

Spatiotemporal imputation of missing aerosol optical depth using hybrid machine learning with downscaling

Abula Tuheti ^{a,b}, Zipeng Dong ^{c,*}, Guanghua Li ^{a,b}, Shunxi Deng ^{a,b}, Zhengqiang Li ^{d,e,f}, Li Li ^{d,e,f}

^a School of Water and Environment, Chang'an University, Xi'an, 710064, China

^b Key Laboratory of Subsurface Hydrology and Ecological Effects in Arid Region, Ministry of Education, Chang'an University, Xi'an, 710064, China

^c Meteorological Institute of Shaanxi Province, Xi'an, 710014, China

^d State Environmental Protection Key Laboratory of Satellite Remote Sensing, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, 100101, China

^e Key Laboratory of Remote Sensing and Digital Earth, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, 100101, China

^f Aerospace Information Technology University, Jinan, Shandong, 250200, China

HIGHLIGHTS

- Developed a robust RF-AutoResnet method combining Random Forest and Autoencoder's Deep Residual Neural Network for estimating missing MAIAC AOD data.
- The RF model downscales MERRA-2 AOD more accurately than Bilinear and Nearest algorithms, achieving better R and RMSE values.
- The RF-AutoResnet model captures spatiotemporal heterogeneity of AOD effectively from 2019 to 2022 in the HCHO + SO₂+NO₂ scenario from 2019 to 2022.
- The method outperformed other models with higher R² and lower RMSE and MAE values.

ARTICLE INFO

Keywords:

MODIS
Aerosol optical depth
Gap-filling
Autoencoder's deep residual neural network
PM2.5

ABSTRACT

The Multi-Angle Implementation of Atmospheric Correction (MAIAC) provides high spatiotemporal resolution Aerosol Optical Depth (AOD), which is crucial for characterizing aerosol evolution on a spatial scale and supports exposure assessments and health research. The influence of clouds and high surface reflectance causes the loss of MAIAC AOD on a spatial scale, leading to inaccuracies in fine particulate matter (PM_{2.5}) retrieval and health risk assessments over the Fenwei Plain (FWP). However, the column concentration of PM_{2.5} precursors has not been used to impute missing AOD. Meanwhile, a spatial downscaling model that reduces errors and higher spatial resolution needs to be developed to assist in the AOD gap-filling task. Thus, we propose a robust estimation method for imputing missing MAIAC AOD (RF-AutoResnet) using a combination of Random Forest (RF) and Autoencoder's Deep Residual Neural Network (AutoResnet). The RF model downscales the Modern-Era Retrospective analysis for Research and Applications, Version 2 (MERRA-2) AOD spatially and introduces AutoResnet for AOD imputation. Compared with the Bilinear and Nearest algorithms, the RF downscaling model more accurately reflects the spatial variability of AOD, achieving better R (0.71) and lower RMSE (0.10). Scenario analysis found that in the HCHO+SO₂+NO₂ scenario (Simultaneously adding formaldehyde (HCHO) sulfur dioxide (SO₂), and nitrogen dioxide (NO₂) to the RF-AutoResnet model) scenario, from 2019 to 2022, the RF-AutoResnet model captured the spatiotemporal heterogeneity of AOD on a fine spatial scale for missing daily and monthly average MAIAC AOD. Compared with Bilinear-AutoResnet and Nearest-AutoResnet, the RF-AutoResnet method achieved better performance (average R² = 0.86–0.98; average RMSE = 0.02–0.05; average MAE = 0.01–0.02). Considering the gap-filled AOD estimated by the column concentration of PM_{2.5} precursors, it shows a strong correlation with the AOD of the Sun sky radiometer Observation NETwork (SONET) (R = 0.91, RMSE = 0.06, Bias = 0.01). These results underscore the reliability of this hybrid machine learning approach for AOD estimation. This is highly significant for improving the precision of evaluating PM_{2.5} effects on human health at a finer scale and enhancing the simulation and prediction capabilities of climate models.

* Corresponding author.

E-mail address: dzp2003@126.com (Z. Dong).

1. Introduction

Atmospheric aerosols play a crucial role in influencing the balance of Earth's radiation (Wei et al., 2023b). They influence the Earth's energy budget both indirectly, by regulating cloud formation and atmospheric circulation, and directly, through scattering and absorbing incoming solar radiation. Recent studies have highlighted that this influence is a key driver of climate change and environmental concerns (Bao et al., 2023; Ma et al., 2022). Therefore, obtaining precise measurements of aerosol loadings and their properties is crucial. AOD is a key parameter that quantifies the concentration of aerosols in the atmospheric column and their impact on sunlight transmission (Garay et al., 2016). AOD provides valuable insights into climate change and radiative forcing, as well as opportunities for estimating the concentration of PM_{2.5}. Accurately quantifying the spatiotemporal variability of AOD is crucial, as it informs public health policies and interventions by reliably estimating PM_{2.5} exposures and their impact on human health (Wei et al., 2023a, 2023c).

Currently, there are two methods for obtaining AOD. The long-term AOD data have been collected through both continuous ground-based and satellite measurements, significantly advancing our understanding of atmospheric aerosols. However, in-situ observations have inherent limitations, particularly in terms of spatial coverage. These observations are often unevenly distributed globally, with higher density in populated regions and sparse coverage in remote areas, which presents challenges for spatial extrapolation. In contrast, space-borne remote sensing has proven effective for studying long-term aerosols on a larger scale due to its ability to cover vast areas, including high-reflectivity surfaces like snow and ice (Hu et al., 2021). Therefore, using remote sensing AOD to understand aerosol-climate interactions over large geographic areas has attracted considerable interest. Recently, MAIAC has enhanced the accuracy of cloud detection, aerosol retrievals, and atmospheric correction using time series analysis and pixel and image-based processing (Lyapustin and Wang, 2022). Retrieved MAIAC AOD has been extensively used to examine regional aerosol loading variations and surface PM_{2.5} estimates from a satellite perspective (Lee, 2019; Lin et al., 2015; Meytar et al., 2015; Reid et al., 2021; Song et al., 2014). However, these analyses are limited to clear-sky conditions (Li et al., 2020). The ubiquitous presence of cloud cover (60–70% globally) leads to a substantial data gap and significantly impedes our ability to observe air pollution from space. According to Xiao et al. (2017), the AOD missing rate in the Yangtze River Delta of China was over 60% during 2013–2014. Missing data is inherently non-random and varies across space and time, posing a significant challenge for gap-filling. It exhibited great spatiotemporal heterogeneities determined by diverse factors, making gap-filling more complicated.

To obtain full-coverage AOD, a variety of approaches have been developed using spatial neighborhood information, lowering the cloud screen criteria, and improving the retrieval algorithm (Aaron et al., 2011; Nordio et al., 2012; Yang and Hu, 2018). Geostatistical models, such as inverse distance-weighted interpolation (IDW), kriging interpolation, spline function interpolation, spatiotemporal smoothing model, and generalized additive models (GAM), have also been applied to reconstruct the gap-filled AODs (Chen et al., 2019; Just et al., 2015; Lv et al., 2016; Xiao et al., 2017). These methods enhance AOD coverage by utilizing interpolation techniques without additional data inputs, streamlining the process. Although such approaches increase data availability, their effectiveness may be limited in regions with scarce observational points, potentially leading to suboptimal outcomes. Moreover, they ignore important influences, such as meteorology and cloud fraction (the missing rate is still above 30%) (Guo et al., 2023). Studies on filling the missing values of AOD products with numerical model AOD have shown a growing trend, in establishing functional relationships between remote sensing AOD and numerical model AOD and other geographic factors. The Empirical Orthogonal Functions method was applied to fuse the global GOCART model data to obtain gap-filled

MODIS Level 3 AOD monthly/annual products (Liu and Pinker, 2005). A Multiple Imputation model was established based on the chemical transport model (CTM) AOD data, considering the terrain and meteorological factors, and full-coverage MAIAC AOD data were obtained (Xiao et al., 2017). Moreover, machine learning has achieved good results in the field of AOD gap-filling. A feed-forward network was developed to associate CTM outputs to fill in the missing AODs (Di et al., 2016). A deep residual neural network was developed using AOD in coarse resolution, combining meteorological and terrain factors, and successfully obtained gap-filled AODs (Li et al., 2020). Full-coverage AODs were generated by fusing remote sensing inversion (MAIAC, Himawari-8) and the CTM AOD results based on random forests (Jiang et al., 2021). Gap-free daytime AOD products were generated by combining ground-based measurements, remote sensing retrievals (Himawari-8), and numerical model simulations using optimal interpolation (Ke et al., 2022). A random forest (RF) model was constructed to generate high-resolution, full-coverage AOD maps globally, accounting for meteorological and cloud conditions (Guo et al., 2023). Moreover, an improved extremely randomized trees model was trained to reconstruct the satellite gap-free AODs using a bilinear interpolation approach (Wei et al., 2023b). However, uncertainties arising from model simulations and their inherent spatial resolution hinder accurate aerosol representation.

According to previous studies, terrain, vegetation, emission inventories, and meteorological factors are commonly used as important variables for AOD gap-filling. Importantly, the previous studies neglected the effect of PM_{2.5} precursor concentration on AOD when filling in the missing AOD values. Under low wind speed and high humidity conditions, gaseous precursors such as NO₂, SO₂, and HCHO undergo photochemical reactions to produce nitrate and sulfate aerosols, further elevating PM_{2.5} levels. NO₂ is oxidized in the atmosphere to form nitric acid (HNO₃), which then reacts with ammonia (NH₃) to produce particulate ammonium nitrate (NH₄NO₃), a major source of nitrates in PM_{2.5}. SO₂ is primarily oxidized by hydroxyl radicals (OH) to form sulfuric acid (H₂SO₄), which then reacts with ammonia to produce ammonium sulfate, an important component of PM_{2.5}. SO₂ can also dissolve in cloud water and reacts with hydrogen peroxide (H₂O₂) or O₃, forming H₂SO₄ (Peng et al., 2021). H₂SO₄ plays a critical role as a crucial precursor for the nucleation of new aerosol particles, which subsequently condense onto existing particles, thus increasing PM_{2.5} concentrations (Zhang et al., 2015). Additionally, HCHO reacts with aqueous SO₂ (SO_{2,aq}) to generate hydroxymethyl sulfonate (HMS; HOCH₂SO₃⁻) (Jonathan et al., 2018). HMS is stable and less susceptible to oxidation by hydrogen peroxide (H₂O₂) and O₃ but can be oxidized by OH radicals to form sulfate (Dovrou et al., 2019). Moreover, HCHO reacts with H₂O₂ to produce hydroxymethyl hydroperoxide (HMHP), which accelerates the oxidation of dissolved SO_{2,aq} to sulfate (Dovrou et al., 2019). Overall, atmospheric HCHO enhances sulfate formation in PM_{2.5} by generating HO₂ radicals and HMHP or HMS. Sulfate is a significant precursor of PM_{2.5}. These reactions increase the concentration of organic aerosols, adding to the overall aerosol load. Research has shown rising AOD loadings are typically linked to higher PM_{2.5} concentrations (Wei et al., 2020). Therefore, considering the concentration of PM_{2.5} precursors comprehensively during the process of gap-filling AOD may reduce the estimation error.

Deep learning is widely applied for filling missing AOD values due to its powerful nonlinear fitting capabilities. However, the feed-forward neural networks experience accuracy loss and saturation when applied to regression problems, as the number of hidden layers increases. Incorporating identity mapping via residual connections within neural architectures can enhance learning efficacy, a technique extensively used in convolutional neural network (CNN) frameworks for deep learning (He et al., 2016). Despite their strengths, CNN-based approaches face limitations when confronted with the extensive gaps in MAIAC AOD datasets, as these methods typically require training on complete or minimally impaired image inputs. In contrast, deep residual

networks can better preserve the input information flow and are more suitable for reconstructing the missing image data, owing to their residual blocks that mitigate the optimization challenges of very deep networks that are hard to optimize. The deep residual network framework has shown superior performance in a variety of air quality forecasting applications (Li et al., 2020).

Furthermore, Standardizing the spatiotemporal resolution of variables is an essential step in addressing gaps in AOD data. Previous studies aimed to obtain full-coverage AOD datasets with consistent spatiotemporal resolution, and often utilizing resampling methods to apply low-resolution residuals to high-resolution. A common challenge for spatial interpolation methods, such as IDW and GAM, is their limited applicability (AErosol RObotic NETwork (AERONET) AOD validation R²: 0.18–0.44) or insufficient spatial coverage of estimation (Aaron et al., 2011; Nordio et al., 2012; Xiao et al., 2017). Traditional interpolation techniques, including bilinear and nearest-neighbor interpolation, fail to accurately capture the fine-scale spatial heterogeneity of aerosol concentrations. While computationally simple, these approaches introduce spatial smoothing, obscuring local pollution hot-spots—especially in urban-industrial areas. This lack of fine resolution hinders accurate air quality assessments. Traditional methods, including kriging, primarily rely on spatial autocorrelation and do not integrate aerosol precursor data, which are critical for understanding secondary aerosol formation. As a result, they do not account for the complex nonlinear relationships between precursor gases (HCHO, SO₂, NO₂) and aerosols that influence AOD variability. Thus, for the cases where the original data resolution and the downscaling target resolution differ significantly, the resampling results in some errors in the spatial distribution of high-resolution residuals. Therefore, developing a downscaling model that minimizes such errors and preserves spatial detail is crucial for improving the accuracy and utility of AOD gap-filling tasks.

Accelerated urban development and industrial expansion in the Fenwei Plain (FWP) have significantly deteriorated air quality, as reflected in elevated PM_{2.5} levels, particularly during winter heating periods. The level of coal consumption in FWP is higher than the national average.¹ To improve air quality, a series of policies for controlling air pollution were implemented. Then, the FWP was determined as one of the areas for air pollution prevention by the Ministry of Ecology and Environment in 2018.² Intensive highway transportation and population density further contribute to pollutant emissions in this area. Moreover, the distinct topographical features of the FWP create conditions conducive to the retention of air pollutants. Consequently, this region exhibits some of the highest yearly average concentrations of PM_{2.5} within China's borders. Recent studies have highlighted that PM_{2.5} levels in the FWP surpass those in other major polluted regions (Cao and Cui, 2021). Therefore, to evaluate the efficacy of clean air policies and further optimize strategies for controlling air pollution, elucidating historical patterns in AOD loadings is essential.

To address these challenges, we developed a novel hybrid approach, the RF-AutoResNet model, which addresses key limitations of both traditional interpolation methods and existing machine learning-based techniques for AOD gap-filling and downscaling. Specifically, the RF-AutoResnet model integrates Random Forest (RF) for spatial downscaling and an autoencoder-based Deep Residual Neural Network (AutoResnet) for gap-filling. The RF component of our model significantly enhances spatial resolution by downscaling MERRA-2 AOD from its coarse 50 km grid to a fine-scale 1 km resolution. This capability is crucial for regions such as the FWP, where sharp pollution gradients are observed between densely industrialized urban centers and rural areas. Our approach incorporates column precursor gas concentrations

(HCHO, SO₂, NO₂), which are integral to secondary aerosol formation processes. By integrating these variables, the RF-AutoResnet model accounts for the dynamic chemical processes that contribute to aerosol loading, improving the overall accuracy of AOD imputation. The influence of PM_{2.5} precursors on AOD reconstruction was further assessed through scenario analysis. Finally, the accuracy of the gap-free AOD was validated based on in situ measurement. This study provides a new approach for reconstructing AOD data and has important implications for conducting fine-scale inversion of ground-level PM_{2.5} concentration and health risk assessment over a large scale.

2. Data and method

2.1. Study area

The FWP (33.5°N–38.7°N, 106.3°E–114.2°E) is situated at the junction of Shaanxi Province, Shanxi Province, and Henan Province in central China. The urban cluster of the FWP includes eleven cities: Xi'an (XA), Tongchuan (TC), Baoji (BJ), Xianyang (XA), and Weinan (WN) in Shaanxi Province, and Jinzhong (JZ), Yuncheng (YC), Linfen (LF), and Lvliang (LL) in Shanxi Province, and Luoyang (LF) and Sanmenxia (SMX) in Henan Province. The FWP, which encompasses an area of approximately 7×10^4 km²,³ is the fourth-largest plain in China (Cao and Cui, 2021). The energy structure of the FWP is predominantly coal-based, with coal accounting for nearly 90% of energy consumption, far exceeding the national average of 60%.⁴ Multi-chemical industries, such as alumina production, steel, and coal chemical processes, have led to significant industrial pollutant emissions in this region.⁴ Moreover, the industrial structure, which is dominated by heavy industry, has further increased the traffic jams, exacerbating the problem of transportation-related pollution. Consequently, in 2018, this region was designated a critical area for governance under China's "Three-Year Action Plan for Winning the Blue-Sky Defense".⁵ According to 2020 statistics, the annual PM_{2.5} concentration of FWP exceeded the national average for 337 cities by 45.4%.⁶ Additionally, air quality standards were exceeded on 29.4% of days across all 11 FWP cities, compared to 13.0% nationally.⁷ With the influence of emissions from various sources, secondary atmospheric pollutants are formed via homogeneous and heterogeneous chemical reactions, exacerbating the formation of the "air pollution complex" in China. The current state of air pollution over the FWP is a severe constraint on local economic development, public health, and quality of life. Elucidating the spatiotemporal variations AOD level is therefore essential for policymakers seeking to implement targeted, effective pollution control strategies across the FWP.

2.2. Data and methods

2.2.1. MAIAC AOD

Analysis of AOD changes was based on Moderate Resolution Imaging Spectroradiometer (MODIS) data. The MODIS, launched by the Earth Observing System (EOS) program of the National Aeronautics and Space Administration (NASA), was employed to observe the Earth's terrestrial and oceanic characteristics. The MODIS is capable of monitoring aerosols and clouds with high precision from space in a long-term and continuous manner. Due to its advantages of a short revisit period, wide

¹ The source of its website is https://www.mee.gov.cn/xxgk2018/xxgk/xxgk15/201911/t20191113_742104.html.

² The source of its website is https://www.mee.gov.cn/xxgk2018/xxgk/xxgk15/201901/t20190122_690319.html.

³ The source of its website is http://www.gov.cn/zhengce/content/2018-07/03/content_5303158.htm.

⁴ The source of its website is [P020210526572756184785.pdf \(mee.gov.cn\)](P020210526572756184785.pdf).

⁵ The source of its website is <https://www.mee.gov.cn/hjzl/sthjzk/zghjzkgb/>.

coverage, and high resolution, MODIS has attracted numerous researchers. In this study, AOD retrievals from the MAIAC products (S-AOD), with a high latitudinal and longitudinal resolution of $1\text{ km} \times 1\text{ km}$, were used to investigate the spatial-temporal imputation of aerosols over the FWP from January 2019 to December 2022.

2.2.2. SONET ground-based measurements

A Cimel sunphotometer CE-318 was installed at the top of a building of the Institute of Earth Environment (SONET) (34.2°N , 109°E), Xi'an, Shaanxi Province. The CE-318, manufactured by CIMEL Electronique, is a multi-channel, automatic sun-and-sky scanning radiometer designed to direct solar irradiance and sky radiance (<http://www.sonet.ac.cn/>). It performs direct spectral solar radiation measurements within a 1.2° full field-of-view every 15 minutes across eight standard bands: 340, 380, 440, 500, 675, 870, 1020, and 1640 nm. The CE-318's high-accuracy data are widely used for the validation of satellite remote sensing retrievals and model simulations, ensuring the high reliability of the data used in this study. To match the satellite's overpass time, the 15-minute AOD observations from CE-318 were interpolated to 550 nm using the Ångström power law between the nearest wavelengths 440, 500, and 675 nm. The interpolated 550 nm CE-318 AOD data were further averaged over 60 minutes to enhance temporal alignment. These high-precision, interpolated AOD values serve as the "true" reference to validate the gap-free AOD datasets obtained from satellite observations.

2.2.3. Auxiliary data

Meteorological variables have been proven to affect AOD loadings (Wei et al., 2023b), including temperature, relative humidity, wind speed, wind direction, surface pressure, accumulated evaporation and precipitation, high/middle/low cloud fractions, surface incoming shortwave flux, surface net downward shortwave flux, and top of atmosphere (TOA) net downward shortwave flux were extracted and averaged to daily values. MERRA-2 is the latest version of global atmospheric reanalysis for the satellite era produced by NASA Global Modeling and Assimilation Office (GMAO) using the Goddard Earth Observing System Model (GEOS). MERRA-2 includes total aerosol extinction AOD (M-AOD (550 nm)), the component and column mass density of the PM_{2.5}. Its high temporal resolution ensures consistent and reliable aerosol estimates over long periods, which are indispensable for imputing MAIAC AOD. Importantly, the secondary formation of PM_{2.5} through chemical reactions relies heavily on precursor gases. Among these, formaldehyde is a key intermediate gas in nearly all oxidation chains of non-methane volatile organic compounds (NMVOCs). For this reason, formaldehyde, derived from the TROPospheric Monitoring Instrument (TROPOMI), was used as a proxy for VOCs. The other precursor data(NO₂ and SO₂) are also provided by the TROPOMI. Thus, precursors, the component of PM_{2.5}, and the PM_{2.5} component with column mass density were also involved. Moreover, factors related to fire information, economy, population, land cover, and terrain were also taken into account. These variables were derived from high-relevance satellite remote sensing products with strong availability, including: - Global high-resolution annual LandScan population data - Monthly VIIRS nighttime light data - MOD13A1 monthly normalized difference vegetation index (NDVI) - MYD14A1 daily fire information - Shuttle Radar Topography Mission (SRTM) terrain data. The date choice for all variables is January 2019 to December 2022. Text S1 and Text S3 in the supplementary material introduce the details of the data and the download address. In total, 33 independent variables were included in this study (Supplementary Materials Table S1).

2.2.4. Deep residual network based on autoencoder (AutoResNet)

AutoResNet is a core component for reconstructing gap-filled AODs (Fig. 2). An autoencoder neural network consists of input layers, multiple encoding layers, a latent representation layer, more decoding layers, and output layers. The purpose of the autoencoder is to compress the input data to the size of the latent representation layer in the

encoding layer and to reconstruct the data to the output layer during decoding. Therefore, the decoding layer and the encoding layer have a completely symmetrical structure. In the process of compressing and reconstructing multidimensional data, the architecture of the neural network exhibits a distinct pattern. The encoding phase progressively reduces the number of nodes across successive layers, while the decoding phase incrementally increases the node count as the network depth expands. This kind of autoencoder structure is similar to principal component analysis (PCA), which can effectively reduce the influence of multicollinearity among the input features. However, unlike PCA, which relies on linear transformations, the AutoResNet framework employs non-linear transformative operations, enhancing its ability to capture complex relationships within the data. As shown in Fig. 2, there are 3 encoding layers, and the number of neural network nodes in each encoding layer is 128, 64, and 32, respectively. The latent representation layer comprises 16 neural units, while the trio of decoding stages exhibits a symmetrical architecture in terms of node quantity, mirroring their encoding counterparts.

We trained AutoResNet on $1\text{ km} \times 1\text{ km}$ pixels that had daily S-AOD. As described in the auxiliary data, 32 covariates are included in our AutoResNet, except for S-AOD. The network parameters of AutoResNet are set as shown in Table 1. The loss function of AutoResNet is Mean Squared Error (MSE):

$$\text{Loss} = \frac{1}{n} \sum_{i=1}^n (y_{ip} - y_i)^2 \quad (1)$$

Where n represents the sample numbers, y_{ip} represents the model predicted value, and y_i represents the true value.

2.2.5. Downscaling of MERRA-2 AOD (M-AOD) with random forest model (RF-D)

MERRA-2 data plays a critical role as an input feature in the AutoResNet model, contributing to the interpolation of missing MAIAC AOD data. Selected for its continuous global coverage and regular temporal resolution, MERRA-2 is particularly well-suited for gap-filling high-resolution datasets like MAIAC. Its integration of satellite observations, ground-based measurements, and advanced atmospheric modeling ensures high-quality aerosol estimates, that effectively capture large-scale transport patterns. By combining MERRA-2's broad coverage with MAIAC's finer resolution, the RF-AutoResnet model achieves more accurate AOD imputation, especially in regions like the FWP. Additionally, MERRA-2 provides detailed information on aerosol components, enabling the model to better account for both anthropogenic emissions (e.g., industrial pollutants) and natural events (e.g., dust storms), thereby enhancing the overall accuracy of AOD predictions. To ensure the utility of MERRA-2 data at a finer spatial scale, downscaling is required. The core principle of MERRA-2 AOD (M-AOD) downscaling lies in the assumption that the statistical relationships between M-AOD and relevant regression coefficients remain consistent across spatial scales. This assumption enables the application of these relationships, derived at coarser resolutions, to generate high-resolution estimates. The main formula is as follows:

$$AOD_c^p = f(var_c) \quad (2)$$

$$\Delta = AOD_c - AOD_c^p \quad (3)$$

$$AOD_f^p = f(var_f) + \Delta \quad (4)$$

Where, AOD_c is the low-resolution M-AOD; AOD_c^p is the predicted low-resolution M-AOD; AOD_f^p is the predicted high-resolution M-AOD; var_c is the low-resolution regression factor; var_f is the high-resolution regression factor; and Δ is the fitting residual.

We incorporate M-AOD as a key input in our AOD imputation framework. However, the spatial resolution of M-AOD (50 km) is

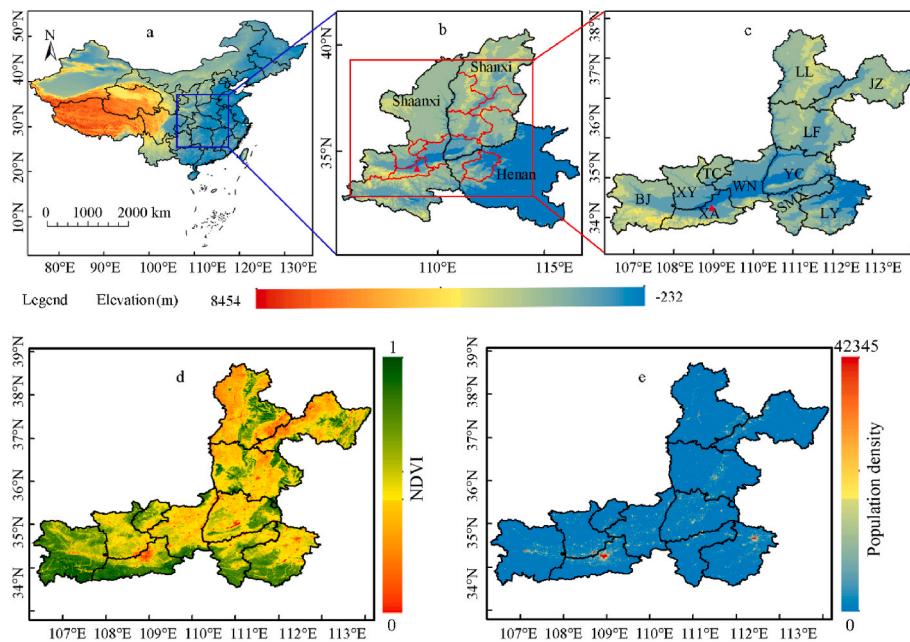


Fig. 1. Geographic location (a, b, and c) and land use and land cover of the FWP. The red triangles in Fig. 1b and c represent the locations of SONET sites. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

significantly coarser than that of our target variable (AOD at 1 km resolution). During the development of RF-D, our initial imputations exhibited spatial patterns resembling the coarse 50 km M-AOD grid. Applying a preprocessing downscaling step enhanced the spatial coherence between the input variables and the high-resolution (1 km) target data.

Thus, we developed an RF algorithm to construct an M-AOD downscaling model, conducting a downscale from M-AOD (50 km) to 1 km spatial resolution. The RF algorithm represents an advanced ensemble learning method based on decision tree methodology. This approach employs bootstrap aggregating (bagging) techniques to generate diverse subsets from the original dataset. These subsets serve as the foundation for constructing multiple decision trees, which are then aggregated to form a cohesive predictive model. Voting scores are a critical step in generating the results of the regression model. The main parameters of the random forest algorithm include the number of decision trees and the number of preselected variables available from the tree nodes. A decision tree is the number of sub-datasets produced from the original dataset by putative back sampling. The number of preselected variables at the tree nodes is the maximum number of features taken into account for constructing the optimal decision tree model. In this study, we used random search to optimize these key parameters. To fully demonstrate the spatial distribution of the high spatial resolution M-AOD after downscaling, the area corresponding to the original resolution M-AOD data should not be too large, which results in a low number of input samples for the RF model. Considering that machine learning models require a minimum of 50 samples, we selected an area of 1337×1114 pixels as the study region to construct the downscaling model.

We consider the effects of terrain, meteorology, and precursor concentrations on M-AOD to account for their comprehensive influence on M-AOD in the RF-D model construction. In our study, we first train a RF-D model using the low-resolution parameters listed in Table 1, and then input the high-resolution parameters into the trained RF model to predict the M-AOD at a finer resolution, based on the principle of “scale invariance”. Moreover, since the regression model cannot explain all the variations of M-AOD, we add residuals to the high-resolution regression results to obtain the final downscaled M-AOD. Auxiliary data for the RF-D downscaling algorithm is described in detail in Supplementary

Material Text S2. Key training parameters for the RF-D model are shown in Table 2.

2.2.6. Scenario experiment design and validation

In this study, we investigate the aerosol composition data using combined datasets of AOD from MODIS and column densities of NO₂, SO₂, and HCHO from TROPOMI. This combination is motivated by the fact that aerosols and these trace gases share significant anthropogenic and biogenic origins. Combustion of fossil fuels generates various aerosol precursor gases, such as nitrogen oxides (NO_x), Sulfur oxides (SO_x), and Non-methane volatile organic compounds (NMVOC), along with primary aerosol particles. These precursor gases subsequently form secondary particles like NO_x, SO_x, and organic aerosols. The combustion of biological materials serves as a substantial source of atmospheric pollutants, including NO_x, and NMVOC, as well as primary carbonaceous aerosols in the form of elemental carbon and organic particulates. Biogenic emissions of isoprene act as a precursor for formaldehyde and secondary organic aerosols (SOA). Furthermore, isoprene emissions are closely associated with other SOA precursors including monoterpenes. Given the simultaneous occurrence of trace gases and aerosols, a temporal and spatial correlation is observed between these components in regions where these sources are predominant, assuming comparable lifetimes.

The typical lifetime of aerosol particles in the lower troposphere is several days, whereas the trace gases employed typically exhibit lifetimes of no more than 24 h, particularly during the summer season. This discrepancy arises due to the complex atmospheric processes that aerosol particles undergo, including settling, wet deposition, condensation, and photochemical reactions, which contribute to their extended lifetimes. In contrast, trace gases typically rapidly diminish through processes such as chemical reactions, wet deposition, and vertical mixing. Due to the differential lifetimes of aerosol particles and trace gases, aerosols can be transported further from their source regions, while tracers tend to have a more limited range of dispersion.

According to the research by Veefkind et al. (2011) the diurnal variations in aerosol particles and trace gases within the troposphere allow for the neglect of the impact of differences in lifetimes through temporal and spatial averaging over at least one month and a large-scale

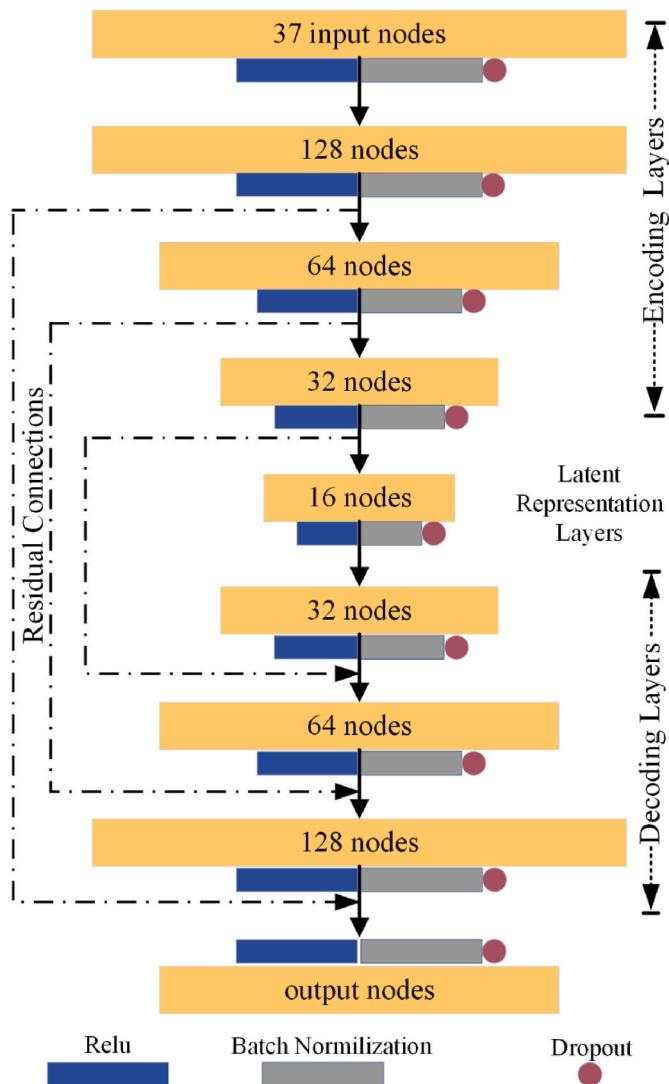


Fig. 2. The network structure of AutoResNet. The dashed lines indicate the residual connections from the encoder to the decoder's three hidden layers. The output of each layer is batch normalized, activated by Relu, and dropouted.

Table 1
The network parameters of AutoResNet.

Parameters	Value	Parameters	Value
Dropout	0	Epochs	100
Learning rate	1×10^{-4}	Shuffle	True in train data loader but false in validate data loader
Batch size	1024	Optimizer	Adam

Table 2
Key training parameters of RF-D model.

Parameters	Parameter Value	Parameters	Parameter Value
n_estimators	35	oob_score	True
max_depth	25	bootstrap	True
min_samples_leaf	15	random_state	10
min_samples_split	13		

region. By employing long-term and wide-ranging spatial averaging, the transport and mixing processes occurring at different times and locations can be comprehensively considered, thereby reducing the

variations caused by individual instantaneous observations. Consequently, this study employed monthly temporal averaging and large-scale spatial averaging to investigate the impact of different trace gas scenarios on AOD interpolation. By smoothing out variations in aerosols and trace gases with different lifetimes and considering them collectively, more accurate and reliable results are obtained.

We utilized daily precursor data for the daily AOD imputation to ensure that the short-term fluctuations in atmospheric conditions and precursor concentrations were accurately captured. This approach allows for a more precise reflection of the temporal dynamics at a daily resolution, which is important for estimating day-to-day variations in AOD.

For the monthly AOD imputation, we applied monthly averages of precursor data to align with the temporal resolution of the AOD data. This approach was based on the need to capture broader trends in the data over a longer period. By using monthly averages, we smooth out short-term variability, thereby providing a clearer understanding of the overall spatial distribution of AOD on a monthly scale. This distinction between daily and monthly imputation processes ensures that the model appropriately reflects the temporal characteristics of the data being imputed. Regarding the spatial range, the averaging region corresponds to an image area of 1337×1114 pixels, which includes the FWP and its surrounding regions.

To examine the effects of precursor concentration and the downscaling model on the prediction results, we developed 6 groups of 21 imputation models (Table 3). In the first step, we progressively incorporate NO_2 , SO_2 , and HCHO species into the AutoResNet model using bilinear interpolation, then assessed the influence of precursor concentrations on model estimation (referred to as the Bilinear-AutoResNet model). Secondly, we combined the bilinear interpolation with the results of our developed downscaling model and sequentially added NO_2 , SO_2 , and HCHO species to AutoResNet, and evaluated the impact of precursor concentration on model estimation (referred to as the RF-AutoResNet model). The third step is similar to the first step, except that bilinear interpolation was replaced by a nearest neighbor interpolation (referred to as the Neighbor-AutoResNet). It should be noted that in Table 3, NO_2 represents the scenario in which only NO_2 is added to the AutoResNet input, with all other auxiliary variables held constant. Similarly, SO_2 or HCHO represents the scenario where only SO_2 or HCHO , respectively, is added to the AutoResNet input, with all other auxiliary variables held constant. NO_2+SO_2 indicates that both NO_2 and SO_2 are input into the prediction model simultaneously, with all other

Table 3
Scenario analysis parameter settings based on AutoResNet.

Approaches	Species
Bilinear	NO_2 SO_2 HCHO $\text{HCHO} + \text{NO}_2$ $\text{HCHO} + \text{SO}_2$ $\text{SO}_2 + \text{NO}_2$ $\text{HCHO} + \text{SO}_2 + \text{NO}_2$
Nearest neighbor	NO_2 SO_2 HCHO $\text{HCHO} + \text{NO}_2$ $\text{HCHO} + \text{SO}_2$ $\text{SO}_2 + \text{NO}_2$ $\text{HCHO} + \text{SO}_2 + \text{NO}_2$
Bilinear + RF-D (Our downscaling model)	NO_2 SO_2 HCHO $\text{HCHO} + \text{NO}_2$ $\text{HCHO} + \text{SO}_2$ $\text{SO}_2 + \text{NO}_2$ $\text{HCHO} + \text{SO}_2 + \text{NO}_2$

auxiliary variables held constant. $\text{NO}_2+\text{SO}_2+\text{HCHO}$ means that NO_2 , SO_2 , and HCHO are all input into the prediction model simultaneously, with all other auxiliary variables unchanged. To handle bad values (e.g., missing or erroneous data) for the NO_2 , SO_2 , and HCHO species, we employed the following data preprocessing steps: In our study, the source data is filtered based on Quality Assurance (QA) values to ensure that only high-quality measurements were used, as per the European Space Agency guidelines (<https://sentiswiki.copernicus.eu/web/s5p-products>). Specifically: For NO_2 , we filtered the tropospheric NO_2 column number density band, retaining pixels with QA values of at least 75%. For HCHO datasets, we applied a QA threshold at 50%. For SO_2 datasets, because of noise on the data, negative vertical column values are often observed in particular over clean regions or for low SO_2 emissions. It is recommended not to filter these values except for outliers, i.e. for vertical columns lower than -0.001 mol/m^2 according to European Space Agency (<https://sentiswiki.copernicus.eu/web/s5p-products>). For missing values, we applied a spatial interpolation techniques. Spatially, we used Bilinear interpolation method to estimate missing values.

To improve the accuracy of the resampling results, we selected a larger area for the training samples. Therefore, the training sample size was 1337×1114 . To evaluate the generalization ability of AutoResNet, we used holdout cross-validation. We randomly divided the sample data into three parts: 70% of the samples were used for model training, 15% of the samples were used for model testing and the remaining 15% were used for model validation. The loss, correlation coefficient (R), coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE) of the validation set were used as the statistical summary of the validation overall performance. Moreover, we compare the RF-D AOD downscaling results to those obtained by a typical approach including bilinear, nearest neighbor.

$$R = \frac{\sum_{i=1}^n (AOD_p - \overline{AOD_p}) \times (AOD_t - \overline{AOD_t})}{\sqrt{\sum_{i=1}^n (AOD_p - \overline{AOD_p})^2 \times \sum_{i=1}^n (AOD_t - \overline{AOD_t})^2}} \quad (5)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (AOD_p - AOD_t)^2}{\sum_{i=1}^n (AOD_t - \overline{AOD_t})^2} \quad (6)$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^n (AOD_p - AOD_t)^2}{n}} \quad (7)$$

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |AOD_p - AOD_t| \quad (8)$$

Where, n represents the sample number, AOD_p represent the model predicted MAIAC AOD in model validation, and AOD_t represents the true MAIAC AOD.

AOD acquired by a SONET at the Institute of Earth Environment, Chinese Academy of Sciences ($34.22^\circ\text{N}, 109.0^\circ\text{E}$) from January 2019 to March 2020 was employed to assess the uncertainty of original MODIS AOD and gapless MODIS AOD data in this paper. Due to the different AOD products with spatial and temporal resolutions, data from divergent platforms are pairwise compared using different matching strategies. In detail, 1) the mean MODIS AOD over a $50 \text{ km} \times 50 \text{ km}$ window centered on the SONET site was compared with the temporal-averaged SONET AOD measured within ± 30 min of satellite overpass time. We evaluated the model performance using R, R^2 , RMSE, and Bias between the SONET and imputed AOD data.

$$R = \frac{\sum_{i=1}^n (AOD_{Sat} - \overline{AOD}_{Sat}) \times (AOD_{SONET} - \overline{AOD}_{SONET})}{\sqrt{\sum_{i=1}^n (AOD_{Sat} - \overline{AOD}_{Sat})^2 \times \sum_{i=1}^n (AOD_{SONET} - \overline{AOD}_{SONET})^2}} \quad (9)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (AOD_{Sat} - AOD_{SONET})^2}{\sum_{i=1}^n (AOD_{SONET} - \overline{AOD}_{SONET})^2} \quad (10)$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^n (AOD_{SONET} - AOD_{Sat})^2}{n}} \quad (11)$$

$$\text{Bias} = \frac{1}{n} \sum_{i=1}^n (AOD_{Sat} - AOD_{SONET}) \quad (12)$$

Where, n represents the sample size, AOD_{Sat} represents the MAIAC AOD or downscaling product of MERRA-2 AOD, and AOD_{SONET} represents the true value of AOD.

The validation process encompassed the following methodological assessments.

- 1) Non-imputed MAIAC AOD vs. SONET AOD
- 2) Imputed AOD with sensitivity experiment of best performance as shown in Table 3 vs. SONET AOD.

A technical flowchart was developed based on our research objectives (Fig. 3). We developed a downscaling model using RF to adjust additional variables to a 1 km resolution. Simultaneously, we obtained 1-km resolution results for MERRA-2 AOD using Bilinear and Nearest-neighbor interpolation methods. The accuracy of the RF-D model was determined to be superior to other methods on a spatial scale through comparison against ground-based AOD retrievals. Subsequently, by integrating RF and a deep residual network, we robustly estimated the long-term time series of missing MAIAC AOD in large heterogeneous regions. Additionally, we assessed the impact of $\text{PM}_{2.5}$ precursor concentrations on AOD gap-filling using the sensitivity analysis method described in section 2.2.6. Finally, we validated the accuracy of gap-free AOD based on SONET data.

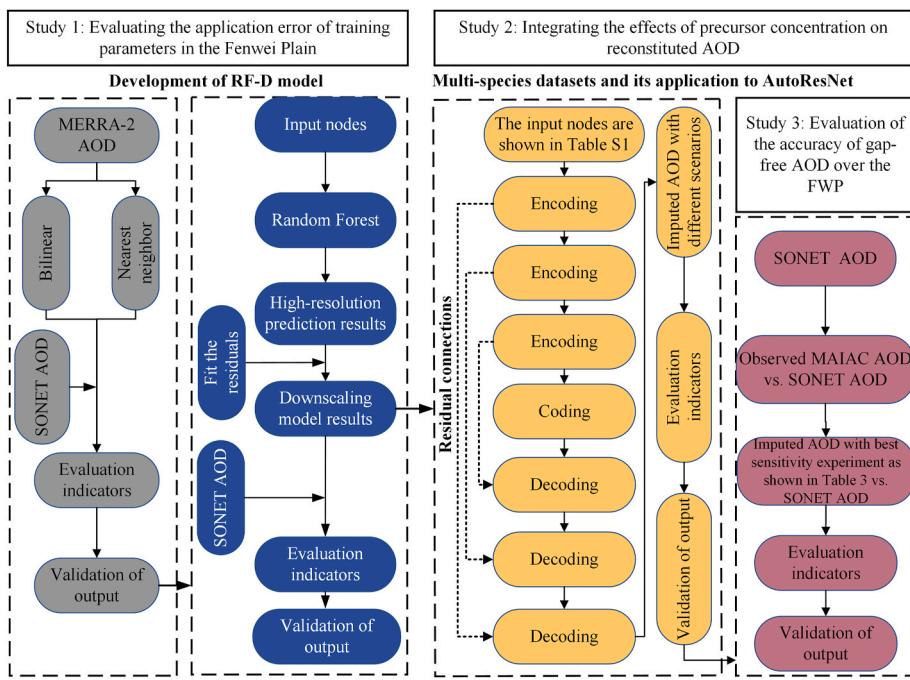
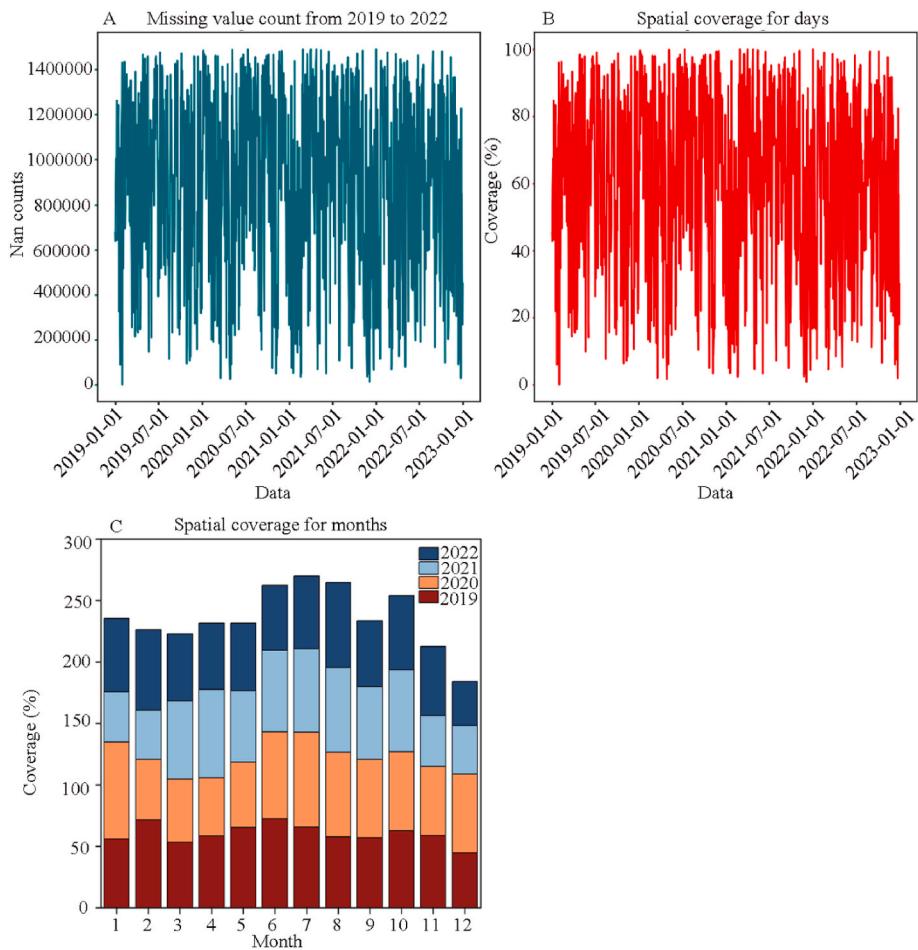
3. Results and discussions

3.1. Coverage of satellite AOD

The missing value counts and spatial coverage of MAIAC AOD are shown in Fig. 4. To study the coverage of missing values, we calculated ratio of missing AOD pixels to the total number of pixels. The number of missing AOD pixels reached as high as 1489355 on April 2, 2021 (Fig. 4A). The lowest AOD missing pixels (13,286 pixels) occurred on December 5, 2021 (Fig. 4A). The AOD missing rate reached 100% on February 28, 2021 and January 1, 2022 (Fig. 4B). The AOD missing rate on July 11, 2021 was close to 100% (Fig. 4B). Throughout the investigation period, MAIAC AOD exhibits a mean pixel-wise data unavailability of 58.93%. Notably, December 2022 demonstrates significantly higher data completeness with only 36.21% missing observations, in stark contrast to December 2020, which records a 63.98% missing rate (Fig. 4C). The monthly mean MAIAC AOD was 0.273 [standard deviation (s.d.): 0.094], with higher values in January 2020 [0.428; s.d. 0.194] and lower in June 2022 [0.167; s.d. 0.079] (Table S2).

3.2. Correcting modeled AOD by downscaling algorithms

Fig. 5 presents the spatial distribution of MERRA-2 AOD ($\sim 50 \text{ km}$)

**Fig. 3.** The technical flowchart of this study.**Fig. 4.** The total missing pixels and coverage of missing values for days and months over the FWP.

and high-resolution AOD (1 km) derived using bilinear interpolation, nearest-neighbor interpolation, and a random forest-based downscaling model (RF-D). The results indicate that the high-resolution AOD obtained via bilinear and nearest-neighbor interpolation is generally consistent with the original data. However, the spatial variability of AOD derived from these two interpolation methods is significantly insufficient compared to the complex geographical environment and human activity distribution in the FWP, failing to accurately reflect the actual spatial distribution characteristics within the region. Specifically, in the Qinling Mountains south of XA and BJ, both bilinear and nearest-neighbor interpolation tend to overestimate AOD values, while they underestimate in WN, inconsistent with the real-world situation. The forested areas along the edges of the Qinling Mountains typically exhibit lower AOD values. The low AOD values in mountainous regions can be attributed to several factors: first, the low population density in these areas results in lower energy consumption and traffic emissions; second, the rich vegetation in mountainous areas effectively reduces atmospheric particulate matter concentrations through absorption and deposition processes; and third, the topography of these regions facilitates pollutant dispersion, further lowering local AOD levels. In contrast, the interpolation methods underestimate AOD in the WN area, diverging from actual observations. According to China's 2020 Air Quality Rankings, WN ranked 15th from the bottom among all cities in the country, indicating relatively high AOD levels in this area.⁸ The elevated AOD in WN is closely linked to factors such as its industrial structure, urbanization, and geographic location. For example, the region's coal mining, thermal power plants, and chemical industries contribute significantly to particulate matter emissions. As of 2022, WN had 34 coal mines with a total designed capacity of 34.14 million tons per year.⁹ Additionally, WN's location in the eastern FWP makes it susceptible to the transmission of air pollutants from upstream urban clusters, further increasing AOD levels.

In comparison, the RF-D model more accurately captures the spatial heterogeneity of high-resolution AOD across the FWP. The RF-D model accurately identifies and represents the low AOD values characteristic of the Qinling Mountain region, which aligns well with real-world observations. The model not only accounts for topographical factors but also likely integrates multidimensional information such as vegetation coverage and population density, allowing a more precise depiction of AOD distribution in mountainous areas. In urban areas such as northern XA, WN, and SMX, where population density is higher, the RF-D model effectively captures AOD differences between cities, reflecting variations in industrial, commercial, and residential areas. In transitional zones between urban and rural areas, as well as between plains and mountainous regions, the RF-D model offers a more realistic representation of AOD distribution, avoiding the abrupt changes often observed in simple interpolation methods. This smooth transition better aligns with the gradual variations in AOD found in real-world geographical environments. Therefore, compared to traditional interpolation methods, the RF-D model not only accurately represents the spatial variability of AOD but also provides a more suitable input for gap-filling MAIAC AOD data.

Fig. 6 shows the fit of the AOD between the simulated AOD, Bilinear, Nearest neighbor, and our downscaling model (RF-D) to the AOD of the SONET sites. We compared the original AOD values as well as the downscaled AOD values obtained using Bilinear and Nearest neighbor methods against the in-situ AOD from SONET sites. We found that the Bias remained unchanged, while the R values ranged from 0.62 and 0.65 (Fig. 6 a, b, and c). Moreover, we find that RF-D achieves the highest R (0.71) with the AOD of SONET sites. The accuracy of the spatial distribution of AOD obtained by RF-D is improved by 8.4%–12.7% compared

to others. This also indicates that the results from the RF-D downscaling model were more accurate. It is noteworthy that the correlations in Fig. 6a, b, and c are nearly identical. This is primarily because the bilinear and nearest interpolation methods have hardly altered the original data, resulting in minimal changes in Fig. 6a, b, and c. However, our downscaling model has modified the spatial variations of the original AOD data, making it more representative of the actual FWP conditions (Fig. 5J, K, and L). Consequently, the correlation between our downscaling model and the SONET data is higher than the correlation between the downscaled AOD derived from bilinear and nearest interpolation methods and the SONET data.

Fig. 7 shows the test metrics for each month of the RF-D model from 2019 to 2022. It can be found that the MAE ranges from 0.03 to 0.07, the MSE ranges from 0.001 to 0.014, the RMSE ranges from 0.03 to 0.12 and the R^2 ranges from 0.55 to 0.89. The highest R^2 value was observed in the April 2019 ($R^2 = 0.89$), while the lowest was observed in the November 2019 ($R^2 = 0.71$). This may be due to the lower AOD missing rate in April 2019 and the higher missing rate in November 2019. The MAE, MSE, and RMSE values in October, November, and January were relatively high, which may be owing to an increase in snow cover extent and thick clouds, which absorb and scatter solar radiation, leading to a higher AOD missing rate (Guo et al., 2023). The performance of the AOD downscaling algorithm is potentially influenced by multiple factors beyond data availability. These include the spatiotemporal resolution of variables known to affect AOD, as well as unaccounted-for parameters. Notably, anthropogenic sources such as industrial activities and vehicular emissions may contribute to model discrepancies (Li et al., 2021). In addition, the MAE, RMSE, and MSE in March were all higher than 0.04, 0.06, and 0.004, respectively. This can be attributed to two factors. Firstly, the higher AOD in March resulted in relatively higher MAE, RMSE, and MSE from a statistical perspective. Secondly, the generation of dust storms in spring leads to the presence of dust aerosols in the region, which differs from the mixed-type aerosols assumed by the MODIS inversion algorithm in that area. These dust aerosols differ from the mixed-type aerosols assumed by the MODIS inversion algorithm in the area, contributing to significant biases in the retrieved AOD values. Overall, the developed model exhibits satisfactory accuracy, with RMSE, MAE, and MSE values well below 0.12, 0.07, and 0.008, respectively. These accuracy statistics are within the acceptable range as described by (Li et al., 2021).

To substantiate the robustness of our downscaling model, we conducted a comparative analysis of the RF-D downscaling model's performance against previous studies (Table 4). In the study conducted by Li et al. (2020), an autoencoder-based deep learning approach was proposed for the imputation of MAIAC AOD in California, demonstrating superior performance compared to conventional neural networks or nonlinear generalized additive models. The enhanced performance of the autoencoder-based method over the one presented in this paper may be attributed to the differences in the downscaling algorithms, with Li et al. (2020) employing a residual network approach. Deep residual networks are particularly adept at handling tasks with large datasets and high-dimensional features, such as imaging. In contrast, RF performs well with datasets comprising multiple feature types. Wang et al. (2022) developed the Artificial Neural Network Sequential Downscaling Method (ASDM) with Transfer Learning Enhancement (ASDMTE) model, which, when combined with high-resolution topographic data, showcased its downscaling capabilities in comparison to other deep-learning downscaling methods. However, the R^2 value obtained by Wang et al. (2022) was lower than that of this work, primarily due to the use of only one high-resolution explanatory variable (DEM), which may have limited the model's accuracy. While alternative statistical approaches such as GAM (Li et al., 2020) and Linear Models (LM) (Wang et al., 2022) are applicable in this context, their capacity for generalization is constrained when confronted with extensive training datasets. In contrast, machine learning methodologies typically exhibit superior adaptability under these circumstances. Taskinen et al. (2022) proposed

⁸ <https://www.mee.gov.cn/hjzl/sthjzk/zghjzkgb/202105/P020210526572756184785.pdf>.

⁹ <http://www.jlnyx.com/meishengchan/116903.html>.

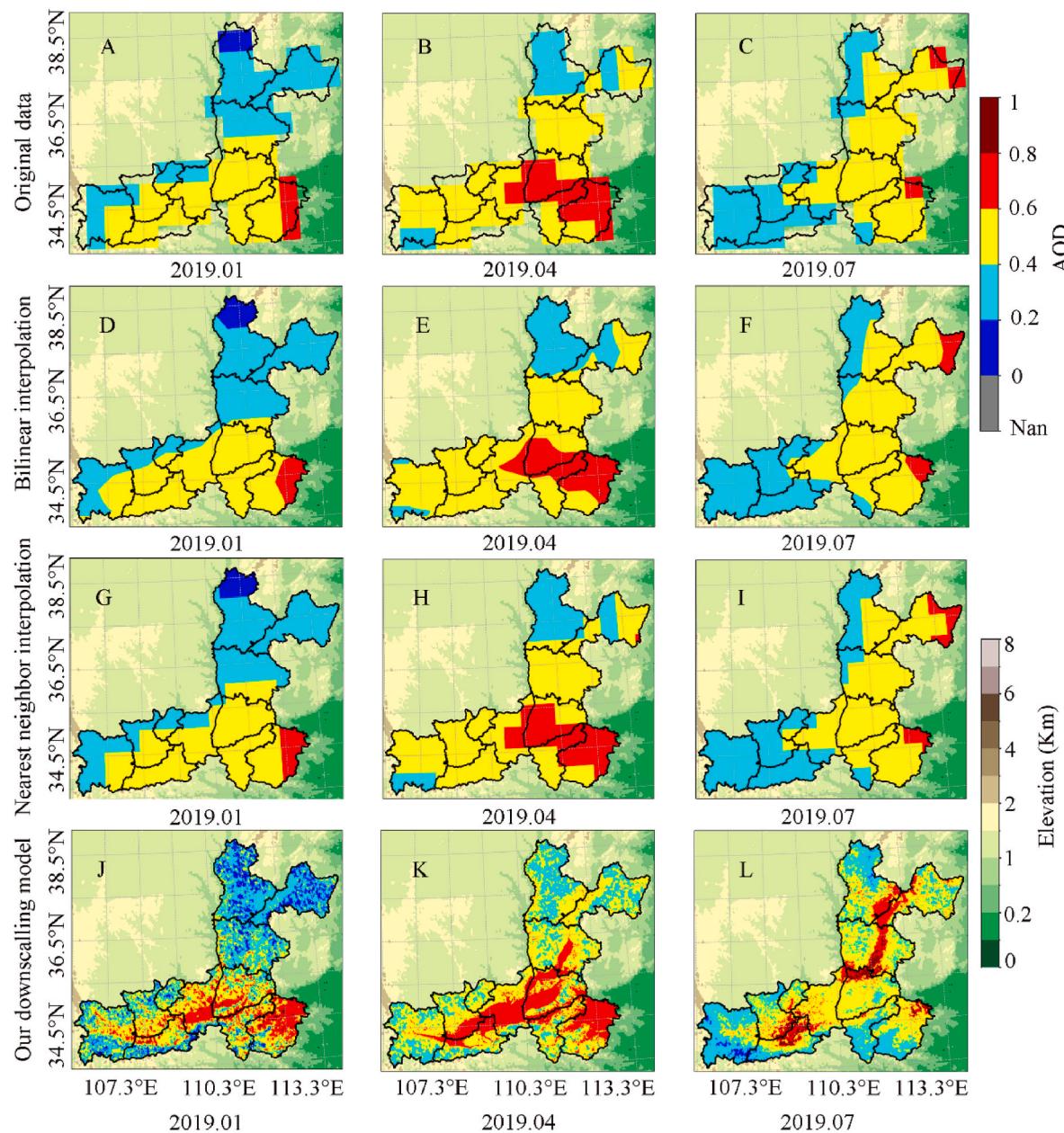


Fig. 5. Comparison of the original Modeled AOD (A, B, and C), downscaled AOD by Bilinear interpolation (D, E, and F), downscaled AOD by Nearest neighbor interpolation (G, H, and I), and downscaled AOD by our downscaling model (J, K, and L) (RF-D) over the FWP.

a downscaling method for coarse spatial resolution AOD using neural networks, enhancing the spatial resolution to 250 m and thereby improving validation accuracy. This model yielded commendable training results, with an R^2 value surpassing that of this work, largely because the method integrated machine learning with physical constraints, significantly enhancing the quality of satellite AOD data and its correlation with ground-based observations. Although they proposed an optimized algorithm, high-resolution data processing, and bias correction still require substantial computational resources and time, which may present bottlenecks when applied on a large scale. The training phase of the model provided by Taskinen et al. (2022) ([GitHub-TUT-ISI/DRAGONcorr at v1.0.0](#)) is time-intensive due to multiple factors: repeated trials with varied seeds, cross-validation across multiple subsets, high epoch limits (up to 10,000), small batch sizes, and computationally heavy distance calculations between MODIS and AERONET observations. Each of these factors contributes to the model's robustness but also significantly increases runtime. The RF-D model also

faces challenges. Its data preprocessing relies on large pandas DataFrames, which can slow down processing and cause memory issues. Additionally, hyperparameter tuning via GridSearchCV introduces further computational load. Liang et al. (2022) introduced the Bias-corrected Geostatistical Inverse Model (BGIM) downscaling algorithm effectively improving the resolution and accuracy of MODIS AOD seasonal products. However, due to the seasonal variability in aerosol component sources and errors in surface reflectance estimation, the accuracy of AOD products also varied seasonally, potentially affecting the overall performance of the algorithm. Consequently, the R^2 result of the BGIM downscaling algorithm was lower than that of the RF-D model. In comparison, our downscaling algorithm also employs a RF model, capturing the local AOD heterogeneity at a fine spatial scale. Our findings underscore that utilizing a RF model can achieve better downscaling performance relative to more traditional methods such as GAM.

Differences in spatial and temporal resolution among AOD test datasets used in various studies can significantly affect the predictive

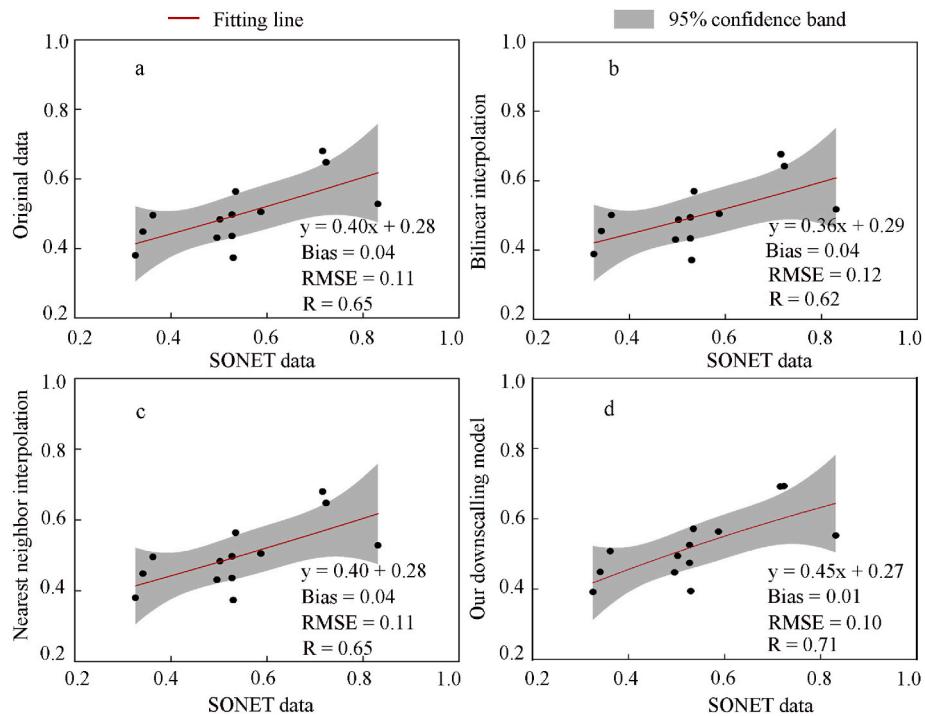


Fig. 6. Scatter plots between the monthly average original AOD, downscaled AOD generated using Bilinear methods, Nearest neighbor methods, and our downscaling model (RF-D) against the in-situ monthly average AOD from SONET sites.

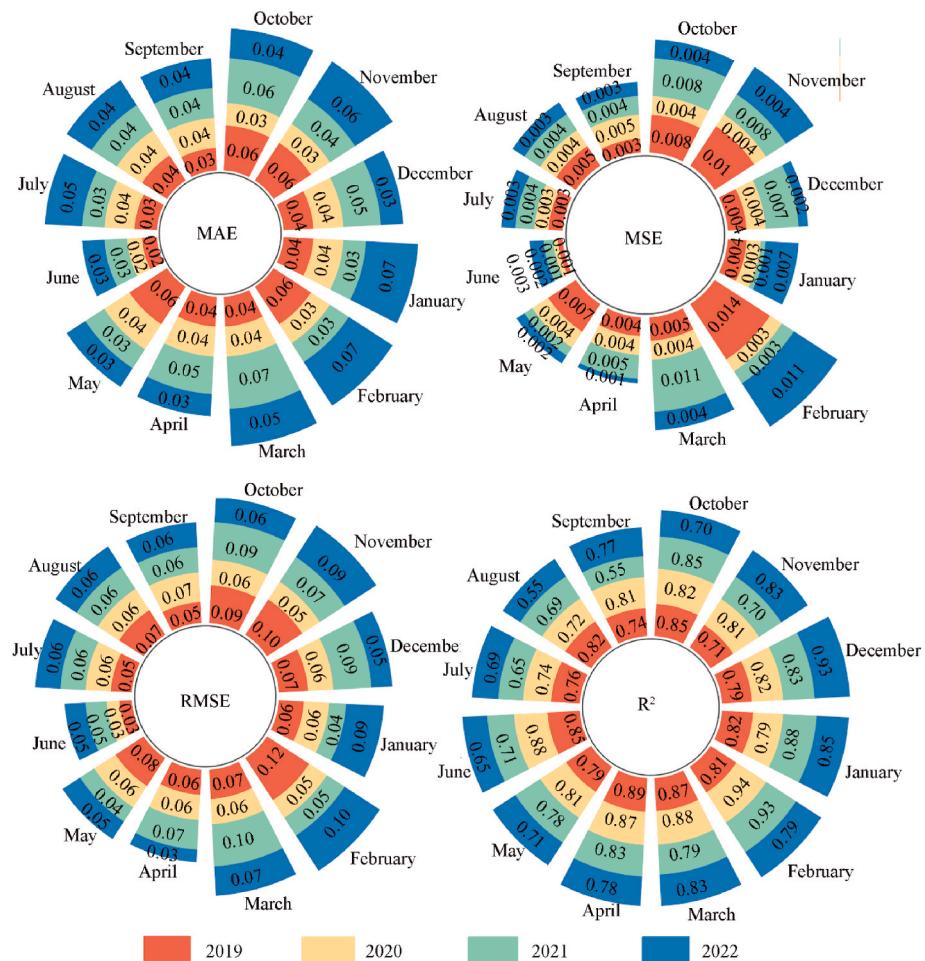


Fig. 7. Performance of RF-D model for monthly downscaling of modeled AOD.

Table 4

Comparison of downscaling model performance metrics with previous study.

Model name	RMSE	R ²	References
Residual network	0.001–0.01	0.23–0.99	Li et al. (2020)
GAM	0.003–0.22	0.22–0.98	Li et al. (2020)
The downscaling model combined with Bias Correction (SDBC)	0.060	0.85	Zhang et al. (2022a)
ASDMTE	0.075–0.010	0.44–0.75	Wang et al. (2022)
GAM	0.009–0.058	0.11	Wang et al. (2022)
LM	0.013–0.059	0.03–0.17	Wang et al. (2022)
Neural Network Models	0.105	0.86	Taskinen et al. (2022)
BGIM	0.21	0.79	Liang et al. (2022)
RF-D	0.030–0.120	0.55–0.93	This study

accuracy of models. For instance, lower-resolution datasets may fail to capture local air quality variations, leading to discrepancies in model predictions. Specifically, models using coarser spatial resolutions often show better performance metrics, such as R², at broader scales, while high-resolution datasets reveal finer details that positively influence evaluation metrics like R². For example, Higher resolution datasets, MAIAC AOD, capture more local variability, which may lead to lower errors (e.g., higher R² (0.86) and lower RMSE (0.105) in study of Taskinen et al. (2022)) and providing better insight into fine-scale processes. Conversely, the model from Wang et al. (2022), although potentially more accurate in capturing small-scale spatial heterogeneity, may exhibit lower R² values (0.44–0.75) due to its sensitivity to localized fluctuations.

The topographical complexity of different geographic regions (such as plains versus mountainous areas) can also impact the predictive capability of the models. For instance, models may exhibit lower predictive accuracy in mountainous regions compared to their performance in flatter terrains. Models tested in different regions—such as mountainous versus flat terrain—may exhibit varying performance. For instance, our RF-D model performs better in the complex terrain of the FWP, which includes the Qinling Mountains and Loess Plateau. Models trained on flatter regions might have different error distributions and accuracy metrics due to simpler atmospheric dynamics. However, Taskinen et al. (2022) use the Washington, D.C.—Baltimore, Maryland, USA, region as their region of interest. Yielding commendable training results with an R² value of 0.86, which surpasses that of this work.

Different studies may involve varying meteorological conditions, including the effects of extreme weather events, such as dust storms or heavy rainfall, on AOD measurements. In our research, which is based on the FWP, we are significantly influenced by dust storms originating from the Loess Plateau. These extreme weather conditions contribute to a lower R² range (0.55–0.93) for our model due to the increased variability in AOD measurements during such events. Similarly, another study conducted in a region affected by dust storms, specifically in the Middle East, reported lower R² range (0.44–0.75) values (Wang et al., 2022). This highlights the challenges that extreme weather poses to AOD predictions in such regions. Conversely, studies conducted in more temperate regions, such as Beijing, California and Washington, D. C., benefit from relatively stable meteorological conditions, which are less impacted by extreme weather events (Li et al., 2020; Taskinen et al., 2022; Zhang et al., 2022a). As a result, the models developed in these regions tend to exhibit improved performance metrics, characterized by higher R² (0.86–0.99) values and lower RMSE (0.001–0.105). Different preprocessing methods (e.g., interpolation, noise reduction) can also affect model performance. Preprocessed data may lead to artificially improved performance in some cases, making direct comparisons between studies using different methods more challenging.

3.3. Imputation and scenario analysis results

3.3.1. Daily imputation

A well-known limitation of satellite-derived AOD datasets is their vulnerability to observational gaps, primarily caused by cloud cover and snow (Li et al., 2021). These spatial deficiencies in AOD data pose challenges for applications in climate and epidemiological research (Li et al., 2020; Ma et al., 2022). Due to the inconsistent spatiotemporal resolution of multimodal data, it is necessary to unify the spatiotemporal resolution of the multimodal data for input into the residual network (AutoResnet). A significant portion of the original data consists of missing pixels (data availability <80%). To address this issue, we trained residual networks with 3 scenarios for January 30, 2019, and August 19, 2020 (When considering HCHO, SO₂, and NO₂ (HCHO+SO₂+NO₂)), spatial downscaling is performed using bilinear interpolation (Bilinear-AutoResnet); when considering HCHO + SO₂+NO₂, spatial downscaling is done using nearest-neighbor interpolation (Nearest-AutoResnet); and when considering HCHO + SO₂+NO₂, spatial downscaling is conducted based on a random forest downscaling algorithm combined with bilinear interpolation (RF-AutoResnet) (Fig. 8). Fig. 8 displays the gap-filled AOD results from Bilinear-AutoResnet, Nearest-AutoResnet, and RF-AutoResnet algorithms when considering HCHO + SO₂+NO₂. While all three scenarios produced reasonable AOD results, there are discernible differences among them (highlighted by the red rectangles in Fig. 8). For the original data on January 30, 2019 (Fig. 8A), both Bilinear-AutoResnet (Fig. 8C) and Nearest-AutoResnet (Fig. 8E) failed to capture the high AOD values present in LL city as seen in the original data. However, the RF-AutoResnet algorithm accurately replicated these high AOD values (Fig. 8G). Additionally, in the southern regions of BJ and XA, higher AOD values were predicted using Bilinear-AutoResnet (Fig. 8C) and Nearest-AutoResnet (Fig. 8E), whereas the RF-AutoResnet algorithm (Fig. 8G) predicted the opposite. For the original data on August 30, 2020, lower AOD values were predicted in the southern regions of BJ and XA using Bilinear-AutoResnet (Fig. 8D) and RF-AutoResnet (Fig. 8H), in contrast to the Nearest-AutoResnet algorithm, which predicted higher values (Fig. 8F). Thus, the AOD predicted by the RF-AutoResnet algorithm more accurately reflects the spatiotemporal heterogeneity of AOD in the actual atmosphere.

The progressive reduction in the loss function across increasing training iterations signifies the gradual convergence of the model, indicating an improvement in predictive accuracy. The increase in R² shows that the model's performance is gradually improving and can better explain the changes in the target variable. The rapid decrease in loss and the rapid increase in R² indicates that the model has better fitting capabilities. As shown in Fig. 9, all three models exhibit a trend of decreasing loss and increasing R² with the increase in epochs. Compared to the other two models, RF-AutoResnet shows a faster rate of loss reduction and R² increase. The RF-AutoResnet model demonstrates enhanced capability in discerning complex associations between input features and target variables. As a result, the model's predictions more closely align with the true values, implying higher accuracy and predictive power in estimating AOD at a fine scale.

In response to concerns regarding the RF-AutoResnet model's loss curve oscillations at higher epochs (Fig. 9), we extended the training from 100 to 500 epochs and found that the loss values and R² metrics stabilized after approximately 100 iterations, indicating convergence. We also experimented with different learning rates (LRs) to assess their impact on these oscillations. A lower LR slightly reduced oscillations but slowed convergence, while a higher LR led to significant instability and degraded model performance (see Text S4, Table S3, Fig. S1, Fig. S2, Fig. S3, and Fig. S4 of the Supplementary materials). The original LR of 0.0001 provided the best balance between convergence speed and stability across all three models (RF-AutoResnet, Bilinear-AutoResnet, and Nearest-AutoResnet), with minimal oscillations and the highest R² values. Additionally, we conducted multiple runs and reported the mean

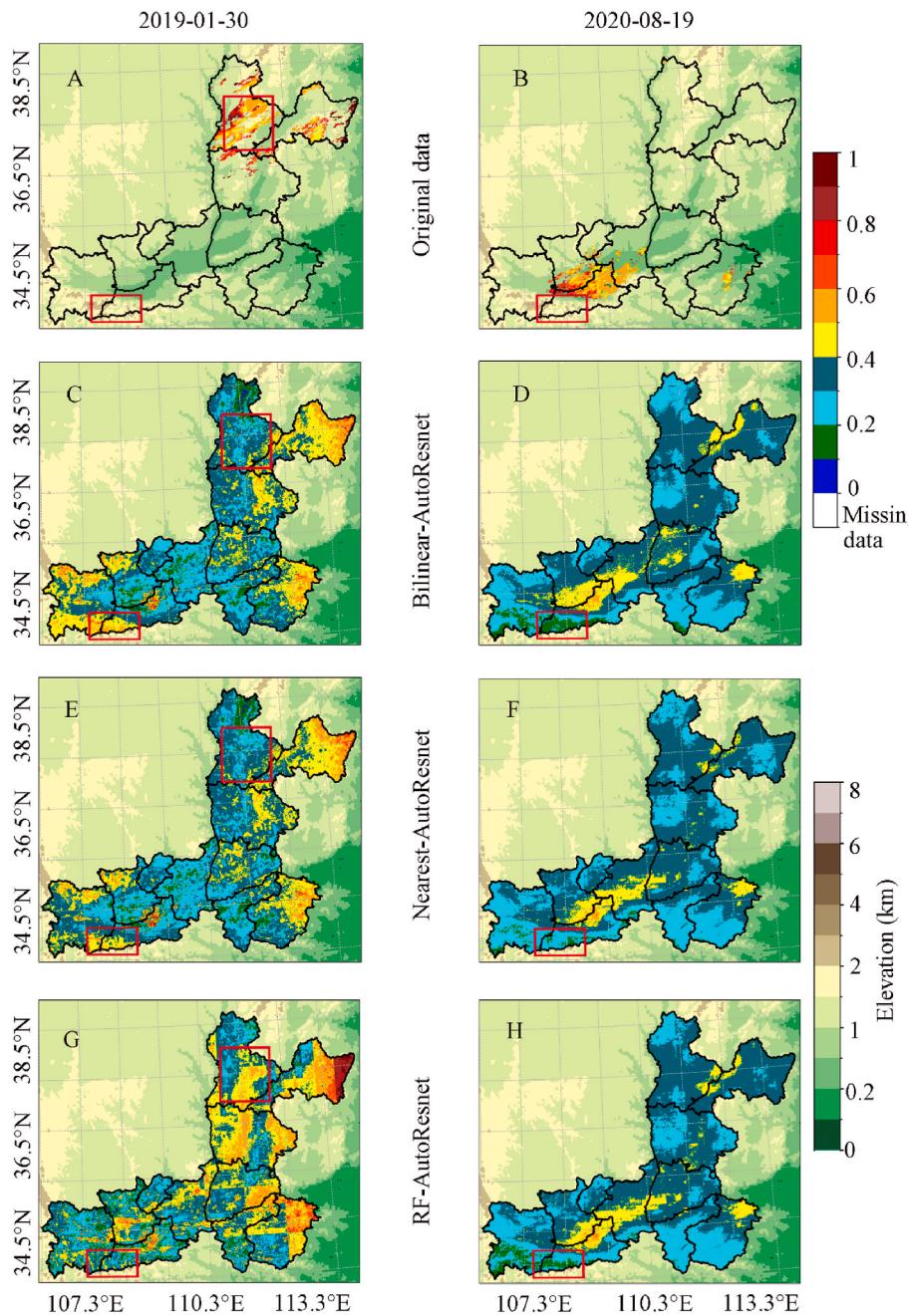


Fig. 8. Grid surfaces of the missing (A and B) and imputed MAIAC AOD for 2 days of different scenarios (C–H). Image with terrain in the background.

and variance of key metrics, confirming that the original LR produced the most stable results. These findings suggest that the original LR is optimal for ensuring both convergence and stability in the RF-AutoResnet model.

3.3.2. Monthly imputation and scenario analysis

From January 2019 to December 2022, as indicated in Table 3, a total of 1008 monthly residual networks were trained, each under different scenarios for every month. Fig. 10, Fig. S5, and Fig. S6 respectively present the gap-filled AOD results for different months of 2019 obtained by the RF-AutoResnet, Bilinear-AutoResnet, and Nearest-AutoResnet algorithms under various scenarios. Although all three scenarios yielded reasonably accurate AODs, there are clear differences among them (Red rectangles show the effect of different species combinations on gap-filled AOD in different scenarios). Overall, the RF-AutoResnet algorithm performed better than the others. Specifically,

for the original data of January 2019 (Fig. 10A, Fig. S5A, and Fig. S6A), the gap-filled AOD obtained by the RF-AutoResnet algorithm under different scenarios is closely matched original image. This is mainly reflected in the successful capture of high AOD values in the northern part of XA, WN, and the central part of YC and LF. In contrast, the Bilinear-AutoResnet (Fig. S5 D, G, J, M, P, S, and V) and the nearest-Neighbor-AutoResnet (Fig. S6 D, G, J, M, P, S, and V) underestimated the high AOD values present in the original data. This discrepancy mainly attributed to the failure of the Bilinear and Nearest interpolations to reflect the high AOD values when unifying the spatial resolution (Fig. 5D–I).

For the original data of April 2019 (Fig. 10B, Fig. S5B, and Fig. S6B), all three algorithms yielded relatively good gap-filled AODs under different scenarios. However, subtle differences were observed between the results obtained by different algorithms. For instance, compared to other algorithms, the RF-AutoResnet predicted more accurate gap-filled

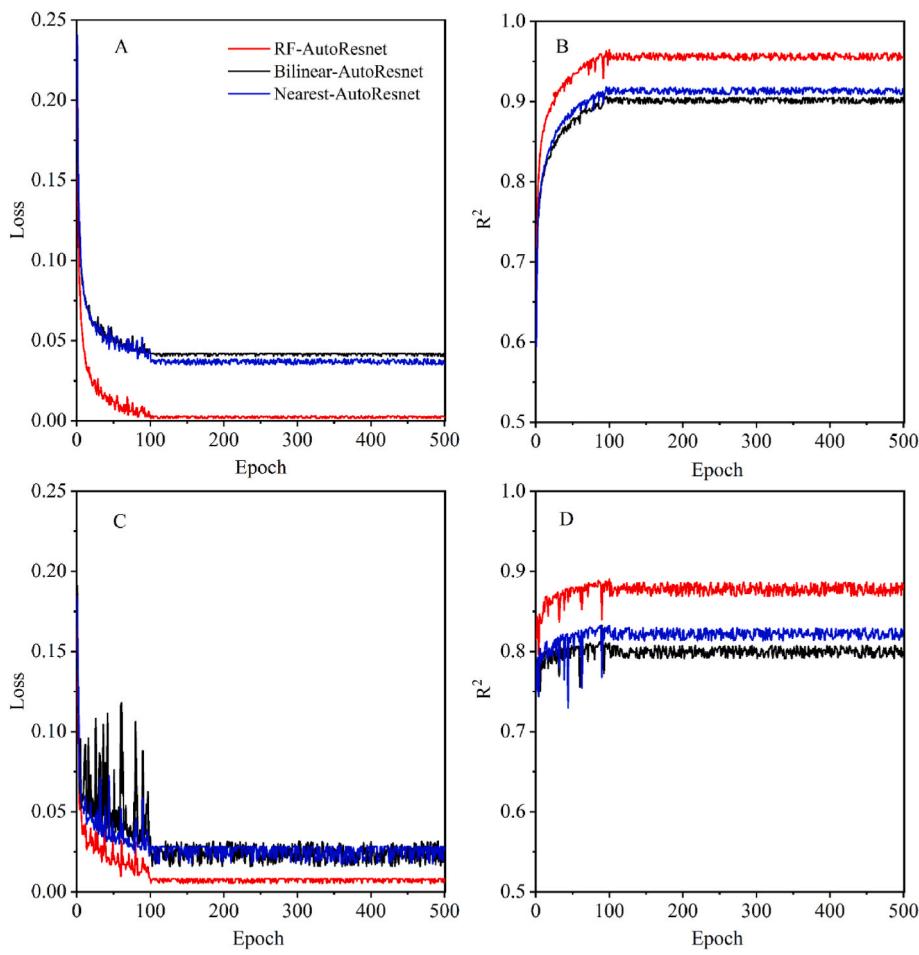


Fig. 9. Trends of loss (A and C) and R^2 (B and D) for the test samples with an increase of the epoch for two days: 2019-01-30 (A and B) and 2020-08-19 (C and D).

AODs under the HCHO + SO₂+NO₂ scenario, mainly reflected in the successful capture of high AOD values in XA and WN (Fig. 10N), which is consistent with the original image (Fig. 10B, Fig. S5B, and Fig. S6B). Similarly, the RF-AutoResnet, Bilinear-AutoResnet, and Nearest-AutoResnet algorithms all demonstrated sensitivity to the NO₂ scenario, with each successfully capturing high AOD values in XA (Fig. 10Q). The gap-filled AODs for October 2019 obtained by the three algorithms under different scenarios were significantly higher than those for April, especially in XA, WN, and SMX. Overall, the AOD predicted by the RF-AutoResnet under the HCHO+SO₂+NO₂ scenario better reflects the spatiotemporal heterogeneity of AOD in the actual atmosphere.

Through cloud and snow detection, MAIAC AOD provides quality assurance indicators to identify invalid or highly uncertain AOD values. Despite these quality assurance indicators ensuring high-quality MAIAC AOD, previous studies have indicated that in areas with high cloud and snow coverage, satellite-derived AOD is occasionally overestimated compared to AERONET AOD (Li et al., 2015). Our predictive model also observed this overestimation phenomenon in high-altitude areas where AOD is missing (Fig. 10O, southern XA region). The FWP region is characterized by complex terrain, variable elevation, and multifaceted aerosol emission sources, including vehicular emissions, aeolian dust, and photochemical processes. This intricate interplay of factors, combined with the region's heterogeneous meteorological conditions, engenders substantial spatiotemporal variability in aerosol distributions. In addition, the FWP region's energy structure is predominantly coal-based, transportation relies heavily on highways, and industrial pollution emissions are large. Affected by the blocking of special terrain mountains and the sinking of leeward slope airflow, the FWP region is

prone to form an anticyclonic airflow stagnation area. Periods of heavy pollution, pollutants are easily trapped in the Fen and Wei river valleys and are not easily dispersed. Furthermore, clouds, snow, or high surface reflectance are the primary causes of AOD missing. The FWP area has more clouds, rain, and snow, which leads to large absences of AOD and an overestimation of AOD.

Fig. 11 presents scatter plots of AOD with vertical column density (VCD) of different tropospheric species and their combinations over the FWP region. The slope differences among different combinations provide indirect information regarding the aerosol composition and potential emission sources. In all satellite-observed scenarios shown in Fig. 11, a strong correlation is observed between AOD and tropospheric NO₂ VCD, indicating that sources emitting NO_x drive AOD (Fig. 11M-O). Moreover, nitrate, as an important component of PM_{2.5}, is a key contributor to haze formation (Wei et al., 2022). Nitrate is formed by the oxidation of NO_x, which is mainly emitted from combustion sources such as vehicles, industries, and power plants. NO emitted during the day mainly reacts with O₃ to form NO₂. O₂ is oxidized by OH radicals to form nitrate (Peng et al., 2021). In environments with abundant ammonia, the gas-phase nitric acid readily partitions into the particulate phase through neutralization reactions, forming ammonium nitrate aerosols (Zhang et al., 2015). Therefore, the VCD NO₂ is particularly sensitive to AOD interpolation.

SO₂ mainly comes from the combustion of fossil fuels. The relatively low slope observed between AOD and SO₂ (Fig. 11 P-R) can be attributed to the installation of desulfurization equipment at many power plants in the FWP region, which helps to reduce SO₂ emissions. In addition, the LY region is a gathering area for coal-fired power plants with a capacity ≥ 300 MW. However, as shown in Fig. 10 S and U, the model fails to

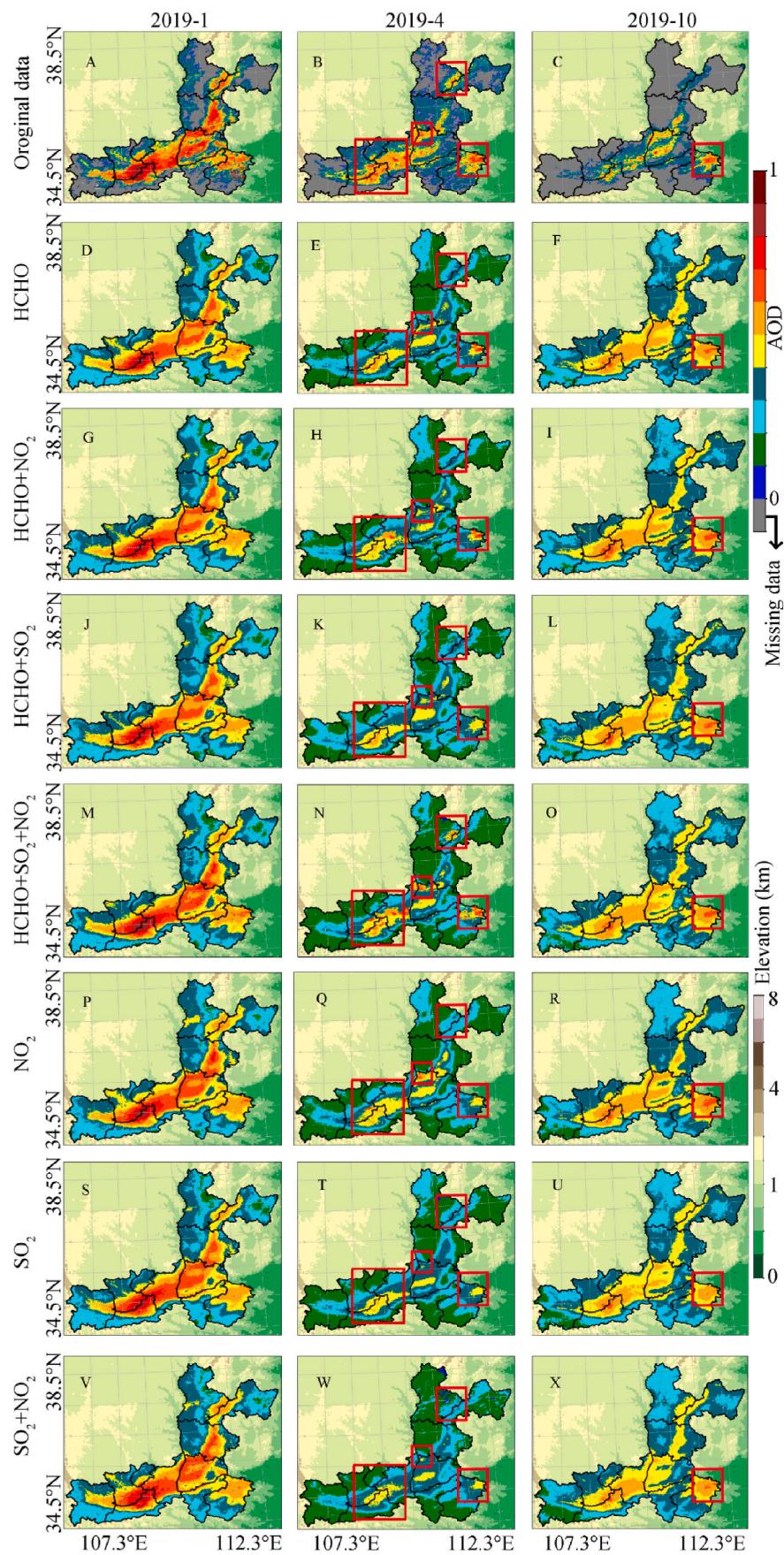


Fig. 10. Grid surfaces of the missing (A, B, and C) and gap-filled MAIAC AOD generated by RF-AutoResnet algorithm with different scenarios (D–X).

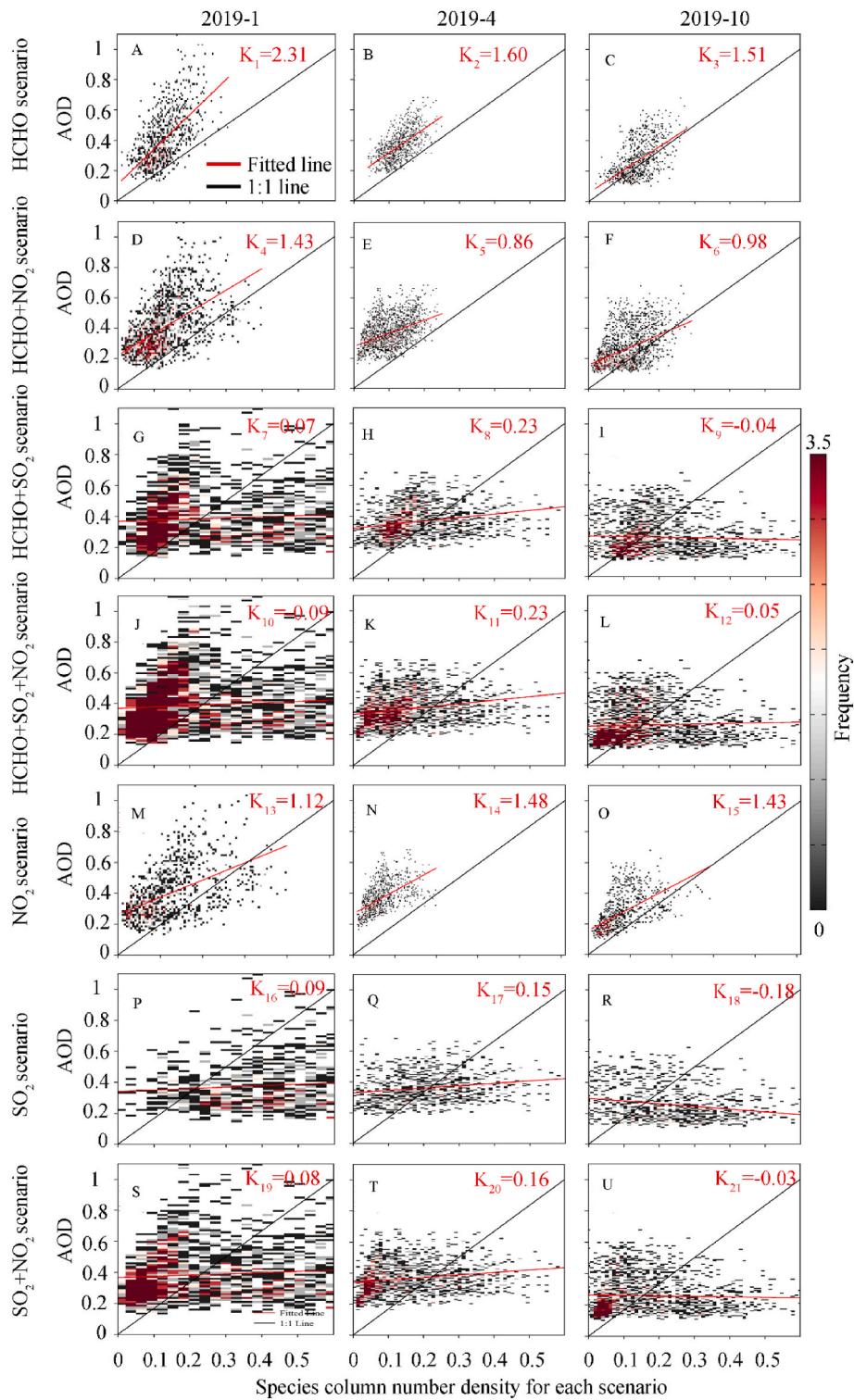


Fig. 11. Scatter plots between the AOD and specious vertical column density vertical column density (mol/m^2) with different scenarios. For example, 11A denotes a scatter plot between the AOT and HCHO vertical column density.

predict high AOD values in the LY region, indicating that the vertical column concentration of SO₂ is not sensitive to AOD interpolation.

While HCHO serves as a proxy for NMVOC emissions, is associated with a complex, non-linear relationship to SOA formation (Veefkind et al., 2011). The observed association between HCHO VCD and AOD measurements over the FWP region suggests a significant contribution of SOA formation to the overall particulate matter burden (Fig. 11A–C). NO₂ and SO₂ are predominantly emitted from anthropogenic activities,

particularly industrial processes, whereas HCHO exhibits a significant biogenic origin. Although the lifetime of formaldehyde is short, the complex chemical process that forms it may require more time, which means that its concentration may be higher downwind of emission sources and exhibit a better correlation with longer-lived aerosol particles (Veefkind et al., 2011). Furthermore, NMVOCs enhance the atmospheric oxidation capacity of FWP, promoting the oxidation process of NO_x to nitric acid in this region (Xu et al., 2023). Due to the reduction

of SO_2 emissions, the increase of ammonia further converts gaseous HNO_3 into particulate nitrates, thereby significantly increasing the concentration of nitrates and $\text{PM}_{2.5}$ in this region (Liu et al., 2018). The atmospheric SOx is oxidized to sulfuric acid, which then reacts with ammonia to form ammonium sulfate particles. When the emission of sulfur dioxide decreases, the concentration of sulfate in the atmosphere also decreases accordingly. This makes the ammonia that originally reacted with sulfuric acid become excessive, because there is not enough sulfuric acid to react with ammonia to form ammonium sulfate (Liu et al., 2019; Liu et al., 2018). Therefore, we observe that $\text{HCHO} + \text{NO}_2$ combination is sensitive to gap-filled AOD. Although the reduction of SO_2 emissions leads to a decrease in sulfate aerosols, we find that $\text{HCHO} + \text{SO}_2 + \text{NO}_2$ scenarios obtained gap-filled AOD close to the original image. This is mainly manifested in Fig. 10N, where the RF-AutoResnet model predicted high values of AOD in the central part of WN and the eastern regions of LY and LL.

In terms of model training parameters, the RF-AutoResnet consistently demonstrated better convergence and overall performance compared to the Bilinear-AutoResnet and Nearest-AutoResnet algorithms (Figss. 12–14 and Figs. S7–S24). For January with the $\text{HCHO} + \text{SO}_2 + \text{NO}_2$ scenario, the RF-AutoResnet achieved a higher mean test R and R^2 (0.98, 0.96), and lower mean test RMSE and MAE (0.03, 0.01) than Bilinear-AutoResnet (R , R^2 , RMSE, and MAE are 0.94, 0.87, 0.04 and 0.03 respectively) and Nearest-AutoResnet (R , R^2 , RMSE and MAE are 0.93, 0.85, 0.05 and 0.03 respectively) (Figss. 12–14). For October under the $\text{HCHO} + \text{SO}_2 + \text{NO}_2$ scenario, the RF-AutoResnet had a higher mean test R^2 (0.92) and lower mean test RMSE (0.02) than Bilinear-AutoResnet (R^2 and RMSE are 0.83 and 0.04 respectively) and Nearest-AutoResnet R^2 and RMSE are 0.8 and 0.05 respectively) (Fig. 12).

For August with the $\text{HCHO} + \text{SO}_2 + \text{NO}_2$ scenario, the RF-AutoResnet improved test mean R^2 (0.86) by 19.4% relative to the Bilinear-AutoResnet (0.72) and by 30.3% over the Nearest-AutoResnet (0.66) (Figss. 12–14). There are varying degrees of missing MAIC AOD observations in the three sample months, with varying spatial patterns of missing data. The surface of observed and estimated for these three sample months shows that AutoResnet can provide robust imputation results despite a significant proportion of missing data (>50%) (Fig. 10).

The analysis of the RF-AutoResnet model performance in the FWP reveals that May exhibits the highest MAE, followed by January, August, September, and March. Understanding the contributing factors to these discrepancies can provide valuable insights into the model's limitations and areas for improvement.

The elevated MAE in May is primarily attributed to seasonal transitions and increased photochemical activity. During this period, higher temperatures and increased solar radiation promote the formation of secondary aerosols such as sulfate and nitrate, leading to greater AOD changes. While the RF-AutoResnet model effectively handles precursor concentrations, it may encounter challenges in accurately capturing the rapid photochemical reactions that drive aerosol formation.

Variability in meteorological conditions plays a significant role in influencing aerosol loadings and, consequently, AOD predictions. For instance, in January, the FWP experiences frequent temperature inversions, which restrict vertical mixing and trap pollutants near the surface. This leads to elevated levels of $\text{PM}_{2.5}$, complicating accurate AOD retrievals.

Conversely, during August and September, increased precipitation and higher humidity levels enhance aerosol removal processes through

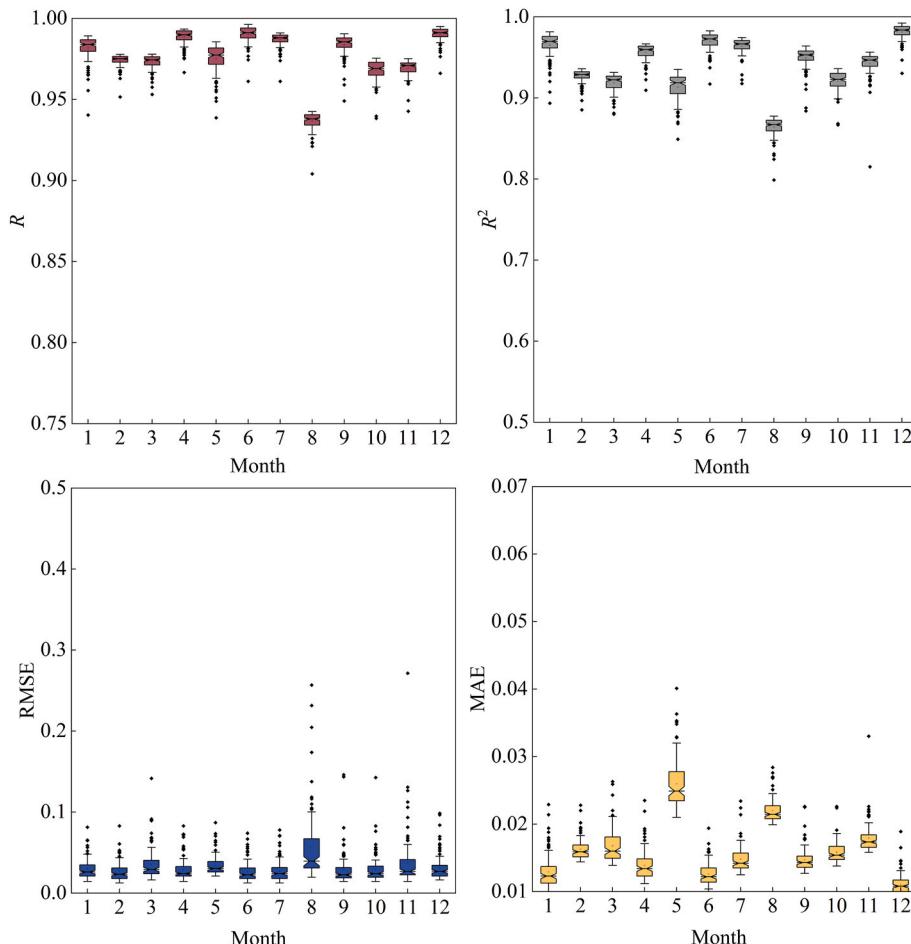


Fig. 12. Leave-one-county cross test (R , R^2 , RMSE and MAE) for RF-AutoResnet imputation with $\text{HCHO} + \text{SO}_2 + \text{NO}_2$ scenario during 2019

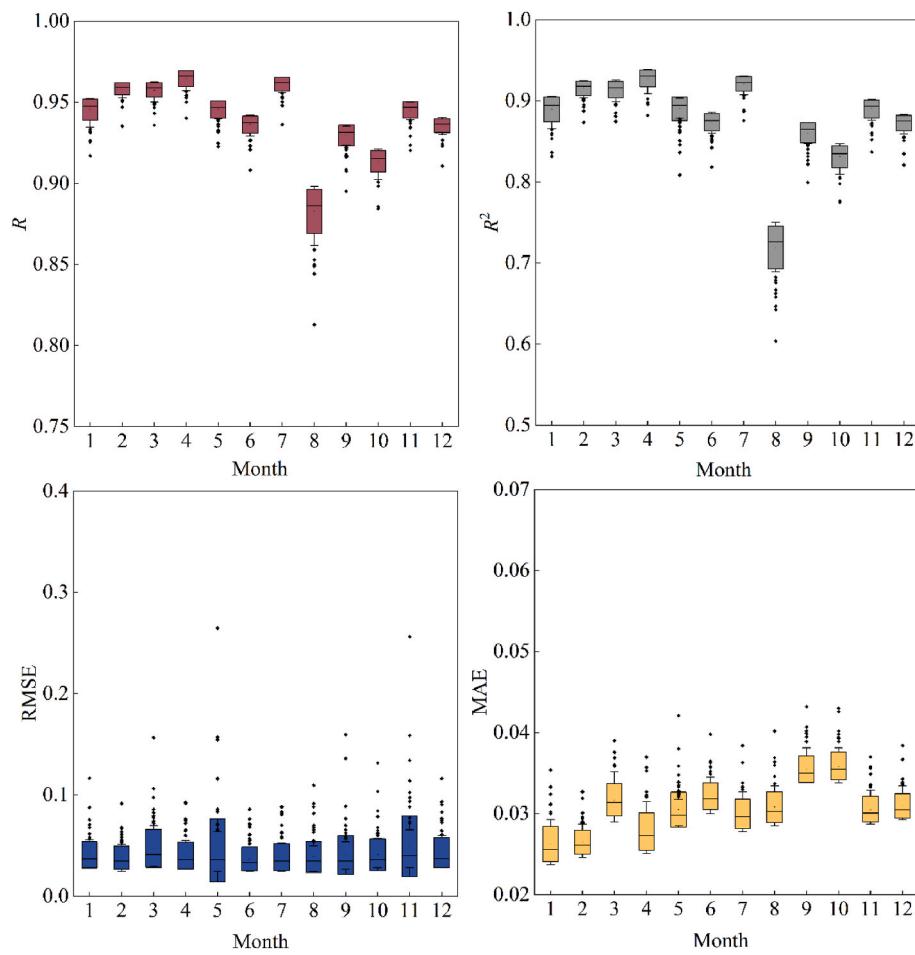


Fig. 13. Leave-one-county cross test (R , R^2 , RMSE and MAE) for Bilinear-AutoResnet imputation with $\text{HCHO} + \text{SO}_2 + \text{NO}_2$ scenario during 2019

wet deposition. The anomalously strong Northwestern Pacific anti-cyclone, Bohai Sea high-pressure system, and Balkhash Lake trough are key circulation systems influencing autumn rainfall in Shaanxi Province. A robust and persistent Northwestern Pacific anomalous anticyclone, directs with the Bohai Sea high-pressure system, jointly steers moisture transport from the Western Pacific toward western China. Meanwhile, the intensification of the Balkhash Lake trough facilitates the southward movement of cold air masses, driving their convergence with warm, moist air over Shaanxi. This dynamic interplay results in prolonged autumn rainfall and an increased frequency of precipitation events (Na et al., 2023). The RF-AutoResnet model relies on sufficient high-quality input data for accurate imputation. However, the reduced availability of AOD observations in August and September, primarily caused by precipitation events, introduces uncertainties into the gap-filling process.

In March, the FWP is occasionally impacted by dust storms originating from desert areas to the northwest. These storms introduce large quantities of coarse particulate matter (PM_{10}), which differ from the $\text{PM}_{2.5}$ typically modeled by the RF-AutoResnet framework. Dust particles, due to their size and scattering properties, contribute to high AOD values. However, the model's emphasis on $\text{PM}_{2.5}$ precursors such as SO_2 , NO_2 , and HCHO does not fully account for the influence of coarse aerosols during dust events. This mismatch between the aerosol types modeled and the aerosols present during dust storms can lead to higher imputation errors.

Another contributing factor to higher errors during January is the severe haze pollution over the Guanzhong region. During this period, elevated aerosol concentrations often cause AOD to be misclassified as cloud by satellite sensors, leading to retrieval inaccuracies. Additionally, the high surface brightness in January can also influence the accuracy of

AOD retrievals. The RF-AutoResnet model depends on sufficient high-quality input data for accurate imputation, and these misclassifications and surface-related issues introduce uncertainties into the gap-filling process, further complicating the AOD retrieval under these conditions.

As shown in Figs. S25–27, all three models exhibit a trend of decreasing loss as the number of epochs increase. However, RF-AutoResnet demonstrates a more rapid decline in loss compared to the other two models, indicating higher accuracy and predictive power in estimating the missing MAIAC AOD. This implies that the RF-AutoResnet model may be more effective in filling in missing AOD data, thereby enhancing the ability to assess the impact of $\text{PM}_{2.5}$ on human health at a finer scale and improving the simulation and predictive performance of climate models. Furthermore, sensitivity experiments conducted for AOD imputation in 2021 using the $\text{HCHO} + \text{SO}_2 + \text{NO}_2$ scenario (Fig. S28) revealed high AOD values in LL during April, likely due to natural factors such as dust storms. The growth of aerosol hygroscopicity due to high precipitation in summer led to high AOD values in the northern part of XA. Moreover, under conditions of strong solar radiation, sulfuric acid generated in the gas phase can nucleate to form new nanoparticles or react with NH_3 to produce ammonium sulfate, significantly influencing AOD (Seinfeld and Pandis, 2016). The photolysis and gas-phase oxidation of VOCs play key roles in SOA formation, with organic compounds forming new aerosols through condensation or nucleation processes, thereby contributing to higher AOD values (Schervish and Donahue, 2020).

One of the key distinctions between the RF-AutoResnet model and the Bilinear-AutoResnet and Nearest-AutoResnet models lies in the variables utilized during the downscaling process. The RF-D model not

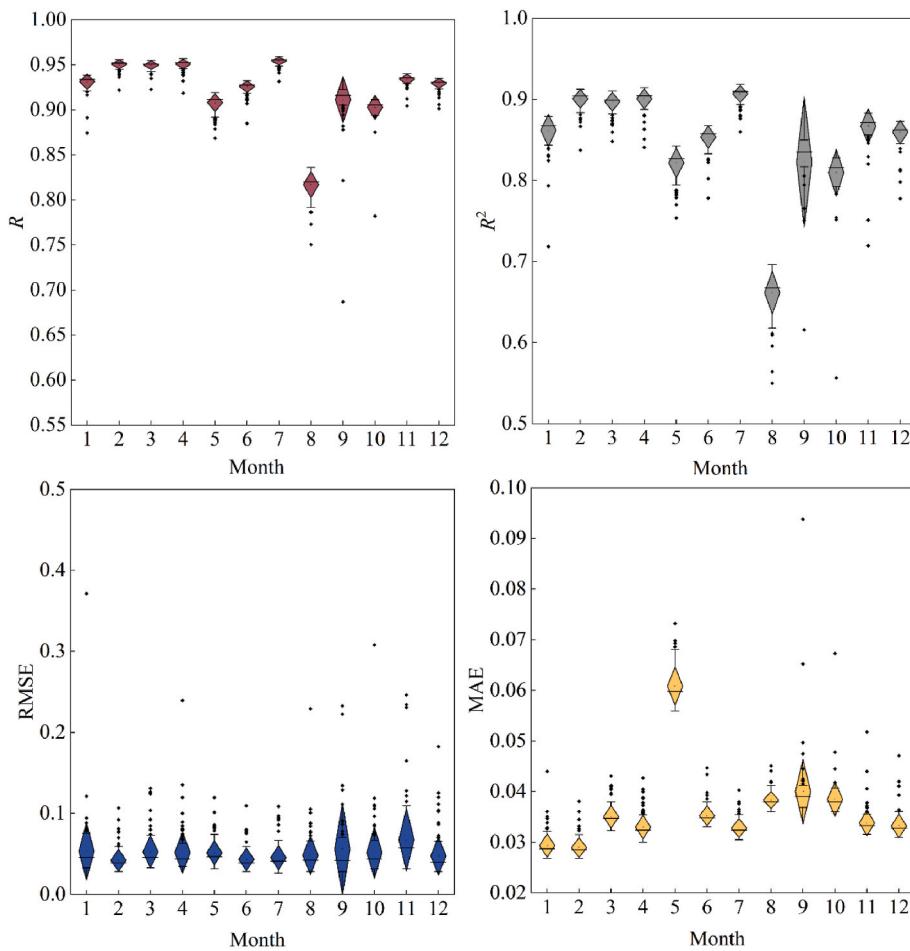


Fig. 14. Leave-one-county cross test (R , R^2 , RMSE and MAE) for Nearest-AutoResnet imputation with $\text{HCHO} + \text{SO}_2 + \text{NO}_2$ scenario during 2019.

only downscales MERRA-2 AOD to 1 km resolution but also captures the non-linear relationships between AOD and multiple covariates. By incorporating variables such as NDVI, DEM, NO_2 , and population density, the RF-D model better represents the spatial heterogeneity of aerosols is accurately represented, especially in regions characterized by diverse surface types and complex terrain. The RF algorithm's ability to process large datasets and integrate multiple variables provides it a distinct advantage over interpolation-based methods, which often oversimplify spatial variations.

While Bilinear-AutoResnet and Nearest-AutoResnet models rely on interpolation techniques to unify the spatial resolution of input variables, they do not incorporate covariates or account for the inherent spatiotemporal heterogeneity of MERRA-2 AOD. These interpolation-based methods are based on the assumption of smooth spatial transitions between data points, which is not suitable for regions like the FWP, where sharp gradients in aerosol concentrations exist due to variations in terrain, vegetation, and industrial emissions.

3.4. Validation with SONET data

The validation dataset comprised hourly measurements collected from a SONET site over the 2019–2020 period. These high-temporal resolution observations were aggregated into monthly means to evaluate both the original and reconstructed MAIAC AOD products. Scenario analysis indicated that $\text{HCHO} + \text{SO}_2 + \text{NO}_2$ scenario exhibited higher sensitivity to AOD interpolation. Therefore, we selected the $\text{HCHO} + \text{SO}_2 + \text{NO}_2$ scenario, based on the RF-Autorenset algorithm, as the final result for comparison with the SONET site observations.

The monthly evaluation results are shown in Fig. 15. For the original

AOD data compared with the SONET AOD at coincident pixel-point estimates, the monthly evaluation metrics were $R=0.83$, $R^2=0.69$, RMSE =0.08, and Bias=0.01 (Fig. 15A). The presence of missing values, primarily caused by cloud cover, contributed to lower correlations and higher RMSE values. After applying the downscaling approach integrated with the AutoResnet algorithm, the gap-filled AOD demonstrated improved agreement with in-situ SONET AOD observations at coincident pixel points. The metrics for the reconstructed AOD were $R=0.91$, $R^2=0.82$, RMSE=0.06, and Bias=0.01 (Fig. 15B). Compared to the original data, the R^2 value increased by 15.8%, highlighting the enhanced reliability of the RF-Autorenset approach. This robust model performance can be attributed to the incorporation of auxiliary data and the capability of the RF-AutoResnet method to effectively capture nonlinear relationships. These advantages make the method suitable for regional-scale applications across diverse landscapes and surface types.

The comparison of monthly averages in this study primarily aims to evaluate the effectiveness of the interpolation method when applied to incomplete data. Monthly average AOD values derived via RF-AutoResnet interpolation demonstrate a significantly enhanced alignment with ground-based observations compared to the original monthly averages. This improvement suggests that the original satellite aerosol products are influenced by factors such as cloud contamination and high surface reflectance, which introduce deviations in climate statistics relative to ground-based remote sensing. The interpolation approach reduces these discrepancies, yielding climate metrics that better correspond with ground observations, thereby providing more robust data for accurately characterizing long-term aerosol trends and their climatic impacts.

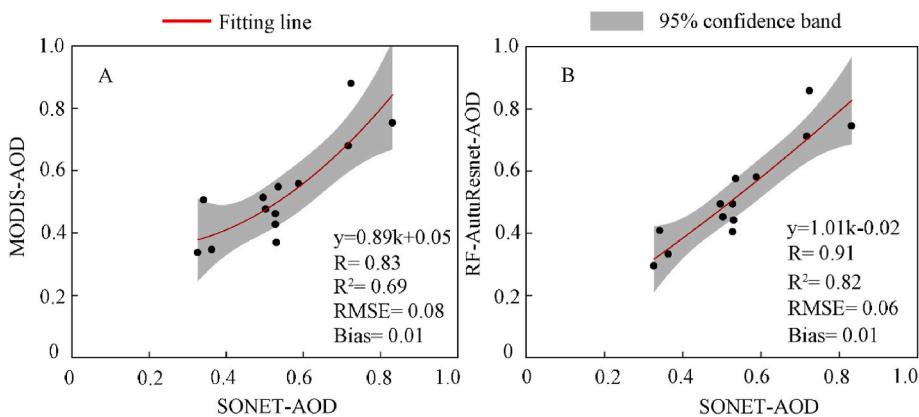


Fig. 15. Scatter plots between the original and imputed monthly average AOD generated using the RF-Autorsenet algorithm against the in-situ monthly average AOD from SONET sites.

3.5. Comparison of metrics with previous AOD interpolation studies

In the exploration of AOD interpolation, we discovered that the RF-AutoRsenet algorithm yielded superior results under the HCHO + SO₂+NO₂ scenario. Consequently, this study employed the RF-AutoRsenet algorithm within the aforementioned scenario and conducted a comprehensive evaluation of its performance metrics (Fig. 16). By comparing these results with existing research, we not only validated the model's effectiveness but also demonstrated its superiority across several key indicators. This section will elaborate on the comparative analysis between the RF-AutoRsenet algorithm and previous research findings, particularly emphasizing advancements in AOD interpolation (Fig. 16). Li et al. (2021) and Guo et al. (2023) applied the RF method to derive global imputed AOD, where the non-random missing rate of global AOD was substantially higher than that in the FWP, culminating in an R² inferior to our findings. Jiang et al. (2021), utilizing a two-stage RF model, ascertained the full-coverage AOD spatial distribution across China, corroborating the outcomes. The absence of downscaling correction for the model AOD resulted in a lower R compared to our results. Zhang et al. (2022b) introduced a spatio-temporal fitting algorithm, achieving an R of 0.79, which outperformed Interpolation-based correlation weighting (ICW) at 0.66 and IDW at 0.67. This enhancement is attributed to the algorithm's consideration of the spatio-temporal dynamics between adjacent pixels, unlike the IDW algorithm, which only accounts for spatial correlations. Li et al. (2020) innovated with

ResNet to produce gap-filled AOD, comparing its performance against traditional neural networks and GAM models. Their findings indicated that ResNet more accurately captured the spatio-temporal heterogeneity of AOD, evidenced by an R² of 0.98. Nonetheless, our R² and RMSE did not reach the benchmark set by Li et al. (2020). This discrepancy likely arises from the geographical differences, with California's predominantly plain terrain contrasting with the FWP's encirclement by the Qinling Mountains and the Loess Plateau, coupled with extensive cloud cover, which amplifies predictive errors. Ke et al. (2022) pioneered multimodal data fusion for full-coverage AOD production, addressing inaccuracies in model AOD due to emission inventory uncertainties, marking a significant leap in the domain. Despite the superior R yielded by our RF-Autorsenet algorithm compared to Ke et al. (2022), the limited availability in-situ AOD measurements over the FWP (a single SONET site) may introduce a bias in our results. Similarly, Zeng et al., (2023) applied a Light Gradient Boosting Machine (Light-GBM) to generate daily full-coverage AOD in the North China Plain, achieving an R² of 0.90. This result is similar to ours, potentially due to the model's exclusion of coarse-resolution data such as model AOD. Moreover, it mitigated the errors associated with the CTM simulations induced by the emission inventory, thereby reducing the error influx into Light-GBM. Goldberg et al. (2019) employed an LM model to reconstruct the gapless AOD in the eastern United States. Their gap-filled product, compared to AERONET data, yielded an R² of 0.52. This highlights the limitations of linear models in capturing the complex nonlinear relationships between AOD and atmospheric variables, potentially diminishing predictive accuracy. Additionally, Xiao et al. (2017) and Lv et al. (2016) employed Kriging interpolation to derive imputed AOD values, yielding R of 0.44 and 0.18, respectively. These results are substantially lower than those reported in our study. These values are significantly lower than those reported in our study, emphasizing the need for more advanced models that can better capture the nonlinear dynamics inherent in atmospheric data to improve predictive performance. Collectively, these studies underscore the formidable potential of machine learning within the environmental science sphere.

3.6. Limitations

The dataset for comparing the original AOD with SONET AOD is limited, which is likely because conditions preventing MAIAC from generating valid observations often coincide with those that prevent SONET from measuring AOD (e.g., cloudy weather). The scenarios where MAIAC data is missing but SONET data is available are frequent, resulting in a smaller sample size for validation and potential uncertainty in the validation metrics. Additionally, the limited number of in-situ and samples for comparing the imputed AOD with SONET AOD also contributes to the uncertainty of these validation metrics. While this

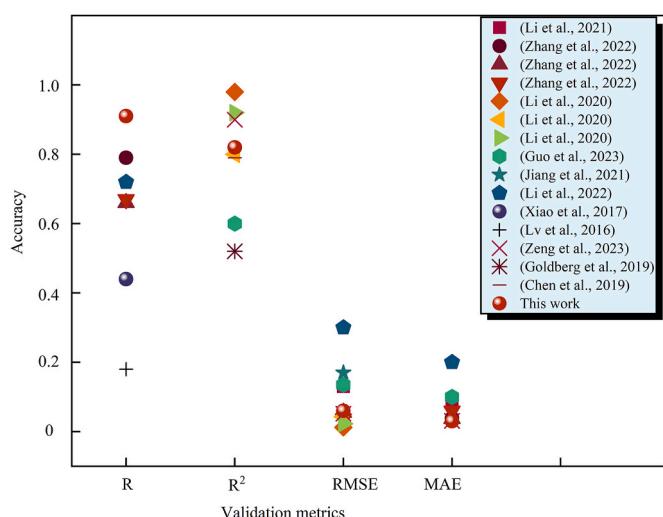


Fig. 16. Scatter plots comparing the imputed AOD generated using the RF-Autorsenet algorithm with AOD from previous imputation studies.

study employed daily precursor data for daily imputation and monthly averages for monthly imputation, we acknowledge that using a moving average approach for the precursors could offer additional benefits. A moving average would smooth short-term fluctuations and may enhance the accuracy of the imputation, especially for daily imputation tasks. The model is prone to overfitting, especially with complex datasets and high-dimensional inputs, despite applying techniques like bootstrapping, regularization, and cross-validation. Future work should consider feature selection and tree pruning to further mitigate this risk. Additionally, the model's computational complexity presents a challenge for large-scale applications, as processing a single image (1337×1114 pixels) can take up to 4.5 hours on high-performance hardware, highlighting the need for both hardware and code optimizations. Lastly, higher MAE during months with extreme meteorological events (e.g., dust storms and monsoons) suggests sensitivity to seasonal variations, which increases computational demands. Future research should focus on improving the model's efficiency in handling such variability without sacrificing accuracy.

4. Conclusion

Satellite-based AOD data often has limited coverage. This paper develops a robust estimation method, RF-AutoResnet, for missing MAIAC AOD by combining an RF model with AutoResnet. The RF model downscale MERRA-2 AOD simulations and processes key input variables according to the resolution of the target variable. Compared to the Bilinear and Nearest algorithms, the RF downscaling model achieved a higher R (0.71) and lower RMSE (0.10), significantly improving the spatial representativeness of the MERRA-2 AOD downscaling estimates. Scenario analysis revealed that in the HCHO+SO₂+NO₂ scenario, the RF-AutoResnet model captured the spatiotemporal heterogeneity of missing daily and monthly average MAIAC AOD at fine spatial scales and improved the prediction performance. The RF-AutoResnet method outperformed Bilinear-AutoResnet and Nearest-AutoResnet with higher average R² (0.87–0.90) and lower RMSE and MAE (0.02–0.03, 0.01–0.04). Furthermore, the performance metrics of RF-AutoResnet surpassed most previous estimation studies. Validation against data from SONET sites showed a strong correlation (R = 0.91), compared to the original MAIAC AOD, demonstrating the reliability of our estimation method. Importantly, this research can reduce measurement errors in improving PM_{2.5} exposure estimates, ensuring more reliable assessments of PM_{2.5}-related health impacts.

CRediT authorship contribution statement

Abula Tuheti: Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Zipeng Dong:** Writing – review & editing, Validation, Supervision, Resources, Methodology, Funding acquisition, Conceptualization. **Guanghua Li:** Visualization, Software. **Shunxi Deng:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition. **Zhengqiang Li:** Validation, Supervision, Data curation. **Li Li:** Resources, Investigation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank NASA for providing aerosol, aerosol fraction, fire point, NDVI, nighttime light, terrain, and cloud fraction data. We thank ESA for providing the Sentinel-5P data. We thank ERA5 for providing meteorological data. We thank NOAA for providing the RTMA data. We also

thank the Institute of Earth Environment, Chinese Academy of Sciences provided SONET data to evaluate our developed model.

This work is supported by National Natural Science Foundation of China (41905016), the Key Research and Development Program of Shaanxi (2024SF-YBXM-583, 2021ZDSL-05-07) and the National Key Research and Development Program of China (2017YFC0212206).

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.atmosenv.2024.120989>.

Data availability

Data and data downloads are in the supplementary material

References

- Aaron, V.D., Martin, Randall V., Levy, Robert C., da Silva, Arlindo M., Michal, K., Chubarova, Natalia E., Eugenia, S., Cohen, Aaron J., 2011. Satellite-based estimates of ground-level fine particulate matter during extreme events: a case study of the Moscow fires in 2010. *Atmos. Environ.* 45 (34), 6225–6232.
- Bao, F., Huang, K., Wu, S., 2023. The retrieval of aerosol optical properties based on a random forest machine learning approach: exploration of geostationary satellite images. *Remote Sens. Environ.* 286, 113426–113448.
- Cao, J.J., Cui, L., 2021. Current status, characteristics and causes of particulate air pollution in the Fenwei Plain, China: a review. *J. Geophys. Res. Atmos.* 126 (11), e2020JD034472.
- Chen, Z., Zhang, T., Zhang, R., Zhu, Z., Yang, J., Chen, P., Ou, C., Guo, Y., 2019. Extreme gradient boosting model to estimate PM_{2.5} concentrations with missing-filled satellite data in China. *Atmos. Environ.* 202, 180–189.
- Di, Q., Kloog, I., Koutrakis, P., Lyapustin, A., Wang, Y., Schwartz, J., 2016. Assessing PM_{2.5} exposures with high spatiotemporal resolution across the continental United States. *Environ. Sci. Technol.* 50 (9), 4712–4721.
- Dovrov, E., Lim, C.Y., Canagaratna, M.R., Kroll, J.H., Worsnop, D.R., Keutsch, F.N., 2019. Measurement techniques for identifying and quantifying hydroxymethanesulfonate (HMS) in an aqueous matrix and particulate matter using aerosol mass spectrometry and ion chromatography. *Atmos. Meas. Tech.* 12 (10), 5303–5315.
- Garay, M.J., Kalashnikova, O.V., Bull, M.A., 2016. Development and assessment of aHighSpatial resolution (4.4 km)MISR aerosol product using AERONET-DRAGON data. *Atmos. Chem. Phys.* 569, 569–595.
- Goldberg, D.L., Gupta, P., Wang, K., Jena, C., Zhang, Y., Lu, Z., Streets, D.G., 2019. Using gap-filled MAIAC AOD and WRF-Chem to estimate daily PM_{2.5} concentrations at 1 km resolution in the Eastern United States. *Atmos. Environ.* 199, 443–452.
- Guo, B., Wang, Z., Pei, L., Zhu, X., Chen, Q., Wu, H., Zhang, W., Zhang, D., 2023. Reconstructing MODIS aerosol optical depth and exploring dynamic and influential factors of AOD via random forest at the global scale. *Atmos. Environ.* 315, 120159–120172.
- He, K., Zhang, X., Ren, S., Sun, J., 2016. Identity mappings in deep residual networks. *Pt Iv* 9908, 630–645.
- Hu, Z., Jin, Q., Ma, Y., Pu, B., Ji, Z., Wang, Y., Dong, W., 2021. Temporal evolution of aerosols and their extreme events in polluted Asian regions during Terra's 20-year observations. *Remote Sens. Environ.* 263, 112541–112560.
- Jiang, T., Chen, B., Nie, Z., Ren, Z., Xu, B., Tang, S., 2021. Estimation of hourly full-coverage PM_{2.5} concentrations at 1-km resolution in China using a two-stage random forest model. *Atmos. Res.* 248, 105146–105158.
- Jonathan, M., Eleni, D., Loretta J., M., Frank N., K., Yuan, C., Daniel J., J., Jingkun, J., Meng, L., William, M., Xiaohui, Q., Qiang, Z., 2018. Contribution of hydroxymethane sulfonate to ambient particulate matter: a potential explanation for high particulate sulfur during severe winter haze in Beijing. *Geophys. Res. Lett.* 45 (21), 11969–11979.
- Just, A.C., Wright, R.O., Schwartz, J., Coull, B.A., Baccarelli, A.A., Tellez-Rojo, M.M., Moody, E., Wang, Y., Lyapustin, A., Kloog, I., 2015. Using high-resolution satellite aerosol optical depth to estimate daily PM_{2.5} geographical distribution in Mexico city. *Environ. Sci. Technol.* 49 (14), 8576–8584.
- Ke, L., Kaixu, B., Zhengqiang, L., Jianping, G., Ni-Bin, C., 2022. Synergistic data fusion of multimodal AOD and air quality data for near real-time full coverage air pollution assessment. *J. Environ. Manag.* 302, PtB, 114121–111435.
- Lee, H.J., 2019. Benefits of high resolution PM_{2.5} prediction using satellite MAIAC AOD and land use regression for exposure assessment: California examples. *Environ. Sci. Technol.* 53 (21), 12774–12783.
- Li, J., Carlson, B.E., Lacis, A.A., 2015. How well do satellite AOD observations represent the spatial and temporal variability of PM_{2.5} concentration for the United States? *Atmos. Environ.* 102, 260–273.
- Li, L., Franklin, M., Girguis, M., Lurmann, F., Wu, J., Pavlovic, N., Breton, C., Gilliland, F., Habre, R., 2020. Spatiotemporal imputation of MAIAC AOD using deep learning with downscaling. *Remote Sens. Environ.* 237 (237), 111584–111601.
- Li, X., Liu, K., Tian, J., 2021. Variability, predictability, and uncertainty in global aerosols inferred from gap-filled satellite observations and an econometric modeling approach. *Remote Sens. Environ.* 261, 112501–112514.

- Liang, Y., Zou, B., Feng, W., Liu, N., 2022. Seasonal deviation correction enhanced BGIM downscaling algorithm for remote sensing AOD products. *Natl. Rem. Sens. Bull.* 26 (8), 1602–1613.
- Lin, C.Q., Li, Y., Yuan, Z.B., Lau, A.K.H., Li, C.C., Fung, J.C.H., 2015. Using satellite remote sensing data to estimate the high-resolution distribution of ground-level PM_{2.5}. *Remote Sens. Environ.* 156, 117–128.
- Liu, M., Huang, X., Song, Yu, Tang, J., Cao, J.J., Zhang, X.Y., Zhang, Q., Wang, S.X., Xu, T.T., Kang, L., Cai, X.H., Zhang, H.S., Yang, F.M., Wang, H.B., Yu, J., Lau, A.K.H., He, L., Huang, X.F., Duan, L., Ding, A.J., Xue, L.K., Gao, J., Liu, B., Zhu, T., 2019. Ammonia emission control in China would mitigate haze pollution and nitrogen deposition, but worsen acid rain. *Proc. Natl. Acad. Sci. USA* 116 (16), 7760–7765.
- Liu, M.X., Huang, X., Song, Y., Xu, T.T., Wang, S.X., Wu, Z.J., Hu, M., Zhang, L., Zhang, Q., Pan, Y.P., Liu, X.J., Zhu, T., 2018. Rapid SO₂ emission reductions significantly increase tropospheric ammonia concentrations over the North China Plain. *Atmos. Chem. Phys.* 18, 17933–17943.
- Liu, H., Pinker, R.T., 2005. A global view of aerosols from merged transport models, satellite, and ground observations. *J. Geophys. Res. Atmos.* 110, 4695–4701.
- Lv, B., Hu, Y., Chang, H.H., Russell, A.G., Bai, Y., 2016. Improving the accuracy of daily PM_{2.5} distributions derived from the fusion of ground-level measurements with aerosol optical depth observations, a case study in North China. *Environ. Sci. Technol.* 50 (9), 4752–4759.
- Lyapustin, A., Wang, Y., 2022. MODIS Multi-Angle Implementation of Atmospheric Correct (MAIAC) Data User's Guide, pp. 1–24, Version 3.1.
- Ma, Z., Dey, S., Christopher, S., Liu, R., Bi, J., Balyan, P., Liu, Y., 2022. A review of statistical methods used for developing large-scale and long-term PM_{2.5} models from satellite data. *Remote Sens. Environ.* 269, 112827–112842.
- Meytar, S.-H., Itai, K., Petros, K., Anthony, W.S., Robert, C., Ayala, C., William, L.R., Broday, David M., 2015. Assessment of PM_{2.5} concentrations over bright surfaces using MODIS satellite observations. *Remote Sens. Environ.* 163, 180–185.
- Na, W., Junhu, Z., Panxing, L., 2023. Analysis of the circulation characteristics and driving factors of the super strong autumn flood in Shaanxi Province in 2021. *Chin. J. Atmos. Sci.* 47 (4), 1283–1294.
- Nordio, F., Coull, B.A., Schwartz, J., Itai, K., 2012. Incorporating local land use regression and satellite aerosol optical depth in a hybrid model of spatiotemporal PM_{2.5} exposures in the Mid-Atlantic states. *Environ. Sci. Technol.* 46 (21), 11913–11921.
- Peng, J., Hu, M., Shang, D., Wu, Z., Du, Z., Tan, T., Wang, Y., Zhang, F., Zhang, R., 2021. Explosive secondary aerosol formation during severe haze in the North China Plain. *Environ. Sci. Technol.* 55 (4), 2189–2207.
- Reid, C.E., Considine, E.M., Maestas, M.M., Li, C., 2021. Daily PM_{2.5} concentration estimates by county, ZIP code, and census tract in 11 western states 2008–2018. *Sci. Data* 8 (1), 112–127.
- Schervish, M., Donahue, N.M., 2020. Peroxy radical chemistry and the volatility basis set. *Atmos. Chem. Phys.* 20 (2), 1183–1199.
- Seinfeld, J., Pandis, S., 2016. Atmospheric Chemistry and Physics: from Air Pollution to Climate Change. Wiley.
- Song, W., Jia, H., Huang, J., Zhang, Y., 2014. A satellite-based geographically weighted regression model for regional PM_{2.5} estimation over the Pearl River Delta region in China. *Remote Sens. Environ.* 154, 1–7.
- Taskinen, H., Väistönen, A., Hatakka, L., Virtanen, T.H., Lähivaara, T., Arola, A., Kolehmainen, V., Lippinen, A., 2022. High-resolution post-process corrected satellite AOD. *Geophys. Res. Lett.* 49 (18), e2022GL099733.
- Veefkind, J.P., Boersma, K.F., Wang, J., Kurosu, T.P., Krotkov, N., Chance, K., Levelt, P. F., 2011. Global satellite analysis of the relation between aerosols and short-lived trace gases. *Atmos. Chem. Phys.* 11 (3), 1255–1267.
- Wang, M., Franklin, M., Li, L., 2022. Generating fine-scale aerosol data through downscaling with an artificial neural network enhanced with transfer learning. *Atmosphere* 13 (2), 255–273.
- Wei, J., Li, Z., Chen, X., Li, C., Sun, Y., Wang, J., Lyapustin, A., Brasseur, G.P., Jiang, M., Sun, L., Wang, T., Jung, C.H., Qiu, B., Fang, C., Liu, X., Hao, J., Wang, Y., Zhan, M., Song, X., Liu, Y., 2023a. Separating daily 1 km PM_{2.5} inorganic chemical composition in China since 2000 via deep learning integrating ground, satellite, and model data. *Environ. Sci. Technol.* 57 (46), 18282–18295.
- Wei, J., Li, Z., Cribb, M., Huang, W., Xue, W.H., Sun, L., Guo, J.P., Peng, Y.R., Li, J., Lyapustin, A., Liu, L., Wu, H., Song, Y.M., 2020. Improved 1 km resolution PM_{2.5} estimates across China using enhanced space-time extremely randomized trees. *Atmos. Chem. Phys.* 20 (6), 3273–3289.
- Wei, J., Li, Z., Lyapustin, A., Wang, J., Dubovik, O., Schwartz, J., Sun, L., Li, C., Liu, S., Zhu, T., 2023b. First close insight into global daily gapless 1 km PM_{2.5} pollution, variability, and health impact. *Nat. Commun.* 14 (1), 8349–8362.
- Wei, J., Liu, S., Li, Z.Q., Liu, C., Qin, K., Liu, X., Pinker, R.T., Dickerson, R.R., Lin, J., Boersma, K.F., Sun, L., Li, R., Xue, W.H., Cui, Y.Z., Zhang, C.X., Wang, J., 2022. Ground-level NO₂ surveillance from space across China for high resolution using interpretable spatiotemporally weighted artificial intelligence. *Environ. Sci. Technol.* 56 (14), 9988–9998.
- Wei, J., Wang, J., Li, Z., Kondragunta, S., Anenberg, S., Wang, Y., Zhang, H., Diner, D., Hand, J., Lyapustin, A., Kahn, R., Colarco, P., da Silva, A., Ichoku, C., 2023c. Long-term mortality burden trends attributed to black carbon and PM_{2.5} from wildfire emissions across the continental USA from 2000 to 2020: a deep learning modelling study. *Lancet Planet. Health* 7 (12), e963–e975.
- Xiao, Q., Wang, Y., Chang, H.H., Meng, X., Geng, G., Lyapustin, A., Liu, Y., 2017. Full-coverage high-resolution daily PM_{2.5} estimation using MAIAC AOD in the Yangtze River Delta of China. *Remote Sens. Environ.* 199, 437–446.
- Xu, J., Lin, J., Luo, G., Adeniran, J., Kong, H., 2023. Foreign emissions exacerbate PM_{2.5} pollution in China through nitrate chemistry. *Atmos. Chem. Phys.* 23 (7), 4149–4163.
- Yang, J., Hu, M., 2018. Filling the missing data gaps of daily MODIS AOD using spatiotemporal interpolation. *Sci. Total Environ.* 633, 677–683.
- Zeng, Q., Li, Y., Tao, J., Fan, M., Chen, L., Wang, L., Wang, Y., 2023. Full-coverage estimation of PM_{2.5} in the Beijing-Tianjin-Hebei region by using a two-stage model. *Atmos. Environ.* 309, 119956–119967.
- Zhang, H., Zou, B., Liu, N., Li, S., 2022a. A downscaling model for satellite AOD product improvement in spatial resolution and accuracy. *China Environ. Sci.* 42 (9), 4033–4042.
- Zhang, R., Wang, G., Guo, S., L. Zamora, M., Ying, Q., Lin, Y., Wang, G.W., Hu, M., Wang, Y., 2015. Formation of urban fine particulate matter. *Chem. Rev.* 115 (10), 3803–3855.
- Zhang, T., Zhou, Y., Zhao, K., Zhu, Z., Asrar, G.R., Zhao, X., 2022b. Gap-filling MODIS daily aerosol optical depth products by developing a spatiotemporal fitting algorithm. *GIScience Remote Sens.* 59 (1), 762–781.