RESEARCH ARTICLE

Open Access

Spatiotemporal aerosol prediction model based on fusion of machine learning and spatial analysis

Kwon-Ho Lee[1,2]* , Seong-Hun Pyo[3] and Man Sing Wong[4]

Abstract

This study examined long-term aerosol optical thickness (AOT) data from the Moderate Resolution Imaging Spectroradiometer (MODIS) to quantify aerosol conditions on the Korean Peninsula. Time-series machine learning (ML) techniques and spatial interpolation methods were used to predict future aerosol trends. This investigation utilized AOT data from Terra MODIS and meteorological data from Automatic Weather System (AWS) in eight selected cities in Korea (Gangneung, Seoul, Busan, Wonju, Naju, Jeonju, Jeju, and Baengyeong) to assess atmospheric aerosols from 2000 to 2021. A machine-learning-based AOT prediction model was developed to forecast future AOT using long-term observations. The accuracy analysis of the AOT prediction results revealed mean absolute error of 0.152 ± 0.15, mean squared error of 0.048 ± 0.016, bias of 0.002 ± 0.011, and root mean squared error of 0.216 ± 0.038, which are deemed satisfactory. By employing spatial interpolation, gridded AOT values within the observation area were generated based on the ML prediction results. This study effectively integrated the ML model with point-measured data and spatial interpolation for an extensive analysis of regional AOT across the Korean Peninsula. These findings have substantial implications for regional air pollution policies because they provide spatiotemporal AOT predictions.

Keywords Satellite, Aerosol, Machine learning, Spatial analysis, Remote sensing

1 Introduction

Atmospheric aerosols vary in size and are composed of a wide range of materials, including dust, soot, sea salt, and sulfate particles. These aerosols play an important role in Earth's climate system, air quality, and environmental processes, making them a topic of considerable scientific interest (IPCC 2021, 2023). The importance of atmospheric aerosols extends beyond climate change and human health considerations (Lelieveld et al., 2015). They influence atmospheric chemistry, contribute to cloud formation and precipitation, affect visibility, and play roles in nutrient transport in terrestrial and aquatic ecosystems. Many studies in this field continue to evolve, providing insights into the intricate interactions among aerosols, climate, and the environment, which are vital for understanding and addressing the complex challenges of global climate change and air quality management (Li et al., 2022; Menon et al., 2008).

In addition, long-range transport of aerosols and photochemical reactions in the atmosphere have complex effects (Chakraborty et al., 2021; Oh et al., 2015). Aerosol particles can remain in the atmosphere for a few minutes to more than a week, making it difficult to accurately quantify their spatiotemporal distribution. Comprehensive instrumental observation and analysis

*Correspondence:
Kwon-Ho Lee
kwonho.lee@gmail.com
[1] Department of Atmospheric and Environmental Sciences, Gangneung-Wonju National University (GWNU), Gangneung 25457, Republic of Korea
[2] Research Institute of Radiation and Satellite, Gangneung-Wonju National University (GWNU), Gangneung 25457, Republic of Korea
[3] National Institute of Meteorological Sciences, Seogwipo 63568, Republic of Korea
[4] Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Hong Kong, China

Lee *et al. Asian Journal of Atmospheric Environment*   (2024) 18:9

Page 2 of 15

of their chemical species and physical properties are necessary to identify the sources or origins of aerosols and assess their optical properties, as well as their impact on climate. In particular, physical quantities, such as atmospheric aerosol optical thickness (AOT), have long been used because they contain information on the total quantity and optical properties of particles in the atmosphere (Lee et al., 2009, 2022; Li et al., 2009). AOT can be derived using satellite- and ground-based observations. The representative ground-based observation network (AERONET) provides reliable continuous AOT values from observation points worldwide (Holben et al., 1998), but the limited number of observation points makes regional analysis difficult. Satellites have been launched for various purposes and are widely used to study spatiotemporal changes in aerosol properties (Lee et al., 2022; Li et al., 2021). Therefore, it is possible to estimate geospatial states and evaluate air quality levels using multi-platform monitoring.

To date, a large amount of data has been accumulated through various Earth observation satellites, and there is a need to quickly reprocess and generate the final products (Lee et al., 2022). Existing mathematical or physical data analysis methods have limitations due to the application of complex formulas and repetitive calculation processes. Recent machine learning (ML) and deep learning (DL) techniques have enabled more efficient data analyses. Therefore, the demand for research using these techniques in the field of remote sensing is increasing and various algorithms are being developed (Adegun et al., 2023; Murphy, 2012). However, spatial analysis techniques allow homogeneous values to be estimated from spatially inhomogeneous measurements, thereby enabling the evaluation of continuous information over a region of interest. Examples include spatial air quality estimation studies using spatial analysis techniques (Wong et al., 2021; Jerette et al. 2005), spatiotemporal analysis using reanalysis model data over long periods of time (Sun et al., 2019), and spatiotemporal distribution and variation analysis of AOT using remotely sensed measurements (Lee, 2018; Torres et al., 2002; Yu et al., 2022).

In this study, long-term AOT data from Moderate Resolution Imaging Spectroradiometer (MODIS) observations were analyzed to statistically quantify the aerosol status of the Korean Peninsula and obtain AOT observation data for areas without ground observation networks. To estimate the future status of regional aerosols, we applied time-series ML and spatial interpolation techniques to evaluate the AOT prediction accuracy and future trends in the studied areas. These results are expected to serve as a guide for the current status and future outlook of aerosols on the Korean Peninsula.

2 Data and methodology

Long-term satellite- and ground-based meteorological data provide important information regarding aerosol composition, emission sources, and transport pathways. Satellites provide global or regional data, whereas ground-based stations provide local, detailed insights. The integration of both datasets ensured a comprehensive understanding of aerosol dynamics across various spatial scales. Satellites track changes in aerosol distribution and behavior over time by providing frequent snapshots of atmospheric conditions, whereas ground-based meteorological data capture short-term fluctuations and trends at a high resolution. Both satellite- and ground-based observations assist in identifying aerosol emission sources, whether natural or anthropogenic. Combining data from multiple sources is crucial for understanding their impacts on air quality and climate. These data are essential for tracking the long-range transport of aerosols and for identifying pathways. Furthermore, integrated observations provide comprehensive data to improve the accuracy and reliability of atmospheric models, understanding aerosol dynamics, and predictive capabilities of atmospheric models.

The study area is a region between 20°N and 30°N latitude and 120°E and 130°E longitude, and includes South Korea (Fig. 1). Eight points were selected in the study area as points of interest: Gangneung (128.89°E, 37.75°N), Wonju (127.94°E, 37.34°N), Seoul (126. 69°E, 37.57°N), Baengnyeong (124.71°E, 37.97°N), Jeonju (127.11°E, 35.84°N), Naju (126.90°E, 35.02°N), Busan (129.03°E, 35.10°N), and Jeju (126.53°E, 33.514°N). The population of each region is listed in the following order: Seoul (9,417,469) > Busan (3,295,760) Jeonju (665,884) > Wonju (361,810) > Gangneung (215,128) > Naju (114,785) > Baengnyeong (5000) (population data are available from the Korea Statistical Information Server [https://kosis.kr/index/index.do]).

Figure 2 shows a flowchart of the data processing and analysis used in this study. In these data processing steps, satellite data and ground-based Automatic Weather System (AWS) observation data at eight selected locations were collected, and preprocessing was performed to modify the data for integrated analysis. The spatial collocation between the two datasets was performed by selecting the satellite pixel data closest to the latitude and longitude of the AWS observations. The temporal matching is done by averaging the AWS observations for ± 1 h before and after the satellite observation time. In the analysis step, we analyzed the AOT ranges according to local weather conditions to investigate the influence of weather characteristics on aerosols and compared the characteristics at each point. After the time-series forecasting model utilizing the ML technique was built and the performance accuracy of the model was verified, the

Lee *et al. Asian Journal of Atmospheric Environment*   (2024) 18:9

Page 3 of 15
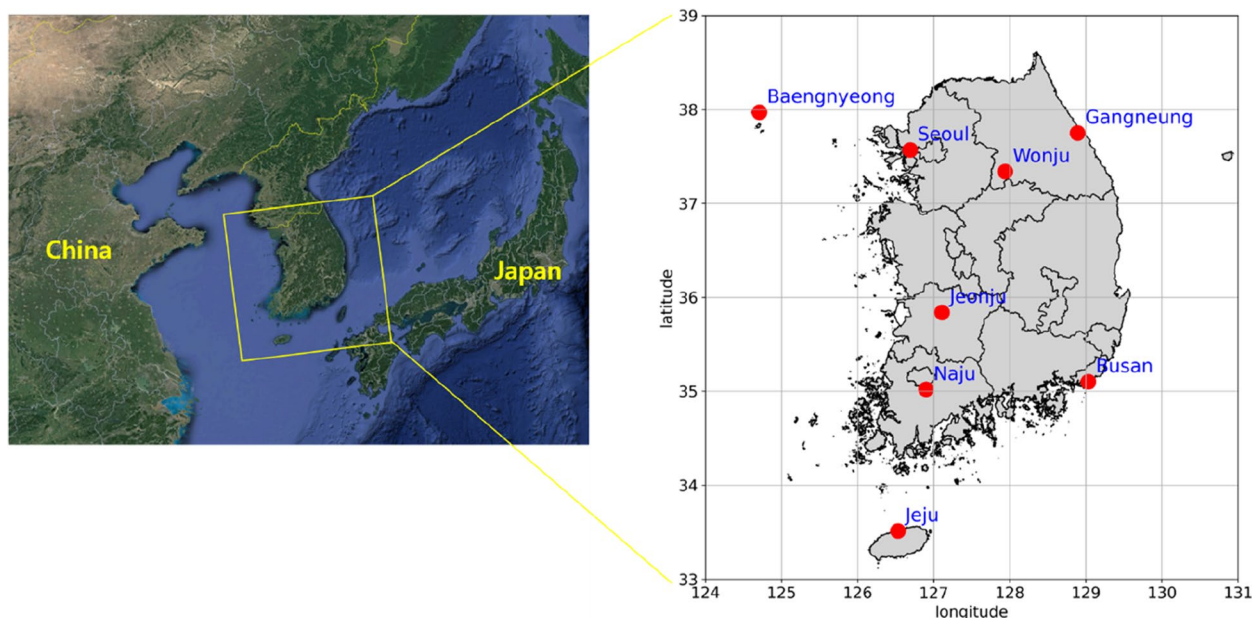
Fig. 1 Map of study area. Selected eight points of interest show as red circles (Gangneung [128.89°E, 37.75°N], Wonju [127.94°E, 37.34°N], Seoul [126.69°E, 37.57°N], Baengnyeong [124.71°E, 37.97°N], Jeonju [127.11°E, 35.84°N], Naju [126.90°E, 35.02°N], Busan [129.03°E, 35.10°N], and Jeju [126.53°E 33.514°N])
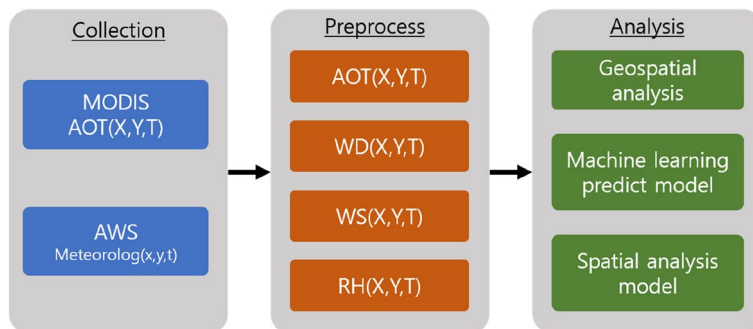
Fig. 2 Schematic diagram of data processing flow for AOT status analysis and time-series machine learning technique. The AWS, WD, WS, and RH stand for ground-based Automatic Weather System, wind direction, wind speed, and relative humidity, respectively

future outlook for aerosols was predicted. Since the predicted values are for each selected point, a reanalysis was performed with gridded values across the observation area through spatial extrapolation.

2.1 Data

Satellite observations provide information on environmental changes and status over large areas and are therefore useful for obtaining data on the spatiotemporal distribution and characteristics of aerosols. The satellite data used in this study were the Terra MODIS level 2 aerosol product (version 6.1) (Levy et al., 2017) and wind direction, wind speed, and relative humidity (RH) data from AWS observations. Table 1 lists the data used in this study, including Terra MODIS aerosol and AWS observation data.

The MODIS AOT data used in this study are among the most widely used satellite products. MODIS is a multipurpose sensor aboard the oldest Earth observation satellites currently in operation: the EOS-AM1 (Terra) (launched December 18, 1999) and EOS-PM1 (Aqua) (launched May 4, 2002). Among the various Earth observation satellites, MODIS measures multispectral radiation in the visible to long-wave infrared wavelength region. The most common aerosol-related output from satellite observations is AOT, which is the amount of radiation attenuated by aerosol particles in the atmosphere. In general, the AOT is determined by analyzing

Table 1 List of the parameters, resolution, and duration of satellite and AWS data used in the study

Instrument	Parameters	Resolution	Period	Source
Terra/MODIS	Aerosol optical thickness (AOT)	1–2 times/day, 10 km^2/ pixel	2000–2021	https://ladsweb. modaps.eosdis.nasa. gov/
Automatic Weather System (AWS)	Wind speed (WS), wind direction (WD), relative humidity (RH)	Hourly	2000–2021	https://data.kma.go.kr

the radiative transfer process of sunlight-reflected light in the visible region, which can be retrieved using Eq. (1).

$$\rho_{\text{TOA}}(\tau, \lambda) = \rho_{\text{Aer}}(\tau, \lambda) + \rho_{\text{Ray}}(p, t, \lambda) + \frac{T_0(\tau_{\text{gas}}, \lambda) \cdot T_S(\tau_{\text{gas}}, \lambda) \cdot \rho_{\text{Sfc}}(\lambda)}{1 - \rho_{\text{Sfc}}(\lambda) \cdot r_h(\tau, \lambda)} \tag{1}$$

where ρ_{TOA}, ρ_{Aer}, ρ_{Ray}, and ρ_{Sfc} are the reflectances observed by the satellite sensor at the top of the atmosphere, atmospheric aerosols, molecules in the atmosphere, and ground surface, respectively. T_0 and T_S are the atmospheric transmittances corresponding to the path from the target point observed by the satellite to the Sun and the satellite, respectively. r_h is the reflectivity of the atmospheric hemisphere. λ and τ represent wavelength and AOT, respectively. Basically, the AOT is determined from ρ_{Aer} acquired by deducting the molecular and surface reflectance terms from ρ_{TOA}. According to Eq. (1), satellite-observed radiance is strongly controlled by surface reflectance and atmospheric transmission. Therefore, the AOT can be inversely calculated using background images (such as the clearest images with very high atmospheric transmittance) or dark surface reflectance in specific channels (such as near-infrared channels in seawater and blue and red channels in areas with dense vegetation) to remove (or minimize) the effects of surface reflection and atmospheric transmission. Current MODIS aerosol retrieval methods of "Deep Blue" and "Dark Target"

use the concept of background image and dark surface, respectively (Hsu et al., 2006; Levy et al., 2013; Remer et al., 2005). This study used Terra MODIS Level-2 aerosol products (codename: MOD04, version 6.1) with a spatial resolution of 10 km for nadir observations. To obtain the spatiotemporal coincidence of satellite and ground observations, the satellite data were averaged over a 10 km radius centered on the ground observation point. The ground observation data for the period closest to the satellite overpass time were then used.

2.2 ML

After conducting a statistical analysis of the AOT and weather conditions at each location, the ML technique was used to characterize the time-series changes and estimate future predictions. The AOT and meteorological data observed over a long period comprise a series of observations arranged in chronological order. A time-series analysis assumes that the prediction of future values depends on variables observed in the past. In Fig. 3, the architecture of AOT prediction using the ML model is shown. The challenge in prediction using ML is to build a prediction model that generalizes
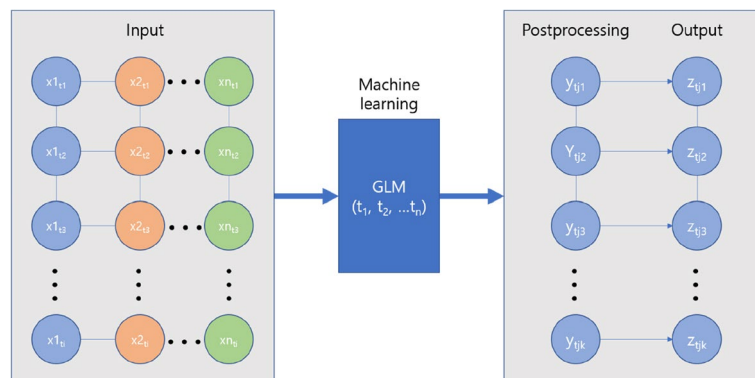
Fig. 3 Machine learning architecture for AOT prediction. The input data is the multi-dimensional time-series observations for each geographical location, and a machine learning model is performed to produce predicted values over time

well to new or nonlinear data. Overfitting is a serious problem that can occur if the model is too complex or if it fits the training data too closely. Generally, regularization is used to prevent overfitting; however, it has limitations in multi-dimensional datasets. To reduce overfitting, a generalized linear model (GLM) (Friedman et al., 2010; Liboschik et al., 2017) was used in the ML model. The GLM estimates via penalized maximum likelihood, which uses recursive coordinate descent to analyze and predict data by iterating until the objective function is optimized. The GLM for modeling discrete time-series data is shown in Eq. (2) (Bosowski & Manolakis, 2017).

$$g(\lambda_t) = \beta_0 + \sum_{k=1}^{p} \beta_0 \widetilde{g}(y_{t-ik}) + \sum_{l=1}^{q} \alpha_l g(\lambda_{t-jl}) + \eta^T X_t \tag{2}$$

where discrete time-series data Y_t with $t \in N$, the conditional mean $E(Y_t | F_t - 1)$ from discrete time-series data, for example, λ_t and $t \in N$. $g: R^+ \to R$ is the link function and $\widetilde{g}: N_0 \to R$ is a transformation function, a vector parameter $\eta = (\eta_1, ..., \eta_r)^T$. $g(\lambda_t)$ is a linear predictor and the regression can be used for the past time response variables defined as $P = \{i_1, ... i_p\}$ and i is integer $0 < i_1 ... < i_p < \infty$, with $p \in N_0$. In Eq. (2), the conditional mean for discrete time-series data represents the expected value of the response variable given a set of predictor variables and the model parameters. The GLM framework extends the linear regression model to handle non-normal distributions and non-constant variance, making it suitable for analyzing various types of data, including discrete outcomes.

The MODIS AOT data observed from 2000 to 2020 were divided into training and testing datasets at 10-year time steps. The training and test datasets comprised 80% and 20% of the input data, and the test data were used to estimate the prediction accuracy. The optimal model was then used to predict the AOT for the next 24 months. An accuracy analysis of the forecast results for each location was performed to evaluate the forecast accuracy of the models for predicting the AOT. This process involves the fundamental assumption that a model with a small error is suitable for predicting future values. To determine whether the model used was suitable for explaining AOT, we analyzed the performance of the model using the Mean Absolute Error (MAE), Mean Square Error (MSE), Mean Absolute Percentage Error (MAPE), Bias, and Root MSE (RMSE), as shown in Eqs. (3)–(7). Therefore, the optimal model had the lowest MAE, MSE, MAPE, Bias, and RMSE, indicating that it had the best predictive capability.

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \widehat{y}_i| \tag{3}$$

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 \tag{4}$$

$$MAPE = \frac{100}{n} \sum_{i=1}^{n} \left| \frac{y_i - \widehat{y}_i}{y_i} \right| \tag{5}$$

$$Bias = \frac{1}{n} \sum_{i=1}^{n} (\widehat{y}_i - y_i) \tag{6}$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2} \tag{7}$$

where y is the observed value; $\widehat{y}$ is the predicted value; and n is the total number of data points.

2.3 Spatial analysis

In the previous section, methods for analyzing the current aerosol situation at a selected point and predicting future aerosol changes were described. However, these analytical methods cannot generate results in areas where data are not available due to the absence of observation equipment or loss of data. Therefore, to analyze the AOT values of points where observations could not be obtained, the results were generated from spatial modeling using the kriging technique. To evaluate the AOT values at given points where ground observations do not exist, spatially interpolated AOT values were estimated based on the AOT values predicted using ML for 2021–2022 in the previous section.

Kriging is a technique that uses weighted linear combinations to make predictions about data in space using Eq. (8). The weights in kriging are determined as a function of distance, such that the error between the predicted and actual values is minimized. In kriging, variograms are used to represent the spatial autocorrelation.

$$Z'(s_0) = \sum_{i=1}^{n} (\lambda_i \cdot Z(s_i)) \tag{8}$$

where Z' and Z represent the predicted unobserved and measured values, respectively, where N is the number of points to be predicted. λ_i is a weighting factor. Kriging uses a variogram to represent the spatial autocorrelation. A variogram is a measure of the squared difference between data points separated by a certain distance, which indicates the similarity between data points.

$$2\gamma \left(h' \right) = \frac{1}{n} \sum_{i=1}^{n} \{z(U_i) - z(U_i + h)\}^2 \tag{9}$$

where $z(U_i)$ and $z(U_i + h)$ are values at a given point in time and at distance h. n is the number of pairs at distance h.

3 Result

3.1 Status of atmospheric aerosols on the Korean Peninsula

To analyze the regional aerosol distribution, MODIS AOT data observed over 22 years (2000–2021) were divided into three bins using a probability density function. Fig. 4 shows a comparison of the probability density functions for the AOT observations at each point. The level of AOT in each location was categorized into three clusters, with low-level AOT locations, including Gangneung, Jeju, and Busan, showing similar distributions. Seoul had the highest AOT distribution. The remaining locations had a moderate AOT distribution. Accordingly, the AOTs at each point can be categorized according to their probability distribution characteristics.

The bins were classified into three categories as low AOT (mean-$\sigma<$AOT$\leq$mean), moderate AOT (mean$<$AOT$\leq$mean$+\sigma$), and high AOT (mean$+\sigma<$AOT). Table 2 summarizes the statistical breakdown of the MODIS AOT data for the study points. The average AOT value for all points of interest ranged from 0.471 ± 0.427, with maximum and minimum values of 0.633 ± 0.591 for Seoul and 0.362 ± 0.312 for Gangneung. The mean and standard deviation values at each point varied widely, which may have been due to the combined effects of emission sources, weather conditions, and terrain. A study by Pyo et al. (2021) showed that the contributions of aerosols from outside Seoul and Gangneung were high, mainly due to westerly winds, at 67%

and 83%, respectively. Thus, it is necessary to analyze the differences in AOT values in the remaining regions due to changes in weather conditions, such as wind direction and wind speed.

Atmospheric particles can change their physical properties and chemistry depending on the weather conditions. It is well known that wind conditions play an important role in the transport and dispersion of air pollutants. Wind patterns significantly influence the distribution and concentration of atmospheric aerosols. Fig. 5 shows wind rose plots as a function of the three AOT categories. The RH directly affects the hygroscopic growth of particles, which can change their size or refractive index and affect their optical properties (Gassó et al., 2000; Hegg et al., 2002; Markowicz et al., 2003). These mechanisms affect the optical properties of aerosol particles and, as a result, cause changes in the measurement of AOT (Bian et al., 2009). However, these relations can vary depending on a variety of environmental factors and aerosol types; therefore, more precise experiments and analyses are required to generalize the observations to specific regions or conditions. To estimate the association between AOT and RH, the average RH ranges at three different AOT levels in each study area were analyzed. Fig. 6 shows boxplots of the RH distribution versus the AOT distribution at each point.

The results of analyzing the meteorological conditions using the three AOT bins at each location were as follows. First, westerly winds are the predominant winds in Gangneung, but the AOT tends to increase when northern or eastern winds prevail, as shown in Fig. 5a. Interestingly, the easterly winds (wind direction ranging from
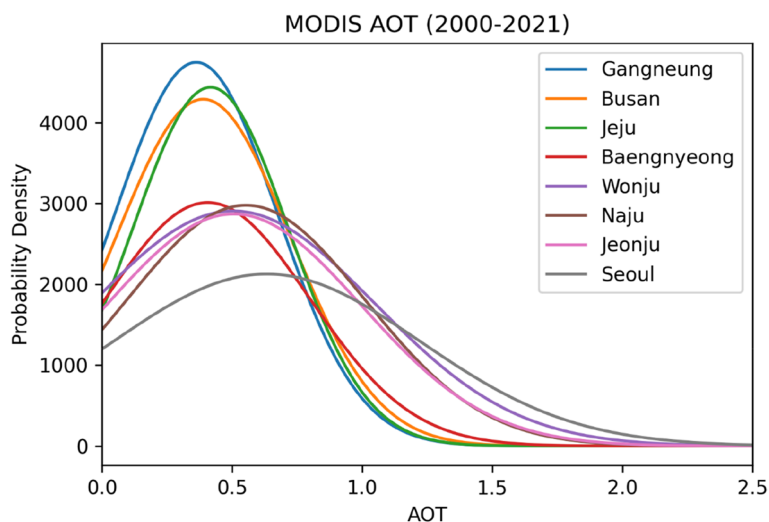

Fig. 4 Probability density distribution of AOT over Gangneung, Wonju, Seoul, Baengnyeong, Jeonju, Naju, Busan, and Jeju during 2000–2021

Lee *et al. Asian Journal of Atmospheric Environment*   (2024) 18:9

Page 7 of 15

Table 2 List of the number of data and average and standard deviation of AOT by region

Site	Number of data (*n*)	Mean (*m*)	Standard deviation (*σ*)	m − σ	m + σ
Gangneung	3712	0.362	0.312	0.050	0.674
Seoul	3150	0.633	0.591	0.042	1.224
Busan	3581	0.389	0.333	0.056	0.721
Wonju	3964	0.504	0.544	0.010	1.044
Naju	3419	0.553	0.458	0.100	1.010
Jeonju	3525	0.506	0.490	0.016	0.996
Jeju	3326	0.417	0.299	0.108	0.716
Baengnyeong	2959	0.406	0.392	0.014	0.798
Average	[a]27,636	0.471	0.427	0.050	0.898

[a] Total number

20 to 90°) accounted for 4.6% for low AOT, 13.5% for moderate AOT, and 22.5% for high AOT cases, respectively. Gangneung faces mountains to the west and the coast to the east. Certain weather conditions can cause atmospheric stagnation, and complex terrain, including mountains and oceans, can cause inversions. In particular, easterly winds blowing landward from the ocean can increase moisture and salt from the sea surface and increase the production of sea salt particles. Thus, easterly winds contribute to atmospheric stagnation and transport sea salt particles, thereby affecting aerosol loads in the local atmosphere. The complex interactions between meteorological and geographical factors determine the magnitude of these effects, emphasizing the need for a comprehensive understanding of regional atmospheric dynamics. Fig. 6a shows that mean RH values in Gangneung are 43.68 ± 15.72%, 54.77 ± 16.39%, and 58.38 ± 16.12% related to the low, moderate, and high AOT categories, respectively. These results indicate that the AOT in Gangneung increased as the RH increased, and the higher the contribution of the east wind, the higher the AOT value.

In Fig. 5b, the frequency distribution of observed wind direction in Wonju showed largest values of 45.3% and 54.7% for east and west winds, respectively. In addition, it was found that the AOT increased as the southwesterly winds increased. This process can lead to an influx of air pollutants from neighboring metropolitan areas. Interestingly, the frequency of southwest winds increased by 9.9% for low AOT, 13.9% for moderate AOT, and 17.6% for high AOT cases. In contrast, winds from the northeast and east lowered the AOT. Wind direction ranged from 50 to 90° accounted for 26.9% of low AOT, 18.9% of moderate AOT, and 15.2% of high AOT cases, respectively. In Fig. 6b, the mean RH values are 57.80 ± 12.04% in the low AOT, 61.80 ± 11.35% in the moderate AOT,

and 65.64 ± 10.02% in the high AOT conditions. Summarizing the observations in Wonju, the AOT increased with increasing RH, and the wind direction increased with increasing contribution from the southwest wind.

The prevailing winds in Seoul are westerly, with approximately 65.5% of the total observations being highly influenced by atmospheric aerosols from the west (Fig. 5c). Higher AOTs in Seoul were associated with southwesterly winds, with the frequency of southwesterly winds increasing to 2.9% for low AOT, 7.0% for moderate AOT, and 13.0% for high AOT. In addition, RH by AOT bin in Fig. 6c was 51.96 ± 11.11% at low AOT, 58.77 ± 10.94% at moderate AOT, and 63.18 ± 10.74% at high AOT. Thus, the AOT tended to increase as the RH and/or southwest wind increased in Seoul.

Baengnyeong, the westernmost island of South Korea, experiences a variety of weather conditions. The AOT increased as the RH and the contribution of the southerly and easterly winds increased, as shown in Fig. 5d. Southerly winds increased from 4.8% at low AOT to 13.7% at high AOT, which means that an approximately three times increase in the frequency of southerly winds is associated with an increase in AOT. Eastern winds ranging from 90 to 110° increased from 7.6% for low AOT to 15.1% for high AOT. Fig. 6d shows the AOT and RH values for Baengnyeong. RH was 62.19 ± 13.30% at low AOT, 69.52 ± 15.07% at moderate AOT, and 73.78 ± 13.94% at high AOT.

Jeonju and Naju are small cities located in the southwestern part of the Korean Peninsula, where high RH and westerly winds contribute to a high AOT (Fig. 5e, f). In Jeonju, the westerly and southeasterly winds were associated with increased AOT. Wind frequencies for the westerly winds were 10.1% (low AOT), 18.4% (moderate AOT), and 23.2% (high AOT), respectively. However, in Naju, the increase and decrease in AOT were reversed under the southwest and northeast wind conditions. In detail, the wind frequencies were 21.6% in the northeast and 74.2% in the southwest. In Fig. 6e, the mean RH values in Naju were 59.13 ± 11.36% at low AOT, 63.32 ± 11.59% at moderate AOT, and 66.80 ± 10.52% at high AOT. Similarly, the mean RH values in Jeonju were 61.07 ± 11.36% at low AOT, 64.66 ± 11.59% at moderate AOT, and 68.62 ± 10.52% at high AOT as shown in Fig. 6f.

As a port city located in the southeastern part of the Korean Peninsula, the AOT in Busan was found to increase with an increase in southerly winds, whereas an increase in northwesterly winds was associated with a decrease in AOT (Fig. 5g). Under high AOT conditions, the frequency of northwesterly winds was 6.5%, but the frequency of southerly winds increased to 44.8%. In addition, the maximum frequency values for wind were found in the north (16.5%) and west winds (13.6%) under low
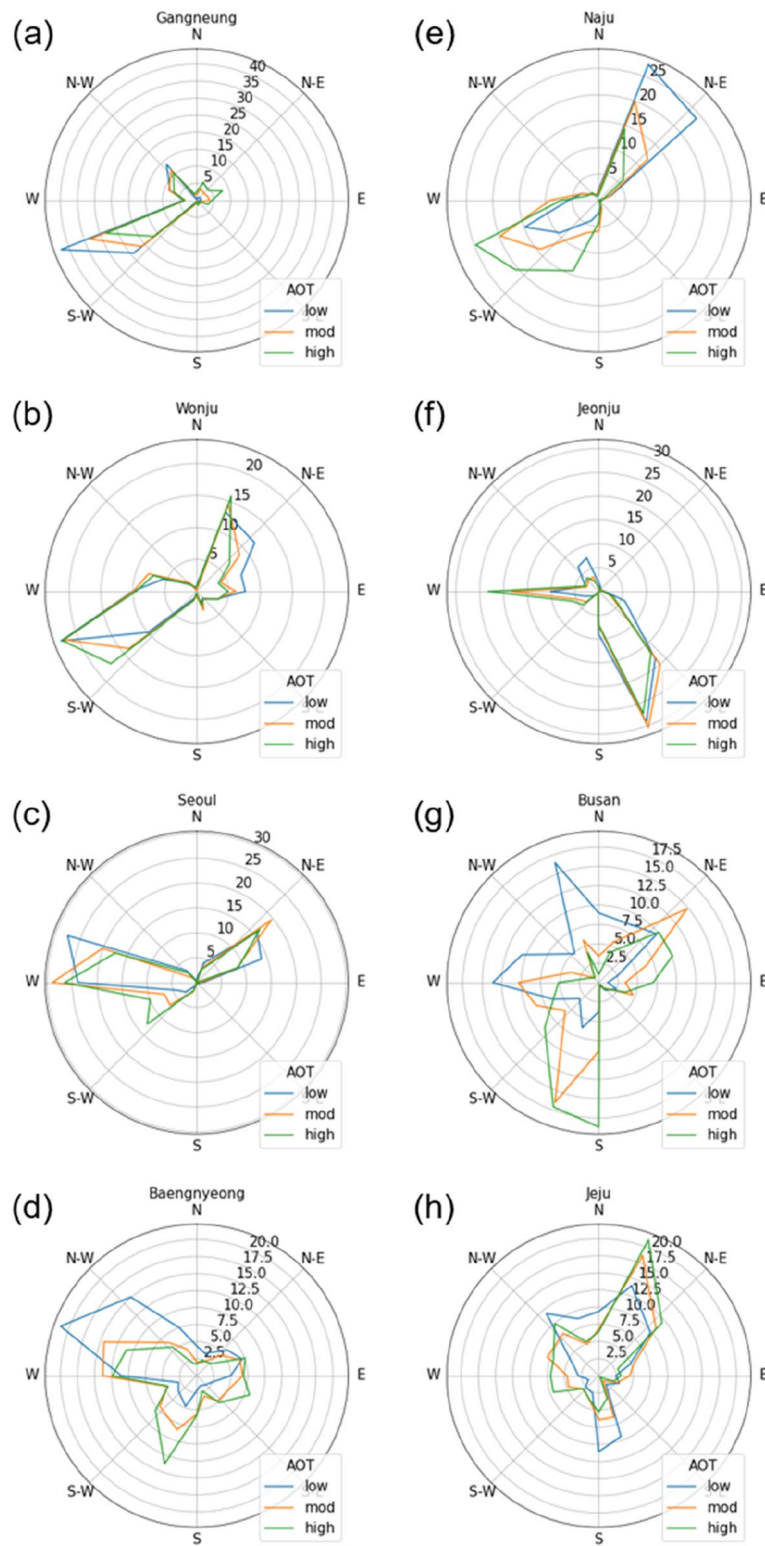
Fig. 5 Distribution of wind direction by three AOT categories over Gangneung, Wonju, Seoul, Baengnyeong, Jeonju, Naju, Busan, and Jeju during 2000–2021
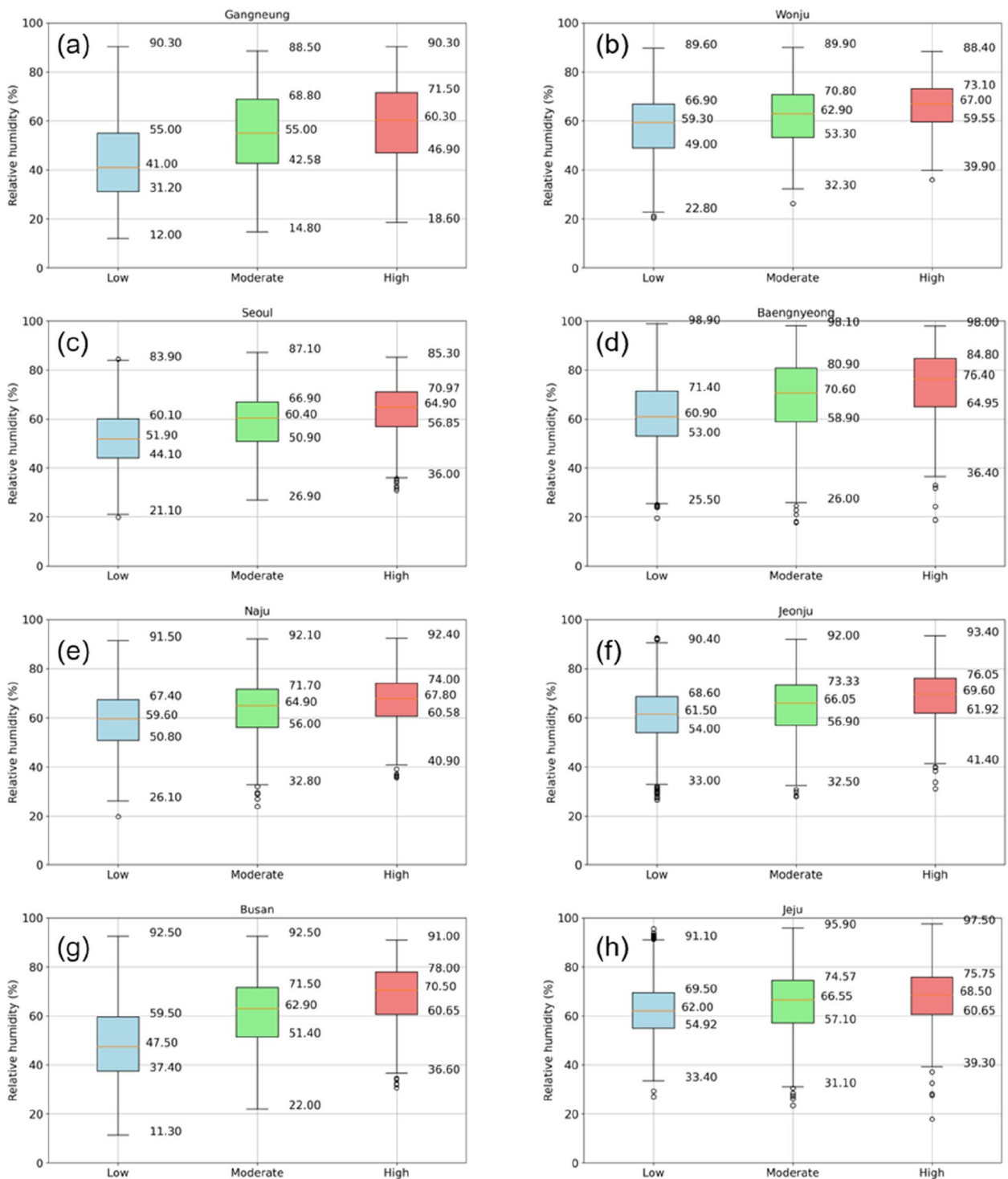
Fig. 6 Boxplots of relative humidity (%) by three AOT categories (low: mean-σ < AOT ≤ mean, medium: mean < AOT ≤ mean + σ, high: mean + σ < AOT)

AOT conditions. Fig. 6g shows that mean RH values were 48.72 ± 14.749% at low AOT in Busan, 61.22 ± 13.56% at moderate AOT, and 68.41 ± 12.35% at high AOT.

Finally, Jeju, an island in the southern part of the Korean Peninsula, tended to have higher AOTs owing to increased prevailing northeastern winds, as shown in

Fig. 5h. Under high AOT conditions, the frequency of northwesterly wind was 33% at maximum. In addition, the lower the AOT value, the greater the frequency of southerly winds. Fig. 6h shows that mean RH values were 62.24 ± 11.03% at low AOT, 65.41 ± 12.46% at moderate AOT, and 67.73 ± 11.64% at high AOT.

3.2 Aerosol prediction with ML

In the previous chapter, the current status of the atmosphere, using long-term observed MODIS AOT data and meteorological data, explained a variety of AOT distributions, and meteorological conditions were found at each location. In aerosol-related research, it is important to obtain information about aerosols in places other than aerosol observation points and to make predictions using long-term observations of aerosols and trend information. The ML technique was used to achieve this goal. As an example, the model used input data with AOT data observed from 2011 to 2020 at eight locations, as shown in Fig. 1. Data from 2011 to 2018 and 2019–2020 were used as the training and testing data, respectively. Accuracy was estimated using the test dataset. Figs. 7 and 8 show the ML results of the AOT time series at the eight study sites.

The metrics of bias and RMSE were used to evaluate the prediction performance of the ML model. Table 3 summarizes the performance results of the ML model for eight sites. The validation indices were MAE, MSE, MAPE, bias, and RMSE, with lower values indicating relatively low error levels. The average values of the validation indices for all points are MAE = 0.152 ± 0.025, MSE = 0.048 ± 0.016, MAPE = 77.325 ± 10.889%, Bias = −0.002 ± 0.011, and RMSE = 0.216 ± 0.038, respectively. Relatively low validation indices were found in Gangneung, Busan, Jeju, and Baengnyeong. These locations are mainly close to the ocean, and thus have low AOT levels. In contrast, the validation indices were relatively high in larger cities, such as Seoul, Wonju, and Jeonju. The size range of the inputs in the ML model can affect the forecast results (Ahsan et al., 2021; Shrestha & Mahmood, 2019). The data scale can directly affect the performance of the forecasting model. If the magnitude range of the input data is small, the model attempts to predict small variations in the data, and if the magnitude range of the input data is large, the model attempts to predict large variations in the data. In other words, a model trained with large-magnitude range data will not have difficulty predicting highly variable data, but a model trained with small-magnitude range data will have difficulty. Our results indicate that as the observational ranges of the MODIS

AOT values over the model period increased, the forecast accuracy decreased.

In addition, the standard deviations of the model inputs and predicted values were compared. This was verified using the *F*-test (or the ratio of variances test), where the ratio approached one if the variances of the two datasets were not significantly different. In this result, the standard deviation of the inputs is 0.431 ± 0.112, and the standard deviation of the predictions is 0.175 ± 0.044. The *F* statistic is 0.166 ± 0.017, which means that the variance of the predicted results is within approximately 34% of the input values, indicating a narrow error band. Consequently, the RMSE values were smaller than the standard deviations of the actual observations for all regions. Therefore, the accuracy of the model predictions was considered sufficient. After ensuring the accuracy of the model performance, AOT predictions from 2021 to 2022 were performed. Forecast results show mean RMSE and bias are 0.033 ± 0.012 and 0.036 ± 0.022 which is less than the standard deviation of the observations, which is accurate enough to make AOT predictions from 2021 to 2022.

3.3 Spatial analysis of aerosol distribution

After accurately estimating the AOT prediction at the eight selected points, the predicted AOT values were acquired for each location. However, these predicted values are limited to point locations and cannot be estimated over large areas. Thus, AOT values within the study area were estimated using the kriging spatial interpolation method. This technique was achieved by applying a spatial Gaussian distribution model with a resolution of 0.01° at both the latitude and longitude.

Figure 9 shows spatially interpolated AOT map by using the ML model and MODIS observations at eight selected locations. Full-covered MODIS AOT maps are also shown as the gridded at 0.1° within the study area of interest. Compared with that of 2021, the predicted AOT data for 2022 showed decreases of 10.5%, 2.9%, 8.3%, 5.3%, 7.1%, 8.5%, 10%, 6.8%, and 7.4% at sites of Gangneung, Wonju, Seoul, Baengnyeong Island, Naju, Jeonju, Busan, Jeju, and the national average, respectively. Generally, the higher the spatial resolution of the observations, the more accurate the predictions because the kriging technique is a weighted linear extrapolation of the interval values. Therefore, obtaining satellite-measured AOT values from a broader range of locations outside the study area may provide more accurate AOT predictions for the Korean Peninsula.

Figure 10 shows scatterplot illustrating the relationship between spatially interpolated AOT and full MODIS AOT. The spatially interpolated AOT data represent two distinct datasets are from the dataset used in Fig. 8. Two
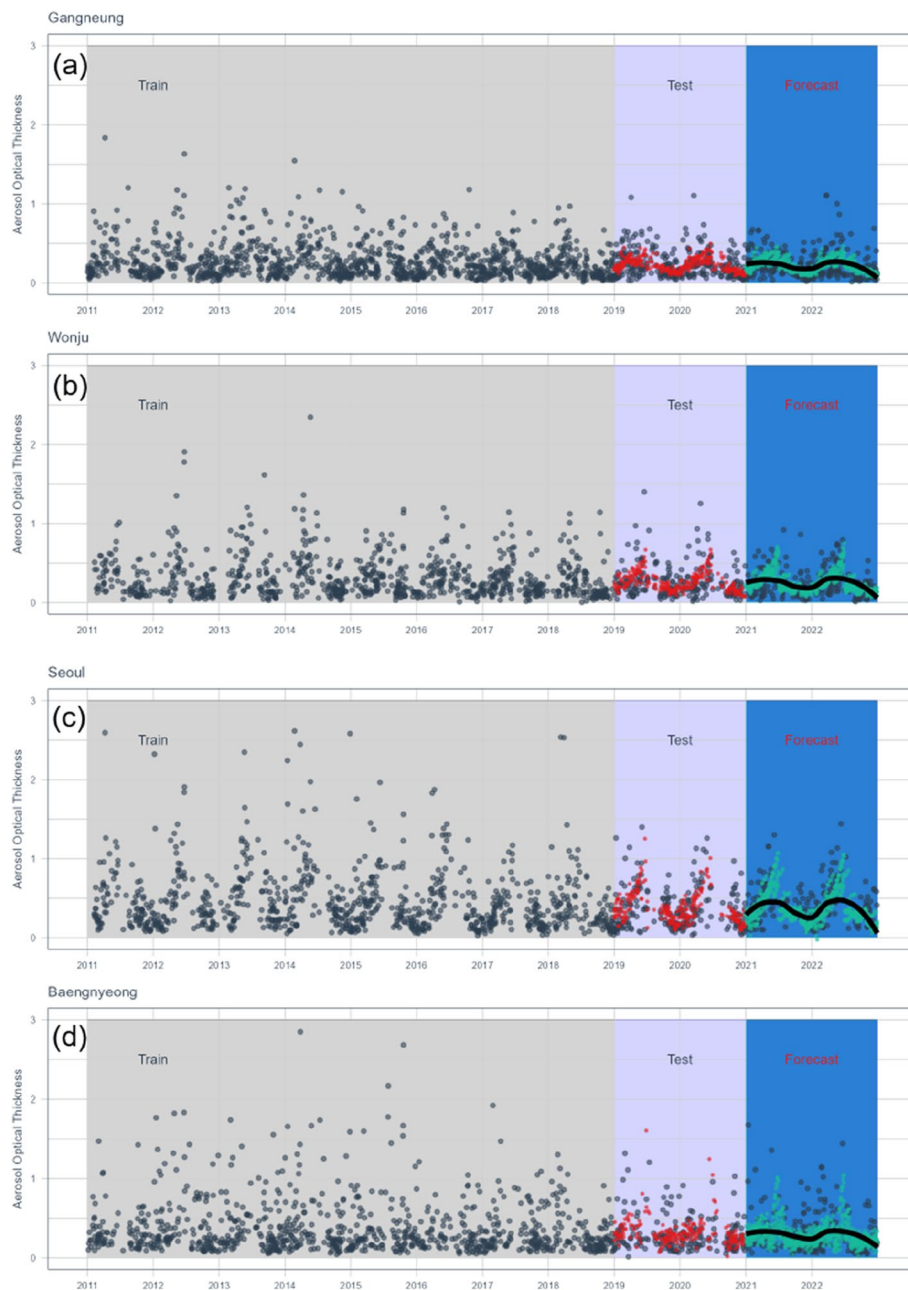
Fig. 7 Machine learning analysis results of AOT for Gangneung, Wonju, Seoul, and Baengnyeong. Data periods used for training (2011–2018), testing (2019–2020), and prediction (2021–2022) are shown with different colors of backgrounds

scatterplots are overlaid with the linear regression lines affirming the strength of the linear association between the variables. Derived linear regression equations are $y = 1.33x - 0.07$ $(r = 0.87)$ and $y = 1.18x - 0.04$ $(r = 0.88)$. The correlation coefficients (r) indicate a strong positive linear relationship between the two variables. This suggests that as the independent variable increases, the

dependent variable tends to increase as well, and vice versa. A comparison of the modeled and MODIS AOTs shows RMSE = 0.08 and bias = 0.02, confirming that these results are within the prediction error presented earlier. Again, the scatterplot with the regression line demonstrates the close alignment of the data points with the predicted values.
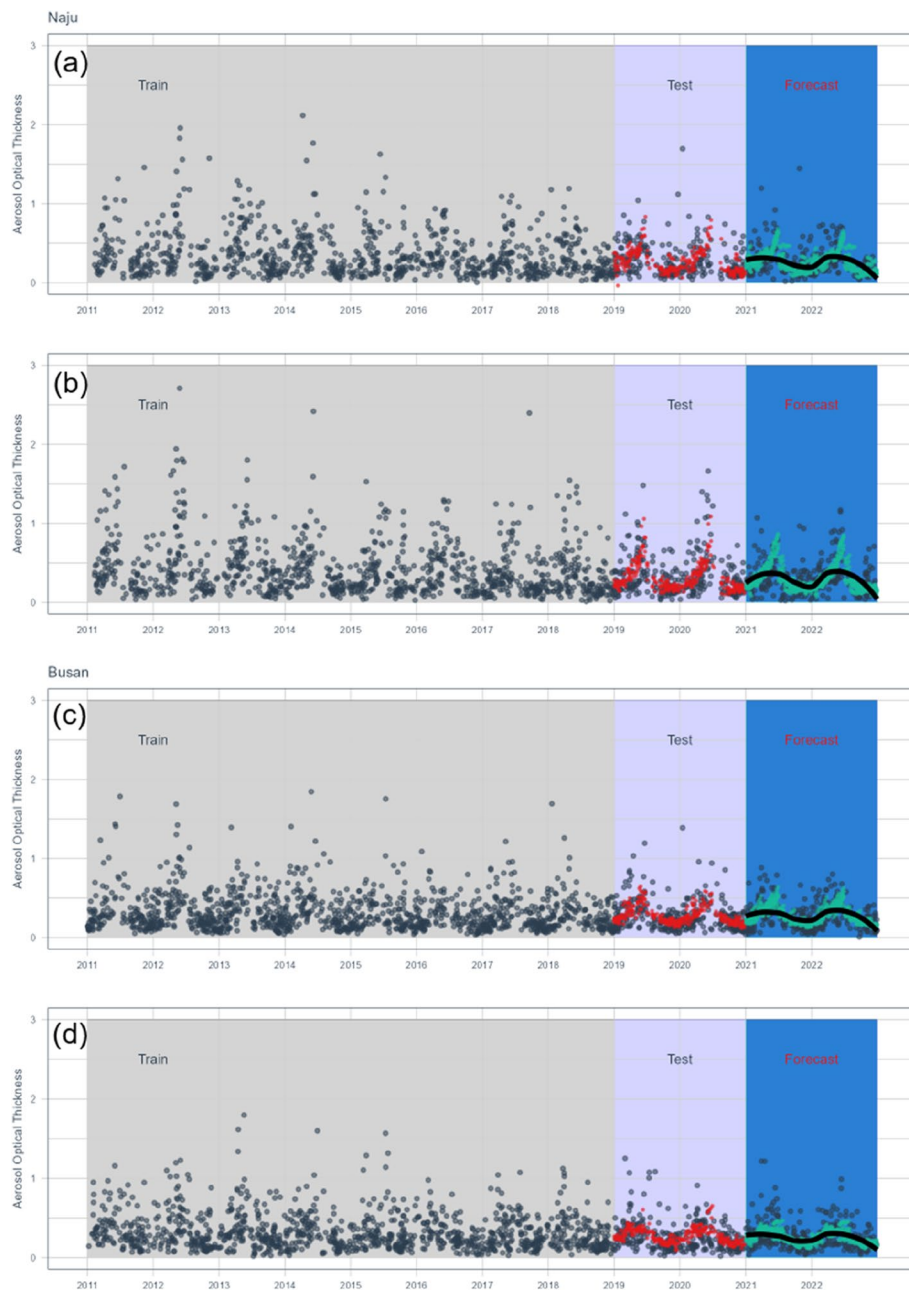
Fig. 8 Machine learning analysis results of AOT for Naju, Jeonju, Busan, and Jeju. Data periods used for training (2011–2018), testing (2019–2020), and prediction (2021–2022) are shown with different colors of backgrounds

4 Summary and conclusion

Column aerosol observation data contain information on the total amount and optical properties of aerosol particles in Earth's atmosphere, which is important for research on aerosol composition, emission sources, and transport. The current status of atmospheric aerosols in Korea was estimated using long-term satellite-observed AOT data from Terra MODIS and meteorological data from AWS. The main conclusions are summarized as follows:

First, AOT levels were analyzed in eight selected cities (Gangneung, Wonju, Seoul, Baengnyeong Island, Jeonju, Naju, Busan, and Jeju) in Korea using MODIS AOT data. The mean and standard deviation results were 0.362 ± 0.312 (Gangneung), 0.544 ± 0.504 (Wonju), 0.633 ± 0.591 (Seoul), 0.406 ± 0.392 (Baengnyeong),

Table 3 Accuracy results for the site-specific time-series forecast models, including MAE, MSE, MAPE, Bias, and RMSE

Site	MAE	MSE	MAPE (%)	Bias	RMSE
Gangneung	0.122	0.028	73.592	− 0.012	0.167
Seoul	0.189	0.068	86.201	0.013	0.261
Busan	0.126	0.033	63.078	0.001	0.180
Wonju	0.144	0.040	85.684	− 0.007	0.201
Naju	0.151	0.050	68.513	− 0.011	0.225
Jeonju	0.174	0.062	71.101	− 0.018	0.249
Jeju	0.133	0.034	74.644	0.006	0.186
Baengnyeong	0.178	0.067	95.788	0.010	0.259
Average	0.152	0.048	77.325	− 0.002	0.216
Standard deviation	0.025	0.016	10.889	0.011	0.038

0.553 ± 0.458 (Naju), 0.506 ± 0.49 (Jeonju), 0.389 ± 0.333 (Busan), and 0.419 ± 0.299 (Jeju), with the highest values in Seoul and the lowest values in Jeju.

Second, the AOT was classified into three bins, i.e., low AOT ($avg - \sigma < AOT \leq avg$), moderate AOT ($avg < AOT \leq avg + \sigma$), and high AOT ($avg + \sigma < AOT$) at the eight regions. Wind direction, wind speed, and humidity were evaluated for each of the three AOT bins.

At each location, the relation between AOT and meteorological parameters exhibited distinct characteristics. Gangneung, located in the northeast of the Korean Peninsula, has experienced an increase in AOT owing to the double frequency of easterly winds. Southwesterly winds in Wonju and Seoul were strongly associated with increased AOT, with the intensity increasing from 6.8 to 12.9% from moderate to high AOT. On Baengnyeong Island, the east and south winds typically increase by 2.6 and 1.5 times as the AOT increased, demonstrating distinctive features from locations at similar latitudes. Jeonju and Naju experienced increased AOT due to the frequency of westerly and southerly winds further south. The AOT increase in Jeju, the southernmost island region, was caused by an increase in westerly winds (approximately 3.3 times) and a reduction in southerly winds (43%).

In addition, regional-scale AOT values for the entire Korean Peninsula were acquired by combining ML with a spatial interpolation method using satellite-observed AOT and meteorological observation data. A time-series forecasting model using a GLM was built for future AOT data. The RMSE was lower than the standard deviation of the actual observations for all points in the test
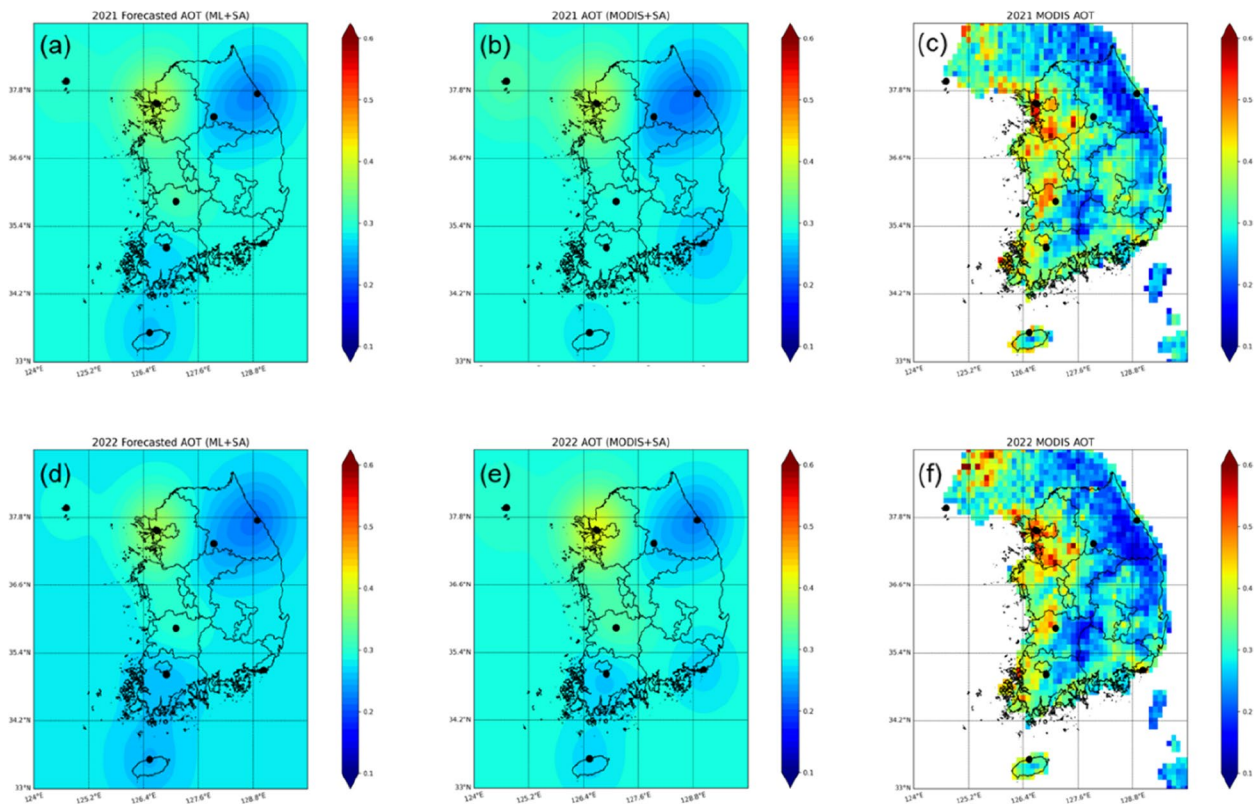
Fig. 9 Spatially interpolated AOT map in **a** and **b** 2021 and **c** and **d** 2022 by using **a** and **d** model estimated and **b** and **e** MODIS observations (black dots: selected eight locations of measurements). **c** and **f** MODIS AOT maps gridded at 0.1° within the study area of interest
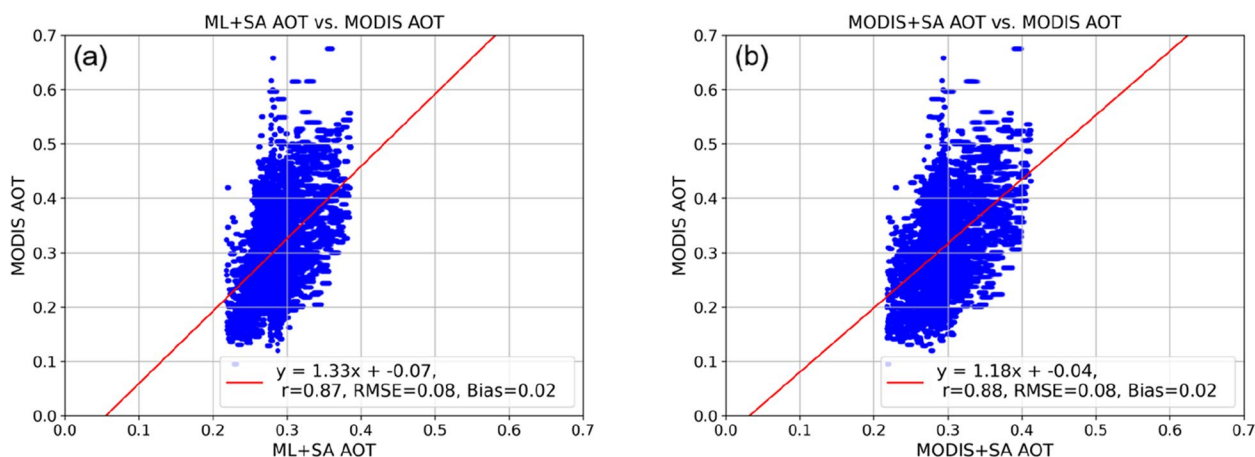
Fig. 10 Scatterplots depicting the relationship between MODIS observations and spatially interpolated AOT with a linear regression line(red line). Spatially interpolated AOT data **a** forecasted by machine learning model and **b** MODIS observations at the selected eight locations

set, confirming that the trained prediction model had an acceptable level of prediction accuracy. Forecasting was conducted for a 24-month period between 2022 and 2023 to determine the future outlook of AOT values, and the results showed a gradual increase. The spatially interpolated AOT map showed that an accurate gridded AOT could be acquired with a higher spatial resolution over the study area. These outcomes could benefit regional air pollution policies.

Authors' contributions
Kwon-ho Lee: research planning, data collection, analysis and visualization, writing, editing manuscript. Seong-Hun Pyo: provided surface in situ data, reviewing, and editing. Man Sing Wong: review and editing. All the authors have read and agreed to the published version of this manuscript.

Availability of data and materials
The datasets generated and analyzed in the current study are available from the corresponding author upon reasonable request.

Declarations

Competing interests
The authors declare that they have no competing interests. In accordance with Springer policy and ethical obligation as a researcher, K.H. Lee discloses that he is an associate editor on this journal.

References
Ahsan, M. M., Mahmud, M. A. P., Saha, P. K., Gupta, K. D., & Siddique, Z. (2021). Effect of data scaling methods on machine learning algorithms and model performance. *Technologies,9*(3), 52. https://doi.org/10.3390/technologies9030052

Bian, H., Chin, M., Rodriguez, J. M., Yu, H., Penner, J. E., & Strahan, S. (2009). Sensitivity of aerosol optical thickness and aerosol direct radiative effect to relative humidity. *Atmospheric Chemistry and Physics,9*(7), 2375–2386. https://doi.org/10.5194/acp-9-2375-2009

Chakraborty, S., Guan, B., Waliser, D. E., da Silva, A. M., Uluatam, S., & Hess, P. (2021). Extending the atmospheric river concept to aerosols: Climate and air quality impacts. *Geophysical Research Letters,48*(GL091827), e2020. https://doi.org/10.1029/2020GL091827

Gassó, S., Hegg, D. A., Covert, D. S., Collins, D., Noone, K. J., Öström, E., Schmid, B., Russell, P. B., Livingston, J. M., Durkee, P. A., & Jonsson, H. (2000). Influence of humidity on the aerosol scattering coefficient and its effect on the upwelling radiance during ACE-2. *Tellus b: Chemical and Physical Meteorology,52*(2), 546–567. https://doi.org/10.3402/tellusb.v52i2.16657

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

Hsu, N. C., Tsay, S.-C., King, M. D., & Herman, J. R. (2006). Deep blue retrievals of Asian aerosol properties during ACE-Asia. *IEEE Transactions on Geoscience and Remote Sensing,44*(11), 3180–3195. https://doi.org/10.1109/TGRS.2006.879540

Jerrett, M., Burnett, R. T., Ma, R., Pope, C. A., 3rd., Krewski, D., Newbold, K. B., Thurston, G., Shi, Y., Finkelstein, N., Calle, E. E., & Thun, M. J. (2005). Spatial analysis of air pollution and mortality in Los Angeles. *Epidemiology,16*(6), 727–736. https://doi.org/10.1097/01.ede.0000181630.15826.7d

Lee, K. H. (2018). Estimation and validation of collection 6 moderate resolution imaging spectroradiometer aerosol products for East Asia. *Asian Journal of Atmospheric Environment,12*(3), 193–203. https://doi.org/10.5572/ajae.2018.12.3.193

Lee, K. H., Wong, M. S., & Li, J. (2022). Review of atmospheric environmental change from earth observing satellites. *Asian Journal of Atmospheric Environment,16*(1), 1–13. https://doi.org/10.5572/ajae.2021.147

Lelieveld, J., Evans, J. S., Fnais, M., Giannadaki, D., & Pozzer, A. (2015). The contribution of outdoor air pollution sources to premature mortality on a global scale. *Nature,525*, 367–371. https://doi.org/10.1038/nature15371

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

Lee *et al. Asian Journal of Atmospheric Environment*   (2024) 18:9

Page 15 of 15

Li, Z., Zhao, X., Kahn, R., Mishchenko, M., Remer, L., Lee, K.-H., Wang, M., Laszlo, I., Nakajima, T., & Maring, H. (2009). Uncertainties in satellite remote sensing of aerosols and impact on monitoring its long-term trend: A review and perspective. *Annales Geophysicae,27*(7), 2755–2770. https://doi.org/10.5194/angeo-27-2755-2009

Li, J., Wong, M. S., Lee, K. H., Nichol, J., & Chan, P. W. (2021). Review of dust storm detection algorithms for multispectral satellite sensors. *Atmospheric Research,250*, 105398. https://doi.org/10.1016/j.atmosres.2020.105398

Li, J., Carlson, B. E., Yung, Y. L., Lv, D., Hansen, J., Penner, J. E., Liao, H., Ramaswamy, V., Kahn, R. A., Zhang, P., Dubovik, O., Ding, A., Lacis, A. A., Zhang, L., & Dong, Y. (2022). Scattering and absorbing aerosols in the climate system. *Nature Reviews Earth and Environment,3*, 363–379. https://doi.org/10.1038/s43017-022-00296-7

Oh, H.-R., Ho, C.-H., Kim, J.-W., Chen, D., Lee, S.-M., Choi, Y.-S., Chang, L.-S., & Song, C.-K. (2015). Long-range transport of air pollutants originating in China: A possible major cause of multi–day high–PM10 episodes during cold season in Seoul, Korea. *Atmospheric Environment,109*, 23–30. https://doi.org/10.1016/j.atmosenv.2015.03.005

Remer, L. A., Kaufman, Y. J., Tanré, D., Mattoo, S., Chu, D. A., Martins, J. V., Li, R.-R., Ichoku, C., Levy, R. C., Kleidman, R. G., Eck, T. F., Vermote, E., & Holben, B. N. (2005). The MODIS aerosol algorithm, products, and validation. *Journal of the Atmospheric Sciences,62*, 947–973. https://doi.org/10.1175/JAS3385.1

Shrestha, A., & Mahmood, A. (2019). Review of deep learning algorithms and architectures. *IEEE Access,7*, 53040–53065. https://doi.org/10.1109/ACCESS.2019.2912200

Sun, E., Xu, X., Che, H., Tang, Z., Gui, K., An, L., Lu, C., & Shi, G. (2019). Variation in MERRA-2 aerosol optical depth and absorption aerosol optical depth over China from 1980 to 2017. *Journal of Atmospheric and Solar-Terrestrial Physics,186*, 8–19. https://doi.org/10.1016/j.jastp.2019.01.019

Torres, O., Bhartia, P. K., Herman, J. R., Sinyuk, A., & Holben, B. (2002). A long-term record of aerosol optical thickness from TOMS observations and comparison to AERONET measurements. *Journal of the Atmospheric Sciences,59*, 398–413. https://doi.org/10.1175/1520-0469_2002_059_0398_altroa_2.0.co_2

Yu, X., Nichol, J. E., Lee, K. H., Li, J., & Wong, M. S. (2022). Analysis of long-term aerosol optical properties combining AERONET sunphotometer and satellite-based observations in Hong Kong. *Remote Sensing,14*, 5220. https://doi.org/10.3390/rs14205220

Adegun, A. A., Viriri, S., & Tapamo, J. R. (2023). Review of deep learning methods for remote sensing satellite images classification: Experimental survey and comparative analysis. *Journal of Big Data*, *10*(1). https://doi.org/10.1186/s40537-023-00772-x

Bosowski, N., & Manolakis, D. (2017). *Generalized linear models for count time series IEEE International Conference on Acoustics, Speech and Signal processing (ICASSP)*. https://doi.org/10.1109/ICASSP.2017.7952962

Friedman, J. H., Hastie, T., Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*, *33*(1), 1–22. https://doi.org/10.18637/jss.v033.i01

Hegg, D. A., Covert, D. S., Crahan, K., & Jonsson, H. (2002). The dependence of aerosol light–scattering on RH over the Pacific Ocean. *Geophysical Research Letters*, *29*(8), 60–1–60–4. https://doi.org/10.1029/2001GL014495

Lee, K. H., Li, Z., Kim, Y. J., & Kokhanovsky, A. (2009). Atmospheric aerosol monitoring from satellite observations: A history of three decades. In Y. J. Kim, U. Platt, M. B. Gu, & H. Iwahashi (Eds.). *Atmospheric and biological environmental monitoring*, (pp. 13–38). https://doi.org/10.1007/978-1-4020-9674-7_2. Berlin, Heidelberg: Springer

Levy, R., Hsu, C., Sayer, A., Mattoo, S., & Lee, J. (2017). *MODIS Atmosphere L2 aerosol product. NASA MODIS adaptive processing system*. Goddard Space Flight Center. https://doi.org/10.5067/MODIS/MYD04_L2.061

Liboschik, T., Fokianos, K., & Fried, R. (2017). Tscount: An R package for analysis of count time series following generalized linear models. *Journal of Statistical Software*, *82*(5). https://doi.org/10.18637/jss.v082.i05

Markowicz, K. M., Flatau, P. J., Quinn, P. K., Carrico, C. M., Flatau, M. K., Vogelmann, A. M., Bates, D., Liu, M., & Rood, M. J. (2003). Influence of relative humidity on aerosol radiative forcing: An ACE-Asia experiment perspective. *Journal of Geophysical Research: Atmospheres*, *108*(D23). https://doi.org/10.1029/2002JD003066

Intergovernmental Panel on Climate Change (IPCC). (2023) Synthesis report. contribution of working groups I. Summary for policymakers. In *Climate Change* Core Writing Team, H. Lee, & J. Romero (Eds.). Geneva, Switzerland: IPCC, *II and III* to the Sixth assessment rReport of the

intergovernmental panel on climate change, (1–34). https://doi.org/10.59327/IPCC/AR6-9789291691647.001

Intergovernmental Panel on Climate Change (IPCC). (2021) The physical science basis. Contribution of working group I to the Sixth assessment report of the intergovernmental panel on climate change (pp. 1–2391) (Masson-Delmotte, V., P. Zhai, A. Pirani, S.L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M.I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J.B.R. Matthews, T.K. Maycock, T. Waterfield, O. Yelekçi, R. Yu and B. Zhou Eds) *Climate Change*. https://doi.org/10.1017/9781009157896. Cambridge, United Kingdom and New York, NY: Cambridge University Press

Menon, S., Unger, N., Koch, D., Francis, J., Garrett, T., Sednev, I., Shindell, D., & Streets, D. (2008) Aerosol climate effects and air quality impacts from 1980 to 2030, Environmental Research Letter, 3, doi:https://doi.org/10.1088/1748-9326/3/2/024004.

Murphy, K. P. (2012) *Machine learning: a probabilistic perspective*. Cambridge, Mass: MIT press

Pyo, S.-H., Lee, K.-H., & Lee, K.-T. (2021). Estimation of column aerosol contribution in Seoul and Gangneung using machine learning clustering technique. *Journal of Korean Society for Atmospheric Environment*, *37*(6), 931–945. (in Korean with English abstract). https://doi.org/10.5572/KOSAE.2021.37.6.931

Wong, M. S., Zhu, R., Yin Tung Kwok, C., Kwan, M., Santi, P., Ho Liu, C., Qin, K., Lee, K. H., Heo, J., Li, H., & Ratti, C. (2021). Association between NO_2 concentrations and spatial configuration: A study of the impacts of COVID-19 lockdowns in 54 US cities. *Environmental Research Letters*, *16*(5). https://doi.org/10.1088/1748-9326/abf396

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.