Full-coverage 250 m monthly aerosol optical depth dataset (2000-2019) emended with environmental covariates by the ensemble machine learning model over the arid and semi-arid areas, NW China

Xiangyue Chen[1], **Hongchao Zuo**[1], **Zipeng Zhang**[2], **Xiaoyi Cao**[1], **Jikai Duan**[1], **Chuanmei Zhu**[2], **Zhe Zhang**[2], **Jingzhe Wang**[3]

[1]College of Atmospheric Sciences, Lanzhou University, Lanzhou, 730000 China
[2]Key Laboratory of Oasis Ecology, Xinjiang University, Xinjiang Urumqi 830046, China
[3]School of Artificial Intelligence, Shenzhen Polytechnic, Shenzhen, 518055, China

Correspondence: Hongchao Zuo（zuohch@lzu.edu.cn) and Jingzhe Wang (jzwang@szpt.edu.cn)

Abstract

Aerosols are a complex compound with a great effect on the global radiation balance and climate system even human health, and concurrently are a large uncertain source in the numerical simulation process. The arid and semi-arid area has a fragile ecosystem, with abundant dust, but lacks related aerosol data or data accuracy. To solve these problems, we use the bagging trees ensemble model, based on 1 km aerosol optical depth (AOD) data and multiple environmental covariates, to produce monthly advanced-performance, full-coverage, and high-resolution (250 m) AOD products (named FEC AOD, Fusing Environmental Covariates AOD) in the arid and semi-arid areas. Then, based on FEC AOD, we analyzed the spatiotemporal pattern of AOD and further discussed the interpretation of environmental covariates to AOD. The result shows that the bagging trees ensemble model has a good performance, with its verification R^2 always keeping at 0.90 and the R^2 being 0.79 for FEC AOD compared with AERONET AOD. The high AOD areas are located in the Taklimakan Desert and the Loess Plateau, and the low AOD areas are concentrated in the south of Qinghai Province. The higher the AOD is, the stronger the interannual variability. Interestingly, the AOD indicates a dramatic decrease in Loess Plateau and an evident increase in the southeast Taklimakan Desert, while the AOD in the southern Qinghai Province almost shows no significant change between 2000 and 2019. The annual variation characteristics present that AOD is the largest in spring (0.267 ± 0.200) and the smallest in autumn (0.147 ± 0.089); the AOD annual variation pattern shows a different feature, with two peaks in March and August respectively over Gansu Province, but only one peak in April in other provinces/autonomous regions. The farmland and construction land are at high AOD levels compared with other land cover types. The meteorological factors demonstrate a maximum interpretation of AOD on all set temporal scales, followed by the terrain factors, and the surface properties are the smallest, i.e., 77.1%, 59.1%, and 50.4% respectively on average. The capability of the environmental covariates for explained AOD varies with season, with a sequence being winter (86.6%) > autumn (80.8%) > spring (79.9%) > summer (72.5%). In this research, we pathbreakingly provide high spatial resolution (250 m) and long time series (2000-2019) FEC AOD dataset in arid and semi-arid regions to support the atmosphere and related study in northwest China, with the full data available at https://doi.org/10.5281/zenodo.5727119 (Chen et al., 2021a).

Keywords: Aerosol optical depth, Spatial downscaling, Machine learning, Gap filling, Arid areas

1 Introduction

Aerosols are a type of complex substance dispersed in the atmosphere that can be natural or anthropogenic sources (Kaufman et al., 2002). Aerosols can affect the global radiation balance and climate system directly, indirectly, or semi-indirectly by absorbing or scattering solar radiation (Myhre et al., 2013). Concurrently, aerosols seriously endanger human health by mixing, reacting, and dispersing dangerous compounds (Chen et al., 2020; Lelieveld et al., 2019). As one of the most significant optical characteristics of aerosols, the aerosol optical depth (AOD) is the integral of aerosol extinction coefficient in the vertical direction and indicates the attenuation impact of aerosols on solar energy (Chen et al., 2021b). AOD is frequently adopted to depict air pollution and also indirectly calculate various atmospheric parameters, such as particulate matter 2.5/10, with an extensive application in atmospheric environment-related research (Goldberg et al., 2019; He et al., 2020).

Generally, the primary AOD acquisition method is in-situ observation, which has high precision. However, in-situ observation is restricted by the distribution of observation stations, so the data lacks spatial continuity, which makes it difficult to meet the objectives of growing regional atmospheric environmental studies (Zhang et al., 2019). Remote sensing (RS) is an effective tool for collecting AOD information over a wide range of spatial scales, significantly offsetting the deficiency of in-situ observation. RS can tackle difficulties connected to insufficient data and an uneven geographical distribution to a certain extent (Chen et al., 2020). Nonetheless, RS is not always a silver bullet for AOD acquirement, with some problems, such as low spatial resolution and data missing in some particular situations (Li et al., 2020). Commonly utilized AOD satellite products derived from various sensors have different emphases in use (Table S1). Yet, the common point is that spatial resolution is coarse, and even has a large number of nodata values (Chen et al., 2022; Sun et al., 2021; Chen et al., 2021b; Wei et al., 2021). All these restrict the application of satellite AOD products on a regional scale, especially on a local scale. Furthermore, the AOD spatial resolution scale often inevitably affects the following atmospheric pollutant prediction (Yang and Hu, 2018). These issues not just affect AOD analysis, but also mislead numerous pertinent uses of AOD data.

Although methods for resolving AOD RS data deficiency have been studied, previous research has not addressed the problem completely (Li et al., 2020; Zhao et al., 2019). Considerable related work concentrates on multi-source AOD dataset fusion or AOD gap filling using different models. The initial and most extensive method is interpolations, but the AOD shows high spatiotemporal variability, thus it is not suitable to apply the approach to anticipating AOD missing data (Singh et al., 2017). Another widely used method is merging multiple AOD products, which can improve data quality but often fails to eliminate completely pixel value missing phenomenon, even bringing about offsetting consequences (Bilal et al., 2017; Ali and Assiri, 2019; Wei et al., 2021). Some statistical models such as linear regression and additive are also employed to fill the pixel values missing and improve the spatial resolution of the AOD products. However, the performance of these models is often dubious due to their simple structure (Xiao et al., 2017). Most current methods for high-resolution AOD forecasts are focused on the individual model technique, which relies on a set of assumptions that are not frequently met, leading to inaccurate predictions (Li et al., 2017; Zhang et al. 2018). As computing technology advances, ensemble machine learning methods, by training multiple models through resampling the training data with the corresponding environmental covariates from their original distribution, provide new considerations and ways, which are less constrained by the hypothesis in a single model, with less over-fitting and outliers (Li et al., 2018). The strong data mining ability of the ensemble machine learning methods is good for fitting multisource data, and it can achieve higher precision at the same time (Zhao et al., 2019). As a result, the present research attempts to adopt ensemble machine learning methods to explore the production of advanced-performance, high-resolution, full-coverage AOD dataset in arid and semi-arid areas.

Currently, many previous studies have focused on AOD research in various regions and scales, which are concentrated on the eastern coastal areas and lack related exploration in arid and semi-arid areas. Arid and semi-arid areas, as important components of the earth's geography units, have extremely fragile bio-system and are extremely sensitive to climate change and human activities (Huang et al., 2017). Since the complex surface situation in arid and semi-arid areas, especially having huge deserts, many AOD retrieval algorithms are not suitable there. Although a minority of algorithms can acquire AOD in arid and semi-arid areas, such as the deep blue (DB) algorithm and multiangle implementation of atmospheric correction (MAIAC) algorithm, which still is limited by coarse resolution, high uncertainty, or a large no-data phenomenon, so these AOD productions are hard to meet the needs of arid and semi-arid areas atmosphere environmental research (Wei et al., 2021). However, arid and semi-arid areas are crucial dust sources, with strong variability in the aspects of aerosol loading and optical characteristics. As a typical dust source and AOD data-scarce areas, the AOD variety in arid and semi-arid areas has significant influences on global climate change and model simulation. Therefore, manufacturing a higher-quality AOD dataset in arid and semi-arid areas is necessary for local and even global atmosphere environment research.

To better solve the lack of AOD data in arid and semi-arid areas, this research aims to acquire advanced-performance, high-resolution, full-coverage AOD datasets that will serve as the foundation for future studies. To achieve this goal, the main work of this study includes: (1) based on MAIAC AOD, combined with multiple environmental covariates, utilized a machine learning method, FEC AOD is obtained for the periods 2000–2019; (2) Aerosol Robotic Network (AERONET) ground observation data and the MCD19A2 and MxD04L2 AOD satellite products were collected to verify the applicability of FEC AOD; (3) the FEC AOD spatiotemporal patterns is analyzed; and (4) the dominant environmental covariates of FEC AOD are explored.

2 Materials and methods

2.1 Study area

Figure 1 shows the arid and semi-arid areas in northwest China (E 73°25’ - 110°55’, N 31°35’ - 49°15’), a typical arid and semi-arid region on the globe, in terms of the spatial location, surface cover, and the environmental problem (Ge et al., 2016). As a dust source and an ecosystem fragile area, the regional difference in climate is significant, which is perennial in drought and less precipitation (< 400 mm) conditions

(Ding and Xingming, 2021). Furthermore, the area is extremely sensitive to climate change and human activities and has a large AOD variability, which brings great difficulty to global climate simulation and radiation balance quantification. With the development of society and technology, the force of people to change nature is increasing. More and more unreasonable human activities (deforestation, soil salinization) and poor land management policies (reclamation, water resources utilization) bring about regional vegetation degradation, desertification, rapid glacier melting, and frequent dust weather, which eventually lead to the fast deterioration of the ecological environment in the whole arid and semi-arid areas.

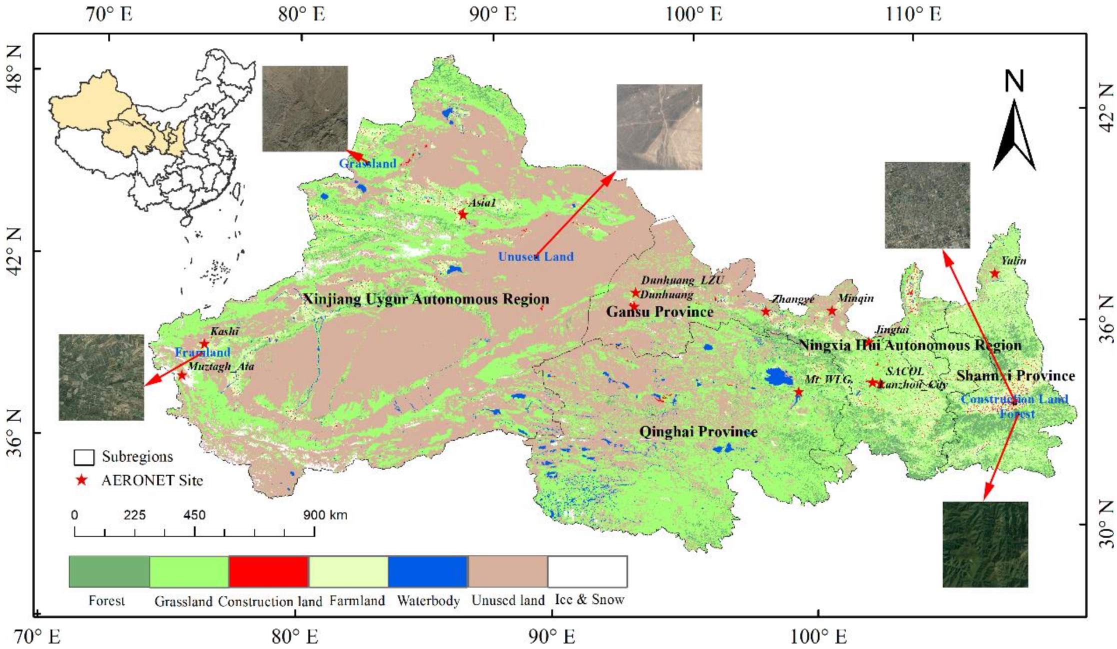

Figure 1. Study area. The figure shows typical arid and semi-arid areas and AERONET site distribution, five provinces/autonomous regions in northwest China.

2.2 MODIS MAIAC data

MAIAC AOD, which is named MCD19A2, is based on MODIS onboard Terra and Aqua, combined with the MAIAC algorithm produced. The MAIAC algorithm is an advanced AOD retrieval method, using time-series analysis and image-based spatial processing, which can acquire AOD data from densely vegetated areas as well as bright desert regions (Lyapustin et al., 2018; Lyapustin et al., 2011). The MAIAC AOD product's temporal and spatial resolutions are 1 day and 1 km × 1 km respectively, which is the highest spatial resolution in existing AOD products. The MAIAC AOD product also offers a long-time-series AOD collection, which has been intended for air quality research on regional and even global scales. Compared with former AOD products, the MAIAC AOD data performance on bright surfaces and heavy AOD loadings areas generally is considered to make a significant improvement (Li et al., 2018; Chen et al., 2021b). In this paper, we acquired MAIAC AOD for the entire study region from the NASA website (https://search.earthdata.nasa.gov/) over 20 years, from March 2000 to February 2020. Using the python tool, we preprocessed the data and computed the daily average AOD by combining the 550 nm AOD data from Terra and Aqua.

2.3 MODIS MxD04L2 data

MYD04L2 and MOD04L2 are the level 2 atmospheric aerosol products from Aqua and Terra respectively, where spatial and temporal resolutions are 10 km × 10 km and 1 day respectively (Zhao et al., 2021). The MxD04L2 AOD product mainly provides two algorithms, the Dark Target (DT) and Deep Blue (DB) algorithms, to retrieve global AOD distribution. Based on the MODIS Collection 6.1, we chose 550 nm combined DT and DB AOD to validate FEC AOD. It is noted that the Aqua and Terra launch time is different, so we can acquire MOD04L2 data from March 2000 to February 2020, but as for MYD04L2, we only acquire data from July 2002 to February 2020. All processes are realized through downloading from NOAA website (https://ladsweb.modaps.eosdis.nasa.gov/) and calculating and analyzing local computer, and main works, including geometric correction, projection conversion, image mosaics, clipping, computing daily and monthly mean of AOD, and numerical extraction, perform in MODIS Reprojection Tool (MRT), ENVI, and ArcGis software.

2.4 AERONET data

AERONET (Aerosol Robotic Network) is a network that monitors aerosols on the ground, providing 0.340-1.060 m aerosol optical characteristics at a high temporal resolution (15 min) (Holben et al., 1998). AERONET currently includes more than 500 sites and covers major regions of the world with a long time series. AERONET AOD has low uncertainty (0.01–0.02), which is considered the highest accuracy AOD data and is widely used in RS AOD products validation as a reference (Almazroui, 2019). In this study, a total of 12 AERONET site data are selected in northwest China, most of which are from the third version of Level 2.0 AERONET AOD, except Mt_WLG station (Level 1.5) (Yan et al., 2022; Giles et al., 2019). Related information about these AERONET sites is available in Table S2 and Figure 1. Satellite products mostly provide 550 nm wavelength AOD, so the AERONET AOD at 550 nm is computed via the Ångström exponent algorithm to better match the AOD observed by satellite (Ångström, 1964). In the temporal dimension, we compute the average of AERONET AOD over Aqua and terra overpass period. In the spatial dimension, we match the satellite and in-situ observed AOD over a 3 × 3 pixels spatial window (Tao et al., 2017). The AERONET data and related information can be found at https://aeronet.gsfc.nasa.gov.

2.5 Environmental covariates

Environmental covariates selected in this study contain 12 covariates in three categories (meteorological parameters, surface properties, and terrain factors). Covariates are selected based on two criteria: first, each variable is considered important to AOD and has a vital influence on AOD formation, accumulation, and migration process, referring to existing research and expert experience (Zhao et al., 2019; Chen et al., 2020; Yan et al., 2022); the second, the data is released to the public for free, which means that the data set is freely available on the national or global scale (Li et al., 2020). The detailed information is listed in Table 1. In this study, we compute two sets of spatial resolution of environment variable data (1 km and 250 m). The 1 km spatial resolution data aim to model with MAIAC 1 km AOD, and 250 m spatial resolution data is the target resolution of FEC AOD. To normalize the covariables on this basis, we interpolated the geo-datasets to 1 km and 250 m in ArcGIS (the bilinear method is used for continuous covariates and the nearest neighbor method is used for classified covariates) and reprojected them onto the 1984 coordinate system of the World Geodetic System (WGS). The environmental covariates can be divided into static and dynamic variables. Static variables are defined as those that do not change essentially with time, i.e., slow change factors. As for dynamic covariates, the average method is adopted to obtain monthly average data. Static variables, similar to a baseline condition, play an initial constraint role in the downscaling of monthly AOD, while dynamic variables play a more dynamic evolution role (Yan et al., 2022). It is noted that the relevant operations are not limited to ArcGIS, and relevant open-source software such as QGIS can also be implemented.

2.5.1 Meteorological parameters

The meteorological parameters include temperature, precipitation, evapotranspiration, and wind speed. The temperature and precipitation data are obtained from the national Tibet Plateau data center (TPDC), whose temporal and spatial resolution is 1 month and 1 km × 1 km respectively. The evapotranspiration (ET) data is from TPDC's terrestrial evapotranspiration dataset across China, whose temporal and spatial resolution is 1 month and 0.1° × 0.1° respectively (Szilagyi et al., 2019). For ET data, we use a downscaling algorithm proposed by Ma (2017) to transform it into 1 km. The wind speed data is from National Earth System Science Data Center, whose temporal and spatial resolution is 1 month and 1 km × 1 km respectively (Sun et al., 2015). As for the four meteorological parameters, we have calculated the monthly average state every year for the next research.

2.5.2 Surface properties

The surface properties mainly employ land use and land cover (LUCC), normalized difference vegetation index (NDVI), and temperature vegetation dryness index (TVDI) to describe. LUCC data selects in the median of the whole study time, 2010, which is from Resource and Environment Science and Data Center. The LUCC data set is obtained by manual visual interpretation of the Landsat Series data as the data source. It includes 6 categories (farmland, forest, grassland, waterbody, construction land, and unused land) and 25 subcategories, with a spatial resolution of 30 m. The LUCC is often likely to indicate the intensity of human activity and is closely related to aerosol emissions, transport, and dustfall (Fan et al., 2020; Li et al., 2022). NDVI data is obtained from NASA Global Inventory, Monitoring, and Modelling Studies (GIMMS) NDVI3g v1, whose temporal and spatial resolution is 15 days and $0.083^{\circ} \times 0.083^{\circ}$ respectively. NDVI, the same as ET, is downscaled to 1 km. TVDI is a soil moisture inversion method based on NDVI and surface temperature. It can better monitor drought and be used to study the spatial variation characteristics of drought degree. TVDI temporal and spatial resolution is 1 month and 1 km × 1 km respectively.

2.5.3 Terrain factor

The elevation is from Shuttle Radar Topography Mission 90 m Digital Elevation Model (SRTM). DEM is highly correlated with surface pressure, and always used to represent the dispersion condition of aerosols (Xue et al., 2021; Fan et al., 2020). Based on elevation, geomorphology is realized under Geographic Resource Analysis Support System extension named r.geomorphon modular (Jasiewicz and Stepinski, 2013). Using System for Automated Geoscientific Analyses soft (https://sourceforge.net/projects/saga-gis/), plan curvature, slope length and slope steepness, and topographic wetness index is computed.

Table 1. Environmental covariates for AOD modeling

Type	Name	Abbreviation	Resolution	Sources
Dynamic covariates				
Meteorological parameters	Temperature	Tem	1 km × 1 km	http://data.tpdc.ac.cn/
	Precipitation	Pre	1 km × 1 km	http://data.tpdc.ac.cn/
	Wind speed	WS	1 km × 1 km	http://www.geodata.cn/
	Evapotranspiration	ET	0.1° × 0.1°	http://data.tpdc.ac.cn/
Surface properties	Normalized difference vegetation index	NDVI	0.083° × 0.083°	https://ecocast.arc.nasa.gov/data/pub/
	Temperature vegetation dryness index	TVDI	1 km × 1 km	http://www.geodata.cn/
Static covariates				
Surface properties	Land use and land cover	LUCC	30 m × 30 m	http://www.resdc.cn/
Terrain factors	Elevation	Elev	90 m × 90 m	http://srtm.csi.cgiar.org/srtmdata/
	Geomorphology	Geoms	90 m × 90 m	
	plan curvature	Curpln	90 m × 90 m	
	slope length and slope steepness	LS	90 m × 90 m	
	topographic wetness index	TWI	90 m × 90 m	

2.6 Bagging trees ensemble

The ensemble machine learning methods according to whether there exists dependency relation between learners are mainly divided into two categories, Boosting and Bagging (Figure S1) (González et al., 2020). If there is a strong dependency between individual weak learners and a series of individual weak learners needs to be generated serially (That means that the following weak learner is affected by the former weak learner), which is Boosting. In contrast, if there is no dependency between individual weak learners, and a series of individual learners can be generated in parallel (There is no constraint relationship between each learner), which is Bagging. The typical representative and extensive use algorithms of Boosting and Bagging are Gradient Boosting Decision Tree (GBDT) and Random Forest (RF) respectively (Zounemat-Kermani et al., 2021). Compared with Boosting, Bagging reduces the difficulty in training and has a strong generalization.

Bagging (namely bootstrap aggregating) as a simple but powerful ensemble algorithm to obtain an aggregated predictor is more accurate than any single model (Breiman, 1996). Bagging is through multiple base learners or individual learners (such as decision trees, neural networks, and other basic learning algorithms) to construct a robust learner under certain combined strategies (Li et al., 2018). Generally, the bagging algorithm includes bootstrap resampling, decision tree growing, and out-of-bag error estimate. The main steps of the Bagging are as follows: (1) Bootstrap resampling, a random sample (return sampling) under abundant individual weak learners. (2) Model training, based on the origin samples to training for abundant individual weak learners in accordance with the self-serving sample set. (3) Result output, based on the decision tree, and calculates the average of all the regression results to obtain regression results. Therefore, bagging reduces the overfitting problem and prediction errors in decision trees and variance, thereby significantly improving the accuracy. Simultaneously, the influence of noise on the Bagging algorithm is comparatively less than the other available machine learning algorithms for AOD (Liang et al., 2021).

In this study, we use 12 environmental covariates (1 km) as downscaling method (bagging trees ensemble algorithms) input to acquire AOD-environmental covariates (AODe) model in 1 km and utilize AODe model and 250 m environmental covariates to acquire FEC AOD. Specifically, the basic idea for downscaling AOD with bagging trees ensemble machine learning (ML) models is to train the relationships between MAIAC AOD and the auxiliary environmental variables at coarse resolution (1 km) using ML algorithms. We then apply the trained relationships to generate a high-resolution FEC AOD product at a fine resolution (250 m) (Duveiller et al., 2020; Yang et al., 2020; Ma et al., 2017). In case of the lack of environmental covariates in some periods, we use the multi-year monthly average to replace them. The reason why the 250 m target resolution is selected is that existing studies show that aerosol RS research at the scale of 250 - 500 m spatial resolution is appropriate, which can better capture aerosols feature (Wang et al., 2021; Chen et al., 2020). Secondly, most high-resolution product data in the global are 250 m, especially soil, which is more convenient for peer comparison and further research and application (De Sousa et al., 2020; Hengl et al., 2017). The model was built monthly from March 2000 to February 2020 to assure the model's accuracy in the inference process, whose specific parameters set include the 10 cross-validation folds, the number of learners ($N = 30$), and the minimum leaf size ($L_{min} = 8$). Each base learner was developed using a bootstrap sample generated individually from the input data. All steps were implemented in Matlab R2021a (Figure 2). Definitely, all modeling and application processes can also be implemented in R or Python.

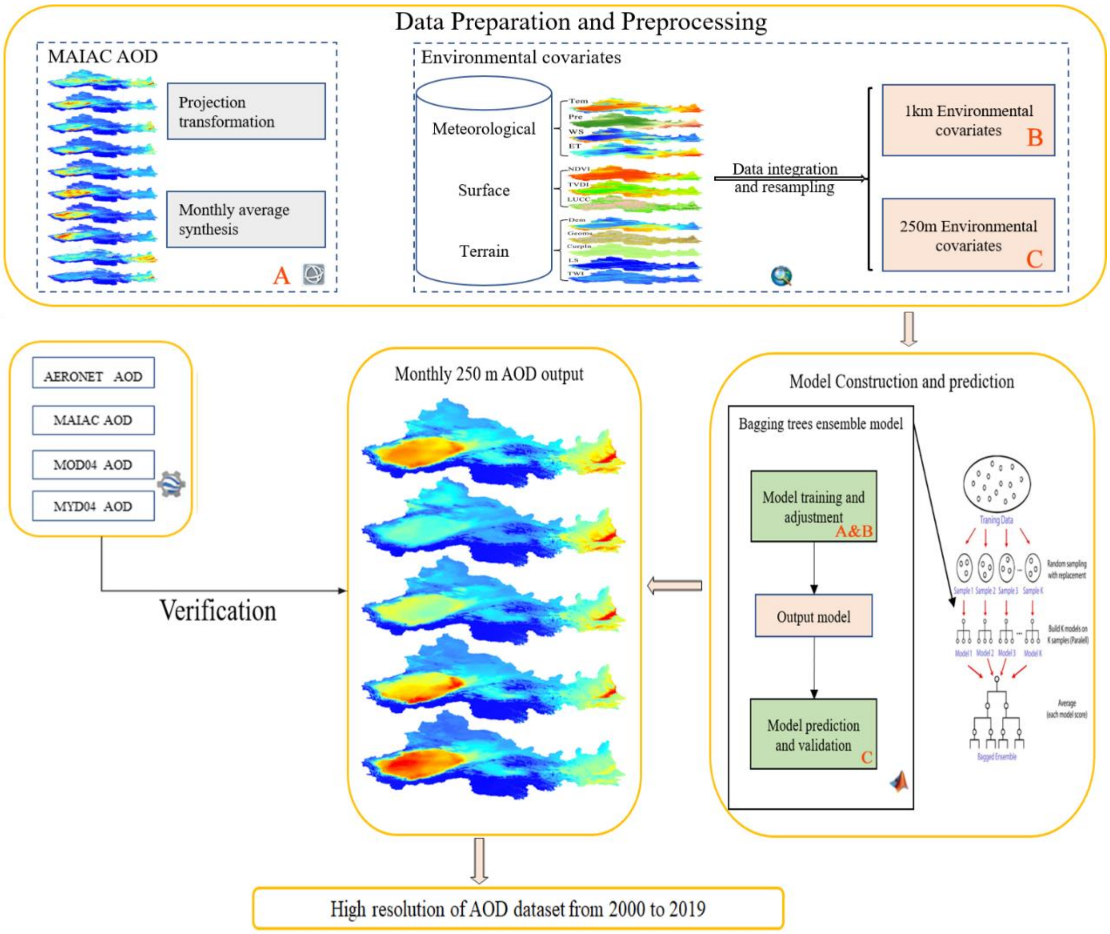

Figure 2. Flow chart of experiment and model calculation process.

3 Results and analysis

3.1 Performance evaluation based on in-situ observation

To verify the performance of the FEC AOD over arid and semi-arid areas, based on AERONET AOD data as reference, some generalized parameters are chosen to assess the performance of FEC AOD, such as the decision coefficient (R^2), root mean square error (RMSE), expected error (EE), etc. (Levy et al., 2010; Ali et al., 2019; Feng et al., 2021). When R^2 is higher and RMSE is lower, the performance of the FEC AOD is better. EE can evaluate the degree of overestimation and underestimation of FEC AOD via three situations (within EE, above EE, and below EE). To examine the high resolution and full coverage FEC AOD performance, we computed the monthly average

AOD at each AERONET site in the whole study region. Specifically, we check data time range and data usability at every site, as for the daily scale, we only compute the average AOD from local time 9:00 am to 2:00 pm as the daily mean (if the valid data number in a day is less than 18, daily mean is considered missing). As for the monthly scale, if the number of the effective daily mean is less than 20 days, the monthly mean is considered missing, so 180 effective matching samples were obtained. As shown in Figure 3a, FEC AOD was highly correlated with AERONET AOD ($R^2 = 0.787$), with MAE of 0.049 and RMSE of 0.061. Approximately 83.9% of monthly collections fell within the EE, with RMB of 1.018 and Bias of 0.005, which means the FEC AOD products almost overcome some problems of overestimation and underestimation. Concurrently, the MAIAC AOD (Figure 3b), MOD04 L2 (Figure 3c), and MYD04 L2 (Figure 3d) also conduct a comparison with AERONET AOD for the same period. MAIAC AOD is superior to the MxD04L2 AOD, and FEC AOD has obvious improvements compared with MAIAC AOD, within EE from 65.0% to 83.9%. It is clear to find that the performance of FEC AOD obviously outperforms other AOD products in terms of the number of valid data, consistency, and deviation. In addition, compared with previous studies, the FEC AOD also has a better applicability advantage (Chen et al., 2021b; Wei et al., 2019).

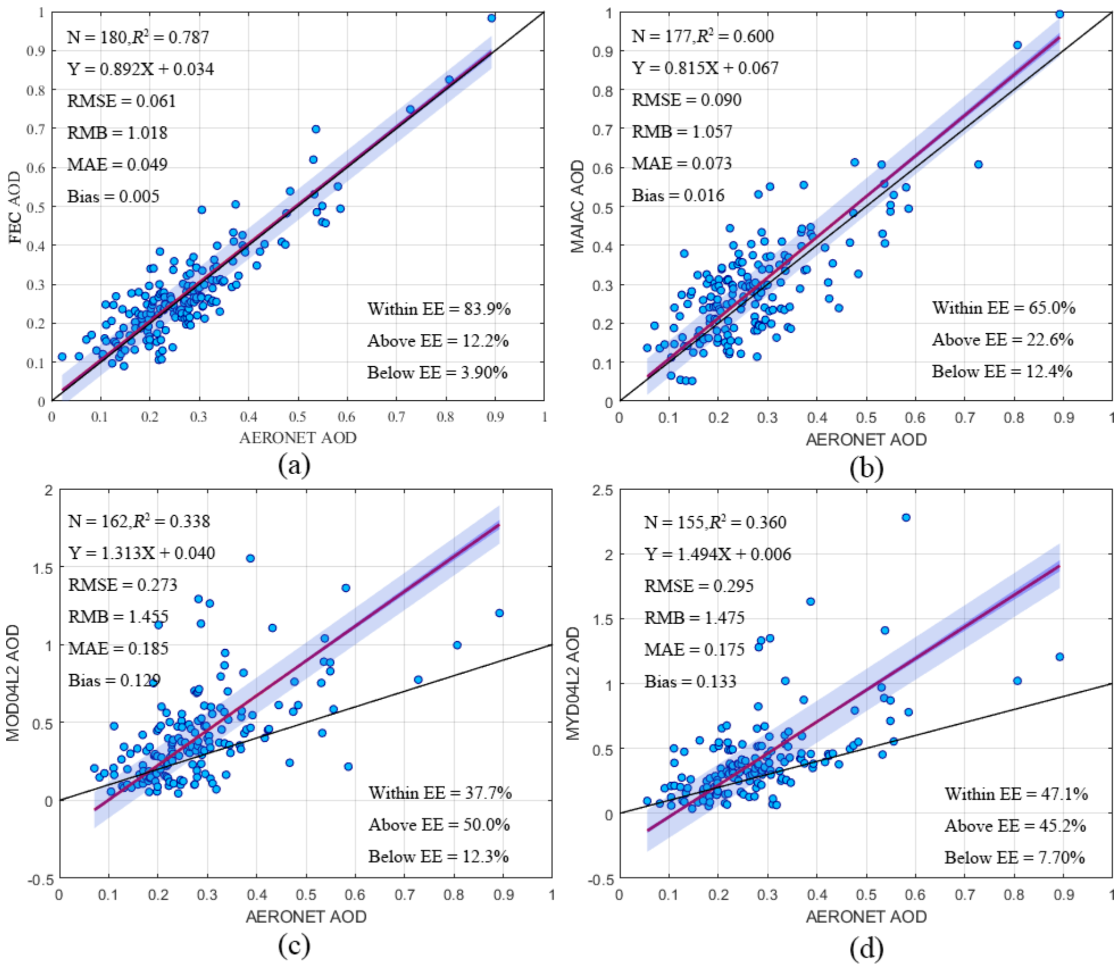

Figure 3. Comparison with AERONET AOD. (a) FEC AOD, (b) MAIAC AOD, (c) MOD04L2 AOD, (d) MYD04L2 AOD. The red line denotes the regression line, the black line shows the 1:1 line, and the blue area indicates the 95% prediction interval.

3.2 Comparison with satellite AOD products

The multi-year average AOD spatial distribution of FEC AOD, MAIAC AOD, MOD04L2 AOD, and MYD04L2 AOD was calculated (Figure 4). The AOD spatial pattern has high consistency among these, and the high AOD is located in Taklimakan Desert and Loess Plateau, and the low AOD is distributed in high altitude areas (such as the mountain zone and Qinghai Province). To further validate the FEC AOD performance, we calculated the monthly, seasonal, and yearly average AOD from 2000 to 2019 (Figure S2-S5). In terms of monthly scale (Figure S2), we can find that many high AOD values appear in March, April and May, concentrated in Taklimakan Desert and its downwind. Generally, FEC AOD, MAIAC AOD, MOD04L2 AOD, and

MYD04L2 AOD have a similar monthly spatial distribution, especially FEC AOD and MAIAC AOD. The monthly correlations of FEC and MAIAC AOD are all above 0.78 in the study area, most of which are higher than 0.9 (R_{mean} = 0.928, N = 240, P < 0.001) (Figure S3). A similar spatial pattern also appears in the multi-year seasonal average AOD (Figure S4). Spring witnesses the broadest high AOD value distribution, followed by summer, while autumn and winter witness the high AOD concentrated in Loess Plateau. At the same time, the multi-year yearly average AOD also has a strong similarity in spatial pattern (Figure S5). Therefore, we can robustly conclude FEC has a strong consistency with MAIAC AOD, MOD04L2 AOD, and MYD04L2 AOD in the monthly, seasonally, and yearly average AOD spatial pattern.

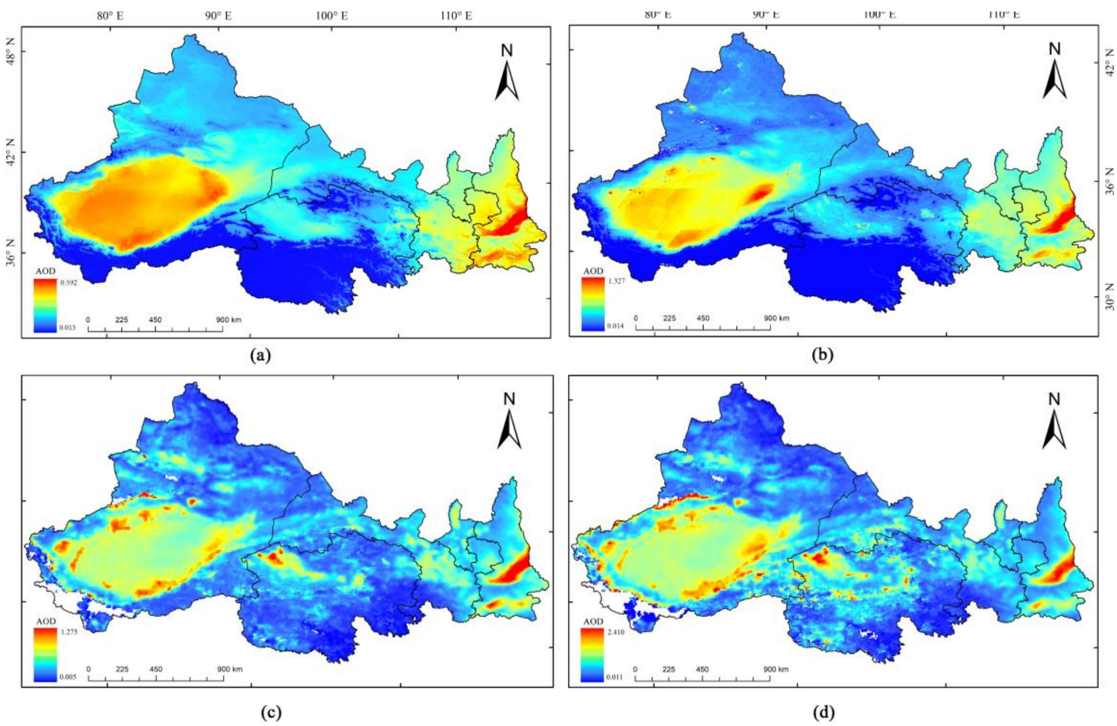

Figure 4. The muti-years spatial average AOD from 2000 to 2019. (a) FEC AOD, (b) MAIAC AOD, (c) MOD04L2 AOD, and (d) MYD04L2 AOD.

Considering that the ability in capturing long-term trends is an important element for a dataset, we compare the FEC AOD, MAIAC AOD, MOD04L2 AOD, and MYD04L2 AOD to further validate FEC AOD. From January to December, the multi-year monthly average of four AOD products shows a similar change trend, increasing and decreasing alternately, reaching the lowest value in November (Figure S6). Of course, there are some differences in the AOD magnitude and fluctuation range, which are mainly due to the difference in AOD retrieval algorithms. To further analyze the consistency and difference, we also compared four products on monthly scale by removing seasonal cycles (Figure 5). Firstly, the four AOD data change in a highly similar manner, and the MxD04L2 AOD fluctuation range is significantly higher than that of FEC AOD and MAIAC AOD. It is noted that FEC AOD and MAIACAOD are substantially consistent with a R^2 of 0.953. In addition, we also computed the monthly and seasonal change trend by removing the seasonal cycles on pixel scale. Because the MxD04L2 has a large missing data, and the detrend results show nodata, we mainly discuss the FEC AOD and MAIAC AOD spatial change trend in the following on monthly and seasonal scales. From Figure S7-S8, we can find that the monthly and seasonal change trend has a good consistency between FEC AOD and MAIAC AOD. The long-term trend based on monthly AOD data between 2000 and 2019 shows a similar spatial pattern in the effective pixel of both FEC AOD and MAIAC AOD (Figure 6), with a significant decrease in the Loess Plateau and a significant increase in the southeastern Taklimakan Desert. Moreover, the long-term trend of FEC AOD is significant ($P < 0.05$) in most areas.

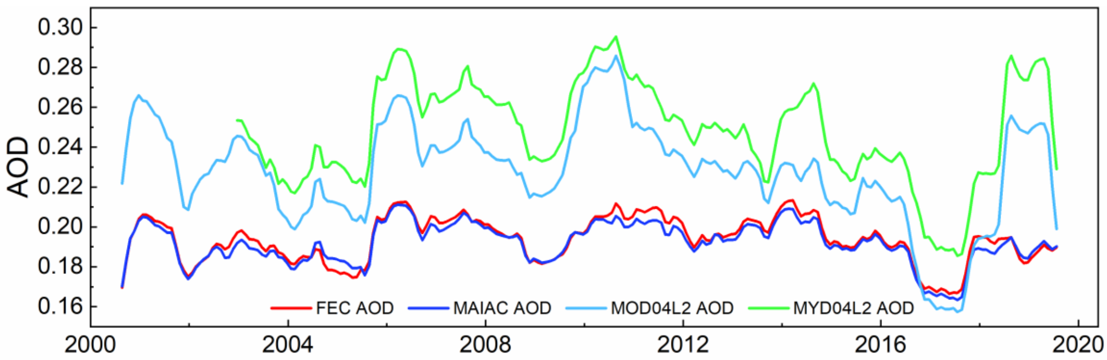

Figure 5. The long-term change trends of four AOD products by removing seasonal cycles.

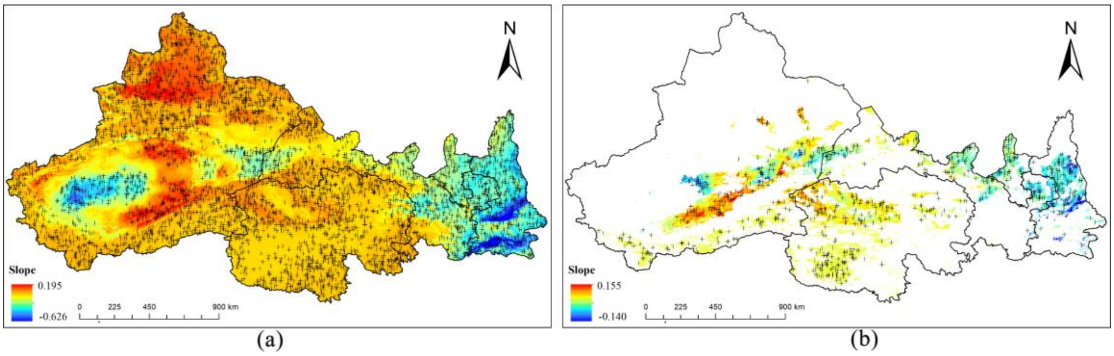

Figure 6. Spatial patterns of AOD trends by removing seasonal cycles between 2000 and 2019. (a) FEC AOD (10^{-3}), (b) MAIAC AOD (10^{-2}). The label '+' indicates statistically significant trends ($p < 0.05$).

As we all know, the effect of scale is a scientific problem in remote sensing, so we further discuss the ability of FEC AOD to describe finer spatial resolution features. Firstly, we create a 10 km × 10 km fishnet; then, a single LUCC is chosen as a corresponding ecosystem; finally, five different ecological zones (forest, grassland, farmland, construction land, and unused land) are selected to further quantify the local performance of FEC AOD (Figure 1). We can find FEC AOD and MAIAC reveal good consistency in the long-term trend, and MxD04L2 AOD show a larger deviation (Figure 7). The FEC AOD and MAIAC AOD has a close relationship in unused land ($R = 0.959$) and farmland, followed by construction land and forest, and grassland ($R = 0.675$) has the lowest relationship. Therefore, the above can evidence FEC AOD also is reliable in the fine spatial resolution long-term trend capture over single surface coverage.

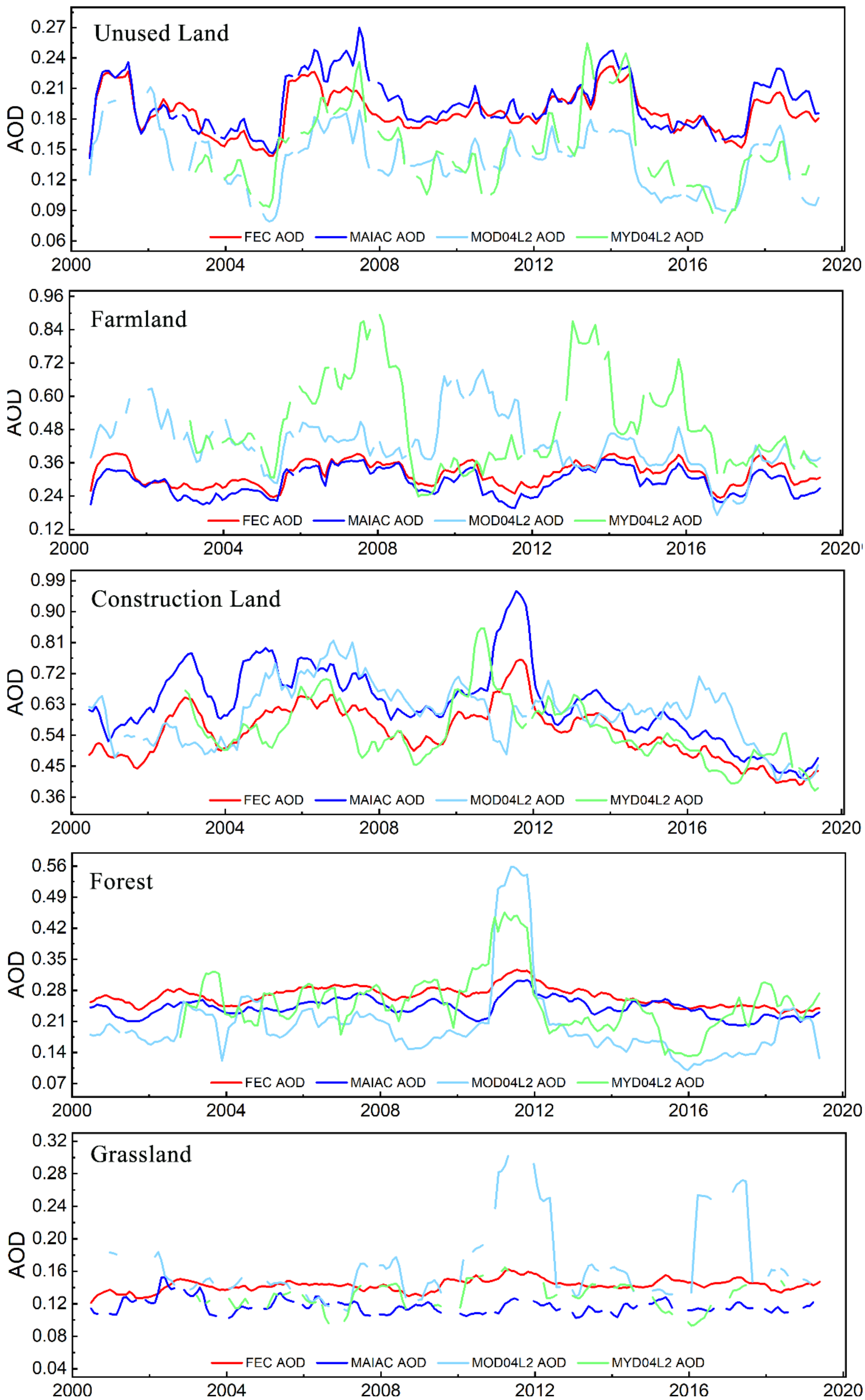

Figure 7. The long-term change trends of four AOD products over five ecological zones by removing seasonal cycles.

3.3 Spatiotemporal patterns of FEC AOD from 2000 to 2019

Generally, spatial patterns of FEC AOD are consistent over different years (Figure S5), where the highest AOD are found in the south of Xinjiang Uyghur Autonomous Region of China (Hereinafter referred to as Xinjiang) and the center of Shaanxi Provinces, mainly due to special meteorological conditions, unique topography and surface coverages. AOD is low in other areas, especially in the south of Qinghai Province. The multi-year mean AOD is 0.193 ± 0.124 for the whole of the study areas. The spatial patterns of AOD greatly differ at the seasonal level (Figure S4). In autumn, AOD is the lightest, with an average AOD value of 0.147 ± 0.089 and most AOD values < 0.2. By contrast, AOD is most severe in spring, with most AOD values > 0.2 (average = 0.267 ± 0.200). The summer and winter have similar spatial patterns and the former is higher than the latter, with AOD values being 0.198 ± 0.134 and 0.159 ± 0.103 respectively. To further investigate the spatiotemporal variety feature of AOD, the concepts of information entropy are introduced, which are temporal information entropy (TIE) and time-series information entropy (TSIE) respectively (Ebrahimi et al., 2010). TIE and TSIE are time series indicators that can depict the changing intensity and trend information of AOD. Generally, the higher (lower) the TIE is, the stronger (weaker) the changing intensity of AOD in the temporal dimension. As for TSIE, if TSIE >0, the shows AOD is increasing in this period, whereas TSIE <0 denotes a downward trend. Furthermore, the bigger the absolute value of TSIE is, the more significant the increasing (decreasing) trend. Figure 8 depicts the TIE and TISE of AOD from 2000 to 2019 over the whole study area. We find that the overall change intensity of AOD over the past 20 years is large, especially in the south of Xinjiang (Taklimakan Desert) and Shannxi Province (Loess Plateau). The areas with low variation intensity are mainly distributed in high elevations (mountainous areas and grassland areas). The characteristic of changing intensity is similar to the AOD change, which means the higher AOD is, the larger the multi-year change. The AOD in Xinjiang is increasing, with the most obvious increases occurring around the Taklimakan Desert and the north of Xinjiang, whereas that in the east is decreasing, mainly concentrated in Shannxi

Province and southeast of Gansu Province. Considering TIE and TSIE together, we can find that AOD has strongly increased in southeastern Taklimakan Desert while slightly increasing in northern Xinjiang and the northwestern Qinghai Province. The AOD in the south of Qinghai Province shows almost no change. The dramatic decrease can be found in the east, mainly distributed in the Shannxi Province, Ningxia Hui Autonomous Region, and southeastern Gansu Province. A possible reason for this finding is that the Loess Plateau is experiencing greening, and the vegetation keeps increasing under artificial intervention. All these various characteristics are in good agreement with the de-trending long-term variation results (Figure 6).

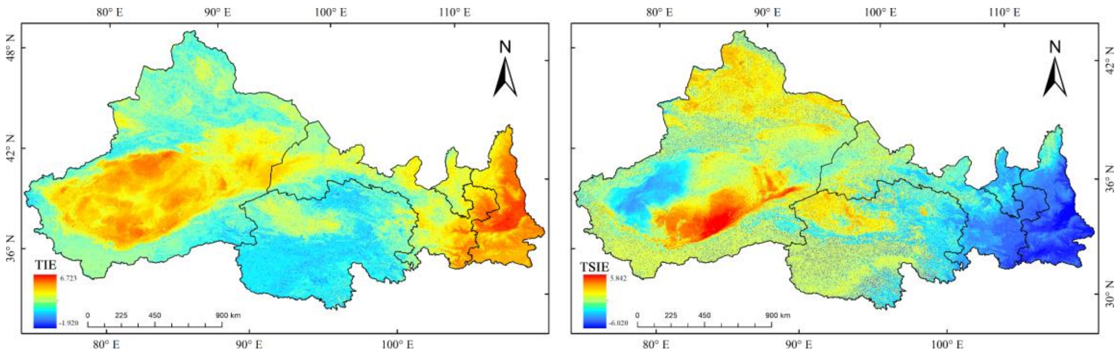

Figure8. Temporal information entropy (TIE) and time-series information entropy (TSIE) of AOD distribution.

The FEC AOD products with high spatial resolution and full coverage over arid and semi-arid areas provided new possible data sources to further research air pollution in scarce data areas on fine scales. Based on the FEC AOD, we explore the regional distribution characteristics under different areas and surface coverage types. Figure 9 shows that AOD in Gansu Province is the highest in all months, and AOD in Qinghai Province is the lowest. From January to December, the AOD almost shows a trend of increasing at first and decreasing next, reaching a peak in March and April. It is noted that except for the Gansu Province, where AOD is bimodal, other provinces/autonomous regions are unimodal. Figure 10 describes the AOD seasonal distribution under seven different land cover types (forest, grassland, waterbody, ice and snow, construction land, unused land, and farmland). The AOD over the ice and snow is the smallest and keeps decreasing from spring to winter. AOD is at a high level over farmland and construction land, which is mainly related to human activities. Despite the land cover type, AOD in spring is still the highest. Except for ice and snow and unused land, else land cover type keeps a similar seasonal distribution, with decrease and then increase, and autumn is the bottom.

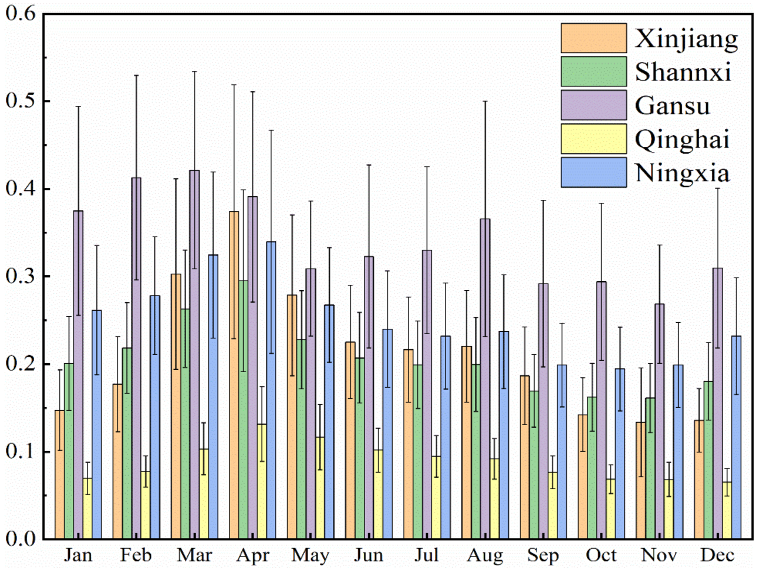

Figure 9. The monthly distribution characteristics of AOD in different provinces/autonomous regions. The error bars represent the standard errors.

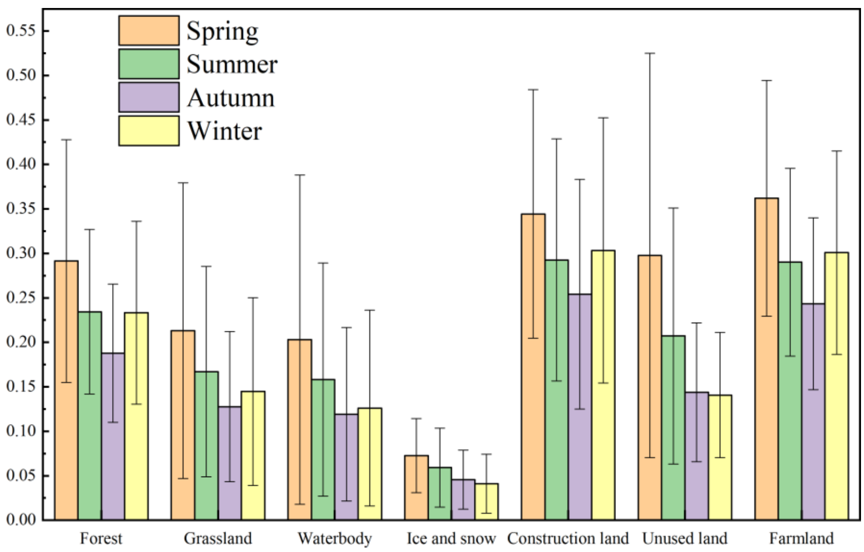

Figure 10. AOD seasonal distribution under different land cover types. The error bars represent the standard errors.

3.4 Variation partitioning of FEC AOD

To examine the contribution of environmental covariates to the FEC AOD dynamic, the redundancy analysis (RDA) was used to explore the association between different seasons of FEC AOD and the environmental covariates. The twelve environmental covariates were divided into three groups, meteorological parameters, surface properties, and terrain factors. The variance proportion driving the variation of FEC AOD on different temporal scales was tested from the environmental covariate groups. The variation of FEC AOD can be interpreted by every group of environmental covariates individually or the combined variation owing to two or more covariates set, and the residual represents the unexplained proportion. The variance partitioning results can be described as Venn's diagram makes by R language (Waits et al., 2018). From Table 2 and Figure 11, the variation partitioning analysis reveals that the meteorological factors still explain a maximal proportion of variance of FEC AOD on different temporal scales, followed by terrain factor, and the surface properties are the smallest, i.e., 77.1%, 59.1%, and 50.4% respectively on average. In different seasons, the environmental covariates have different abilities to explain FEC AOD, with the sequence being winter (86.6%) > autumn (80.8%) > spring (79.9%) > summer (72.5%). Except for winter, the largest variance is explained by three groups' environmental covariates, with 40.7%, 38.9%, and 45.4% respectively. In winter, the largest variance is explained by meteorological and terrain factors (39.1%). From spring to winter, the explanatory ability of the three groups of covariates is always the highest in autumn, and meteorological parameters, surface properties, and terrain factors reach the lowest in summer, winter, and spring respectively.

Table 2. Three groups of environmental covariates for AOD variation partitioning

Variance proportion	Spring	Summer	Autumn	Winter	Average
Meteorological parameters	78.8%	70.4%	80.5%	74.8%	77.1%
Surface properties	44.5%	37.9%	52.5%	31.4%	50.4%
Terrain factor	48.7%	49.5%	62.6%	62.8%	59.1%
Residual	20.1%	27.5%	19.2%	13.4%	21.8%

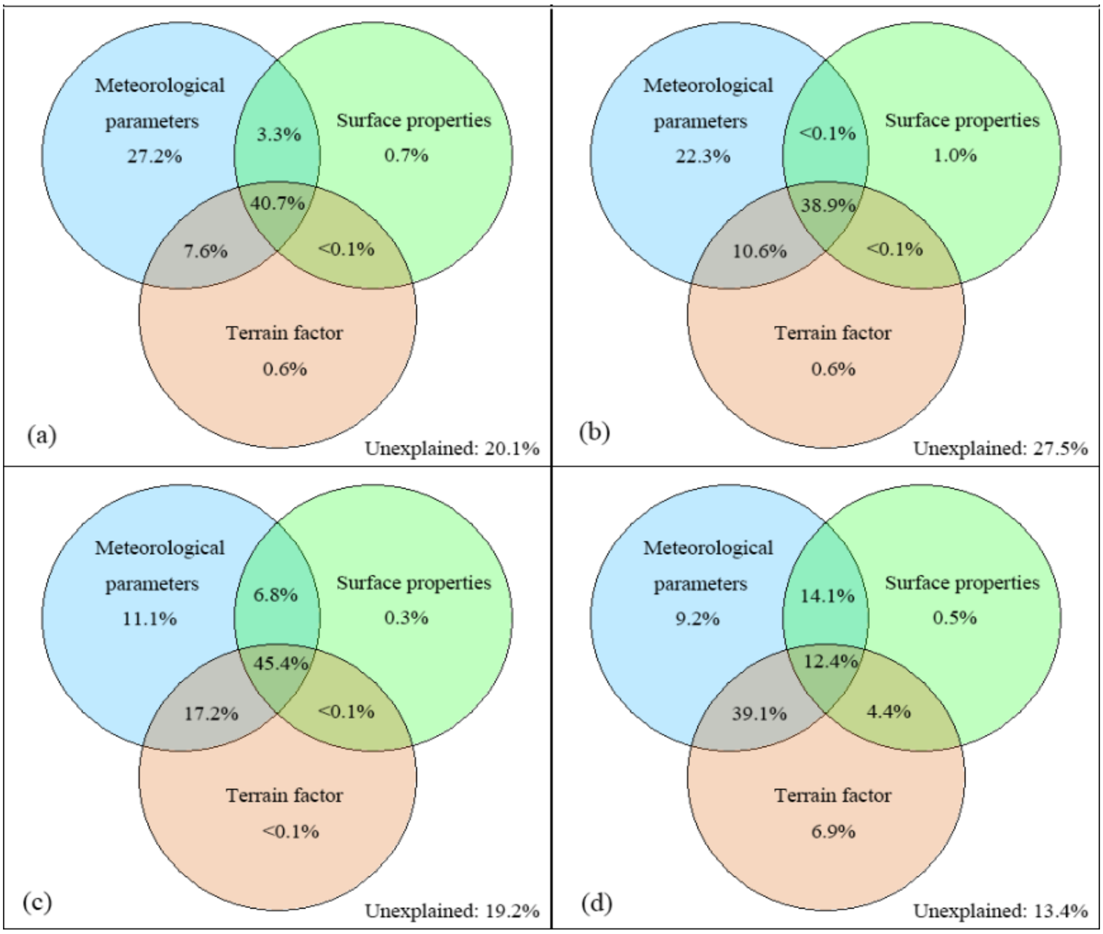

Figure 11. Variation partitioning for seasons and average AOD explained by (a) spring; (b) summer; (c) autumn. (d) winter.

4 Discussion

4.1 Model uncertainty

This study, based on MAIAC AOD and 12 environmental covariates data, adopting bagging trees ensemble approaches, produces monthly advanced-performance, full-coverage, and high-resolution FEC AOD in northwest China. The bagging trees ensemble approach has a strong advantage in characteristics modeling and prediction, but also there exists some problems, for example, most of the base learners are black box, which means the explanation is limited (Zounemat-Kermani et al., 2021). Concurrently, the model uncertainty that is also an issue to be considered possibly arises from the setting of hyperparameters and base learner and sample number selection (Khaledian and Miller, 2020). Therefore, the model robustness is critical to modeling and predicting. Simultaneously, providing mapping uncertainty helps users better understand the quality of FEC AOD in different regions, which further promotes users' reasonable use of AOD products. To check the reliability and stability of the AOD simulated model and consider the computing efficiency simultaneously, one month's data were randomly selected (August 2010), and we conducted 100 times 10-fold cross-validation, that is, 100 times prediction for each pixel, and the final prediction result is the average of 100 times (Rodriguez et al., 2010; Wei et al., 2021; Zhang et al., 2021; Ma et al., 2022). Then, we calculate model uncertainty, specifically, through the standard deviation, upper and lower limits 95% confidence interval to realize (Text S1). From 100 experiments, the validated R^2 still remains at 0.90, and the RMSE and MAE range in 0.058 - 0.057 and 0.0319 - 0.0317 respectively. Concurrently, the case average and uncertainty results are shown in Figure 12. The FEC AOD mainly concentrates on the range 0 - 0.6, accounting for 96.2%, and the maximum distribution is 0.1 - 0.2 (36.8%). The uncertainty mainly concentrates on the range 0.2 - 0.6, accounting for 80.0%, and the maximum distribution is 0.4 - 0.5 (38.1%). We also calculated the average uncertainty corresponding to different levels of FEC AOD (Figure 13). The uncertainty is lower than 0.5, accounting for 77.3% of the region, and the lowest uncertainty (0.3) corresponds to the largest proportion of FEC AOD (0.1 - 0.2). With the AOD increasing, the uncertainty also remains on rise, in other words, the high AOD areas often feature high uncertainty.

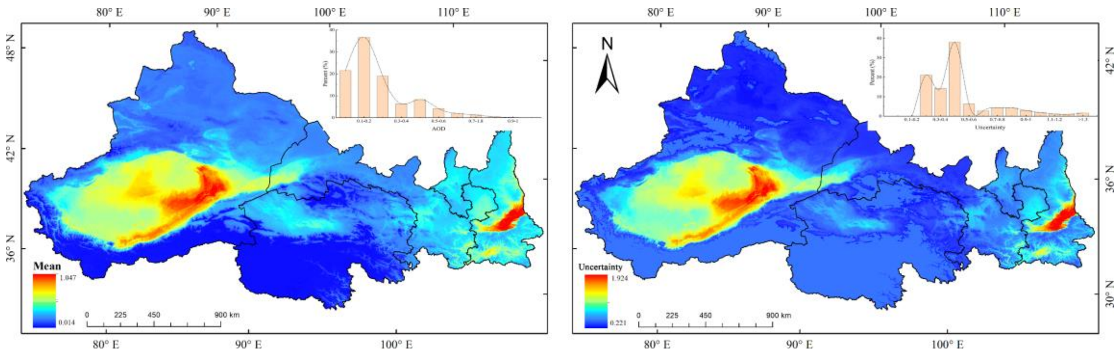

Figure 12. Distribution of mean and uncertainty in the prediction model of AOD.

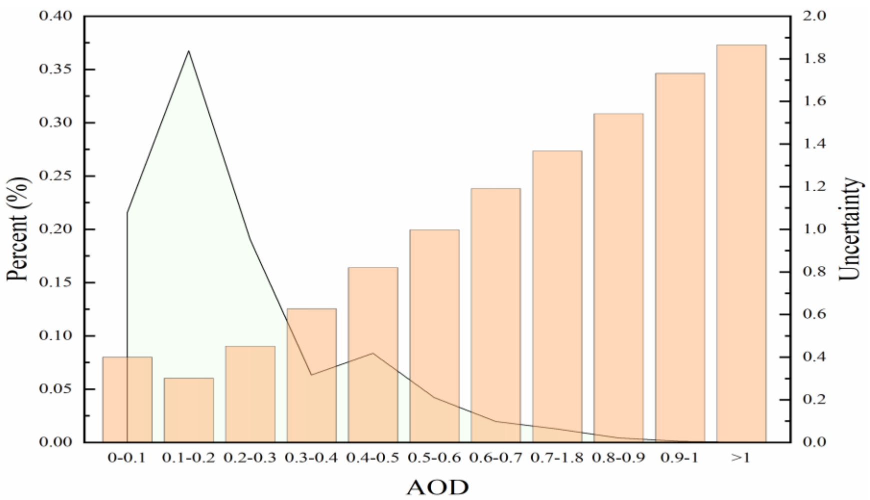

Figure 13. The average uncertainty corresponding to different levels of AOD. The light-colored area surrounded by black lines is the AOD percentage, and the histogram is the uncertainty.

4.2 AOD as affected by environmental covariates

The bagging trees ensemble method performance generally is affected by the selection of environmental covariates (Khaledian and Miller, 2020). Prediction accuracy is dependent on input variables, static variables are underpinning, and meteorological factors (dynamics variables) explain most of the variation in AOD (Yan et al., 2022). Despite our selection of 12 environmental covariates that can explain most AOD variation, there are always about 13.4% - 27.5% that can not be well explained, and there are differences in the interpretation of environmental covariates. Therefore, there is much space for improvement in the optimization of environmental covariates. There is no doubt that the meteorological parameter is the most significant contributor because of the temperature, precipitation, evapotranspiration, and wind speed through direct or indirect interaction to effectively influence AOD in the air (Chen et al., 2020). At the same time, the effect of terrain factors can not be ignored, which affects the propagation, diffusion, and settlement of AOD. The surface factors through the surface cover and soil wetness affect dust generation and reduction. However, there are also some questions that need further research, such as surface properties performance to explain AOD in summer lower spring, and the terrain factors having a higher AOD variance analytical power in autumn and winter compared with spring and summer. It is preliminarily speculated that this may be related to multi-factor interaction, which needs further analysis. In the following research, we consider introducing more related environmental covariates to try to improve prediction accuracy. In addition, we plan to further explore the internal correlation between various covariates and the relative contribution of individual covariates to AOD. Of course, the high spatial resolution and accuracy of environmental covariates are also necessary to take into consideration (add or replace).

4.3 FEC AOD for local information characterize over complex underlying surface

The spatial heterogeneity, as the 2nd law of geography, is the scale effect source. As the result, the richness of feature information varies in accordance with spatial scales in remote sensing data, and in most cases certain patterns are only found on specific scales (Miller et al., 2015). The complex underlying surfaces are often accompanied by strong spatial heterogeneity and scale effect, which brings a great challenge to high spatial resolution remote sensing observation and product generation. In this research, FEC AOD, which is generated by the way in which MAIAC AOD is constrained by combining dynamic and static variables, is well consistent with MAAC AOD on the whole. Specially, the monthly correlations are all above 0.78 in the study area, and most of these are higher than 0.9 (N = 240, Rmean = 0.928, P < 0.001, Figure S3). In addition, the FEC AOD also is evidenced to be reliable in the fine resolution long-term trend capture on a single ecosystem. However, the performance of FEC AOD on complex surfaces needs further exploration. Two typical cities (Urumqi and Lanzhou) and two months (April and October) are randomly selected to analyze the FEC AOD applicability in a complex underlying surface, and Shaybak District and Chengguan District are randomly selected for magnification in Urumqi and Lanzhou cities respectively (Figure 14). Obviously, MOD04L2 and MYD04L2 AOD products are not suitable for local air quality research, because it is difficult to characterize the detailed features of AOD due to the coarse spatial resolution and too many nodata values. However, there are also some evident differences between FEC AOD and MAIAC AOD, especially in April 2010 over the southeast of Urumqi. To this end, we have quantitatively analyzed the difference between FEC AOD and MAIAC AOD in April 2010 over Urumqi (Figure S9). FEC AOD and MAIAC AOD are close in the northwest (±0.05, close to the magnitude of one standard deviation), while are obviously different in the southeast. Accordingly, we have carefully compared multiple AOD products in April 2010 over Urumqi to try to find the reasons for the evident difference and determined its rationality. From Figure 15, we have found significant heterogeneity in some areas, and the portrayal of local AOD features vary from product to product, for example, FEC, MERRA-2, MERIS, MOD04L2, and MOD08 AOD show high value in the southeast of Urumqi. Therefore, we think the main reasons for the evident difference between FEC AOD and MAIAC AOD in the southeast of Urumqi may be as follows:(1) Limitations of the algorithm. The MAIAC algorithm assumes that the surface state is stable over a short period of time, resulting in a large number of high AOD records not being detected in MAIAC AOD (Lyapustin et al., 2018; Lyapustin et al., 2011). Certainly, our model and the selection of environmental covariates also introduce some uncertainty, which has been systematically discussed above; (2) Scale effect and spatial heterogeneity. As we all know, scale effects are common phenomena in remote sensing, which are inevitable and hard to eliminate. Once scale effects overlaying spatial heterogeneity, it may be difficult to process for the AOD retrieval algorithm under the existing technology level. In this situation, most modes may have fuzzed and smoothed the AOD extremum and so have not well captured the local information. Despite the significant differences in April 2010 over the southeast of Urumqi, we found that FEC AOD still has a good ability to capture long-term trends in Urumqi (Figure S10-S11). The FEC AOD and MAIAC AOD has a close relationship in Midong District ($R = 0.811$) and Dabancheng District, and Shaybak District ($R = 0.620$) has the lowest relationship. In summary, the evident differences between FEC AOD and MAIAC AOD in some highly heterogeneous areas are objective and reasonable in some way, but there is still much research to be done to say which AOD products are more reliable in the local feature portrayal.

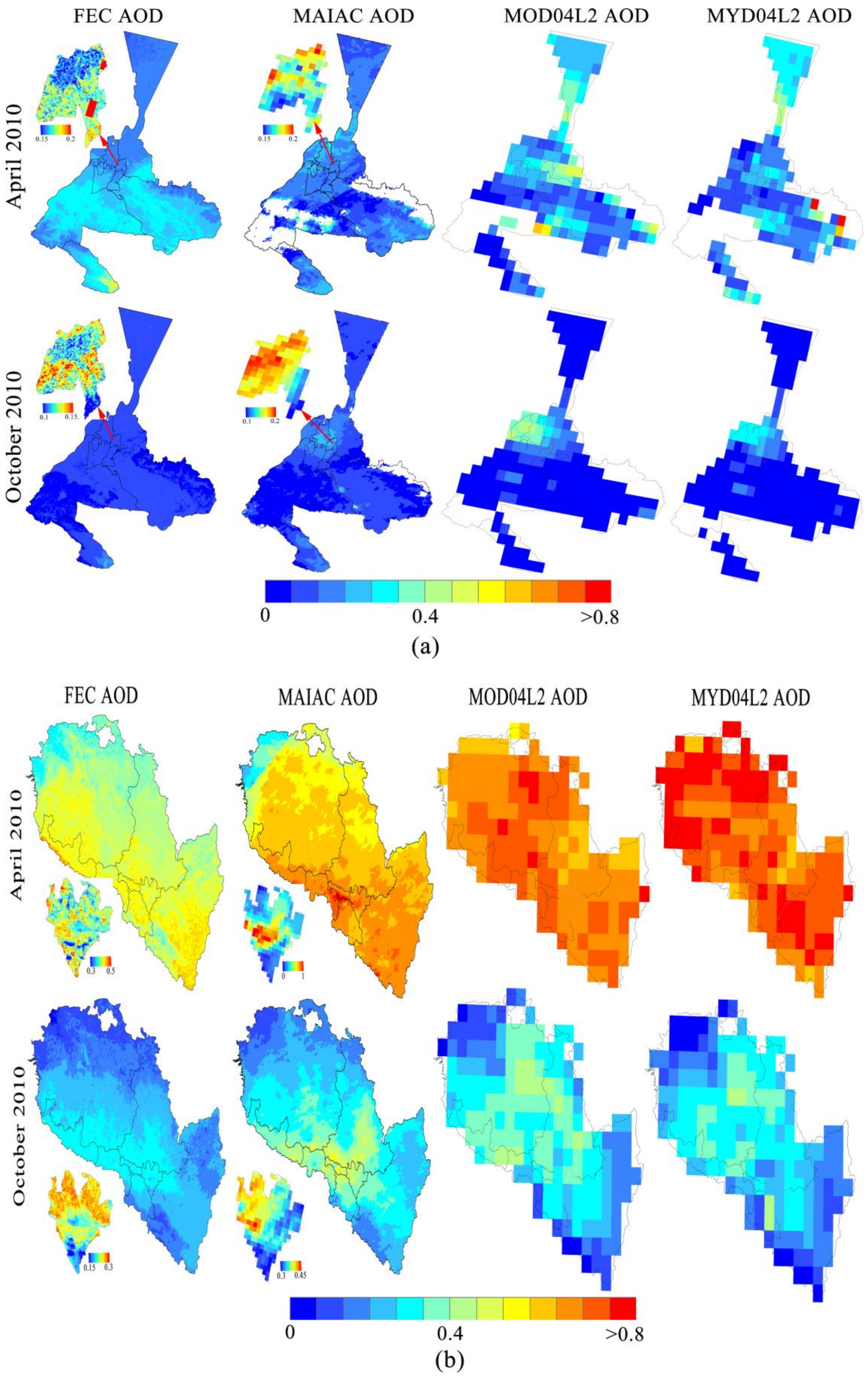

Figure 14. The spatial patterns of four AOD products in April and October 2010. (a): Urumqi, (b) Lanzhou.

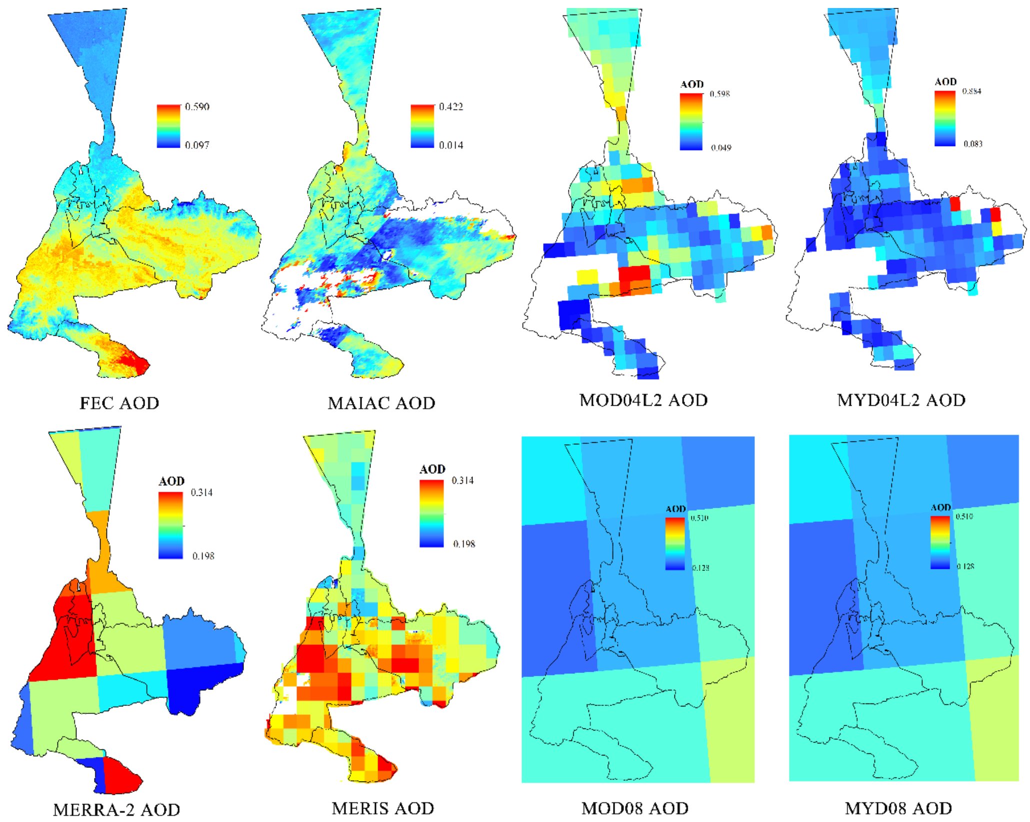

Figure 15. The spatial patterns between FEC and other AOD products in April 2010 over Urumqi.

5 Data availability

This monthly advanced-performance, full-coverage, high-resolution AOD dataset (FEC AOD) over northwest China is freely available via https://doi.org/10.5281/zenodo.5727119(Chen et al., 2021a).

6 Conclusion

In this paper, the monthly advanced-performance, full-coverage, high-resolution AOD dataset, based on MAIAC AOD and multiple environmental covariates, and utilizing a machine learning method, is produced from 2000 to 2019 in the northwest region of China. AERONET and MODIS AOD data were collected to verify the applicability of FEC AOD. Then, the FEC AOD spatiotemporal change is analyzed and the interpretation of environmental covariates to FEC AOD is explored. The result shows that the FEC AOD effectively compensates for the deficiency and constraints of in-situ observation and satellite AOD products. Meanwhile, FEC AOD products demonstrate a reliable performance and ability to capture local information, even superior to MAIAC and MxD04L2 AOD products, which has also indicated the necessity of the high spatial resolution of AOD data. The spatial patterns are consistent among different years and greatly differ at the seasonal level. The higher the AOD is, the stronger the time variability. The AOD shows a dramatic decrease in Loess Plateau and an evident increase in the southeast Taklimakan Desert between 2000 and 2019. The farmland and construction land are at high AOD levels in comparison with other land cover types. The meteorological factors demonstrate a maximum interpretation of AOD on all set temporal scales, and the capability of the environmental covariates for the explained AOD varies with season.

Author contribution: Xiangyue Chen designed and developed the methodology and software, conducted analysis and validation, and wrote the paper. Hongchao Zuo supported and supervised the study. Zipeng Zhang developed the methodology and reviewed the paper. Xiaoyi Cao and Jikai Duan made investigation and developed methodology. Chuanmei Zhu and Zhe Zhang made conceptualization and investigation. Jingzhe Wang supported and supervised the study and reviewed the paper.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgments: This work was jointly supported by the Second Tibetan Plateau Scientific Expedition and Research Program (STEP) (Grant No. 2019QZKK0103), Basic Research Program of Shenzhen (20220811173316001), Guangdong Basic and Applied Basic Research Foundation (No. 2020A1515111142) and Key Laboratory of Spatial Data Mining & Information Sharing of Ministry of Education, Fuzhou University (No.2022LSDMIS05). We are grateful to the Atmosphere Archive and Distribution System (https://search.earthdata.nasa.gov) and AERONET (http://aeronet.gsfc.nasa.gov) for providing much data support for our research.

References

Ali, G., Bao, Y., Boiyo, R., Tang, W., Lu, Q., and Min, J.: Evaluating MODIS and MISR aerosol optical depth retrievals over environmentally distinct sites in Pakistan, Journal of Atmospheric and Solar-Terrestrial Physics, 183, 19-35, https://doi.org/10.1016/j.jastp.2018.12.008, 2019.

Ali, M. A. and Assiri, M.: Analysis of AOD from MODIS-Merged DT–DB Products Over the Arabian Peninsula, Earth Systems and Environment, 3, 625-636, https://doi.org/10.1007/s41748-019-00108-x, 2019.

Almazroui, M.: A comparison study between AOD data from MODIS deep blue collections 51 and 06 and from AERONET over Saudi Arabia, Atmospheric Research, 225, 88-95, https://doi.org/10.1016/j.atmosres.2019.03.040, 2019.

Ångström, A.: The parameters of atmospheric turbidity, Tellus, 16, 64-75, https://doi.org/10.1016/0038-092X(65)90225-2, 1964.

Bilal, M., Nichol, J. E., and Wang, L.: New customized methods for improvement of the MODIS C6 Dark Target and Deep Blue merged aerosol product, Remote Sensing of Environment, 197, 115-124, https://doi.org/10.1016/j.rse.2017.05.028, 2017.

Breiman, L.: Bagging predictors, Machine Learning, 24, 123-140, https://doi.org/10.1007/BF00058655, 1996.

Chen, B., Song, Z., Pan, F., and Huang, Y.: Obtaining vertical distribution of $PM_{2.5}$ from CALIOP data and machine learning algorithms, Science of The Total Environment, 805, 150338, https://doi.org/10.1016/j.scitotenv.2021.150338, 2022.

Chen, X., Ding, J., Liu, J., Wang, J., Ge, X., Wang, R., and Zuo, H.: Validation and comparison of high-resolution MAIAC aerosol products over Central Asia, Atmospheric Environment, 251, 118273, https://doi.org/10.1016/j.atmosenv.2021.118273, 2021b.

Chen, X., Ding, J., Wang, J., Ge, X., Raxidin, M., Liang, J., Chen, X., Zhang, Z., Cao, X., and Ding, Y.: Retrieval of Fine-Resolution Aerosol Optical Depth (AOD) in Semiarid Urban Areas Using Landsat Data: A Case Study in Urumqi, NW China, Remote Sensing, 12, 467, https://doi.org/10.3390/rs12030467, 2020.

Chen, X., Zuo, H., Zhang, Z., Cao, X., Duan, J., Wang, J., Zhu, C., Zhang, Z.: High-resolution and full coverage AOD downscaling based on the bagging model over the arid and semi-arid areas, NW China [Data set], https://doi.org/10.5281/zenodo.5727119, 2021a.

De Sousa, L., Poggio, L., Batjes, N., Heuvelink, G., Kempen, B., Riberio, E., and Rossiter, D.: SoilGrids 2.0: producing quality-assessed soil information for the globe, https://doi.org/10.5194/soil-2020-65, 2020.

Ding, H. and Xingming, H.: Spatiotemporal change and drivers analysis of desertification in the arid region of northwest China based on geographic detector, Environmental Challenges, 4, 100082, https://doi.org/10.1016/j.envc.2021.100082, 2021.

Duveiller, G., Filipponi, F., Walther, S., Köhler, P., Frankenberg, C., Guanter, L., and Cescatti, A.: A spatially downscaled sun-induced fluorescence global product for enhanced monitoring of vegetation productivity, Earth Syst. Sci. Data, 12, 1101–1116, https://doi.org/10.5194/essd-12-1101-2020, 2020

Ebrahimi, N., Soofi, E. S., and Soyer, R.: Information Measures in Perspective, International Statistical Review, 78, 383-412, https://doi.org/10.1111/j.1751-5823.2010.00105.x, 2010.

Fan, W., Qin, K., Cui, Y., Li, D., and Bilal, M.: Estimation of Hourly Ground-Level $PM_{2.5}$ Concentration Based on Himawari-8 Apparent Reflectance, IEEE Transactions on Geoscience and Remote Sensing, 59, 76-85, https://doi.org/10.1109/TGRS.2020.2990791, 2020.

Feng, F. and Wang, K.: Merging ground-based sunshine duration observations with satellite cloud and aerosol retrievals to produce high-resolution long-term surface solar radiation over China, Earth System Science Data, 13, 907–922, https://doi.org/10.5194/essd-13-907-2021, 2021.

Ge, Y., Abuduwaili, J., Ma, L., Wu, N., and Liu, D.: Potential transport pathways of dust emanating from the playa of Ebinur Lake, Xinjiang, in arid northwest China, Atmospheric Research, 178-179, 196-206, https://doi.org/10.1016/j.atmosres.2016.04.002, 2016.

Giles, D. M., Sinyuk, A., Sorokin, M. G., Schafer, J. S., Smirnov, A., Slutsker, I., Eck, T. F., Holben, B. N., Lewis, J. R., Campbell, J. R., Welton, E. J., Korkin, S. V., and Lyapustin, A. I.: Advancements in the Aerosol Robotic Network (AERONET) Version 3 database – automated near-real-time quality control algorithm with improved cloud screening for Sun photometer aerosol optical depth (AOD) measurements, Atmospheric Measurement Techniques, 12, 169-209, https://doi.org/10.5194/amt-12-169-2019, 2019.

Goldberg, D. L., Gupta, P., Wang, K., Jena, C., Zhang, Y., Lu, Z., and Streets, D. G.: Using gap-filled MAIAC AOD and WRF-Chem to estimate daily $PM_{2.5}$ concentrations at 1 km resolution in the Eastern United States, Atmospheric Environment, 199, 443-452, https://doi.org/10.1016/j.atmosenv.2018.11.049, 2019.

González, S., García, S., Del Ser, J., Rokach, L., and Herrera, F.: A practical tutorial on bagging and boosting based ensembles for machine learning: Algorithms, software tools, performance study, practical perspectives and opportunities, Information Fusion, 64, 205-237, https://doi.org/10.1016/j.inffus.2020.07.007, 2020.

He, Q., Gu, Y., and Zhang, M.: Spatiotemporal trends of $PM_{2.5}$ concentrations in central China from 2003 to 2018 based on MAIAC-derived high-resolution data, Environment International, 137, 105536, https://doi.org/10.1016/j.envint.2020.105536, 2020.

Hengl, T., de Jesus, J. M., Heuvelink, G. B. M., Gonzalez, M. R., Kilibarda, M., Blagotic, A., Shangguan, W., Wright, M. N., Geng, X., Bauer-Marschallinger, B., Guevara, M. A., Vargas, R., MacMillan, R. A., Batjes, N. H., Leenaars, J. G. B., Ribeiro, E., Wheeler, I., Mantel, S., and Kempen, B.: SoilGrids250m: Global gridded soil information based on machine learning, Plos One, 12, e0169748, https://doi.org/10.1371/journal.pone.0169748, 2017.

Holben, B. N., Eck, T. F., Slutsker, I., Tanre, D., Buis, J., Setzer, A., Vermote, E., Reagan, J. A., Kaufman, Y., and Nakajima, T.: AERONET—A federated instrument network and data archive for aerosol characterization, Remote sensing of environment, 66, 1-16, https://doi.org/10.1016/S0034-4257(98)00031-5, 1998.

Huang, J., Yu, H., Dai, A., Wei, Y., and Kang, L.: Drylands face potential threat under 2 C global warming target, Nature Climate Change, 7, 417-422, https://doi.org/10.1038/nclimate3275, 2017.

Jasiewicz, J. and Stepinski, T. F.: Geomorphons — a pattern recognition approach to classification and mapping of landforms, Geomorphology, 182, 147-156, https://doi.org/10.1016/j.geomorph.2012.11.005, 2013.

Kaufman, Y. J., Tanré, D., and Boucher, O.: A satellite view of aerosols in the climate system, Nature, 419, 215, https://doi.org/10.1038/nature01091, 2002.

Khaledian, Y. and Miller, B. A.: Selecting appropriate machine learning methods for digital soil mapping, Applied Mathematical Modelling, 81, 401-418, https://doi.org/10.1016/j.apm.2019.12.016, 2020.

Lelieveld, J., Klingmüller, K., Pozzer, A., Burnett, R. T., Haines, A., and Ramanathan, V.: Effects of fossil fuel and total anthropogenic emission removal on public health and climate, Proceedings of the National Academy of Sciences, 116, 7192-7197, https://doi.org/10.1073/pnas.1819989116, 2019.

Levy, R. C., Remer, L. A., Kleidman, R. G., Mattoo, S., Ichoku, C., Kahn, R., and Eck, T. F.: Global evaluation of the Collection 5 MODIS dark-target aerosol products over land, Atmospheric Chemistry and Physics, 10, 10399-10420, https://doi.org/10.5194/acp-10-10399-2010, 2010.

Li, K., Bai, K., Ma, M., Guo, J., Li, Z., Wang, G., and Chang, N.-B.: Spatially gap free analysis of aerosol type grids in China: First retrieval via satellite remote sensing and big data analytics, ISPRS Journal of Photogrammetry and Remote Sensing, 193, 45-59, https://doi.org/10.1016/j.isprsjprs.2022.09.001, 2022.

Li, L., Franklin, M., Girguis, M., Lurmann, F., Wu, J., Pavlovic, N., Breton, C., Gilliland, F., and Habre, R.: Spatiotemporal imputation of MAIAC AOD using deep learning with downscaling, Remote Sensing of Environment, 237, 111584, https://doi.org/10.1016/j.rse.2019.111584, 2020.

Li, L., Lurmann, F., Habre, R., Urman, R., Rappaport, E., Ritz, B., Chen, J. C., Gilliland, F., and Wu, J.: Constrained Mixed-Effect Models with Ensemble Learning for Prediction of Nitrogen Oxides Concentrations at High Spatiotemporal Resolution, Environmental Science & Technology, 51, 9920-9929, https://doi.org/10.1021/acs.est.7b01864, 2017.

Li, L., Zhang, J., Meng, X., Fang, Y., Ge, Y., Wang, J., Wang, C., Wu, J., and Kan, H.: Estimation of $PM_{2.5}$ concentrations at a high spatiotemporal resolution using constrained mixed-effect bagging models with MAIAC aerosol optical depth, Remote Sensing of Environment, 217, 573-586, https://doi.org/10.1016/j.rse.2018.09.001, 2018.

Liang, T., Sun, L., and Li, H.: MODIS aerosol optical depth retrieval based on random forest approach, Remote Sensing Letters, 12, 179-189, https://doi.org/10.1080/2150704X.2020.1842540, 2021.

Lyapustin, A., Wang, Y., Korkin, S., and Huang, D.: MODIS Collection 6 MAIAC algorithm, Atmospheric Measurement Techniques, 11, 5741-5765, https://doi.org/10.5194/amt-11-5741-2018, 2018.

Lyapustin, A., Wang, Y., Laszlo, I., Kahn, R., Korkin, S., Remer, L., Levy, R., and Reid, J. S.: Multiangle implementation of atmospheric correction (MAIAC): 2. Aerosol algorithm, Journal of Geophysical Research Atmospheres, 116, 0148-0227, https://doi.org/10.1029/2010JD014986, 2011.

Ma, R., Ban, J., Wang, Q., Zhang, Y., Yang, Y., Li, S., Shi, W., Zhou, Z., Zang, J., and Li, T.: Full-coverage 1 km daily ambient $PM_{2.5}$ and O_3 concentrations of China in 2005–2017 based on a multi-variable random forest model, Earth System Science Data, 14, 943–954, https://doi.org/10.5194/essd-14-943-2022, 2022.

Ma, Z., Shi, Z., Zhou, Y., Xu, J., Yu, W., and Yang, Y.: A spatial data mining algorithm for downscaling TMPA 3B43 V7 data over the Qinghai–Tibet Plateau with the effects of systematic anomalies removed, Remote Sensing of Environment, 200, 378-395, https://doi.org/10.1016/j.rse.2017.08.023, 2017.

Miller, B.A., Koszinski, S., Wehrhan, M., and Sommer, M.: Impact of multi-scale predictor selection for modeling soil properties, Geoderma, 239-240, 97-106, https://doi.org/10.1016/j.geoderma.2014.09.018, 2015.

Myhre, G., Samset, B. H., Schulz, M., Balkanski, Y., Bauer, S., Berntsen, T. K., Bian, H., Bellouin, N., Chin, M., and Diehl, T.: Radiative forcing of the direct aerosol effect from AeroCom Phase II simulations, Atmospheric Chemistry and Physics, 13, 1853, https://doi.org/10.5194/acp-13-1853-2013, 2013.

Rodriguez, J. D., Perez, A., and Lozano, J. A.: Sensitivity Analysis of k-Fold Cross Validation in Prediction Error Estimation, IEEE Transactions on Pattern Analysis and Machine Intelligence, 32, 569-575, https://doi.org 10.1109/TPAMI.2009.187, 2010.

Singh, M. K., Venkatachalam, P., and Gautam, R.: Geostatistical Methods for Filling Gaps in Level-3 Monthly-Mean Aerosol Optical Depth Data from Multi-Angle Imaging SpectroRadiometer, Aerosol and Air Quality Research, 17, 1963-1974, https://doi.org/10.4209/aaqr.2016.02.0084, 2017.

Sun, J., Gong, J., and Zhou, J.: Estimating hourly $PM_{2.5}$ concentrations in Beijing with satellite aerosol optical depth and a random forest approach, Science of The Total Environment, 762, 144502, https://doi.org/10.1016/j.scitotenv.2020.144502, 2021.

Sun, W., Song, X., Mu, X., Gao, P., Wang, F., and Zhao, G.: Spatiotemporal vegetation cover variations associated with climate change and ecological restoration in the Loess Plateau, Agricultural and Forest Meteorology, 209-210, 87-99, https://doi.org/10.1016/j.agrformet.2015.05.002, 2015.

Szilagyi, J., Yinsheng, Z., Ning, M., and Wenbin, L.: Terrestrial evapotranspiration dataset across China (1982-2017), National Tibetan Plateau Data Center [dataset], https://doi.org/10.11888/AtmosPhys.tpe.249493.file, 2019.

Tao, M., Chen, L., Wang, Z., Wang, J., Che, H., Xu, X., Wang, W., Tao, J., Zhu, H., and Hou, C.: Evaluation of MODIS Deep Blue Aerosol Algorithm in Desert Region of East Asia: Ground Validation and Intercomparison, Journal of Geophysical Research: Atmospheres, 122, 10,357-310,368, https://doi.org/10.1002/2017JD026976, 2017.

Waits, A., Emelyanova, A., Oksanen, A., Abass, K., and Rautio, A.: Human infectious diseases and the changing climate in the Arctic, Environment International, 121, 703-713, https://doi.org/10.1016/j.envint.2018.09.042, 2018.

Wang, Z., Deng, R., Ma, P., Zhang, Y., Liang, Y., Chen, H., Zhao, S., and Chen, L.: 250-m Aerosol Retrieval from FY-3 Satellite in Guangzhou, Remote Sensing, 13, 920, https://doi.org/10.3390/rs13050920, 2021.

Wei, J., Peng, Y., Guo, J., and Sun, L.: Performance of MODIS Collection 6.1 Level 3 aerosol products in spatial-temporal variations over land, Atmospheric Environment, 206, 30-44, https://doi.org/10.1016/j.atmosenv.2019.03.001, 2019.

Wei, J., Li, Z., Lyapustin, A., Sun, L., Peng, Y., Xue, W., Su, T., and Cribb, M.: Reconstructing 1-km-resolution high-quality $PM_{2.5}$ data records from 2000 to 2018 in China: spatiotemporal variations and policy implications, Remote Sensing of Environment, 252, 112136, https://doi.org/10.1016/j.rse.2020.112136, 2021.

Wei, X., Bai, K., Chang, N.-B., and Gao, W.: Multi-source hierarchical data fusion for high-resolution AOD mapping in a forest fire event, International Journal of Applied Earth Observation and Geoinformation, 102, 102366, https://doi.org/10.1016/j.jag.2021.102366, 2021.

Xiao, Q., Wang, Y., Chang, H. H., Meng, X., Geng, G., Lyapustin, A., and Liu, Y.: Full-coverage high-resolution daily $PM_{2.5}$ estimation using MAIAC AOD in the Yangtze River Delta of China, Remote Sensing of Environment, 199, 437-446, https://doi.org/10.1016/j.rse.2017.07.023, 2017.

Xue, W., Wei, J., Zhang, J., Sun, L., Che, Y., Yuan, M., and Hu, X.: Inferring Near-Surface $PM_{2.5}$

Concentrations from the VIIRS Deep Blue Aerosol Product in China: A Spatiotemporally Weighted Random Forest Model, Remote Sensing, 13, 505, https://doi.org/10.3390/rs13030505, 2021.

Yan, X., Zang, Z., Li, Z., Luo, N., Zuo, C., Jiang, Y., Li, D., Guo, Y., Zhao, W., Shi, W., and Cribb, M.: A global land aerosol fine-mode fraction dataset (2001–2020) retrieved from MODIS using hybrid physical and deep learning approaches, Earth System Science Data, 14, 1193–1213, https://doi.org/10.5194/essd-14-1193-2022, 2022.

Yang, J. and Hu, M.: Filling the missing data gaps of daily MODIS AOD using spatiotemporal interpolation, Science of The Total Environment, 633, 677-683, https://doi.org/10.1016/j.scitotenv.2018.03.202, 2018.

Yang, Q., Yuan, Q., Li, T., and Yue, L.: Mapping PM2.5 concentration at high resolution using a cascade random forest based downscaling model: Evaluation and application, Journal of Cleaner Production, 277, 123887, https://doi.org/10.1016/j.jclepro.2020.123887, 2020.

Zhang, R., Di, B., Luo, Y., Deng, X., Grieneisen, M. L., Wang, Z., Yao, G., and Zhan, Y.: A nonparametric approach to filling gaps in satellite-retrieved aerosol optical depth for estimating ambient $PM_{2.5}$ levels, Environmental Pollution, 243, 998-1007, https://doi.org/10.1016/j.envpol.2018.09.052, 2018.

Zhang, Z., Wu, W., Fan, M., Wei, J., Tan, Y., and Wang, Q.: Evaluation of MAIAC aerosol retrievals over China, Atmospheric Environment, 202, 8-16, https://doi.org/10.1016/j.atmosenv.2019.01.013, 2019.

Zhang, Z., Ding, J., Zhu, C., Chen, X., Wang, J., Han, L., Ma, X., and Xu, D.: Bivariate empirical mode decomposition of the spatial variation in the soil organic matter content: A case study from NW China, Catena, 206, 105572, https://doi.org/10.1016/j.catena.2021.105572, 2021.

Zhao, C., Liu, Z., Wang, Q., Ban, J., Chen, N. X., and Li, T.: High-resolution daily AOD estimated to full coverage using the random forest model approach in the Beijing-Tianjin-Hebei region, Atmospheric Environment, 203, 70-78, https://doi.org/10.1016/j.atmosenv.2019.01.045, 2019.

Zhao, H., Gui, K., Ma, Y., Wang, Y., Wang, Y., Wang, H., Zheng, Y., Li, L., Zhang, L., Che, H., and Zhang, X.: Climatological variations in aerosol optical depth and aerosol type identification in Liaoning of Northeast China based on MODIS data from 2002 to 2019, Science of The Total Environment, 781, 146810, https://doi.org/10.1016/j.scitotenv.2021.146810, 2021.

Zounemat-Kermani, M., Batelaan, O., Fadaee, M., and Hinkelmann, R.: Ensemble machine learning paradigms in hydrology: A review, Journal of Hydrology, 598, 126266, https://doi.org/10.1016/j.jhydrol.2021.126266, 2021.