0.68	0.21	−0.16	0.55
RF	0.92	0.26	−0.28	0.76
RF (With Satellite)	0.91	0.39	0.00	0.81
XGBoost	0.87	−0.12	−0.16	0.77
XGBoost (With Satellite)	0.93	0.47	0.16	0.81
MLP	0.50	−16000	−0.36	0.60
MLP (With Satellite)	0.73	0.16	0.01	0.55

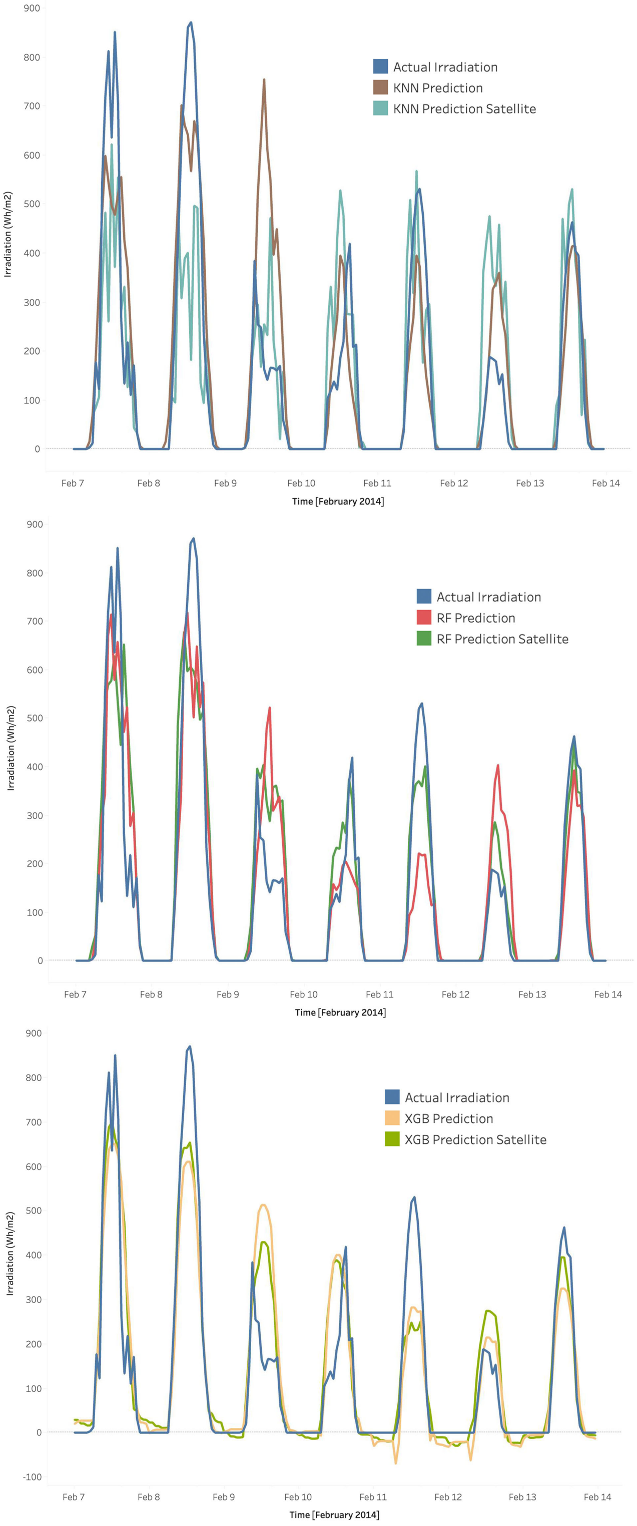
presents the performance comparison for the first test case. For simplicity, only R^2 is reported for each weather parameter.

The best results were obtained using XGBoost with satellite information. The imputation of pressure and wind speed were considered the most challenging, with many negative R^2 values. In Figure 2, the results for the first week for imputing global horizontal radiation are presented. Each plot presents the imputation using one model with and without satellite data along with the actual radiation values for that week.

The graphs highlight that XGBoost and RF are able to capture the global horizontal radiation trend best for this week using satellite information. ANN tends to underestimate the radiation trend with and without satellite information.

The second test case scenario was between June 15 and June 21, 2014. Table 7 compares the R^2 for imputing each of the weather parameters for this week using machine learning models.

Figure 2
Comparison of actual and machine learning imputed radiation with and without satellite information



The best performance for all weather attributes is obtained using satellite information. ANN performed the best for pressure and RF performed the best for temperature, wind speed, and radiation. The visualization for all four models in predicting temperature for this week is presented in Figure 3.

Figure 2
(Continued)

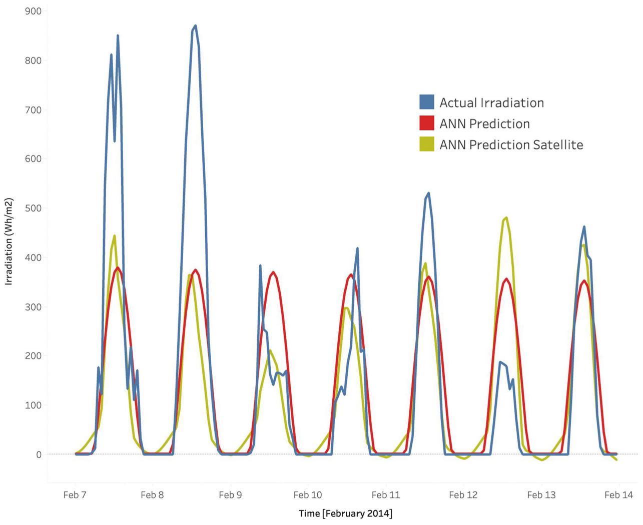


Table 7
Test case 2 R^2 comparison

	Temperature	Pressure	Wind	Radiation
K-NN	−0.84	−1.08	−0.43	0.84
K-NN (With Satellite)	0.59	0.37	0.09	0.82
RF	0.29	−1.87	−0.29	0.81
RF (With Satellite)	0.73	−4.79	0.71	0.91
XGBoost	0.45	−0.04	−0.19	0.77
XGBoost (With Satellite)	0.61	0.66	0.26	0.86
MLP	−0.69	−91000	−0.12	0.67
MLP (With Satellite)	0.57	0.79	0.23	0.79

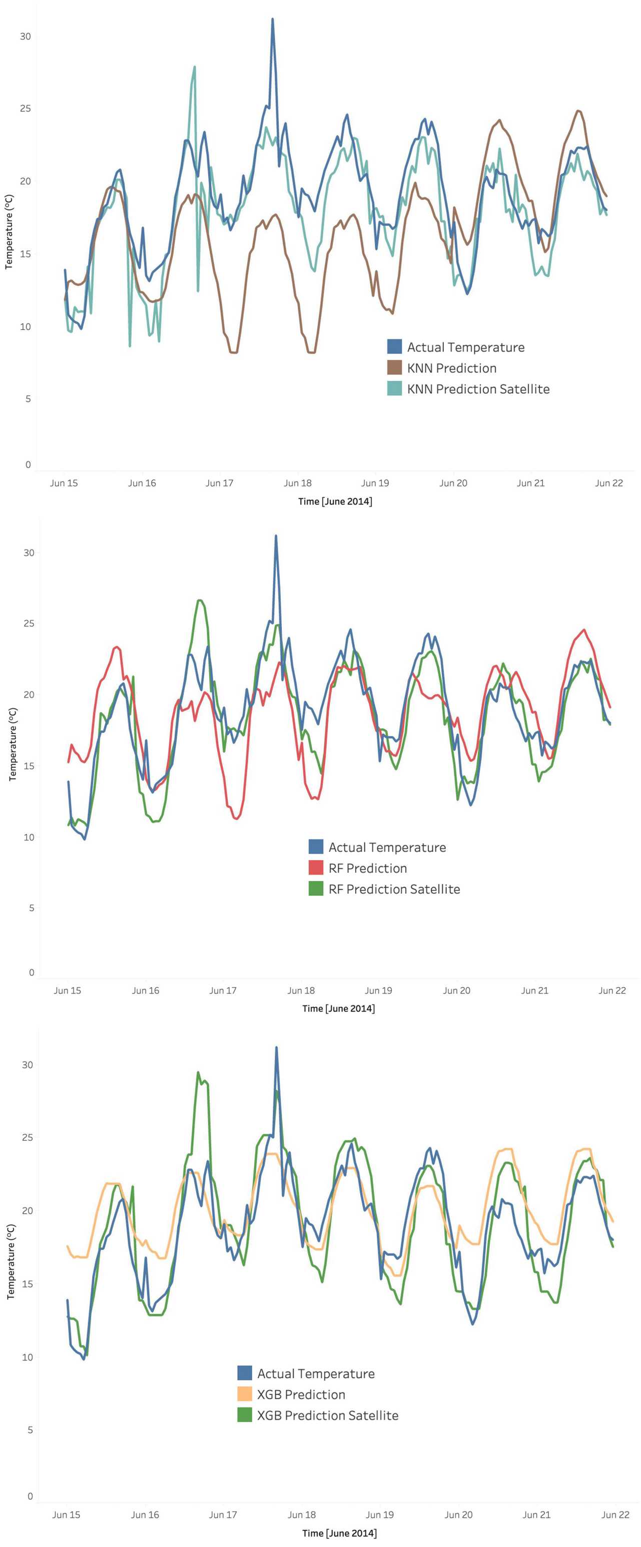
The predictions made by K-NN, RF, and ANN overestimate the temperature without satellite information. On the other hand, the predictions made by XGBoost without satellite information slightly overestimate the temperature. The graphs indicate that XGB was able to predict the temperature most accurately for this week using satellite information.

The third test case scenario was between October 23 and October 29, 2014. The R^2 scores for all the models are summarized in Table 8.

The best performance for all weather parameters is obtained using satellite information. RF provided the best performance for temperature, wind speed, and radiation. The best performance for pressure was obtained using XGBoost. The visualization for predicting wind speed for this week is provided in Figure 4.

The graphs show that without using satellite information, the models tend to underpredict the wind speed. Moreover, the performance of RF, XGB, and ANN with satellite information is very close to the actual wind speed for this week.

Figure 3
Comparison of actual and machine learning imputed temperature with and without satellite information



5. Discussion

The results from both experiments indicate that machine learning algorithms are more effective in imputing meteorological and solar data using satellite information as input features. Furthermore, the imputation performance for temperature and global radiation is overall better as compared to pressure and wind speed. The results also indicate that the ensemble machine learning models, RF and XGBoost, perform better than ANN and K-NN. Next, some of the implications of the proposed work and future work are outlined.

Figure 3
(Continued)

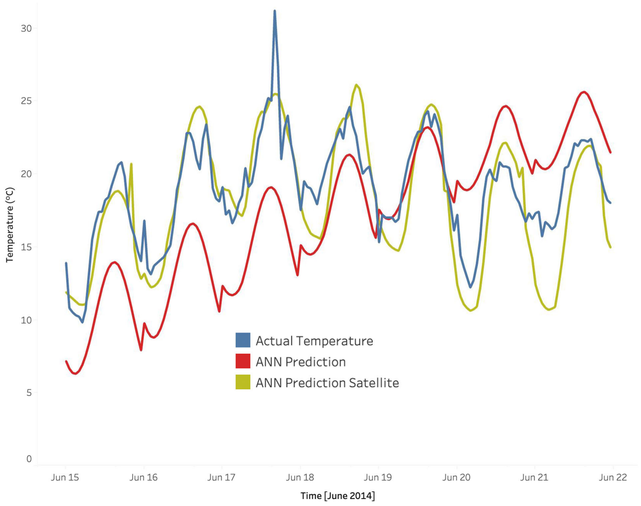


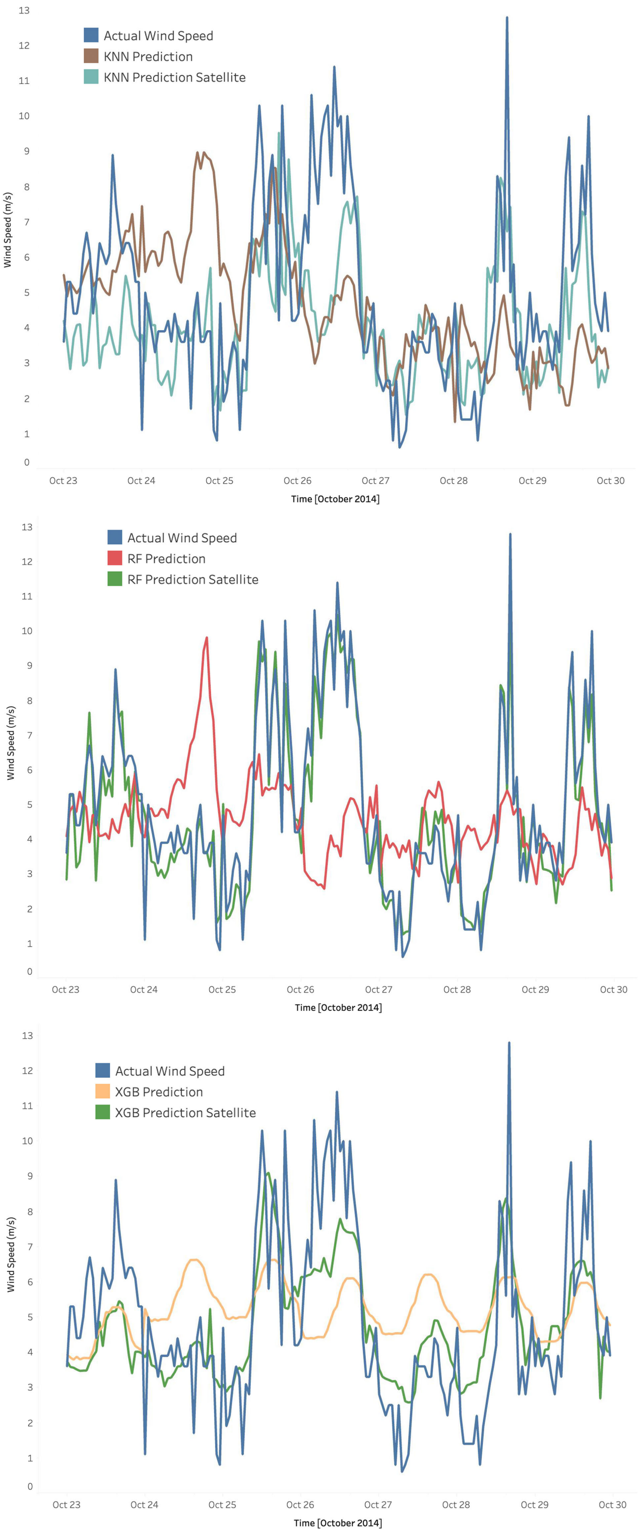
Table 8
Test case 3 R^2 comparison

	Temperature	Pressure	Wind	Radiation
K-NN	−1.01	−1.39	−0.26	0.83
K-NN (With Satellite)	0.54	0.41	0.20	0.60
RF	−0.01	−2.75	−0.23	0.74
RF (With Satellite)	0.75	0.13	0.90	0.94
XGBoost	−0.51	−0.16	0.04	0.60
XGBoost (With Satellite)	0.38	0.87	0.54	0.67
MLP	−0.90	−37000	0.06	0.53
MLP (With Satellite)	0.34	0.86	0.48	0.67

5.1. Implications

Despite being important measures for various applications, meteorological data from ground stations are prone to unreliability. Consequently, the proposed work highlights the effectiveness of utilizing satellite information for replacing unreliable data. The proposed work has several implications. Firstly, it demonstrates the effectiveness of using machine learning for weather data imputation. This will likely encourage organizations and researchers to utilize machine learning approaches for cleaning historical weather data. Second, the paper highlighted the usefulness of satellite information for imputing weather data. This will further draw the attention of organizations and researchers to utilize satellite data for cleaning historical weather data. Moreover, it will also facilitate research interest in using satellite data for other applications in conjunction with machine learning algorithms like smart grid monitoring [41]. Finally, it is hoped that the proposed work draws more research attention to the need for reliable meteorological and solar data, particularly in light of the global climate change concern.

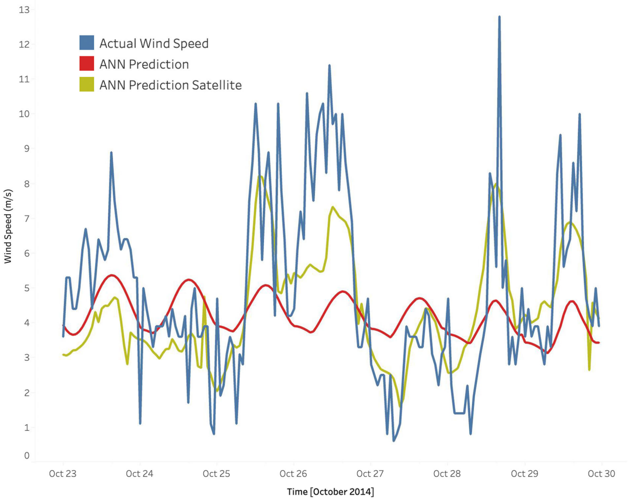
Figure 4
Comparison of actual and machine learning imputed wind speed with and without satellite information



5.2. Future research directions

Utilizing data from neighboring weather stations has been shown to be effective in weather data imputation [42, 43]. Although this approach may not be feasible for weather stations without any surrounding stations and due to the complexity of collaboration, future research should explore a combination of surrounding weather stations and satellite data as input features.

Figure 4
(Continued)



A combination of nearby weather stations and satellites will likely provide a more accurate imputation performance. In this work, extensive hyperparameter tuning was not performed because the objective was to simply compare the difference in performance. Therefore, future research should consider hyperparameter tuning to study the impact on model performance with and without satellite information. Also, the proposed research did not utilize more complex deep learning models, including convolutional and recurrent neural networks. These advanced networks may be able to provide a better imputation for wind speed and pressure data, which was not as effective using the models in the proposed work. Finally, the potential of large language models like ChatGPT and BARD [44] to provide better imputation in future research using weather data should be explored.

6. Conclusion

Reliable meteorological and solar data are necessary for studying climate change trends and renewable energy production. Missing values and inaccurate readings need to be replaced or imputed. This research investigated the use of machine learning algorithms and the impacts of using satellite information for weather data imputation. Results from two experiments demonstrate that machine learning algorithms are more effective for imputation using satellite information. Overall, ensemble machine learning models performed better than K-NN and ANN. Additionally, the imputation of temperature and solar radiation was more effective than pressure and wind speed. The implications and future research directions were also highlighted.

Funding Support

This work was supported by Zayed University under the research grant RIF Grant 23069.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The data that support the findings of this study are not publicly available due to privacy concerns. However, anonymous data are available on request from the corresponding author on reasonable request.

Author Contribution Statement

Kadhim Hayawi: Conceptualization, Validation, Formal analysis, Resources, Data curation, Visualization, Supervision, Funding acquisition. **Sakib Shahriar:** Methodology, Software, Investigation, Writing – original draft, Writing – review & editing. **Hakim Hacid:** Validation, Resources, Visualization, Supervision, Project administration.

References

[1] Molitor, M. R. (2023). The United Nations Climate Change Agreements. In N. J. Vig, & R. S. Axelrod (Eds.), *The global environment* (pp. 210–235). Routledge. https://doi.org/10.4324/9781003421368-13

[2] Kolstad, C. D., & Moore, F. C. (2020). Estimating the economic impacts of climate change using weather observations. *Review of Environmental Economics and Policy*, *14*(1), 1–24. https://doi.org/10.1093/reep/rez024

[3] Neumann, J. E., Willwerth, J., Martinich, J., McFarland, J., Sarofim, M. C., & Yohe, G. (2020). Climate damage functions for estimating the economic impacts of climate change in the United States. *Review of Environmental Economics and Policy*, *14*(1), 25–43. https://doi.org/10.1093/reep/rez021

[4] Gelfan, A., Kalugin, A., Krylenko, I., Nasonova, O., Gusev, Y., & Kovalev, E. (2020). Does a successful comprehensive evaluation increase confidence in a hydrological model intended for climate impact assessment? *Climatic Change*, *163*(3), 1165–1185. https://doi.org/10.1007/s10584-020-02930-z

[5] Rasp, S., Dueben, P. D., Scher, S., Weyn, J. A., Mouatadid, S., & Thuerey, N. (2020). WeatherBench: A benchmark data set for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems*, *12*(11), e2020MS002203. https://doi.org/10.1029/2020MS002203

[6] Masson, V., Heldens, W., Bocher, E., Bonhomme, M., Bucher, B., Burmeister, C., . . . , & Zeidler, J. (2020). City-descriptive input data for urban climate models: Model requirements, data sources and challenges. *Urban Climate*, *31*, 100536. https://doi.org/10.1016/j.uclim.2019.100536

[7] Neofytou, H., Nikas, A., & Doukas, H. (2020). Sustainable energy transition readiness: A multicriteria assessment index. *Renewable and Sustainable Energy Reviews*, *131*, 109988. https://doi.org/10.1016/j.rser.2020.109988

[8] Kalogirou, S. A. (2023). *Solar energy engineering: Processes and systems*. Netherlands: Elsevier.

[9] Stawowy, M., Olchowik, W., Rosiński, A., & Dąbrowski, T. (2021). The analysis and modelling of the quality of information acquired from weather station sensors. *Remote Sensing*, *13*(4), 693. https://doi.org/10.3390/rs13040693

[10] Vargas, J., Alsweiss, S., Toker, O., Razdan, R., & Santos, J. (2021). An overview of autonomous vehicles sensors and their vulnerability to weather conditions. *Sensors*, *21*(16), 5397. https://doi.org/10.3390/s21165397

[11] Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., & Tabona, O. (2021). A survey on missing data in machine learning. *Journal of Big Data*, *8*(1), 140. https://doi.org/10.1186/s40537-021-00516-9

[12] Afrifa-Yamoah, E., Mueller, U. A., Taylor, S. M., & Fisher, A. J. (2020). Missing data imputation of high-resolution temporal climate time series data. *Meteorological Applications*, *27*(1), e1873. https://doi.org/10.1002/met.1873

[13] Abu Romman, Z., Al-Bakri, J., & Al Kuisi, M. (2021). Comparison of methods for filling in gaps in monthly rainfall series in arid regions. *International Journal of Climatology*, *41*(15), 6674–6689. https://doi.org/10.1002/joc.7219

[14] Dong, Y., Qin, S. J., & Boyd, S. P. (2022). Extracting a low-dimensional predictable time series. *Optimization and Engineering*, *23*(2), 1189–1214. https://doi.org/10.1007/s11081-021-09643-x

[15] Phan, T. T. H., Poisson Caillault, É., Lefebvre, A., & Bigand, A. (2020). Dynamic time warping-based imputation for univariate time series data. *Pattern Recognition Letters*, *139*, 139–147. https://doi.org/10.1016/j.patrec.2017.08.019

[16] Doreswamy, H., Gad, I., & Manjunatha, B. R. (2017). Performance evaluation of predictive models for missing data imputation in weather data. In *International Conference on Advances in Computing, Communications and Informatics,* 1327–1334. https://doi.org/10.1109/ICACCI.2017.8126025

[17] Jiang, H., Wan, C., Yang, K., Ding, Y., & Xue, S. (2022). Continuous missing data imputation with incomplete dataset by generative adversarial networks–based unsupervised learning for long-term bridge health monitoring. *Structural Health Monitoring*, *21*(3), 1093–1109. https://doi.org/10.1177/14759217211021942

[18] Zhang, Y., Zhou, B., Cai, X., Guo, W., Ding, X., & Yuan, X. (2021). Missing value imputation in multivariate time series with end-to-end generative adversarial networks. *Information Sciences*, *551*, 67–82. https://doi.org/10.1016/j.ins.2020.11.035

[19] Mendelsohn, R., Kurukulasuriya, P., Basist, A., Kogan, F., & Williams, C. (2007). Climate analysis with satellite versus weather station data. *Climatic Change*, *81*(1), 71–83. https://doi.org/10.1007/s10584-006-9139-x

[20] Chivers, B. D., Wallbank, J., Cole, S. J., Sebek, O., Stanley, S., Fry, M., & Leontidis, G. (2020). Imputation of missing sub-hourly precipitation data in a large sensor network: A machine learning approach. *Journal of Hydrology*, *588*, 125126. https://doi.org/10.1016/j.jhydrol.2020.125126

[21] Boomgard-Zagrodnik, J. P., & Brown, D. J. (2022). Machine learning imputation of missing Mesonet temperature observations. *Computers and Electronics in Agriculture*, *192*, 106580. https://doi.org/10.1016/j.compag.2021.106580

[22] Gad, I., Hosahalli, D., Manjunatha, B. R., & Ghoneim, O. A. (2021). A robust deep learning model for missing value imputation in big NCDC dataset. *Iran Journal of Computer Science*, *4*(2), 67–84. https://doi.org/10.1007/s42044-020-00065-z

[23] Kiani, K., & Saleem, K. (2017). K-nearest temperature trends: A method for weather temperature data imputation. In *Proceedings of the International Conference on Information System and Data Mining,* 23–27. https://doi.org/10.1145/3077584.3077592

[24] Bou-Rabee, M., Sulaiman, S. A., Saleh, M. S., & Marafi, S. (2017). Using artificial neural networks to estimate solar radiation in Kuwait. *Renewable and Sustainable Energy Reviews*, *72*, 434–438. https://doi.org/10.1016/j.rser.2017.01.013

[25] Hussain, S., & AlAlili, A. (2017). A hybrid solar radiation modeling approach using wavelet multiresolution analysis and artificial neural networks. *Applied Energy*, *208*, 540–550. https://doi.org/10.1016/j.apenergy.2017.09.100

[26] Kazem, H. A., Yousif, J. H., & Chaichan, M. T. (2016). Modeling of daily solar energy system prediction using support vector machine for Oman. *International Journal of Applied Engineering Research*, *11*(20), 10166–10172.

[27] Tuohy, A., Zack, J., Haupt, S. E., Sharp, J., Ahlstrom, M., Dise, S., . . . , & Collier, C. (2015). Solar forecasting: Methods, challenges, and performance. *IEEE Power and Energy Magazine*, *13*(6), 50–59. https://doi.org/10.1109/MPE.2015.2461351

[28] Notton, G., Nivet, M. L., Voyant, C., Paoli, C., Darras, C., Motte, F., & Fouilloy, A. (2018). Intermittent and stochastic character of renewable energy sources: Consequences, cost of intermittence and benefit of forecasting. *Renewable and Sustainable Energy Reviews*, *87*, 96–105. https://doi.org/10.1016/j.rser.2018.02.007

[29] Fathi, M., Haghi Kashani, M., Jameii, S. M., & Mahdipour, E. (2022). Big data analytics in weather forecasting: A systematic review. *Archives of Computational Methods in Engineering*, *29*(2), 1247–1275. https://doi.org/10.1007/s11831-021-09616-4

[30] Chantry, M., Christensen, H., Dueben, P., & Palmer, T. (2021). Opportunities and challenges for machine learning in weather and climate modelling: Hard, medium and soft AI. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *379*(2194), 20200083. https://doi.org/10.1098/rsta.2020.0083

[31] Government of Canada. (2025). *Engineering climate datasets.* Retrieved from: https://climate.weather.gc.ca/prods_servs/engineering_e.html

[32] Gelaro, R., McCarty, W., Suárez, M. J., Todling, R., Molod, A., Takacs, L., . . . , & Zhao, B. (2017). The modern-era retrospective analysis for research and applications, version 2 (MERRA-2). *Journal of Climate*, *30*(14), 5419–5454. https://doi.org/10.1175/JCLI-D-16-0758.1

[33] Siu, C. Y., & Liao, Z. (2020). Method for converting CWEEDs weather files to EPW format for multiyear simulation of building thermal dynamics. *MethodsX*, *7*, 101016. https://doi.org/10.1016/j.mex.2020.101016

[34] Bishop, S. (n.d.). *Pytz–World timezone definitions for Python.* Retrieved from: https://pythonhosted.org/pytz/

[35] McKinney, W. (2011). Pandas: A foundational Python library for data analysis and statistics. In *Python for High Performance and Scientific Computing*, *14*(9), 1–9.

[36] Fenoglio, A., Wagner, T., Auclair, P., & Langhals, B. (2022). Effect of trigonometric transformations on the machine learning prediction and quality control of air temperature. In *World Congress in Computer Science, Computer Engineering, and Applied Computing*.

[37] Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, *13*(1), 21–27. https://doi.org/10.1109/TIT.1967.1053964

[38] Shahriar, S., Al-Ali, A. R., Osman, A. H., Dhou, S., & Nijim, M. (2020). Machine learning approaches for EV charging behavior: A review. *IEEE Access*, *8*, 168980–168993. https://doi.org/10.1109/ACCESS.2020.3023388

[39] Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, *2*(3), 18–22.

[40] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* 785–794. https://doi.org/10.1145/2939672.2939785

[41] Ramesh, J., Shahriar, S., Al-Ali, A. R., Osman, A., & Shaaban, M. F. (2022). Machine learning approach for smart distribution transformers load monitoring and management system. *Energies*, *15*(21), 7981. https://doi.org/10.3390/en15217981

[42] Mott, P., Sammis, T. W., & Southward, G. M. (1994). Climate data estimation using climate information from surrounding climate stations. *Applied Engineering in Agriculture*, *10*(1), 41–44. https://doi.org/10.13031/2013.25825

[43] Ashraf, M., Loftis, J. C., & Hubbard, K. G. (1997). Application of geostatistics to evaluate partial weather station networks. *Agricultural and Forest Meteorology*, *84*(3–4), 255–271. https://doi.org/10.1016/S0168-1923(96)02358-1

[44] Hayawi, K., Shahriar, S., & Mathew, S. S. (2024). The imitation game: Detecting human and AI-generated texts in the era of ChatGPT and BARD. *Journal of Information Science*. Advance online publication. https://doi.org/10.1177/01655515241227531

How to Cite: Hayawi, K., Shahriar, S., & Hacid, H. (2025). Climate Data Imputation and Quality Improvement Using Satellite Data. *Journal of Data Science and Intelligent Systems*, *3*(2), 87–97. https://doi.org/10.47852/bonviewJDSIS42022857