EEET ECOLOGICAL ENGINEERING & ENVIRONMENTAL TECHNOLOGY

Ecological Engineering & Environmental Technology, 2026, 27(8), 147–160
https://doi.org/10.12912/27197050/225181
ISSN 2719–7050, License CC-BY 4.0

Received: 2026.06.17
Accepted: 2026.07.10
Published: 2026.08.01

Development of physically interpretable principal component regression models for estimating dust pollution over Baghdad City

Safaa A. Alkinani[1,2*], Ali Mohammed Al-Salihi[2], Jasim M. Rajab[2]

[1] Scientific Research Commission, Baghdad, Iraq
[2] Department of Atmospheric Sciences, College of Science, Mustansiriyah University, Baghdad, Iraq
* Corresponding author's e-mail: safaaabdulrazzaq@uomustansiriyah.edu.iq

ABSTRACT

The aerosol optical depth retrieved from a satellite is basically provides an inexpensive choice for regular observing of particulate matter (PM) concentration, but there are challenges present in estimation accuracy. This study developed a physically interpretable principal component analysis (PCA) – regression to estimate PM2.5-related dust column mass density (DCMD-PM2.5) and surface PM2.5 levels in Baghdad using the MERRA-2 aerosol and ERA5 reanalysis data from 2014 to 2025. The meteorological data collected from ERA5 consisted of: temperature, dew point, wind components, cloud cover, precipitation, radiation, boundary-layer, and aerosol-related variables from MERRA-2; aerosol extinction, black carbon surface mass concentration, and scattering indicators. These variables were checked to assess their impacts on DCMD-PM2.5 and surface PM2.5 values. Varimax rotation identified five PCs Baghdad, explaining 75.63% of the total variance and representing thermal-radiative mixing, cloud dynamics, aerosol optical loading, precipitation-related removal, and moisture conditions. The PC-based regression models were reconstructed using representative driver variables: temperature, high cloud cover, aerosol optical depth (AOD), total precipitation, and dew point. During model development, the PC-based DCMD-PM2.5 model explained 80.6% of the variance, compared with 53.4% for the surface PM2.5 model. Independent validation showed stronger performance for DCMD-PM2.5, with R = 0.799, RMSE = 4.82×10^{-5} kg m^{-2}, MAE = 3.86×10^{-5} kg m^{-2}, and a very small negative bias. The surface PM2.5 model showed weaker but still informative performance, with R = 0.448, RMSE = 25.29 µg m^{-3}, MAE = 20.27 µg m^{-3}, and slight positive bias. The results indicate that DCMD-PM2.5 and surface PM2.5 are related but physically distinct aerosol indicators. The proposed framework provides a transparent approach for representing reanalysis-based dust and PM2.5 variability in data-limited urban environments.

Keywords: fine particulate matter, dust column mass density, principal component regression, aerosol optical depth, dust pollution, Baghdad.

INTRODUCTION

Fine particulate matter (PM2.5) is of great importance for air quality due to its ability to remain suspended in the atmosphere and easily reach the lower respiratory system after inhalation. Either long- or short-term exposure to PM2.5 has been linked with respiratory, heart, and blood vessel diseases as well as higher premature mortality. The health relevance of PM2.5 has made its monitoring a major scientific and environmental priority, especially in the regions where air-quality observations are limited (Cohen et al., 2017; World Health Organization, 2021).

Mineral dust emissions represent one of the most important natural sources of atmospheric aerosols in arid and semi-arid districts (Ibrahim et al., 2026). The Middle East lies within the global dust belt, where dust emissions are regularly enhanced by dry surfaces, limited vegetation, strong winds, and seasonal atmospheric circulation (Prospero et al., 2002; Kifah et al., 2020). Previous studies have demonstrated that Iraq and its neighboring countries are significant sources of dust activity in southwestern Asia, and the dust events can strongly affect urban particulate pollution. In Iraq, these conditions are intensified by extensive arid surfaces, seasonal dryness, limited vegetation cover, and frequent dust-storm activity, making dust a persistent contributor to atmospheric aerosol variability (Albarakat and Lakshmi, 2019; Moridnejad et al., 2015; Prospero et al., 2002).

Baghdad represents an important urban environment for examining the interaction between regional dust activity and near-surface particulate pollution. It is affected by both transported mineral dust and local urban emissions, while the availability of continuous ground-based PM2.5 observations remains limited. This makes satellite-derived aerosol products and reanalysis-based meteorological variables particularly valuable for assessing PM variability over the city (Abdulrahman, 2025; Hamad et al., 2015; Khadom et al., 2024). At the regional scale, previous studies have also shown that Iraq and Kuwait experience elevated PM2.5 levels associated with the combined influence of desert dust and urban pollution (Li et al., 2021). Similar interactions between mineral dust and anthropogenic particulate sources have also been reported in other dust-affected arid urban environments (Fomba et al., 2024; Myasar et al., 2026).

Satellite-derived aerosol optical depth (AOD) has been commonly used to support PM2.5 estimation, since it provides spatially continuous information on aerosol loading (Handschuh et al., 2024; Ma et al., 2022). Empirical, statistical, and machine-learning approaches have also linked AOD with ground-level PM2.5 in different environments (Kong et al., 2016; T. Li et al., 2020; Bagheri, 2022). The MODIS aerosol products are widely used in aerosol studies (Levy et al., 2013; Lyapustin et al., 2018), while reanalysis datasets, such as ERA5 and MERRA-2, offer consistent meteorological and aerosol variables over long periods (Gelaro et al., 2017; Hersbach et al., 2020; Randles et al., 2017). However, AOD represents column-integrated aerosol extinction, whereas surface PM2.5 represents near-surface particle concentration. Thus, their relationship is affected by boundary-layer mixing, humidity, precipitation, aerosol vertical distribution, and other variables (Handschuh et al., 2024; Ma et al., 2022).

In this study, DCMD-PM2.5 refers to the PM2.5-related fraction of dust column mass density, representing a column-integrated dust-mass indicator, whereas surface PM2.5 represents near-ground particulate concentration. Several studies have shown that the combination of aerosol indicators and meteorological variables are improves the PM2.5 estimation instead of using just AOD. The particle emission, transport, vertical dilution, hygroscopic growth, and wet removal are influenced by several meteorological variables such as temperature, humidity, dew point, precipitation, wind, cloud conditions, and boundary-layer processes (Ge et al., 2024; Qi et al., 2024; Talepour et al., 2024). This is very important in dust-prone urban areas, where surface PM2.5 reacts differently from column aerosol or dust loading.

Numerous studies utilized satellite, reanalysis, and statistical approaches to estimate or interpret PM2.5 variability at different spatial scales. Li et al. (2021) estimated ambient PM2.5 over Iraq and Kuwait using machine learning and remote sensing data, while Abdulrahman (2025) focused on daily ground-level PM2.5 estimation over Baghdad using MODIS AOD. Khadom et al. (2024) applied machine-learning and deep-learning approaches to predict PM2.5 and AQI levels in Baghdad. Other studies have demonstrated the importance of aerosol optical properties and reanalysis products for PM2.5 estimation (Handschuh et al., 2024; Ma et al., 2022). High-resolution and machine-learning approaches further support the use of integrated predictors for PM2.5 mapping (T. Li et al., 2020; Meng et al., 2024). However, these studies mainly addressed surface PM2.5 estimation or regional aerosol characterization, while the physical distinction between column-related dust mass indicators and near-surface PM2.5 has received less attention within a unified interpretable modeling.

Despite the growing use of satellite and reanalysis data for PM2.5 assessment, most previous studies in Iraq have focused mainly on estimating surface PM2.5 or describing air-quality status. Less attention has been given to comparing column-related dust mass indicators with surface PM2.5 within the same statistical framework. This distinction is important because DCMD-PM2.5 and surface PM2.5 are related aerosol variables, but they do not necessarily represent the same atmospheric process. DCMD-PM2.5 is more closely linked to column dust loading, while surface PM2.5 is strongly affected by near- surface mixing, moisture, removal processes, and local emissions (Ge et al., 2024; T. Li et al., 2020; Qi et al., 2024; Talepour et al., 2024).

In addition, atmospheric predictors are often highly correlated, making direct regression models less stable and harder to interpret (Abdi and Williams, 2010; Jolliffe and Cadima, 2016). Therefore, this study bridged this gap by combining PCA-regression with representative driver-variable reconstruction to compare DCMD-PM2.5 and surface PM2.5 under the same reduced and physically meaningful structure. Based on this framework, two physically interpretable PCA-regression models were developed for Baghdad City: one for DCMD-PM2.5 and another for surface PM2.5. The study used satellite-derived aerosol indicators and reanalysis-based meteorological variables to identify the dominant physical drivers, construct predictive equations, and compare the responses of column-related dust loading and near-surface particulate concentration.

The novelty of this study lies in developing a physically interpretable PCA-regression framework that not only estimates DCMD-PM2.5 and surface PM2.5, but also compares their dominant atmospheric controls within the same reduced driver-variable structure. This allows the two aerosol indicators to be evaluated as related but physically distinct responses to meteorological, cloud, moisture, and aerosol-loading conditions.

Accordingly, the specific objectives were to: (i) identify the dominant atmospheric components controlling dust and PM2.5 variability using PCA; (ii) develop regression models linking the extracted components to DCMD-PM2.5 and surface PM2.5; (iii) reconstruct the PCA-based models using representative driver variables; and (iv) evaluate the predictive performance of both models using an independent validation year (2025).

STUDY AREA

Baghdad City was selected as the case study for this research. It is located in central Iraq at approximately 33.3152 °N and 44.3661 °E, on the banks of the Tigris River, as shown in Figure 1. The city lies on a broad alluvial plain with low elevation, and the Tigris River divides it into the eastern part and the western part. Baghdad is the largest urban center in Iraq, which makes it an important location for studying particulate pollution under both urban and regional atmospheric influences (Abdulrahman, 2025; Khadom et al., 2024). Baghdad lies within Iraq's hot arid-to-semi-arid climatic setting, characterized by very hot dry summers and limited seasonal rainfall. These conditions enhance the susceptibility of the city and its surroundings to dust activity (Faten et al., 2024; World Bank Group, 2026).

Baghdad is affected by natural and anthropogenic sources of PM2.5. Dust storms, especially those associated with the western desert of Iraq and the Arabian Peninsula, represent an important natural source of particulate pollution. Further, urban sources such as traffic emissions, diesel electrical generators, fuel combustion, and others also impact the level of PM2.5 in the city (Hamad

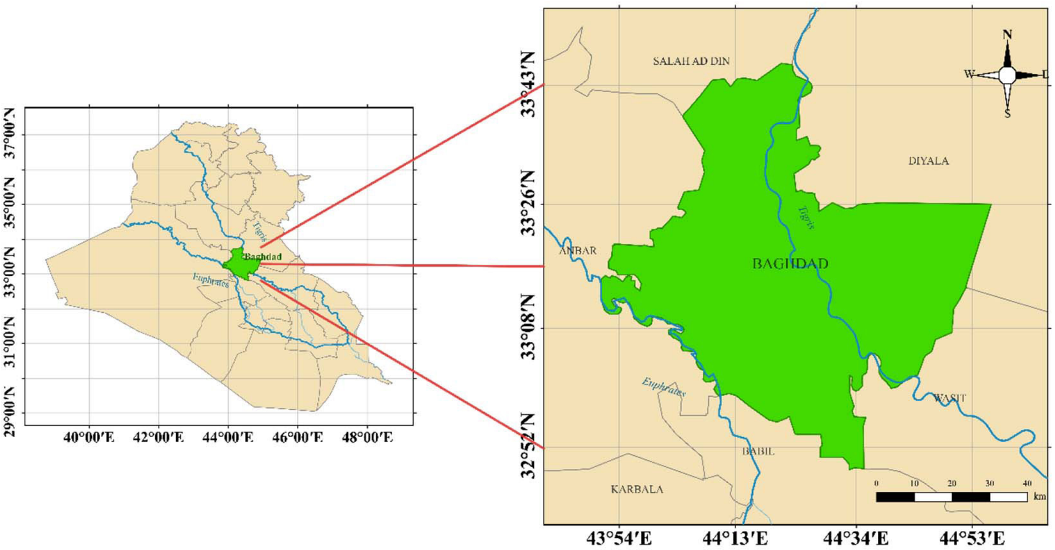


Figure 1. Location of Baghdad City within Iraq and the selected study domain, prepared by the authors using GIS

et al., 2015; J. Li et al., 2021). All these sources make Baghdad an appropriate place to study the difference between the column-related dust loading and the near-surface dust concentration.

In addition, Baghdad was chosen due to the limited availability of PM2.5 ground-based monitoring in Iraq. Previous studies reported that the regional monitoring network is sparse, which limits the ability to assess long-term PM2.5 exposure using only ground observations (Abdulrahman, 2025; Khadom et al., 2024). Therefore, satellite-derived aerosol products and reanalysis-based meteorological as well as dust variables provide valuable support for air-quality modeling in Baghdad and similar data-limited urban environments. Accordingly, Baghdad provides a suitable case for evaluating physically interpretable PCA-regression models to distinguish between column-related dust loading and near-surface PM2.5 concentration in a data-limited urban environment (J. Li et al., 2021; Ma et al., 2022).

DATA AND PRE-PROCESSING

A multi-source daily dataset was prepared in Baghdad during the period 2014–2025. Two target variables were considered: PM2.5-related dust column mass density (DCMD-PM2.5) and surface PM2.5 concentration. Predictor variables were obtained from three main data sources. ERA5 reanalysis was used to provide the main meteorological predictors, including temperature, dew point, wind components, cloud cover, precipitation, radiation, and boundary-layer-related variables. MERRA-2 reanalysis was utilized to collect dust and aerosol-related variables, including aerosol extinction, black carbon surface mass concentration, and scattering indicators (Gelaro et al., 2017; Randles et al., 2017). Aerosol optical indicators, such as Ångström exponent and AOD at 550 nm, were retrieved using MODIS/Giovanni products (Levy et al., 2013; Lyapustin et al., 2018). The combination of these datasets constituted the primary atmospheric processes affecting dust and particulate matter variability over Baghdad.

Satellite-based AOD retrievals can be contaminated by clouds, affected by bright-surface issues, and problematic when they fail to retrieve, all of which can impact the estimation of PM2.5 if not handled carefully (Fu et al., 2020). Two target variables for model development and validation were used: DCMD-PM2.5 and surface PM2.5, which were used as reference values from the MERRA-2 aerosol reanalysis. The MERRA-2 targets were interpreted as references and not direct observations as previous validation studies have demonstrated region- and season-dependent biases of reanalysis-derived particulate concentrations against ground-based observations (Ali et al., 2022).

Model calibration was performed for the period 2014-2024, whereas 2025 was kept as an independent validation year. All datasets were resampled to common 24-hour daily scale and converted to Iraq local time (IST) using UTC+3. The continuous variables of atmosphere and aerosol were calculated as daily means, and the precipitation was calculated as a daily accumulated variable. If both Terra and Aqua MODIS data were available for the same day, then MODIS Terra and Aqua were combined to get one daily AOD observation.

A consistent gap-filling procedure was used for missing values. Temporal interpolation was applied to short gaps (up to three consecutive days), and statistically consistent relationships to available correlated predictors were applied to longer gaps.

Before PCA, all the predictor variables were normed with z-score normalization. Only the calibration period data (2014-2024) were used for the calculation of the mean and standard deviation, and then applied to the independent validation year (2025), to minimize information leakage and comparability between variables with different scales (Abdi and Williams, 2010; Jolliffe and Cadima, 2016).

METHODOLOGY

The methodology aimed to build physically interpretable PCA-regression models to estimate DCMD-PM2.5 and surface PM2.5 over Baghdad. The model was divided into three steps: firstly, a multi collinear reduction of the predictors of the atmospheric model (principal component analysis, PCA) was used to remove the multicollinearity among the atmospheric predictors (Abdi and Williams, 2010; Jolliffe and Cadima, 2016; Abed et al., 2026); secondly, linear regression was applied to relate the PCA extracted PCs to each target variable; thirdly the PCA-regression models were reconstructed using specific driver variables to build simpler and physically meaningful equations. Data from 2014 to 2024 were used to calibrate the models, and data from 2025 were used to validate the models independently. The whole process is summarized in Figure 2.

Principal component analysis

The predictor variables were standardized by using z-score normalization before applying PCA to place variables with different units and magnitudes on a comparable scale. PCA was then used as a dimension-reduction technique to transform the correlated atmospheric variables into a smaller set of uncorrelated components. Each principal component represents a linear combination of the original variables, with the first component explaining the largest possible variance and subsequent components explaining the remaining variance while remaining uncorrelated with the previous ones. This approach helps reduce multicollinearity and improves the stability and interpretability of the subsequent regression models (Abdi and Williams, 2010; Jolliffe and Cadima, 2016). Similar statistical-learning frameworks have also been used in atmospheric modeling applications (James et al., 2021; Faten et al., 2026).

To assess the appropriateness of the data set for PCA, the Kaiser-Meyer-Olkin (KMO) measure and Bartlett's test of sphericity were used. The retained components were interpreted based on their atmospheric meaning, rotated loadings, explained variance, and eigenvalues. Based on the dominant rotated loadings, five components were then used as independent predictors in the PC-based regression models for DCMD-PM2.5 and surface PM2.5.

Regression model development

After PCA extraction, multiple linear regression was applied separately for the two target variables: DCMD-PM2.5 and surface PM2.5. In both models, the principal component scores (PCs) were used as independent predictors.

$$Y = \beta_0 + \sum_{i=1}^{5} \beta_i PC_i \tag{1}$$

where: Y is the target variable, β_0 is the intercept, β_i is the regression coefficient of the i-th principal component, and PC_i is the corresponding PCs.

This formulation follows the standard principal component approach (Jasim et al., 2012; Lim et al., 2022).

Driver – variable reconstruction

Although PCs – based regression improves model stability, the resulting equations may be difficult to interpret because each component combines many original variables. Therefore, a representative driver-variable approach was used to reconstruct the models in a simpler form. One physically meaningful variable was selected to

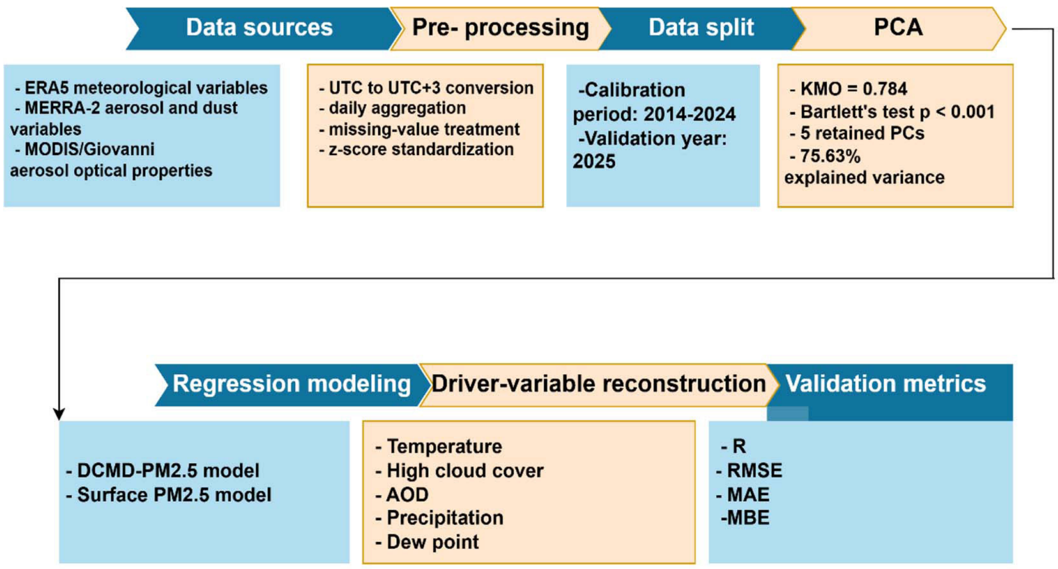


Figure 2. Methodological workflow of the data preparation, principal component analysis, regression modeling, driver-variable reconstruction, and validation stages

represent each principal component based on loading magnitude, physical relevance, non-redundancy, and contribution stability.

The reduced driver-based model was expressed using standardized variables as

$$Y \approx C + \sum_{i=1}^{5} w_i Z_i \tag{2}$$

where: Y is the predicted target variable, C is the intercept term, w_i is the final weight of each selected driver variable, and Z_i is the standardized value of the corresponding predictor.

The final weights (w_i) were estimated separately for each target variable using ordinary least-squares regression during the calibration period (2014–2024). In this step, each target variable was regressed against the five standardized representative drivers: temperature, high cloud cover, AOD 550 nm, total precipitation, and dew point. Therefore, the reported weights represent the fitted regression coefficients of the reduced driver-based model, rather than the PCA loadings themselves. This formulation gives a simplified approximation of the PC-based regression model while preserving physical interpretability (James et al., 2021).

Model validation

The final reduced models were validated on the independent validation year 2025. The following performance indicators were applied: the correlation coefficient (R), the root mean square error (RMSE), the mean absolute error (MAE), and the mean bias error (MBE). The degree of agreement between reference and predicted temporal variability was checked by R, and MBE was used to determine the overall error bias, whether it is over or under. For the assessment of environmental models, the error metrics used were the RMSE and MAE (Willmott and Matsuura, 2005; Abed et al., 2025).

RESULTS AND DISCUSSION

In this section, two reduced PCA-regression models were examined. The first model is used to estimate the PM2.5-related fraction of the dust column mass density (DCMD-PM2.5), and the second model is used to estimate the near-surface PM2.5 mass concentration. The models were both derived from the same set of atmospheric predictors, although they are tested separately since the processes that control column dust loading differ somewhat from those that control near-ground particle concentration. As a result, the statistical structure from the PCA is related to the physical behaviors of dust, aerosol loading, moisture, clouds, and particulate matter levels at the surface of Baghdad.

The PCA analysis revealed five PCs, which accounted for 75.63% of the total variance in the standardized atmospheric data set. This suggests that a small number of independent meteorological, radiative, cloud, and aerosol predictors accounted for most of the variance in the meteorological data. The Kaiser-Meyer-Olkin value was 0.784, and Bartlett's test was significant ($p < 0.001$), confirming that the dataset had a suitable correlation structure for PCA.

PC1 explained the largest part of the variance (38.72%) and was mainly associated with thermal, radiative, and atmospheric mixing

Table 1. Retained principal components, initial and rotated explained variance, dominant rotated loadings, and physical interpretation

Principal component	Initial eigenvalue	Initial variance (%)	Rotated variance (%)	Dominant variables	Physical interpretation
PC1	8.132	38.72	27.307	Downward UV radiation; boundary-layer height; temperature	Thermal-radiative conditions and atmospheric mixing
PC2	3.981	18.96	16.752	High, total and medium cloud cover	Cloud dynamics and circulation
PC3	1.637	7.79	14.730	AOD 550 nm; total aerosol extinction; dust scattering	Aerosol optical loading and dust-related activity
PC4	1.097	5.23	9.421	Low cloud cover; total precipitation	Precipitation-related and wet-removal conditions
PC5	1.035	4.93	7.421	Evaporation; dew point	Moisture and near-surface saturation conditions

conditions. It showed strong positive loadings for temperature, boundary layer height, downward UV radiation, and forecast albedo, while relative humidity, surface pressure, and black carbon surface mass concentration showed negative loadings. This pattern suggests that PC1 represents the general energy balance and atmospheric mixing state over Baghdad.

PC2 (18.96%) was primarily associated with processes and the circulation of clouds. There was good positive loading for high and total cloud cover and negative loading for the U-wind component. PC3 (7.79%) was dominated by aerosol optical variables, particularly AOD 550, total aerosol extinction, and dust scattering, and it showed the aerosol loading and optical activity of the dust. Low cloud cover and total precipitation were the primary factors influencing PC4 (5.23%), implying a precipitation and wet removal component. Finally, PC5 (4.93%) was associated with dew point and evaporation, which are both related to moisture and near-surface saturation.

Overall, the PCs show that the variability in dust and PM2.5 over Baghdad is due to a combination of five main atmospheric processes: thermal-radiative mixing, cloud dynamics, aerosol optical loading, precipitation, and moisture conditions. PC1 had the largest variance, but PC3 is significant for the current study because it is directly related to the optical parameters of aerosols (AOD, extinction, and dust scattering). Table 1 summarizes the retained principal components, explained variance (initial and rotated), dominant rotated loadings, and physical interpretations.

Selection of representative driver variables

To make the PCA-regression models more physically interpretable, one representative driver variable was selected for each retained principal component. The selection was not based only on the highest loading value, but also on physical relevance. This step allowed the PC-based models to be reconstructed into simpler equations using observable and physically meaningful predictors. Temperature was selected to represent PC1 because this component mainly reflected thermal-radiative conditions. Although PC1 also showed strong loadings for boundary layer height, downward UV radiation, forecast albedo, surface pressure, and relative humidity, temperature provides a simple and physically meaningful indicator of surface heating, turbulence, and atmospheric mixing over Baghdad.

High cloud cover was chosen to represent PC2 because this component was dominated by cloud-related variables, specifically high and total cloud cover, making it a suitable proxy for cloud-dynamic conditions that may influence radiation, atmospheric stability, and particle accumulation.

AOD was chosen for PC3 since it had high loading within the aerosol optical component and directly represents column aerosol loading, which is highly relevant for both dust-related column mass and PM2.5 estimation (Shin et al., 2022).

The variables representing PC4, which is related to wet removal processes, were selected as precipitation, and the variable representing PC5, related to near-surface moisture and atmospheric saturation (which can affect particle growth, visibility, and PM2.5 variability), was selected as dew point. Then, the final reduced models in

Table 2. Representative driver variables selected for the reduced PCA-regression models

Principal component	Selected representative driver	Loading of selected driver	Selection rationale
PC1	Temperature	0.783	Represents the thermal-radiative and atmospheric mixing component and is directly linked to surface heating and boundary-layer development.
PC2	High cloud cover	0.850	Represents the strongest cloud-related variable and reflects cloud-dynamic variability.
PC3	AOD 550 nm	0.859	Represents the strongest aerosol optical variable and directly reflects column aerosol loading.
PC4	Total precipitation	0.708	Represents precipitation-related wet removal and atmospheric cleansing processes, despite the stronger loading of low cloud cover.
PC5	Dew point	0.626	Represents a physically meaningful moisture indicator related to near-surface saturation and humidity conditions.

the standardized form were reconstructed using these selected variables. This way, the physical meaning of PCA was maintained, and a complicated equation with too many correlated predictors was avoided.

This summary of representative driver variables for each principal component retained is shown in Table 2. Selection was not only determined by the size of the rotated loading, but also by its physical interpretability and lack of redundancy among the rotated predictors in the reduced regression model.

Final reduced PCA-regression models

The reduced driver-based equations were derived by reconstructing the PC-based regression models using the selected standardized driver variables. For the DCMD-PM2.5 model, all selected drivers showed positive coefficients, suggesting positive statistical associations with column-related dust loading within the reduced standardized model. These coefficients should be interpreted as model-based relationships rather than direct causal effects, especially for variables such as precipitation that may also represent removal-related conditions.

The largest coefficient was associated with AOD, confirming that aerosol optical loading is the strongest driver of the DCMD-PM2.5 model. This is physically reasonable because DCMD-PM2.5 is a column-related dust/aerosol variable and is therefore expected to respond strongly to aerosol optical properties (Shin et al., 2022).

For the surface PM2.5 model, the reduced equation showed both positive and negative coefficients. The negative coefficients of temperature, precipitation, and dew point suggest that surface PM2.5 over Baghdad is strongly affected by near-surface mixing, dilution, moisture-related processes, and wet removal.

The positive high cloud cover coefficient could indicate that in cloudy or synoptic conditions, surface heating is reduced and the effect of vertical mixing on the particles is reduced, so that these particles can stay closer to the ground. The weak negative coefficient of AOD shows that the relationship between column aerosol loading and surface PM2.5 is not direct, because surface PM2.5 also depends on boundary-layer structure, local emissions, and aerosol vertical distribution (Gao et al., 2024; Vovk et al., 2024).

All predictor variables in Equations 3 and 4 are standardized values calculated using the calibration-period mean and standard deviation. The final reduced equations are given as follows:

$$\begin{gathered} DCMD\text{-}PM2.5 \approx 0.0001030864 + \\ + 3.21 \times 10^{-5} Z_{Temp} + 1.60 \times \\ \times 10^{-5} Z_{Hcloud} + 4.81 \times \\ \times 10^{-5} Z_{AOD} + 9.70 \times 10^{-6} Z_{Precip} + \\ + 2.55 \times 10^{-5} Z_{Dewpoint} \end{gathered} \tag{3}$$

$$\begin{gathered} PM2.5 \approx 64.481 - 14.155 Z_{Temp} + \\ + 3.939 Z_{Hcloud} - 0.476 Z_{AOD} - 2.373 Z_{Precip} - \\ - 7.771 Z_{Dewpoint} \end{gathered} \tag{4}$$

Model performance and validation

During model development, the original PC-based regression models showed stronger explanatory performance for DCMD-PM2.5 than for surface PM2.5. The DCMD-PM2.5 PC-based model showed R = 0.898 and R^2 = 0.806, indicating that approximately 80.6% of the variance was explained by the principal components. In comparison, the surface PM2.5 PC-based model showed R = 0.731 and R^2 = 0.534. These calibration statistics describe the original PC-based model formulation and are presented separately from the independent validation of the final reduced equations.

The final reduced equations were independently evaluated for 2025 against MERRA-2-derived reference values. The DCMD-PM2.5 reduced equation showed stronger validation performance, with $R = 0.799$, indicating a strong positive relationship between reference and predicted values. The error values were relatively low, with RMSE = 4.82×10^{-5} kg m^{-2}, MAE =3.86×10^{-5} kg m^{-2}, and MBE = -1×10^{-6} kg m^{-2}. The slightly negative MBE indicates a small underestimation bias, but the magnitude is very small, suggesting that the model is nearly unbiased for daily-scale DCMD-PM2.5 prediction.

For surface PM2.5, the final reduced equation showed weaker but still informative validation performance, with $R = 0.448$. The error values were $RMSE = 25.29$ μg m^{-3}, $MAE = 20.27$ μg m^{-3}, and $MBE = +4.52$ μg m^{-3}. The positive MBE suggests a minor tendency for overestimation. Such weaker performance is expected as surface PM2.5 is affected by local emissions, traffic, generators, urban activities, short-term dust occurrences, and vertical mixing processes, which are

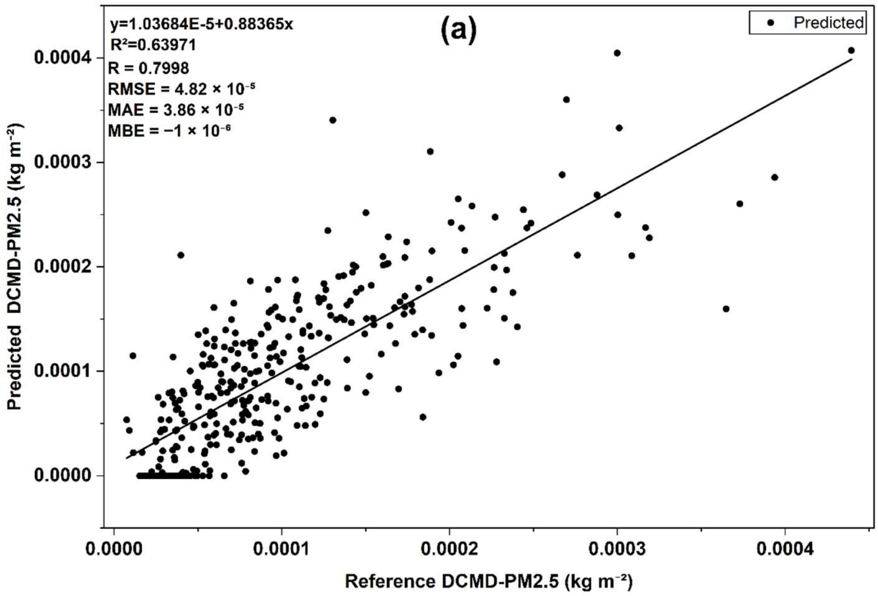
Table 3. Independent validation performance of the final reduced DCMD-PM2.5 and surface PM2.5 equations against MERRA-2-derived reference values during 2025

Model	Validation R	Validation RMSE	Validation MAE	Validation MBE
DCMD-PM2.5	0.799	4.82×10^{-5} kg m^{-2}	3.86×10^{-5} kg m^{-2}	-1×10^{-6} kg m^{-2}
Surface PM2.5	0.448	25.29 μg m^{-3}	20.27 μg m^{-3}	+4.52 μg m^{-3}

Note: * RMSE, MAE, and MBE are interpreted within each target variable because DCMD-PM2.5 and surface PM2.5 have different physical units and numerical scales.

not completely captured by satellite and reanalysis predictors (Vovk et al., 2024). The independent validation metrics for the final reduced models are summarized in Table 3.

This is further shown in the validation scatter plots in Figure 3, which demonstrates the difference in the predictive ability of the two models. The surface PM2.5 model exhibits greater scatter



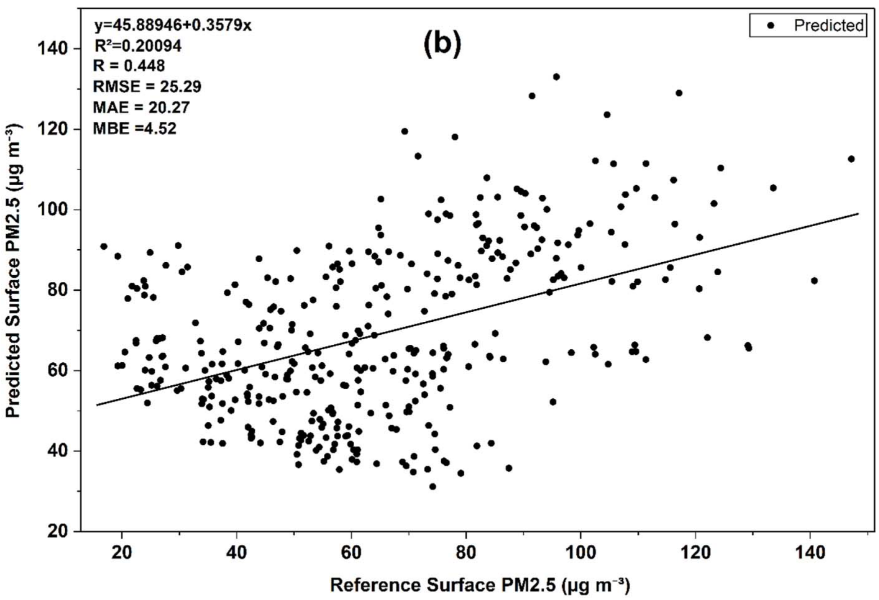


Figure 3. Reference versus predicted values during the independent validation year 2025 for (a) DCMD-PM2.5 and (b) surface PM2.5

between the reference and predicted values as compared to the DCMD-PM2.5 model, which indicates the larger impacts of local and near-surface processes.

In addition to the scatter plot, the time-series comparison in Figure 4 shows how the reference and predicted values change throughout the 2025 validation year. This view is useful because it shows not only the overall agreement between the models and reference data, but also how well the reduced equations followed the daily variability of DCMD-PM2.5 and surface PM2.5.

Comparative interpretation of the two models

The two reduced models did not respond to the selected drivers in the same way. In the DCMD-PM2.5 equation, the coefficients were mainly positive, with AOD and temperature being especially important. This behavior is reasonable for a column-related dust variable. When aerosol optical loading increases, or when surface heating supports dust uplift and vertical exchange, the dust column signal is expected to increase as well. Therefore, the DCMD-PM2.5 model

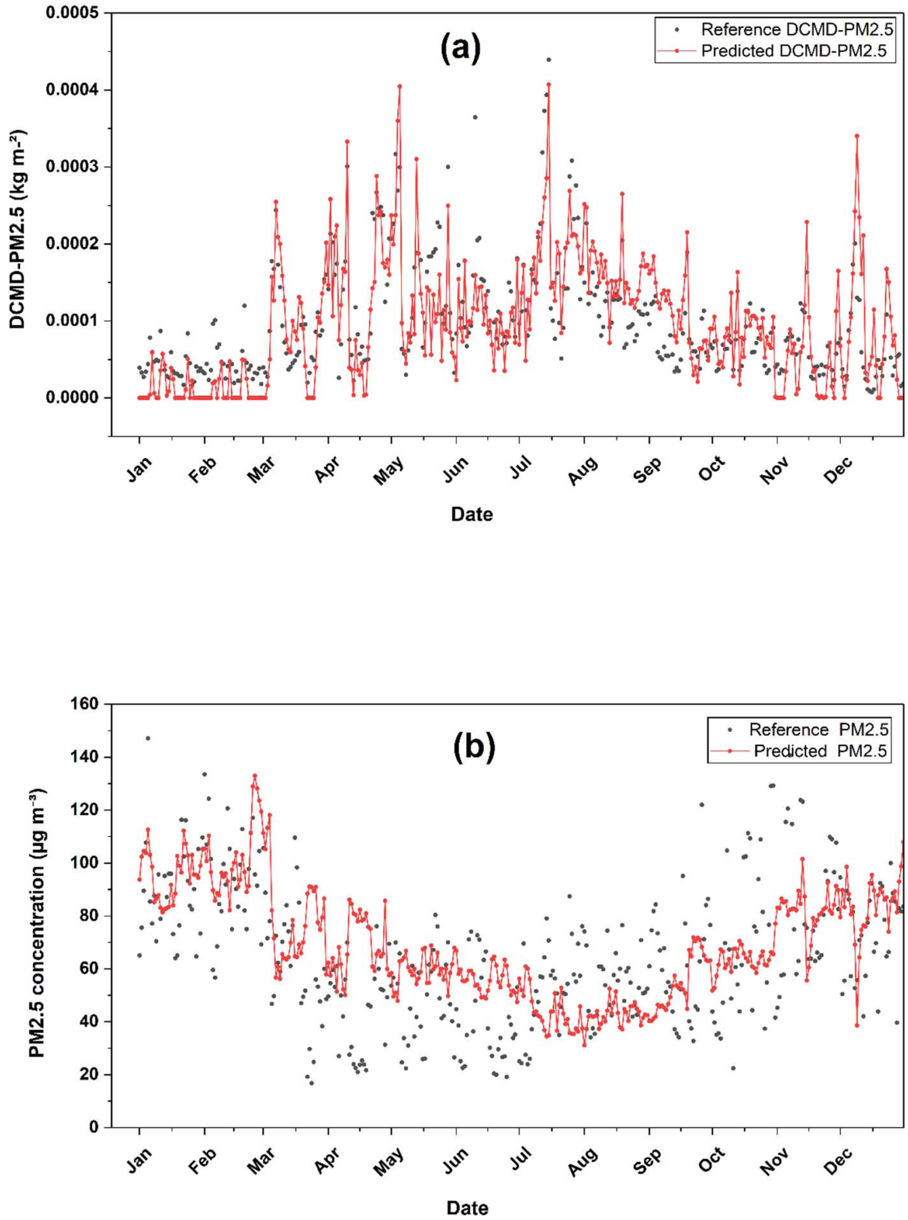


Figure 4. Time-series comparison of daily reference and predicted values during the independent validation period in 2025 for (a) DCMD-PM2.5 and (b) surface PM2.5

mainly reflects large-scale dust and aerosol loading through the atmospheric column.

The surface PM2.5 model showed a different response. Its negative temperature coefficient suggests that warmer conditions may be associated with stronger boundary-layer development and greater vertical dilution near the surface. The negative precipitation coefficient also agrees with the wet scavenging and removal of particles. The reduced equation had a negative coefficient for dew point, suggesting that moisture-related conditions may be a factor in the removal, deposition, or modification of near-surface particle behavior. The positive sign of the high cloud cover coefficient, on the contrary, could indicate that under high cloud cover conditions, the heating or stability of the atmosphere decreases, and that particles are closer to the ground.

The differences suggest that the two (DCMD-PM2.5 and surface PM2.5) cannot be interchanged as indicators. DCMD-PM2.5 is more representative of column-integrated dust and aerosol mass, whereas surface PM2.5 depends strongly on local emissions, vertical mixing, atmospheric stability, and removal processes near the ground. This separation is of significance for the city of Baghdad, as the city is receiving both local and regional transported dust. The two models therefore provide complementary information: the DCMD-PM2.5 model describes the broader column dust signal, while the surface PM2.5 model reflects the more complex near-surface pollution environment.

The validation results should also be interpreted with caution because the reference values were derived from MERRA-2 rather than from continuous ground-based measurements. Thus, the reported performance describes how well the reduced equations reproduced reanalysis-based aerosol variability. Further evaluation against in-situ PM2.5 observations would be needed as longer, more continuous monitoring records become available (Ali et al., 2022). A few raw negative DCMD-PM2.5 predictions appeared when the reference values were very low. This occurred because the reduced linear equation was not constrained to produce only non-negative values. Since dust column mass density cannot be physically negative, these values were set to zero before the final validation statistics and figures were reported.

CONCLUSIONS

This study developed two physically interpretable PCA-regression models for estimating DCMD-PM2.5 and surface PM2.5 over Baghdad using satellite-derived aerosol indicators and reanalysis-based meteorological variables. The PCA reduced the dimensionality of the atmospheric dataset and identified five dominant components representing thermal-radiative mixing, cloud dynamics, aerosol optical loading, precipitation removal, and moisture conditions. These components were linked to the target variables through regression and reconstructed using representative driver variables: temperature, high cloud cover, AOD, precipitation, and dew point.

The results revealed the DCMD-PM2.5 model performed better than the surface PM2.5 model. During model development, the PC-based DCMD-PM2.5 model explained 80.6% of the variance, while the corresponding surface PM2.5 PC-based model explained 53.4%. The validation confirmed stronger predictive capability for DCMD-PM2.5, with $R = 0.799$, low error values, and a very small negative bias. In contrast, the final reduced surface PM2.5 equation showed weaker but still informative validation performance, with $R = 0.448$, higher errors, and a slight overestimation bias.

The different behavior of the two models indicates that DCMD-PM2.5 and surface PM2.5 should not be interpreted as identical aerosol indicators. The DCMD-PM2.5 is mainly controlled by column aerosol loading and dust-related atmospheric mass, with AOD acting as the strongest driver. Surface PM2.5, however, is more strongly influenced by near-surface mixing, atmospheric stability, moisture, wet removal, and local emission conditions. This explains why the same atmospheric predictors may have different coefficient signs in the two models.

Finally, the use of the proposed PCA-regression and driver-variable framework is a simple and transparent, physically meaningful way to estimate the quantity of dust-related column loading and PM2.5 on the surface in Baghdad. This framework is particularly helpful for areas where ground observations are sparse, and satellite as well as reanalysis data can be used for air-quality assessment.

REFERENCES

1. Abdi, H., Williams, L. J. (2010). Principal component analysis. *WIREs Computational Statistics*, *2*(4), 433-459. https://doi.org/10.1002/wics.101
2. Abed, F., Rajab, J., Daway, H. (2026). Modeling surface ozone concentrations in Iraq using principal component analysis and stepwise multiple linear regression method. *Ozone Science and Engineering*, *48*(1), 66–82.
3. Abed, G., Rajab, J., Daway, H. (2025). Integration of stepwise multiple linear regression and artificial neural network models for predicting surface ozone concentration in Iraq. *Papers in Applied Geography*, *11*(4), 402–420. https://doi.org/10.1080/23754931.2025.2455745
4. Abdulrahman, K. R. (2025). Air quality assessment by daily estimation of ground-level PM2.5 concentrations over Baghdad City using MODIS AOD data. *Air Quality, Atmosphere & Health, 18*(7), 2047–2059. https://doi.org/10.1007/s11869-025-01752-1
5. Albarakat, R., Lakshmi, V. (2019). Monitoring dust storms in Iraq using satellite data. *Sensors*, *19*(17), 3687. https://doi.org/10.3390/s19173687
6. Ali, M. A., Bilal, M., Wang, Y., Nichol, J. E., Mhawish, A., Qiu, Z., de Leeuw, G., Zhang, Y., Zhan, Y., Liao, K., Almazroui, M., Dambul, R., Shahid, S., Islam, M. N. (2022). Accuracy assessment of CAMS and MERRA-2 reanalysis PM2.5 and PM10 concentrations over China. *Atmospheric Environment*, *288*, 119297. https://doi.org/10.1016/j.atmosenv.2022.119297
7. Bagheri, H. (2022). A machine learning-based framework for high resolution mapping of PM2.5 in Tehran, Iran, using MAIAC AOD data. *Advances in Space Research*, *69*, 3333–3349. https://doi.org/10.1016/j.asr.2022.02.032
8. Cohen, A. J., Brauer, M., Burnett, R., Anderson, H. R., Frostad, J., Estep, K., Balakrishnan, K., Brunekreef, B., Dandona, L., Dandona, R., Feigin, V., Freedman, G., Hubbell, B., Jobling, A., Kan, H., Knibbs, L., Liu, Y., Martin, R., Morawska, L.,... Forouzanfar, M. H. (2017). Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: An analysis of data from the Global Burden of Diseases Study 2015. *The Lancet*, *389*(10082), 1907–1918. https://doi.org/10.1016/S0140-6736(17)30505-6
9. Faten, A., Jasim, R., Ibtihaj, A., Hwee, L. (2024). Influence of air temperature and interrelationship with greenhouse gases (CO2 and CH4) over Iraq using AIRS data. *Atmospheric Pollution Research*, *15*(11), 102293. https://doi.org/10.1016/j.apr.2024.102293
10. Faten, G., Jasim, R., Hazim, D. (2026). Combining principal component analysis method and artificial neural network model for surface ozone concentration prediction in Baghdad, Iraq. *Communications in Computer and Information Science*, *2653*, 281–296. https://doi.org/10.1007/978-3-032-13026-6_20
11. Fomba, K. W., Faboya, O. L., Deabji, N., Makhmudov, A., Hofer, J., dos Santos Souza, E. J., Müller, K., Althausen, D., Sharipov, S., Abdullaev, S., Herrmann, H. (2024). The seasonal variation of Asian dust, anthropogenic PM, and their sources in Dushanbe, Tajikistan. *Atmospheric Environment*, *333*, 120667. https://doi.org/10.1016/j.atmosenv.2024.120667
12. Fu, D., Song, Z., Zhang, X., Xia, X., Wang, J., Che, H., Wu, H., Tang, X., Zhang, J., Duan, M. (2020). Mitigating MODIS AOD non-random sampling error on surface PM2.5 estimates by a combined use of Bayesian maximum entropy method and linear mixed-effects model. *Atmospheric Pollution Research*, *11*, 482–490. https://doi.org/10.1016/j.apr.2019.11.020
13. Gao, Z., Do, K., Li, Z., Jiang, X., Maji, K. J., Ivey, C. E., Russell, A. G. (2024). Predicting PM2.5 levels and exceedance days using machine learning methods. *Atmospheric Environment*, *323*, 120396. https://doi.org/10.1016/j.atmosenv.2024.120396
14. Ge, S., Su, J., Zhao, P., Li, J., Liu, S., Qiu, Y., Pu, W., Ma, Z. (2024). Characteristics of PM2.5 hygroscopicity and the influences of water-soluble ions during haze events in Beijing. *Atmospheric Environment*, *322*, 120382. https://doi.org/10.1016/j.atmosenv.2024.120382
15. Gelaro, R., McCarty, W., Suárez, M. J., Todling, R., Molod, A., Takacs, L., Randles, C. A., Darmenov, A., Bosilovich, M. G., Reichle, R. (2017). The modern-era retrospective analysis for research and applications, Version 2 (MERRA-2). *Journal of Climate*, *30*(14), 5419–5454. https://doi.org/10.1175/JCLI-D-16-0758.1
16. Hamad, S. H., Schauer, J. J., Heo, J., Kadhim, A. K. H. (2015). Source apportionment of PM2.5 carbonaceous aerosol in Baghdad, Iraq. *Atmospheric Research*, *156*, 80–90. https://doi.org/10.1016/j.atmosres.2014.12.017
17. Handschuh, J., Erbertseder, T., Baier, F. (2024). On the added value of satellite AOD for the investigation of ground-level PM2.5 variability. *Atmospheric Environment*, *331*, 120601. https://doi.org/10.1016/j.atmosenv.2024.120601
18. Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 1999–2049. https://doi.org/10.1002/qj.3803
19. Ibrahim, M., Mahdi, B., Darwesh, D., Jasim, R., Ali, M., Hwee, L. (2026). Pollution assessment

and ecological risk of five commonly monitored heavy metals in urban dust from Duhok City, Iraq: A seasonal and isotopic perspective. *Environmental Monitoring and Assessment*, *198*, 579. https://doi.org/10.1007/s10661-026-15416-9
20. James, G., Witten, D., Hastie, T., Tibshirani, R. (2021). An introduction to statistical learning. *Springer*. https://doi.org/10.1007/978-1-0716-1418-1
21. Jasim, R., Mohd, J., Hwee, L., Khiruddin, A. (2012). Regression analysis in modeling of air surface temperature and factors affecting its value in Peninsular Malaysia. *Optical Engineering*, *51*(10), 101702. https://doi.org/10.1117/1.OE.51.10.101702
22. Jolliffe, I. T., Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202
23. Khadom, A. A., Albawi, S., Abboud, A. J., Mahood, H. B., Hassan, Q. (2024). Predicting air quality index and fine particulate matter levels in Baghdad city using advanced machine learning and deep learning techniques. *Journal of Atmospheric and Solar-Terrestrial Physics*, *262*, 106312. https://doi.org/10.1016/j.jastp.2024.106312
24. Kifah, B. A.-T., Jasim, M. R., Al-Salihi, A. M. (2020). Climatology and classification of aerosols based on optical properties over selected stations in Iraq. *AIP Conference Proceedings, 2290*(1), 050041. https://doi.org/10.1063/5.0031471
25. Kong, L., Xin, J., Zhang, W., Wang, Y. (2016). The empirical correlations between PM2.5, PM10 and AOD in the Beijing metropolitan region and the PM2.5, PM10 distributions retrieved by MODIS. *Environmental Pollution*, *216*, 350–360. https://doi.org/10.1016/j.envpol.2016.05.085
26. Levy, R. C., Mattoo, S., Munchak, L. A., Remer, L. A., Sayer, A. M., Patadia, F., Hsu, N. C. (2013). The Collection 6 MODIS aerosol products over land and ocean. *Atmospheric Measurement Techniques*, *6*(11), 2989–3034. https://doi.org/10.5194/amt-6-2989-2013
27. Li, J., Garshick, E., Hart, J. E., Li, L., Shi, L., Al-Hemoud, A., Huang, S., Koutrakis, P. (2021). Estimation of ambient PM2.5 in Iraq and Kuwait from 2001 to 2018 using machine learning and remote sensing. *Environment International*, *151*, 106445. https://doi.org/10.1016/j.envint.2021.106445
28. Li, T., Shen, H., Yuan, Q., Zhang, L. (2020). Geographically and temporally weighted neural networks for satellite-based mapping of ground-level PM2.5. *ISPRS Journal of Photogrammetry and Remote Sensing*, *167*, 178–188. https://doi.org/10.1016/j.isprsjprs.2020.06.019
29. Lim, S., Rajab, J., Al-Salihi, A., Zainab, S., MatJafri, M. (2022). A statistical model to predict and analyze air surface temperature based on remotely sensed observations. *Environmental Science and Pollution Research*, *29*, 9755–9765. https://doi.org/10.1007/s11356-021-16321-z
30. Lyapustin, A., Wang, Y., Korkin, S., Huang, D. (2018). MODIS collection 6 MAIAC algorithm. *Atmospheric Measurement Techniques*, *11*(10), 5741–5765. https://doi.org/10.5194/amt-11-5741-2018
31. Ma, Z., Dey, S., Christopher, S., Liu, R., Bi, J., Balyan, P., Liu, Y. (2022). A review of statistical methods used for developing large-scale and long-term PM2.5 models from satellite data. *Remote Sensing of Environment*, *269*, 112827. https://doi.org/10.1016/j.rse.2021.112827
32. Meng, L., Xu, X., Huang, X., Li, X., Chang, X., Xu, D. (2024). High-resolution estimation of PM2.5 concentrations across China using multiple machine learning approaches and model fusion. *Atmospheric Pollution Research*, *15*, 102110. https://doi.org/10.1016/j.apr.2024.102110
33. Moridnejad, A., Karimi, N., Ariya, P. A. (2015). Newly desertified regions in Iraq and its surrounding areas: Significant novel sources of global dust particles. *Journal of Arid Environments*, *116*, 1-10. https://doi.org/10.1016/j.jaridenv.2015.01.008
34. Myasar, I., Berivan, M., Dalshad, D., Jasim, R., Ali, M., Hwee, L. (2026). Seasonal dynamics, pollution indices, and health risks of heavy metals isotopes (Ni, Cu, Zn, Mn) in urban dust from Duhok City, Iraq. *Water, Air, & Soil Pollution*, *237*, 847. https://doi.org/10.1007/s11270-026-09495-0
35. Prospero, J. M., Ginoux, P., Torres, O., Nicholson, S. E., Gill, T. E. (2002). Environmental characterization of global sources of atmospheric soil dust identified with the Nimbus 7 total ozone mapping spectrometer absorbing aerosol product. *Reviews of Geophysics*, *40*(1), 1002. https://doi.org/10.1029/2000RG000095
36. Qi, L., Zheng, H., Ding, D., Wang, S. (2024). A comparison of meteorological normalization of PM2.5 by multiple linear regression, general additive model, and random forest methods. *Atmospheric Environment*, *338*, 120854. https://doi.org/10.1016/j.atmosenv.2024.120854
37. Randles, C. A., da Silva, A. M., Buchard, V., Colarco, P. R., Darmenov, A., Govindaraju, R., Smirnov, A., Holben, B., Ferrare, R., Hair, J., Shinozuka, Y., Flynn, C. J. (2017). The MERRA-2 aerosol reanalysis, 1980 onward. Part I: System description and data assimilation evaluation. *Journal of Climate*, *30*(17), 6823-6850. https://doi.org/10.1175/JCLI-D-16-0609.1
38. Shin, J., Sim, J., Dehkhoda, N., Joo, S., Kim, T., Kim, G., Müller, D., Tesche, M., Shin, S.-K., Shin, D., Noh, Y. (2022). Long-term variation study of

fine-mode particle size and regional characteristics using AERONET data. *Remote Sensing*, *14*(18), 4429. https://doi.org/10.3390/rs14184429

39. Talepour, N., Tahmasebi Birgani, Y., Kelly, F. J., Jaafarzadeh, N., Goudarzi, G. (2024). Analyzing meteorological factors for forecasting PM10 and PM2.5 levels: A comparison between MLR and MLP models. *Earth Science Informatics*, *17*, 5603–5623. https://doi.org/10.1007/s12145-024-01468-3
40. Vovk, T., Kryza, M., Werner, M. (2024). Using random forest to improve EMEP4PL model estimates of daily PM2.5 in Poland. *Atmospheric Environment*, *332*, 120615. https://doi.org/10.1016/j.atmosenv.2024.120615
41. Willmott, C. J., Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, *30*, 79–82. https://doi.org/10.3354/cr030079
42. World Bank Group. (2026, May 22). Iraq - climatology (CRU) - *Climate Change Knowledge Portal*. Climate Change Knowledge Portal. https://climateknowledgeportal.worldbank.org/country/Iraq/climate-data-historical
43. World Health Organization (2021). WHO global air quality guidelines: Particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide. World Health Organization. https://www.who.int/publications/i/item/9789240034228