Scientific Research Publishing

Open Journal of Air Pollution, 2025, 14(4), 73-100
https://www.scirp.org/journal/ojap
ISSN Online: 2169-2661
ISSN Print: 2169-2653

Using Machine Learning Models to Predict Daily PM_{10} Concentration in the Wet Savanna of Lamto Station in Côte d'Ivoire

Touré E. N'Datchoh[1*], Money Ossohou[1,2], Adama Bamba[1], Yoman C. E. Etien[2], Kouakou Kouadio[1,3], Fidèle Yoroba[1,3,4], Madina Doumbia[1,5], Mohamed L. Kassamba-Diaby[1,6], Sylvain Gnamien[1], Véronique Yoboué[1,6]

[1]UFR SSMT, LASMES, University Félix Houphouët-Boigny, Abidjan, Côte d'Ivoire
[2]Department of Physics, University of Man, Man, Côte d'Ivoire
[3]Station Geophysic of Lamto, N'Douci, Côte d'Ivoire
[4]UFR Sciences of Technology, Department of Physics, University Alassane Ouattara of Bouaké, Bouaké, Côte d'Ivoire
[5]University Peleforo Gon Coulibaly, Korhogo, Côte d'Ivoire
[6]University Nangui Abrogoua, Abidjan, Côte d'Ivoire
Email: *ndatchoh.toure01@ufhb.edu.ci, *Evelyne.Toure@outlook.com

How to cite this paper: N'Datchoh, T.E., Ossohou, M., Bamba, A., Etien, Y.C.E., Kouadio, K., Yoroba, F., Doumbia, M., Kassamba-Diaby, M.L., Gnamien, S. and Yoboué, V. (2025) Using Machine Learning Models to Predict Daily PM_{10} Concentration in the Wet Savanna of Lamto Station in Côte d'Ivoire. *Open Journal of Air Pollution*, **14**, 73-100.
https://doi.org/10.4236/ojap.2025.144006

Received: August 31, 2025
Accepted: November 2, 2025
Published: November 5, 2025

Abstract

Despite its significant impacts on climate, environment, and public health, air pollution monitoring in Africa is still sparse. This study developed a machine learning framework of four models—namely SARIMAX, Random Forest, XGBoost, and LightGBM—that were trained on a daily basis using 25 meteorological variables from ERA5 reanalysis to reproduce daily PM_{10} surface concentrations in Lamto, Côte d'Ivoire. A total of 2225 daily recorded concentrations from 2017 to 2023 were used, with 80% (1780 days) allocated to training the models and 20% (445 days) for prediction. An assessment of ERA5 reanalysis data against *in-situ* measurements highlighted its ability to reproduce the pattern of meteorological variables such as temperature and relative humidity, despite some biases. Results show that all models were able to reproduce the observed PM_{10} variations, although they slightly overestimated concentrations during the main wet season (March-April-May-June-July) and underestimated high pollution events in the main dry season (November-December-January-February). Finally, performance analysis revealed that the Random Forest model outperformed, with an R^2 of 0.78 and RMSE of 23 $\mu g/m^3$, outperforming the other models. This framework successfully demonstrates the utility of machine learning for air quality prediction in West Africa, with potential for future improvements through bias correction and model combination.

Keywords

Lamto, Prediction, PM_{10}, Machine Learning, Air Pollution, Performance

1. Introduction

Human beings, through their activities, have strongly impacted and shaped their environment, including the Earth's atmospheric composition. Since the Industrial Revolution in the 18th century, numerous compounds have been released into the atmosphere. These atmospheric compounds are responsible for current global warming, which induces climate change [1] [2]. They contain aerosols, defined as particulate matter in a solid or liquid state suspended in the atmosphere. These aerosols are reported to play a significant role in the climate through their direct, indirect, and semi-direct effects [3] [4]. Their size distribution in the total suspended particles is often classified into three (3) modes: coarse (PM_{10}), fine ($PM_{2.5}$), and ultra-fine (PM_1) particles. Their lifetime in the atmosphere is strongly dependent on their mode [5] and generally ranges from a few hours to several weeks. The coarse mode particles tend to be deposited close to their sources, while fine and ultra-fine particles, which contribute to particulate matter (PM), have a longer lifetime and can undergo long-range transport. These particles contribute to air quality deterioration and have numerous impacts on climate, ecosystems, and human health [6]-[8].

Air quality and its impacts in West Africa have drawn attention in several works, revealing increasing air pollution affecting population health [9] [10]. In urban areas, they are often investigated using ground-based observations [11]-[15], satellite [16], and modeling [17] [18]. During the Dynamics-Aerosol-Chemistry-Cloud Interactions in West Africa (DACCIWA) project, numerous studies have highlighted that anthropogenic pollutants emitted from the coastal region of the Guinean Coast were transported to remote areas [19] [20]. As urban areas' air quality received considerable attention, only a few studies have focused on air quality in rural areas. Investigating daily Aerosol Optical Depth (AOD) in 11 stations across various biomes of Southwest Africa, Niamien [21] found that their AOD ranged from 0.35 in the Sahel area to 0.49 in the Guinea Coast area, and a maximum of 0.53 was observed in the savanna area. Moreover, the International Network to study Deposition and Atmospheric Chemistry in Africa (INDAAF, https://indaaf.obs-mip.fr), has allowed researchers to focus attention on the atmospheric chemical composition of various biomes and rural areas in Africa [22]-[25]. For example, Lamto atmospheric composition was reported to be influenced by emissions from both local and regional sources, particularly from mesoscale sources with a strong anthropogenic signature [26].

Another important aspect of air quality degradation is its strong impact on the environment and human health. According to the World Health Organization (WHO), air pollution is responsible for more than 7 million premature deaths

every year, particularly affecting urban and peri-urban areas in developing countries [27]. In Africa, a report on the current situation of air quality highlighted an increase in air pollution across the continent, which was responsible for 250,000 premature deaths in 2018 [28]. Among these air pollutants, aerosols, particularly PM_{10}, are recognized as some of the most harmful components for both human health and ecosystems. For instance, PM_{10} is linked to respiratory and cardiovascular diseases and contributes to climate imbalance [29]. Moreover, prolonged exposure to high concentrations of PM_{10} has been associated with increased hospital admissions for respiratory conditions such as asthma, obstructive lung disease, and pulmonary infections. A connection between peaks in PM_{10} levels and deaths induced by cardiovascular disease has been indicated through strong correlations, highlighting an increase in myocardial infarction, stroke, and hypertension [30]. Despite their hazardous nature, monitoring of these particles remains limited, especially in West Africa, where the high cost of referenced instruments and the scarce infrastructure obstruct continuous monitoring. This lack of data compromises the implementation of effective public policies for air quality management. However, certain scientific initiatives, such as the INDAAF program, have enabled the regular and long-term collection of atmospheric chemistry data from multiple sites in Africa [22]-[25]. Over Lamto, an INDAAF supersite located in the center of Côte d'Ivoire, the atmospheric chemistry composition is monitored. The composition of the atmosphere at this station is the result of complex interactions between natural and anthropogenic sources in a rural area [26]. This makes Lamto an ideal site for studying how particulate compounds interact with meteorological conditions. However, spatiotemporal coverage remains too limited for full operational monitoring of this area. In response to these structural limitations, the integration of new technological approaches is emerging as a promising alternative. Among them, Artificial Intelligence (AI), and more specifically Machine Learning (ML) methods, offer powerful tools to compensate for the lack of data and improve predictive capabilities. The ML algorithms usually can analyze data from various sources, including air quality sensors, satellite imaging, and weather data, to provide real-time information on air quality [31]-[33]. Information about air pollution sources, pollution levels, and attenuation actions can be provided by these algorithms under the conditions they have been trained to recognize patterns and trends in the data [32]. However, all these algorithms do not perform similarly, and their choice may be highly dependent on the nature of pollutants and the purpose of the study. For example, in their health-related research, Dobrea [34] compared other AI-oriented techniques/models for air pollutants and suggested Support Vector Regression (SVR) and AutoRegressive Integrated Moving Average (ARIMA) as the best-performing techniques for time series analysis of PM_{10} and $PM_{2.5}$. Moreover, some works emphasized that hybrid models have better performance and present several advantages for environmental monitoring policy and decision-making [35] [36]. Therefore, by exploiting existing data, these techniques enable the anticipation of changes in pollutant levels

such as PM_{10}, thereby supporting environmental management efforts, even in low-resource settings. Their capacity to identify complex, non-linear relationships makes them particularly suitable for environments like Lamto, where multiple factors influence atmospheric processes. In this context, the present study aims to explore for the first time the potential of Artificial Intelligence (AI) models to reproduce and predict PM_{10} concentrations at Lamto over seven years (2017-2023). To achieve this, four AI models were employed, and their performances were assessed.

2. Materials and Methods

2.1. Sampling Site

Lamto is a station located at 6˚13'N and 5˚2'W, approximately 120 km from Abidjan (Côte d'Ivoire), situated within the humid savanna (**Figure 1**). The Lamto reserve was established in 1962, covering about 2,617 ha, and has been the subject of numerous research studies [37]-[42]. The reserve contains various measurement observation sites for seismology, infrasound, climatology, Greenhouse Gas (CarboAfrica European project, http://www.carboafrica.eu/), atmospheric mercury, atmospheric chemistry (INDAAF), and ecology. The reserve is often burned around mid-January to remove epigenetic phytomass and conduct ecological experiments [41] [42]. The climate of the Lamto reserve is governed by the West African Monsoon (WAM) and the seasonal shift of the Intertropical Convergence Zone (ITCZ). The annual mean rainfall ranges between approximately 1240 and 1590 mm [40], with four main seasons: a dry season (November to February), a wet season (March to July), a slightly dry season in August, due to monsoon jump, and a slightly wet season from September to November (monsoon retreat). Temperatures are relatively hot throughout the year, averaging around 27˚C, while

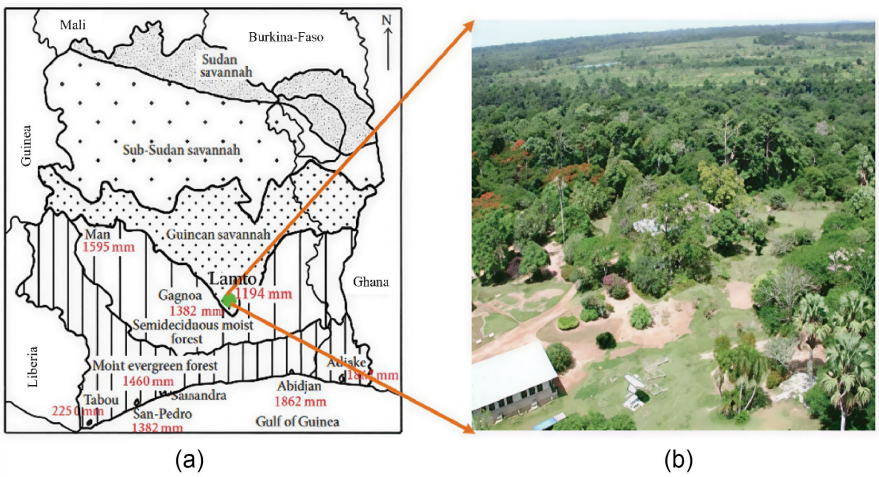

Figure 1. (a) Localization map of the Lamto-(Côte d'Ivoire) reserve (6˚31'N-5˚02'W). The green point indicates the Lamto location; numbers in red are the mean annual precipitation (1962-1997) of some synoptic stations, and from north to south, are located the types of vegetation [43]; and (b) zoom on the vegetation of Lamto [44].

relative humidity varies between 65% (minimum in January during the main dry season) and 82% (maximum in June during the main and first rainy season).

2.2. Data

2.2.1. PM_{10}

The PM_{10} concentration data used in our study were retrieved using the TEOM (Tapered Element Oscillating Microbalance) 1400A, covering seven years from January 1, 2017 to December 31, 2023, at the Lamto station [45]. Lamto is one of the super sites of the INDAAF program (https://indaaf.obs-mip.fr/). INDAAF is a National Observation Service (SNO) of the Institut National des Sciences de l'Univers (INSU) of the Centre National de Recherche Scientifique (CNRS) and is supported by Institut de Recherche pour le Développement (IRD). These PM_{10} concentration data are available on the INDAAF program website at an hourly scale in micrograms per cubic meter ($\mu g/m^3$). To reduce the number of missing values, we aggregated the PM_{10} data to compute daily averages. This approach led to 16.43% missing values across the entire dataset.

2.2.2. Meteorological and Physical Data

Hourly weather and climate parameter data were extracted from the ERA5 database [46] at the grid point closest to the Lamto station. ERA5 is the fifth generation of reanalyses from the European Centre for Medium-Range Weather Forecasts (ECMWF). The data, available from 1940 to the present day, relate to climate and weather conditions on a global scale over the last eight decades. For the present work, 25 meteorological variables (**Table 1**) have been selected from January 1, 2017 to December 31, 2023. These variables were chosen because they are known to represent weather conditions and atmospheric dynamics. Indeed, these variables relate to wind, temperature, humidity, the planetary boundary layer, and precipitation, which are known to influence the dispersion, accumulation, and leaching of aerosols [47]-[50].

Table 1. List of ERA5 meteorological variables used for the machine learning model training and prediction of PM_{10} concentration.

Parameters	Description
Zonal wind at 100 m	East-west component of wind at 100 m
Meridional wind at 100 m	North-south component of wind at 100 m
Zonal wind at 10 m	East-west component of wind at 10 m
Meridional wind at 10 m	North-south component of wind at 10 m
Air temperature at 2 m	Near-surface air temperature
Dew point temperature at 2 m	Temperature at which air becomes saturated
Maximum daily air temperature at 2 m	Daily maximum air temperature near the surface
Minimum daily air temperature at 2 m	Daily minimum air temperature near the surface

Continued

Planetary boundary layer height	Height of the atmospheric boundary layer
Turbulent surface stress (east)	Surface stress in the eastward direction
Turbulent surface stress (north)	Surface stress in the northward direction
Frictional velocity	Surface friction velocity
Mean evaporation rate	Average evaporation rate
Mean sea level pressure	Atmospheric pressure at sea level
Surface pressure	Atmospheric pressure at the surface
Mean solar radiation flux downward at the surface	Incoming solar radiation at the surface
Net solar radiation flux at the surface	Net radiation balance at the surface
Surface latent heat flux	Latent heat exchange at the surface
Surface sensible heat flux	Sensible heat exchange at the surface
Soil temperature-surface layer	Soil temperature near the surface
Soil volumetric water content (layer 1)	Soil moisture content in the first layer
Total column water (precipitation)	Integrated water vapour column
Total cloud cover	Fraction of sky covered by clouds
Total precipitation	Precipitation amount
Integrated vertical divergence of humidity	Vertical divergence of moisture flux

2.3. Models, Training, and Evaluation

2.3.1. Models

To achieve the goals of this study, four machine learning models were used (Table 2). Their performances are based on their ability to process linear and non-linear relationships between variables, regression and classification, boosting skills, memory efficiency, etc.

Table 2. Brief description of the different models used in this study.

Models	Description	Output/analyzed parameters
SARIMAX model	Linear relationship between the variable to be predicted and its past values (autocorrelation)	Temporal dynamics of PM_{10}
Random forest model	A multitude of decision trees [51] Handling complex, noisy, and non-linear relationships Classification and regression tasks	Temporal dynamics of PM_{10}
XGBoost model	Boosting technique: chaining together several weak (shallow) decision trees to build a powerful model.	Temporal dynamics of PM_{10}
LightGBM model	Optimized for speed and memory efficiency, enforced particularly by its training strategy based on leaf-wise growth rather than level-wise growth.	Temporal dynamics of PM_{10}

2.3.2. Model Training and cross-Validation Building Sets

For this study, the data were split into two distinct parts. The first part, represent-

ing 80% of the total data length, constitutes a training set. This training set contains 80% of the oldest observations, *i.e.*, starting from January 1, 2017. The second part of the data, representing the remaining 20% of the total data length, lies in the most recent period, *i.e.*, closest to December 31, 2023. This split follows a strict temporal logic in which no random permutation of the data (shuffle) was carried out, which is essential in time series to preserve the coherence of the chronology of events. Respecting the order of the dates avoids any leakage of information from the future to the past at the learning stage, which would distort the model's actual performance under prediction conditions. The choice of the 80/20 ratio is widely recognised as a good compromise between the quantity of data for training and the robustness of the evaluation [52].

- **Model training and cross-validation**

Model training is a decisive stage in the modelling chain. The model training objective is to adjust the internal parameters of the algorithms so that they learn, from the training data, to predict the PM_{10} concentration as a function of the meteorological variables selected. The rigor of this stage directly conditions the quality and robustness of future predictions, as well as the reliability of the assessment. In keeping with the temporal nature of the data, the models were trained solely on the training set (representing 80% of the data), with the remaining 20% being reserved for the test. No random resampling (shuffle) was carried out, in accordance with good practice in time series modelling [53], in order to preserve the chronological order essential for model consistency.

When modelling time series, it is essential to assess the robustness of the models while respecting the temporal structure of the data. Unlike the classic cross-validation method, which assumes the independence of observations and uses random partitions, time series require a specific approach to avoid the biases linked to time dependency. To meet this requirement, the cross-validation method adapted to time series was used, as described by scikit-learn in its official documentation. This method, called TimeSeriesSplit, divides the data into several segments while preserving the chronological order. Each successive segment uses a larger portion of the training data, followed by a test set that is immediately later in time. The use of TimeSeriesSplit provides a more realistic assessment of model performance on unseen data, while avoiding the data leakage problems that could arise with inappropriate cross-validation methods for time series.

The exploratory analysis shows that the year 2023 has 58.63% missing data. Aware of this irregularity, we chose to restrict the study period to the time segment running from 14 January 2017 to 16 February 2023, a period during which data coverage is much more satisfactory. This delimitation made it possible to significantly reduce the overall rate of missing values to 8.99%, a proportion generally considered acceptable for modern imputation methods [54].

- **Statistical evaluation**

Performance evaluation is a crucial stage in the predictive modelling process, as it enables us to judge the accuracy, robustness, and practical usefulness of the

model. Three metrics commonly adopted in time series prediction work were used, namely, the mean absolute error (MAE), the root mean square error (RMSE), and the coefficient of determination (R^2).

MAE (Equation (1)) measures the average absolute deviation between the predicted and observed values. It is simple to interpret and not very sensitive to outliers, but it does not consider the dispersion of errors.

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

where:

- y_i : actual value at time *i*.
- $\hat{y}_i$: predicted value at time *i*.
- n : total number of observations.

RMSE is the square root of the average of the squared deviations between the actual and predicted values. It is shown in Equation (2):

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

This metric gives greater weight to large errors. It is, therefore, particularly useful for detecting models that fail to correctly predict peaks in pollution, a critical aspect of air quality.

Coefficient of determination (R^2)

The coefficient of determination, R^2 given in Equation (3), measures the proportion of the variance in the observed data that the model explains.

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

where $\bar{y}$ is the mean of the actual values. An R^2 value close to 1 indicates that the model explains the variability of PM_{10} concentrations well, while a negative value means that the model performs worse than a simple average.

3. Results and Discussion

3.1. Assessment of ERA5 Data over Lamto

Time series of daily air temperature at 2 m (**Figure 2(a)**), daily minimum (**Figure 2(b)**) and maximum (**Figure 2(c)**) temperature, as well as relative humidity at the synoptic hours 06:00 UTC (**Figure 2(d)**), 12:00 UTC (**Figure 2(e)**), and 18:00 UTC (**Figure 2(f)**) from 14 January 2017 to 16 February 2023 for the ERA5 and *in-situ* are compared. Results show that the diurnal cycle of these selected variables is quite well reproduced in the ERA5 reanalysis. However, ERA5 is underestimating the 2 m daily air temperature as well as the daily maximum temperature. The air temperature at 2 m fluctuates between 21.3˚C and 27.9˚C for ERA5 and between 22.7˚C and 33.8˚C for the *in-situ*. Mean daily air temperature is about 24.5˚C and 28.7˚C for ERA5 and *in-situ* measurements. Daily maximum temperature varies between 21.5˚C and 27.8˚C with a mean of 24.5˚C for ERA5, and

22.0˚C and 41.5˚C with a mean of 34.2˚C for *in-situ* measurements, respectively. When considering the daily minimum temperature, ERA5 is slightly warmer than the *in-situ* measured data at Lamto. ERA5 exhibits maximum temperature ranging from 21.2˚C to 27.5˚C with a mean of 24.3˚C, while the *in-situ* values range from 16 to 34.9˚C. It can be noticed that the extreme events characterized by a sharp increase or decrease in the temperature magnitude are poorly reproduced in the ERA5 data. This indicates a poor performance of using ERA5 to investigate extreme events over the Lamto station.

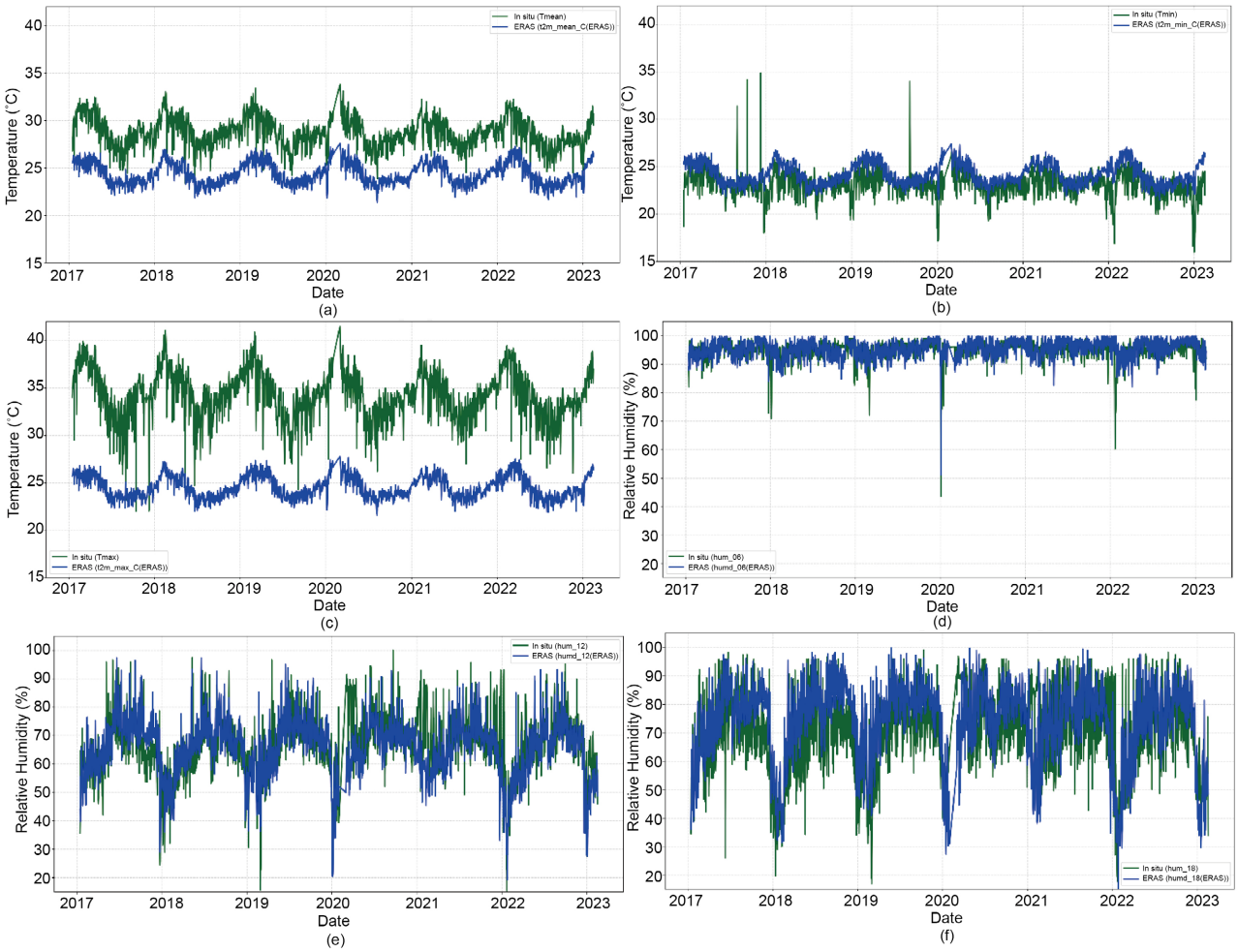

Figure 2. Timeseries comparison of daily air temperature at 2 m (a), daily minimum (b), and maximum (c) temperature, and relative humidity at the synoptic hours 06:00 UTC (d), 12:00 UTC (e), and 18:00 UTC (f) for the ERA5 and *in-situ* data between 14th January 2017 and 16th February 2023.

Analyzing the relative humidity comparison between ERA5 and *in-situ* recorded at synoptic hours, it can be observed that the main features (diurnal cycle, seasonality, etc.) match and the magnitude between the two datasets is in the same range. At 06:00 UTC, it varies from 51% to 100% with a mean of 95% for ERA5 and from 43.6% to 100% with a mean of 95% for the *in-situ* data. At 12:00 UTC, relative humidity fluctuates from 19.2% to 97% with a mean of 65% for ERA5, and

for the *in-situ*, values range from 15.3% to 100% with a mean value of 66.7%. The mean relative humidity from the ERA5 reanalysis at 18:00 UTC is about 73.6% (range of 13.5% to 99.9%), and for the *in-situ*, the mean humidity is about 69.7% (range of 16.9% to 99.1%).

Figure 3 shows the Spearman correlation between the *in-situ* and ERA5 variables. Quite good and strong correlations (r varying from 0.48 to 0.73) are found between the temperature variables, and quite moderate correlations (r varying from 0.22 to 0.65) are found between the relative humidity variables, significant with p-values less than 0.01. These correlations between ERA5 and *in-situ* measured variables were like those reported by Bodjrènou [55] when assessing ERA5 hydrological variables with *in-situ* measurements over Benin. There is also moderate to strong anti-correlation between temperature and relative humidity, indicating that the relative humidity at Lamto station tends to decrease when temperature increases, as warmer air has a greater capacity to hold water vapour than cooler air [56].

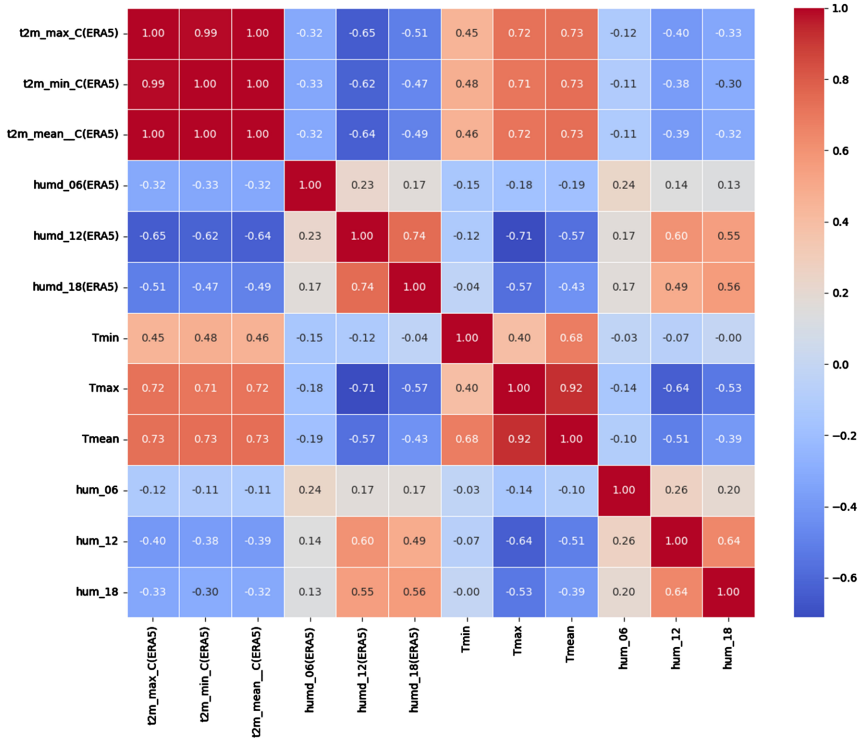

Figure 3. Heat map of the Spearman correlation between the *in-situ* and ERA5 of daily mean air temperature at 2 m, daily minimum and maximum temperature, relative humidity at the synoptic hours 06:00 UTC, 12:00 UTC, and 18:00 UTC between 14th January 2017 and 16th February 2023.

Overall, the ERA5 data capture the diurnal and seasonal cycles of climate and meteorological features of Lamto. As an evaluation of ERA5 data over the Lamto station using *in-situ* observed data did not exist in the literature, it appeared important for us to perform this quick evaluation on the ERA5 data, as it offers the

advantage of a wider range of meteorological variables needed for the use of Artificial Intelligence (AI) tools in this work.

3.2. Exploratory Analysis

Figure 4 shows the time series evolution of daily PM_{10} concentrations at the Lamto station, marked by a strong seasonal cycle, with a peak occurring during the dry season (November to March). Minimum PM_{10} concentrations are recorded during the rainy season (April to October). In contrast to a bimodal rainfall pattern observed at Lamto, the PM_{10} concentration pattern is characterized by a single peak occurring during the dry season. Moreover, the highest PM_{10} concentration peaks were recorded in 2018, with values reaching 888.9 µg/m^3 on 1st January. The year 2021 recorded relatively low PM_{10} concentrations during the dry season compared to the other years. The dry season in Lamto is often impacted by pollutants from both local and long-range sources. During the dry season, Lamto station and surrounding areas are subject to biomass burning fires [57], which are set by different groups of people for different reasons (research purposes, maintaining savanna landscapes, grazing and agriculture, hunting, ...). Moreover, N'Datchoh [58] showed that biomass burning is prevalent over the West African region from November to December, while analyzing satellite fire products from SPOT VEGETATION. In addition to these local and long-range sources from biomass burning activities, this region of West Africa is also reported to be impacted by dust aerosols transported from the Sahel and Saharan region to the Guinean Gulf by the Harmattan winds. During the main rainy season, minimum PM_{10} concentrations are recorded with daily concentration values well below 50 µg/m^3. Furthermore, these PM_{10} daily concentration values maintain their lowest values during the little dry season. This little dry season in the region along the Guinean coast is often due to the WAM jump, which allows the WAM rain belt to be located over the Sahel region [59].

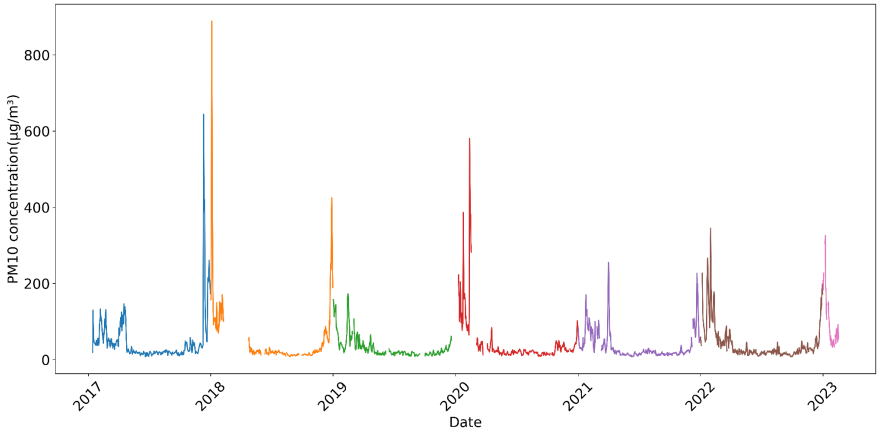

Figure 4. Daily PM_{10} concentration time series, with different colours corresponding to a specific year.

Table 3 summarizes the descriptive statistics for the raw PM_{10} database at

Lamto. A total of 2225 days were analysed. An average of daily PM_{10} concentrations of about 42.79 μg/m^3 is associated with a very high standard deviation of about 59.5 μg/m^3, as well as a median of 22.15 μg/m^3, indicating a high variability and lack of balance in PM_{10} concentrations in Lamto. This type of distribution, asymmetrical and spread out towards the high values, reflects the presence of occasional high pollution events.

Table 3. Daily PM_{10} concentration at Lamto dataset basic statistics overview between 2017 and 2023.

Parameter	Values (μg/m^3)
Mean	42.79
Standard deviation	59.49
Minimum	7.31
First quartile	16.35
Median	22.15
Third quartile	43.55
Maximum	888.90

3.3. Decomposition of the PM_{10} Time Series

Figure 4 presents the original PM_{10} concentration time series (**Figure 5(a)**) and its decomposition into trend (**Figure 5(b)**), seasonality (**Figure 5(c)**), noise (**Figure 5(d)**), and their transformation into a logarithmic function (**Figures 5(e)-(h)**). The logarithmic function was applied to the daily PM_{10} concentration for the purpose of smoothing and attenuating the existing high variability observed in the data, making variation easier to analyse, as suggested by Hyndman and Athanasopoulos [60]. The overall trend (**Figure 5(b)** and **Figure 5(f)**) exhibits phases of high and low daily PM_{10} concentrations, which are the result of complex interactions between environmental, climatic, and meteorological factors as well as human activities. A comparison between the original (**Figure 5(b)**) and logarithmic function (**Figure 5(f)**) shows a smoothing, allowing detection of changes that are crucial in the use of linear models. The seasonal decomposition (**Figure 5(c)** and **Figure 5(g)**) shows the seasonal cycle in the daily PM_{10} concentrations, which are dominated by higher values of daily PM_{10} concentration during the dry season compared to the wet season. Here, the use of the logarithmic transformation allows smoothing of the seasonal cycle while capturing the annual and seasonal cycles. Residual or noise components (**Figure 5(d)** and **Figure 5(h)**) characterize all daily PM_{10} concentration data, which cannot be explained either by trend or seasonality. Since SARIMA needs stability to provide accurate results in terms of its output, the logarithmic transformation allows smoothing of the data and facilitates its machine learning process, which can be affected by

very high values such as extreme events [61]. The presence of structured trends and well-defined seasonality favored the use of classical models such as SARIMAX, which can take explicit account of such trend and seasonality. However, the unpredictable nature of certain variations and the possible non-linear interactions between daily PM_{10} concentration and meteorological variables favoured the consideration of more flexible machine learning models, such as XGBoost, Random Forest, or LightGBM. These latter models are known for their robustness in dealing with noisy and highly variable data and their ability to capture complex relationships [62] between variables.

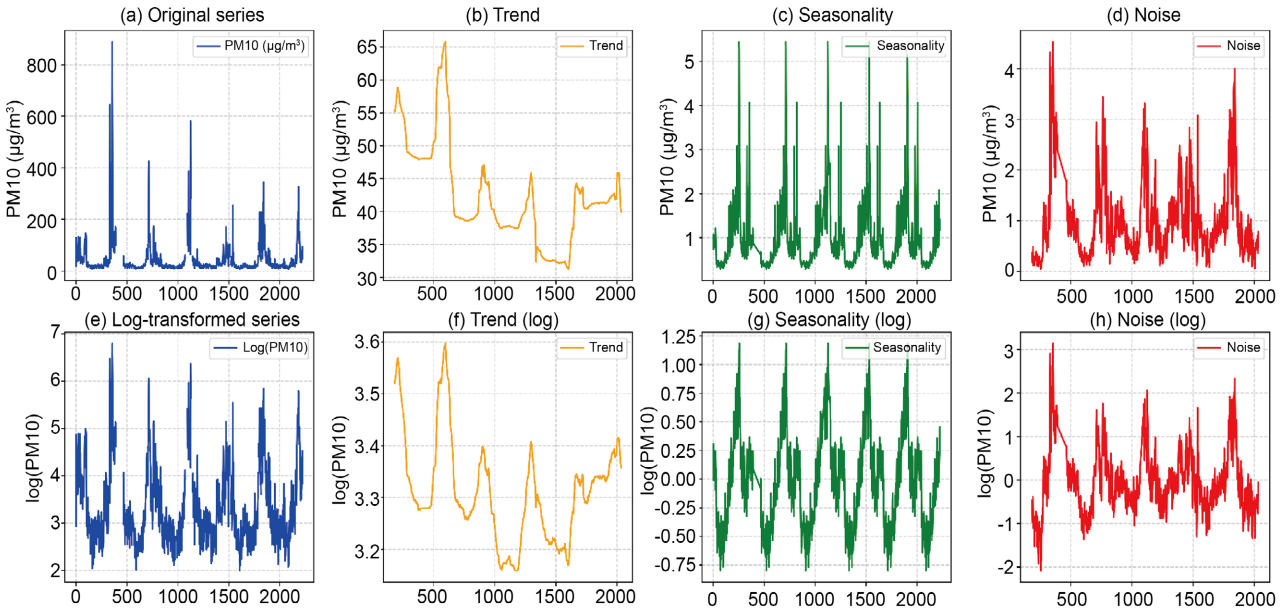

Figure 5. Time series of PM_{10} concentrations and logarithmic transformations (log): Original ((a) and (e)), Trend ((b) and (f)), Seasonality ((c) and (g)), and Noise ((d) and (h)).

A Shapiro-Wilk test recommended for moderate to large samples was conducted in order to check whether the daily PM_{10} follows a normal distribution. The Shapiro-Wilk test results reveal that all the variables considered here deviated significantly from normality, with a very low statistic for PM_{10} ($W = 0.51$) and a p-value close to zero ($\approx 1.12 \times 10^{-59}$) ($W = 0.51$) and a p-value close to zero ($\approx 1.12 \times 10$, implying that classic parametric tests were less appropriate for this work. Therefore, a non-parametric Kruskal-Wallis test was conducted to check monthly and yearly variability significance in the daily PM_{10} dataset. This allows confirmation of the existence of strong monthly variability and seasonality of all variables considered. An ADF (Augmented Dickey-Fuller) stationarity test, which is a must for any time series study [63], was conducted. These test results indicate that the time series are stationary, meaning that their statistical properties (mean, variance) remain constant over time.

Figure 6 presents the seasonal pattern for PM10 concentration climatology between 2017 and 2023.

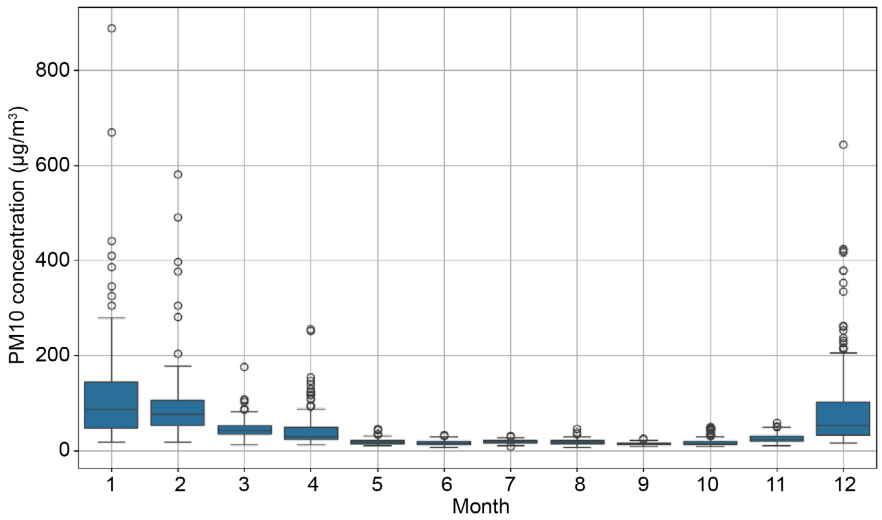

Figure 6. Boxplots showing the seasonal pattern of PM_{10} concentrations from 2017 to 2023.

3.4. Explanatory Variables

The final selection of explanatory variables and their optimal lags is presented in **Table 4**. Time lags of 1 to 3 days [PM_{10}(t-1), PM_{10}(t-2), PM_{10}(t-3)] were included as additional explanatory variables. This decision was motivated by two preliminary analyses. (i) An examination of the autocorrelation (ACF) and partial autocorrelation (PACF) functions revealed significant dependence up to a lag of three days. (ii) A comparative performance evaluation showed that adding lags beyond 3 days lag (d-3) did not significantly improve RMSE and R^2 metrics, while increasing the risk of overfitting. This choice is also consistent with several previous work [64] that have demonstrated the significant contribution of delayed PM_{10} values in predicting future concentrations. To avoid any leakage of information from the future to the past, which would distort the predictions, a breakdown of the time series into two sets was performed: the training data (Train) covers the period from 14 January 2017 to 28 November 2021, *i.e.*, 1780 days, and the test data runs from 29 November 2021 to 16 February 2023, representing 445 days. This breakdown respects the natural temporal logic of the series, a fundamental criterion in the context of forecasting. This breakdown can be seen in **Figure 7**. **Figure 7(a)** shows the separation of the daily PM10 concentration series into train (blue color) and test before filling the missing values. **Figure 7(b)** shows the daily PM_{10} concentration with the missing values filled using the median of similar days in previous and following years to consider the strong seasonality observed in the dataset. For example, the missing values on 10 July 2019 are interpolated using the median for 10 July in 2017, 2018, 2020, etc. This method inspired by the work of Weed [65] makes it possible to retain seasonal effects without introducing time distortion. Although this method is not the most common in the literature, it is based on a simple and solid logic, which takes advantage of seasonal repeatability from one year to the next. It is particularly well suited to environmental data, where seasonal effects are often very marked.

Table 4. Explanatory variables and optimal time lags used in this work.

Variables	Full name of the variables	Time lag
var18_swvl1	Soil water content-surface layer	1 day
var5_mx2t	Maximum temperature at 2 m	1 day
var6_tcwv	Total precipitable water in the air column	1 day
Month	Month of the year	1 day
var14_slhf	Latent surface heat flux	1 day
var11_msl	Atmospheric pressure at mean sea level	1 day
var8_blh	Height of the atmospheric boundary layer	1 day
var3_tcc	Total cloud cover	6 days
var4_d2m_C	Dew point temperature at 2 m	1 day
var7_avg_snswrf	Net short-wave solar flux at the surface	2 days
PM10_lag1	PM_{10} surface concentration of the previous day	1 day
PM10_lag2	PM_{10} surface concentration of the two previous days	2 days
PM10_lag3	PM_{10} surface concentration of the 3 previous days	3 days

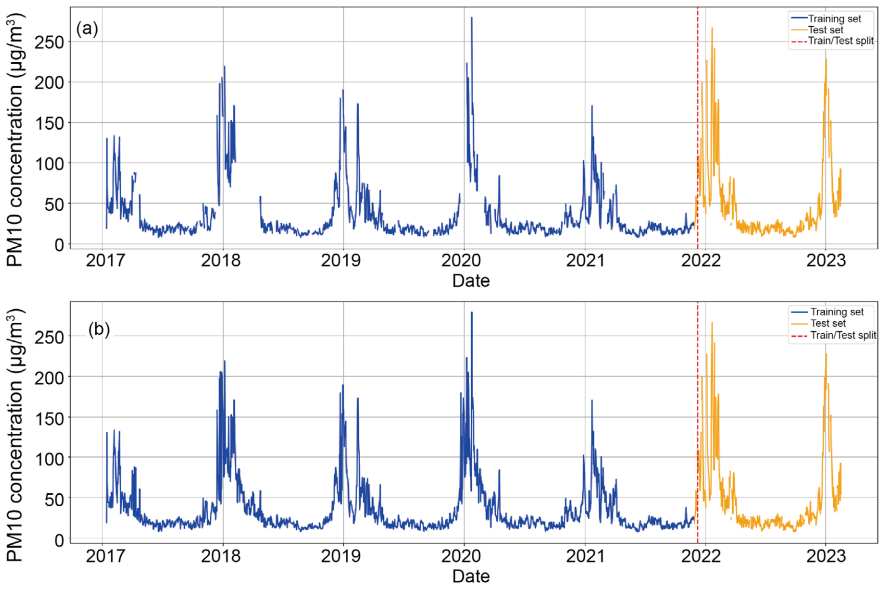

Figure 7. Daily PM_{10} surface concentration breakdown into train from 14th January 2017 to 28th November 2021 (1780 days) and test runs from 29th November 2021 to 16th February 2023 (445 days) for (a) original data and (b) missing data fill-up.

To handle missing values in daily PM_{10} concentrations, we used the median of observations from the same calendar day in other years of the study. This method was selected for its proven ability to preserve the temporal characteristics of time series. However this way of handling the missing data may potentially bias the model training by reducing the natural variance of the data and smoothing out actual pollution peaks. This may lead the model to underestimate the amplitude

of surface PM_{10} concentration peaks during high pollution events. As a result, test errors could be underestimated during validation, as the model would be evaluated on smoothed data that does not fully reflect the diversity and intensity of the observed conditions, which could limit its ability to generalize to new data.

3.5. Separation of Time Series and Handling of Missing Values

To assess the reliability and ability of the Machine Learning models to generalize well the predictions of PM_{10} concentrations, a cross-validation was done. The model was trained on past data to make predictions about a future period, without ever using information from the future. Three models were used, namely Random Forest, XGBoost, and LightGBM. For each model, performance was measured using the root mean square error (RMSE). Mean errors obtained are summarized in **Table 3**, highlighting that the Random Forest model performs better, with an average RMSE of 15.94. It is closely followed by LightGBM (16.44), then XGBoost (17.03). These scores indicate that Random Forest seems to better capture the dynamics of the daily PM_{10} concentration, while remaining stable across the different periods tested. However, the SARIMAX model was not included in this cross-validation. The SARIMAX model relies heavily on the temporal dependence between observations, and therefore was evaluated separately, using a simple chronological breakdown between training and testing.

3.6. Final Evaluation and Model Performance

Figure 8 shows the original (blue) and predicted time series modelled (orange) by daily PM_{10} surface concentrations from January 2022 to March 2023 for SARIMAX (**Figure 8(a)**), Random Forest (**Figure 8(b)**), XGBoost (**Figure 8(c)**), and LightGBM (**Figure 8(d)**). The hyperparameters used for each model are presented in the supplementary material (**Table S1**). It should be noted, that no performance optimisation tests were carried out on in this present work. We used default configurations for each of the models. Overall, all four models were able to capture the general trend of the PM_{10} surface concentration in Lamto. However, a close analysis of Figure 8 reveals some slight differences between the models themselves and between the models and the *in-situ* observations. For example, the PM_{10} concentrations range from 10.19 (minimum) to 210.01 (maximum) $\mu g/m^3$ for the SARIMAX model, 10.48 (minimum) to 171.88 (maximum) $\mu g/m^3$ for the Random Forest model, 12.47 (minimum) to 190.37 (maximum) $\mu g/m^3$ for the XGBoost model, and 11.68 (minimum) to 171.20 (maximum) $\mu g/m^3$ for the LightGBM model, while the observations are within 8.00 (minimum) and 266.57 (maximum) $\mu g/m^3$. All the models tend to overestimate the PM_{10} minimum concentration, while they underestimate the maximum compared to the observations. Moreover, all models tend to slightly overestimate PM_{10} concentrations compared to observations during the wet season. Furthermore, in terms of capturing the high variability in PM_{10} concentration peaks associated with pollution events, some slight lags seem to be produced in all the models except the Random Forest, which seems

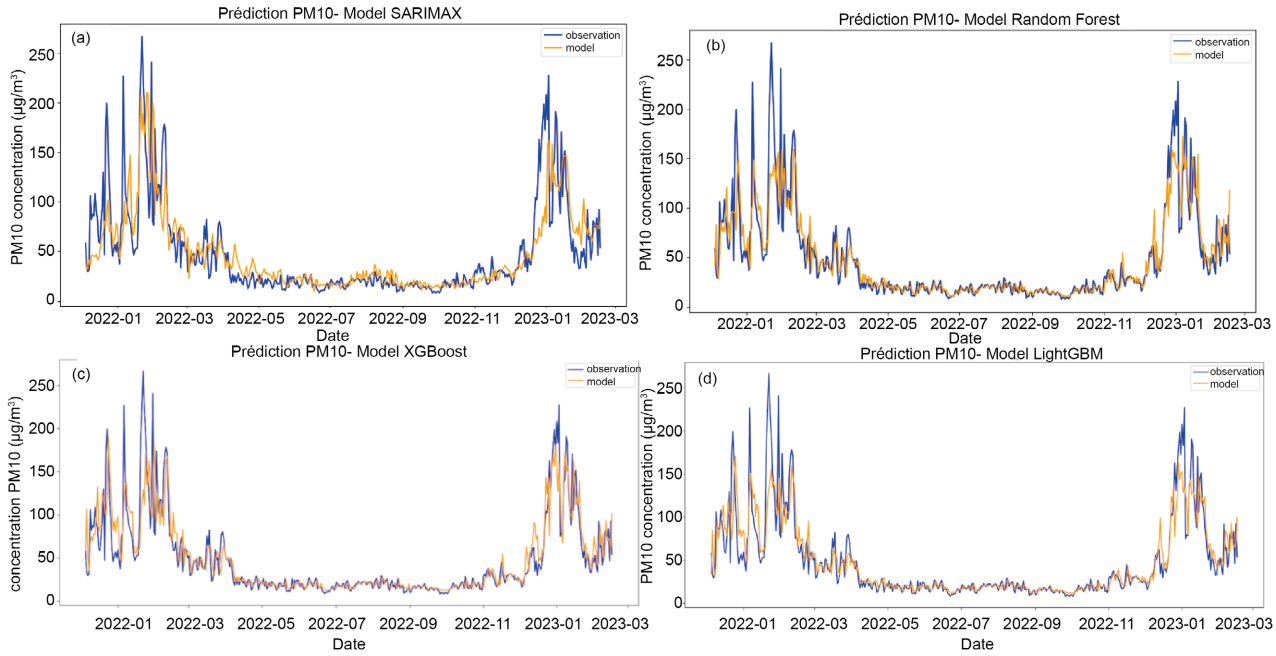

Figure 8. Original (blue) and predicted (orange) time series simulated by the (a) SARIMAX, (b) Random Forest, (c) XGBoost, and (d) LightGBM models of daily PM_{10} surface concentrations from January 2022 to March 2023.

Table 5. Statistical indicators RMSE, MAE, and R^2 provided at annual and seasonal time scales for each of the four models used in this study.

	Random forest	XGBoost	LightGBM	SARIMAX
RMSE (µg/m³)	23.12	24.31	23.70	28.52
MAE (µg/m³)	12.35	13.23	12.78	16.52
R^2	0.78	0.75	0.76	0.64
Main dry season RMSE (µg/m³)	33.98	35.41	35.66	39.93
Main dry season MAE (µg/m³)	22.38	23.28	24.08	29.84
Main dry season R^2	0.61	0.58	0.57	0.46
Main wet season RMSE (µg/m³)	7.79	8.11	7.98	16.31
Main wet season MAE (µg/m³)	5.20	5.46	5.36	12.99
Main wet season R^2	0.73	0.71	0.72	−0.18
Little dry season RMSE (µg/m³)	4.82	4.81	5.13	17.97
Little dry season MAE (µg/m³)	4.02	3.96	4.38	15.67
Little dry season R^2	0.13	0.13	0.01	−2.81
Little wet season RMSE (µg/m³)	3.58	3.68	3.72	8.47
Little wet season MAE (µg/m³)	2.65	2.87	2.90	6.94
Little wet season R^2	0.32	0.28	0.26	−11.11

to follow more closely. Therefore, it appears necessary to evaluate the performances of the different models that are measured using indicators such as RMSE, MAE, and R^2. These indicators are summarized in Table 5 for the entire predicted timeseries but also per season. From Table 5, it can be seen that the SARIMAX model is the least performant model with a limited capacity to capture changes in the daily PM_{10} concentrations at Lamto, as it scored lower performance, with a RMSE of 28.52 μg/m^3, a MAE of 16.52 μg/m^3, and a R^2 of 0.64. These relatively poor scores underline that the present models (Random Forest, XGBoost, LightGBM, and SARIMAX) have poor capacity in reproducing daily PM_{10} surface concentration, which may be explained by its high variance and the complex interaction between meteorology, local, and long-range factors that influence the PM_{10} loaded in the Lamto atmosphere. PM_{10} concentrations in Lamto are influenced by complex non-linear phenomena as well as irregular exogenous factors, which are difficult to model within a purely statistical framework.

Conversely, the Machine Learning models show a potential possibility to be adapted and to integrate complex atmospheric processes involving local and long-range factors under the condition of providing good and robust observations, which may be used to train and test them. Therefore, the Random Forest model scored a better overall performance with an RMSE of 23.12 μg/m^3, an MAE of 12.35 μg/m^3, and an R^2 of 0.78. Furthermore, an RMSE of 23.12 indicates an average prediction error of 23.12 μg/m^3 compared with the actual values, while an MAE of 12.35 underlines that the average absolute difference between the predicted and observed values is 12.35 μg/m^3. The coefficient of determination R^2 = 0.78 indicates that 78% of the variability in the daily PM_{10} surface concentrations may be explained by the model, which is satisfactory but also suggests margins of error. Therefore, in general, the pattern of the daily PM_{10} concentration is better reproduced by the random forest, despite some bias. For example, the predicted PM_{10} values during the dry season (November to March) are less important than those observed. These biases are associated with specific events of PM_{10} pollution, which are more frequent during the dry season. The relatively low daily PM_{10} surface concentration during the wet season is quite well reproduced in both pattern and magnitude, highlighting the ability of machine learning to better assimilate periods characterized by weak changes.

Variables permutation importance for the Random Forest, LightGBM, and XGBoost are shown in Figure 9. This post-modeling analysis confirms the relevance of preliminary selection of variables based on mutual information and cross-correlation. The results of that PM10_lag1 predominance highlights a temporal self-dependence as the predominant mechanism in predicting PM_{10} surface concentration in Lamto. Also, the PM10_lag2 and PM10_lag3 delays confirm this multi-day persistence, may be related to the dry season where rainfall is less frequent, less washout of atmosphere occurring and more pollutant particularly PM_{10} holds into atmosphere. When considering, meteorological variables, boundary layer height (var8_blh_lag1), humidity (var4_d2m_C_lag1),

temperature (var5_mx2t_lag1), and latent heat flux (var14_slhf_lag1) emerge as consistent secondary modulators, influencing vertical dispersion, hygroscopic particle growth, convection, and turbulence, respectively. Finally, the Month_lag1 variable robustly captures the seasonality documented in Lamto. This multi-model analysis validates both the robustness of the predictors and the complementarity of our variable selection approaches. Finally, because of the SARIMAX model's structure, variable permutation is not possible; therefore, permutation importance could not be computed and was excluded from this analysis.

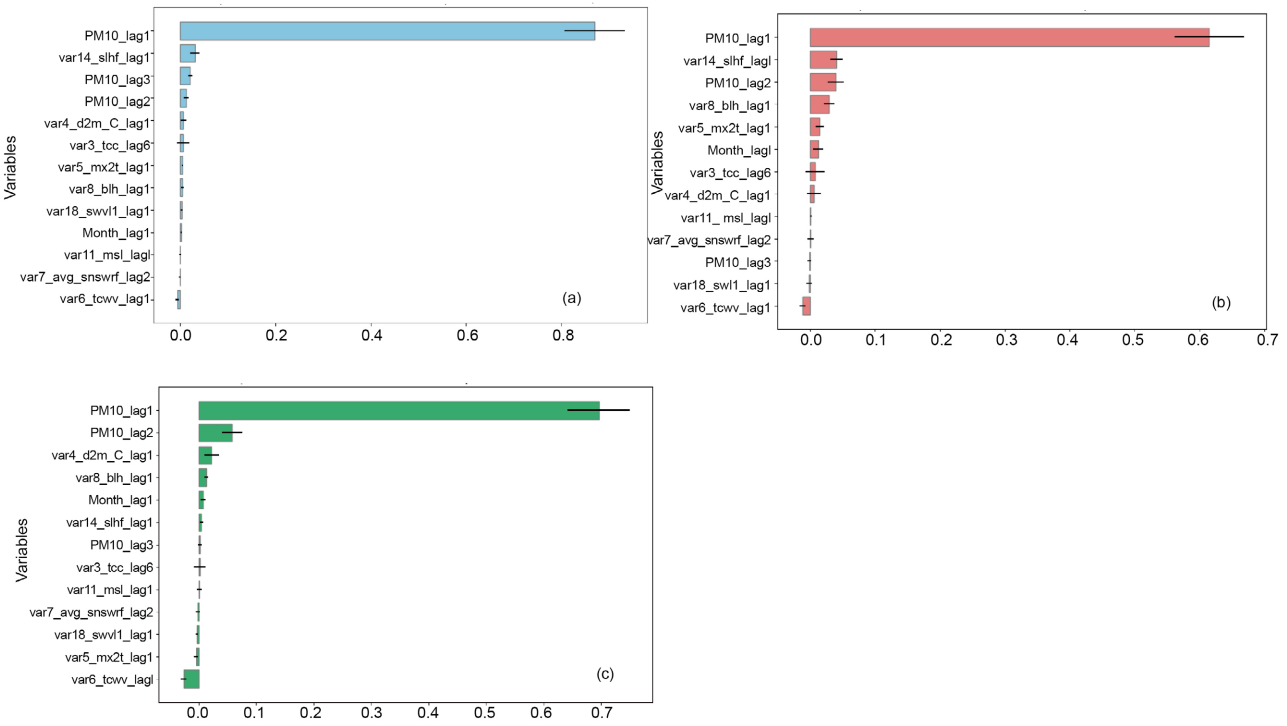

Figure 9. Variables permutation importance for (a) Random Forest; (b) XGBoost; and (c) LghtGBM.

3.7. Discussion

This work focused on the use of Machine Learning (ML) tools for predicting PM_{10} concentrations. To achieve this goal, the methodology applied consisted of four ML models, namely SARIMAX, Random Forest, XGBoost, and LightGBM models, to train them over 80% of the total filtered PM_{10} data length. The remaining 20% of the observed PM_{10} data was used for testing the model outputs. The Random Forest was identified as the best-performing model among these four models, with estimated biases of 23.12 $\mu g/m^3$. These observed biases between original and predicted daily PM_{10} surface concentration at Lamto may be explained, on the one hand, by the complex mixture of sources influencing its chemical composition [22] [26] [40]. Lamto chemical composition is the result of a combination of local (biomass burning, domestic fires, biogenic emission, ...) and remote (long-range transport of dust from the Northern Hemisphere of Africa during the dry season,

biomass burning, sea salt, ...). The dry season months are characterized by high variability in pollution events, as all local and long-range transport of Sahel and Saharan dust transportation toward the Guinean Gulf [66] are also prevalent. On the other hand, biases observed between ERA5 and the *in-situ* measurement may have contributed to all models underestimating pollutant events maxima. A combination of *in-situ* measurement (when these are available) and ERA5 data for the machine learning models may contribute to reducing these observed maxima underestimations observed in all models used. Thus, this low performance of the model during the dry season may be related to the high variability in daily PM_{10}, which contributes directly to the increase in model error and cold biases observed with the ERA5 meteorological temperature variables. During the wet season, the washout of the atmosphere by rain contributes to fewer high air pollution events, thus PM_{10} loads into the atmosphere. This led to the higher scores observed by all models, especially the Random Forest. This finding highlights the current difficulties and limitations of applying logical tools such as mathematics through modelling to complex environmental processes involving meteorology and climate, which are highly impacted by human socio-economic activities such as air pollution. Arowosegbe [67] used Random Forest models to impute missing daily PM_{10} concentrations in South Africa between 2010 and 2017 using spatio-temporal predictors including meteorological variables, land cover, elevation, and proximity to roads. They found an average R^2 of 0.78 at the national level, 0.70 at the provincial level, and 0.55 at the level of each site, thus demonstrating the effectiveness of the Random Forest model in an African context despite the scarcity of data. The temporal component of the national models explained up to 78% of the variability in concentrations. Moreover, Adnane [68] in Morocco used Nonlinear Autoregressive Neural Networks with Exogenous Inputs (NARX) models to predict hourly PM_{10} concentrations in Agadir, at 1 h and 24 h horizons. The best results were obtained with the 24h horizon model, combining meteorological variables and secondary pollutants (CO, SO_2, O_3) as inputs. The model achieved a correlation coefficient of 93.75%, *i.e.*, an approximate R^2 of 0.88, illustrating the added value of pollution data combined with meteorological variables in improving predictive performance.

As hybrid models have been reported to have better performance for environmental monitoring policy and decision-making [35] [36], such methods need to be evaluated in the limited data context of sub-Saharan Africa. As AI is emerging as a necessary transversal tool in many research areas, the African continent must integrate such a tool in environmental process monitoring and management. Applying AI to their AirQo low-cost sensors in Africa, Bainomugisha [69] highlighted a set of digital solutions for the environmental air pollution challenges from custom-designed low-cost air quality monitors, deployment methodology, AI-powered digital tools, and a framework for citizens' and leaders' engagement. Finally, leveraging AI and data science to mitigate the respiratory health impacts in the context of climate change in Africa, Sowunmi [70] emphasized pilot AI-

driven health monitoring systems in major cities with clear indicators, such as improved weather forecasts, reductions in respiratory symptoms or hospitalizations, and improved healthcare access.

4. Conclusions

Daily PM_{10} surface concentrations in Lamto over seven years (2017-2023) were investigated using a machine learning tool. To achieve this, firstly, the ERA5 daily data were assessed with *in-situ* measurements recorded at the Lamto station. This assessment shows that, despite some biases, ERA5 can reproduce a fairly diurnal cycle of temperature and relative humidity for Lamto. Good and strong correlations (r varying from 0.48 to 0.73) are found between the ERA5 and *in-situ* measurements of temperature variables. Also, moderate correlations (0.22 to 0.65) exist between the relative humidity variables of these two datasets.

Using 25 meteorological variables from daily ERA5 and PM_{10} surface concentration, we applied machine learning to 80% (14th January 2017 to 28th November 2021, *i.e.*, 1780 days) of the daily PM_{10} surface concentration to train four machine learning models (SARIMAX, Random Forest model, XGBoost, and LightGBM), while 20% (from 1st January 2022 to March 2023, *i.e.*, 445 days) were used to validate their predictions. All the models reasonably reproduced trends and patterns of PM_{10} daily concentration compared to the *in-situ* observations, despite some biases. These biases consisted of a slight overestimation of the daily PM_{10} concentration during the wet season, while underestimating high pollution days during the dry season. The performances of these four machine learning models were evaluated using indicators such as RMSE, MAE, and R^2. The Random Forest model achieved an R^2 of 0.78 and an RMSE of 23 μg/m^3, surpassing the performance of the SARIMAX model (R^2 of 0.64 and an RMSE of 28.54 μg/m^3), XGBOOST (R^2 of 0.74 and an RMSE of 24.54 μg/m^3), and LightGBM (R^2 of 0.76 and an RMSE of 23.70 μg/m^3). This model's performance helped identify the Random Forest as the best performing model in reproducing the daily PM_{10} surface concentration at Lamto. This performance was better during the wet season, when daily PM_{10} presents less variability with low pollution levels. Despite the high variability associated with pollution events depending on local and remote factors, the pattern was captured despite a lower high pollution magnitude. Moreover, the low performance of the machine learning models during the dry season may be related to this high variability in daily PM_{10} and biases observed with ERA5 meteorological data, though we did not investigate the error sources and their impacts on the work. Future work will focus on the contribution of bias correction into ERA5 as well as the combination of models to improve the machine learning model performance in capturing environmental processes over Lamto and other places of the country and the continent.

Acknowledgements

The authors would like to acknowledge the INDAAF (International Network to

study Deposition and Atmospheric chemistry in Africa) project. INDAAF is supported by the INSU (Institut National des Sciences de l'Univers)/CNRS (Centre National de la Recherche Scientifique), IRD (Institut de Recherche pour le Développement) and the research infrastructure ACTRIS-FR registered on the Roadmap of the French Ministry of Research. We also express our gratitude to all INDAAF local technicians for their maintenance and sampling work. We would also like to thank the JEAI PATI (Jeune Equipe Associée de l'IRD Physico-chimie Atmosphérique et Impacts) project under the sponsorship of the IRD for their generous financial support in the publication of this article.

Data Availability

The INDAAF data can be retrieved from the website https://indaaf.obs-mip.fr/. Interested persons may contact corinne.galy-lacaux@aero.obs-mip.fr or beatrice.marticorena@lisa.ipsl.fr for these data.

Author Contributions

Conceptualization and methodology, N.E.T., M.G.O., and C.E.; Analysis experience, N.E.T., M.G.O., C.E., and A.B.; writing—original draft preparation, N.E.T., M.G.O., A.B., C.E., K.K., F.Y., M.D., M.L.D., G.N.S.K., and V.Y.; writing—review and editing, N.E.T., M.G.O., A.B., C.E., K.K., F.Y., M.D., M.L.D., G.N.S.K. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] IPCC (2001) Climate Change 2001: The Scientific Basis. Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, 881.

[2] IPCC (2014) Climate Change 2014: Mitigation of Climate Change. Contribution of Working Group III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press.

[3] Lohmann, U. and Feichter, J. (2001) Can the Direct and Semi-Direct Aerosol Effect Compete with the Indirect Effect on a Global Scale? *Geophysical Research Letters*, **28**, 159-161. https://doi.org/10.1029/2000gl012051

[4] Nabat, P., Somot, S., Mallet, M., Michou, M., Sevault, F., Driouech, F., *et al.* (2015) Dust Aerosol Radiative Effects during Summer 2012 Simulated with a Coupled Regional Aerosol-Atmosphere-Ocean Model over the Mediterranean. *Atmospheric Chemistry and Physics*, **15**, 3303-3326. https://doi.org/10.5194/acp-15-3303-2015

[5] Seinfeld, J.H. and Pandis, S.N. (1998) From Air Pollution to Climate Change. *Atmospheric Chemistry and Physics*, 1326.

[6] Knippertz, P., Coe, H., Chiu, J.C., Evans, M.J., Fink, A.H., Kalthoff, N., *et al.* (2015) The DACCIWA Project: Dynamics-Aerosol-Chemistry-Cloud Interactions in West Africa. *Bulletin of the American Meteorological Society*, **96**, 1451-1460. https://doi.org/10.1175/bams-d-14-00108.1

[7] Fiore, A.M., Naik, V. and Leibensperger, E.M. (2015) Air Quality and Climate Connections. *Journal of the Air & Waste Management Association*, **65**, 645-685. https://doi.org/10.1080/10962247.2015.1040526

[8] Nairobi, 2022 United Nations Environment Programme. https://www.ccacoalition.org/sites/default/files/resources/files/Summary%20for%20Decision%20Makers_Integrated%20Assessment%20of%20Air%20Pollution%20and%20Climate%20Change%20for%20Sustainable%20Development%20in%20Africa.pdf

[9] De Longueville, F., Hountondji, Y., Henry, S. and Ozer, P. (2010) What Do We Know about Effects of Desert Dust on Air Quality and Human Health in West Africa Compared to Other Regions? *Science of the Total Environment*, **409**, 1-8. https://doi.org/10.1016/j.scitotenv.2010.09.025

[10] Mir Alvarez, C., Hourcade, R., Lefebvre, B. and Pilot, E. (2020) A Scoping Review on Air Quality Monitoring, Policy and Health in West African Cities. *International Journal of Environmental Research and Public Health*, **17**, Article 9151. https://doi.org/10.3390/ijerph17239151

[11] Bahino, J., Giordano, M., Beekmann, M., Yoboué, V., Ochou, A., Galy-Lacaux, C., *et al.* (2024) Temporal Variability and Regional Influences of $PM_{2.5}$ in the West African Cities of Abidjan (Côte d'Ivoire) and Accra (Ghana). *Environmental Science: Atmospheres*, **4**, 468-487. https://doi.org/10.1039/d4ea00012a

[12] Djossou, J., Léon, J., Akpo, A.B., Liousse, C., Yoboué, V., Bedou, M., *et al.* (2018) Mass Concentration, Optical Depth and Carbon Composition of Particulate Matter in the Major Southern West African Cities of Cotonou (Benin) and Abidjan (Côte d'Ivoire). *Atmospheric Chemistry and Physics*, **18**, 6275-6291. https://doi.org/10.5194/acp-18-6275-2018

[13] Anand, A., Touré, N.E., Bahino, J., Gnamien, S., Hughes, A.F., Arku, R.E., *et al.* (2024) Low-Cost Hourly Ambient Black Carbon Measurements at Multiple Cities in Africa. *Environmental Science & Technology*, **58**, 12575-12584. https://doi.org/10.1021/acs.est.4c02297

[14] Gnamien, S., Liousse, C., Keita, S., Silué, S., Bahino, J., Gardrat, E., *et al.* (2023) Chemical Characterization of Urban Aerosols in Abidjan and Korhogo (Côte d'Ivoire) from 2018 to 2020 and the Identification of Their Potential Emission Sources. *Environmental Science: Atmospheres*, **3**, 1741-1757. https://doi.org/10.1039/d3ea00131h

[15] Silué, S., Kouassi, A.A., N'datchôh, E.T. and Yoboué, V. (2021) Meteorological Factors Contribution over Two Urban Sites in Côte d'Ivoire. *RAMReS Science des Structure de la Matiere*, **3**, 12-39.

[16] Léon, J., Akpo, A.B., Bedou, M., Djossou, J., Bodjrenou, M., Yoboué, V., *et al.* (2021) $Pm_{2.5}$ Surface Concentrations in Southern West African Urban Areas Based on Sun Photometer and Satellite Observations. *Atmospheric Chemistry and Physics*, **21**, 1815-1834. https://doi.org/10.5194/acp-21-1815-2021

[17] Gnamien, S., Liousse, C., Keita, S., Kumar, R. and Yoboué, V. (2024) High-resolution Modeling of Air Quality in Abidjan (Côte d'Ivoire) Using a New Urban-Scale Inventory. *Atmosphere*, **15**, Article 758. https://doi.org/10.3390/atmos15070758

[18] de Coëtlogon, G., Deroubaix, A., Flamant, C., Menut, L. and Gaetani, M. (2023) Impact of the Guinea Coast Upwelling on Atmospheric Dynamics, Precipitation and Pollutant Transport over Southern West Africa. *Atmospheric Chemistry and Physics*, **23**, 15507-15521. https://doi.org/10.5194/acp-23-15507-2023

[19] Flamant, C., Knippertz, P., Fink, A.H., Akpo, A., Brooks, B., Chiu, C.J., *et al.* (2018) The Dynamics-Aerosol-Chemistry-Cloud Interactions in West Africa Field Cam-

paign: Overview and Research Highlights. *Bulletin of the American Meteorological Society*, **99**, 83-104. https://doi.org/10.1175/bams-d-16-0256.1

[20] Deroubaix, A., Menut, L., Flamant, C., Brito, J., Denjean, C., Dreiling, V., *et al.* (2019) Diurnal Cycle of Coastal Anthropogenic Pollutant Transport over Southern West Africa during the DACCIWA Campaign. *Atmospheric Chemistry and Physics*, **19**, 473-497. https://doi.org/10.5194/acp-19-473-2019

[21] Niamien, A.F., Léon, J., Adon, M., Rajot, J., Feron, A. and Yoboué, V. (2024) Variability of Aerosol Optical Depth and Altitude for Key Aerosol Types over Southern West Africa via CALIPSO/CALIOP Observations. *Atmosphere*, **15**, Article 396. https://doi.org/10.3390/atmos15040396

[22] Ossohou, M., Galy-Lacaux, C., Yoboué, V., Hickman, J.E., Gardrat, E., Adon, M., *et al.* (2019) Trends and Seasonal Variability of Atmospheric NO_2 and HNO_3 Concentrations across Three Major African Biomes Inferred from Long-Term Series of Ground-Based and Satellite Measurements. *Atmospheric Environment*, **207**, 148-166. https://doi.org/10.1016/j.atmosenv.2019.03.027

[23] Adon, M., Galy-Lacaux, C., Yoboué, V., Delon, C., Lacaux, J.P., Castera, P., *et al.* (2010) Long Term Measurements of Sulfur Dioxide, Nitrogen Dioxide, Ammonia, Nitric Acid and Ozone in Africa Using Passive Samplers. *Atmospheric Chemistry and Physics*, **10**, 7467-7487. https://doi.org/10.5194/acp-10-7467-2010

[24] Galy-Lacaux, C., Laouali, D., Descroix, L., Gobron, N. and Liousse, C. (2009) Long Term Precipitation Chemistry and Wet Deposition in a Remote Dry Savanna Site in Africa (Niger). *Atmospheric Chemistry and Physics*, **9**, 1579-1595. https://doi.org/10.5194/acp-9-1579-2009

[25] Galy-Lacaux, C. and Delon, C. (2014) Nitrogen Emission and Deposition Budget in West and Central Africa. *Environmental Research Letters*, **9**, Article ID: 125002. https://doi.org/10.1088/1748-9326/9/12/125002

[26] Yoboué, V., Galy-Lacaux, C., Lacaux, J.P. and Silué, S. (2005) Rainwater Chemistry and Wet Deposition over the Wet Savanna Ecosystem of Lamto (Côte d'Ivoire). *Journal of Atmospheric Chemistry*, **52**, 117-141. https://doi.org/10.1007/s10874-005-0281-z

[27] World Health Organization (2021) WHO Global Air Quality Guidelines: Particulate Matter ($PM_{2.5}$ and PM_{10}), Ozone, Nitrogen Dioxide, Sulfur Dioxide and Carbon Monoxide. World Health Organization.

[28] United Nations Environment Programme (2023) Integrated Assessment of Air Pollution and Climate Change for Sustainable Development in Africa.

[29] Lelieveld, J., Evans, J.S., Fnais, M., Giannadaki, D. and Pozzer, A. (2015) The Contribution of Outdoor Air Pollution Sources to Premature Mortality on a Global Scale. *Nature*, **525**, 367-371. https://doi.org/10.1038/nature15371

[30] Burnett, R., Chen, H., Szyszkowicz, M., Fann, N., Hubbell, B., Pope, C.A., *et al.* (2018) Global Estimates of Mortality Associated with Long-Term Exposure to Outdoor Fine Particulate Matter. *Proceedings of the National Academy of Sciences of the United States of America*, **115**, 9592-9597. https://doi.org/10.1073/pnas.1803222115

[31] Garbagna, L., Babu Saheer, L. and Maktab Dar Oghaz, M. (2025) AI-Driven Approaches for Air Pollution Modelling: A Comprehensive Systematic Review. *Environmental Pollution*, **373**, Article ID: 125937. https://doi.org/10.1016/j.envpol.2025.125937

[32] Olawade, D.B., Wada, O.Z., Ige, A.O., Egbewole, B.I., Olojo, A. and Oladapo, B.I. (2024) Artificial Intelligence in Environmental Monitoring: Advancements, Challenges, and Future Directions. *Hygiene and Environmental Health Advances*, **12**, Ar-

ticle ID: 100114. https://doi.org/10.1016/j.heha.2024.100114

[33] Zheng, T., Bergin, M., Wang, G. and Carlson, D. (2021) Local $PM_{2.5}$ Hotspot Detector at 300 M Resolution: A Random Forest-Convolutional Neural Network Joint Model Jointly Trained on Satellite Images and Meteorology. *Remote Sensing*, **13**, Article 1356. https://doi.org/10.3390/rs13071356

[34] Dobrea, M., Badicu, A., Barbu, M., Subea, O., Balanescu, M., Suciu, G., *et al.* (2020) Machine Learning Algorithms for Air Pollutants Forecasting. 2020 *IEEE 26th International Symposium for Design and Technology in Electronic Packaging* (*SIITME*), Pitesti, 21-24 October 2020, 109-113. https://doi.org/10.1109/siitme50350.2020.9292238

[35] Subramaniam, S., Raju, N., Ganesan, A., Rajavel, N., Chenniappan, M., Prakash, C., *et al.* (2022) Artificial Intelligence Technologies for Forecasting Air Pollution and Human Health: A Narrative Review. *Sustainability*, **14**, Article 9951. https://doi.org/10.3390/su14169951

[36] Fu, L., Li, J. and Chen, Y. (2023) An Innovative Decision Making Method for Air Quality Monitoring Based on Big Data-Assisted Artificial Intelligence Technique. *Journal of Innovation & Knowledge*, **8**, Article ID: 100294. https://doi.org/10.1016/j.jik.2022.100294

[37] Le Roux, X., Abbadie, L., Lensi, R. and Serça, D. (1995) Emission of Nitrogen Monoxide from African Tropical Ecosystems: Control of Emission by Soil Characteristics in Humid and Dry Savannas of West Africa. *Journal of Geophysical Research*: *Atmospheres*, **100**, 23133-23142. https://doi.org/10.1029/95jd01923

[38] Le Roux, X., Gauthier, H., Bégué, A. and Sinoquet, H. (1997) Radiation Absorption and Use by Humid Savanna Grassland: Assessment Using Remote Sensing and Modelling. *Agricultural and Forest Meteorology*, **85**, 117-132. https://doi.org/10.1016/s0168-1923(97)00002-6

[39] Abbadie, L., Gignoux, J., Roux, X. and Lepage, M. (2006) Lamto: Structure, Functioning, and Dynamics of a Savanna Ecosystem (Vol. 179). Springer Science and Business Media.

[40] Ossohou, M., Galy-Lacaux, C., Yoboué, V., Adon, M., Delon, C., Gardrat, E., *et al.* (2020) Long-term Atmospheric Inorganic Nitrogen Deposition in West African Savanna over 16 Year Period (Lamto, Côte d'Ivoire). *Environmental Research Letters*, **16**, Article ID: 015004. https://doi.org/10.1088/1748-9326/abd065

[41] N'Dri, A.B., Gignoux, J., Dembele, A. and Konate, S. (2012) Short Term Effects of Fire Intensity and Fire Regime on Vegetation Dynamic in a Tropical Humid Savanna (Lamto, Central Côte d'Ivoire). *Natural Science*, **4**, 1056-1064. https://doi.org/10.4236/ns.2012.412134

[42] N Dri, A.B. and Konan, L.N. (2018) Does the Date of Burning Affect Carbon and Nutrient Losses in a Humid Savanna of West Africa? *Environment and Natural Resources Research*, **8**, 102-116. https://doi.org/10.5539/enrr.v8n3p102

[43] Tiemoko, D.T., Yoroba, F., Diawara, A., Kouadio, K., Kouassi, B.K. and Yapo, A.L.M. (2020) Understanding the Local Carbon Fluxes Variations and Their Relationship to Climate Conditions in a Sub-Humid Savannah-Ecosystem during 2008-2015: Case of Lamto in Cote d'Ivoire. *Atmospheric and Climate Sciences*, **10**, 186-205. https://doi.org/10.4236/acs.2020.102010

[44] Tiemoko, D.T., Yoroba, F., Paris, J., Diawara, A., Berchet, A., Pison, I., *et al.* (2020) Source-Receptor Relationships and Cluster Analysis of CO_2, CH_4, and CO Concentrations in West Africa: The Case of Lamto in Côte d'Ivoire. *Atmosphere*, **11**, Article 903. https://doi.org/10.3390/atmos11090903

[45] Bouet, C., Yoboué, V., Marticorena, B., Rajot, J.L., Allègre, M., Féron, A., Gaimoz, C., Maisonneuve, F., Siour, G., Valorso, R., Ki, A.F., Konaté, I. and Ouattara, A. (2021) PM10 Concentration, Lamto, Côte d'Ivoire. https://doi.org/10.25326/279

[46] Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., *et al.* (2021) ERA5-Land: A State-of-the-Art Global Reanalysis Dataset for Land Applications. *Earth System Science Data*, **13**, 4349-4383. https://doi.org/10.5194/essd-13-4349-2021

[47] Seibert, P. (2000) Review and Intercomparison of Operational Methods for the Determination of the Mixing Height. *Atmospheric Environment*, **34**, 1001-1027. https://doi.org/10.1016/s1352-2310(99)00349-0

[48] Petäjä, T., Järvi, L., Kerminen, V., Ding, A.J., Sun, J.N., Nie, W., *et al.* (2016) Enhanced Air Pollution via Aerosol-Boundary Layer Feedback in China. *Scientific Reports*, **6**, Article No. 18998. https://doi.org/10.1038/srep18998

[49] Stull, R.B. (2015) Practical Meteorology: An Algebra-Based Survey of Atmospheric Science. University of British Columbia Press.

[50] Volná, V. and Hladký, D. (2020) Detailed Assessment of the Effects of Meteorological Conditions on PM10 Concentrations in the Northeastern Part of the Czech Republic. *Atmosphere*, **11**, Article 497. https://doi.org/10.3390/atmos11050497

[51] Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-32. https://doi.org/10.1023/a:1010933404324

[52] Géron, A. (2022) Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly Media, Inc.

[53] Bergmeir, C., Hyndman, R.J. and Koo, B. (2018) A Note on the Validity of Cross-Validation for Evaluating Autoregressive Time Series Prediction. *Computational Statistics & Data Analysis*, **120**, 70-83. https://doi.org/10.1016/j.csda.2017.11.003

[54] Schafer, J.L. and Graham, J.W. (2002) Missing Data: Our View of the State of the Art. *Psychological Methods*, **7**, 147-177. https://doi.org/10.1037/1082-989x.7.2.147

[55] Bodjrènou, R., Sintondji, L.O., N'Tcha, Y.M.P., Germain, D., Azonwade, F.E., Sohindji, F., *et al.* (2024) Assessment of Hydrologic Data Estimates from ERA5 Reanalyses in Benin, West Africa. *Geoscience Data Journal*, **12**, e288. https://doi.org/10.1002/gdj3.288

[56] Humphreys, W.J. (1929) Physics of the Air. McGraw-Hill Book Company.

[57] Soro, T.D., Koné, M., N'Dri, A.B. and N'Datchoh, E.T. (2021) Identified Main Fire Hotspots and Seasons in Côte d'Ivoire (West Africa) Using MODIS Fire Data. *South African Journal of Science*, **117**, Article No. 7659. https://doi.org/10.17159/sajs.2021/7659

[58] N'Datchoh, E.T., Konaré, A., Diedhiou, A., Diawara, A., Quansah, E. and Assamoi, P. (2015) Effects of Climate Variability on Savannah Fire Regimes in West Africa. *Earth System Dynamics*, **6**, 161-174. https://doi.org/10.5194/esd-6-161-2015

[59] Sultan, B. and Janicot, S. (2003) The West African Monsoon Dynamics. Part II: The "Preonset" and "Onset" of the Summer Monsoon. *Journal of Climate*, **16**, 3407-3427. https://doi.org/10.1175/1520-0442(2003)016<3407:twamdp>2.0.co;2

[60] Hyndman, R.J. and Athanasopoulos, G. (2018) Forecasting: Principles and Practice. OTexts.

[61] Shumway, R.H. and Stoffer, D.S. (2017) ARIMA Models. In: Shumway, R.H. and Stoffer, D.S., Eds., *Time Series Analysis and Its Applications*, Springer, 75-163. https://doi.org/10.1007/978-3-319-52452-8_3

[62] Chen, T. and Guestrin, C. (2016) XGBoost: A Scalable Tree Boosting System. *Proceed-*

ings of the 22*nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, 13-17 August 2016, 785-794. https://doi.org/10.1145/2939672.2939785

[63] Dickey, D.A. and Fuller, W.A. (1979) Distribution of the Estimators for Autoregressive Time Series with a Unit Root. *Journal of the American Statistical Association*, **74**, 427-431. https://doi.org/10.1080/01621459.1979.10482531

[64] Taşpınar, F. (2015) Improving Artificial Neural Network Model Predictions of Daily Average PM_{10} Concentrations by Applying Principle Component Analysis and Implementing Seasonal Models. *Journal of the Air & Waste Management Association*, **65**, 800-809. https://doi.org/10.1080/10962247.2015.1019652

[65] Weed, L., Lok, R., Chawra, D. and Zeitzer, J. (2022) The Impact of Missing Data and Imputation Methods on the Analysis of 24-Hour Activity Patterns. *Clocks & Sleep*, **4**, 497-507.

[66] Touré, N.E., Konaré, A. and Silué, S. (2012) Intercontinental Transport and Climatic Impact of Saharan and Sahelian Dust. *Advances in Meteorology*, **2012**, Article ID: 157020. https://doi.org/10.1155/2012/157020

[67] Arowosegbe, O.O., Röösli, M., Künzli, N., Saucy, A., Adebayo-Ojo, T.C., Jeebhay, M.F., *et al.* (2021) Comparing Methods to Impute Missing Daily Ground-Level PM_{10} Concentrations between 2010-2017 in South Africa. *International Journal of Environmental Research and Public Health*, **18**, Article 3374. https://doi.org/10.3390/ijerph18073374

[68] Adnane, A., Leghrib, R., Chaoufi, J. and Chirmata, A. (2022) Prediction of PM10 Concentrations in the City of Agadir (Morocco) Using Non-Linear Autoregressive Artificial Neural Networks with Exogenous Inputs (NARX). *Materials Today*: *Proceedings*, **52**, 146-151. https://doi.org/10.1016/j.matpr.2021.11.340

[69] Bainomugisha, E., Adrine Warigo, P., Busigu Daka, F., Nshimye, A., Birungi, M. and Okure, D. (2024) AI-Driven Environmental Sensor Networks and Digital Platforms for Urban Air Pollution Monitoring and Modelling. *Societal Impacts*, **3**, Article ID: 100044. https://doi.org/10.1016/j.socimp.2024.100044

[70] Sowunmi, A.O., Eze, O.I., Osadolor, U., Iseolorunkanmi, A. and Adeloye, D. (2024) Leveraging AI and Data Science to Mitigate the Respiratory Health Impacts of Climate Change in Africa: Organisation, Costs, and Challenges. *Journal of Global Health*, **14**, Article ID: 03051. https://doi.org/10.7189/jogh.14.03051

Supplement Material

Table S1. Hyperparameter information used for each of the model use in this study. Note that, nosystematic investigation for hyperparameters (using grid or cross-validation) was performed, as we focused in this study on basic comparison of models with their default configuration.

Model	Hyperparameter	Final value or setting	Reference / comment
Random Forest	n_estimators	100	Number of trees in the forest
	max_depth	10	Maximum depth of each tree
	random_state	42	Reproducibility
XGBoost	n_estimators	100	Number of boosting rounds
	learning_rate	0.1	Learning rate
	max_depth	3	Maximum depth of each tree
	random_state	42	Random seed for reproducibility
LightGBM	objective	regression	Regression task
	metric	rmse	Evaluation function
	boosting_type	gbdt	Type of boosting algorithm
	num_boost_round	100	Number of learning iterations
	early_stopping_rounds	10	Early stopping if no improvement
SARIMAX	order (p, d, q)	(1, 0, 2)	Autoregressive and moving average components
	seasonal_order (P, D, Q, s)	(2, 0, 2, 12)	Seasonal components (s = 12 for monthly annual cycle)
	enforce_stationarity	False	No forcing of stationarity
	enforce_invertibility	False	No forcing of invertibility