Environ Monit Assess (2023) 195:343
https://doi.org/10.1007/s10661-023-10961-z

Data-driven techniques for temperature data prediction: big data analytics approach

Adamson Oloyede · Simeon Ozuomba · Philip Asuquo · Lanre Olatomiwa · Omowunmi Mary Longe

Received: 5 September 2022 / Accepted: 23 January 2023

Abstract For extrapolation, climate change and other meteorological analysis, a study of past and current weather events is a prerequisite. NASA (National Aeronautics and Space Administration) has been able to develop a model capable of predicting various weather data for any location on the Earth, including locations lacking weather stations, weather satellite coverage, and other weather measuring instruments. This paper evaluates the prediction accuracy of the NASA temperature data with respect to NiMet (Nigerian Meteorological Agency) ground truth measurement, using Akwa Ibom Airport as a case study. Exploratory data analysis (descriptive and diagnostic analyses) of temperature retrieved from NiMet and NASA was performed to give a clear path to follow for predictive and prescriptive analyses. Using 2783 days of weather data retrieved from NiMet as ground truth, the accuracy of NASA predictions with the corresponding resolution was calculated. Mean absolute error (MAE) of 2.184 °C and root mean square error (RMSE) of 2.579 °C, with a coefficient of determination (R^2) of 0.710 for maximum temperature, then MAE of 0.876 °C, RMSE of 1.225 °C with a coefficient of determination (R^2) of 0.620 for minimum temperature was discovered. There is a good correlation between the two datasets; hence, a model can be developed to generate more accurate predictions, using the NASA data as input. Predictive and prescriptive analyses were performed by employing five prediction algorithms: decision tree regression, XGBoost regression and MLP (multilayer perceptron) with LBFGS (limited-memory Broyden-Fletcher-Goldfarb-Shanno) optimizer, MLP with SGD (stochastic gradient) optimizer and MLP with Adam optimizer. The MLP LBFGS algorithm performed best, by significantly reducing the MAE by 35.35% and RMSE by 31.06% for maximum temperature, accordingly, MAE by 10.05% and RMSE by 8.00% for minimum temperature. Results obtained show that given sufficient data, plugging NASA predictions as input to an LBFGS-MLP model gives more accurate temperature predictions for the study area.

A. Oloyede (✉)
Advanced Space Technology Applications Laboratory Uyo, National Space Research and Development Agency, P.M.B. 437 Abuja, Nigeria
e-mail: ooloyede@astaluyo.gov.ng; ka07krd@gmail.com

A. Oloyede · S. Ozuomba · P. Asuquo
Department of Computer Engineering, University of Uyo, Uyo 520103, Nigeria
e-mail: simeonozuomba@uniuyo.edu.ng

P. Asuquo
e-mail: philipasuquo@uniuyo.edu.ng

L. Olatomiwa · O. M. Longe
Department of Electrical and Electronic Engineering Science, University of Johannesburg, Johannesburg 2006, South Africa
e-mail: olatomiwa.l@futminna.edu.ng

O. M. Longe
e-mail: omowunmil@uj.ac.za

L. Olatomiwa
Department of Electrical and Electronics Engineering, Federal University of Technology Minna, P.M.B. 65, Minna, Nigeria

Published online: 30 January 2023

Keywords Weather prediction · Artificial neural network · Correlation analysis · Descriptive and diagnostic analyses · Predictive and prescriptive analytics

Introduction

Weather prediction remains a vital aspect of our livelihoods, as the current state of the weather has an enormous impact on humans, their livelihood, and their day-to-day activities (Denissen et al., 2008; Dundas & Von Haefen, 2021). Knowledge of the future state of the weather can be used in events and activities planning, disaster monitoring, and management (Kusiak & Shah, 2006; Marson & Legerton, 2021). Usually, real-time hyperlocal weather state and its parametric values are measured using sensors sited on ground, or remotely sensed using sensors mounted on aircrafts, satellites, and radiosondes. These weather data are then pooled into a database, for storage, visualization, analysis, and extrapolation. Past weather data are used together with computer algorithms, to predict future weather conditions. The art and science of extracting meaningful information and discovering patterns from these data is termed data mining, which is currently the most used technique for weather prediction (Twin, 2021; Sheikh et al., 2016).

In advanced countries, weather data sources have increased, as a result of technological advancements in the aspect of miniaturization of sensors and the internet of things. Weather data from varying sources, varying formats, varying structures, and continuous increments are pooled together for storage and analysis. This type of data is termed big data, and it can be cumbersome to deal with using traditional approaches (Fathi et al., 2021). The approach to extracting meaningful information from this type of data is termed big data analytics, and it is categorized into four; descriptive, diagnostic, predictive, and prescriptive analytics (Maydon, 2017). The descriptive and diagnostic analytics clearly defines the path to follow for the predictive and prescriptive analytics to follow (Oloyede et al., 2022).

Nevertheless, the problem of unavailability of weather data from remote areas of developing countries like Nigeria remains a challenge. There are no records of past weather data as a result of limited weather stations, even where there are weather stations, the weather data are full of inconsistencies. This necessitates solutions like the development of meteorological satellites that can capture weather data using satellite remote sensing technologies (Waring & Running, 2007). However, meteorological satellites are only applicable in measurement of land surface temperature (LST) and suitable for LST-related correlation analysis (Oloyede et al., 2021; Nnah et al., 2021). In addition, they have their limitations of low spatial resolution, cloud cover restriction, and in some cases, low temporal resolution depending on the configuration of the satellite. Consequently, intelligent models have been developed, to address these limitations and synergize the effort of meteorological satellites, by predicting meteorological elements of every location on Earth. An example of this is NASA's POWER (National Aeronautics and Space Administration's Prediction of Worldwide Energy Resources) project (NASA, 2022). NASA POWER meteorological elements are derived from a combination of NASA's GEOS 5.12.4 FP-IT (Goddard Earth Observing System, Ver. 5.12.4, Forward Processing – Instrument Teams) and GMAO MERRA-2 (Global Model and Assimilation Office' Modern-Era Retrospective analysis for Research and Applications, Ver. 2) (Gelaro et al., 2017).

In as much as NASA POWER meteorological data are widely used, many reviewers query the use of the data in place of ground-measured data, as they are of the opinion that their predictions are not accurate (Halabi et al., 2017). Some researchers have performed evaluations of the NASA POWER accuracy, using varying meteorological parameters at different geographical locations (Aboelkhair et al., 2019; Osama, 2021; Quansah et al., 2022; Rodrigues & Braga, 2021). Results from their evaluations show that the accuracy of the NASA POWER data varies with the geographical location, and some meteorological parameters are more accurate than others. However, there is a consensus that the NASA POWER data can be useful in cases where there are no ground-measured data and cases where there are missing values in the ground-measured data. Root mean squared error (RMSE) metric, mean absolute error (MAE) metric, and R-squared (R^2) metric are among the reliable indicators employed for assessing such model's performance (Olatomiwa et al., 2015; Xi et al., 2021). Nevertheless, there is a need to evaluate the accuracy of NASA POWER in the selected study area, which

this work addresses, and also proffer solutions to prediction inaccuracy by employing data-driven models to improve prediction accuracy. Research has shown that Artificial Intelligent models are capable of handling such huge and nonlinear data, thereby increasing the accuracy and efficiency of meteorological data predictions (Tan et al., 2021). Prediction of meteorological data is a nonlinear regression problem (Abhishek et al., 2013), and regression, in machine learning falls under supervised machine learning, as the model is fed with an input and a corresponding output, then the model maps a function from input to output (Garbade, 2018).

This work is therefore aimed at developing and evaluating data-driven models that can predict near precise weather data, using NASA weather data as input. Accordingly, the following are the specific implemented objectives:

1. An area, with available ground truth weather data, is selected for use as a case study,
2. In situ weather data of the study area is obtained from NiMet (Nigerian Meteorological Agency), and the GPS (global positioning system) coordinates of the study area are used to retrieve weather data with the corresponding spatiotemporal resolution, from NASA
3. Prediction accuracy of the NASA data is evaluated by performing descriptive, diagnostic and correlation analyses on the NiMet and NASA weather data, using visualization and statistical tools in a Python integrated development environment,
4. Predictive, prescriptive, and correlation analyses are performed using data mining techniques, and then the performances of the techniques are evaluated using statistical tools in a Python integrated development environment.

The rest of this paper is structured in the following pattern: a review of related literature on data mining and weather prediction techniques is presented in the "Review of related works" section. The "Methodology" section presents the methodology employed in accomplishing the specific objectives of this work, while the "Results and discussion" section presents the discussion of the results. The "Conclusion and recommendation" section is a combination of the conclusion and necessary recommendations.

Review of related works

While data assimilation methods combine direct measurements with a model's output, to improve the model's output, data-driven techniques provide faster and computationally cheaper simulations and extrapolations based on analysis of historic data (Kaneko et al., 2020; Nature, 2021). A review of works relating to the applicability of data-driven techniques for weather prediction is presented in this section. Nikam and Meshram (2013) proposed a data-intensive model, using data mining methods, to predict the weather. They were more interested in the prediction of rainfall, so they utilized seven weather features, which play a major role in the prediction of rainfall, out of 36 weather features collected from the Meteorological Department in India. The seven weather features used are temperature, mean sea pressure, station level pressure, vapour pressure, relative humidity, rainfall, and wind speed. A supervised learning method for classification; Bayesian classifier, was employed to build this model. The Bayesian model performed well, and it was observed that it performs better with a large training dataset.

A report on how CART (classification and regression trees) is used for weather prediction was presented by (Petre, 2009). Data were collected between 2002 and 2005, for Hong Kong, the capital city of China. The weather attributes utilized are temperature, average pressure, relative humidity, cloud quantity and rainfall. A machine learning application, WEKA (Waikato Environment for Knowledge Analysis), built with the CART algorithm, was used for data analysis, visualization, and predictive modelling. Their model was able to predict the average temperature for a future month, to a certain accuracy. They recommend that having larger data, with more weather parameters like wind speed, wind direction, and radiation can positively impact the predictive accuracy of the model.

A comparative analysis of two data mining techniques for weather prediction, C4.5 and Naïve Bayes decision tree algorithm, was carried out by (Sheikh et al., 2016). They collected weather for 2 years and simultaneously analyzed the performances of C4.5 and Naïve Bayes decision tree algorithm in a weather prediction model. It was discovered that the C4.5 algorithm performed better, with an accuracy of 88.2% compared to the Naïve Bayes, which did only

54.8%. C4.5 resulted in more true positives than the Naïve Bayes. A larger training dataset and more data attributes impacted positively the performance of the C4.5 algorithm, but negatively impacted the performance of the Naïve Bayes algorithm.

Findawati et al. (2019) further carried out a comparative analysis of Naïve Bayes, C4.5, and K-nearest neighbor (KNN) algorithms for weather prediction. Daily weather data was collected from BMKG (Badan Meteorologi, Klimatologi, dan Geofisika); an Indonesian government agency for meteorology and climatology. The data spanned January 2015 to November 2018, with 1422 entries and 8 weather attributes; minimum temperature, maximum temperature, temperature average, wind speed, relative humidity, wind direction, radiation, and rain intensity. The test result is presented by evaluating the value of accuracy, precision, recall value and f-measure, using the WEKA machine learning tool. From the test result, the KNN method did approximately 71.6% accuracy, with k = 7 and fold = 5, while C4.5 came second, with an accuracy of 69.8%, at fold = 20, and Naïve Bayes with an accuracy of 68.8%, at fold = 3.

Olaiya and Adeyemo (2012) performed extensive research on the application of artificial neural networks (ANNs) and decision tree algorithms, data mining methods in weather prediction and climate change studies. Weather data for January 2000 to December 2009, for Ibadan, a city in Oyo State, Nigeria, was collected from the NiMet Oyo State office. Minimum temperature, maximum temperature, rainfall, sunshine, radiation, cloud form, evaporation, and wind speed are the weather attributes considered in this work. C5 Decision Tree classifier was selected and used, after performance evaluation in comparison with C4.5 and CART algorithms was done. The C5 algorithm was implemented in the See5 data mining environment. For the ANN algorithms, multilayer perceptron (MLP) time-lagged feed-forward network (TLFN) and other recurrent networks are used, and implementation was performed in a neural network software development environment; Neuro Solutions 6. Among the recurrent networks used, the TLFN network, using a TDNN (time delayed neural network) memory component, a hidden layer and eight nodes, trained on the Lavenberg-Marquet learning algorithm, performed best. Good insight into trends and patterns emanated from the See5 rules generated and ANN was used to train a model to detect the relationship between the input variable and its output. These authors believe that a larger weather dataset, collected over decades, will improve the model's performance.

A comparative study of classification and prediction models on weather data from the National Climatic Data Center was carried out by (Gad & Hosahalli, 2022) for regression and classification prediction. A comparison of emerging models and traditional meteorological models was implemented. Linear regression, support vector machine (SVM), linear discriminant analysis (LDA), Gaussian NB, random forest, k-nearest neighbors (KNNs), AdaBoost, an ensemble method named Extreme Gradient Boosting (XGBoost), decision tree (CART), artificial neural networks multilayer perceptron (MLP), and deep learning are amongst the models considered for evaluation. The AdaBoost, CART, and XGB models outperform other classification algorithms, while linear regression performed best for the prediction task.

A multiple linear regression-based model for daily average temperature prediction was presented by (Gupta et al., 2022). Datasets utilized were created using Weather Underground's API web service, with 997 daily instances of mean temperature, mean dew point temperature, mean humidity, and precipitation. Python "corr" function of Pandas library was employed to calculate the coefficient of determination, which resulted in 0.6 and above in all meteorological elements considered. Results generated show that their model was able to predict the mean temperature of a day, with an error range of 2.8 °C.

Soft computing techniques for future weather forecasting were presented by (Khajure & Mohod, 2016). The artificial neural network used was trained using a combination of weather parameters, which include temperature, pressure, humidity, dew point, wind speed, and visibility. From their results, they concluded that a neural network is an important tool for weather forecasting capable of modelling a weather forecast system, with a combination of fuzzy inference systems to enhance accuracy. An artificial neural network was also employed by (Bhardwaj & Duhoon, 2018), in a comparative study of multilayer perceptron, support vector regression, linear regression and Gaussian process are evaluated, to attain the highest efficiency in the prediction of parameters affecting the weather. They stated that climate differs from weather, as the latter focuses on the short term, while the former is on the long term. RMSE, root relative

squared error (RRSE), relative absolute error (RAE), MAE, and coefficient of correlation (CC) are the performance metrics employed. Temperature, relative humidity, rainfall, wind speed, sunshine, and evaporation are the meteorological parameters considered. Support vector regression and multilayer perceptron performed best out of the four models considered.

Markovics and Mayer (2022) carried out a comparison of machine learning techniques for another important weather element, photovoltaic (PV) capacity. Twenty-four machine learning models were tested and evaluated for photovoltaic power forecasting. Dataset from 16 PV plants in Hungary, with 15-min resolution, for 2 years was used for the analysis. Results show that multilayer perceptron and kernel ridge regression performed best out of all models considered, which was able to decrease the RMSE by 13.9% when compared with the baseline model of linear regression. They underscored the importance of selecting the right predictor, as substituting the basic numerical weather prediction data with sun position angles and irradiance values had a positive impact on the prediction system.

Data-driven and machine learning approach were also used to build a model for air quality index prediction (Xi et al., 2021). Climatic data with multiple features from 31 regions and provinces in China was used to build a Bayesian network model for monitoring and predicting air quality index. A Bayesian Network with two layers was developed for analyzing factors influencing various air pollutants, forecasting spatiotemporal changes and revealing the extent of impact incurred from each factor. Results generated were compared with results from other machine learning models, and they were able to conclude that the Bayesian Network model was able to reach a 90% accuracy in prediction.

A deep spatiotemporal forecasting model was proposed by (Kong et al., 2022), for multiple site weather prediction by using spatial and temporal information. Their research takes into consideration spatial and temporal parameters, and obtains forecasts of multiple weather stations simultaneously, using the same framework. In addition, the impact of changes in season and topographical variations on the accuracy of the model was considered. A convolutional neural network-based, deep spatiotemporal forecasting model was used for short-term weather forecasts at 226 weather stations in Beijing, China. Experiments indicate that the model has strong stability and high prediction accuracy.

An ensemble of spatial–temporal attention networks and multilayer perceptron for weather forecasting using was proposed, modelled, evaluated, and presented by (Li et al., 2021). A hybrid of the multilayer perceptron and the spatial–temporal attention network was employed to build a model for forecasting humidity, surface temperature, wind speed, and direction at a total of 24 weather stations in Beijing, the capital of China. The report showed that this ensemble model performed better than the numerical weather prediction model and other algorithms.

Research has shown that reanalysis of data assimilation methods are not good for trend studies, due to changes in observing systems. This extensive review reveals that data-driven techniques can be effective and efficient when employed for weather data prediction and forecasting. In addition, to improve prediction accuracy, the authors emphasized the need for a large training historical dataset, which has proven to be difficult to obtain or non-existent in many locations. Consequently, this work explores the use of the NASA POWER dataset and a few ground measurement, to train the system, and subsequently plug any NASA POWER data as an input feature to generate predictions that are more accurate.

Methodology

The flowchart describing systematically the objectives followed to achieve the aim of this research work is presented in Fig. 1.

Study area extraction

The availability of ground truth weather data, sufficient enough to train our prediction model, is key in selecting the study area. Consequently, the Victor Attah International Airport is selected, since there is a Nigerian Meteorological Agency (NiMet) weather station at the location. The Victor Attah International Airport is sited in the Uruan LGA (local government area) of Akwa Ibom State, Nigeria, at coordinates 4.8714° N, 8.0916° E. Figure 2 is a satellite image of the Victor Attah International Airport, composed using the Google Earth Pro application.

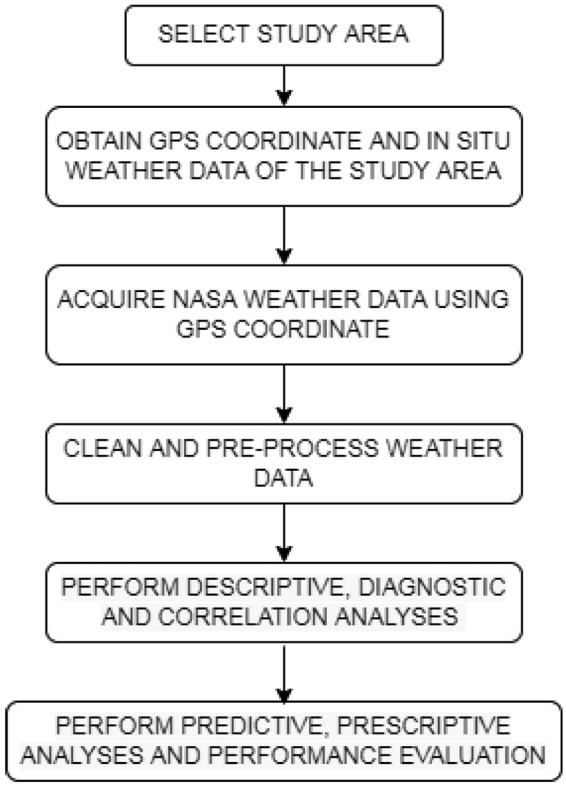

Fig. 1 Methodology flowchart

This location is pointed, composed using QGIS (Quantum Geographic Information System), as shown in Fig. 3

Data retrieval, cleaning, and pre-processing

Step 1: retrieve and view dataset

In situ weather data of the study area is retrieved from NiMet (NiMet, 2022), in Microsoft Word document format, and is converted to comma-separated values format for analysis in the Python environment. To get an understanding of the dataset, and ensure the right data are imported into the Python environment, the Python head and tail function is employed to view the first five and last five rows. Figure 4 shows the first five and last five rows of the NiMet dataset.

It is discovered that the dataset spans from 01/01/2010 to 14/08/2017, with a total of 2783 entries for each feature. The dataset is a constituent of eight features: date, maximum temperature (°C), minimum temperature (°C), rain (mm), and relative humidity (%) at 0600Z, 0900Z, 1200Z, and 1500Z. This implies that temperature, rain, and relative humidity are the three weather parameters captured in the dataset. Also,

Fig. 2 Satellite image of the Victor Attah International Airport

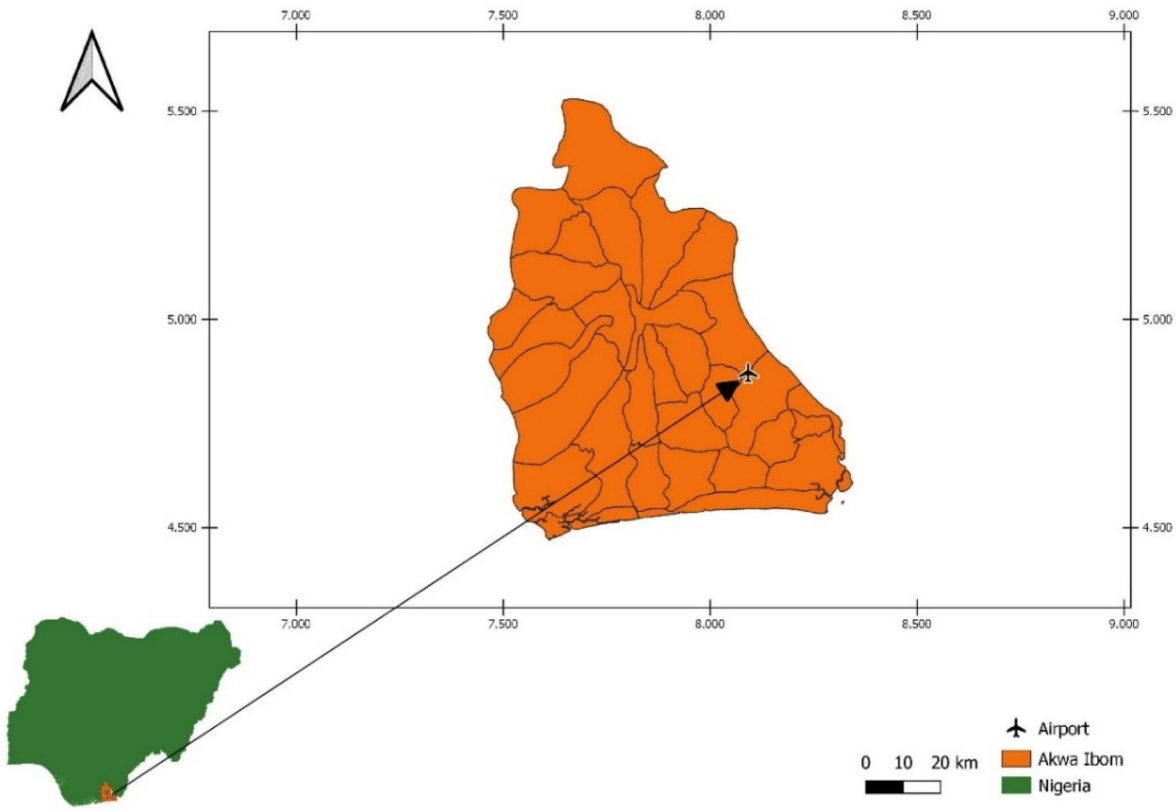

Fig. 3 Map of Akwa Ibom showing the study location

it is observed that there are some unwanted columns in the dataset and other weather features that are not required. These columns are filtered out.

The GPS coordinates of the study area (4.8714° N, 8.0916° E) are used to retrieve weather data with corresponding meteorological, temporal, and spatial features from the NASA repository (NASA, 2022). Python head and tail function is used to view the dataset. Figure 5 shows the first five and last five rows of the NASA dataset.

The dataset is constituted of the required features for this analysis. However, the date format differs from the NiMet date format. The date format will be converted to conform to the NiMet date format.

Step 2: check for missing values and unwanted columns

A check for missing values in the NiMet and NASA datasets is implemented. It is observed that the NASA dataset contains no missing data points; however, some missing data points and some unwanted features were recorded in the NiMet dataset. It is usual practice to overlook missing data points, but it is argued that this has a negative influence on model learning performance because the missing data points may include important information (Gill et al., 2007; Romero-Fiances et al., 2022; Faybishenko et al., 2022). There are several approaches to filling missing data points in time series data: forward filling, backward filling, linear interpolation, and mean of nearest neighbor (Johnson et al., 2021; Sher, 2020; Zhang & Thorburn, 2022). The unwanted data columns are filtered out, and the mean of nearest neighbor approach is employed to fill the missing data points. The following equation explains the approach.

$$\overline{X} = \frac{\sum X}{N} \tag{1}$$

Fig. 4 First five and last five rows of the NiMet dataset

	DATE	MAX TEMP	MIN TEMP	RAIN (mm)	RH% 0600Z	RH% 0900Z	RH% 1200Z	RH% 1500Z
0	01/01/2010	32.5	24.5	0	97	88	61	52
1	02/01/2010	34	23.7	0	97	81	55	40
2	03/01/2010	33.2	23.8	0	96	88	58	42
3	04/01/2010	32.1	24.1	0	99	88	61	49
4	05/01/2010	32.5	24.3	0	98	84	58	47
...	...	...	...	...	...	...	...	...
2778	10/08/2017	28	22.3	1.2	98	93	85	85
2779	11/08/2017	29.5	23.3	TR	96	92	81	88
2780	12/08/2017	26.1	23.5	19.2	100	95	92	90
2781	13/08/2017	27.2	-	0.8	98	84	82	96
2782	14/08/2017	26.8	23	41.5	96	90	95	94

where $\overline{X}$ represents mean of nearest neighbors, $\sum X$ represents sum of all nearest neighbor datapoints; N is the number of data points.

Table 1 lists the extracted features taken into account for this work, along with a description and the international system of units (SI unit) that correspond to them.

Descriptive, diagnostic, and correlation analyses

Data visualizations for the NASA and NiMet dataset, using a time series plot is employed for descriptive analyses, aimed at gaining insights into what has happened in the past in aspects of trends, patterns, and most importantly uncover the variation between the NASA predictions and NiMet ground truth measurements. Before that, the input features and the target feature need to be well defined.

	YEAR	MO	DY	T2M	T2M_MAX	T2M_MIN	RH2M
0	2010	1	1	26.70	31.05	23.30	79.38
1	2010	1	2	26.26	30.87	22.85	78.06
2	2010	1	3	25.14	29.96	20.08	62.25
3	2010	1	4	25.39	30.40	20.49	62.56
4	2010	1	5	26.71	31.03	22.96	76.94
...	...	...	...	...	...	...	...
2778	2017	8	10	24.87	27.47	22.92	90.81
2779	2017	8	11	25.05	27.86	23.25	92.19
2780	2017	8	12	24.76	26.67	23.73	94.19
2781	2017	8	13	24.33	25.95	23.21	94.94
2782	2017	8	14	24.34	26.18	23.26	94.38

Fig. 5 First five and last five rows of the NASA dataset

Step 1: set features and target

Maximum temperature Considering maximum temperature as the weather parameter for prediction, a new data frame is created, with NASA daily average temperature, daily maximum temperature, daily minimum temperature, and daily relative humidity as features, with NiMet maximum daily temperature as the target. Features are independent variables that act as input in a model, while the target is the resultant output. Figure 6 shows a screenshot of the newly created data frame.

Minimum temperature When considering minimum temperature as a weather parameter for prediction, the data frame constituents will be NASA daily average temperature, daily maximum temperature, daily minimum temperature, and daily relative humidity as features, with NiMet minimum daily temperature as the target. Figure 7 shows a screenshot of the newly created data frame.

Step 2: time series plots

This section is a statistical analysis of the NiMet and NASA datasets, presented in data visualization form. This helps in gaining insights and understanding of trends, patterns, and seasonality in the datasets and also helps to compare the NASA model's

Table 1 Extracted features for this work

S/N	Feature	Feature description	SI unit
1.	Date	Date each data was logged	Second
2.	Daily maximum temperature	Maximum temperature value per date	°C
3.	Daily minimum temperature	Minimum temperature value per date	°C
4.	Daily average temperature	Average temperature value per date	°C
5.	Daily relative humidity	Relative humidity value per date	%

values with the NiMet ground truth values. Figure 8 is a time series plot of the NiMet daily maximum temperature and the NASA daily maximum temperature, in a comparative form. Figure 9 further clearly shows the difference in the NiMet and NASA curves, by resampling the data points on a monthly average basis.

Comparatively, the time series plot for NASA is different from that of NiMet. There is a definite indication that the NiMet daily maximum temperature and ground truth measurement differ from the data points for daily maximum temperature predicted by NASA. This is a result of having weather data values that are inaccurate when compared with the ground truth values.

Figure 10 is a time series plot of the NiMet daily minimum temperature and the NASA daily minimum temperature, in a comparative form, and Fig. 11 presents same on a monthly average basis.

Comparatively, the time series plot for NASA minimum temperature is slightly different from that of NiMet minimum temperature. Also, visually, the NASA model is close to the NiMet ground truth measurement, but this will further be estimated statistically in the next section.

Step 3: evaluate the mean absolute error and the root mean squared error

The MAE and the RMSE are model evaluation metrics associated with regression models. These metrics are used to evaluate the extent of error between the NASA predictions and the NiMet ground truth values, as the time series plots cannot tell the extent.

While a prediction error is an evaluation of the discrepancy between the ground truth value and the predicted value of an instance, MAE is the mean of prediction errors over all instances.

Fig. 6 Features and maximum temperature as target

	DATE	TEMP	MAX_TEMP	MIN_TEMP	RH_DAILY	NIMET_MAX_TEMP
0	01/01/2010	26.70	31.05	23.30	79.38	32.5
1	02/01/2010	26.26	30.87	22.85	78.06	34.0
2	03/01/2010	25.14	29.96	20.08	62.25	33.2
3	04/01/2010	25.39	30.40	20.49	62.56	32.1
4	05/01/2010	26.71	31.03	22.96	76.94	32.5
...	...	...	...	...	...	...
2778	10/08/2017	24.87	27.47	22.92	90.81	28.0
2779	11/08/2017	25.05	27.86	23.25	92.19	29.5
2780	12/08/2017	24.76	26.67	23.73	94.19	26.1
2781	13/08/2017	24.33	25.95	23.21	94.94	27.2
2782	14/08/2017	24.34	26.18	23.26	94.38	26.8

Fig. 7 Features and minimum temperature as target

	DATE	TEMP	MAX_TEMP	MIN_TEMP	RH_DAILY	NIMET_MIN_TEMP
0	01/01/2010	26.70	31.05	23.30	79.38	24.500000
1	02/01/2010	26.26	30.87	22.85	78.06	23.700000
2	03/01/2010	25.14	29.96	20.08	62.25	23.800000
3	04/01/2010	25.39	30.40	20.49	62.56	24.100000
4	05/01/2010	26.71	31.03	22.96	76.94	24.300000
...	...	...	...	...	...	...
2778	10/08/2017	24.87	27.47	22.92	90.81	22.300000
2779	11/08/2017	25.05	27.86	23.25	92.19	23.300000
2780	12/08/2017	24.76	26.67	23.73	94.19	23.500000
2781	13/08/2017	24.33	25.95	23.21	94.94	23.311191
2782	14/08/2017	24.34	26.18	23.26	94.38	23.000000

$$MAE = \frac{\sum_{i=1}^{n} abs(y_i - \lambda(x_i))}{n} \tag{2}$$

where y_i is the ground truth value for an instance x_i; $\lambda(x_i)$ is the predicted value for an instance x_i; n is the number of instances.

While mean squared error (MSE) is the squared of prediction errors over all instances, RMSE is the square root of the MSE.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \lambda(x_i))^2}{n}} \tag{3}$$

where y_i is the ground truth value for an instance x_i; $\lambda(x_i)$ is the predicted value for an instance x_i; n is the number of instances.

Note that MAE and RMSE are relative metrics and differ from one criterion to another. The value ranges from 0 to $+\infty$, with the best values being the lowest

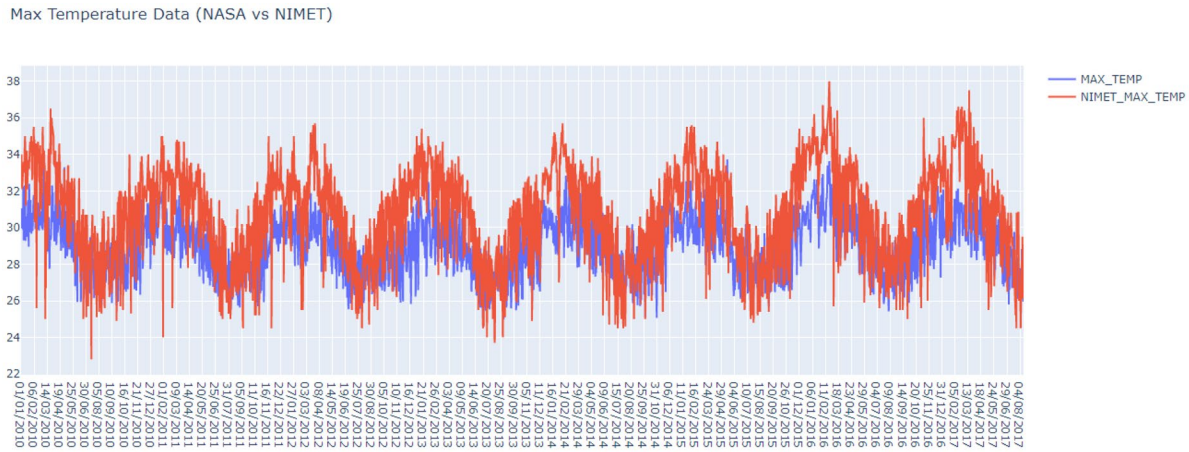

Fig. 8 NiMet and NASA maximum temperature time series plot

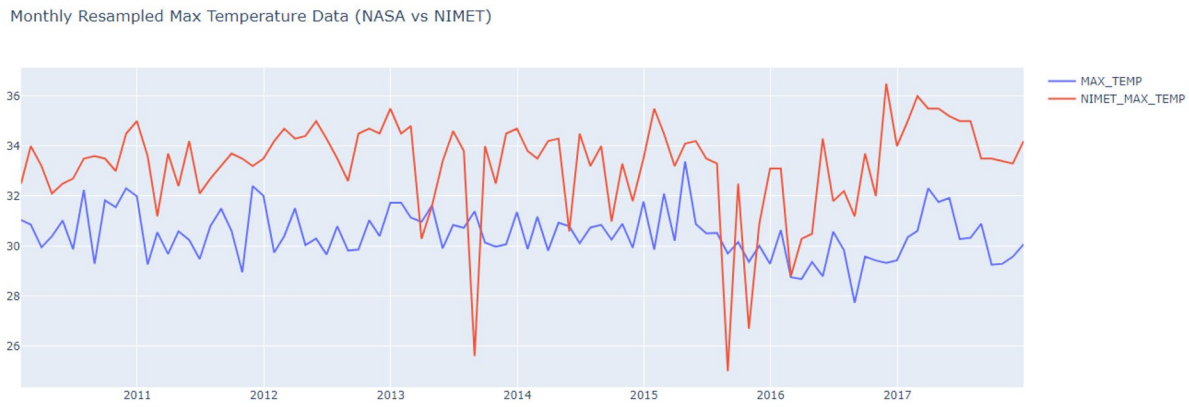

Fig. 9 Monthly resampled NiMet and NASA maximum temperature time series plot

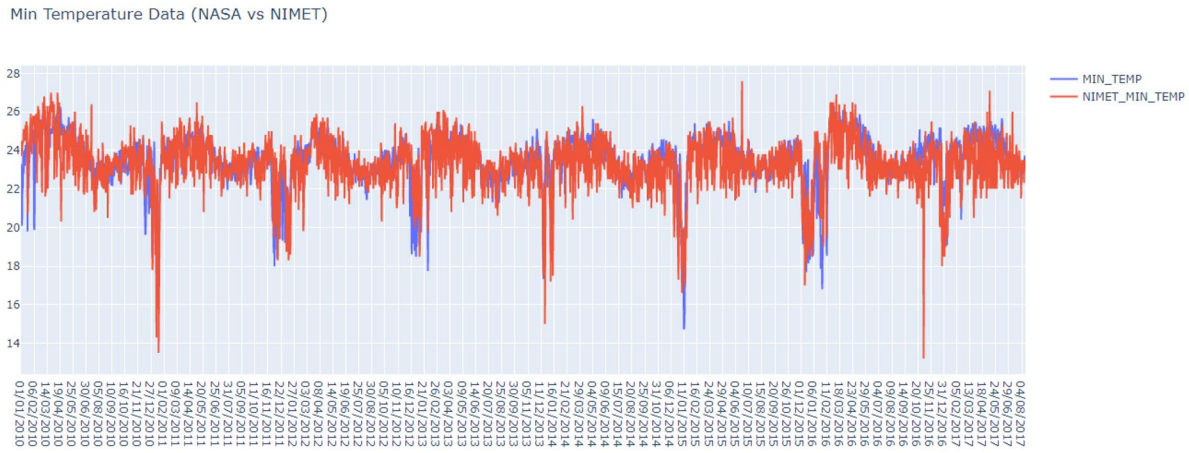

Fig. 10 NiMet and NASA minimum temperature time series plot

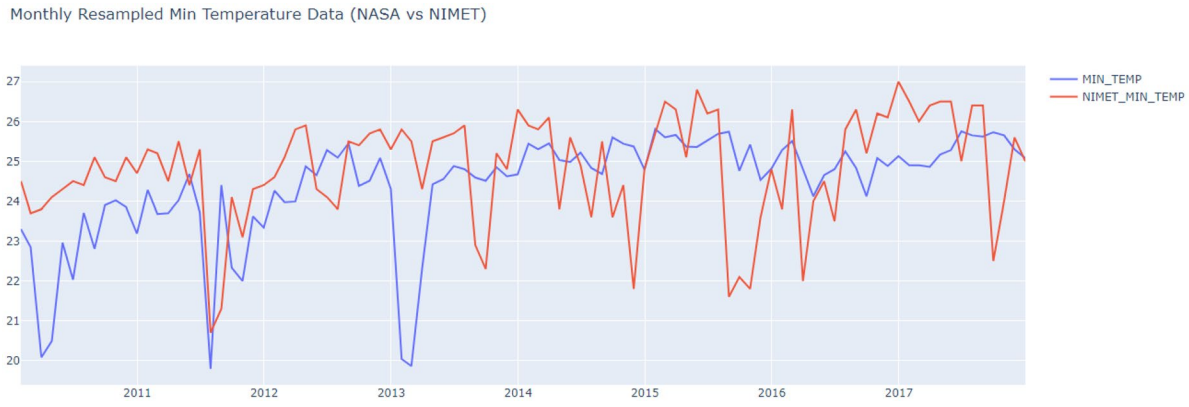

Fig. 11 Monthly resampled NiMet and NASA minimum temperature time series plot

(Idrissi et al., 2019). The lower the MAE and RMSE, the better the model (Elsaraiti & Merabet, 2021; Vulova et al., 2021).

The MAE and RMSE for daily maximum temperature are approximately 2.184 °C and 2.579 °C, respectively, while that of daily minimum temperature are approximately 0.876 °C and 1.225 °C, respectively. This implies that there are approximately 2.184 °C MAE and 2.579 °C RMSE between the NASA weather data and the NiMet ground truth measurement for maximum temperature in the study area, also approximately 0.876 °C MAE and 1.225 °C RMSE between both datasets for minimum temperature.

The extent of NASA prediction error from the NiMet ground truth measurement has been established statistically, but we also need to see and evaluate the extent of correlation between the NASA values and the NiMet values. This will evaluate the suitability of employing both datasets to build a prediction model with lower prediction error from ground truth measurement, hence, better prediction performance. The next section deals with the correlation analysis.

Step 4: correlation analysis

Correlation analysis or exploratory data analysis is employed, by creating a correlation heat map, to ascertain the suitability of the NASA data as an input feature for the development of the prediction model. Also, it will identify the feature with the highest correlation with the target and evaluates the extent of correlation. Figure 12 is a heat map of the correlation between the features and NiMet ground truth maximum temperature measurement, as the target.

The results from the correlation heat map show that there is a good correlation between the NASA maximum daily temperature (MAX_TEMP) and the NiMet maximum daily temperature (NIMET_MAX_TEMP), and it is, therefore, suitable for use as input in the prediction model development.

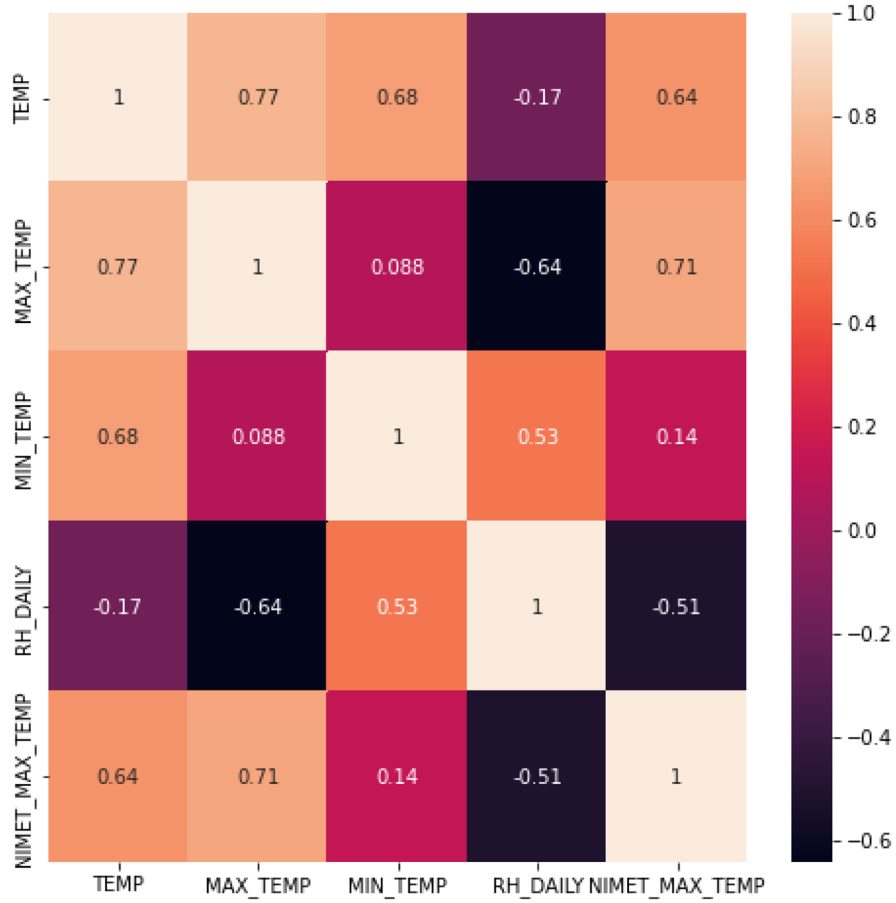

Fig. 12 Correlation heat map for maximum temperature

The NASA maximum daily temperature (MAX_TEMP) feature and the target, which is the ground truth NiMet maximum daily temperature (NIMET_MAX_TEMP), have the highest positive correlation, with a 0.71 correlation coefficient value. This is followed by the NASA average daily temperature (TEMP), with a 0.64 correlation coefficient value. Conversely, the NASA daily relative humidity (RH_DAILY) feature has a negative correlation with the target, with a value of − 0.51. Furthermore, the NASA minimum daily temperature (MIN_TEMP) feature is dropped, as it has a very low coefficient of determination value of 0.14 with the target. Experimentally, it is discovered that features with low correlation usually have a negative impact on the model training.

Accordingly, a correlation heat map showing the correlation between the features and NiMet ground truth minimum temperature measurement, as the target, is generated and shown in Fig. 13.

The NASA minimum daily temperature (MIN_TEMP) feature and the target, which is the ground truth NiMet minimum daily temperature (NIMET_MIN_TEMP), have the highest positive correlation, with a 0.62 coefficient of determination value. This is followed by the NASA average daily temperature (TEMP), with a 0.44 coefficient of determination value and NASA daily relative humidity, with a 0.36 coefficient of determination value. Furthermore, the NASA maximum daily temperature (MAX_TEMP) feature with a coefficient of determination of 0.07 is dropped, as it would have a negative impact when training the model.

Predictive, prescriptive, and performance analyses

Generally, there are two classes of machine learning algorithms: supervised and unsupervised learning. In this work, we have the NASA dataset as input,

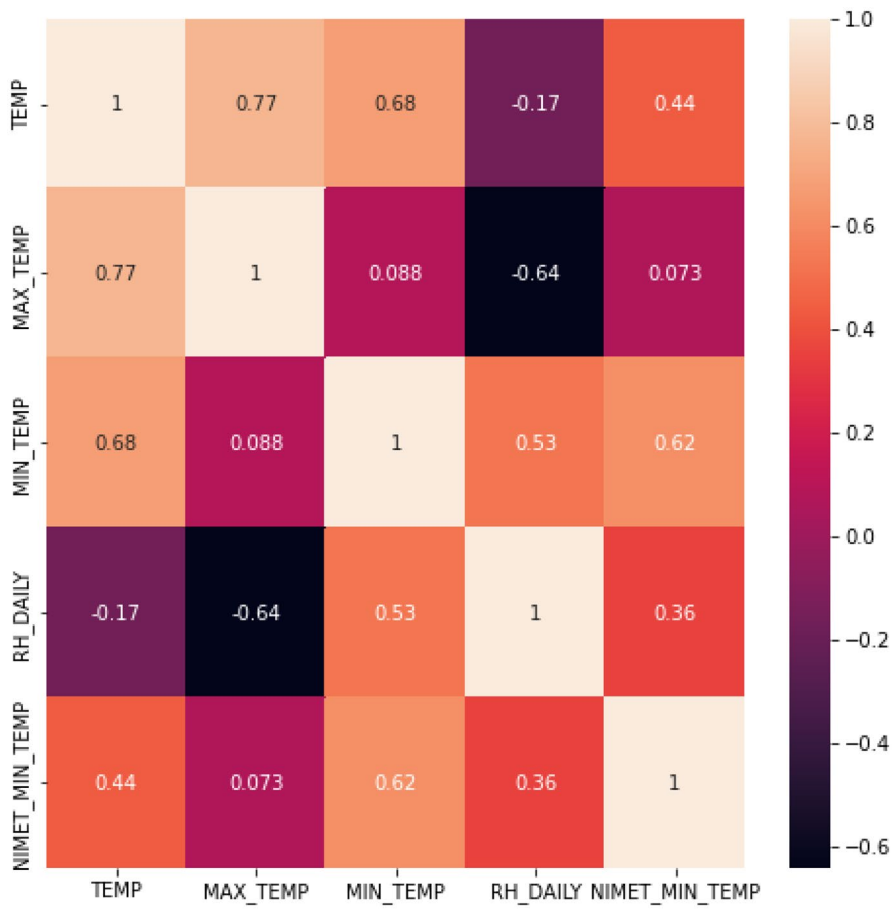

Fig. 13 Correlation heat map for minimum temperature

then the NiMet dataset as output and the process of training the system to learn the function that maps the NASA input to the NiMet output is a supervised learning task. This task is termed prediction. Prediction is involved with the evaluation of outputs from inputs, and this is achieved by fitting a model to a training data set, which generates an estimator $f(X)$ that has the capability of making predictions for new samples of X (Doring, 2018). It is important to note that this task is a regression (non-linear), as it involves the prediction of numerical values given some numerical input.

Mathematically,

$$Y = f(X) + C \tag{4}$$

where Y represent the output; f represents the relationship between the input and the output; X represents the input; C represents the random error.

A couple of processes were implemented, in preparation for data training, and eventual testing. The following are highlights of the processes:

1. Recall that features with very low coefficient of determination values were dropped, resulting from the correlation analysis. The remaining features were separated from the target, accordingly.
2. The dataset is split into a training set and a test set, in an 80:20 ratio, with the training set having the first 80% of the dataset, and the remaining 20% reserved for testing. The training set contains 2226 data samples, while the test set is 557 samples.
3. Data standardization is applied to the training and test set. Standardization is an important requirement for predictive modelling algorithms that sets all features to be on the same scale and minimize loss functions.

In furtherance of the previously implemented processes, and since we are dealing with a non-linear regression problem, algorithms designed to deal with such problems would be employed. As such, decision tree regression, XGBoost CART decision tree and multilayer perceptron algorithms are employed and implemented, and then their performances are evaluated using RMSE and R-squared performance metrics.

Decision tree regression

The decision tree regression is a supervised learning regression algorithm. It works in a tree-like form, by transiting from observing and input (branches) to drawing conclusions about the input (leaves). Implementation of the decision tree regression algorithm resulted in predictions, with an approximate RMSE value of 1.863 °C, MAE value of 1.471 °C, and an R-squared value of 0.531, for maximum temperature, then an approximate RMSE value of 1.150 °C, MAE value of 0.805 °C, and an R-squared value of 0.253, for minimum temperature. The results are presented in Figs. 14 and 15.

XGBoost regression

XGBoost (Extreme Gradient Boosting) is an optimized gradient boosting supervised learning algorithm that uses CART (Classification and Regression Trees). CART are trees containing real value scores in each leaf, irrespective of whether they are employed for regression or classification. Approximate RMSE and MAE values of 1.822 °C and 1.444 °C, respectively, with an R-squared value of 0.551 was recorded for maximum temperature, also approximate RMSE

	actual-nimet-max-temp	predicted-nimet-max-temp
0	35.5	32.875000
1	35.5	32.520067
2	34.9	33.424277
3	34.5	32.520067
4	34.6	32.067568
...	...	...
552	28.0	28.486667
553	29.5	28.105293
554	26.1	27.979651
555	27.2	26.893458
556	26.8	26.893458

```
MAE of Decision Tree Regressor: 1.471197
RMSE of Decision Tree Regressor: 1.863311
R Squared Decision Tree Regressor: 0.530982
```

Fig. 14 Maximum temperature predictions, RMSE, and R-squared evaluation using decision tree regression

	actual-nimet-min-temp	predicted-nimet-min-temp
0	24.000000	23.834522
1	24.900000	23.834522
2	24.000000	23.834522
3	24.500000	23.320124
4	24.200000	20.012500
...	...	...
552	22.300000	22.954252
553	23.300000	23.320124
554	23.500000	23.320124
555	23.303721	23.320124
556	23.000000	23.320124

```
MAE of Decision Tree Regressor: 0.805247
RMSE of Decision Tree Regressor: 1.150441
R Squared Decision Tree Regressor: 0.252656
```

Fig. 15 Minimum temperature predictions, RMSE, and R-squared evaluation using decision tree regression

and MAE values of 1.182 °C and 0.816 °C, respectively, with an *R*-squared value of 0.211 was recorded for minimum temperature. The results are presented in Figs. 16 and 17.

	actual-nimet-max-temp	predicted-nimet-max-temp
0	35.5	33.591656
1	35.5	32.588371
2	34.9	33.055836
3	34.5	32.705227
4	34.6	32.511822
...	...	...
552	28.0	28.093462
553	29.5	28.006184
554	26.1	28.098515
555	27.2	27.377220
556	26.8	27.489985

```
MAE of XGBoost Regressor: 1.444088
RMSE of XGBoost Regressor: 1.822444
R Squared XGBoost Regressor: 0.551330
```

Fig. 16 Maximum temperature predictions, RMSE, and R-squared evaluation using XGBoost regression

	actual-nimet-min-temp	predicted-nimet-min-temp
0	24.000000	22.991674
1	24.900000	23.572309
2	24.000000	24.053396
3	24.500000	23.423731
4	24.200000	21.340387
...	...	...
552	22.300000	22.956404
553	23.300000	23.115887
554	23.500000	23.308804
555	23.303721	22.802366
556	23.000000	22.822910

```
MAE of XGBoost Regressor: 0.815597
RMSE of XGBoost Regressor: 1.182266
R Squared XGBoost Regressor: 0.210736
```

Fig. 17 Minimum temperature predictions, RMSE, and R-squared evaluation using XGBoost regression

Accordingly, multilayer perceptron (MLP) algorithm is employed. MLP is a feed-forward ANN (artificial neural network) with more than one perceptron, with varying hyper-parameters, that learns the relationship between linear and nonlinear data. Choosing the right parameters is crucial to the performance of the MLP neural network.

An MLP is constituted of one or multiple hidden layers that connect between the input layer and the output layer, with an activation function that transforms the output of the hidden layer(s). However, research has shown that the default activation function recommendation for modern neural networks, including MLP, is the rectified linear unit (ReLU) activation function (Brownlee, 2021; Goodfellow et al., 2016). While the activation function controls how well the network model learns the training dataset, optimizers are used to tune the parameters of the network model to reduce errors. Research shows that there is no defined guideline for selecting an optimizer, but many authors argue that the Adam optimizer has been developed for large datasets and is a good choice to start with (Stack Exchange Network, 2018; Kingma & Ba, 2015; Okewu et al., 2019). Consequently, different MLP neural architecture, using available optimizers in the Scikit-learn machine learning library

for python integrated development environment, is employed for this work. The following is an implementation of MLP with different neural architectures, and their respective performance evaluation. It is important to note that there is no rule to evaluate the number of hidden layers/nodes in a multilayer perceptron. Although many authors suggest that 2 hidden layers are enough but research has shown that systematic experimentation is the approach to discovering what works best (Chavan, 2013). Consequently, varying architectures were experimented and we found that the architecture with 2 hidden layers, first with 4 nodes and second with 2 nodes, performed optimally.

MLP with LBFGS optimizer

LBFGS (limited-memory Broyden-Fletcher-Goldfarb-Shanno) optimizer, with two hidden layers: the first hidden layer with four nodes, while the second hidden layer with two nodes. For regularization and to control overfitting/underfitting, alpha (a parameter that controls the size of weights in the hidden layers) is set to 0.00001. The activation function used is ReLU. MAE and RMSE evaluation resulted in 1.412 °C and 1.778 °C, respectively, with an *R*-squared value of 0.573, for maximum temperature.

	actual-nimet-max-temp	predicted-nimet-max-temp
0	35.5	33.968471
1	35.5	33.001041
2	34.9	33.276980
3	34.5	33.055468
4	34.6	32.337246
...	...	...
552	28.0	28.639041
553	29.5	28.193822
554	26.1	27.811368
555	27.2	27.071183
556	26.8	27.088545

```
MAE of LBFGS MLP Regressor: 1.412065
RMSE of LBFGS MLP Regressor: 1.777780
R Squared LBFGS MLP Regressor: 0.573052
```

Fig. 18 Maximum temperature predictions, MAE, RMSE, and R-squared evaluation using LBFGS MLP

	actual-nimet-min-temp	predicted-nimet-min-temp
0	24.000000	23.271720
1	24.900000	23.661682
2	24.000000	23.809720
3	24.500000	23.283484
4	24.200000	21.827789
...	...	...
552	22.300000	22.916439
553	23.300000	23.160132
554	23.500000	23.364923
555	23.303721	23.082189
556	23.000000	23.077392

```
MAE of LBFGS MLP Regressor: 0.788126
RMSE of LBFGS MLP Regressor: 1.127602
R Squared LBFGS MLP Regressor: 0.282035
```

Fig. 19 Minimum temperature predictions, MAE, RMSE, and R-squared evaluation using LBFGS MLP

MAE and RMSE evaluation resulted in 0.788 °C and 1.127 °C, respectively, with an *R*-squared value of 0.282, for minimum temperature. The results are presented in Figs. 18 and 19.

MLP with SGD optimizer

SGD (stochastic gradient) optimizer, with two hidden layers; the first hidden layer with four nodes, while the second hidden layer with two nodes. For regularization and to control overfitting/underfitting, alpha is set to 0.00001. The activation function used is ReLU. MAE and RMSE evaluation resulted in 1.424 °C and 1.786 °C, respectively, with an *R*-squared value of 0.569, for maximum temperature. MAE and RMSE evaluation resulted in 0.794 °C and 1.126 °C, respectively, with an *R*-squared value of 0.284, for minimum temperature. The results are presented in Figs. 20 and 21.

MLP with Adam optimizer

Adam optimizer, with two hidden layers; the first hidden layer with four nodes, while the second hidden layer with two nodes. For regularization and to control

	actual-nimet-max-temp	predicted-nimet-max-temp
0	35.5	33.641967
1	35.5	32.806663
2	34.9	33.401182
3	34.5	32.648589
4	34.6	32.954214
...	...	...
552	28.0	28.481014
553	29.5	28.735572
554	26.1	28.026865
555	27.2	27.204876
556	26.8	27.275940

```
MAE of SGD MLP Regressor: 1.424171
RMSE of SGD MLP Regressor: 1.786375
R Squared SGD MLP Regressor: 0.568914
```

Fig. 20 Maximum temperature predictions, MAE, RMSE, and R-squared evaluation using SGD MLP

overfitting/underfitting, alpha is set to 0.00001. The activation function used is ReLU. MAE and RMSE evaluation resulted in 1.428 °C and 1.797 °C, respectively, with an R-squared value of 0.564, for maximum temperature. MAE and RMSE evaluation resulted in 0.795 °C and 1.128 °C, respectively, with an R-squared value of 0.282, for minimum temperature. The results are presented in Figs. 22 and 23.

	actual-nimet-min-temp	predicted-nimet-min-temp
0	24.000000	23.309587
1	24.900000	23.666925
2	24.000000	23.766366
3	24.500000	23.323213
4	24.200000	21.677778
...	...	...
552	22.300000	22.891011
553	23.300000	23.161258
554	23.500000	23.365541
555	23.303721	23.058272
556	23.000000	23.048999

```
MAE of SGD MLP Regressor: 0.794293
RMSE of SGD MLP Regressor: 1.126182
R Squared SGD MLP Regressor: 0.283842
```

Fig. 21 Minimum temperature predictions, MAE, RMSE, and R-squared evaluation using SGD MLP

	actual-nimet-max-temp	predicted-nimet-max-temp
0	35.5	33.625305
1	35.5	33.103925
2	34.9	33.569347
3	34.5	32.991443
4	34.6	33.003385
...	...	...
552	28.0	28.582274
553	29.5	28.658506
554	26.1	27.940183
555	27.2	27.093066
556	26.8	27.171503

```
MAE of ADAM MLP Regressor: 1.427722
RMSE of ADAM MLP Regressor: 1.796539
R Squared ADAM MLP Regressor: 0.563994
```

Fig. 22 Maximum temperature predictions, MAE, RMSE, and R-squared evaluation using Adam MLP

	actual-nimet-min-temp	predicted-nimet-min-temp
0	24.000000	23.392786
1	24.900000	23.786466
2	24.000000	23.905462
3	24.500000	23.425179
4	24.200000	21.541489
...	...	...
552	22.300000	22.819485
553	23.300000	23.104213
554	23.500000	23.268100
555	23.303721	23.354565
556	23.000000	23.299830

```
MAE of ADAM MLP Regressor: 0.795217
RMSE of ADAM MLP Regressor: 1.127506
R Squared ADAM MLP Regressor: 0.282156
```

Fig. 23 Minimum temperature predictions, MAE, RMSE, and R-squared evaluation using Adam MLP

Table 2 Predictive models employed and their performance evaluation

Feature	Predictive algorithm	MAE (°C)	RMSE (°C)	*R*-squared
Maximum temperature	NASA	2.184	2.579	0.710
	Decision tree regression	1.471	1.863	0.531
	XGBoost regression	1.444	1.822	0.551
	LBFGS multilayer perceptron	1.412	1.778	0.573
	SGD multilayer perceptron	1.424	1.786	0.569
	Adam multilayer perceptron	1.428	1.797	0.564
Minimum temperature	NASA	0.876	1.225	0.620
	Decision tree regression	0.805	1.150	0.253
	XGBoost regression	0.816	1.182	0.211
	LBFGS multilayer perceptron	0.788	1.127	0.282
	SGD multilayer perceptron	0.794	1.126	0.284
	Adam multilayer perceptron	0.795	1.128	0.282

Results and discussion

Tables 2 and 3 show, in summary, the NASA prediction and other prediction algorithms employed, with their corresponding performance evaluation.

From the results generated, models built with the multilayer perceptron algorithm performed better than models built with decision tree algorithms. Although XGBoost regression performed better than decision tree regression in predicting the daily maximum temperature for the study area, it could not outperform its counterpart in the prediction of the study area's daily minimum temperature. Accordingly, amongst the multilayer perceptron, the LBFGS multilayer perceptron slightly outperformed its counterparts in predicting both the maximum and minimum temperatures of the study area, except for one occasion where the SGD multilayer perceptron did better in RMSE, by 0.08%.

Generally, RMSE tells the performance of the model by evaluating the average difference between the predictions and actual measurements; the lower the difference, the better the model. While *R*-squared shows the performance of the model by evaluating how well the model fits the ground truth data values. *R*-squared does not tell how good a model is alone; this explains why it is usually used alongside other performance metrics (Chicco et al., 2021). However, a coefficient of determination value of ≥ 0.5 stipulates a good correlation between the model and the ground truth measurement. Since this is a regression analysis problem, which implies that we are more interested in the prediction values, the *R*-squared value will not matter.

Table 3 Predictive models employed and their performance improvement WRT NASA prediction

Feature	Predictive algorithm	MAE	RMSE
Maximum temperature	NASA	2.184	2.579
	Decision tree regression	32.65%	27.76%
	XGBoost regression	33.88%	29.35%
	LBFGS multilayer perceptron	35.35%	31.06%
	SGD multilayer perceptron	34.80%	30.75%
	Adam multilayer perceptron	34.62%	30.32%
Minimum temperature	NASA	0.876	1.225
	Decision tree regression	8.11%	6.12%
	XGBoost regression	6.85%	3.51%
	LBFGS multilayer perceptron	10.05%	8.00%
	SGD multilayer perceptron	9.36%	8.08%
	Adam multilayer perceptron	9.25%	7.92%

Conclusion and recommendation

Descriptive and diagnostic analyses of NiMet and NASA maximum temperature datasets for the study area were performed, with trends and patterns identified. Using the NiMet data as ground truth, performance evaluation of the NASA data reported a mean absolute error (MAE) of 2.184 °C and root mean squared error (RMSE) of 2.579 °C, while correlation analysis reported a coefficient of determination (R^2) value of 0.710. This is an indication of a good correlation between the two datasets, in addition, a model can be developed to predict more accurately, weather data for the study area, using the NASA data as input. Predictive and prescriptive analyses were performed by employing five prediction algorithms, from the decision tree class and artificial neural network class. The five prediction algorithms employed are decision tree regression, XGBoost regression, MLP (multilayer perceptron) with LBFGS (limited-memory Broyden-Fletcher-Goldfarb-Shanno) optimizer, MLP with SGD (stochastic gradient) optimizer and MLP with Adam optimizer. MLP with LBFGS optimizer performed best, by reducing the MAE from 2.184 to 1.412 °C (35.35%) and RMSE from 2.579 to 1.778 °C (31.06%) for maximum temperature. Accordingly, a reduction in MAE from 0.876 to 0.788 °C (10.05%) and RMSE from 1.225 to 1.127 °C (8.00%) for minimum temperature. Further improvement can still be achieved by employing ensemble models.

Acknowledgements Profound gratitude to the management and staff of the Advanced Space Technology Applications Laboratory (ASTAL), the ASTAL Digital Image Processing Laboratory (DIPL), and the National Space Research and Development Agency (NASRDA) for providing an enabling environment for this work. Appreciation to Kanda Weather Group LLC and the Nigerian Meteorological Agency (NiMet) for provision of ground truth weather data.

Author contribution All authors contributed to the study conceptualization and methodology. Data collection was carried out by SO. Data processing and analysis was carried out by AO. First draft of the manuscript was written by AO, and all authors commented on all versions of the manuscript. All authors read and approved the final manuscript.

Data availability Processed data are available upon request to the authors.

Declarations

Conflict of interest The authors declare no competing interests.

References

Abhishek, S., Neeta, V., & Tripathi, K. (2013). A review study of weather forecasting using artificial neural network approach. *International Journal of Engineering Research & Technology, 2*(11), 2029–2035.

Aboelkhair, H., Morsy, M., & El Afandi, G. (2019). Assessment of agroclimatology NASA POWER reanalysis datasets for temperature types and relative humidity at 2 meter against ground observations over Egypt. *Advances in Space Research, 64*, 129–142. https://doi.org/10.1016/j.asr.2019.03.032

Bhardwaj, R., & Duhoon, V. (2018). Weather forecasting using soft computing techniques. *2018 International Conference on Computing, Power and Communication Technologies (GUCON)* (pp. 1111–1115). New Delhi: IEEE. https://doi.org/10.1109/GUCON.2018.8675088

Brownlee, J. (2021). *How to choose an activation function for deep learning*. (Machine Learning Mastery) Retrieved June 21, 2022, from https://machinelearningmastery.com/choose-an-activation-function-for-deep-learning/

Chavan, P. (2013). *How to decide the number of hidden layers and nodes in a hidden layer?* Retrieved from Researchgate: https://www.researchgate.net/post/How-to-decide-the-number-of-hidden-layers-and-nodes-in-a-hidden-layer

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science, 7*, 623.

Denissen, J., Butalid, L., Penke, L., & Aken, M. (2008). The effects of weather on daily mood: A multilevel approach. *Emotion (Washington, D.C.), 8,* 662–7. https://doi.org/10.1037/a0013497

Doring, M. (2018). *Prediction vs forecasting*. Retrieved July 9, 2022, from Data Science Blog: https://www.datascienceblog.net/post/machine-learning/forecasting_vs_prediction/

Dundas, S. J., & Von Haefen, R. H. (2021). The importance of data structure and nonlinearities in estimating climate impacts on outdoor recreation. *Natural Hazards, 107*(3), 2053–2075. https://doi.org/10.1007/s11069-020-04484-w

Elsaraiti, M., & Merabet, A. (2021). A comparative analysis of the ARIMA and LSTM predictive models and their effectiveness for predicting wind speed. *Energies, 14*(20). https://doi.org/10.3390/en14206782

Fathi, M., Haghi Kashani, M., & Jemeii, S. (2021). Big data analytics in weather forecasting: A systematic review. *Archives of Computational Methods in Engineering*. https://doi.org/10.1007/s11831-021-09616-4

Faybishenko, B., Versteeg, R., & Pastorello, G. (2022). Challenging problems of quality assurance and quality control (QA/QC) of meteorological time series data. *Stochastic*

Environmental Research and Risk Assessment, 36, 1049–1062. https://doi.org/10.1007/s00477-021-02106-w

Findawati, Y., Indra Astutik, I., Fitroni, A., Indrawati, I., & Yuniasih, N. (2019). Comparative analysis of Naïve Bayes, K Nearest Neighbor and C.45 method in weather forecast. *Journal of Physics: Conference Series*, 1–6. https://doi.org/10.1088/1742-6596/1402/6/066046

Gad, I., & Hosahalli, D. (2022). A comparative study of prediction and classification models on NCDC weather data. *International Journal of Computers and Applications, 44*(5), 414–425. https://doi.org/10.1080/1206212X.2020.1766769

Garbade, M. (2018). *Regression versus classification machine learning: What's the difference?* (Medium) Retrieved December 7, 2021, from https://medium.com/quick-code/regression-versus-classification-machine-learning-whats-the-difference-345c56dd15f7

Gelaro, R., McCarty, W., Suárez, M., 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.

Gill, M., Asefa, T., Kaheil, Y., & Mckee, M. (2007). Effects of missing data on performance of learning algorithms for hydrologic predictions. *Advancing Earth and Space Science*, 50–62.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning (adaptive computation and machine learning series)*. MIT Press.

Gupta, I., Mittal, H., Rikhari, D., & Singh, A. K. (2022). *MLRM: a multiple linear regression based model for average temperature prediction of a day*. (arXiv) Retrieved July 09, 2022, from https://arxiv.org/abs/2203.05835

Halabi, M. L., Mekhilef, S., Olatomiwa, L., & Hazelton, J. (2017). Performance analysis of hybrid PV/diesel/battery system using HOMER: A case study Sabah, Malaysia. *Energy Conversion and Management*, 322–339. https://doi.org/10.1016/j.enconman.2017.04.070

Idrissi, E. L., & T., Idri, A., & Bakkoury, Z. (2019). Systematic map and review of predictive techniques in diabetes self-management. *International Journal of Information Management, 46*, 263–277. https://doi.org/10.1016/j.ijinfomgt.2018.09.011

Johnson, T. F., Isaac, N. J., Paviolo, A., & González-Suárez, M. (2021). Handling missing values in trait data. *Global Ecology and Biogeography, 30*(1), 51–62. https://doi.org/10.1111/geb.13185

Kaneko, A., Zhu, X. -H., & Lin, J. (2020). Data Assimilation. In A. Kaneko, X. -H. Zhu, & J. Lin, *Coastal Acoustic Tomography* (pp. 95–106). Taizhou: Elsevier. https://doi.org/10.1016/B978-0-12-818507-0.00008-1

Khajure, S., & Mohod, S. W. (2016). Future weather forecasting using soft computing techniques. *Procedia Computer Science*, *78*, 402–407. Nagpur. https://doi.org/10.1016/j.procs.2016.02.081

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. *3rd International Conference for Learning Representations.* San Diego.

Kong, W., Li, H., Yu, C., Xia, J., Kang, Y., & Zhang, P. (2022). A deep spatio-temporal forecasting model for multi-site weather prediction post-processing. *Communications in Computational Physics, 31*, 131–153.

Kusiak, A., & Shah, S. (2006). Data-mining-based system for prediction of water chemistry faults. *IEEE Transactions on Industrial Electronics, 53*(2), 593–603. https://doi.org/10.1109/TIE.2006.870706

Li, Y., Lang, J., Ji, L., Zhong, J., Wang, Z., Guo, Y., & He, S. (2021). Weather forecasting using ensemble of spatial-temporal attention network and multi-layer perceptron. *Asia-Pacific Journal of Atmospheric Sciences, 57*, 533–546. https://doi.org/10.1007/s13143-020-00212-3

Markovics, D., & Mayer, M. J. (2022). Comparison of machine learning methods for photovoltaic power forecasting based on numerical weather prediction. *Renewable and Sustainable Energy Reviews, 161*(112364). https://doi.org/10.1016/j.rser.2022.112364

Marson, S., & Legerton, M. (2021). Disaster diaspora and the consequences of economic displacement and climate disruption, including Hurricanes Matthew (October 8, 2016) and Florence (September 14, 2018) in Robeson County. *North Carolina. Natural Hazards, 107*(3), 2247–2262. https://doi.org/10.1007/s11069-021-04529-8

Maydon, T. (2017). *The 4 Types of Data Analytics*. (KD Nuggets). Retrieved April 04, 2022, from https://www.kdnuggets.com/2017/07/4-types-data-analytics.html

NASA. (2022). *NASA Power Data Access Viewer*. Retrieved October 18, 2021, from https://power.larc.nasa.gov/data-access-viewer/

Nature. (2021). The rise of data-driven modelling. *Nature Reviews Physics, 3*(6), 383. https://doi.org/10.1038/s42254-021-00336-z

Nikam, V., & Meshram, B. (2013). Modeling rainfall prediction using data mining method: A Bayesian approach. *2013 Fifth International Conference on Computational Intelligence, Modelling and Simulation.* Seoul. Retrieved March 8, 2013, from https://doi.org/10.1109/CIMSim.2013.29

NiMet. (2022). *Nigerian meteorological agency*. Retrieved October 8, 2021, from https://www.nimet.gov.ng/

Nnah, B. C., Okenwa, A. I., Oloyede, O. A., Nwaibe, O., & Agbu, A. U. (2021). Geospatial assessment of urban heat island in Port Harcourt L.G.A, Rivers State, Nigeria. *International Journal of Sciences: Basic and Applied Research (IJSBAR)*, 33–55.

Okewu, E., Adewole, P., & Sennaike, O. (2019). Experimental comparison of stochastic optimizers in deep learning. *International Conference on Computational Science and Its Applications* (pp. 704–715). Saint Petersburg: Springer. https://doi.org/10.1007/978-3-030-24308-1_55

Olaiya, F., & Adeyemo, A. (2012). Application of data mining techniques in weather prediction and climate change studies. *I.J. Information Engineering and Electronic Business, 1*, 51–59. https://doi.org/10.5815/ijieeb.2012.01.07

Olatomiwa, L., Mekhilef, S., Shamshirband, S., Mohammadi, K., Petković, D., & Sudheer, C. (2015). A support vector machine–firefly algorithm-based model for global solar radiation prediction. *Solar Energy, 115*, 632–644. https://doi.org/10.1016/j.solener.2015.03.015

Oloyede, A. O., Olatunbosun, D. E., Asuquo, P. M., Udo, U. E., & Essien, I. O. (2021). Correlation Analysis of Vegetation and Land Surface Temperature in Uyo, Nigeria Using Satellite Remote Sensing and Python-Based Geographic Information System. *Science and Technology Publishing*, 1126–1133.

Oloyede, A., Ozuomba, S., & Asuquo, P. (2022). Descriptive and diagnostic analysis of NASA and NiMet big weather data.

2022 IEEE Nigeria 4th International Conference on Disruptive Technologies for Sustainable Development (NIGERCON). Abuja. https://doi.org/10.1109/NIGERCON54645.2022.9803015

Osama, M. A. (2021). Assessment of global warming in Al Buraimi, sultanate of Oman based on statistical analysis of NASA POWER data over 39 years, and testing the reliability of NASA POWER against meteorological measurements. *Heliyon,* (3), 1–19. https://doi.org/10.1016/j.heliyon.2021.e06625

Petre, E. G. (2009). A decision tree for weather prediction. *BMIF, LXI*(1), 77–82.

Quansah, A. D., Dogbey, F., Asilevi, P. J., Boakye, P., Darkwah, L., Oduro-Kwarteng, S., & Mensah, P. (2022). Assessment of solar radiation resource from the NASA-POWER reanalysis products for tropical climates in Ghana towards clean energy application. *Scientific Reports*, 1–10. https://doi.org/10.1038/s41598-022-14126-9

Rodrigues, G. C., & Braga, R. P. (2021). Evaluation of NASA POWER reanalysis products to estimate daily weather variables in a hot summer Mediterranean climate. *Agronomy, 11*(6), 1207. https://doi.org/10.3390/agronomy11061207

Romero-Fiances, I., Livera, A., Theristis, M., Makrides, G., Stein, J. S., Nofuentes, G., & Georghio, G. E. (2022). Impact of duration and missing data on the long-term photovoltaic degradation rate estimation. *Renewable Energy, 181*, 738–748. https://doi.org/10.1016/j.renene.2021.09.078

Sheikh, F., Karthick, S., Malathi, D., Sudarsan, J., & Arun, C. (2016). Analysis of data mining techniques for weather prediction. *Indian Journal of Science and Technology, 9*(38), 1–9. https://doi.org/10.17485/ijst/2016/v9i38/101962

Sher, V. (2020). *Time series analysis using pandas in Python*. (Towards Data Science) Retrieved December 21, 2021, from https://towardsdatascience.com/time-series-analysis-using-pandas-in-python-f726d87a97d8

Stack Exchange Network. (2018). *Stack exchange network*. (Stack Exchange Inc.) Retrieved June 21, 2022, from https://datascience.stackexchange.com/questions/10523/guidelines-for-selecting-an-optimizer-for-training-neural-networks

Tan, L., Guo, J., Mohanarajah, S., & Zhou, K. (2021). Can we detect trends in natural disaster management with artificial intelligence? A review of modeling practices. *Natural Hazards, 107*(3), 2389–2417. https://doi.org/10.1007/s11069-020-04429-3

Twin, A. (2021). *Data Mining*. (Investopedia) Retrieved November 30, 2021, from https://www.investopedia.com/terms/d/datamining.asp

Vulova, S., Meier, F., Rocha, A. D., Quanz, J., Nouri, H., & Kleinschmit, B. (2021). Modeling urban evapotranspiration using remote sensing, flux footprints, and artificial intelligence. *Science of The Total Environment, 786*. https://doi.org/10.1016/j.scitotenv.2021.147293

Waring, R. H., & Running, S. W. (2007). Spatial Scaling Methods for Landscape and Regional Ecosystem Analysis. In *Forest Ecosystems (Third Edition)* (p. 225). Academic Press.

Xi, X., Zuo, J., Dooling, T. A., & Mohanarajah, S. (2021). Bayesian network reasoning and machine learning with multiple data features: Air pollution risk monitoring and early warning. *Natural Hazards, 107*(3), 2555–2572. https://doi.org/10.1007/s11069-021-04504-3

Zhang, Y., & Thorburn, P. J. (2022). Handling missing data in near real-time environmental monitoring: A system and a review of selected methods. *Future Generation Computer Systems, 128*, 63–72. https://doi.org/10.1016/j.future.2021.09.033