IBIMA Publishing
Communications of International Proceedings
https://ibimapublishing.com/p-articles/47AI/2026/4724426/
Vol. 2026 (12), Article ID 4724426

Scene-Level Meteorological Predictors Are Insufficient for AOD550 Estimation: A Baseline Study Toward Hyperspectral Data Fusion*

Dawid JURCZYŃSKI

Silesian University of Technology in Gliwice, Poland

Correspondence should be addressed to: Dawid JURCZYŃSKI, dawid.jurczynski@polsl.pl

*Presented at the 47th IBIMA International Conference, 29-30 June 2026, Madrid, Spain

Abstract

This study investigates whether coarse scene-level meteorological variables alone are sufficient for estimating aerosol optical depth at 550 nm (AOD550), an important indicator of atmospheric aerosol burden widely used in climate, air quality, and environmental monitoring research. Although meteorological variables are frequently incorporated into aerosol-related machine learning frameworks, their standalone explanatory capability remains insufficiently quantified, particularly in simplified scene-level representations. This study addresses this gap by evaluating the predictive limits of meteorology-only modelling and establishing a baseline for future hyperspectral data fusion approaches.

The methodology was based on the integration of three data sources: HYPSO hyperspectral satellite scene metadata, CAMS AOD550 products, and ERA5 meteorological reanalysis variables. A dataset containing 33 satellite scenes was constructed, where each scene was represented by aggregated meteorological descriptors, including air temperature, surface pressure, wind components, and total column water vapour. CAMS-derived scene-level mean AOD550 values were used as the regression target. Two baseline regression approaches, Linear Regression and Random Forest, were applied to evaluate whether the selected meteorological predictors could explain aerosol optical variability.

The results demonstrated that meteorological predictors contained only weak to moderate relationships with AOD550, with total column water vapour and meridional wind showing the strongest associations. Both regression models exhibited limited predictive capability, substantial regression-to-the-mean behaviour, and low or negative coefficients of determination. The findings indicate that scene-level meteorological descriptors do not provide sufficient physical information for reliable AOD550 estimation when used in isolation. Consequently, the study confirms the necessity of incorporating richer spectral and optical information, particularly hyperspectral observations, into future aerosol retrieval and data fusion frameworks.

Keywords: Aerosol Optical Depth (AOD550); ERA5; CAMS; HYPSO-1; hyperspectral remote sensing; aerosol retrieval; meteorological predictors; machine learning; Random Forest; atmospheric modelling.

Introduction

Atmospheric aerosols are a major component of the Earth system because they affect the radiative balance, cloud microphysics, atmospheric chemistry, visibility, air quality, and human health[1]. One of the most widely used integrated indicators of aerosol burden is aerosol optical depth (AOD), which quantifies the attenuation of solar radiation in the atmospheric column caused by scattering and absorption by aerosol particles[2]. As a column-

[1] IPCC. *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, 2021.
[2] Kaufman, Y. J., Tanré, D., Boucher, O. A satellite view of aerosols in the climate system. *Nature*, 2002, 419(6903), 215–223. https://doi.org/10.1038/nature01091

Cite this Article as: Dawid JURCZYŃSKI Vol. 2026 (12) "Scene-Level Meteorological Predictors Are Insufficient for AOD550 Estimation: A Baseline Study Toward Hyperspectral Data Fusion " Communications of International Proceedings, Vol. 2026 (12), Article ID 4724426, https://doi.org/10.5171/2026.4724426

integrated optical property, AOD has become a key variable in climate studies, air pollution assessment, and environmental monitoring[3]. Over the past two decades, satellite remote sensing has made large-scale aerosol monitoring operationally feasible[4]. Multispectral instruments and associated retrieval frameworks have enabled global or near-real-time observation of aerosol-related products and substantially advanced research on aerosol climatology, transport, and pollution events. At the same time, aerosol retrieval from space remains inherently difficult. The top-of-atmosphere signal depends not only on aerosol loading, but also on surface reflectance, cloud contamination, viewing geometry, aerosol type assumptions, and vertical aerosol distribution. As a consequence, retrieval uncertainty remains regionally variable, and the interpretation of AOD products requires caution, especially when they are used in downstream environmental analyses[5]. This broader methodological context is well established in the aerosol remote sensing literature and motivates continued efforts to improve the representation and estimation of aerosol optical properties. In parallel with the development of satellite retrievals, atmospheric reanalyses and composition services have become an important source of aerosol-related information. Products generated within frameworks such as CAMS and ERA5 provide globally consistent spatio-temporal fields describing atmospheric composition and meteorological state[6]. Their major advantage lies in temporal continuity, operational availability, and internal physical consistency, which makes them attractive for integrated environmental modelling. However, these products do not play the same role. CAMS aerosol fields provide a model-based representation of aerosol conditions, while ERA5 variables describe the broader thermodynamic and dynamic background of the atmosphere[7]. This distinction is important, because meteorological context may help explain part of aerosol variability, but it does not directly encode the optical properties of aerosol particles themselves. For this reason, meteorological variables such as air temperature, pressure, wind, humidity, and water vapour are frequently incorporated into aerosol-related studies. They are physically relevant because they modulate aerosol transport, dispersion, dilution, hygroscopic growth, and removal. In machine learning frameworks, such variables are often used as auxiliary predictors to improve retrieval performance, reconstruct missing AOD fields, or explain temporal variability in aerosol indicators[8]. Nevertheless, their role should not be overstated. The relationship between AOD and meteorology is indirect, nonlinear, and strongly conditioned by aerosol composition, source regions, surface properties, and vertical atmospheric structure[9]. In other words, meteorology provides context, but not a complete optical description of aerosol burden. This limitation becomes especially important when the predictor space is reduced to scene-level aggregated meteorological descriptors. Aggregation simplifies data integration and creates a convenient tabular representation for statistical modelling, but it also removes sub-scene spatial variability, local gradients, and fine-scale atmospheric heterogeneity. Such simplification may be useful for establishing a baseline, yet it can also severely restrict the information available to the model. In the case of AOD, which reflects the combined influence of aerosol amount, particle properties, moisture conditions, and vertical structure, coarse scene-level meteorology may be too weakly informative to

---

[3] 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., Lavenu, F., Jankowiak, I., Smirnov, A. AERONET—A federated instrument network and data archive for aerosol characterization. *Remote Sensing of Environment*, 1998, 66(1), 1–16. https://doi.org/10.1016/S0034-4257(98)00031-5

[4] Levy, R. C., Mattoo, S., Munchak, L. A., Remer, L. A., Sayer, A. M., Patadia, F., Hsu, N. C. The Collection 6 MODIS aerosol products over land and ocean. *Atmospheric Measurement Techniques*, 2013, 6(11), 2989–3034. https://doi.org/10.5194/amt-6-2989-2013

[5] Vogel, A., Alessa, G., Scheele, R., et al. Uncertainty in aerosol optical depth from modern aerosol-climate models, reanalyses, and satellite products. *Journal of Geophysical Research: Atmospheres*, 2022, 127(2), e2021JD035483. https://doi.org/10.1029/2021JD035483

[6] Inness, A., Ades, M., Agustí-Panareda, A., Barré, J., Benedictow, A., Blechschmidt, A.-M., Dominguez, J. J., Engelen, R., Eskes, H., Flemming, J., Huijnen, V., Jones, L., Kipling, Z., Massart, S., et al. The CAMS reanalysis of atmospheric composition. *Atmospheric Chemistry and Physics*, 2019, 19(6), 3515–3556. https://doi.org/10.5194/acp-19-3515-2019

[7] Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., et al. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 2020, 146(730), 1999–2049. https://doi.org/10.1002/qj.3803

[8] Zhao, C., Liu, Z., Wang, Q., Ban, J., Chen, N., Li, T. High-resolution daily AOD estimated to full coverage using the random forest model approach in the Beijing–Tianjin–Hebei region. *Atmospheric Environment*, 2019, 203, 70–78. https://doi.org/10.1016/j.atmosenv.2019.01.033

[9] Chang, H. H., Hu, X., Liu, Y., et al. Calibrating MODIS aerosol optical depth for predicting daily PM2.5 concentrations via statistical downscaling. *Journal of Exposure Science & Environmental Epidemiology*, 2014, 24(4), 398–404. https://doi.org/10.1038/jes.2013.90

support reliable estimation on its own. At the same time, recent years have seen growing interest in data-driven aerosol estimation and in multi-source fusion approaches that combine physical context with more observationally rich inputs. Machine learning and deep learning models have been used to retrieve AOD from radiometric measurements, fuse multiple aerosol products, fill spatio-temporal gaps, and estimate aerosol-related quantities from auxiliary variables. These studies show that predictive performance depends not only on the regression architecture, but also on the physical information content of the input variables. If the predictors do not directly capture aerosol-specific optical or spectral characteristics, even flexible nonlinear models may provide only limited gains. This observation is particularly relevant in the context of emerging hyperspectral Earth observation systems. Compared with conventional multispectral data, hyperspectral imagery provides denser spectral sampling and may better capture subtle atmospheric and surface-atmosphere interaction signatures[10]. Such information has the potential to improve the separability of aerosol effects from background reflectance and to support more physically meaningful inference strategies. In this context, missions such as HYPSO-1 are of interest because they extend the availability of compact hyperspectral observations for Earth observation research[11]. However, before developing a full fusion framework that combines hyperspectral observations with atmospheric context, it is necessary to establish a clear baseline showing what can, and cannot, be explained by scene-level meteorology alone. The present study addresses this baseline question. Using hyperspectral satellite scenes only as a common spatio-temporal reference frame, it evaluates whether scene-level meteorological variables derived from ERA5 can explain scene-level AOD550 values derived from CAMS. The purpose of the study is not to develop a high-performance operational estimator, but to quantify the explanatory limits of coarse meteorological predictors when used in isolation. In this sense, the work is intentionally constrained and should be interpreted as a baseline experiment rather than as a complete aerosol retrieval framework. The central assumption is that scene-level meteorological predictors are insufficient for reliable AOD550 estimation because they do not contain direct aerosol-specific optical information. Testing this assumption is scientifically useful for two reasons. First, it provides an empirical verification of a physically expected limitation within a clearly defined modelling setup. Second, it establishes the methodological justification for subsequent data fusion studies in which richer spectral information, including hyperspectral observations, can be incorporated together with atmospheric background variables. Thus, the contribution of this paper is not the demonstration of strong predictive performance, but the quantitative identification of the limits of a simplified predictor space and the resulting motivation for more information-rich fusion strategies.

Materials and Methods

Study Design

The study was designed as a baseline experiment to evaluate whether scene-level meteorological descriptors alone contain sufficient information to explain scene-level aerosol optical depth at 550 nm (AOD550). Rather than attempting to construct a full aerosol retrieval framework, the analysis deliberately adopted a simplified predictor space in order to isolate the explanatory value of coarse atmospheric background variables. The methodological objective was therefore diagnostic rather than operational: to test the limits of meteorology-only modelling in a clearly defined scene-based setting.

To achieve this, a tabular dataset was constructed by integrating three information sources: hyperspectral satellite scene metadata, CAMS aerosol optical depth data, and ERA5 meteorological reanalysis variables12. The analytical unit was the individual satellite scene. For each scene, a single target value and a small set of aggregated meteorological predictors were derived using consistent temporal and spatial matching rules. This scene-level

---

[10] Van der Meer, F., van der Werff, H., van Ruitenbeek, F., Hecker, C., Bakker, W., Noomen, M., van der Meijde, M., Carranza, E. J. M., de Smeth, B., Woldai, T. Multi- and hyperspectral geologic remote sensing: A review. *International Journal of Applied Earth Observation and Geoinformation*, 2012, 14(1), 112–128. https://doi.org/10.1016/j.jag.2011.08.002

[11] Bakken, S., Henriksen, M. B., Birkeland, R., Garrett, J. L., Prentice, E. F., Honoré-Livermore, E., Langer, D. D., et al. HYPSO-1 CubeSat: First Images and In-Orbit Characterization. *Remote Sensing*, 2023, 15(3), 755. https://doi.org/10.3390/rs15030755

[12] Inness, A., Ades, M., Agustí-Panareda, A., Barré, J., Benedictow, A., Blechschmidt, A.-M., Dominguez, J. J., Engelen, R., Eskes, H., Flemming, J., Huijnen, V., Jones, L., Kipling, Z., Massart, S., et al. The CAMS reanalysis of atmospheric composition. *Atmospheric Chemistry and Physics*, 2019, 19(6), 3515–3556. https://doi.org/10.5194/acp-19-3515-2019

design follows a common strategy in Earth observation studies where heterogeneous gridded products are harmonized to a shared spatio-temporal support in order to enable comparative statistical modelling[13].

Scene Reference Framework Based on HYPSO Metadata

The common spatial and temporal reference for data integration was derived from scenes acquired by the HYPSO satellite mission. HYPSO-1 is a 6U CubeSat equipped with a hyperspectral imager operating in the visible and near-infrared range and developed to support compact hyperspectral Earth observation applications[14]. In the present study, however, the radiometric content of the hyperspectral imagery was not used directly for modelling. Instead, scene metadata were used exclusively to define the temporal and geographic context for linking meteorological and aerosol products. For each scene, the following metadata were used: acquisition timestamp, spatial extent expressed as a geographic bounding box, and scene identifier. These attributes provided a stable integration framework for cross-referencing atmospheric and meteorological products that differ in structure, resolution, and native dimensionality. The use of scene metadata in this role is methodologically justified because it allows the experiment to establish a controlled baseline before introducing direct spectral information in subsequent studies. A total of 33 scenes were retained in the final dataset. Each scene constituted a single observation in the analytical table. Only cases with complete and numerically valid data available across all integrated sources were included in the final analysis.

Target Variable: CAMS AOD550

The target variable was aerosol optical depth at 550 nm (AOD550), obtained from the Copernicus Atmosphere Monitoring Service (CAMS) reanalysis. CAMS provides globally consistent aerosol and atmospheric composition fields through an integrated modelling and data assimilation framework and is widely used in environmental and atmospheric studies requiring continuous spatio-temporal coverage[15]. In this study, CAMS AOD550 was treated as the reference aerosol descriptor associated with each scene. For each HYPSO scene, the CAMS product temporally closest to the acquisition timestamp was identified. A spatial subset corresponding to the scene bounding box was then extracted. Because the analytical unit of the study was the scene rather than the native grid cell, the subset was reduced to a single scalar by computing the mean AOD550 over the selected area. The resulting value represents a scene-level average of aerosol optical depth and serves as the regression target in the subsequent modelling stage. This aggregation step was intentional and consistent with the baseline nature of the experiment. It enabled direct comparison with similarly aggregated meteorological predictors, while simultaneously reducing the complexity of cross-source integration. At the same time, this simplification necessarily suppresses sub-scene spatial variability, which is important when interpreting the limits of the resulting models.

Predictor Variables: ERA5 Meteorological Data

Meteorological predictors were obtained from the ERA5 global reanalysis produced by the European Centre for Medium-Range Weather Forecasts (ECMWF) within the Copernicus Climate Change Service. ERA5 is one of the most widely used modern atmospheric reanalysis datasets and provides physically consistent global estimates of atmospheric state variables at high temporal resolution. Its broad adoption in environmental modelling makes it a suitable source of contextual meteorological information for scene-based aerosol analysis. Five variables were selected as candidate predictors: 2 m air temperature (t2m), surface pressure (sp), 10 m zonal wind component (u10), 10 m meridional wind component (v10), and total column water vapour (tcwv). These variables were chosen because they describe atmospheric conditions that may influence aerosol transport, dispersion, moisture-related particle growth, and optical behaviour[16]. Although none of them directly represents aerosol optical properties, they provide a compact characterization of the meteorological background associated with each scene. For each scene, ERA5 data were matched using the same temporal and spatial logic as applied to CAMS. The

---

[13] Prakash, M., Ramage, S., Kavzoglu, T., et al. Data integration in remote sensing and GIS. *International Journal of Image and Data Fusion*, 2010, 1(1), 3–15. https://doi.org/10.1080/19479830903561929

[14] Bakken, S., Henriksen, M. B., Birkeland, R., Garrett, J. L., Prentice, E. F., Honoré-Livermore, E., Langer, D. D., et al. HYPSO-1 CubeSat: First Images and In-Orbit Characterization. *Remote Sensing*, 2023, 15(3), 755. https://doi.org/10.3390/rs15030755

[15] Vogel, A., Alessa, G., Scheele, R., et al. Uncertainty in aerosol optical depth from modern aerosol-climate models, reanalyses, and satellite products. *Journal of Geophysical Research: Atmospheres*, 2022, 127(2), e2021JD035483. https://doi.org/10.1029/2021JD035483

[16] Zhang, W., Liu, S., Chen, X., et al. Estimation of All-Day Aerosol Optical Depth in the Beijing–Tianjin–Hebei Region Using Ground Air Quality Data. *Remote Sensing*, 2024, 16(8), 1410. https://doi.org/10.3390/rs16081410

temporally nearest available record was selected, the geographic subset corresponding to the scene bounding box was extracted, and the values were averaged over the scene extent. This produced one scalar value per variable for each scene. No additional spatial derivatives, higher-order interactions, or temporal lag variables were introduced. This decision was deliberate and reflects the role of the present study as a transparent baseline rather than an optimized predictive workflow.

Construction of the Final Analytical Dataset

The final analytical dataset consisted of 33 observations, each representing one hyperspectral acquisition scene. For every observation, the table included the scene identifier, acquisition time, scene bounding box, mean CAMS AOD550, and the five aggregated ERA5 variables. The entire dataset was therefore organized in a compact scene-level format suitable for exploratory statistics and supervised regression modelling. This data structure offers clear practical advantages. It simplifies the integration of heterogeneous sources, reduces dimensional mismatch between products, and makes the experiment directly interpretable. However, the adopted representation also imposes important constraints. By reducing both the target and the predictors to scene means, the approach removes local gradients, within-scene heterogeneity, and fine-scale spatial relationships that may be relevant for aerosol variability. These limitations are not incidental. They are central to the experimental design, because the purpose of the study is precisely to determine how much explanatory power remains when only coarse scene-level meteorological summaries are available.

Statistical Analysis

The first stage of the analysis was descriptive and exploratory. Summary inspection of the AOD550 distribution was performed in order to assess the empirical range, central tendency, and variability of the target variable across scenes. Pairwise relationships between AOD550 and the selected meteorological predictors were then examined using scatter plots and Pearson correlation coefficients. This step was intended to determine whether any individual meteorological variables exhibited visible linear association with the target and to provide an interpretable context for the subsequent regression modelling. In addition, a correlation matrix was used to inspect dependencies among predictors themselves. This was important because even in a small baseline dataset, moderate inter-variable correlation may affect the apparent contribution of individual predictors and complicate interpretation of linear model behaviour. The statistical analysis was therefore not used to establish causality, but to characterize the strength, direction, and structure of the observed associations within the adopted scene-level representation.

Baseline Regression Models

To evaluate whether the selected meteorological variables could support quantitative estimation of AOD550, two baseline regression models were applied: Linear Regression and Random Forest Regressor[17]. These models were intentionally chosen to provide a simple contrast between a fully linear and an inherently nonlinear approach. Linear Regression was used as the most transparent benchmark. It tests whether the relationship between scene-level meteorological predictors and AOD550 can be approximated by a linear combination of the available variables. Because of its simplicity and interpretability, it is a useful starting point in exploratory environmental modelling, especially when the main objective is to assess whether any meaningful signal exists in the predictor space rather than to maximize predictive accuracy. Random Forest was selected as a flexible nonlinear ensemble method capable of capturing interactions and non-monotonic relationships without requiring explicit functional assumptions[18]. Random Forest has been widely used in remote sensing and environmental prediction because of its robustness, relatively low sensitivity to distributional assumptions, and practical interpretability through feature importance analysis[19]. In the present study, however, it was not intended as a fully optimized predictive solution, but as a stronger baseline against which the limitations of the input space could be evaluated. The Random Forest Regressor was implemented using the scikit-learn library. The model was configured with 100 decision trees (n_estimators = 100) and a fixed random_state value of 42 to ensure reproducibility. Other parameters were kept at their default settings. No extensive hyperparameter optimization was performed because the primary objective of the study was to establish a transparent baseline model rather than maximize predictive performance. In

[17] Hastie, T., Tibshirani, R., Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed., Springer, New York, 2009. https://doi.org/10.1007/978-0-387-84858-7

[18] Breiman, L. Random forests. *Machine Learning*, 2001, 45, 5–32. https://doi.org/10.1023/A:1010933404324

[19] Chen, L., Ren, C., Zhang, B., et al. Mapping Spatial Variations of Structure and Function Parameters for Forest Condition Assessment of the Changbai Mountain National Nature Reserve. *Remote Sensing*, 2019, 11(24), 3004. https://doi.org/10.3390/rs11243004

addition, no feature normalization or standardization was applied, since tree-based methods are generally insensitive to feature scaling.

Training and Evaluation Procedure

The dataset was divided into training and testing subsets using a 70/30 split. Model fitting was performed on the training subset, while predictive performance was evaluated on the held-out test subset. Given the small size of the dataset, the resulting estimates should be interpreted as indicative rather than definitive. The purpose of this split was not to provide a final operational assessment, but to maintain a minimal separation between fitting and evaluation in line with standard predictive modelling practice. Model performance was assessed using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2)[20]. MAE quantifies the average absolute prediction error and provides an intuitive measure of model deviation from observed values. RMSE gives greater weight to larger errors and is therefore more sensitive to occasional strong prediction failures. R^2 indicates the proportion of target variance explained by the model relative to a naive mean-based predictor. In the context of this study, negative R^2 values are particularly informative because they indicate that the model performs worse than simply predicting the mean AOD550 for all scenes. To further interpret the nonlinear model, feature importance values were extracted from the Random Forest. These values were used only as heuristic indicators of relative predictor contribution within the fitted ensemble. They were not interpreted causally, but rather as an additional diagnostic tool showing which meteorological descriptors contributed most strongly to variance reduction under the adopted modelling setup.

Methodological Scope and Limitations

The methodological scope of the study is intentionally narrow. First, all analyses are conducted at the scene level, with both predictors and target represented by spatial means. This makes the dataset compact and internally consistent, but eliminates fine-scale variability that may be relevant for aerosol processes. Second, the predictor space contains no direct spectral, radiometric, or textural information from hyperspectral imagery. As a result, the models are not expected to capture aerosol-specific optical signatures, surface–atmosphere interactions visible in spectral space, or other forms of information that are central to more advanced retrieval strategies. These limitations are not weaknesses of implementation but deliberate features of the experimental design. The study is intended to quantify what can be learned from coarse meteorological context alone and, by doing so, to provide a justified baseline for future work involving richer feature spaces. In this sense, limited predictive performance should be interpreted not merely as model failure, but as evidence of insufficient physical information content in the simplified predictor representation.

An important limitation of the present study is the relatively small sample size consisting of 33 scenes. Because the dataset covers only a limited number of observations, the resulting regression metrics may be sensitive to the particular train–test split and individual scene variability. Consequently, the reported performance should be interpreted as indicative rather than statistically definitive. The primary purpose of the experiment was not to establish a robust operational prediction framework, but rather to evaluate whether coarse scene-level meteorological descriptors contain sufficient explanatory information for AOD550 estimation. In this context, the observed instability and limited predictive performance are themselves informative, as they highlight the limitations of the simplified predictor representation.

Results

Overview of Scene-Level AOD550 Variability

---

[20] Willmott, C. J., Matsuura, K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, 2005, 30, 79–82. https://doi.org/10.3354/cr030079

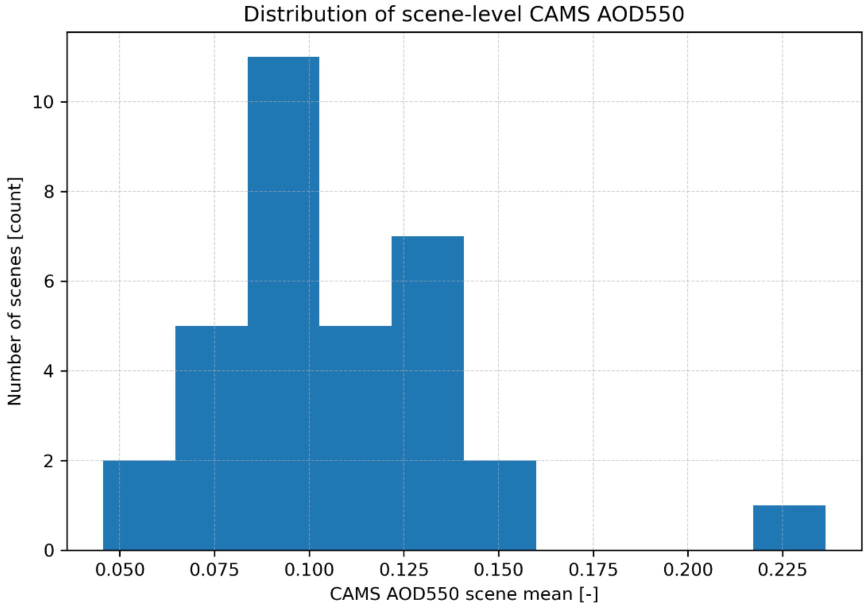

Figure 1 Distribution of scene-level aerosol optical depth at 550 nm (AOD550) derived from CAMS data.

The scene-level AOD550 values derived from CAMS showed a relatively narrow distribution, with most observations concentrated within a low-to-moderate aerosol loading range. The majority of scenes clustered approximately between 0.08 and 0.13, while only a limited number of cases extended toward higher values, producing a right-skewed distribution with a single clearly elevated episode (Figure 1). This indicates that the analysed dataset is dominated by typical background or moderately enhanced aerosol conditions rather than by strongly polluted or optically extreme situations.

From an analytical perspective, this distribution is important because it constrains the learning problem. The predictive models were trained primarily on scenes representing moderate aerosol regimes, whereas the number of high-AOD cases remained small. As a result, the dataset is appropriate for evaluating whether scene-level meteorological descriptors contain a stable baseline signal under ordinary conditions, but it is less informative for assessing model behaviour under rare or strongly elevated aerosol episodes.

The temporal evolution of AOD550 was also characterized by moderate variability and several local increases, but without a clear monotonic or seasonal trend across the analysed scenes. This suggests that the observed aerosol variability was driven primarily by episodic and irregular atmospheric conditions rather than by a simple large-scale temporal pattern. Such behaviour is inherently difficult to capture using static scene-level predictors alone, because it may reflect transient interactions between transport, moisture, accumulation, and aerosol source variability that are not explicitly encoded in the adopted feature space.

Pairwise Relationships Between AOD550 and Meteorological Predictors

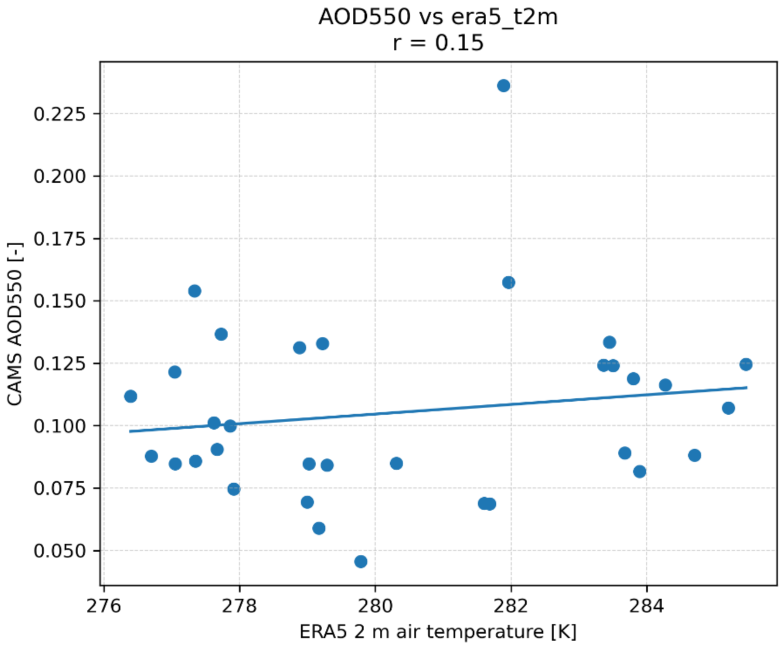

Figure 2 Scatter plot of AOD550 versus ERA5 2 m air temperature (t2m) with linear regression fit (r = 0.15)

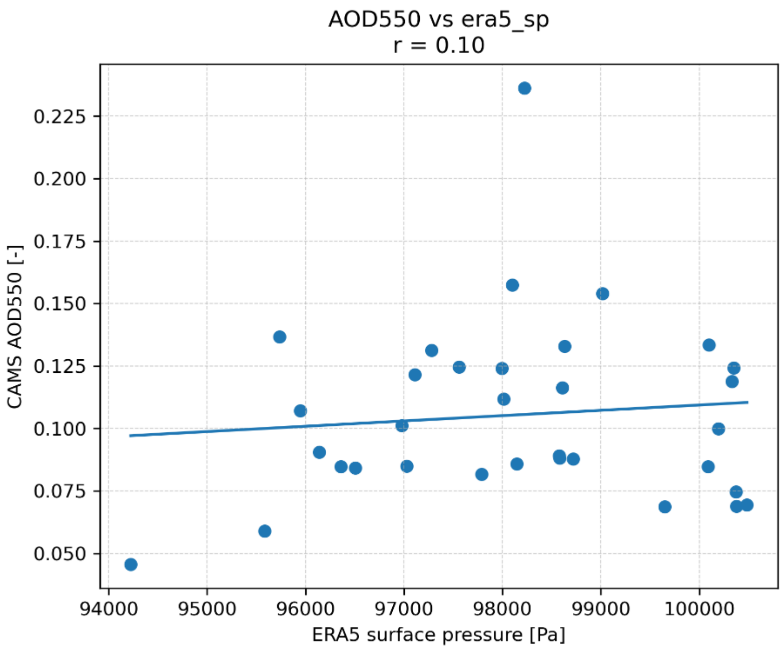

Figure 3 Scatter plot of AOD550 versus ERA5 surface pressure (sp) with linear regression fit (r = 0.10).

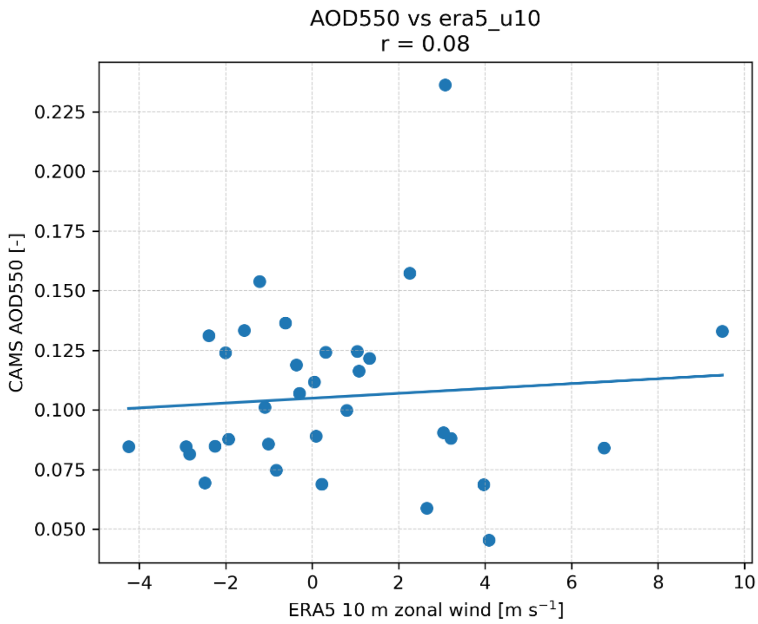

Figure 4 Scatter plot of AOD550 versus ERA5 10 m zonal wind (u10) with linear regression fit (r = 0.08).

The pairwise analysis showed that most meteorological variables had only weak direct relationships with scene-level AOD550. Near-surface air temperature (t2m) exhibited a weak positive association, indicating that higher temperatures were not consistently linked to substantially higher aerosol optical depth (Figure 2). Surface pressure (sp) showed an almost negligible relationship with the target variable, suggesting that pressure, when treated as an isolated scene-level descriptor, carried little explanatory value for AOD550 variability (Figure 3). A similarly weak pattern was observed for the zonal wind component (u10), whose correlation with AOD550 remained close to zero (Figure 4).

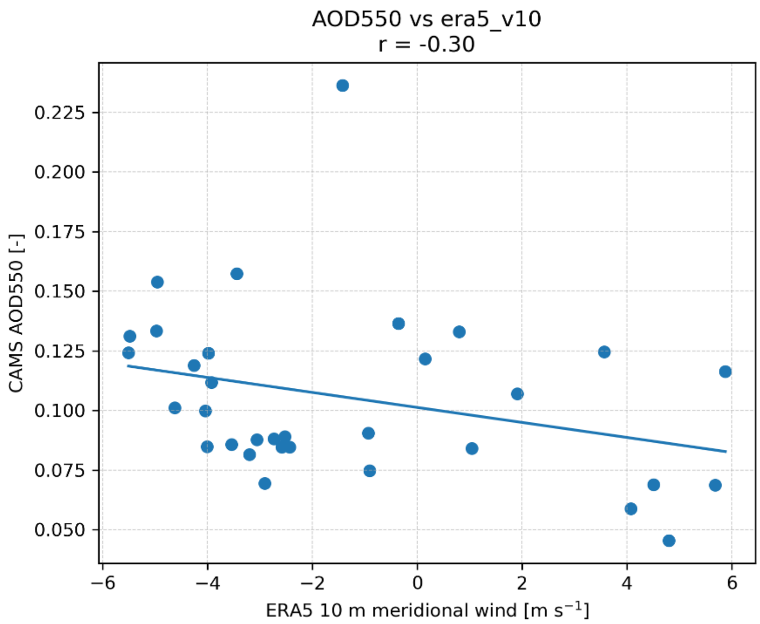

Figure 5 Scatter plot of AOD550 versus ERA5 10 m meridional wind

(v10) with linear regression fit (r = −0.30).

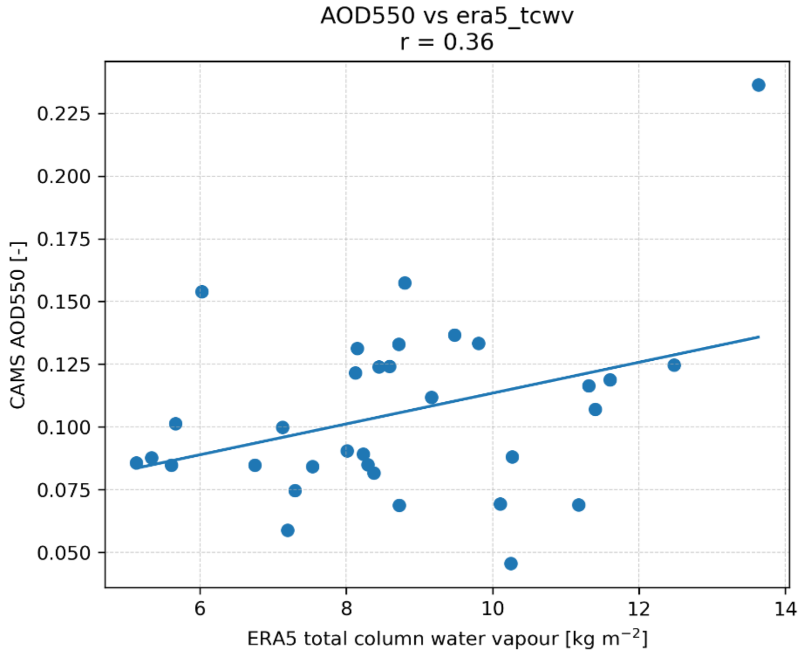

Figure 6 Scatter plot of AOD550 versus ERA5 total column water vapour (tcwv) with linear regression fit (r = 0.36).

Among all analysed predictors, the meridional wind component (v10) and total column water vapour (tcwv) displayed the clearest associations with AOD550. The relationship with v10 was negative and of moderate magnitude, indicating that stronger meridional flow tended to coincide with lower AOD550 values (Figure 5). Although the dispersion of observations remained substantial, this pattern suggests that north–south transport or ventilation-related processes may contribute to scene-level aerosol variability more strongly than the other meteorological variables considered individually. In contrast, tcwv showed the strongest positive association with AOD550, implying that scenes with higher atmospheric moisture generally corresponded to higher aerosol optical depth (Figure 6). This is physically plausible, since increased water vapour may enhance aerosol optical thickness through hygroscopic particle growth and may also be associated with atmospheric conditions favouring pollutant persistence. Despite these two relatively stronger signals, none of the individual pairwise relationships approached the level of a dominant explanatory dependency. The scatter remained broad in all cases, which indicates that even the most informative meteorological variables explain only part of the observed variability. Taken together, these results suggest that meteorological context influences AOD550 in the analysed scenes, but does not determine it in a sufficiently direct or stable way to support accurate estimation from isolated scene-level descriptors alone.

Correlation Structure of the Predictor Space

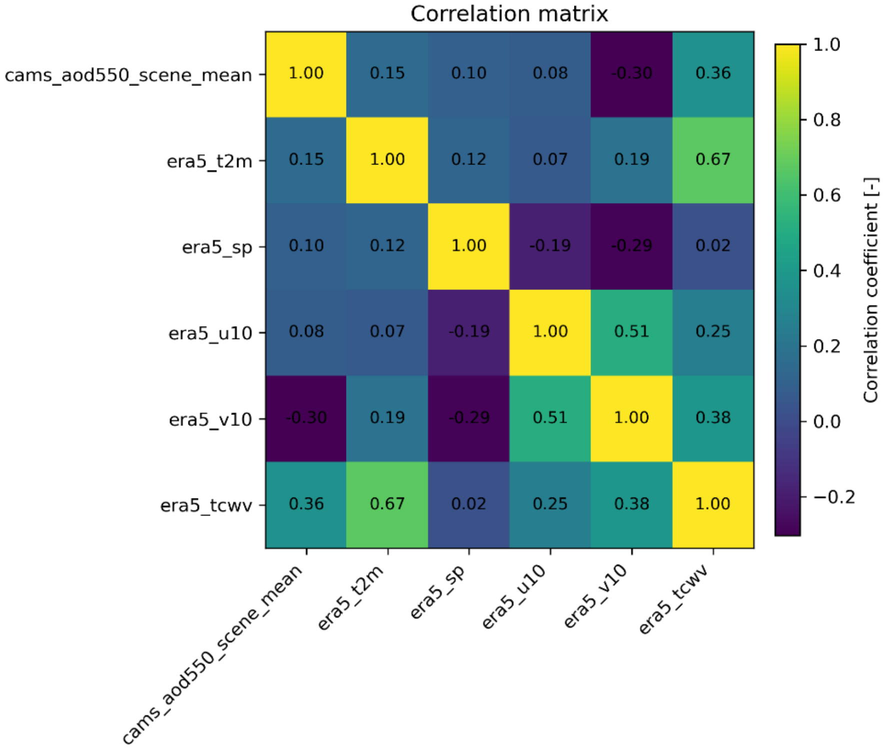

Figure 1 Correlation matrix of AOD550 and selected ERA5 variables

showing weak relationships with AOD and moderate inter-variable dependencies.

The correlation matrix confirmed the weak coupling between AOD550 and most scene-level meteorological variables (Figure 7). The strongest positive relationship with the target was observed for tcwv, while the clearest negative relationship was found for v10. The remaining variables showed only weak or negligible associations with AOD550, consistent with the visual interpretation of the scatter plots. At the same time, the matrix revealed moderate dependencies among the predictors themselves. In particular, t2m and tcwv were positively correlated, indicating that warmer scenes were generally associated with higher atmospheric moisture content. A moderate relationship was also visible between the two wind components, which suggests some coherence in the circulation patterns across the analysed observations. These interdependencies are relevant for regression modelling because they imply that the predictor set is not fully independent. Consequently, part of the apparent contribution of one variable may overlap with information already carried by another. Overall, the correlation structure supports the central interpretation of the study. The predictor space contains some physically meaningful atmospheric signal, especially in relation to moisture and transport, but this signal is weak, partially redundant, and insufficiently rich to reconstruct aerosol optical depth with high fidelity at the scene level.

Spatial Distribution of Scenes

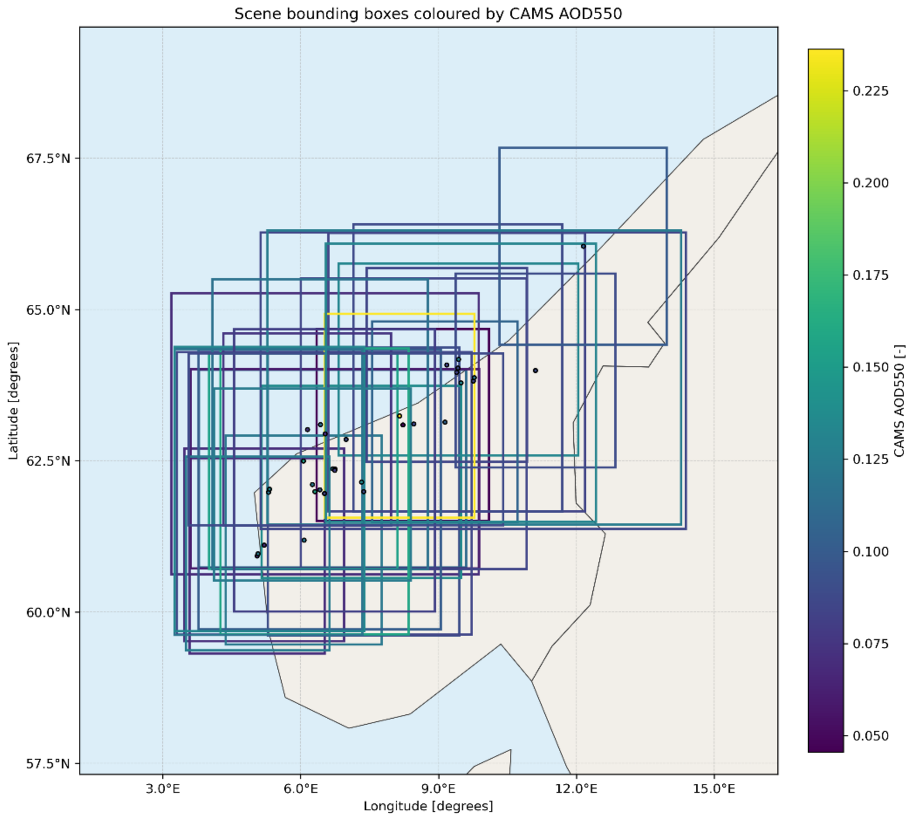

Figure 2 Spatial distribution of analysed satellite scenes represented as bounding boxes, coloured according to CAMS AOD550 values.

The analysed scenes were concentrated within a relatively consistent geographic domain, with substantial overlap among bounding boxes (Figure 8). This repeated coverage of similar areas is methodologically useful because it reduces the confounding effect of large spatial shifts between observations and allows the analysis to focus more directly on changing atmospheric conditions across comparable locations. The distribution of AOD550 across the mapped scenes was broadly consistent with the histogram and time-series analysis. Most scenes corresponded to low or moderate aerosol optical depth, while higher-AOD cases appeared only occasionally and did not form a distinct spatial cluster. Instead, elevated values occurred within the same general observation region as lower values. This suggests that the observed aerosol variability was not dominated by a fixed geographic hotspot, but rather by temporally varying atmospheric conditions such as transport episodes, humidity changes, or short-term accumulation regimes. This repeated observation of comparable spatial domains under different aerosol conditions is valuable from the standpoint of future work. It indicates that part of the variability captured in the dataset reflects changing atmospheric states over recurrently sampled areas rather than purely spatial contrasts between different locations.

Baseline Regression Performance

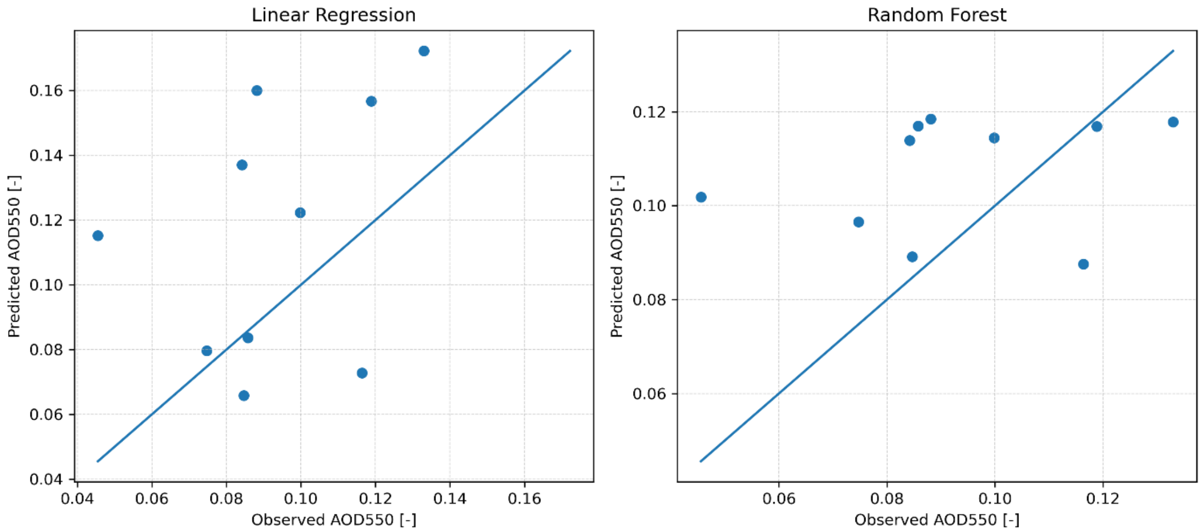

Figure 3 Observed versus predicted AOD550 for linear regression and Random Forest models, showing limited predictive performance and deviation from the 1:1 reference line.

To evaluate the predictive usefulness of scene-level meteorological variables, two baseline regression models were applied: Linear Regression and Random Forest (Figure 9). Both models used the same ERA5-derived predictors: t2m, sp, u10, v10, and tcwv. In both cases, the predictive performance was limited, indicating that the selected meteorological variables did not contain sufficient information for reliable estimation of AOD550. The Linear Regression model showed weak agreement between observed and predicted values. Predictions remained concentrated in a narrow range and deviated substantially from the 1:1 reference line, indicating that the linear model was unable to reproduce the amplitude of the observed aerosol variability. This behaviour suggests that the relationship between meteorological background conditions and scene-level AOD550 cannot be captured adequately by a simple linear combination of the selected predictors. The Random Forest model provided only a modest qualitative improvement. Some predictions were closer to the reference line than in the linear case, indicating that limited nonlinear structure was present in the predictor space. However, the model still showed a strong regression-to-the-mean effect: higher observed AOD550 values were generally underestimated, while lower values were often overestimated. The predicted values remained compressed relative to the observed range, demonstrating that even the more flexible nonlinear model was unable to recover the full variability of the target variable. This pattern is particularly important because it suggests that the primary limitation does not lie only in model form. If the weak performance were mainly due to linearity assumptions, a nonlinear ensemble should have produced a more substantial gain. Instead, the limited improvement of Random Forest indicates that the main constraint is the restricted physical information content of the predictors themselves. Where available, the numerical evaluation metrics should be reported directly here. In the present formulation, both MAE and RMSE indicated non-negligible prediction error, while the coefficient of determination remained low and, in at least part of the analysis, negative. A negative R^2 is especially informative in this context because it means that the fitted model performs worse than a naive predictor that simply returns the mean AOD550 for all scenes. This confirms that the meteorology-only scene representation is not sufficiently informative for reliable baseline estimation of the target variable.

Random Forest Feature Importance

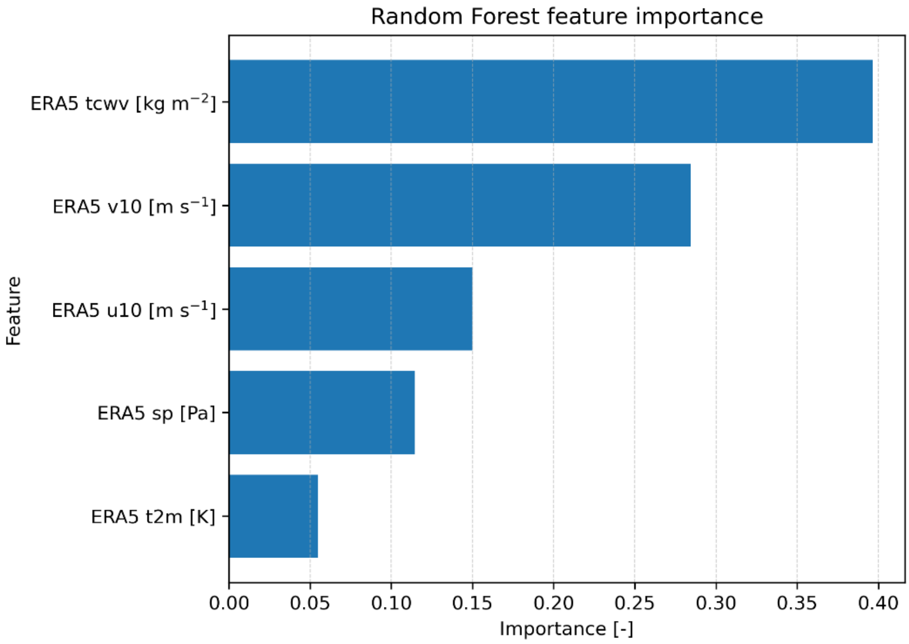

Figure 4 Random Forest feature importance showing tcwv and v10 as the most influential predictors of AOD550.

The Random Forest feature importance analysis showed that total column water vapour was the most influential predictor, followed by the meridional wind component. The zonal wind component occupied an intermediate position, whereas surface pressure and especially near-surface air temperature contributed less strongly. This ranking is consistent with the earlier pairwise analysis and suggests that the most informative meteorological signal in the dataset is associated with moisture conditions and transport-related processes (Figure 10). However, the feature importance results should not be overinterpreted. The identification of relatively more important variables within a poorly performing model does not imply that these predictors are strong in an absolute sense. Rather, it shows only that, within an already limited feature space, tcwv and v10 contribute more to variance reduction than the remaining variables. Their relative importance is therefore diagnostically useful, but it does not alter the broader conclusion that the overall predictive content of the meteorological predictor set remains insufficient.

Synthesis of Results

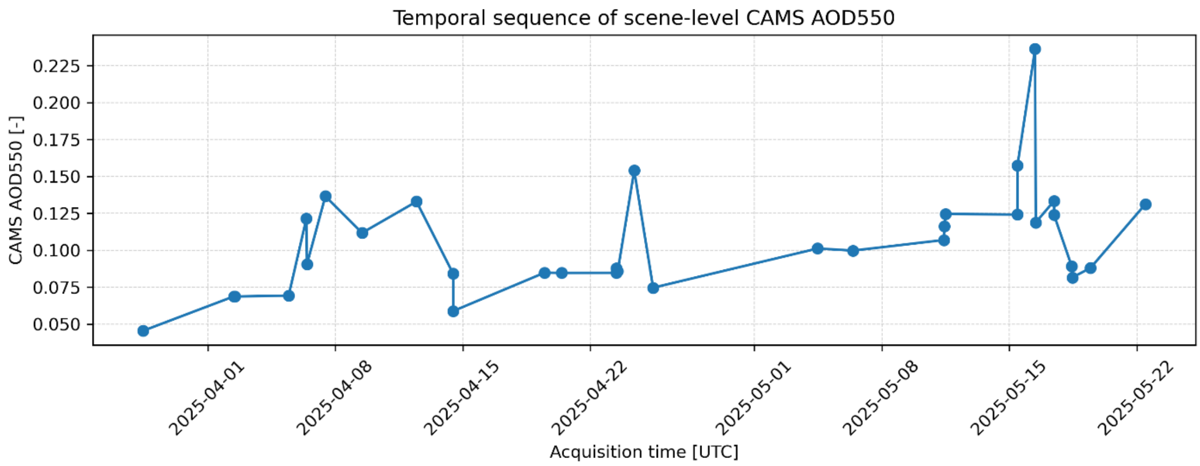

Figure 5 Time series of scene-level AOD550 showing moderate variability and occasional peaks without a clear temporal trend.

Taken together, the results converge on a consistent interpretation. Scene-level meteorological variables do contain weak and physically plausible information related to aerosol optical depth, particularly through moisture-related effects and meridional transport (Figure 11). However, this information is incomplete and too coarse to support reliable quantitative estimation of AOD550. The pairwise analysis revealed only weak to moderate associations, the predictor space showed partial redundancy, and both baseline regression models failed to reproduce the observed variability with satisfactory fidelity. These findings support the central premise of the study: scene-level meteorological predictors are insufficient for AOD550 estimation when used in isolation. In this sense, the negative modelling outcome is not merely a failed regression exercise, but an informative empirical result. It demonstrates that the adopted predictor representation lacks the physical richness required to explain aerosol optical behaviour and thereby provides a justified baseline for future studies incorporating more informative optical or spectral data sources.

Study Limitations

One of the main limitations of the present study is the relatively small dataset consisting of only 33 scenes. Because of the limited number of observations, the resulting regression metrics may be sensitive to individual scenes and to the particular train–test split applied during evaluation. Consequently, the reported predictive performance should be interpreted as exploratory rather than statistically definitive. An additional limitation results from the scene-level aggregation strategy adopted throughout the study. Both the target variable and the meteorological predictors were reduced to spatial mean values, which simplified cross-source integration but simultaneously removed fine-scale spatial heterogeneity, local gradients, and sub-scene atmospheric variability. As a result, potentially important spatial relationships relevant to aerosol behaviour were not represented in the final predictor space. The study is also limited by the intentionally simplified feature representation. Only coarse meteorological variables derived from ERA5 were used as predictors, while no direct spectral, radiometric, or textural information from hyperspectral imagery was incorporated into the regression models. Consequently, the models had no access to aerosol-specific optical signatures that are likely required for more accurate AOD550 estimation. Another important limitation concerns the restricted geographic and environmental diversity of the analysed scenes. Most observations represented relatively low or moderate aerosol conditions, while strongly elevated aerosol episodes were rare. Therefore, the presented results may not generalize to highly polluted environments or to substantially different atmospheric regimes. Finally, the methodological objective of the study was intentionally diagnostic rather than operational. The models were designed as transparent baseline experiments intended to evaluate the explanatory limits of meteorology-only predictors. Future work should therefore focus on larger and more diverse datasets, cross-validation strategies, and the integration of physically informative hyperspectral features within multi-source data fusion frameworks.

References

- IPCC. Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, 2021.
- Kaufman, Y. J., Tanré, D., Boucher, O. A satellite view of aerosols in the climate system. Nature, 2002, 419(6903), 215–223. https://doi.org/10.1038/nature01091
- 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., Lavenu, F., Jankowiak, I., Smirnov, A. AERONET—A federated instrument network and data archive for aerosol characterization. Remote Sensing of Environment, 1998, 66(1), 1–16. https://doi.org/10.1016/S0034-4257(98)00031-5
- Levy, R. C., Mattoo, S., Munchak, L. A., Remer, L. A., Sayer, A. M., Patadia, F., Hsu, N. C. The Collection 6 MODIS aerosol products over land and ocean. Atmospheric Measurement Techniques, 2013, 6(11), 2989–3034. https://doi.org/10.5194/amt-6-2989-2013
- Vogel, A., Alessa, G., Scheele, R., et al. Uncertainty in aerosol optical depth from modern aerosol-climate models, reanalyses, and satellite products. Journal of Geophysical Research: Atmospheres, 2022, 127(2), e2021JD035483. https://doi.org/10.1029/2021JD035483
- Inness, A., Ades, M., Agustí-Panareda, A., Barré, J., Benedictow, A., Blechschmidt, A.-M., Dominguez, J. J., Engelen, R., Eskes, H., Flemming, J., Huijnen, V., Jones, L., Kipling, Z., Massart, S., et al. The CAMS reanalysis of atmospheric composition. Atmospheric Chemistry and Physics, 2019, 19(6), 3515–3556. https://doi.org/10.5194/acp-19-3515-2019
- Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., et al. The ERA5 global reanalysis. Quarterly Journal of the Royal Meteorological Society, 2020, 146(730), 1999–2049. https://doi.org/10.1002/qj.3803
- Zhao, C., Liu, Z., Wang, Q., Ban, J., Chen, N., Li, T. High-resolution daily AOD estimated to full coverage using the random forest model approach in the Beijing–Tianjin–Hebei region. Atmospheric Environment, 2019, 203, 70–78. https://doi.org/10.1016/j.atmosenv.2019.01.033
- Chang, H. H., Hu, X., Liu, Y., et al. Calibrating MODIS aerosol optical depth for predicting daily PM2.5 concentrations via statistical downscaling. Journal of Exposure Science & Environmental Epidemiology, 2014, 24(4), 398–404. https://doi.org/10.1038/jes.2013.90
- Van der Meer, F., van der Werff, H., van Ruitenbeek, F., Hecker, C., Bakker, W., Noomen, M., van der Meijde, M., Carranza, E. J. M., de Smeth, B., Woldai, T. Multi- and hyperspectral geologic remote sensing: A review. International Journal of Applied Earth Observation and Geoinformation, 2012, 14(1), 112–128. https://doi.org/10.1016/j.jag.2011.08.002
- Bakken, S., Henriksen, M. B., Birkeland, R., Garrett, J. L., Prentice, E. F., Honoré-Livermore, E., Langer, D. D., et al. HYPSO-1 CubeSat: First Images and In-Orbit Characterization. Remote Sensing, 2023, 15(3), 755. https://doi.org/10.3390/rs15030755

- Inness, A., Ades, M., Agustí-Panareda, A., Barré, J., Benedictow, A., Blechschmidt, A.-M., Dominguez, J. J., Engelen, R., Eskes, H., Flemming, J., Huijnen, V., Jones, L., Kipling, Z., Massart, S., et al. The CAMS reanalysis of atmospheric composition. Atmospheric Chemistry and Physics, 2019, 19(6), 3515–3556. https://doi.org/10.5194/acp-19-3515-2019

- Prakash, M., Ramage, S., Kavzoglu, T., et al. Data integration in remote sensing and GIS. International Journal of Image and Data Fusion, 2010, 1(1), 3–15. https://doi.org/10.1080/19479830903561929

- Bakken, S., Henriksen, M. B., Birkeland, R., Garrett, J. L., Prentice, E. F., Honoré-Livermore, E., Langer, D. D., et al. HYPSO-1 CubeSat: First Images and In-Orbit Characterization. Remote Sensing, 2023, 15(3), 755. https://doi.org/10.3390/rs15030755
- Vogel, A., Alessa, G., Scheele, R., et al. Uncertainty in aerosol optical depth from modern aerosol-climate models, reanalyses, and satellite products. Journal of Geophysical Research: Atmospheres, 2022, 127(2), e2021JD035483. https://doi.org/10.1029/2021JD035483

- Zhang, W., Liu, S., Chen, X., et al. Estimation of All-Day Aerosol Optical Depth in the Beijing–Tianjin–Hebei Region Using Ground Air Quality Data. Remote Sensing, 2024, 16(8), 1410. https://doi.org/10.3390/rs16081410
- Hastie, T., Tibshirani, R., Friedman, J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed., Springer, New York, 2009. https://doi.org/10.1007/978-0-387-84858-7
- Breiman, L. Random forests. Machine Learning, 2001, 45, 5–32. https://doi.org/10.1023/A:1010933404324

- Chen, L., Ren, C., Zhang, B., et al. Mapping Spatial Variations of Structure and Function Parameters for Forest Condition Assessment of the Changbai Mountain National Nature Reserve. Remote Sensing, 2019, 11(24), 3004. https://doi.org/10.3390/rs11243004
- Willmott, C. J., Matsuura, K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate Research, 2005, 30, 79–82. https://doi.org/10.3354/cr030079