 doi:10.20944/preprints202209.0186.v1

Article

Machine Learning-Based Improvement of Aerosol Optical Depth from CHIMERE Simulations using MODIS Satellite Observations

Farouk Lemmouchi[1]*, Juan Cuesta[1], Mathieu Lachatre[2], Julien Brajard[3], Adriana Coman[1], Matthias Beekman[4] and Claude Derognat[2]

[1] Univ. Paris Est Creteil and Université Paris Cité, CNRS, LISA, 94010 Créteil, France; farouk.lemmouchi@lisa.ipsl.fr, cuesta@lisa.ipsl.fr, coman@lisa.ipsl.fr
[2] ARIA Technologies, Boulogne-Billancourt, France; mathieu.lachatre@suez.com, claude.derognat@suez.com.
[3] Nansen Environmental and Remote Sensing Center (NERSC), Bergen, Norway; julien.brajard@locean.ipsl.fr.
[4] Université de Paris Cité and Univ. Paris Est Creteil, CNRS, LISA, 75013 Paris, France; beekman@lisa.ipsl.fr.
* Correspondence: farouk.lemmouchi@lisa.ipsl.fr.

Abstract: We develop a machine learning (ML) approach for improving the accuracy of the horizontal distribution of the aerosol optical depth (AOD) simulated by the CHIMERE chemistry-transport model over Northern Africa using Moderate-Resolution Imaging Spectroradiometer (MODIS) AOD satellite observations. These observations are used during the training phase of the ML method for deriving a regional bias correction of AODs simulated by CHIMERE. The results are daily maps of regional bias corrected AODs with full horizontal coverage over Northern Africa. We test four types of ML models: multiple linear regression (MLR), random forests (RF), gradient boosting (XGB), and multiple layer perceptron networks (NN). We perform comparisons with satellite and independent ground-based observations of AOD that are not used in the training phase. They suggest that all models have overall comparable performances with a slight advantage of the RF model which expresses less spatial artifacts. While the method slightly underestimates the very high AODs, it significantly reduces biases and absolute errors, and clearly enhances linear correlations with respect to independent observations. This improvement for deriving the AOD is particularly relevant for high dust pollution regions like the Sahara Desert, which dramatically lack ground-based measurements for validations of chemistry-transport modeling which currently remains challenging and imprecise.

Keywords: mineral dust; North African dust; Saharan dust; Bodélé depression; bias correction; machine learning; aerosol optical depth; chemistry-transport model; aerosols; particulate matter

1. Introduction

Particulate matter suspended in the air, called aerosol, has major impacts on the environment. Scattering and absorption of radiation by aerosols (e.g., desert dust and black carbon) significantly modify the Earth's radiative balance, consequently affecting the climate system [1–3]. Aerosols are also the major air pollutant, they represent the greatest environmental threat for human health, causing more than 3 million premature deaths worldwide every year [4]. It's therefore of a great importance to estimate and forecast the spatial distribution and variability of aerosols and how they interact with radiation. This last aspect is described by their optical properties, such as the aerosol optical depth (AOD).

Satellite measurements, such as those derived from the Moderate-Resolution Imaging Spectroradiometer (MODIS) spaceborne sensor [5], play a fundamental role for observing the spatial distribution of aerosols at the global scale. These remote sensing observations are mainly derived in terms of AOD vertically integrated over the whole atmospheric column, which describe their horizontal distribution. However, they are available

 doi:10.20944/preprints202209.0186.v1

only for cloud free conditions and are limited by the overpass time of the (polar orbiting) satellite. On the other hand, chemistry-transport models (CTM) such as CHIMERE [6] numerically simulate the hourly 3D evolution of aerosol plumes in the atmosphere, regardless of the cloudiness. The accuracy of the CTM simulations depend on the precision of the inputs e.g., emissions of atmospheric constituents, wind, vertical velocity fields and also on the assumed aerosol properties (e.g., microphysical and optical properties). Therefore, simulations of the aerosol spatial distribution are prone to bias compared to observations [7,8]. The biases are linked to uncertainties in the physical parameterizations of the model, input data, and numerical approximations [9]. The modeling errors are more noticeable in regions lacking ground-based stations for validating and constraining the simulations, which is the case of the African continent.

Several approaches make use of AOD remote observations to assess the surface PM concentrations e.g., [10,11]. Other works combine the advantages of the spatial and temporal continuity of modeling-derived aerosol distribution and the good accuracy of the satellite observations e.g., [12–14]. Most model constraining techniques use ground *in situ* measurements and rather recently satellite data such as MODIS AOD is also used e.g., [15,16] in north african mineral dust emissions inversion and [17] for the Copernicus Atmosphere Monitoring Service (CAMS). They are mainly based on data assimilation techniques such as variational or filtering approaches [18] which are costly in terms of computation. On the other hand, new approaches use machine learning (ML) techniques for correcting model systematic biases with respect to observations, which are quick and computationally efficient. These methods are increasingly being used thanks to the advancement in the development of ML hardware and technology. Until now, these ML bias correction techniques are mainly used for chemistry-transport model simulations of trace gases e.g., [19] or using in situ surface data for aerosols e.g., [20,21], or for post-processing forecasts e.g., [22,23].

In this work, we develop a new ML-based bias correction for correcting the bias of the CHIMERE simulated AOD maps at the wavelength of 550 nm at 13:00 LT (local time). The method is applied in a post-processing manner. The correction relies on the good accuracy of AOD satellite measurements from MODIS taken at the same wavelength and approximately the same time. Observations are only used in the training phase, so the method provides full coverage, daily maps of the corrected AODs over the whole CHIMERE simulation domain. The method is developed for North Africa using data from the year 2021 that we detail in section 2. We discuss in section 3 the performance of four different ML models: multiple linear regression, neural networks, random forest, and gradient boost model against independent ground-based and satellite observations.

2. Materials and Methods

The principle of the AOD bias correction method is to use AOD and other variables such as meteorological fields simulated by CHIMERE as inputs of a trained ML-based model. The latter derives *a posteriori* bias corrected AODs that are in better agreement with AOD observations from MODIS than those simulated by the model, which are called hereafter "raw" CHIMERE AOD. This is achieved by "teaching" the model how to correct the AOD using MODIS measurements during a supervised training stage. A schematic description of this process is shown in Figure 1. The inference of AOD, also called predictions in other contexts is done in a pixel-by-pixel fashion for all the ground pixels, no surrounding regional information is used for the AOD correction. In the following subsections, we describe the dataset used for the training (section 2.1.1 & 2.1.2), the data preparation (section 2.1.3) and then the ML implementation for correcting AOD regional biases (section 2.2).

To do so, we assume that the AOD is a function of the input features, and we try to approximate the relationship that maps the features inputs to the correct AOD using statistical modeling. Statistical models are data driven i.e. require a large database of known solutions that covers a wide range of possible situations.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 September 2022 doi:10.20944/preprints202209.0186.v1

Figure 1. Flowchart of the pixel-by-pixel bias correction approach of the CHIMERE AOD estimation.

2.1 Inputs

2.1.1 MODIS satellite observations

The satellite observations used for training the ML model are derived from the MODIS spaceborne multi-wavelength radiometer onboard the AQUA platform with an overpass time around 13:30 LT. This satellite follows a near-polar sun-synchronous low orbit, hovering at 705 km of altitude and within the A-Train constellation. MODIS is observing structures at 36 spectral bands at a horizontal resolution that ranges from 250 to 1000 m and a nadir swath 2330 km wide, covering the majority of the Earth's surface on a daily basis [24]. The MODIS AOD product used in this work is the collection 6.1 MYD04_L2 with 10 km resolution [25,26]. This product is the combination of AOD products derived with the Dark Target [27] and Deep Blue [28] algorithms. The Dark Target algorithm is suitable over ocean and dark land (e.g., vegetation), while the Deep Blue algorithm covers the entire land areas including both dark and bright surfaces. This MODIS dataset is interpolated at the horizontal resolution of the CHIMERE model (0.45°x0.45°) using the mean value method. All MODIS AODs mentioned hereafter refers to the observations from the AQUA platform, at the wavelength of 550 nm, unless mentioned otherwise.

The quality of the reference dataset for supervision is critical for accurate ML modeling; therefore, we filter out the low-quality observations by only keeping retrievals with the highest quality assurance flag. We verify the quality of the MODIS AODs used here by comparison (Figure 2a) with daily averages of direct measurements of the AOD from ground-based sun photometers of the AErosol RObotic NETwork [29]. This network imposes standardization of instruments, calibration and processing which assure the best quality. The comparison (Figure 2a) shows that the AOD observations of MODIS for the year 2021 correlates well with the measurements of the eight AERONET stations (the location of each station in Figure 3). The Pearson correlation coefficient (r) is 0.88 and the root mean squared error (RMSE) is rather low (0.13). The AERONET data being used for the comparison are of level 1.5; we convert the AOD from 500 nm to 550 nm using the 440-675 nm Angstrom exponent provided by the same station.

2.1.2 CHIMERE simulations

CHIMERE is an Eulerian CTM that simulates the formation, deposition, and transport of aerosols and other atmospheric species [30]. It is capable of simulating phenomena from the local scale e.g., urban heat island up to hemispherical scale. The principle is to use the available information about the earth's atmospheric composition, and the earth surface including source emissions, then simulate the evolution of the atmospheric species taking into account the internal forcing e.g., wind, and external forcing e.g., shortwave incoming radiation. The computation of the evolution of the plumes is done by numerically solving the transport equation and using a chemistry interaction scheme. CHIMERE is used as a tool to forecast and analyze the daily air quality in terms of particulate and gaseous pollution [31]. It is widely used for aerosols, and aerosol precursors research e.g., [32–36].

The CHIMERE simulations used as inputs for the ML model are derived for the 12 months of the year 2021 using CHIMERE-2017 version [37] over the region of North Africa and Arabian peninsula (10 to 38°N 19°W to 53°E). The horizontal resolution is 0.45° x 0.45°

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 September 2022 doi:10.20944/preprints202209.0186.v1

on a regular grid; while the vertical resolution is composed of 20 layers of increasing thickness, from approximately 30 m at the surface level to 675 m (upper limit around 500 hPa). Aerosols are distributed into 10 size-classes (also denoted as bins), from 0.05 μm to 40 μm. The AOD is calculated for all the aerosol species considered in the simulation under an external mixing hypothesis. This is calculated using online Fast-JX (version 7.0b) photolysis computations [38] providing the optical properties [39] and the aerosols concentration. The computation is done at five wavelengths (200, 300, 400, 600, and 999 nm); in this work we use the interpolated 400 nm AOD to 550 nm using the 400-600 nm Angstrom exponent.

The simulation boundary and initial conditions are taken from the Laboratoire de Météorologie Dynamique general circulation model coupled with the Interaction with Chemistry and Aerosols (LMDz-INCA) [40]. The CHIMERE simulation is run in offline mode using National Oceanic and Atmospheric Administration (NOAA) meteorological final analysis data and Weather Research and Forecasting Model (WRF) version 4.1.1 [41]. MELCHIOR2 (Modèle Lagrangien de Chimie de l'Ozone à l'échelle Régionale) [42] is used as chemistry interaction scheme. A 2015 based emission inventory is used from the Emissions Database for Global Atmospheric Research EDGARv5.0 [43] for the gaseous and particulate matter anthropogenic emissions. The dust emission scheme implemented in CHIMERE requires the knowledge of soil properties and wind conditions. This module allows for the computation of dust aerosols emissions and their size distribution, by modeling the processes of sandblasting and saltation [44,45]. It takes into account uplifts of both silty and sandy soils, which are emitted from the north Algerian region and the Sahara respectively [46]; but a non-negligible uncertainty remains in the modeling of Saharan dust emission and transport [47].

An evaluation of the accuracy of this AOD simulated by CHIMERE with respect to AERONET reference measurements is shown in Figure 2b. It shows that for high pollution cases, the CHIMERE AOD is clearly overestimated with respect to the AOD measured by the AERONET sun photometers the mean bias (MB) is equal to 0.16, but the inverse is observed in the background AODs. The correlation coefficient is less than for the comparison of MODIS with AERONET AOD (0.63, compared to 0.88 for MODIS). The RMSE is greater, being 0.46 for CHIMERE AOD against 0.13 for MODIS. The same behavior is noticed for the mean absolute error (MAE). Other chemistry-transport models have also shown biases in the AOD simulated in north africa, with comparable overestimation e.g., [48].

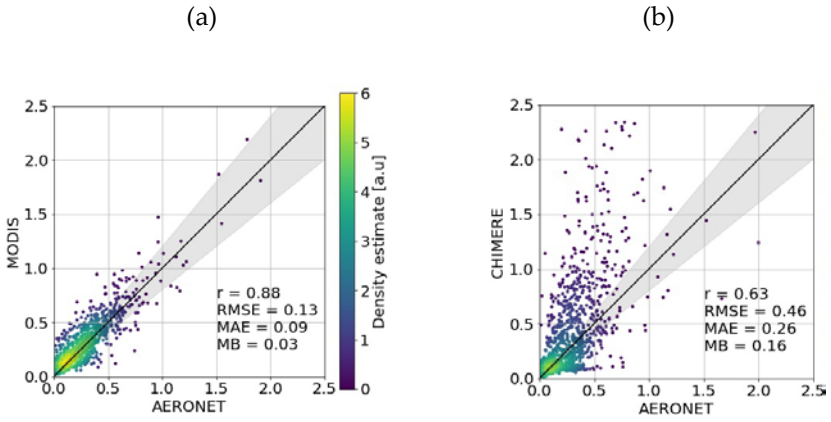

Figure 2. Scatter plot of 966 collocated data points comparing the AOD550 measured by 8 AERONET sun photometers for the year 2021 with (a) MODIS satellite observations and (b) simulated AODs by the CHIMERE chemistry-

 doi:10.20944/preprints202209.0186.v1

transport model. The black line is the y=x line with the gray shaded areas representing the ±20% interval of its slope.

The map of the annual median difference between the AOD from CHIMERE and MODIS (Figure 3) shows that the major overestimation is located over most of the Saharan desert. The positive bias of CHIMERE AOD is most pronounced over a region northeast of the lake of Tchad named the Bodélé depression, where half of the days from 2021 express a bias larger than unity. Mineral dust is frequently emitted from this region (Supplementary data, Figure S1b), around 100 times per year [49]. Yearly emissions from this single spot (17.8°E; 16.9°N) go up to 40 million tons of dust. The large positive biases of raw CHIMERE simulations are possibly linked to the uncertainties in the inputs of the dust emission model e.g., near surface wind speed and friction velocity, but it can also be due to the refractive index used in the AOD computation.

On the contrary, over the Arabian Peninsula and Northern Egypt, the AOD is slightly underestimated by CHIMERE, as compared to MODIS. We see a low correlation (0.53) between the two estimations, and a high standard deviation with an RMSE of 0.62 and a MAE of 0.35.

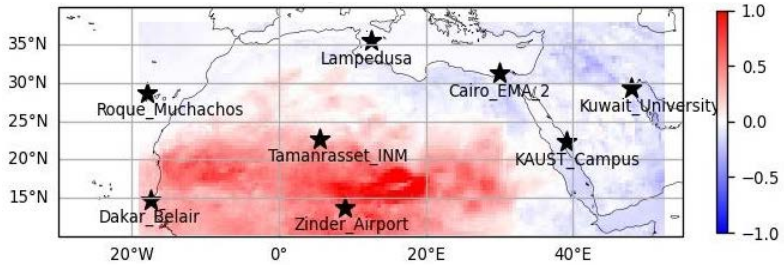

Figure 3. The median of the difference between CHIMERE and MODIS AOD_{550} for the year 2021 in the simulation domain of north Africa. The stars show the 8 AERONET stations used for the comparisons in Figure 2.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 September 2022 doi:10.20944/preprints202209.0186.v1

The bias of CHIMERE raw AOD simulation output with respect to MODIS AOD is assessed for different raw AOD intensities (Figure 4) for the whole year. We clearly see the higher the AOD of CHIMERE the more overestimating it is; the linear fit has a slope of 0.79 and the intercept value is low (-0.2). The linear correlation of the bias with the raw AOD is the highest correlation compared to other CHIMERE variables like PM25, PM10 (not shown). The exact reason for this linear relationship is unknown, an investigation on the cause of the bias is under consideration. But the method that we describe in this paper is a ML model designed for improving these biases *a posteriori*.

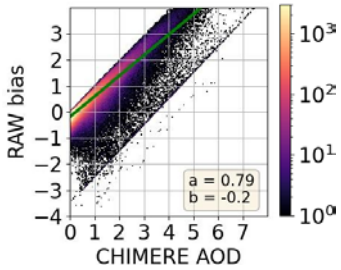

Figure 4. Scatterplot showing the linear relationship between CHIMERE AOD_{550} bias toward MODIS observations for different AOD_{550} levels represented by the CHIMERE AOD_{550}. The colorbar represents the occurrence in the year 2021. The green line represents the linear fit with the correlation coefficient a and the intercept b.

2.1.3 Dataset preparation

We selected the days used for the training following the common convention in the machine learning field; i.e. splitting the whole 12 months of 2021: 66% is used in the training and validation stage, and the other 33% is kept for testing the performance of the models.

The training data size and diversity is important for a good generalization of a ML model. In order to avoid the bias induced due to the seasonal variability, we take two thirds of each month of 2021 for building the models, i.e. days from 1st to 20th for the training and validation, which make up 240 days from 2021. We call this dataset D_{Train} and it is composed of around 1.4 million cloud free ground pixels. Then from the 21st to the end of each month for testing the models, which makes up 122 days that are not used in the training process, we call this dataset D_{Test} and it is composed of around 0.7 million CHIMERE simulation ground pixels. Dataset D_{Train} is constituted of sequences of consecutive yet shuffled data, the correlation between a training subset and a validation subset is limited due to the decorrelation time scale of the atmospheric processes. This is a key point to ensure that the test dataset is independent and that the scores computed are robust to evaluate the generalization skill of the model. We excluded the days 23, 24, and 25 of September 2021 because of the unavailability of MODIS observations.

Besides the simulated AOD, the local variables related to the atmospheric composition taken from CHIMERE which are used as inputs of the ML models are the following: vertical concentration profiles of dust, organic carbon, particulate matter (PM_{10} and $PM_{2.5}$ with diameter smaller than respectively 10 and 2.5 µm), carbon monoxide (CO), ammonia (NH_3), toluene, ozone (O_3), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), nitrous acid (HNO), sulfuric acid (H_2SO_4), specific humidity, water droplets, sea salt, nitric acid (HNO_3), ROOH, non-methane hydrocarbons (NMHC), OH, and the two compound families NO_x and NO_y. For these vertical profiles, we consider 4 simulation levels out of the 20 available, corresponding to approximately 967, 920, 797 and 560 hPa pressure levels. The other CHIMERE outputs used as inputs for the ML are related to surface properties and meteorological conditions: surface albedo, shortwave radiation flux, soil moisture, surface relative humidity, boundary layer height, surface latent heat flux, surface sensible heat flux, pressure profile, and relative humidity profile. We don't include wind fields because they do not improve the residuals. All CHIMERE and meteorological variables are taken at 13:00 LT which is the closest to the AQUA platform spacecraft overpass time. Therefore, D_{Train} and D_{Test} have a total of 96 features.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 September 2022 doi:10.20944/preprints202209.0186.v1

These variables used as inputs are selected such that each variable should be related directly or indirectly to the aerosol concentration. Choosing a subset of the available variables speeds up computations, mitigates the overfitting and reduces the curse of dimensionality [50]. We keep variables even if they show significant correlations between them e.g., PM_{10} & $PM_{2.5}$, because we find that this slightly improves the performance of the bias correction method. Although most features of D_{Train}, as well as D_{Test} follow a log normal distribution (e.g., PM), the use of the normal logarithm of these variables as inputs of the ML degrades its performances by inducing underestimations of the predicted AODs. Preprocessing by applying standardization/rescaling, and/or outlier removal does not significantly improve the accuracy either. Therefore, we keep the input data, without any transformation.

2.2 Bias correction ML models construction

The fitting problem we try to solve is overdetermined and the high dimensional input variables are not independent. We implement several machine learning approaches, tree based, a neural networks model, and the more classical linear regression model. We discuss their performances in section 3. In the following, we describe how the four models are built (section 2.3.1 to 2.3.4) using python version 3.8 programming language [51] and the web-based interactive computational environment jupyter notebook [52].

2.2.1 Multiple linear regression (MLR)

Multiple linear regression modeling is widely used in several fields thanks to its convenience and its ease of use and interpretation. While it is suitable when the variables are independent, we tested it here to serve as a performance baseline and compare it with the three other more sophisticated models. The training is performed on half of D_{Train} data that have been randomly sampled. We use only half of D_{Train} to keep the comparison fair between the other models which use some data from the other half during a tuning phase.

If **F** is the function that maps the input features X to the outcome denoted y, our objective is to find a function **G** that approximates **F** using a set of known solutions $\{X_i,y_i\}_i^N$. Usually, the function is learned successively by minimizing the expected value of some loss function L(y, G(X)). The residuals are assumed normally distributed and have a constant variance. In light of these assumptions and that **F** is linear, we approximate it as the following equation (1) :

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon \quad (1)$$

where ε represents the error associated with the approximation, β_i the coefficient corresponding to the feature p_i, and n the number of variables in the feature space which equals 96 in our case.

The cost function $\boldsymbol{L}$ (equations 2 and 3) below represents the residuals from the model and the observation is solved using the least squares method.

$$L = \sum_{j=1}^{N} (y_j - y'_j)^2 \quad (2)$$

$$= \sum_{j=1}^{N} (y_j - \beta_0 - \sum_{k=1}^{N} \beta_k X_{j,k})^2 \quad (3)$$

where N is the number of the observations available. For this work, we use the linear regression model from sklearn v1.0.2 library [53].

2.2.2 Feed-forward neural networks (NN)

Neural networks regression is a modeling technique that is increasingly being used in many fields thanks to the advancement of big data and parallelized hardware. The principle is to use perceptrons stacked in layers to approximate a nonlinear relationship [54], where the weight of each perception is learned from a reference dataset by gradient backpropagation [50,55,56]. The technique can model non-linear relationships by using some activation functions [57]. Optimizing multidimensional functions with neural networks is hard because of the proliferation of saddle points [58], but stochastic gradient descent algorithms are good at finding the global minimum [59–61].

 doi:10.20944/preprints202209.0186.v1

The training of the NN AOD bias corrector is done using the Tensorflow library v2.5.0 [62]. First, we search for a suitable network architecture for the problem by performing 100 trials of feed-forward neural networks with dense layers. The architectures are sampled from an arbitrary reasonable hyperparameters search space (Supplementary data, Table 1). We use the random search feature of the Keras-tuner 1.0.3 library [63] to find the best network. All the tested networks use batch normalized input feed [64], and a rectified linear unit [65] as activation function. The use of a dropout layer [66,67] worsens the accuracy so we do not use it here. The optimization of the weights of the network is done using Adam optimizer [68] –which is a variation of stochastic gradient descent– with respect to the mean squared error. Training is run for 100 epochs, we set early stopping in the conditions when the validation loss either exceeds 0.1, or stabilizes at 0.01. We notice that the increase in the number of neurons (perceptrons) does not improve the accuracy of the estimator, and we manage to get the same performance of the best randomly generated NN model using a bottleneck architecture, with less complexity and fewer parameters (Supplementary data, Table 2). The network has 3943 trainable parameters.

2.2.3 Random forest (RF)

A random forest regressor [69] is an ensemble model that is widely used in both classification and regression problems. Each ensemble member called tree [70] is constructed using different randomized parameters; the final prediction is the average of all trees predictions.

The model is tuned for three key hyperparameters: number of estimators (trees), minimum samples in a leaf, and maximum number of features. The tuning is done using the brute-force search method with two cross-fold validations on 10000 randomly selected ground pixel data from D_{Train}. Using only a subset of the available training dataset accelerates the search process and saves memory. We find out that the best performing RF candidate has a maximum of 20 features, minimum samples in a leaf 4, and 100 estimators. After that, we perform training of this best model using half of D_{Train}. The training was performed using sklearn v1.0.2.

2.2.4 Gradient boosting (XGB)

A gradient boosting regressor [71,72] is also a tree-based modeling technique but constructs trees successively in a way that each tree minimizes the preceding trees residuals. The final estimation is the sum of all the trees predictions.

We use the Tree-based Pipeline Optimization Tool (TPOT) library v0.11.7. It is a genetic programming-based automated ML system that optimizes a series of feature preprocessors and ML models to build a better data processing pipeline for supervised tasks [73]. We let TPOT determine the best pipeline for correcting the CHIMERE AOD using 10000 randomly selected data from D_{Train}. The number of generations is set empirically to 20 and the population size is set to 20, increasing further these numbers slows the search process. The best pipeline suggested by TPOT uses XGBoost model (XGB) [72] which is a variation of the gradient boost model that uses L1 and L2 regularization, preceded by the zeros and non-zeros counts as features. The post-processor XGB is made of 100 estimators, with a maximum depth of 9 and a minimum child weight of 17 (see the documentation [74] for more details about the meaning of those hyperparameters). The best pipeline suggested by TPOT is retrained using half of D_{Train} dataset that were randomly sampled.

3. Results and discussion

This section presents an evaluation of the performance of the trained AOD bias correctors, with respect to D_{Test} MODIS observations (section 3.1) and ground based AERONET measurements (section 3.2). While the training is performed partially using a Quadro P620 GPU, the prediction runs only on a single thread of a CPU (i9 2.30GHz). The time cost of the training and the inference of the models is a key indication of the usability of the developed method. The training of the ML models did not take a long time, in fact the MLR takes a few seconds, and NN, RF, and XGB training does not exceed one hour each

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 September 2022 doi:10.20944/preprints202209.0186.v1

(without counting the hyperparameter search step which is much longer). Concerning the inference time, all approaches perform the daily correction in less than half a second (Table 1) which allows for a potential use of these models in real time, we note a slightly quicker inference for MLR and NN.

3.1 Comparison against independent MODIS observations

We compare the estimations of the AOD from raw CHIMERE and the inferences from the ML models with respect to MODIS observations that are not used for their training (Table 1). The table shows that the ML models corrected successfully the raw CHIMERE AODs to a certain extent. They all show a comparable improvement of the AOD correction, with both larger correlation coefficients with MODIS AOD (between 0.62 to 0.71) and smaller RMSE (from 0.19 to 0.21) and MAE. The improvement in AOD estimation is noticeable looking at the raw CHIMERE AODs performance (r of 0.56 and RMSE of 0.65). We note that the XGB and RF are the best-performing bias correctors in terms of correlation coefficient (0.71) and RMSE (0.19). The NN model follows with the same RMSE but slightly lower correlation coefficient (0.69). Lastly, the MLR corrector offers a slightly lower correlation coefficient (0.62) and higher RMSE (0.21).

Table 1. Average performance of each AOD bias corrector and raw CHIMERE on D_{Test} with respect to MODIS AOD. t: daily mean inference time cost; r: Pearson correlation coefficient; RMSE: Root Mean Squared Error; MAE: Mean Absolute Error; Skp: Pearson's Coefficient of Skewness; μ: mean; percentages show the percentiles of the error. The metrics are calculated for the 737129 pixels making D_{Test}.

	t(s)	**r**	**RMSE**	**MAE**	**Skp**	**μ**	**min**	**25%**	**50%**	**75%**	**max**
RAW	N/A	0.56	0.65	0.37	2.55	0.24	-3.57	-0.09	0.03	0.39	6.95
MLR	0.19	0.62	0.21	0.13	-3.9	0	-4.15	-0.06	0.03	0.1	2.49
NN	0.35	0.69	0.19	0.12	-3.18	0	-4.04	-0.06	0.02	0.09	5.09
RF	0.22	0.71	0.19	0.12	-3.45	0.01	-4.21	-0.05	0.03	0.1	2
XGB	0.3	0.71	0.19	0.12	-2.93	0.01	-3.96	-0.06	0.02	0.09	2.47

The AOD correction biases distribution –the AOD estimation bias in this work being the difference between the modeled/estimated value and the MODIS observed value– show that they are bell shaped centered around a small positive value, suggesting an overall positive overcorrection of the raw AOD (Figure 5). This is consistent with the positive values of the medians (50% percentile) and the negative skewness in Table 1. The means on the other hand are near or equal to zero, which is expected since the models are trained by minimizing the mean squared error.

The model's biases spreads are similar as well. The most underestimated AOD value is from the RF model with a minimum bias of -4.21, and the least underestimating model is the XGB with a minimum bias value of -3.96. The maximum bias values on the other hand, which correspond to the extreme underestimation of the AOD is from the NN model with a maximum bias of 5.09; while the best model in terms of underestimation is the RF model with a maximum bias of 2.

We observe that the distribution of the raw AOD biases of CHIMERE (Figure 5) are skewed towards higher values, this is because more background AOD pixels are available which are underestimated by CHIMERE, while in the high AOD episodes, the raw CHIMERE AOD overestimate it. The overestimation occurs in most of the dust emission episodes with very large bias values as shown in Figure 3, which is why the biases RMSE is high, and median bias is positive (0.03) in Table 1.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 September 2022 doi:10.20944/preprints202209.0186.v1

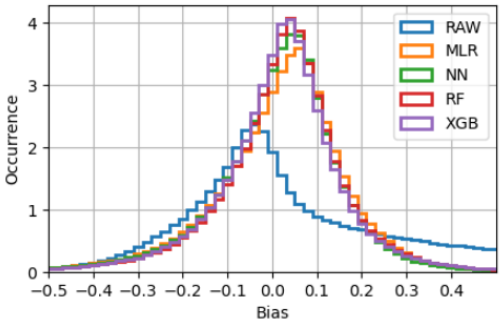

Figure 5. Histogram showing the frequency per 0.02 bin size bias, which is the difference between modeled (raw CHIMERE) or corrected (ML-models) AOD_{550} and satellite observed AOD_{550} from MODIS for D_{Test}.

The most important feature from the training vector is the CHIMERE raw AOD as suggested by the RF and XGB models (not shown). So an analysis of the correction versus the a priori AOD can indicate how impactful the a priori AOD in terms of intensities on the predicted AOD. The bias in the AOD correction of the different ML models is found to be only slightly dependent on the *a priori* (raw) AOD levels (Figure 6). We observe for all of them that in high *a priori* AOD (above 0.5 value), the correction bias can be high, while in low *a priories* the AOD correction is quite good and expresses low bias typically under 0.3 in most of the pixels. This may be attributed to the fact that more pixels available for training have a typical low background AOD level.

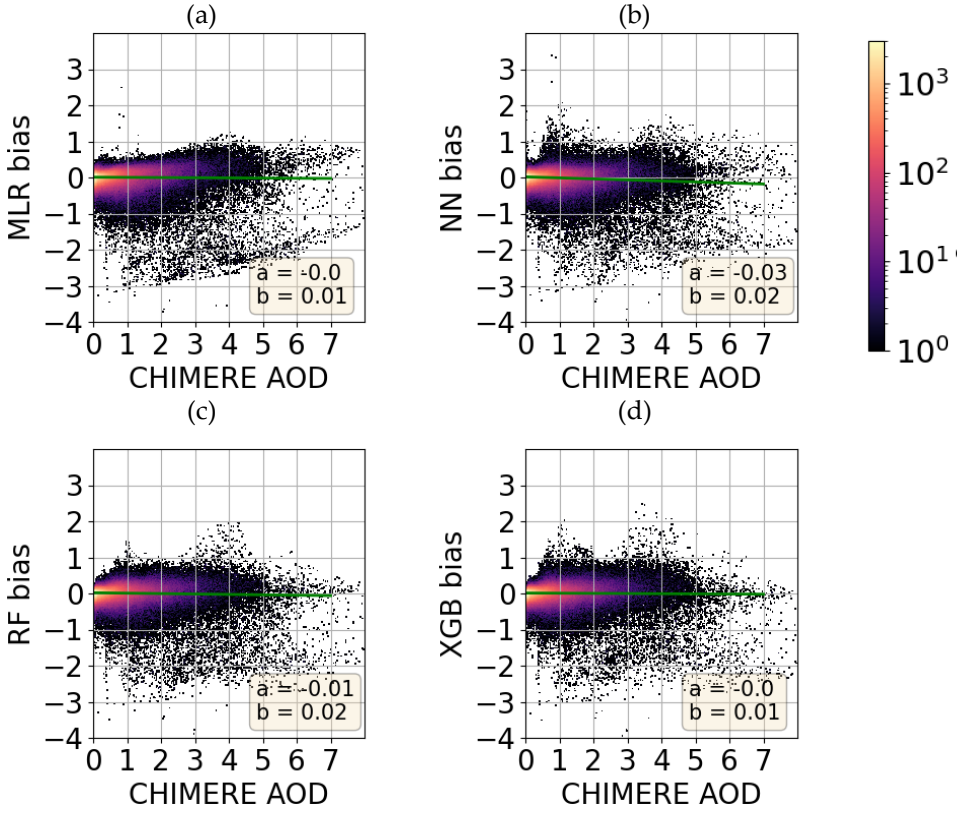

Figure 6. Scatterplots showing the AOD correction bias amplitudes of (a) MLR, (b) NN, (c) RF and (d) MLR in terms of different *a priori* AOD intensities represented by the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 September 2022 doi:10.20944/preprints202209.0186.v1

CHIMERE AOD_{550} for D_{Test}. The green line represents the linear fit with a slope a and the intercept b. Colormap represents the occurrence.

The analysis of the predicted AOD of the different models varies according to the AOD level (Figure 7). In low AOD levels (under 1) the models correct to an acceptable degree the AOD of CHIMERE most of the time as shown with the yellow colormap, with a slight undercorrection for high AOD levels. The MLR model (Figure 7a) expresses some negative AOD pixels, this is an artifact (see also Figure 8c) associated with the use of negative regression coefficients, which improves the overall performance of the MLR model. The best performer among the ML models is the XGB (Figure 7d) with a linear fit slope of 0.52 followed by the RF model.

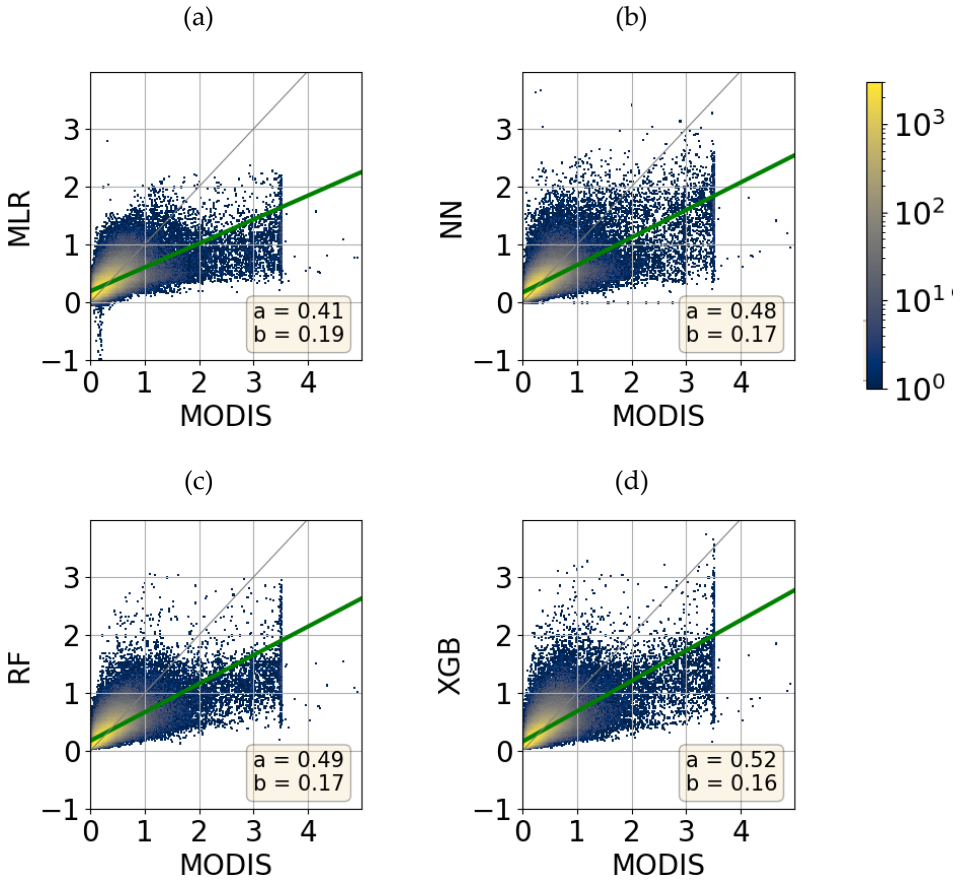

Figure 7. Scatterplots showing the corrected AOD_{550} of (a) MLR, (b) NN, (c) RF and (d) MLR in terms of different AOD levels represented by the MODIS AOD_{550} for D_{Test}. The green line represents the linear fit with a slope a and the intercept b. The gray lines represent the lines $y=x$. Colormap represents the occurrence.

The bias corrected AOD of CHIMERE predicted with the four ML models (Figure 8c-f) are compared with MODIS AOD measurements (Figure 8a) and with CHIMERE raw AOD simulation (Figure 8b). It is worth noting that a clear added-value of the corrected AOD maps is the full geographical coverage of this estimation. They are not affected by the cloud cover nor by the satellite swath size, while offering a better accuracy than the raw CHIMERE simulations. The ML models successfully correct the AOD, by reducing the raw CHIMERE values in the high emission regions such as the Sahara and by increasing the low AOD values over the Northern Arabian Peninsula for better matching MODIS measurements. Despite the independent pixel-by-pixel processing, the horizontal structures of the corrected AOD features are continuous and homogeneous in most of the re-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 September 2022 doi:10.20944/preprints202209.0186.v1

gions. This is found for most of the dates of D_{Test} (see data availability clause). Some exceptions are encountered such as some discontinuities seen near the southwest coast of North Africa on 30 September 2021 for the NN and MLR predictions (Figure 8c-d). Discontinuities in the AOD are observed that do not reflect the natural concentration gradient found in dispersed dust plumes. In the case of MLR (Figure 8c), the issue also seen previously in (Figure 7a) is associated with the use of negative regression coefficients. For the NN model, the issue could be explained by the fact that multiple layer perceptrons may not extrapolate well the non-linear relationships outside the training set [75]. XGB and RF bias correctors do not show such artifacts (Figure 5e-f). Yet we notice that XGB is a bit noisier, suggesting an overfit to the MODIS AOD pixels noise. In fact, [69] suggests that using a random selection of features to split each node in tree based models yields error rates that compare favorably to adaptive boosting (Adaboost), but are more robust with respect to noise.

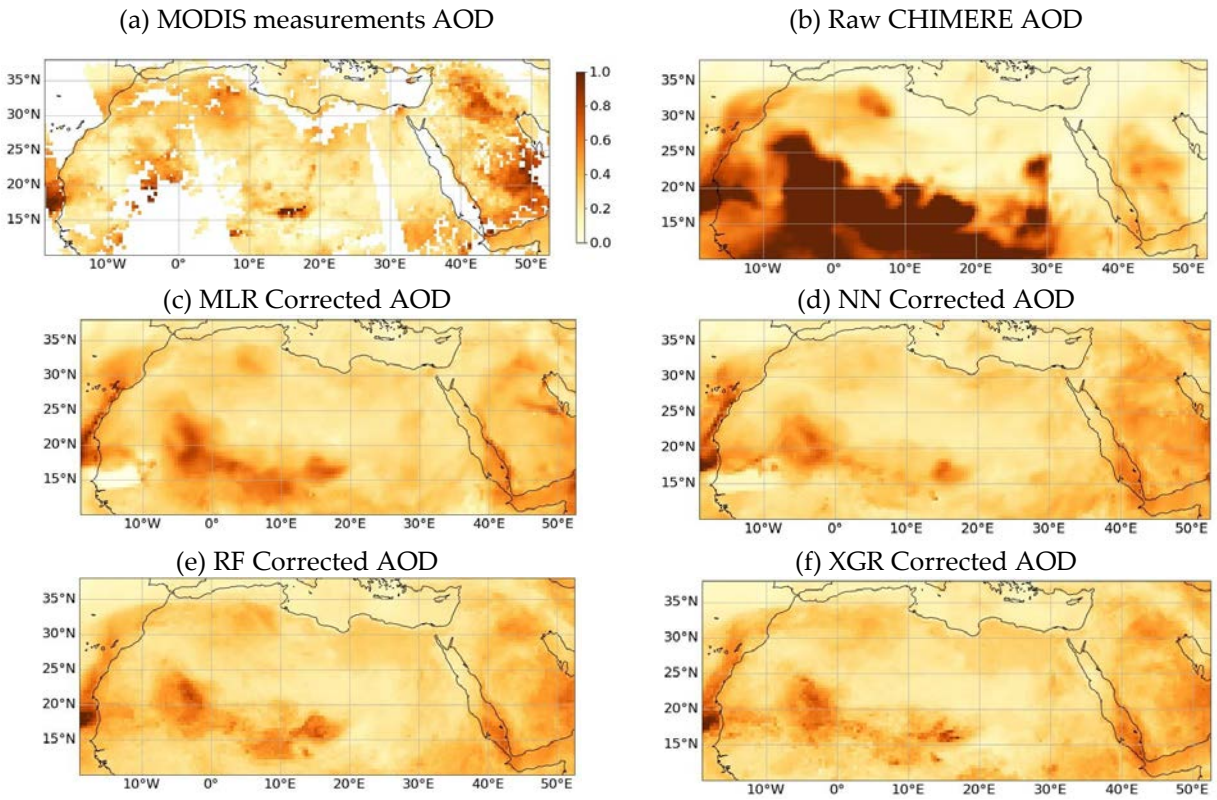

Figure 8. Horizontal distribution of the AOD_{550} on 30 September 2021 (not used training) from (a) MODIS satellite observations, (b) raw CHIMERE simulation, (c) MLR, (d) NN, (e) RF, (f) XGB. The AOD predictions of the rest of the days are available from the link found in the data availability clause.

 doi:10.20944/preprints202209.0186.v1

Figure 9 shows an evaluation of the performance of the four bias correctors in terms of the median bias of the AOD correction over the whole testing dataset, D_{Test}. The median bias does not exceed 0.3 in most of the domain, and it is clearly smaller than that of raw CHIMERE simulations (Figure 3). Similar biases are seen for the four correctors, with some positive, and negative biases for some regions characterized by high annual mean AOD values (supplementary data, Figure S1.b), and they are smaller in absolute terms for RF and XGB particularly over the Arabian Peninsula and the southwestern part of the Sahara. A common behavior for all the ML models is that they highly underestimate the AOD over a zone (10° to 20°E, 14° to 20°N) corresponding to the Bodélé depression (Figure 9).

The Bodélé depression region is a known major source of desert dust [49]. The region is part of what was the paleolake Megachad which was the biggest lake on the planet 7000 years ago [76]. Mineral dust emissions from the Bodélé depression region are generally fine sized dominated by quartz, with admixtures of clay minerals and Fe-oxyhydrates [77]. Therefore, they are brighter (higher reflectance) and whiter (flat reflectance spectrum), compared to the rest of the Sahara plumes [78] (supplementary data, Figure S1.a). Indeed, Algerian Sahara dust [79] and the Bodélé depression region dust [80] have different refractive indices. Unlike the CHIMERE CTM model, the aerosol retrieval algorithm Deep Blue uses two different single scattering albedo for the dust AOD retrieval [78], one for the redder dust, and one for the whiter dust. Furthermore, an upper limit of 3.5 is set for the AOD values (Figure 10) [25].

Improving the accuracy of near surface wind speeds, the representation of particle size and the aerosol optical properties within CHIMERE may reduce the bias in the AOD simulated in North Africa, but this is beyond the scope of this paper.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 September 2022 doi:10.20944/preprints202209.0186.v1

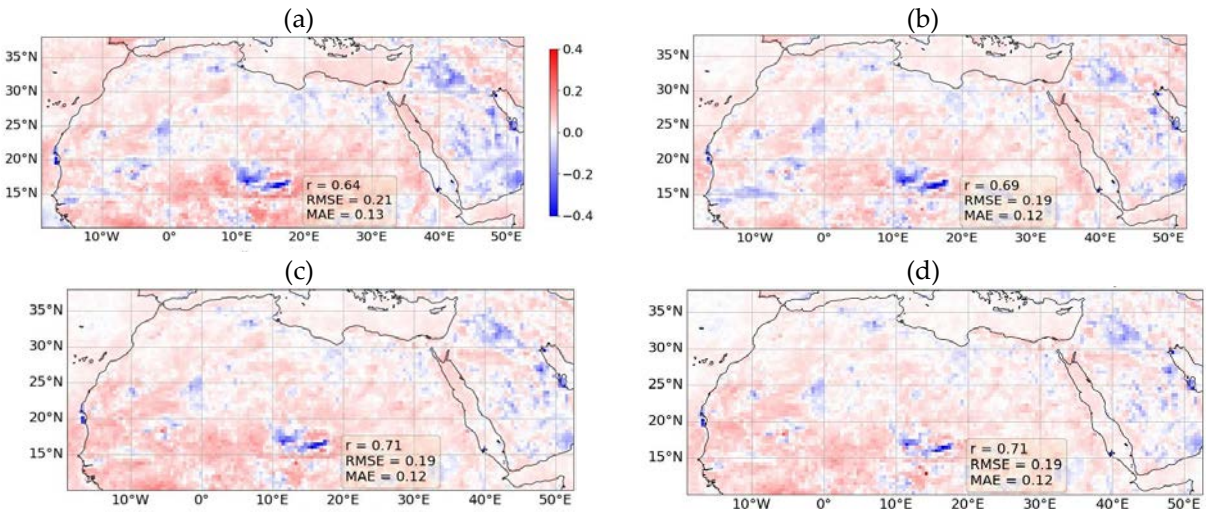

Figure 9. The median of the difference between CHIMERE bias corrected AOD_{550} and MODIS AOD_{550} for the dates of D_{Test} for the (a) MLR, (b) NN, (c) RF, and (d) XGB ML models.

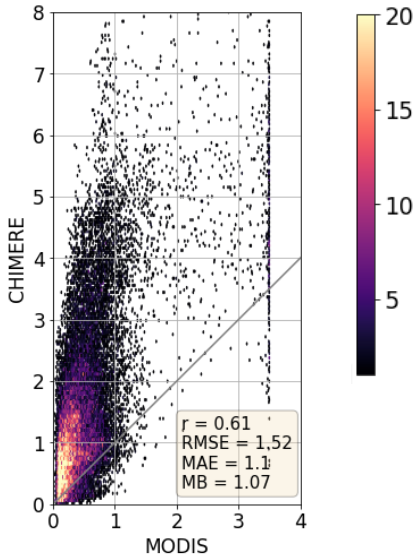

Figure 10. Scatter plot of MODIS and raw CHIMERE AODs in the region of Bodélé depression. Colormap indicates occurrence.

As the AOD correction models are trained using MODIS Aqua data, one need to ensure that the models are temporally stable and usable outside the Aqua overpass time (13:30 LT). For this purpose, we correct the raw CHIMERE AOD outputs at 10:00 LT using the RF model, then compare the inferred AODs to MODIS observations from the satellite Terra which has a local overpass time of 10:30. We choose only the RF model to perform the upcoming analysis because the statistical performance of the other ML models is similar. The comparison shows that the correlation and residuals improvement is comparable to the corrected AODs at 13:00. From Table 2, we see that the residuals (RMSE, MAE, MB) are reduced, e.g., a reduction of 0.68% is observed in the RMSE; the correlation is increased to 0.68 from 0.52. This good AOD correction for a morning hour implies the possibility to use the correction method for more daytime simulation hours, yet we remain conservative

 doi:10.20944/preprints202209.0186.v1

on the use of the AOD correction method during night hours as the ML model will more often have to extrapolate to unseen situations e.g., very low boundary layer height, low temperature, etc.

Table 2. Correlation and residuals of AOD estimations for the year 2021 against MODIS Terra observation. The raw AODs are CHIMERE outputs at 10:00 LT and for the corrected AOD with the RF model. Number of data 3.7 million.

	r	RMSE	MAE	MB
RAW	0.52	0.59	0.34	-0.23
RF-corrected	0.68	0.19	0.12	-0.03

3.2 Comparison with AERONET stations

The performance of the RF model along with that of MODIS is assessed against collocated AERONET measurements in the 8 stations for testing dates (Table 3). We see an improvement of r from 0.54 for the raw AOD to 0.73. The residuals also improved as the RMSE and MAE dropped from 0.45 and 0.27 to around 0.16 and 0.12 respectively. The MB is significantly reduced from 0.18 to around 0.6 for the corrected AOD. Figure 11 provides an illustration of the daily evolution and temporal consistency of the performance of the RF model, note that we also include the days used in the training phase, but we will restrict the analysis on the testing dates. We chose the station of Cairo for this purpose because of its good data coverage of the period, and because it shows situations of positive and negative corrections of the biases (the same figure for the other sites can be found at the link in the data availability clause). A suite of high AOD episodes are seen in March and April and also from 24 to 29 October 2021. In those cases, raw AODs simulated by CHIMERE are clearly overestimated with respect to AERONET measurements. The ML-based correctors reduce the AOD values for clearly approaching the sun photometer measurements, while still depicting the AOD peaks. On the other hand, CHIMERE simulations underestimate the AOD in the periods of 20-27 June and 21 to 30 September. In those cases, the bias corrector increased the AOD to reach the appropriate values close to both AERONET and MODIS measurements. In cases when raw CHIMERE AODs are correct (e.g., on 20-21 May), the bias corrector does not modify the AOD, remaining around 0.24.

Table 3. Correlation and residuals of 330 AOD estimations on testing dates that are collocated with the 8 AERONET stations.

	r	RMSE	MAE	MB
MODIS	0.85	0.12	0.09	0.03
RAW	0.54	0.45	0.27	0.18
RF	0.73	0.16	0.12	0.06

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 September 2022 doi:10.20944/preprints202209.0186.v1

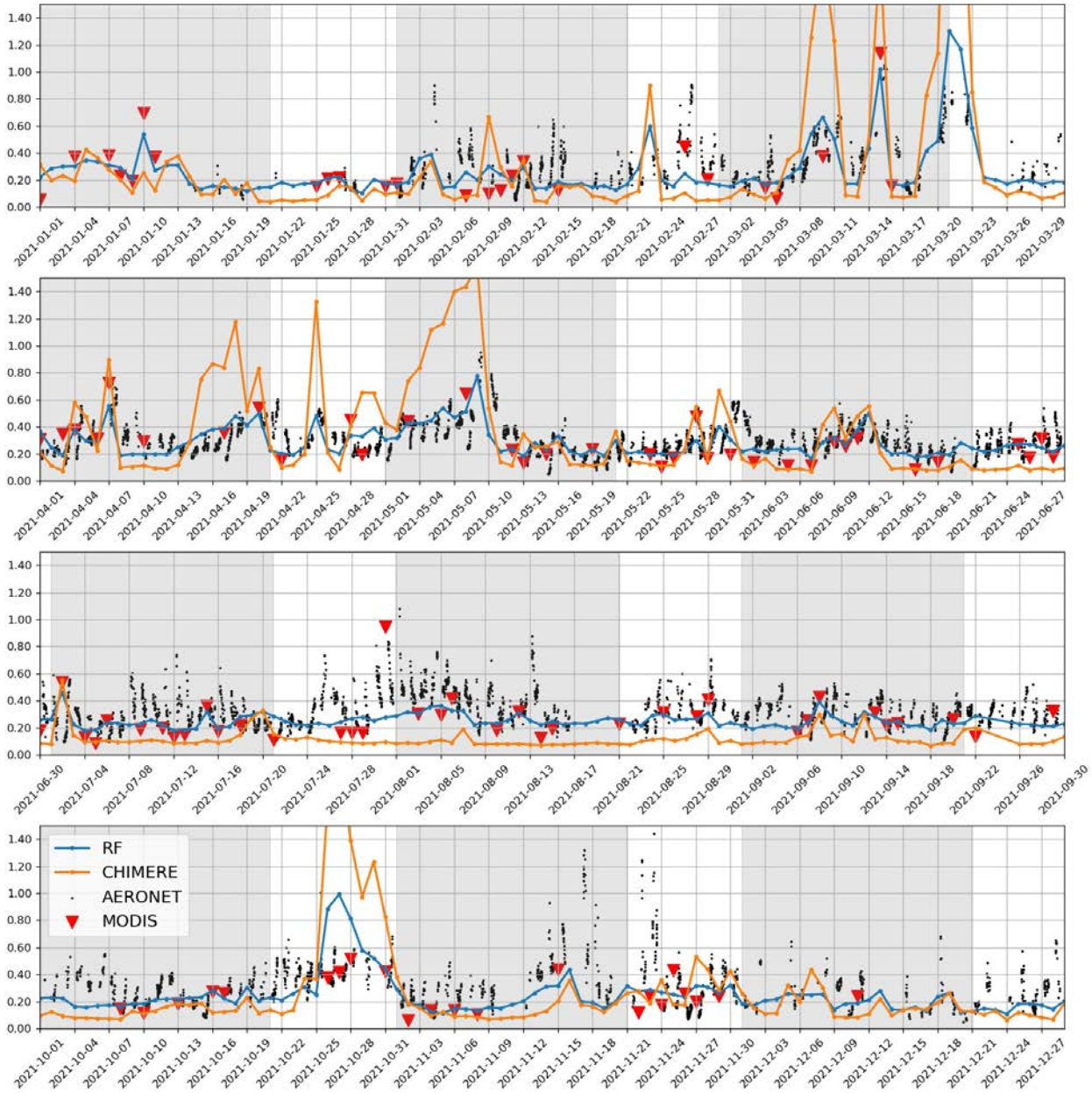

Figure 11. Time series of AOD_{550} in Cairo during the year 2021 from AERONET ground-based sun photometer measurements, raw CHIMERE simulations, ML-corrected AODs and MODIS. The gray areas represent dates used for training the models.

4. Conclusion and perspectives

In this work, we develop a new ML-based model for correcting *a posteriori* the biases of CHIMERE chemistry-transport model simulations of the AOD, by training it with MODIS satellite measurements. This approach provides full coverage daily maps of AODs over North Africa with a clearly better agreement with satellite and ground-based observations than the raw CHIMERE simulations. AODs corrected by the ML models show substantially higher correlations and lower errors (RMSE and MAE) than the raw AOD simulations, as compared to MODIS satellite measurements. The RMSE of the AODs is reduced from 0.65 (raw CHIMERE) to 0.19 (RF bias corrector) and the correlation coefficient r increased from 0.56 to 0.71 (respectively for raw CHIMERE and the RF corrector). The bias corrector reduces the overestimations in the AOD maps over the Saharan desert and the underestimations over the Arabian Peninsula, yet a slight over correction is observed in low pollution regions.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 September 2022 doi:10.20944/preprints202209.0186.v1

Out of the four ML-based bias correctors, the best results are found for the RF regressor which gives a spatially smooth AOD in good agreement with MODIS AODs. The other bias correctors used here (MLR, XGB, and NN) also increase the accuracy of the AOD, but showed spatial artifacts, some of them linked to overfitting. The AOD correction was tested for a time of the day other that the one used for the training. The comparison of the CHIMERE AOD correction at 10:00 LT against that of MODIS Terra which has a close overpass time (10:30 LT), show that the RF correction expresses as good results as against MODIS Aqua (used in the training). This is a step towards validating the AOD correction for a usage at any day time of the CHIMERE simulation raw outputs.

We also notice at high AOD value; the models underestimate the AOD compared to MODIS observation. Evaluation of the daily consistency of the corrected AOD shows that overestimation of peaks and underestimated background values at a given location (shown for Cairo) have been successfully improved.

As the ML models are trained using analysis data, the AOD correction applied on forecasts will probably be less accurate. An adaptation of the ML models for handling forecasts is in our perspective. Furthermore, a deep understanding on the different features contribution will help improve the temporal stability and the spatial artifacts.

We think that the results of our method could_potentially be used as a gap filling system for the global MODIS AOD maps. Additionally, further development will be conducted for correcting the CHIMERE vertical profile of the aerosols concentration using state of the art retrievals of the 3D distribution of aerosols, such as AEROIASI [81] which uses the Infrared Atmospheric Sounding Interferometer (IASI) thermal infrared measurements for coarse particles; AEROS5P [82] which uses TROPospheric Ozone Monitoring Instrument (TROPOMI) measurements for fine particles.

Supplementary Materials: The following supporting information can be downloaded at: www.mdpi.com/xxx/s1.

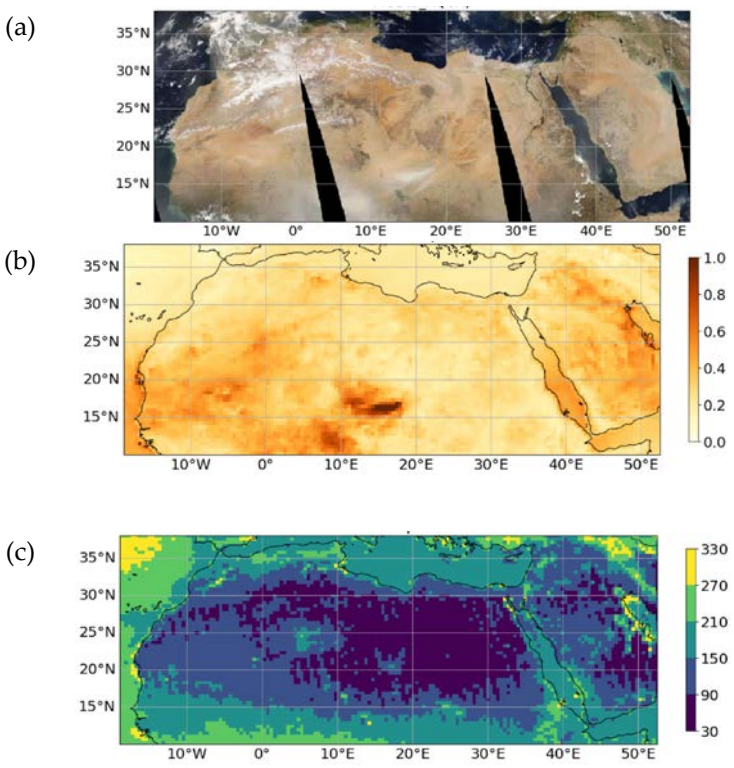

Figure S1. (a) MODIS true color showing the mineral dust emitted from Bodélé depression region on 23 March 2021. (b) Average

 doi:10.20944/preprints202209.0186.v1

MODIS AOD_{550} values for 2021. (c) Number of observations used in the computation of the average AOD_{550} of panel (b)

Table 1. The network architecture search space

Hyperparameter	Minimum value	Maximum value	Increment
Number of layers	10	50	10
Number of perceptrons	10	50	10
Batch size	5000	20000	1000

Table 2 The best architecture for the NN regressor.

Layer type	Output shape	Number of parameters
Batch normalization	96	384
Dense	30	2910
Dense	20	620
Dense	10	210
Dense	1	11

Author Contributions: Conceptualization, formal analysis, methodology, data curation, software, visualization, validation, writing—original draft, FL; conceptualization, funding acquisition, project administration, supervision, writing—review and editing, JC; data curation, writing—review and editing, ML; methodology, writing—review and editing, JB; methodology, writing—review, AC; writing—review and editing, MB; funding acquisition, writing —review, CD; All authors participated in the analysis and have read and agreed to the published version of the manuscript. Please turn to the CRediT taxonomy for the term explanation.

Funding: This work is funded by the Region Ile-de-France in the framework of the Domaine d'Intérêt Majeur Réseau de recherche Qualité de l'air en Ile-de-France (DIM QI2) through the program Paris Region PhD (PRPHD), and ARIA Technologies, and supported by the *Centre National des Etudes Spatiales* (CNES) through the SURVEYPOLLUTION project from TOSCA (*Terre Ocean Surface Continental et Atmosphère*), the *Centre National de Recherche Scientifique—Institute National de Sciences de l'Univers* (CNRS-INSU) and the *Université Paris Est Créteil* (UPEC).

Data Availability Statement: In this link https://bit.ly/3bj35Kr several materials can be found: additional AOD maps figures, the data used for the development and the results of this work, the trained ML models, and a code used for the visualization of some of the figures.

Acknowledgments: This work is part of the Make Our Planet Great Again initiative through the project Make Air Quality Great Again (MAQGA) lead by R. Subramanian within the EFLUVE OSU (Observatoire de Sciences de l'Univers). We acknowledge the free use of MODIS data from NASA's Earth Observing System Data and Information System (EOSDIS) via Worldview application. We acknowledge the free download of data from 8 AErosol RObotic NETwork (AERONET) stations through the online platform. We thank the developers of the python libraries used in this work.

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

References

1. Tsikerdekis, A.; Zanis, P.; Georgoulias, A.K.; Alexandri, G.; Katragkou, E.; Karacostas, T.; Solmon, F. Direct and Semi-Direct Radiative Effect of North African Dust in Present and Future Regional Climate Simulations. *Climate Dynamics* 2019, *53*, 4311–4336.
2. Mahowald, N.M.; Baker, A.R.; Bergametti, G.; Brooks, N.; Duce, R.A.; Jickells, T.D.; Kubilay, N.; Prospero, J.M.; Tegen, I. Atmospheric Global Dust Cycle and Iron Inputs to the Ocean. *Global Biogeochem. Cycles* **2005**, *19*, doi:10.1029/2004gb002402.
3. Klingmüller, K.; Lelieveld, J.; Karydis, V.A.; Stenchikov, G.L. Direct Radiative Effect of Dust–Pollution Interactions. *Atmospheric Chemistry and Physics* 2019, *19*, 7397–7408.
4. Lelieveld, J.; Evans, J.S.; Fnais, M.; Giannadaki, D.; Pozzer, A. The Contribution of Outdoor Air Pollution Sources to Premature Mortality on a Global Scale. *Nature* **2015**, *525*, 367–371.
5. Remer, L.A.; Kaufman, Y.J.; Tanré, D.; Mattoo, S.; Chu, D.A.; Martins, J.V.; Li, R.-R.; Ichoku, C.; Levy, R.C.; Kleidman, R.G.; et al. The MODIS Aerosol Algorithm, Products, and Validation. *J. Atmos. Sci.* **2005**, *62*, 947–973.
6. Menut, L.; Bessagnet, B.; Briant, R.; Cholakian, A.; Couvidat, F.; Mailler, S.; Pennel, R.; Siour, G.; Tuccella, P.; Turquety, S.; et al. The CHIMERE V2020r1 Online Chemistry-Transport Model. *Geoscientific Model Development* 2021, *14*, 6781–6811.
7. Bessagnet, B.; Hodzic, A.; Vautard, R.; Beekmann, M.; Cheinet, S.; Honoré, C.; Liousse, C.; Rouil, L. Aerosol Modeling with CHIMERE—Preliminary Evaluation at the Continental Scale. *Atmospheric Environment* 2004, *38*, 2803–2817.
8. Turquety, S.; Menut, L.; Siour, G.; Mailler, S.; Hadji-Lazaro, J.; George, M.; Clerbaux, C.; Hurtmans, D.; Coheur, P.-F. APIFLAME v2.0 Biomass Burning Emissions Model: Impact of Refined Input Parameters on Atmospheric Concentration in Portugal in Summer 2016. *Geoscientific Model Development* 2020, *13*, 2981–3009.
9. Mallet, V.; Sportisse, B. Uncertainty in a Chemistry-Transport Model Due to Physical Parameterizations and Numerical Approximations: An Ensemble Approach Applied to Ozone Modeling. *Journal of Geophysical Research* 2006, *111*.
10. Maheshwarkar, P.; Sunder Raman, R. Population Exposure across Central India to PM2.5 Derived Using Remotely Sensed Products in a Three-Stage Statistical Model. *Sci. Rep.* **2021**, *11*, 544.
11. Gupta, P.; Christopher, S.A. Particulate Matter Air Quality Assessment Using Integrated Surface, Satellite, and Meteorological Products: 2. A Neural Network Approach. *J. Geophys. Res.* **2009**, *114*, doi:10.1029/2008jd011497.
12. Amraoui, L.E.; El Amraoui, L.; Sič, B.; Piacentini, A.; Marécal, V.; Attié, J.-L.; Frebourg, N. Aerosol Data Assimilation in the Chemical Transport Model MOCAGE during the TRAQA/ChArMEx Campaign: Lidar Observations 2019.
13. Sič, B.; El Amraoui, L.; Piacentini, A.; Marécal, V.; Emili, E.; Cariolle, D.; Prather, M.; Attié, J.-L. Aerosol Data Assimilation in the Chemical Transport Model MOCAGE during the TRAQA/ChArMEx Campaign: Aerosol Optical Depth. *Atmospheric Measurement Techniques* 2016, *9*, 5535–5554.
14. Lee, D.; Byun, D.W.; Kim, H.; Ngan, F.; Kim, S.; Lee, C.; Cho, C. Improved CMAQ Predictions of Particulate Matter Utilizing the Satellite-Derived Aerosol Optical Depth. *Atmospheric Environment* 2011, *45*, 3730–3741.
15. Escribano, J.; Boucher, O.; Chevallier, F.; Huneeus, N. Impact of the Choice of the Satellite Aerosol Optical Depth Product in a Sub-Regional Dust Emission Inversion. *Atmospheric Chemistry and Physics* 2017, *17*, 7111–7126.
16. Escribano, J.; Boucher, O.; Chevallier, F.; Huneeus, N. Subregional Inversion of North African Dust Sources. *J. Geophys. Res. D: Atmos.* **2016**, *121*, 8549–8566.
17. Garrigues; Remy; Chimot; Ades Monitoring Multiple Satellite Aerosol Optical Depth (AOD) Products within the Copernicus Atmosphere Monitoring Service (CAMS) Data Assimilation System. *Atmos. Clim. Sci.* **2022**.
18. Bocquet, M.; Elbern, H.; Eskes, H.; Hirtl, M.; Žabkar, R.; Carmichael, G.R.; Flemming, J.; Inness, A.; Pagowski, M.; Pérez Camaño, J.L.; et al. Data Assimilation in Atmospheric Chemistry Models: Current Status and Future Prospects for Coupled Chemistry Meteorology Models. *Atmospheric Chemistry and Physics* 2015, *15*, 5325–5358.
19. Sayeed, A.; Eslami, E.; Lops, Y.; Choi, Y. CMAQ-CNN: A New-Generation of Post-Processing Techniques for Chemical Transport Models Using Deep Neural Networks. *Atmospheric Environment* 2022, *273*, 118961.
20. Xu, M.; Jin, J.; Wang, G.; Segers, A.; Deng, T.; Lin, H.X. Machine Learning Based Bias Correction for Numerical Chemical Transport Models. *Atmos. Environ.* **2021**, *248*, 118022.
21. Jin; Lin; Segers; Xie Machine Learning for Observation Bias Correction with Application to Dust Storm Data Assimilation. *Chem. Phys. Lipids* **2019**.
22. Rasp, S.; Lerch, S. Neural Networks for Postprocessing Ensemble Weather Forecasts. *Monthly Weather Review* 2018, *146*, 3885–3900.
23. Taillardat, M.; Mestre, O.; Zamo, M.; Naveau, P. Calibrated Ensemble Forecasts Using Quantile Regression Forests and Ensemble Model Output Statistics. *Mon. Weather Rev.* **2016**, *144*, 2375–2393.
24. Https://modis.gsfc.nasa.gov Available online: https://modis.gsfc.nasa.gov/about/specifications.php (accessed on 27 July 2022).
25. Sayer, A.M.; Munchak, L.A.; Hsu, N.C.; Levy, R.C.; Bettenhausen, C.; Jeong, M.-J. MODIS Collection 6 Aerosol Products: Comparison between Aqua's e-Deep Blue, Dark Target, and "Merged" Data Sets, and Usage Recommendations. *J. Geophys. Res.* **2014**, *119*, 13,965-13,989.
26. Wei, J.; Li, Z.; Peng, Y.; Sun, L. MODIS Collection 6.1 Aerosol Optical Depth Products over Land and Ocean: Validation and Comparison. *Atmos. Environ.* **2019**, *201*, 428–440.
27. Gupta; Levy; Mattoo; Remer A Surface Reflectance Scheme for Retrieving Aerosol Optical Depth over Urban Surfaces in MODIS Dark Target Retrieval Algorithm. *Atmos. Clim. Sci.* **2016**.

28. Hsu, N.C.; Jeong, M.-J.; Bettenhausen, C.; Sayer, A.M.; Hansell, R.; Seftor, C.S.; Huang, J.; Tsay, S.-C. Enhanced Deep Blue Aerosol Retrieval Algorithm: The Second Generation. *J. Geophys. Res.* **2013**, *118*, 9296–9315.
29. Holben, B.N.; Eck, T.F.; Slutsker, I.; Tanré, D.; Buis, J.P.; Setzer, A.; Vermote, E.; Reagan, J.A.; Kaufman, Y.J.; Nakajima, T.; et al. AERONET—A Federated Instrument Network and Data Archive for Aerosol Characterization. *Remote Sens. Environ.* **1998**, *66*, 1–16.
30. Menut, L.; Bessagnet, B.; Khvorostyanov, D.; Beekmann, M.; Blond, N.; Colette, A.; Coll, I.; Curci, G.; Foret, G.; Hodzic, A.; et al. CHIMERE 2013: A Model for Regional Atmospheric Composition Modelling. *Geoscientific Model Development* **2013**, *6*, 981–1028.
31. Http://www.prevair.org/ Available online: http://www.prevair.org/ (accessed on 27 July 2022).
32. Ciarelli, G.; Theobald, M.R.; Vivanco, M.G.; Beekmann, M.; Aas, W.; Andersson, C.; Bergström, R.; Manders-Groot, A.; Couvidat, F.; Mircea, M.; et al. Trends of Inorganic and Organic Aerosols and Precursor Gases in Europe: Insights from the EURODELTA Multi-Model Experiment over the 1990--2010 Period, Geosci. Model Dev., 12, 4923--4954 2019.
33. Lachatre, M.; Foret, G.; Laurent, B.; Siour, G.; Cuesta, J.; Dufour, G.; Meng, F.; Tang, W.; Zhang, Q.; Beekmann, M. Air Quality Degradation by Mineral Dust over Beijing, Chengdu and Shanghai Chinese Megacities. *Atmosphere* **2020**, *11*, 708.
34. Cholakian; Beekmann; Colette Simulation of Fine Organic Aerosols in the Western Mediterranean Area during the ChArMEx 2013 Summer Campaign. *Atmos. Clim. Sci.* **2018**.
35. Deroubaix; Flamant; Menut Interactions of Atmospheric Gases and Aerosols with the Monsoon Dynamics over the Sudano-Guinean Region during AMMA. *Atmos. Clim. Sci.* **2018**.
36. Fortems-Cheiney, A.; Dufour, G.; Foret, G.; Siour, G.; Van Damme, M.; Coheur, P.-F.; Clarisse, L.; Clerbaux, C.; Beekmann, M. Understanding the Simulated Ammonia Increasing Trend from 2008 to 2015 over Europe with CHIMERE and Comparison with IASI Observations. *Atmosphere* **2022**, *13*, 1101.
37. Mailler, S.; Menut, L.; Khvorostyanov, D.; Valari, M.; Couvidat, F.; Siour, G.; Turquety, S.; Briant, R.; Tuccella, P.; Bessagnet, B.; et al. CHIMERE-2017: From Urban to Hemispheric Chemistry-Transport Modeling. *Geoscientific Model Development* 2017, *10*, 2397–2423.
38. Bian, H.; Prather, M.J. Fast-J2: Accurate Simulation of Stratospheric Photolysis in Global Chemical Models. *J. Atmos. Chem.* **2002**, *41*, 281–296.
39. Péré, J.C.; Mallet, M.; Pont, V.; Bessagnet, B. Evaluation of an Aerosol Optical Scheme in the Chemistry-Transport Model CHIMERE. *Atmospheric Environment* 2010, *44*, 3688–3699.
40. Hauglustaine, D.A.; Hourdin, F.; Jourdain, L.; Filiberti, M.-A.; Walters, S.; Lamarque, J.-F.; Holland, E.A. Interactive Chemistry in the Laboratoire de Météorologie Dynamique General Circulation Model: Description and Background Tropospheric Chemistry Evaluation. *J. Geophys. Res. D: Atmos.* **2004**, *109*, doi:10.1029/2003JD003957.
41. Skamarock, C.; Klemp, B.; Dudhia, J.; Gill, O.; Liu, Z.; Berner, J.; Wang, W.; Powers, G.; Duda, G.; Barker, D.; et al. A Description of the Advanced Research WRF Model Version 4.1. **2019**, doi:10.5065/1dfh-6p97.
42. Derognat, C. Effect of Biogenic Volatile Organic Compound Emissions on Tropospheric Chemistry during the Atmospheric Pollution Over the Paris Area (ESQUIF) Campaign in the Ile-de-France Region. *Journal of Geophysical Research* 2003, *108*.
43. Oreggioni, G.D.; Monforti-Ferraio, F.; Crippa, M.; Schaaf, E.; Guizzardi, D.; Muntean, M.; Duerr, M.; Vignati, E. EDGAR v5.0: A Tool to Evaluate the Influence of Technology Incorporation and Regulatory Frameworks on Global Greenhouse Gases and Air Pollutant Emissions 2020.
44. Alfaro, S.C.; Gomes, L. Modeling Mineral Aerosol Production by Wind Erosion: Emission Intensities and Aerosol Size Distributions in Source Areas. *Journal of Geophysical Research: Atmospheres* 2001, *106*, 18075–18084.
45. Menut, L.; Schmechtig, C.; Marticorena, B. Sensitivity of the Sandblasting Flux Calculations to the Soil Size Distribution Accuracy. *J. Atmos. Ocean. Technol.* **2005**, *22*, 1875–1884.
46. Gama, C.; Ribeiro, I.; Lange, A.C.; Vogel, A.; Ascenso, A.; Seixas, V.; Elbern, H.; Borrego, C.; Friese, E.; Monteiro, A. Performance Assessment of CHIMERE and EURAD-IM' Dust Modules. *Atmospheric Pollution Research* 2019, *10*, 1336–1346.
47. Menut, L.; Chiapello, I.; Moulin, C. Previsibility of Saharan Dust Events Using the CHIMERE-DUST Transport Model. *IOP Conference Series: Earth and Environmental Science* 2009, *7*, 012009.
48. Abdoul Aziz Saidou Chaibou, Xiaoyan Ma, Kanike Raghavendra Kumar, Hailing Jia, Yaoguo Tang, Tong Sha Evaluation of Dust Extinction and Vertical Profiles Simulated by WRF-Chem with CALIPSO and AERONET over North Africa. *J. Atmos. Sol. Terr. Phys.* **2020**, *199*, 105213.
49. Washington, R.; Bouet, C.; Cautenet, G.; Mackenzie, E.; Ashpole, I.; Engelstaedter, S.; Lizcano, G.; Henderson, G.M.; Schepanski, K.; Tegen, I. Dust as a Tipping Element: The Bodélé Depression, Chad. *Proceedings of the National Academy of Sciences* **2009**, *106*, 20564–20571.
50. Bittner, L. R. Bellman, Adaptive Control Processes. A Guided Tour. XVI 255 S. Princeton, N. J., 1961. Princeton University Press. Preis Geb. $ 6.50. *ZAMM - Zeitschrift für Angewandte Mathematik und Mechanik* 1962, *42*, 364–365.
51. Https://www.python.org/ Available online: https://www.python.org/ (accessed on 22 July 2022).
52. Https://jupyterbook.org (accessed on 22 July 2022).
53. Pedregosa; Varoquaux; Gramfort Scikit-Learn: Machine Learning in Python. *of machine Learning …* **2011**.
54. Bengio, Y. Learning Deep Architectures for AI. *Foundations and Trends® in Machine Learning* **2009**, *2*, 1–127.
55. Rumelhart, D.E.; Hinton, G.E.; Williams, R.J. Learning Representations by Back-Propagating Errors. *Nature* 1986, *323*, 533–536.
56. MacKay, D.J.C. A Practical Bayesian Framework for Backpropagation Networks. *Neural Comput.* **1992**, *4*, 448–472.

57. Dubey, S.R.; Singh, S.K.; Chaudhuri, B.B. Activation Functions in Deep Learning: A Comprehensive Survey and Benchmark. *Neurocomputing* **2022**, *503*, 92–108.
58. Dauphin; Pascanu; Gulcehre Identifying and Attacking the Saddle Point Problem in High-Dimensional Non-Convex Optimization. *Adv. Neural Inf. Process. Syst.*
59. Goodfellow, I.J.; Vinyals, O.; Saxe, A.M. Qualitatively Characterizing Neural Network Optimization Problems. *arXiv [cs.NE]* 2014.
60. Zhou, Y.; Yang, J.; Zhang, H.; Liang, Y.; Tarokh, V. SGD Converges to Global Minimum in Deep Learning via Star-Convex Path. *arXiv [cs.LG]* 2019.
61. Du, S.; Lee, J.; Li, H.; Wang, L.; Zhai, X. Gradient Descent Finds Global Minima of Deep Neural Networks. In Proceedings of the Proceedings of the 36th International Conference on Machine Learning; Chaudhuri, K., Salakhutdinov, R., Eds.; PMLR, 09--15 Jun 2019; Vol. 97, pp. 1675–1685.
62. Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G.S.; Davis, A.; Dean, J.; Devin, M.; et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *arXiv [cs.DC]* 2016.
63. O'Malley, T.; Bursztein, E.; Long, J.; Chollet, F.; Jin, H.; Invernizzi, L.; Others Keras Tuner. *Retrieved May* **2019**, *21*, 2020.
64. Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In Proceedings of the Proceedings of the 32nd International Conference on Machine Learning; Bach, F., Blei, D., Eds.; PMLR: Lille, France, 07--09 Jul 2015; Vol. 37, pp. 448–456.
65. Fukushima, K. Cognitron: A Self-Organizing Multilayered Neural Network. *Biol. Cybern.* **1975**, *20*, 121–136.
66. Hinton, G.E.; Srivastava, N.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R.R. Improving Neural Networks by Preventing Co-Adaptation of Feature Detectors. *arXiv [cs.NE]* 2012.
67. Srivastava; Hinton; Krizhevsky Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *The journal of machine*.
68. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv [cs.LG]* 2014.
69. Breiman, L. Machine Learning 2000, *40*, 229–242.
70. Gordon, A.D.; Breiman, L.; Friedman, J.H.; Olshen, R.A.; Stone, C.J. Classification and Regression Trees. *Biometrics* **1984**, *40*, 874.
71. Friedman, J.H. Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics* 2001, *29*.
72. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; Association for Computing Machinery: New York, NY, USA, August 13 2016; pp. 785–794.
73. Olson, R.S.; Moore, J.H. TPOT: A Tree-Based Pipeline Optimization Tool for Automating Machine Learning. In Proceedings of the Proceedings of the Workshop on Automatic Machine Learning; Hutter, F., Kotthoff, L., Vanschoren, J., Eds.; PMLR: New York, New York, USA, June 24 2016; Vol. 64, pp. 66–74.
74. Https://Xgboost.Readthedocs.Io/En/Stable/Parameter.Html Available online: https://xgboost.readthedocs.io/en/stable/parameter.html (accessed on 2 August 2022).
75. Xu, K.; Zhang, M.; Li, J.; Du, S.S.; Kawarabayashi, K.-I.; Jegelka, S. How Neural Networks Extrapolate: From Feedforward to Graph Neural Networks. *arXiv [cs.LG]* 2020.
76. Drake, N.; Bristow, C. Shorelines in the Sahara: Geomorphological Evidence for an Enhanced Monsoon from Palaeolake Megachad. *Holocene* **2006**, *16*, 901–911.
77. Chudnovsky, A.; Kostinski, A.; Herrmann, L.; Koren, I.; Nutesku, G.; Ben-Dor, E. Hyperspectral Spaceborne Imaging of Dust-Laden Flows: Anatomy of Saharan Dust Storm from the Bodélé Depression. *Remote Sens. Environ.* **2011**, *115*, 1013–1024.
78. Hsu, N.C.; Tsay, S.-C.; King, M.D.; Herman, J.R. Aerosol Properties over Bright-Reflecting Source Regions. *IEEE Transactions on Geoscience and Remote Sensing* 2004, *42*, 557–569.
79. Rocha-Lima; Martins; Remer A Detailed Characterization of the Saharan Dust Collected during the Fennec Campaign in 2011: In Situ Ground-Based and Laboratory Measurements. *Atmos. Clim. Sci.* **2018**.
80. Todd, M.C.; Washington, R.; Martins, J.V.; Dubovik, O.; Lizcano, G.; M'Bainayel, S.; Engelstaedter, S. Mineral Dust Emission from the Bodélé Depression, Northern Chad, during BoDEx 2005. *J. Geophys. Res.* **2007**, *112*, doi:10.1029/2006jd007170.
81. Cuesta, J.; Eremenko, M.; Flamant, C.; Dufour, G.; Laurent, B.; Bergametti, G.; Höpfner, M.; Orphal, J.; Zhou, D. Three-Dimensional Distribution of a Major Desert Dust Outbreak over East Asia in March 2008 Derived from IASI Satellite Observations. *Journal of Geophysical Research: Atmospheres* 2015, *120*, 7099–7127.
82. Lemmouchi, F.; Cuesta, J.; Eremenko, M.; Derognat, C.; Siour, G.; Dufour, G.; Sellitto, P.; Turquety, S.; Tran, D.; Liu, X.; et al. Three-Dimensional Distribution of Biomass Burning Aerosols from Australian Wildfires Observed by TROPOMI Satellite Observations. *Remote Sensing* 2022, *14*, 2582.