

Machine-learning prediction of dust aerosol optical depth at 550 nm over West Africa

Peace Awoleye

peace.awoleye@up.ac.za

University of Pretoria <https://orcid.org/0000-0002-4430-5776>

Babatunde Abiodun

University of Cape Town

Manfred Wendisch

Universität Leipzig

Julius Akinyoola

University of Cape Town

Ayodeji Oluleye

Federal University of Technology Akure

Research Article

Keywords:

Posted Date: August 13th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10674265/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Machine-learning prediction of dust aerosol optical depth at 550 nm over West Africa

Peace Awoleye*^{1,5}, Babatunde Abiodun^{2,3}, Manfred Wendisch⁴, Julius Akinyoola^{2,3},
Ayodeji Oluleye¹

¹Department of Meteorology and Climate Science, Federal University of Technology, Akure,
34001, Nigeria

²Climate System and Analysis Group, Department of Environmental and Geographical
Sciences, University of Cape Town, Rondebosch 7700, Cape Town, South Africa

³Nansen-Tutu Research Centre for Marine Environment, Department of Oceanography,
University of Cape Town, Rondebosch 7700, South Africa

⁴Institute of Meteorology, Faculty of Physics and Earth Sciences, Universität Leipzig,
Leipzig, Stephanstraße 3, 04103, Germany

⁵Department of Geography, Geoinformatics and Meteorology, University of Pretoria,
Lynwood, 0002, South Africa

Correspondence to: Peace Awoleye (peace.awoleye@up.ac.za)

Abstract

Dust aerosol optical depth (AOD) over West Africa is controlled by interacting meteorological, land-surface and seasonal factors. We have developed a machine-learning framework to predict daily dust AOD at 550 nm wavelength over West Africa from 2003 to 2024 using meteorological and land-surface reanalysis variables. Linear Regression, Random Forest and LightGBM were compared using a time-based split for training, validation, and independent test periods. LightGBM gave the strongest validation performance and was therefore selected for independent testing. The model reproduced the north-south dust-AOD gradient and the seasonal structure, although errors remained larger during March-May and June-August. Probability density and cumulative distribution analyses showed that LightGBM captured much of the low-to-moderate dust AOD range and predicted fewer high-dust events. At the 90th percentile, the model detected 41% of high-dust events and produced a frequency bias of 0.76. SHapley Additive exPlanations (SHAP) analysis identified temperature, wind speed, relative humidity and soil moisture as the leading predictors. The spatial self-organising map further showed that the low-to-moderate dust patterns were better represented than the high dust patterns. The results define both the capability and the remaining limitations of the dust AOD predicting framework over West Africa.

1 Introduction

Mineral dust is a major aerosol component over West Africa, where Saharan and Sahelian drylands provide important source regions for regional aerosol loading and transport toward the tropical Atlantic (Prospero et al., 2002; Knippertz and Todd, 2012). Dust particles influence radiation, surface radiative energy balance, visibility and air quality, but the regional impact depends strongly on day-to-day variability in emission, transport, vertical mixing and removal. Dust AOD is therefore important because it represents the column-integrated optical effect of mineral dust, with relevance for solar-energy production, satellite retrieval accuracy, visibility, air-quality assessment and environmental planning. High-dust conditions are favoured in cases of erodible surfaces, dry soils, strong near-surface winds and atmospheric boundary-layer (ABL) conditions. They result in high values of dust-AOD that are substantially higher than background conditions (De Longueville et al., 2010; Schepanski et al., 2009; Solmon et al., 2012; Ogunjobi and Awoleye, 2019; Saidou Chaibou et al., 2020; Awoleye et al., 2025a).

Aerosol reanalyses combine modelled emissions, transport, removal and AOD assimilation, but performance varies as a function of aerosol type, region and season. Copernicus

Atmosphere Monitoring Service (CAMs) derived dust AOD must therefore be considered as a reanalysis target (Benedetti et al., 2009; Randles et al., 2017; Buchard et al., 2017; Inness et al., 2019; Xian et al., 2024; Tuna Tuygun and Elbir, 2024; Paiva et al., 2025). The European Centre for Medium-Range Weather Forecasts Reanalysis v5 (ERA5) provide near-surface data relevant to dust mobilisation, transport and persistence, while ERA5/ERA5-Land soil-moisture information supplies land-surface parameters on dust emission (Hersbach et al., 2020; Muñoz-Sabater et al., 2021). Climate Hazards Group InfraRed Precipitation with Station data (CHIRPS) delivers long-term satellite–gauge precipitation estimates suitable for hydroclimatic analysis in data-sparse regions, including West Africa (Funk et al., 2015; Dembélé and Zwart, 2016; Atiah et al., 2020; Houngnibo et al., 2023). Therefore, these datasets enable the study to assess how much daily CAMs dust-AOD variability can be reconstructed from meteorological and land-surface predictors.

Continuous ground-based dust observations remain sparse across much of West Africa, making it difficult to assess daily dust variability using ground measurements alone. Previous machine-learning applications in aerosol and air-quality research have shown that nonlinear models can represent complex and spatially heterogeneous relationships between aerosols and predictors (Just et al., 2018; Zhong et al., 2021; Liang et al., 2022; Tian et al., 2023; Gilbert et al., 2025). However, these studies do not directly address the specific problem examined here: daily CAMs dust AOD reanalysis over West Africa using combined meteorological and land surface predictors. This study, therefore, combines bulk performance metrics with distributional diagnostics, event-based verification, SHAP attribution and SOM-based spatial-pattern classification. The pattern-classification component follows previous applications of self-organising maps and weather-pattern methods for diagnosing atmospheric states, model behaviour and extreme-event conditions (Cassano et al., 2006; Nigro et al., 2011; Gibson et al., 2017; Sheridan and Lee, 2011).

This study develops a machine-learning framework for daily CAMs reanalysis dust AOD prediction over West Africa using ERA5 atmospheric and land-surface variables, CHIRPS precipitation and lagged predictors. The results of different machine learning approaches, such as Linear Regression as a linear baseline, Random Forest as a bagged decision-tree ensemble, and LightGBM as a gradient-boosted tree model, are compared. The study shows how well daily CAMs dust AOD can be reconstructed from meteorological and land-surface information; where and when prediction errors are largest; and which predictors and recurring dust patterns explain those errors.

2 Data and Methods

2.1 Study area and datasets

This study was conducted over West Africa, covering 20° W - 20° E and 0° - 20° N (**Fig. 1**). The domain includes the southern Saharan margin, the Sahelian dust belt, and the major west-east and north-south dust transport corridors that influence dust-AOD variability across the region.

The analysis covered the daily period from 2003 to 2024, and the target variable was CAMs reanalysis dust aerosol optical depth at 550 nm. The 550 nm wavelength was used because it is a standard visible-wavelength reference for aerosol optical depth in many satellites and reanalysis aerosol products, including CAMs. It provides a consistent measure of the column-integrated optical effect of dust and supports comparison with previous aerosol studies (Awoleye et al. 2025b).

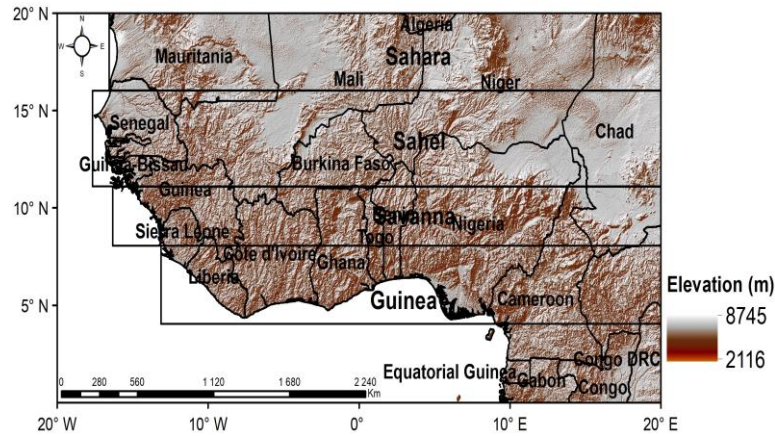


Figure 1: Topographic map of the West African study domain, adapted from Awoleye et al., (2025b), based on the GMTED2010 surface digital elevation model.

2.2 Data preprocessing and predictor construction

Daily dust AOD from the Copernicus Atmosphere Monitoring Service (CAMS) was used as the model target. CAMS provides global atmospheric-composition reanalysis products and has been widely used for aerosol, trace-gas, model-evaluation and climatological studies (Inness et al., 2019).

Meteorological and land-surface predictors were obtained from ERA5, the fifth-generation ECMWF atmospheric reanalysis (Hersbach et al., 2020). The ERA5 variables were 10 m above-ground zonal wind, 10 m meridional wind, 2 m temperature, 2 m dew-point temperature, volumetric soil water in the upper soil layer, and boundary-layer height. These variables were selected because dust mobilisation, atmospheric mixing, transport, and removal are influenced by wind speed, soil moisture, humidity, surface heating and ABL development. The inclusion of wind speed and soil moisture follows the physical basis of dust-emission theory, where emission depends on aerodynamic forcing and surface erodibility (Marticorena and Bergametti, 1995; Shao, 2001).

Daily precipitation data were obtained from CHIRPS, the Climate Hazards Group InfraRed Precipitation with Station data. CHIRPS combines satellite infrared observations with station data to provide long-term gridded precipitation estimates suitable for hydroclimatic analysis (Funk et al., 2015). Precipitation was included because rainfall can suppress dust through wet deposition, enhanced soil moisture and reduced availability of erodible surface material.

Table 1: Description of datasets and variables used from 2003 to 2024 over West Africa.

Dataset group	Variables	Source	Spatial resolution	Temporal resolution
Dust aerosol optical depth	Dust AOD at 550 nm	CAMS	$0.75^\circ \times 0.75^\circ$	3-hourly
Near-surface meteorology	2m temperature, dew-point temperature, u and v wind components	ERA5	$0.25^\circ \times 0.25^\circ$	Hourly
Boundary-layer condition	Boundary-layer height	ERA5	$0.25^\circ \times 0.25^\circ$	Hourly
Land-surface moisture	Volumetric soil water	ERA5	$0.25^\circ \times 0.25^\circ$	Hourly
Rainfall	Precipitation	CHIRPS	$0.05^\circ \times 0.05^\circ$	Daily

All datasets described in **Table 1** were restricted to the common period from 1 January 2003 to 31 December 2024. Time coordinates were harmonised to daily timestamps, and latitude–longitude coordinate names were standardised across datasets. The CAMS dust-AOD grid was used as the reference grid. ERA5 variables were interpolated onto the CAMS grid using linear interpolation, whereas CHIRPS precipitation was interpolated using nearest-neighbour interpolation to preserve the closest rainfall estimate without introducing additional spatial smoothing.

After temporal and spatial harmonisation, all variables were combined into one daily gridded dataset containing the CAMS dust-AOD target and the meteorological, land-surface and rainfall predictors. Rows with missing values in either the predictor variables or the target variable were excluded from the final machine-learning table.

Near-surface wind speed was calculated from the ERA5 10m wind components as;

$$WS = \sqrt{u^2 + v^2} \quad (1)$$

where WS is wind speed, (u) is the zonal wind component, and (v) is the meridional wind component at 10 m above ground.

Relative humidity was derived from temperature and dew-point temperature, both taken from 2 m above ground using the Magnus formula (Alduchov and Eskridge, 1996):

$$RH = 100 \times \frac{\exp\left(\frac{17.625T_d}{243.04+T_d}\right)}{\exp\left(\frac{17.625T}{243.04+T}\right)} \quad (2)$$

where RH is relative humidity, T is 2m temperature in $^{\circ}C$, and T_d is 2m dew point temperature in $^{\circ}C$. Values were constrained to the physical range of 0-100%.

Lagged predictors were included to capture short-term memory in rainfall, soil moisture, wind speed and ABL conditions. The one-day rainfall lag was defined as:

$$P_{lag1}(t) = P(t - 1) \quad (3)$$

The three-day antecedent rainfall accumulation was calculated correspondingly:

$$P_{3day}(t) = \sum_{k=1}^3 P(t - k) \quad (4)$$

The one-day lagged soil moisture, wind speed and boundary-layer height were calculated as:

$$x_{lag1}(t) = x(t - 1) \quad (5)$$

where P is rainfall, x is soil moisture, wind speed, and boundary-layer height.

The final predictor set consisted of zonal and meridional wind, wind speed, relative humidity, temperature, soil moisture, boundary-layer height, rainfall, one-day rainfall lag, three-day antecedent rainfall, one-day soil-moisture lag, one-day wind-speed lag and one-day boundary-layer-height lag.

2.3 Machine-learning table and models

The harmonised gridded dataset was converted into a tabular machine-learning format in which each row represented one daily grid-cell observation. The target variable for each observation was CAMS reanalysis dust AOD:

$$y_i = AOD_{dust,i} \quad (6)$$

The predictor vector for the same observation was defined as:

$$X_i = (x_{i1}, x_{i2}, x_{i3}, \dots, x_{ip}) \quad (7)$$

where y_i is the CAMS dust AOD value for observation i , X_i is the corresponding predictor vector, and p is the total number of predictors.

A time-based split was used to avoid temporal leakage. The data were divided into three independent periods: 2003-2018 for training, 2019-2020 for evaluation, and 2021-2024 for final testing. Candidate models were trained on 2003-2018 and compared using the 2019-2020 evaluation period. It is important to note that the evaluation period and the independent test period serve different purposes. The 2019-2020 evaluation data were used to compare Linear Regression, Random Forest and LightGBM and to select the best-performing model. Because that period influenced model choice, its performance cannot be treated as the final unbiased estimate of model accuracy. After LightGBM was selected, it was retrained using all available pre-test data from 2003-2020 so that the final model could benefit from an additional two years of information. The retrained model was then evaluated once on the 2021-2024 data, which had not been used for fitting, tuning or model selection. The independent test assesses how the selected and refitted LightGBM model performs on years that were not used for model fitting, tuning or selection. Any reduction from the 2019-2020 evaluation performance to the 2021-2024 independent test performance reflects differences in the unseen test period and does not mean that refitting reduced model skill.

Linear regression was used as a baseline model:

$$\hat{y}_i = \beta_0 + \sum_{j=1}^p \beta_j x_{ij} \quad (8)$$

where $\hat{y}_i$ is the predicted dust AOD, β_0 is the intercept, β_j is the coefficient for the predictor j , and x_{ij} is the value of the predictor j for CAMS reanalysis i .

Random forest regression is a nonlinear ensemble benchmark because it averages predictions from multiple decision-tree regressors trained on subsamples of the dataset (Breiman, 2001). The random forest prediction was expressed as:

$$\hat{y}_i = \frac{1}{B} \sum_{b=1}^B f_b(X_i) \quad (9)$$

where B is the number of trees and $f_b(X_i)$ is the prediction from tree b .

LightGBM is an efficient gradient-boosted decision-tree algorithm designed for large datasets and high-dimensional predictor spaces (Ke et al., 2017). The boosted prediction was represented as:

$$\hat{y}_i = F_M(X_i) = \sum_{m=1}^M \eta f_m(X_i) \quad (10)$$

where F_M is the final boosted model after M trees, f_m is the m -th regression tree, and η is the learning rate.

After model evaluation and selection, the final LightGBM model was re-trained using the combined 2003 - 2020 training and validation periods and independently tested over 2021 - 2024.

Table 2: Hyperparameter configuration of the nonlinear models used for daily CAMS dust-AOD prediction.

Parameter	Random Forest	LightGBM
Objective/Loss function	Squared-error regression	Regression
Number of trees/estimators	100	2000
Learning rate	Not applicable	0.03
Number of leaves	Not applicable	63
Maximum depth	20	-1
Minimum leaf/child samples	5	50
Subsample	Bootstrap sampling	0.8
Column subsample	Not specified	0.8
Random state	42	42
Early stopping	Not used	100 rounds using RMSE
Best iteration	Not applicable	Determined from RMSE

Linear Regression is not included in **Table 2** because it was used as an ordinary least-squares baseline and therefore did not require tree-based or boosting hyperparameters. For LightGBM, 2000 represents the maximum number of boosting iterations allowed; the effective number of iterations was determined by early stopping on the validation set.

2.4 Model evaluation and distributional characterisation

Model performance was evaluated using the coefficient of determination (r^2), root-mean-square error (RMSE), mean absolute error (MAE), mean bias (MB), and the Pearson correlation coefficient (r). These metrics quantify explained variance, linear association, average error magnitude and systematic bias between CAMS reanalysis dust AOD and model predictions.

Spatial and seasonal performance was assessed by mapping annual and seasonal means of CAMS dust AOD, LightGBM-predicted dust AOD and prediction error. For each grid cell, the prediction error was calculated as:

$$Error = \hat{y}_t - y_t \quad (11)$$

where y_t is CAMS dust AOD at time t , $\hat{y}_t$ is predicted dust AOD. Positive errors indicate overprediction relative to CAMS, while negative errors indicate underprediction. Seasonal averages were calculated for DJF, MAM, JJA, and SON.

The agreement between the CAMS and LightGBM dust-AOD distributions was evaluated using probability density functions (PDFs) and cumulative distribution functions (CDFs). The PDF and CDF diagnostics were used to assess whether the model reproduced both the central distribution and the high distribution.

2.5 High-dust event analysis

High-dust events were defined using the CAMS reanalysis dust-AOD distribution during the independent 2021–2024 test period. The high-dust threshold was defined as the 90th percentile of CAMS dust AOD:

$$T_{90} = Q_{0.90}(y) \quad (12)$$

where T_{90} is the CAMS 90th percentile dust-AOD threshold and y is the CAMS dust-AOD distribution during the test period. The total numbers of hits, misses, and false alarms were produced to examine where the model captured or failed to capture high-dust.

2.6 Feature importance and SHAP analysis

LightGBM predictor importance was first assessed using gain-based importance, which measures the total reduction in the model objective associated with splits involving each predictor. Because gain importance ranks predictors without showing the direction of their influence, SHAP analysis was used to examine predictor contributions to individual LightGBM predictions (Lundberg and Lee, 2017).

For each prediction, SHAP represents the model output as the sum of a baseline value and predictor-specific contributions:

$$f(X_i) = \phi_0 + \sum_{j=1}^p \phi_{ij} \quad (13)$$

where $f(X_i)$ is the LightGBM prediction for observation i , ϕ_0 is the expected model output, and ϕ_{ij} is the SHAP contribution of predictor j . Positive SHAP values increase predicted dust AOD relative to the baseline, while negative SHAP values reduce it. TreeSHAP was used because it is designed for efficient SHAP estimation in tree-based models, including gradient-boosted decision trees (Lundberg et al., 2020). SHAP summary and dependence plots were used to interpret the influence of the dominant predictors.

2.7 Self-organising map analysis

We first used a 3×3 SOM to group the 12 seasonal mean bias maps produced by Linear Regression, Random Forest and LightGBM for DJF, MAM, JJA and SON according to similarities in their spatial error patterns. A second, 4×4 SOM was trained using daily CAMS dust AOD from 2003-2024 to identify the recurring spatial distributions of dust over West Africa. Both analyses were used for pattern classification and model evaluation, not for prediction. Each daily CAMS dust-AOD field was reshaped into a one-dimensional spatial vector:

$$X_t = (y_{t,1}, y_{t,2}, \dots, y_{t,g}) \quad (14)$$

where X_t is the daily dust AOD spatial pattern on day t , and g is the number of grid cells. The grid-cell time series were standardised before SOM training so that the classification reflected spatial-pattern structure rather than being dominated by the largest-magnitude pixels.

For the daily analysis, each of the 8,036 CAMS dust-AOD maps was reshaped into a spatial vector using the common valid grid cells and transformed using the same mean and standard deviation. The resulting vectors were used to train the 4×4 SOM, producing 16 nodes numbered 1–16. For each day, the best-matching node was identified as the node whose weight vector had the smallest Euclidean distance from the corresponding daily dust-AOD vector:

$$BMU(t) = \arg \min_k \|X_t - W_k\| \quad (15)$$

where W_k is the weight vector of the SOM node k . Each day was assigned to the regime corresponding to its best-matching unit.

For each SOM regime, the mean CAMS dust AOD composite was calculated as:

$$\bar{Y}_r = \frac{1}{N_r} \sum_{t \in r} Y_t \quad (16)$$

where $\bar{Y}_r$ is the mean dust AOD map for the regime r . Y_t is the CAMS dust AOD map on day t , and N_r is the number of days assigned to state (r).

The daily CAMS, Linear Regression, Random Forest and LightGBM dust-AOD distributions for the 2019-2020 validation period were then assigned to the fixed CAMS-derived nodes using the same transformation applied during SOM training. The percentage of days assigned to each node was calculated for CAMS and the three models. These percentages were compared to assess whether the models reproduced the frequency and range of the recurring CAMS patterns, including the less frequent high-AOD patterns (Kohonen, 1990, 2001; Hewitson and Crane, 2002).

3. Results and Discussion

3.1 CAMS dust AOD climatology over West Africa

The monthly climatology shows that dust loading is concentrated within the Saharan-Sahelian zone and changes markedly with the seasonal circulation (Fig. 2). This distribution shows the location of the principal North African dust sources with the combined influence of the Harmattan, Saharan heat low and West African monsoon circulation on dust emission and transport (Prospero et al., 2002; Knippertz and Todd, 2012; Senghor et al., 2017; Schepanski et al., 2017).

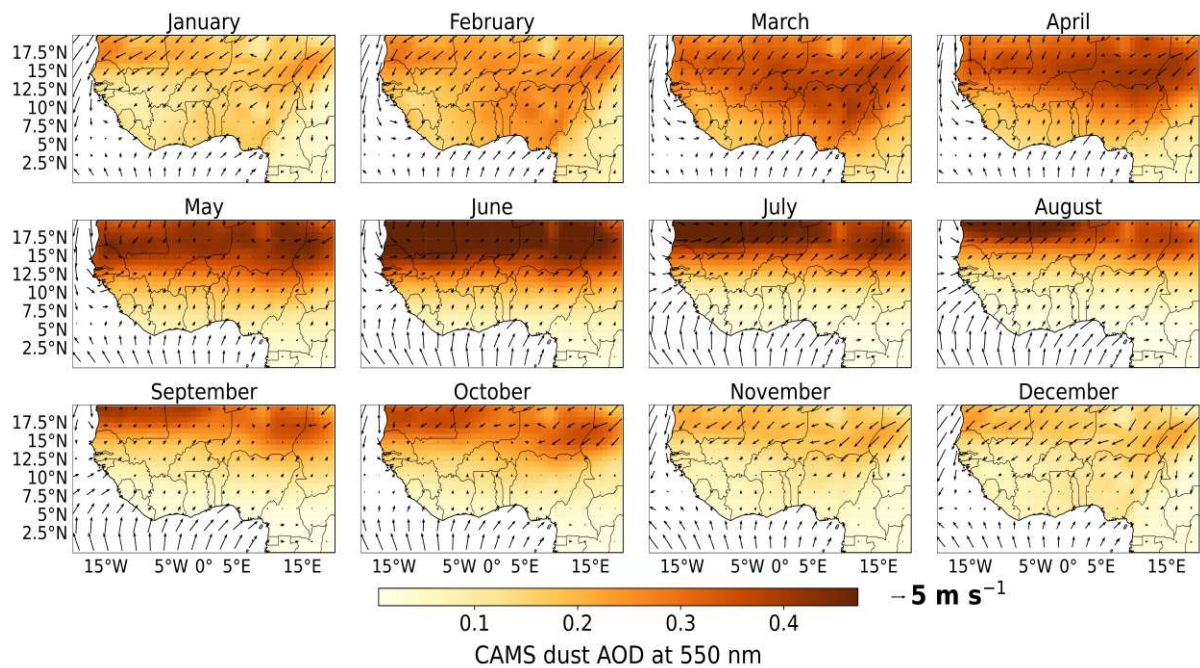


Figure 2: Monthly spatial climatology of CAMS reanalysis dust AOD at 550 nm and ERA5 near-surface winds over West Africa during 2003–2024. Shading represents the long-term monthly mean dust AOD, while vectors show the corresponding 10-m horizontal wind field. The reference vector represents 5 m s⁻¹.

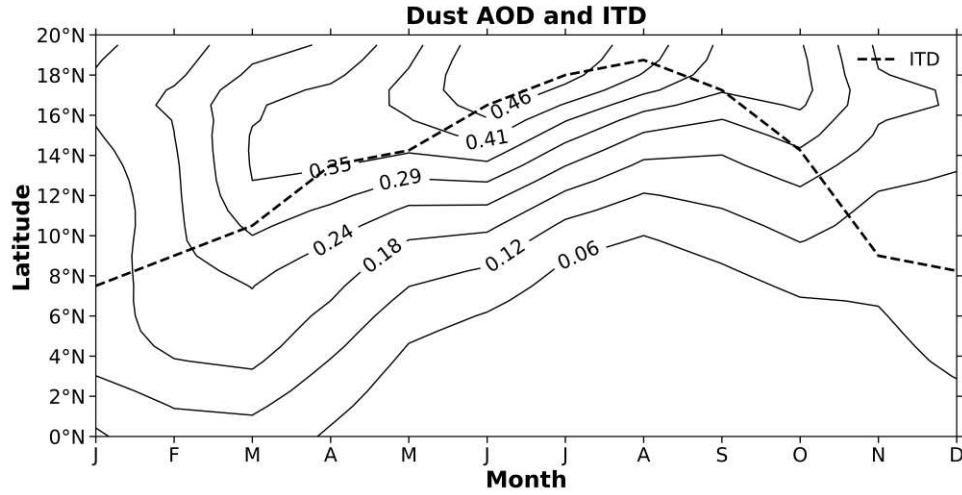


Figure 3. Contours show the monthly climatological CAMS dust AOD at 550 nm, averaged zonally over the West African longitude band. The dashed line indicates the latitudinal ITD position, derived from the near-surface wind-moisture discontinuity using the 10 m meridional wind reversal and the associated dew-point gradient.

The seasonal meridional profiles link this dust evolution to the migration of the low-level moisture and wind transition (**Fig. 3**). In DJF, the ITD is located near 8.3°N, with relatively moist southerly flow to the south and drier Harmattan air to the north. Dust AOD increases northward but remains lower than during the other seasons. By MAM, the ITD advances to approximately 13.4°N and dust AOD rises sharply toward 15–18°N. In JJA, the meridional wind remains southerly throughout the analysis domain, placing the ITD north of 20°N; dust AOD nevertheless increases rapidly north of about 10°N and reaches its maximum near the northern boundary as dew-point temperature declines. During SON, the ITD moves southward to about 15.2°N, and the northern dust maximum weakens. The link between the seasonal displacement of the dry-moist boundary and the latitude of enhanced dust agrees with previous evidence that the annual North African dust cycle is closely linked to meridional convergence. Strong dust mobilisation can occur near the leading edge of the monsoon flow in the ITD region (Engelstaedter and Washington, 2007; Bou Karam et al., 2009).

3.2 Effect of predictor structure and model architecture on validation performance

Linear Regression captured the broad annual cycle in **Fig. 4** but did not reproduce the temporal variability of CAMS dust AOD. Its prediction follows a relatively regular seasonal pattern, while CAMS contains stronger day-to-day fluctuations, irregular peaks and substantial variation within each year. This mismatch is reflected in its low evaluation r^2 of 0.28 and its larger RMSE and MAE of 0.143 and 0.108. Random Forest and LightGBM followed both the seasonal cycle and the shorter-term CAMS variations more closely, increasing evaluation r^2 to 0.54–0.55 and reducing RMSE to about 0.11, although several of the highest CAMS values remained underestimated.

Random Forest produced the strongest training fit, with r^2 of 0.66, but its evaluation r^2 decreased to 0.54. LightGBM retained an r^2 of 0.55 in both periods and achieved the lowest RMSE of 0.113 in evaluation. Mean biases were small and positive for all three models, showing that the main differences arose from their ability to reproduce temporal variability rather than from a persistent overall offset. LightGBM was therefore selected for independent testing because it combined the strongest evaluation skill with the most consistent performance between training and evaluation.

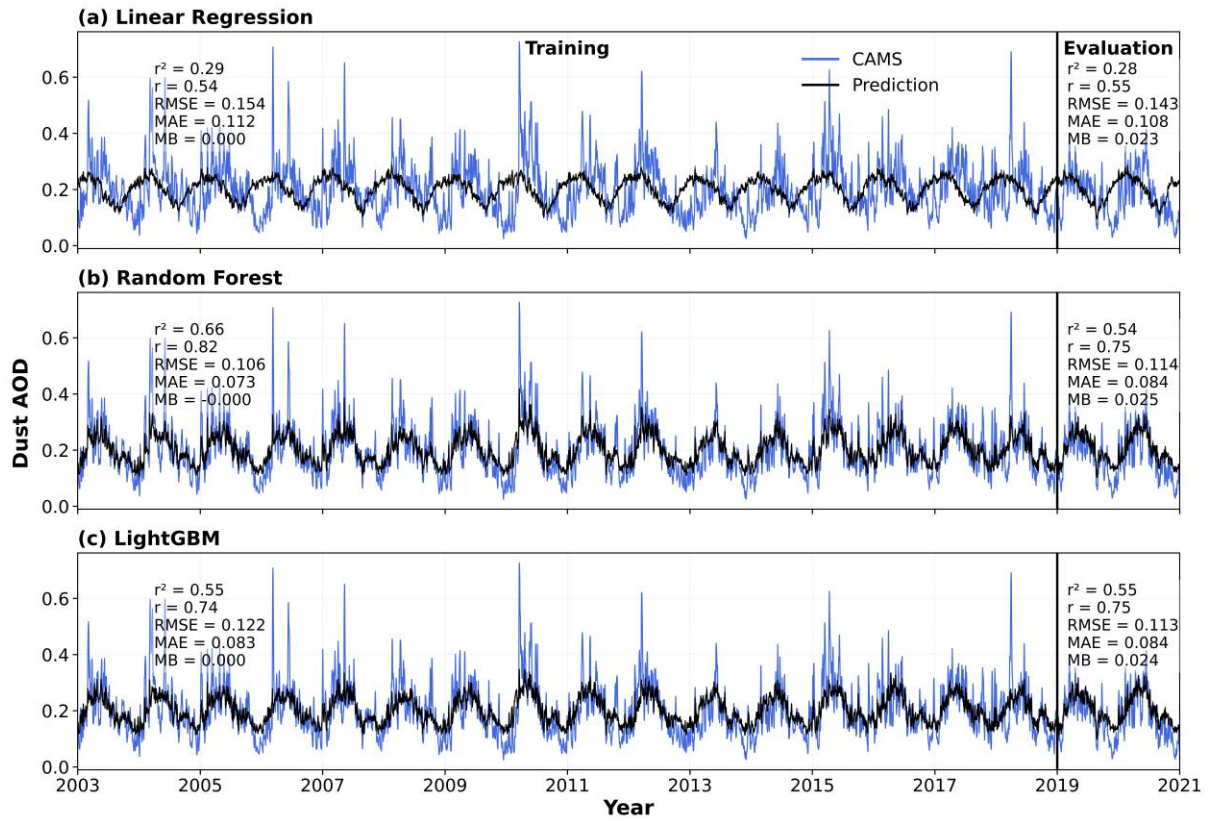


Figure 4: Training and validation time series of spatially averaged CAMS and predicted daily dust AOD over West Africa. The vertical line separates the 2003-2018 training period from the 2019-2020 evaluation period. Statistics reported in each section are the coefficient of determination (r^2), Pearson correlation coefficient (r), root-mean-square error (RMSE), mean absolute error (MAE) and mean bias (MB).

The error maps in **Fig. 5** show that Linear Regression produced stronger systematic errors than the tree-based models. Its annual error field is dominated by overestimation across eastern and southeastern parts of West Africa, together with smaller areas of underestimation in the northern and western sectors. Random Forest and LightGBM substantially reduce these errors and produce closely related spatial patterns, although both retain positive biases over parts of the northern and eastern domain.

The seasonal maps identify where the remaining errors occur. For example, in DJF, Linear Regression strongly overestimates dust AOD across most of West Africa, while Random Forest and LightGBM reduce the magnitude and spatial extent of this bias. During MAM and JJA, all three models underestimate dust AOD over parts of the northern and northwestern dust belt and overestimate it toward the east and south; this contrast is strongest for Linear Regression. In SON, positive errors re-emerge across the northern belt, although they remain weaker for the tree-based models. The seasonal error structure therefore shows that model performance is lowest where the position and intensity of the northern dust maximum change most strongly.

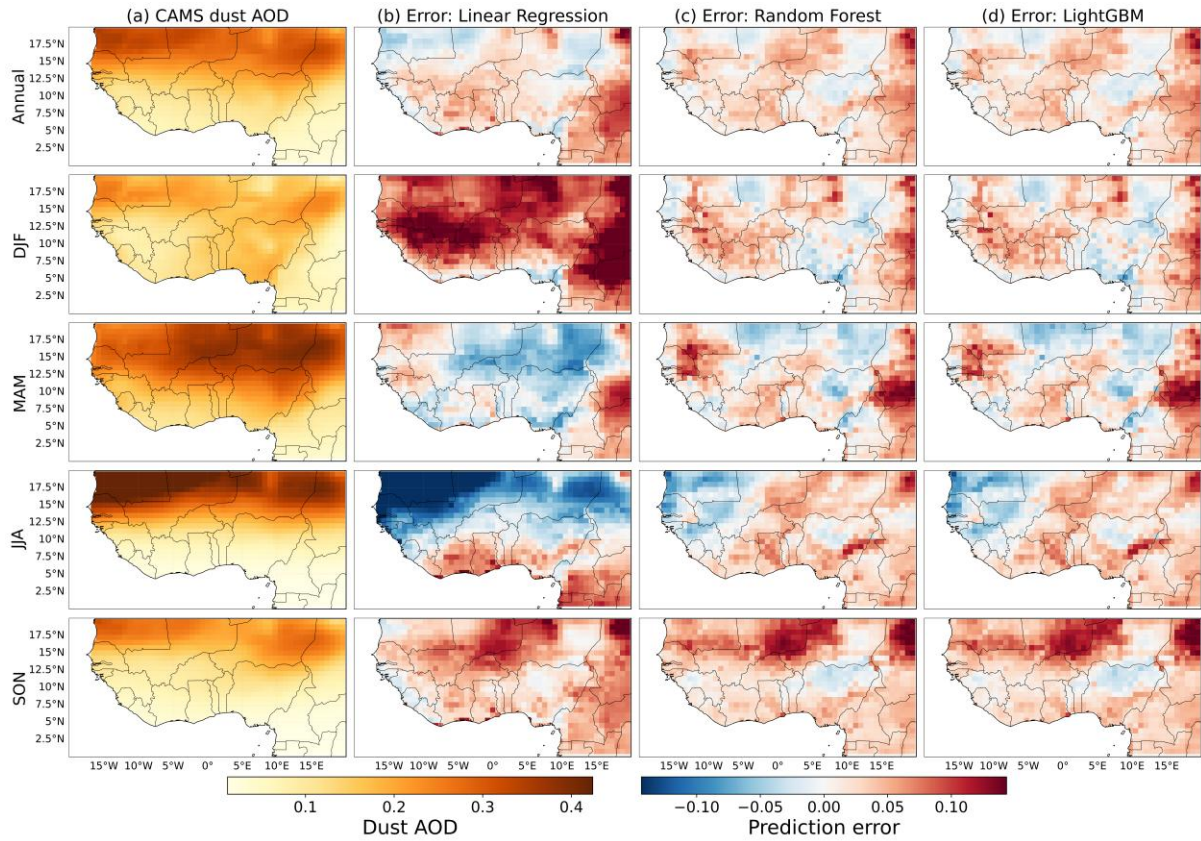


Figure 5: Annual and seasonal CAMS dust-AOD climatologies and spatial prediction errors from Linear Regression, Random Forest and LightGBM during the validation period. Column (a) shows the CAMS dust AOD at 550 nm, while columns (b–d) show the corresponding prediction errors for Linear Regression, Random Forest and LightGBM, respectively. Rows present the annual mean and the DJF, MAM, JJA and SON climatologies. Prediction error is calculated as modelled minus CAMS dust AOD; red indicates overestimation and blue indicates underestimation.

The stronger performance of Random Forest and LightGBM supports the nonlinear controls on mineral-dust emission. Dust mobilisation begins only after a threshold wind stress is exceeded and depends jointly on surface erodibility, roughness, particle-size distribution and soil moisture; saltation bombardment and aggregate disintegration further complicate the response (Marticorena and Bergametti, 1995; Shao, 2001; Kok et al., 2012). These relationships cannot be represented adequately by a single global linear function. Random Forest instead combines many randomised regression trees and averages their predictions, while LightGBM fits gradient-boosted trees sequentially to the remaining prediction errors (Breiman, 2001; Friedman, 2001; Ke et al., 2017). Their higher validation skill therefore indicates that the meteorological controls on dust AOD involve interactions and responses that Linear Regression could not capture.

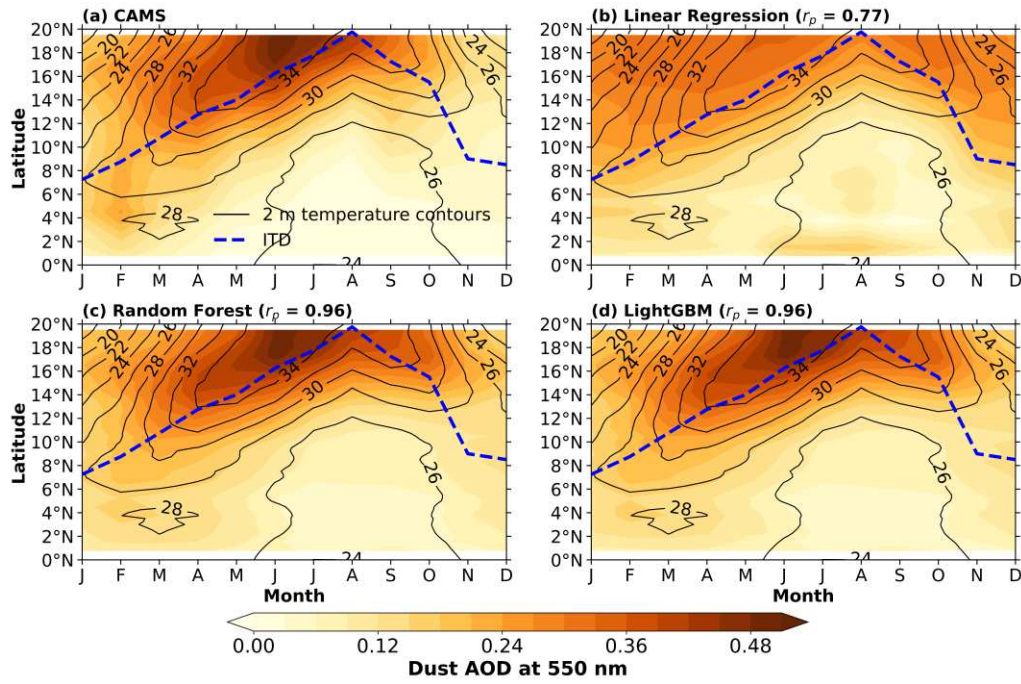


Figure 6: Latitude-month evolution of 2 m temperature and CAMS/model-predicted dust AOD relative to the ITD over West Africa. Shading shows the longitudinally averaged dust AOD from (a) CAMS, (b) Linear Regression, (c) Random Forest, and (d) LightGBM. The black contours show the monthly 2 m temperature, and the blue dashed line marks the latitudinal position of the ITD. The pattern correlation, r_p , measures the similarity between each predicted latitude dust-AOD pattern and CAMS.

Fig. 6 shows that the CAMS field places the annual dust maximum within the warm continental belt through which the ITD migrates. From March, rising temperatures over the Sahel and Sahara accompany the northward movement of the ITD and the expansion of high dust AOD toward northern latitudes. During June-August, the highest AOD occupies the strongly heated northern zone around 14-20°N, while the ITD lies near its northernmost position. This arrangement reflects the seasonal interaction of the dry Harmattan, moist monsoon flow and Saharan heat low; convergence and strong near-surface winds around the ITD can also promote dust emission and uplift over the Sahel (Bou Karam et al., 2009; Schepanski et al., 2017).

Linear Regression reproduces the broad northward shift but gives a weaker and less continuous dust maximum, yielding $r_p = 0.77$. Random Forest and LightGBM reproduce the CAMS structure much more closely, including the concentration of high AOD within the warm northern belt and its movement with the ITD; both attain $r_p = 0.96$. The pattern correlations show that the tree-based models recover the seasonal placement of the dust maximum, whereas Linear Regression represents only its broad progression.

3.3 Independent test performance and structure of LightGBM predictions

The final LightGBM model was retrained on the combined 2003-2020 development dataset and then applied to the completely unseen 2021-2024 period. It achieved $r^2 = 0.50$, $r = 0.71$, RMSE = 0.125, MAE = 0.087 and MB = 0.01 (**Fig. 7**). These values are slightly weaker than those obtained for the 2019–2020 validation period, but the difference should not be attributed to retraining. The model settings were chosen using the validation data, whereas the test years had no role in model development and contained a different mix of dust conditions.

The larger change in RMSE than in MAE shows that the reduction in test performance was driven mainly by a smaller number of large errors rather than by a general increase in error across all days. This matches the later distributional and event-based results, which show that the model reproduced common dust-AOD values more accurately than the highest values. At the same time, the small difference between training and test errors indicates that the model transferred reasonably well to 2021-2024, despite the greater difficulty of the independent period.

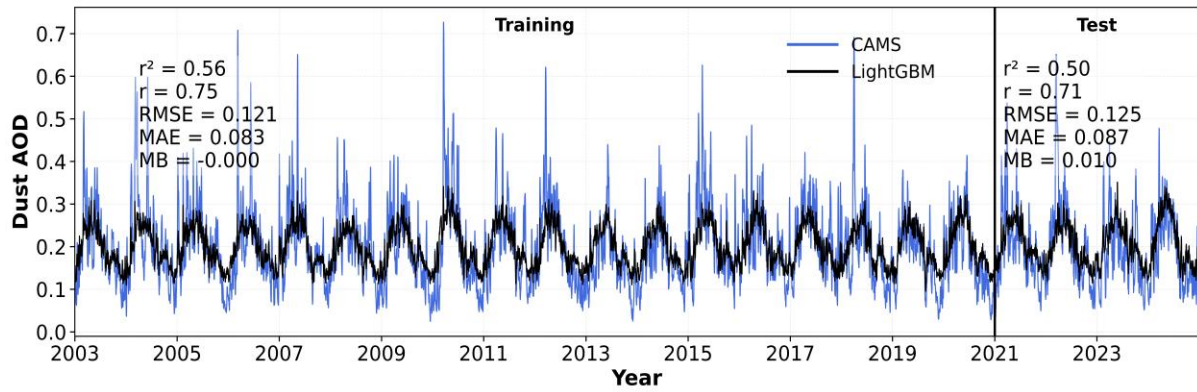


Figure 7: Independent test-period performance of the final LightGBM model for daily dust-AOD prediction over West Africa.

Fig. 8 shows where that limitation appears in distributional terms, with the PDF indicating that CAMS exhibits a stronger concentration of very low dust-AOD values near the lower end of the distribution. In contrast, LightGBM shifts part of that frequency into the low-to-moderate range. In practical terms, the model produced too few of the weakest dust-AOD conditions and too many values around roughly 0.10-0.30. The upper tail is also shortened in the prediction: CAMS extends to larger dust-AOD values, while the LightGBM distribution tapers off earlier. This means the model smoothed both ends of the distribution, especially the highest dust-AOD events

For the CDF, there exists a similar behaviour to the PDF from a cumulative perspective. The CAMS curve rises more quickly at low thresholds, confirming that CAMS contains a larger fraction of very weak dust-AOD values. The two curves converge toward the centre of the distribution, where the model performs best. At the upper end, however, the LightGBM curve reaches its asymptote earlier because the predicted distribution does not extend as far as CAMS. The model therefore reproduces the central bulk of the daily AOD better than the rare AOD peaks.

High-dust events were evaluated using the CAMS 90th-percentile dust-AOD threshold. At the CAMS 90th-percentile threshold of ~ 0.43 , high-dust events accounted for 10% of the test-period grid-cell days, whereas LightGBM classified about 8% as high-dust events (**Table 3**). The frequency bias of ~ 0.76 confirms an overall underprediction of event occurrence, agreeing with the shortened upper tail of the predicted dust-AOD distribution in **Fig. 8**. However, LightGBM detected $\sim 41\%$ of the CAMS events, while $\sim 46\%$ of its predicted events were false alarms. The critical success index of ~ 0.31 shows that fewer than one-third of all observed or predicted high-dust cases were correctly matched.

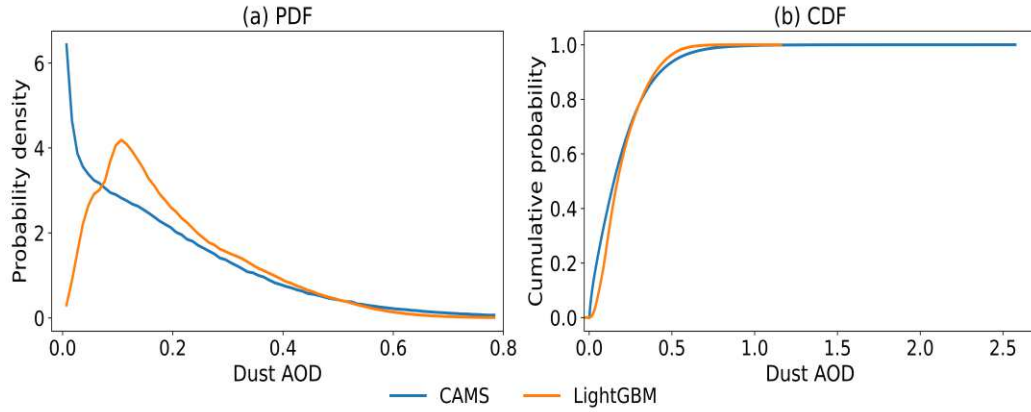


Figure 8: Probability density and cumulative distribution comparison of CAMS and LightGBM dust AOD during the independent test period (2021–2024). (a) shows the probability density function (PDF), and (b) shows the cumulative distribution function (CDF), for CAMS dust AOD and LightGBM-predicted dust AOD over West Africa.

Table 3: Contingency-table evaluation of LightGBM high-dust event detection at the CAMS 90th-percentile threshold during 2021–2024. The table reports the CAMS and predicted high-dust frequencies (HDF), hit frequency (HF), missed-event frequency (MF), false-alarm frequency (FAF), probability of detection (POD), false-alarm ratio (FAR), critical success index (CSI) and frequency bias (FB).

Threshold	Value	CAMS HDF	Pred. HDF	HF	MF	FAF	POD	FAR	CSI	FB
CAMS P90	0.428	0.10	0.076	0.041	0.058	0.034	41.4%	0.455	0.308	0.761

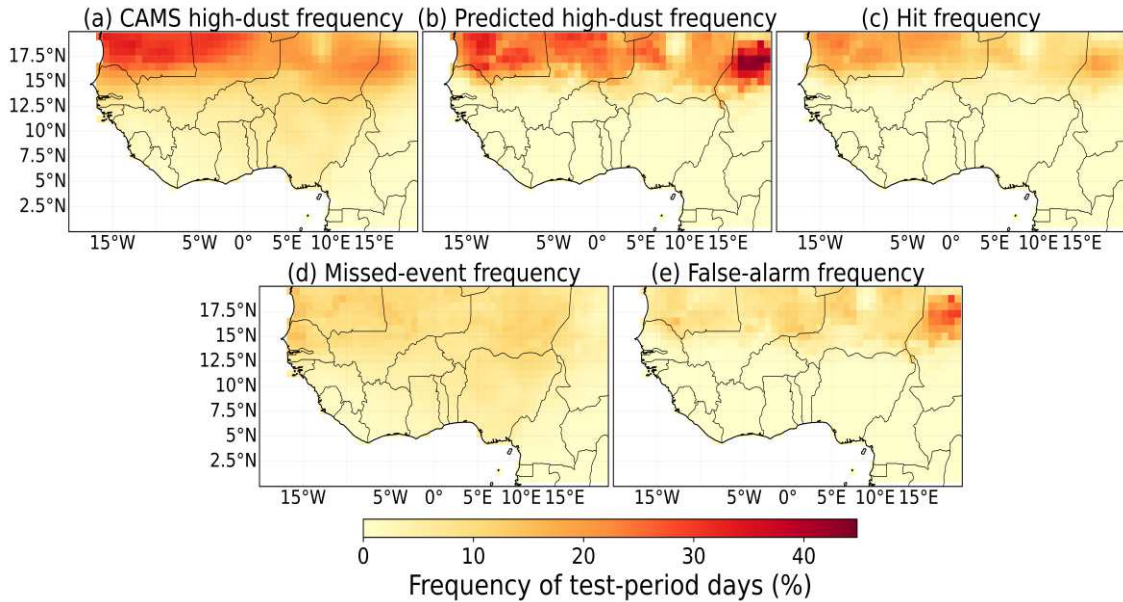


Figure 9: Spatial occurrence and LightGBM detection of CAMS-defined high-dust events over West Africa during the independent test period (2021–2024). High-dust events were defined separately using the 90th percentile of CAMS dust AOD. Panels show the percentage of test-period days classified as (a) CAMS high-dust events, (b) LightGBM-predicted high-dust events, (c) correctly detected events, (d) missed events and (e) false alarms. A hit occurs when CAMS and LightGBM both exceed the threshold; a missed event occurs when only CAMS exceeds it; and a false alarm occurs when only LightGBM exceeds it.

The spatial distributions in **Fig. 9** show that both CAMS and LightGBM place the highest event frequencies within the northern dust belt. The prediction nevertheless redistributes part of this occurrence eastward: missed events are concentrated across sections of the western and north-central dust belt, whereas false alarms are strongest in the far northeast. The model therefore identifies the broad region where high-dust conditions occur but does not consistently reproduce their location and timing. This explains how the domain-wide event frequency can be underestimated even though some northeastern grid cells show excessive predicted occurrence. The event scores were interpreted using the contingency-table verification framework of Jolliffe and Stephenson (2012).

3.4 Predictor influence and physical interpretation of the LightGBM model

The LightGBM importance ranking is dominated by 2 m temperature, which contributes by far the largest total gain, followed by 10-m wind speed, relative humidity, and surface soil moisture (**Fig. 10**). This includes previous-day wind speed, u10, and boundary-layer height, while the rainfall variables contribute little, especially daily rainfall and previous-day rainfall. This ranking shows that the model draws most of its predictive skill from the thermal, dynamical and near-surface moisture fields, not from rainfall alone.

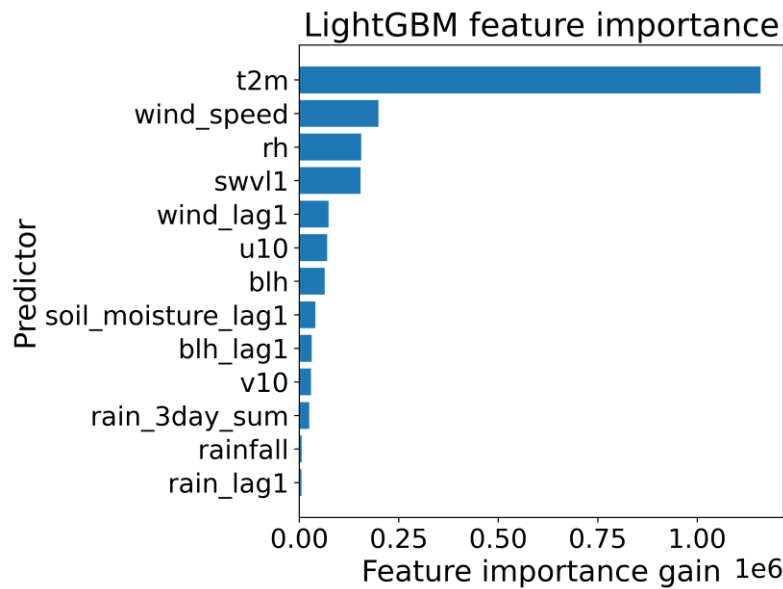


Figure 10: Gain-based predictor importance from the final LightGBM model for daily dust-AOD prediction over West Africa. Bars show the total gain contributed by each predictor across the fitted ensemble. Larger gain indicates a greater contribution to reducing model error.

The SHAP summary in **Fig. 11** and **Table 4** shows how these predictors affect the predicted dust AOD. High t2m values are associated mainly with positive SHAP values, whereas low t2m values are associated with negative contributions. Temperature therefore pushes the prediction upward under warmer conditions and downward under cooler conditions. Relative humidity shows the opposite behaviour: high RH values are concentrated on the negative SHAP side, whereas low RH values are more associated with positive contributions. This is a physically consistent result, because drier conditions favour dust emission and persistence, whereas moister conditions suppress them.

Wind speed and wind_lag1 also show a clear positive influence at higher values. Their SHAP distributions extend strongly into the positive range, indicating that stronger present-day and antecedent winds increase predicted dust AOD. This matches the basic physics of dust

mobilisation, since dust emission depends strongly on wind forcing and increases once the surface stress exceeds the threshold for particle entrainment (Marticorena and Bergametti, 1995; Shao, 2001; Kok et al., 2012). The role of u_{10} is weaker than that of total wind speed, but it still contributes positively in many cases, showing that the directional wind field adds information beyond speed alone.

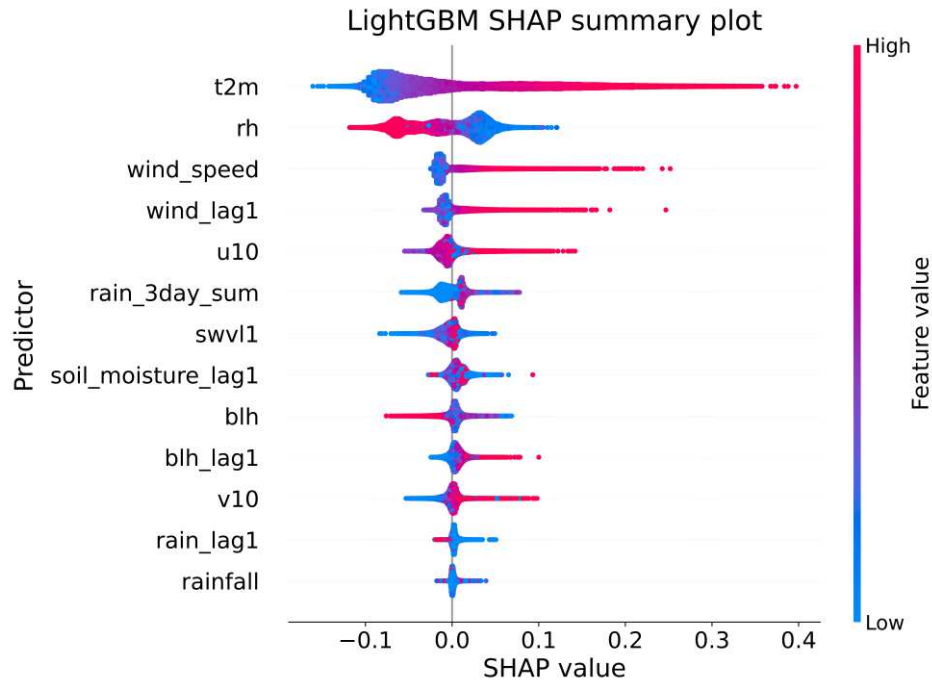


Figure 11: SHAP summary plot for the final LightGBM model where each point represents one test-period sample. The horizontal axis shows the SHAP value, where positive values increase the predicted dust AOD and negative values decrease it. Point colour shows the predictor value from low to high.

The remaining predictors act mainly as secondary modifiers. $swvl_1$, $soil_moisture_lag_1$, blh , blh_lag_1 , v_{10} , and $rain_3day_sum$ mostly cluster closer to zero in the SHAP summary, so their effects are smaller and more context-dependent. That does not make them unimportant; it means that they help adjust the prediction under specific combinations of meteorological conditions rather than driving the prediction on their own. By contrast, $rainfall$ and $rain_lag_1$ contribute very little, which suggests that, in this model, $rainfall$ influences dust AOD more indirectly through its links with moisture and surface conditions than as a strong standalone predictor.

Predictor names are as follows: t_{2m} , 2 m air temperature; $wind_speed$, 10 m wind speed; rh , relative humidity; $swvl_1$, volumetric soil water in layer 1; blh , boundary-layer height; u_{10} , 10-m zonal wind; v_{10} , 10-m meridional wind; $rain_3day_sum$, accumulated 3-day rainfall sum; and lag_1 , previous-day value.

Table 4: Relative importance of predictor variables derived from the LightGBM SHAP analysis.

Rank	Predictor	~ SHAP value range	Effect	Interpretation
1	t2m	-0.12 to 0.42	Very High	Higher temperature strongly increases the model prediction.
2	rh	-0.10 to 0.11	High	Relative humidity has a strong influence with both positive and negative effects.
3	wind_speed	-0.02 to 0.24	Moderate–High	Strong winds generally improve the prediction.
4	wind_lag1	-0.03 to 0.20	Moderate	Previous-day wind speed provides important temporal information.
5	u10	-0.03 to 0.13	Moderate	Zonal wind component contributes moderately.
6	rain_3day_sum	-0.04 to 0.07	Low–Moderate	Accumulated rainfall has a relatively small effect.
7	swvl1	-0.03 to 0.05	Low	Surface soil moisture contributes weakly.
8	soil_moisture_lag1	-0.03 to 0.04	Low	Previous-day soil moisture has limited predictive power.
9	blh	-0.07 to 0.04	Low	ABL has a modest negative influence when low.
10	blh_lag1	-0.03 to 0.04	Very Low	Previous boundary layer height contributes minimally.
11	v10	-0.03 to 0.08	Very Low	Meridional wind component has a relatively weak influence.
12	rain_lag1	-0.02 to 0.03	Negligible	Previous-day rainfall has little predictive effect.
13	rainfall	-0.02 to 0.02	Negligible	Present-day rainfall is the least influential predictor.

The dependence plots in **Figs. 12-13** show why the tree-based model performed better than Linear Regression. Temperature contributions become strongly positive above roughly 28-30°C, while relative humidity shifts progressively toward negative contributions as moisture increases. Wind speed and its one-day lag remain weak at low values but rise sharply above about 3 m s⁻¹, indicating a threshold response. Zonal wind also changes nonlinearly, with stronger positive values producing larger positive contributions. These relationships show that dust-AOD prediction depends on curved and threshold-sensitive responses to heat, dryness and wind forcing, which cannot be represented well by a single linear function (Marticorena and Bergametti, 1995; Shao, 2001; Kok et al., 2012).

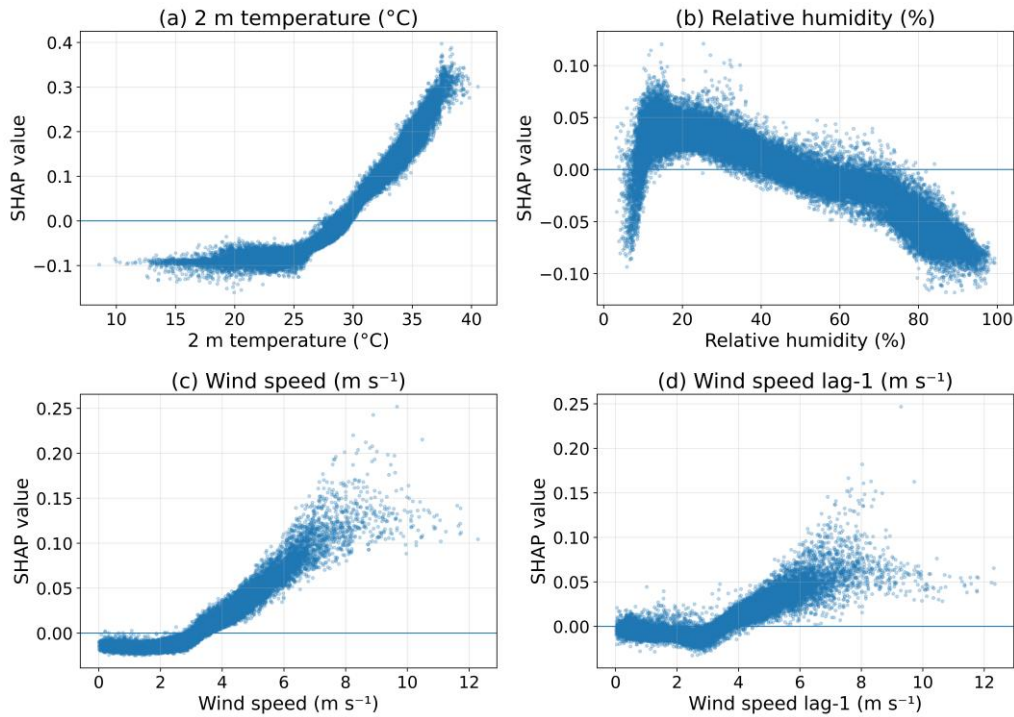


Figure 12: SHAP dependence of the leading meteorological predictors in the final LightGBM model for daily dust-AOD prediction over West Africa. Panels show the dependence of SHAP values on (a) 2 m air temperature, (b) relative humidity, (c) 10 m wind speed, and (d) previous-day 10 m wind speed. Positive SHAP values increase the predicted dust AOD, whereas negative SHAP values decrease it.

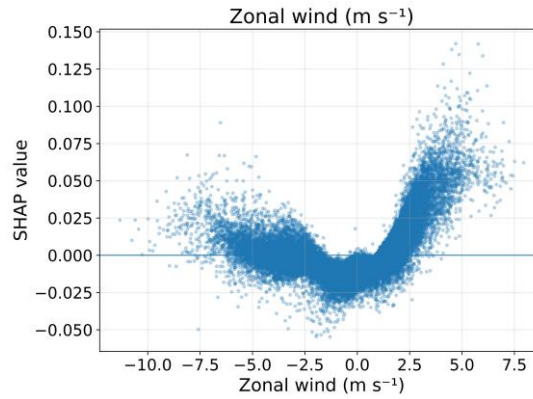


Figure 13: SHAP dependence of the 10-m zonal wind contribution in the final LightGBM model. The plot shows how the contribution of u_{10} changes across its observed range, with positive SHAP values increasing the predicted dust AOD and negative SHAP values decreasing it.

3.5 SOM analysis of spatial dust patterns and seasonal model errors

Fig. 14 groups 12 seasonal bias maps, comprising four seasons for each of the three models. Tags a-d represent Linear Regression in DJF, MAM, JJA and SON; e-h represent Random Forest; and i-l represent LightGBM. Node 5 is empty because none of the 12 bias maps had a spatial error pattern close enough to be assigned to that node. Random Forest and LightGBM are paired in the same node for every season: e and i in node 4 for DJF, f and j in node 9 for MAM, g and k in node 6 for JJA, and h and l in node 3 for SON. Their seasonal bias maps are

therefore very similar. The four Linear Regression maps are assigned to separate nodes 1, 8, 7 and 2 for DJF, MAM, JJA and SON, showing larger seasonal changes in their spatial errors.

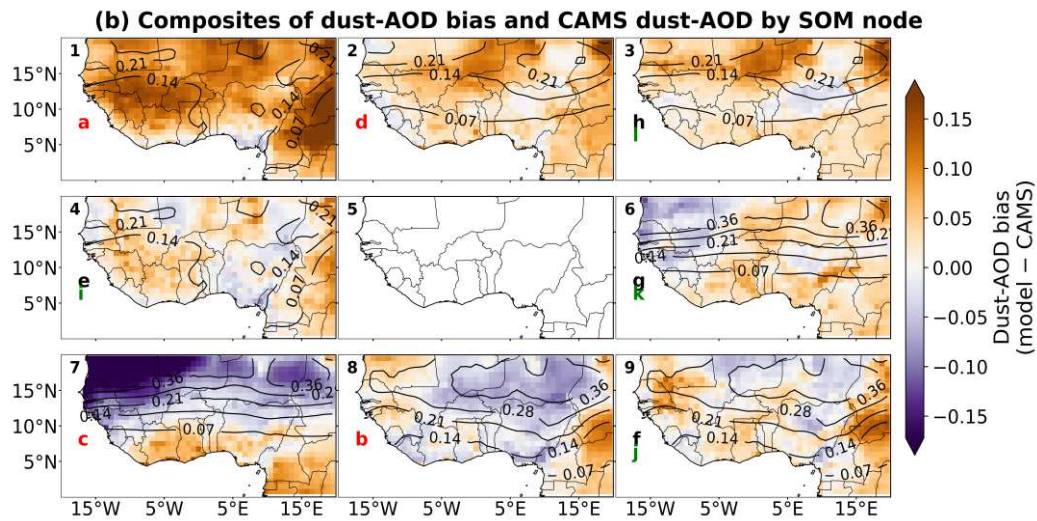


Figure 14: SOM grouping of seasonal ML models dust-AOD bias patterns and their correspondence with CAMS dust loading. Shading represents the composite prediction bias, calculated as model minus CAMS dust AOD, while black contours show the corresponding CAMS dust AOD. Letters identify the model-season tags listed in Table 7; Linear Regression tags are red, Random Forest tags are black, and LightGBM tags are green.

Table 5: Self-organising-map node assignments and error statistics for the seasonal prediction-bias fields from Linear Regression, Random Forest, and LightGBM.

Simulation Tags	SOM		Mean Bias		RMSE	
	Node	Rank	Value	Rank	Value	Rank
b	8	1	-0.009	1	0.147	11
k	6	2	0.017	2	0.112	5
g	6	2	0.018	4	0.113	6
j	9	3	0.017	3	0.139	8
f	9	3	0.019	5	0.14	9
i	4	4	0.021	6	0.103	3
e	4	4	0.022	7	0.103	4
c	7	5	-0.041	9	0.156	12
l	3	6	0.041	8	0.094	1
h	3	6	0.042	10	0.094	2
d	2	7	0.048	11	0.118	7
a	1	8	0.095	12	0.146	10

The quantification of the errors associated with these tags is shown in Table 5. Tag l, representing LightGBM–SON, has the lowest RMSE of 0.094, followed by tag h, RF–SON, with a similar value. Tag c, LR–JJA, has the highest RMSE of 0.156, while tag a, LR–DJF, has the largest positive mean bias of 0.095. Tag b, LR–MAM, has the smallest absolute mean bias at -0.009 but ranks eleventh by RMSE, showing that positive and negative spatial errors cancel in the domain average.

Fig. 15 shows how daily CAMS dust AOD varies in both magnitude and spatial distribution across West Africa. Node 4 is the most frequent pattern, representing 16.8% of all days, followed by nodes 8, 7, 12 and 3, which contribute 12.9%, 9.6%, 8.6% and 8.1%, respectively.

These commonly occurring nodes are characterised mainly by low-to-moderate dust AOD, with the higher values generally confined to the northern part of the region. This suggests that, on most days, dust remains concentrated near the main source areas and does not spread strongly across the southern part of West Africa. In contrast, the dustiest patterns occur much less often. Node 5, which shows high dust AOD across much of the domain, represents only 0.9% of all days, while nodes 9 and 13 each account for about 2.0%. Node 9 is more strongly centred over central and eastern West Africa, whereas node 13 shows a broad band of high dust AOD across the north. The low contribution of these nodes shows that widespread dusty conditions are relatively uncommon compared with the more frequent low-to-moderate patterns. Nodes 6, 10, 11, 14, 15 and 16 represent intermediate conditions, with differences in the strength, width and position of the northern dust band.

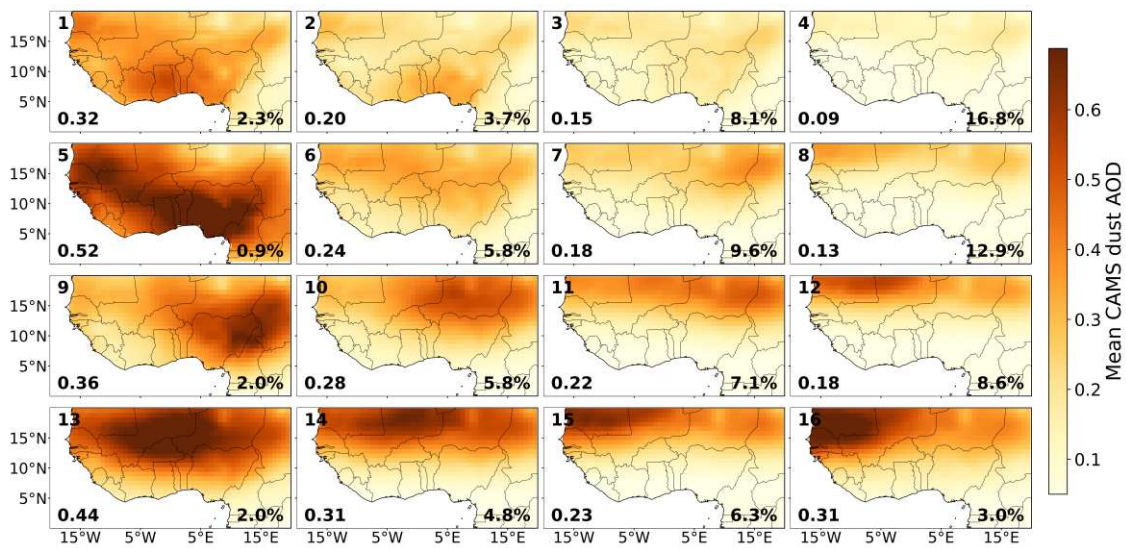


Figure 15: Sixteen-node SOM map of daily CAMS dust aerosol optical depth, where each panel shows the mean dust AOD distribution for all days assigned to that node, with nodes arranged according to their positions. The value in the lower-left of each panel is the domain-mean dust AOD, while the value in the lower-right is the percentage of the 8,036 daily CAMS distributions represented by that node.

Fig. 16 shows clear differences in how the three models reproduce the CAMS dust-AOD patterns during 2019-2020. Linear Regression assigns most of its predictions to nodes 6, 7 and 8, with more than 40% placed in node 6 alone. This shows that its predictions are concentrated within a narrow range of moderate dust conditions, rather than following the wider day-to-day changes in dust amount and location seen in CAMS. Random Forest and LightGBM reproduce a broader range of patterns and give very similar contributions across most nodes. However, both models place very few days in nodes 5, 9 and 13, where CAMS shows some of the highest and most widespread dust loading. These dusty days are likely more difficult to predict because they depend on several factors occurring together, including strong winds, dry surface conditions, dust uplift and regional transport. The predictor set contains meteorological and land-surface information but does not include direct dust-emission estimates or previous-day dust AOD. This may explain why the models reproduce the more common moderate patterns more successfully than the less frequent, very dusty patterns.

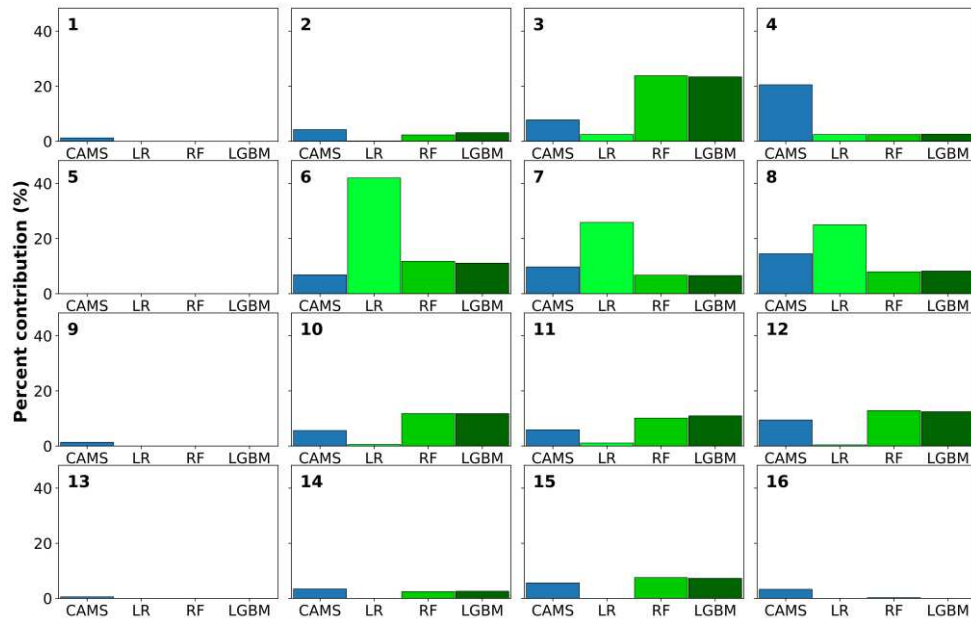


Figure 16: Percentage contribution of the evaluation-period daily CAMS, Linear Regression, Random Forest and LightGBM dust-AOD distributions to the sixteen CAMS-defined SOM nodes during 2019-2020. The SOM nodes were first derived from the full 2003-2024 CAMS record shown in Fig. 15, after which each daily CAMS and model-predicted distribution from the evaluation period was assigned to its closest node. The bars therefore show how frequently each observed or predicted daily spatial pattern resembles each CAMS dust-AOD pattern.

3.6 Synthesis of model behaviour and remaining limitations

The independent diagnostics identify the same limitation in the final LightGBM model. The small mean bias shows that it did not consistently overpredict or underpredict dust AOD across the full domain, but the PDF and CDF reveal a shortened upper tail. At the CAMS P90 threshold, the model predicted fewer high-dust events than observed, detected only 41% of them and produced a frequency bias of 0.76. The SOM analysis gives the spatial counterpart of this result, where LightGBM represented the frequent low-to-moderate CAMS patterns more successfully but assigned almost no predictions to nodes 5, 9, and 13, which contain the rarest and highest dust-AOD values. The main weakness is therefore the underrepresentation of intense dust conditions, not a uniform error across all days.

SHAP analysis shows that the predictions depend mainly on temperature, wind speed, relative humidity and surface soil moisture. These variables describe atmospheric and surface conditions associated with dust mobilisation, but they do not directly specify the amount of erodible material available, active source strength, pre-existing aerosol loading or dust transported from upstream regions. Such information becomes particularly important during intense events because dust emission responds to threshold wind forcing and depends on surface erodibility and moisture, while the resulting aerosol distribution is further controlled by uplift and transport pathways (Marticorena and Bergametti, 1995; Shao, 2001; Kok et al., 2012; Knippertz and Todd, 2012). Future development should therefore examine direct indicators of source activity, upstream aerosol loading and antecedent dust conditions, with particular attention to reducing missed high-dust events.

4 Conclusions

This study demonstrates that daily CAMS dust AOD over West Africa can be estimated from meteorological, land-surface and rainfall predictors without using direct aerosol information.

LightGBM provided the most reliable representation of the spatial and temporal variability, while Linear Regression was unable to capture the nonlinear relationships linking dust AOD to temperature, wind and moisture conditions.

The main contribution is the identification of both the capability and the limit of this approach. For instance, LightGBM reproduced the common dust patterns more successfully than the rare, intense conditions, showing that good overall performance does not guarantee reliable representation of high-dust events. The framework is therefore useful for daily dust-AOD reconstruction and diagnostic analysis, but further improvement requires independent observational evaluation and predictors that describe source activation, antecedent aerosol loading and transport history more directly.

Data availability

The CAMS global reanalysis (EAC4) dust aerosol optical depth data used in this study are publicly available from the Copernicus Atmosphere Data Store (<https://doi.org/10.24381/d58bbf47>). The ERA5 meteorological and land-surface data are publicly available from the Copernicus Climate Data Store (<https://doi.org/10.24381/cds.adbb2d47>). The CHIRPS v2.0 daily precipitation data are publicly available from the Climate Hazards Centre at the University of California, Santa Barbara.

Author contribution

PA conceptualised the study, developed the methodology and software, processed and analysed the data, produced the figures, interpreted the results and prepared the original manuscript. BA and MW contributed to the study design, scientific interpretation and supervision. JA contributed to the methodology and interpretation of the results. AO contributed to the study design, supervision and scientific interpretation. All authors reviewed and edited the manuscript and approved the final version.

Competing interests

The authors declare that they have no conflict of interest.

Acknowledgements

The authors acknowledge the Copernicus Atmosphere Monitoring Service and the Copernicus Climate Change Service for providing access to the CAMS EAC4 and ERA5 datasets, respectively. The authors also acknowledge the Climate Hazards Centre at the University of California, Santa Barbara, for providing access to the CHIRPS v2.0 precipitation data.

Financial support

This research received no external funding.

References

- Alduchov, O. A. and Eskridge, R. E.: Improved Magnus form approximation of saturation vapor pressure, *J. Appl. Meteorol.*, 35, 601–609, [https://doi.org/10.1175/1520-0450\(1996\)035<0601:IMFAOS>2.0.CO;2](https://doi.org/10.1175/1520-0450(1996)035<0601:IMFAOS>2.0.CO;2), 1996.
- Atiah, W. A., Amekudzi, L. K., Aryee, J. N. A., Preko, K., and Danuor, S. K.: Validation of satellite and merged rainfall data over Ghana, West Africa, *Atmosphere*, 11, 859, <https://doi.org/10.3390/atmos11080859>, 2020.
- Awoleye, P. O., Adebisi, A. A., Abiodun, B. J., Balogun, I. A., Akinbobola, A., Oluleye, A., and Ajayi, V. O.: Moisture and dust in motion: the dual role of integrated vapour transport over West Africa, *Theor. Appl. Climatol.*, 156, 666, <https://doi.org/10.1007/s00704-025-05913-1>, 2025a.
- Awoleye, P. O., Ogunjobi, K. O., and Balogun, I. A.: Climatology and dynamics of aerosol optical properties from selected AERONET sites over West Africa, *Discover Atmosphere*, 3, 11, <https://doi.org/10.1007/s44292-025-00036-8>, 2025b.
- Benedetti, A., Morcrette, J. J., Boucher, O., Dethof, A., Engelen, R. J., Fisher, M., Flentje, H., Huneeus, N., Jones, L., Kaiser, J. W., Kinne, S., Mangold, A., Razinger, M., Simmons, A. J., and Suttie, M.: Aerosol analysis and forecast in the European Centre for Medium-Range Weather Forecasts Integrated Forecast System: 2. Data assimilation, *J. Geophys. Res.-Atmos.*, 114, D13205, <https://doi.org/10.1029/2008JD011115>, 2009.
- Bou Karam, D., Flamant, C., Tulet, P., Chaboureaud, J.-P., Dabas, A., and Todd, M. C.: Estimate of Sahelian dust emissions in the intertropical discontinuity region of the West African Monsoon, *J. Geophys. Res.-Atmos.*, 114, D13106, <https://doi.org/10.1029/2008JD011444>, 2009.
- Breiman, L.: Random forests, *Mach. Learn.*, 45, 5–32, <https://doi.org/10.1023/A:1010933404324>, 2001.
- Buchard, V., Randles, C. A., da Silva, A. M., Darmenov, A., Colarco, P. R., Govindaraju, R., Ferrare, R., Hair, J., Beyersdorf, A. J., Ziemba, L. D., and Yu, H.: The MERRA-2 aerosol reanalysis, 1980 onward, Part II: Evaluation and case studies, *J. Climate*, 30, 6851–6872, <https://doi.org/10.1175/JCLI-D-16-0613.1>, 2017.
- Cassano, E. N., Lynch, A. H., Cassano, J. J., and Koslow, M. R.: Classification of synoptic patterns in the western Arctic associated with extreme events at Barrow, Alaska, USA, *Clim. Res.*, 30, 83–97, <https://doi.org/10.3354/cr030083>, 2006.
- De Longueville, F., Hountondji, Y.-C., Henry, S., and Ozer, P.: What do we know about effects of desert dust on air quality and human health in West Africa compared to other regions?, *Sci. Total Environ.*, 409, 1–8, <https://doi.org/10.1016/j.scitotenv.2010.09.025>, 2010.
- Dembélé, M. and Zwart, S. J.: Evaluation and comparison of satellite-based rainfall products in Burkina Faso, West Africa, *Int. J. Remote Sens.*, 37, 3995–4014, <https://doi.org/10.1080/01431161.2016.1207258>, 2016.
- Engelstaedter, S. and Washington, R.: Atmospheric controls on the annual cycle of North African dust, *J. Geophys. Res.-Atmos.*, 112, D03103, <https://doi.org/10.1029/2006JD007195>, 2007.
- Friedman, J. H.: Greedy function approximation: A gradient boosting machine, *Ann. Stat.*, 29, 1189–1232, <https://doi.org/10.1214/aos/1013203451>, 2001.

- 1 Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., Husak, G., Rowland,
2 J., Harrison, L., Hoell, A., and Michaelsen, J.: The climate hazards infrared precipitation with
3 stations—a new environmental record for monitoring extremes, *Sci. Data*, 2, 150066,
4 <https://doi.org/10.1038/sdata.2015.66>, 2015.
- 5 Gibson, P. B., Perkins-Kirkpatrick, S. E., Uotila, P., Pepler, A. S., and Alexander, L. V.: On
6 the use of self-organizing maps for studying climate extremes, *J. Geophys. Res.-Atmos.*, 122,
7 3891–3903, <https://doi.org/10.1002/2016JD026256>, 2017.
- 8 Gilbert, J., Aryee, J. N. A., Adjei, M. J., Mensah, C., and Quagraine, K. T.: Machine learning-
9 based assessment of aerosol optical depth over Ghana, West Africa using MODIS satellite data,
10 *PLOS Clim.*, 4, e0000651, <https://doi.org/10.1371/journal.pclm.0000651>, 2025.
- 11 Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas,
12 J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X.,
13 Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P.,
14 Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A.,
15 Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P.,
16 Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and
17 Thépaut, J.-N.: The ERA5 global reanalysis, *Q. J. Roy. Meteor. Soc.*, 146, 1999–2049,
18 <https://doi.org/10.1002/qj.3803>, 2020.
- 19 Hewitson, B. C. and Crane, R. G.: Self-organizing maps: applications to synoptic climatology,
20 *Clim. Res.*, 22, 13–26, <https://doi.org/10.3354/cr022013>, 2002.
- 21 HOUNGNIPO, M. C. M., MINOUNGOU, B., TRAORE, S. B., MAIDMENT, R. I., ALHASSANE, A., and ALI,
22 A.: Validation of high-resolution satellite precipitation products over West Africa for rainfall
23 monitoring and early warning, *Front. Clim.*, 5, 1185754,
24 <https://doi.org/10.3389/fclim.2023.1185754>, 2023.
- 25 Inness, A., Ades, M., Agustí-Panareda, A., Barré, J., Benedictow, A., Blechschmidt, A.-M.,
26 Dominguez, J. J., Engelen, R., Eskes, H., Flemming, J., Huijnen, V., Jones, L., Kipling, Z.,
27 Massart, S., Parrington, M., Peuch, V.-H., Razinger, M., Rémy, S., Schulz, M., and Suttie, M.:
28 The CAMS reanalysis of atmospheric composition, *Atmos. Chem. Phys.*, 19, 3515–3556,
29 <https://doi.org/10.5194/acp-19-3515-2019>, 2019.
- 30 Jolliffe, I. T. and Stephenson, D. B.: *Forecast Verification: A Practitioner's Guide in*
31 *Atmospheric Science*, 2nd edn., Wiley, Chichester, <https://doi.org/10.1002/9781119960003>,
32 2012.
- 33 Just, A. C., De Carli, M. M., Shtein, A., Dorman, M., Lyapustin, A., and Kloog, I.: Correcting
34 measurement error in satellite aerosol optical depth with machine learning for modeling PM_{2.5}
35 in the Northeastern USA, *Remote Sens.*, 10, 803, <https://doi.org/10.3390/rs10050803>, 2018.
- 36 Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y.: LightGBM:
37 A highly efficient gradient boosting decision tree, *Adv. Neural Inf. Process. Syst.*, 30, 3146–
38 3154, <https://doi.org/10.5555/3294996.3295074>, 2017.
- 39 Knippertz, P. and Todd, M. C.: Mineral dust aerosols over the Sahara: Meteorological controls
40 on emission and transport and implications for modeling, *Rev. Geophys.*, 50, RG1007,
41 <https://doi.org/10.1029/2011RG000362>, 2012.
- 42 Kohonen, T.: The self-organizing map, *Proc. IEEE*, 78, 1464–1480,
43 <https://doi.org/10.1109/5.58325>, 1990.

1 Kohonen, T.: Self-Organizing Maps, 3rd edn., Springer, Berlin, [https://doi.org/10.1007/978-3-](https://doi.org/10.1007/978-3-642-56927-2)
2 [642-56927-2](https://doi.org/10.1007/978-3-642-56927-2), 2001.

3 Kok, J. F., Parteli, E. J. R., Michaels, T. I., and Bou Karam, D.: The physics of wind-blown
4 sand and dust, *Rep. Prog. Phys.*, 75, 106901, <https://doi.org/10.1088/0034-4885/75/10/106901>,
5 2012.

6 Liang, T., Liang, S., Zou, L., Sun, L., Li, B., Lin, H., He, T., and Tian, F.: Estimation of aerosol
7 optical depth at 30 m resolution using Landsat imagery and machine learning, *Remote Sens.*,
8 14, 1053, <https://doi.org/10.3390/rs14051053>, 2022.

9 Lundberg, S. M. and Lee, S.-I.: A unified approach to interpreting model predictions, *Adv.*
10 *Neural Inf. Process. Syst.*, 30, 4765–4774, <https://doi.org/10.48550/arXiv.1705.07874>, 2017.

11 Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R.,
12 Himmelfarb, J., Bansal, N., and Lee, S.-I.: From local explanations to global understanding
13 with explainable AI for trees, *Nat. Mach. Intell.*, 2, 56–67, [https://doi.org/10.1038/s42256-019-](https://doi.org/10.1038/s42256-019-0138-9)
14 [0138-9](https://doi.org/10.1038/s42256-019-0138-9), 2020.

15 Marticorena, B. and Bergametti, G.: Modeling the atmospheric dust cycle: 1. Design of a soil-
16 derived dust emission scheme, *J. Geophys. Res.-Atmos.*, 100, 16415–16430,
17 <https://doi.org/10.1029/95JD00690>, 1995.

18 Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G.,
19 Boussetta, S., Choulga, M., Harrigan, S., Hersbach, H., Martens, B., Miralles, D. G., Piles, M.,
20 Rodríguez-Fernández, N. J., Zsoter, E., Buontempo, C., and Thépaut, J.-N.: ERA5-Land: a
21 state-of-the-art global reanalysis dataset for land applications, *Earth Syst. Sci. Data*, 13, 4349–
22 4383, <https://doi.org/10.5194/essd-13-4349-2021>, 2021.

23 Nigro, M. A., Cassano, J. J., and Seefeldt, M. W.: A weather-pattern-based approach to
24 evaluate the Antarctic Mesoscale Prediction System forecasts: Comparison to automatic
25 weather station observations, *Weather Forecast.*, 26, 184–198,
26 <https://doi.org/10.1175/2010WAF2222444.1>, 2011.

27 Ogunjobi, K. O. and Awoleye, P. O.: Intercomparison and validation of satellite and ground-
28 based aerosol optical depth (AOD) retrievals over six AERONET sites in West Africa, *Aerosol*
29 *Sci. Eng.*, 3, 32–47, <https://doi.org/10.1007/s41810-019-00040-7>, 2019.

30 Paiva, M. S., Franco, M. A., and Rizzo, L. V.: Evaluation of near-surface atmospheric
31 composition reanalysis data in the metropolis of São Paulo, Brazil, *J. South. Hemisph. Earth*
32 *Syst. Sci.*, 75, ES24041, <https://doi.org/10.1071/ES24041>, 2025.

33 Prospero, J. M., Ginoux, P., Torres, O., Nicholson, S. E., and Gill, T. E.: Environmental
34 characterization of global sources of atmospheric soil dust identified with the Nimbus 7 Total
35 Ozone Mapping Spectrometer absorbing aerosol product, *Rev. Geophys.*, 40, 1002,
36 <https://doi.org/10.1029/2000RG000095>, 2002.

37 Randles, C. A., da Silva, A. M., Buchard, V., Colarco, P. R., Darmenov, A., Govindaraju, R.,
38 Smirnov, A., Holben, B., Ferrare, R., Hair, J., Shinozuka, Y., and Flynn, C. J.: The MERRA-2
39 aerosol reanalysis, 1980 onward, Part I: System description and data assimilation evaluation,
40 *J. Climate*, 30, 6823–6850, <https://doi.org/10.1175/JCLI-D-16-0609.1>, 2017.

41 Saidou Chaibou, A. A., Ma, X., and Sha, T.: Dust radiative forcing and its impact on surface
42 energy budget over West Africa, *Sci. Rep.*, 10, 12236, [https://doi.org/10.1038/s41598-020-](https://doi.org/10.1038/s41598-020-69223-4)
43 [69223-4](https://doi.org/10.1038/s41598-020-69223-4), 2020.

1 Schepanski, K., Heinold, B., and Tegen, I.: Harmattan, Saharan heat low, and West African
2 monsoon circulation: modulations on the Saharan dust outflow towards the North Atlantic,
3 Atmos. Chem. Phys., 17, 10223–10243, <https://doi.org/10.5194/acp-17-10223-2017>, 2017.

4 Schepanski, K., Tegen, I., Todd, M. C., Heinold, B., Bönisch, G., Laurent, B., and Macke, A.:
5 Meteorological processes forcing Saharan dust emission inferred from MSG-SEVIRI
6 observations of subdaily dust source activation and numerical models, J. Geophys. Res.-
7 Atmos., 114, D10201, <https://doi.org/10.1029/2008JD010325>, 2009.

8 Senghor, H., Machu, É., Hourdin, F., and Gaye, A. T.: Seasonal cycle of desert aerosols in
9 western Africa: Analysis of the coastal transition with passive and active sensors, Atmos.
10 Chem. Phys., 17, 8395–8410, <https://doi.org/10.5194/acp-17-8395-2017>, 2017.

11 Shao, Y.: A model for mineral dust emission, J. Geophys. Res.-Atmos., 106, 20239–20254,
12 <https://doi.org/10.1029/2001JD900171>, 2001.

13 Sheridan, S. C. and Lee, C. C.: The self-organizing map in synoptic climatological research,
14 Prog. Phys. Geogr., 35, 109–119, <https://doi.org/10.1177/0309133310397582>, 2011.

15 Solmon, F., Mallet, M., Elguindi, N., Giorgi, F., Zakey, A., and Konaré, A.: Radiative and
16 climatic effects of dust over West Africa, as simulated by a regional climate model, Clim. Res.,
17 52, 97–113, <https://doi.org/10.3354/cr01039>, 2012.

18 Tian, Z., Wei, J., and Li, Z.: How important is satellite-retrieved aerosol optical depth in
19 deriving surface PM_{2.5} using machine learning?, Remote Sens., 15, 3780,
20 <https://doi.org/10.3390/rs15153780>, 2023.

21 Tuna Tuygun, G. and Elbir, T.: Comparative analysis of CAMS aerosol optical depth data and
22 AERONET observations in the Eastern Mediterranean over 19 years, Environ. Sci. Pollut. Res.,
23 31, 27069–27084, <https://doi.org/10.1007/s11356-024-32950-6>, 2024.

24 Xian, P., Reid, J. S., Ades, M., Benedetti, A., Colarco, P. R., da Silva, A., Eck, T. F., Flemming,
25 J., Hyer, E. J., Kipling, Z., Rémy, S., Sekiyama, T. T., Tanaka, T., Yumimoto, K., and Zhang,
26 J. L.: Intercomparison of aerosol optical depths from four reanalyses and their multi-reanalysis
27 consensus, Atmos. Chem. Phys., 24, 6385–6411, <https://doi.org/10.5194/acp-24-6385-2024>,
28 2024.

29 Zhong, J., Zhang, X., Gui, K., Wang, Y., Che, H., Shen, X., Zhang, L., Zhang, Y., Sun, J., and
30 Zhang, W.: Robust prediction of hourly PM_{2.5} from meteorological data using LightGBM,
31 Natl. Sci. Rev., 8, nwaa307, <https://doi.org/10.1093/nsr/nwaa307>, 2021.