https://doi.org/10.5194/essd-2021-426
Preprint. Discussion started: 7 April 2022

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], Jingzhe Wang [3], Chuanmei Zhu [2], Zhe Zhang [2]

1 College of Atmospheric Sciences, Lanzhou University, Lanzhou, 730000 China
2 Key Laboratory of Oasis Ecology, Xinjiang University, Xinjiang Urumqi 830046, China
3 MNR Key Laboratory for Geo-Environmental Monitoring of Great Bay Area & Guangdong Key Laboratory of Urban Informatics & Shenzhen Key Laboratory of Spatial Smart Sensing and Services, Shenzhen University, Shenzhen 518060, China

*Correspondence: zuohch@lzu.edu.cn; Tel.: 13119413436

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. The high AOD areas are located in the Taklimakan Desert and the Loess Plateau, and the low AOD area is 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

https://doi.org/10.5194/essd-2021-426
Preprint. Discussion started: 7 April 2022

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) and the smallest in autumn (0.147); the AOD pattern in Gansu province is bimodal, but unimodal in other provinces. 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 an 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-

https://doi.org/10.5194/essd-2021-426
Preprint. Discussion started: 7 April 2022

related research(Goldberg et al., 2019; He et al., 2020).

Generally, the capital 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 a perfect method to acquire AOD, 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 different sensor sources have different emphases in use (Tab.S1). Yet, the common point is that spatial resolution is coarse, and even has a large number of pixel values missing(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 an urban 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). The initial and most extensive method is interpolations, but the AOD has high spatiotemporal variability, thus applying the approach to anticipating AOD missing data isn't very suitable(Singh et al., 2017). Another widely used method is merging multiple AOD products, which can often improve data quality but always not completely eliminate pixel values missing phenomenon, even bringing offsetting consequences(Bilal et al., 2017; Ali and Assiri, 2019). 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, these models'

https://doi.org/10.5194/essd-2021-426
Preprint. Discussion started: 7 April 2022

performance are always doubtful 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). As computing technology advances, ensemble machine learning methods 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.

https://doi.org/10.5194/essd-2021-426
Preprint. Discussion started: 7 April 2022

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 MxD08 AOD satellite products were collected to verify the accuracy of FEC AOD; (3) the FEC AOD spatiotemporal change is analyzed; (4) the dominant environmental covariates of FEC AOD are explored.

2 Materials and methods

2.1 Study area

Fig.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 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 the 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.

https://doi.org/10.5194/essd-2021-426
Preprint. Discussion started: 7 April 2022

Open Access Earth System Science Data Discussions

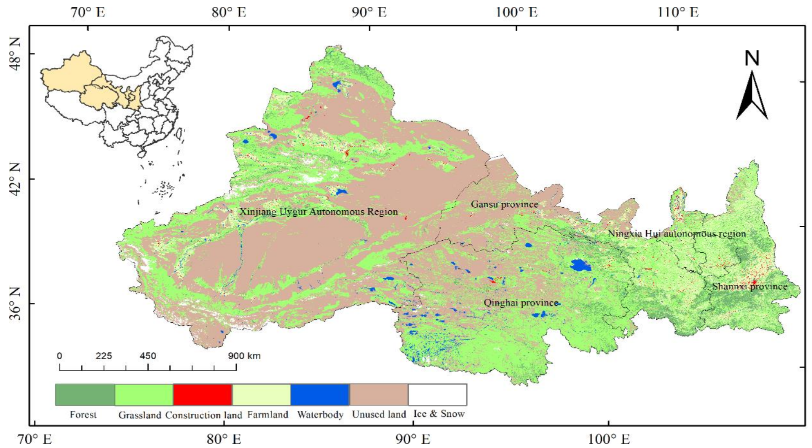


Figure 1 Study area. The figure shows typical arid and semi-arid areas, five provinces in northwest China.

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. Based on the python tool, we preprocessed the data and computed the daily average AOD by combining the 550 nm AOD data from Terra and Aqua.

https://doi.org/10.5194/essd-2021-426
Preprint. Discussion started: 7 April 2022

Open Access Earth System Science Data Discussions

2.3 MODIS MxD08_M3 data

MYD08_M3 and MOD08_M3 are the level 3 atmosphere global products from Aqua and Terra respectively, where spatial and temporal resolutions are 1° × 1° and 1 month respectively. Based on the MODIS Collection 6.1, we chose 550 nm combined land and ocean mean band AOD to validate FEC AOD. It is worth noting that the Aqua and Terra launch time is different, so we can acquire MOD08_M3 data from March 2000 to February 2020, but as for MYD08_M3, we only acquire data from July 2002 to February 2020. All processes are realized through the Google Earth Engine (GEE) cloud computing platform (Also, the data is available independently from https://ladsweb.modaps.eosdis.nasa.gov/). GEE has a multi-petabyte database of an extensive and varied number of reanalysis and satellite imagery data collected over several decades, which allows for the quick and effective processing of numerous years of satellite data using an easy-to-use online cloud computing platform(Hu et al., 2021; Gorelick et al., 2017).

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). 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 (Tab.S2).

https://doi.org/10.5194/essd-2021-426
Preprint. Discussion started: 7 April 2022

2.5 Environmental covariates

Environmental covariates selected in this study contain 12 covariates in three categories (meteorology, surface information, and topography). 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 Tab.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 a 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. As for dynamic covariates, the monthly average method is adopted to obtain the multi-year average data. It is worth pointing out 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 $0.1^{\circ} \times 0.1^{\circ}$ respectively(Szilagyi et al., 2019). To ET data, we use a downscaling algorithm proposed by Ma (2017) to transform

https://doi.org/10.5194/essd-2021-426
Preprint. Discussion started: 7 April 2022

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. 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). 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.

https://doi.org/10.5194/essd-2021-426
Preprint. Discussion started: 7 April 2022

Table 1 Environmental covariates for AOD modeling

Type	Name	Abbreviation	Resolution	Sources
Dynamic covariate				
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 covariate				
Surface properties	Land use and land cover	LUCC	30 m × 30 m	http://www.resdc.cn/
Terrain factor	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	

https://doi.org/10.5194/essd-2021-426
Preprint. Discussion started: 7 April 2022

Open Access Earth System Science Data Discussions

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 (Fig.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

https://doi.org/10.5194/essd-2021-426
Preprint. Discussion started: 7 April 2022

et al., 2021).

In this study, we use 12 environmental covariates (1 km) as bagging trees ensemble algorithms input to acquire AOD environmental covariates model in 1 km, and utilize the model and 250 m environmental covariates to acquire FEC AOD. 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 (Fig.2). Definitely, all modeling and application processes can also be implemented in R or Python.

https://doi.org/10.5194/essd-2021-426
Preprint. Discussion started: 7 April 2022

Open Access Earth System Science Data Discussions

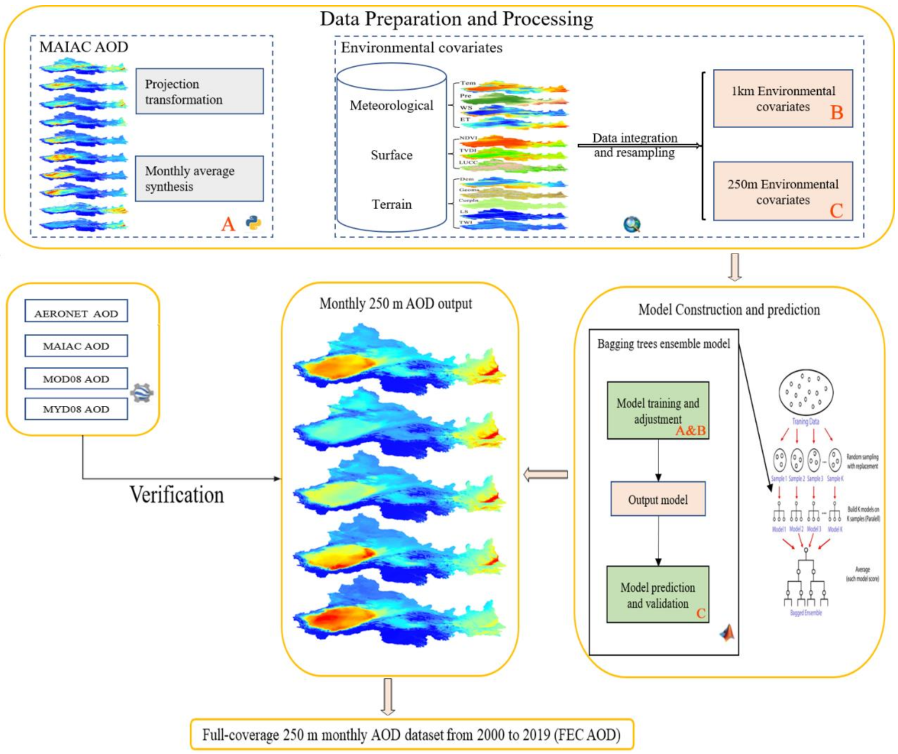


Figure 2 Flow chart of experiment and model calculation process.

3 Results and analysis

3.1 Accuracy validation based on in-situ and satellite

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 model 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 month average AOD at each AERONET site in the whole study region. Specifically, we check data

https://doi.org/10.5194/essd-2021-426
Preprint. Discussion started: 7 April 2022

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 Fig.3, FEC AOD was highly correlated with AERONET monthly AOD ($R^2 = 0.787$), with MAE of 0.049 and RMSE of 0.061. Approximately 64.5% 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. Compared with previous studies, the FEC AOD has better accuracy than MAIAC AOD and MxD08 AOD products(Chen et al., 2021b; Wei et al., 2019).

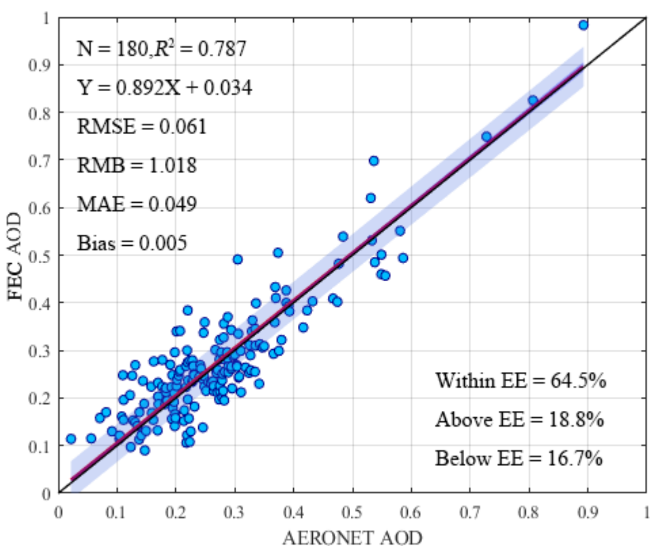


Figure 3 Comparison between the FEC AOD and AERONET AOD. The red line is the regression line, the black dashed line is the 1:1 line, and the blue area indicates the prediction interval.

The multi-year average AOD spatial distribution was calculated (Fig.4). The AOD spatial pattern has high consistency, and the high AOD is located in Taklimakan Desert and Loess Plateau, and the low AOD is distributed in high altitude areas (such as

https://doi.org/10.5194/essd-2021-426
Preprint. Discussion started: 7 April 2022

Open Access Earth System Science Data Discussions

mountain zone and Qinghai). As for MxD08 AOD, the direct feeling is coarse spatial resolution, with lots of missing data. To further explore the improvement of FEC AOD based on MAIAC AOD, two typical cities in arid and semi-arid areas are selected to analyze the use on an urban scale. From Fig. S2 and S3, we can easily find the difference in different AOD satellite products. Obviously, MOD08 and MYD08 AOD products are not suitable for urban air quality research. We randomly select Shaybak and Chengguan districts for magnification in Urumqi and Lanzhou cities respectively. Compared with MAIAC AOD, the FEC AOD has a strong potential to describe local AOD features or fine AOD distribution. Concurrently, the multi-year monthly average of four AOD products (FEC AOD, MAIAC AOD, MOD08 AOD, and MYD08 AOD) was counted (Fig.S4). From January to December, the four AOD products show a trend of increasing first and decreasing next, reaching the lowest value in November. Of course, there are some differences in the AOD magnitude and fluctuation range, which are mainly due to the difference in AOD retrieval algorithms.

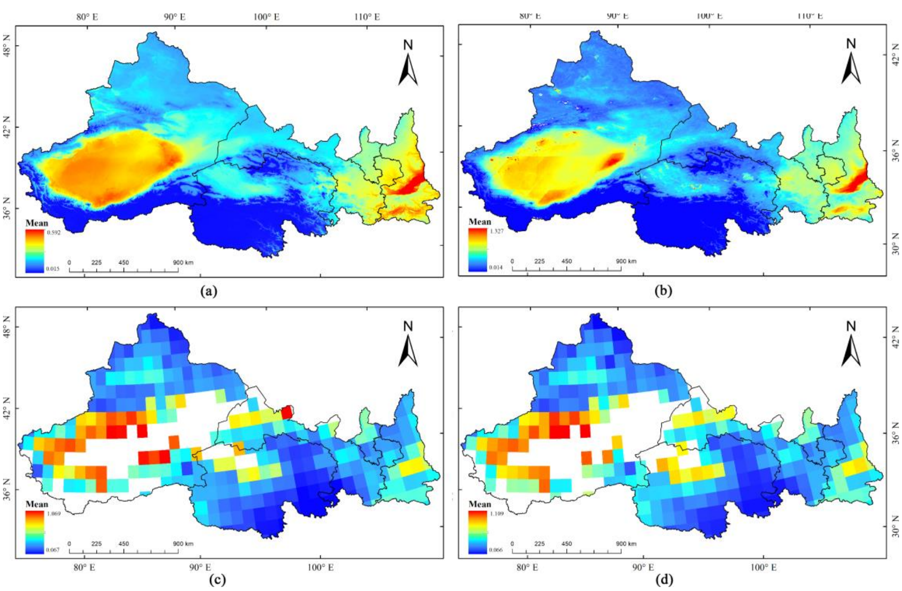


Figure 4 The muti-years spatial average AOD for (a) FEC AOD, (b) MAIAC AOD, (c) MOD08 AOD, and (d) MYD08 AOD.

https://doi.org/10.5194/essd-2021-426
Preprint. Discussion started: 7 April 2022

The seasonal-trend decomposition procedure based on loess (STL) is used in time-series decomposition for four AOD products to further analyze the consistency and difference in time scale (Fig.5). STL decomposes the time series data into additive variation three components: trend, seasonal and remainder(Chen et al., 2021b). Firstly, the four AOD data change in a similar manner, fluctuating and slightly decreasing, and the MxD08 AOD fluctuation range is significantly higher than that of FEC AOD and MAIAC AOD. As for seasonal characteristics, the four AOD products feature significant seasonal cycle variation. The spring and summer AOD remain at a high level, and winter always is the lowest. Then, moving to the general trend after the season effect is removed, the four AOD products show a tortuous rise at first, beginning to decline around 2012, and rebounding about 2017. In summary, we can draw the decision that FEC AOD products demonstrate a reliable accuracy and ability to capture local information, even superior to MAIAC and MxD08 AOD products.

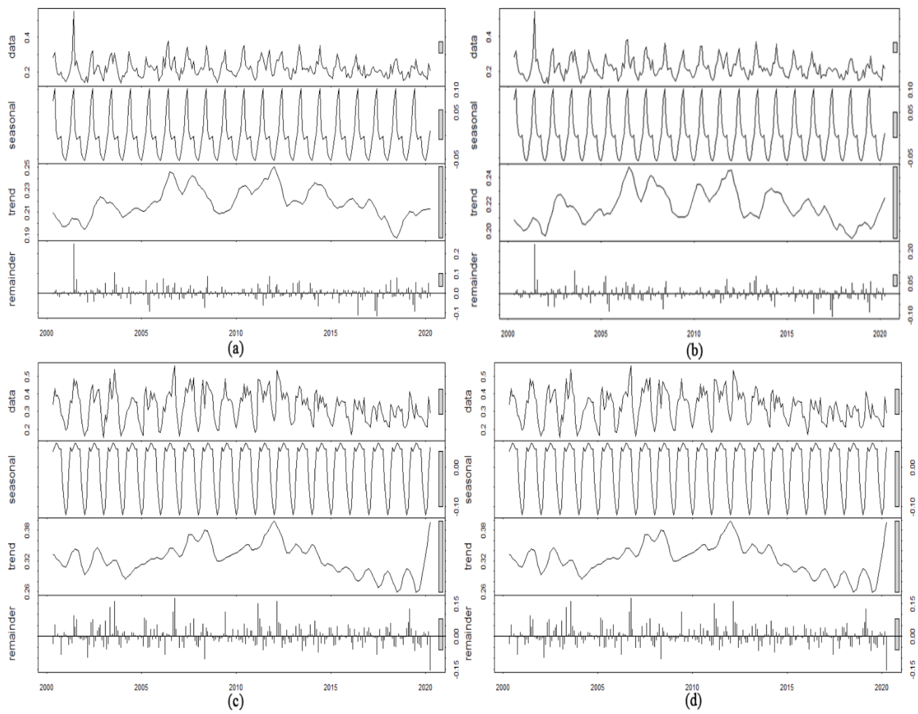


Figure 5 Seasonal and trend decomposition using loess for (a) FEC AOD, (b) MAIAC AOD, (c) MOD08 AOD, and (d) MYD08 AOD.

https://doi.org/10.5194/essd-2021-426
Preprint. Discussion started: 7 April 2022

Open Access Earth System Science Data Discussions

3.2 Spatiotemporal pattern of FEC AOD from 2000 to 2019

Fig.6 shows annual mean FEC AOD maps for each year from 2000 to 2019 and multi-year mean AOD map. In general, spatial patterns are consistent over different years, where the highest AOD are found in the south of 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. Fig.7 illustrates the spatial distributions of seasonal mean AOD from 2000 to 2019. The spatial patterns of AOD greatly differ at the seasonal level. 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. The higher the AOD level is, the stronger the fluctuation of AOD.

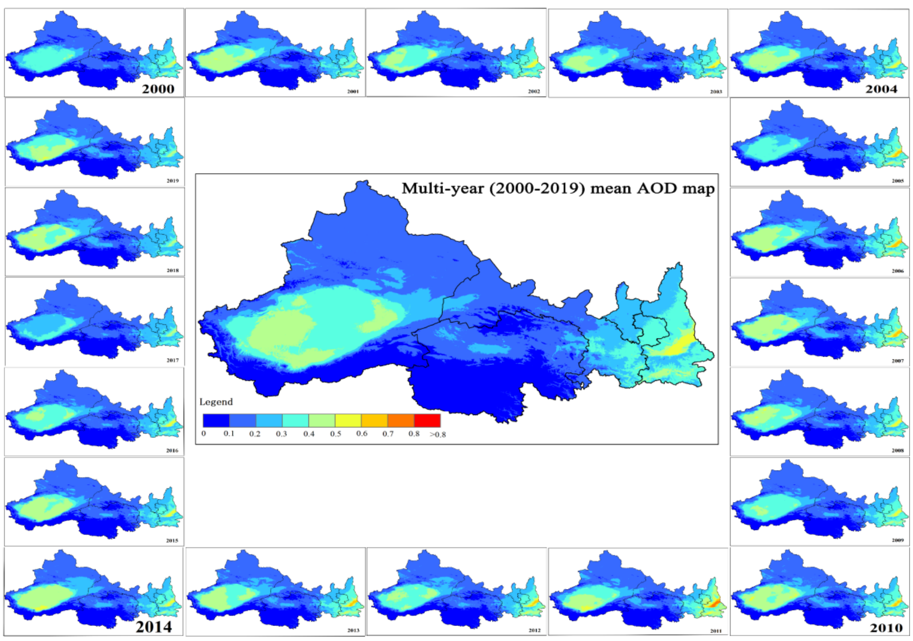


Figure 6 FEC AOD annual mean maps for each year from 2000 to 2019.

https://doi.org/10.5194/essd-2021-426
Preprint. Discussion started: 7 April 2022

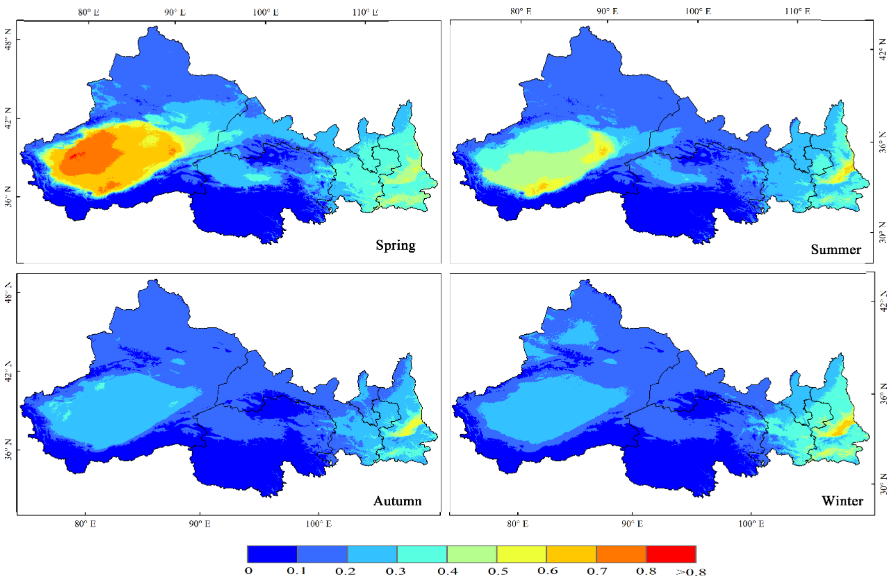


Figure 7 FEC AOD seasonal mean maps averaged over the period 2000-2019.

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. Fig.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 19 years is large, especially in the south of Xinjiang(The Taklimakan Desert) and Shannxi province(The 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 is. The AOD in Xinjiang is increasing,

https://doi.org/10.5194/essd-2021-426
Preprint. Discussion started: 7 April 2022

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, Ningxia, and southeastern Gansu provinces. A possible reason for this finding is that the Loess Plateau is experiencing greening, and the vegetation keeps increasing under artificial intervention.

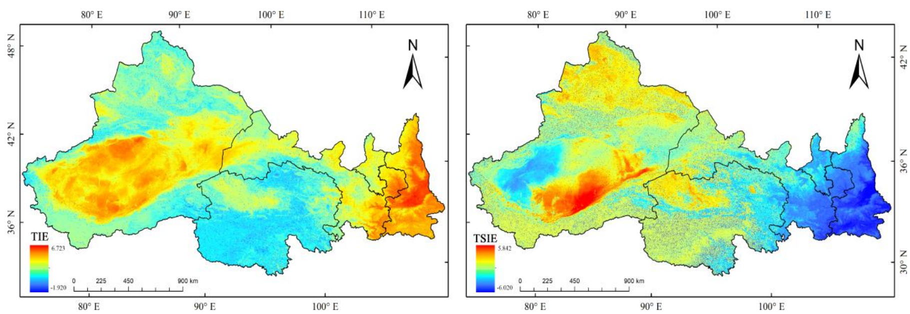


Figure 8 TIE and TSIE of AOD distribution.

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 the 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. Fig.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 worth noting that except for the Gansu province, where AOD is bimodal, other provinces are unimodal. Fig.10 describes the AOD season 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

https://doi.org/10.5194/essd-2021-426
Preprint. Discussion started: 7 April 2022

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 seasons distribution, with decrease and then increase, and autumn is the bottom.

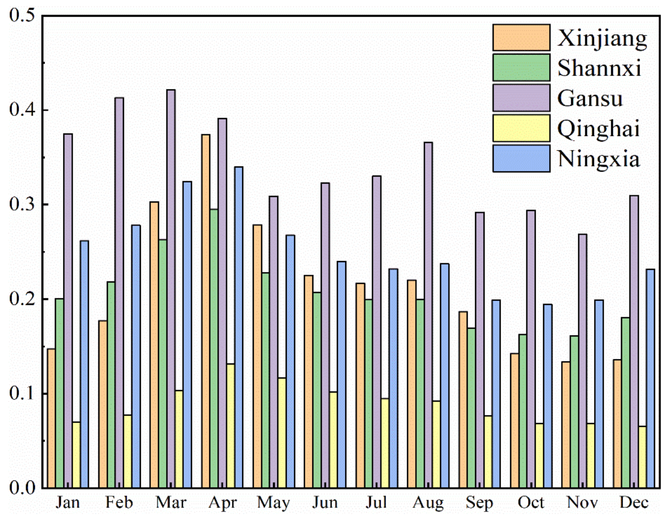


Figure 9 The monthly distribution characteristics of AOD in different provinces.

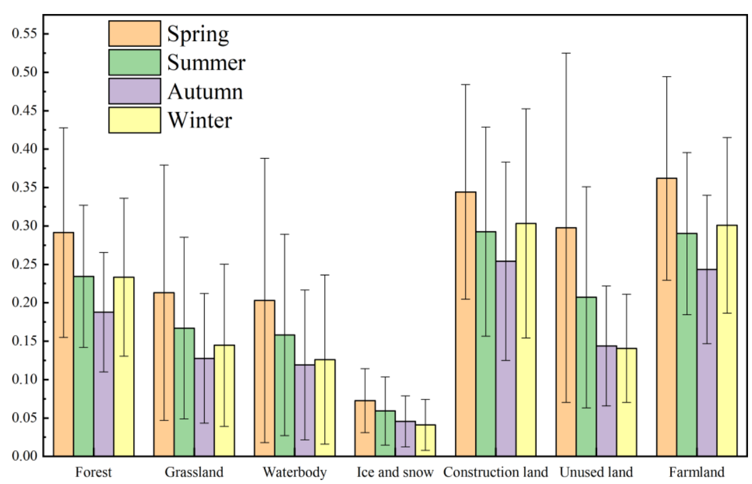


Figure 10 AOD season distribution under different land cover types.

https://doi.org/10.5194/essd-2021-426
Preprint. Discussion started: 7 April 2022

3.3 Variation partitioning of FEC AOD

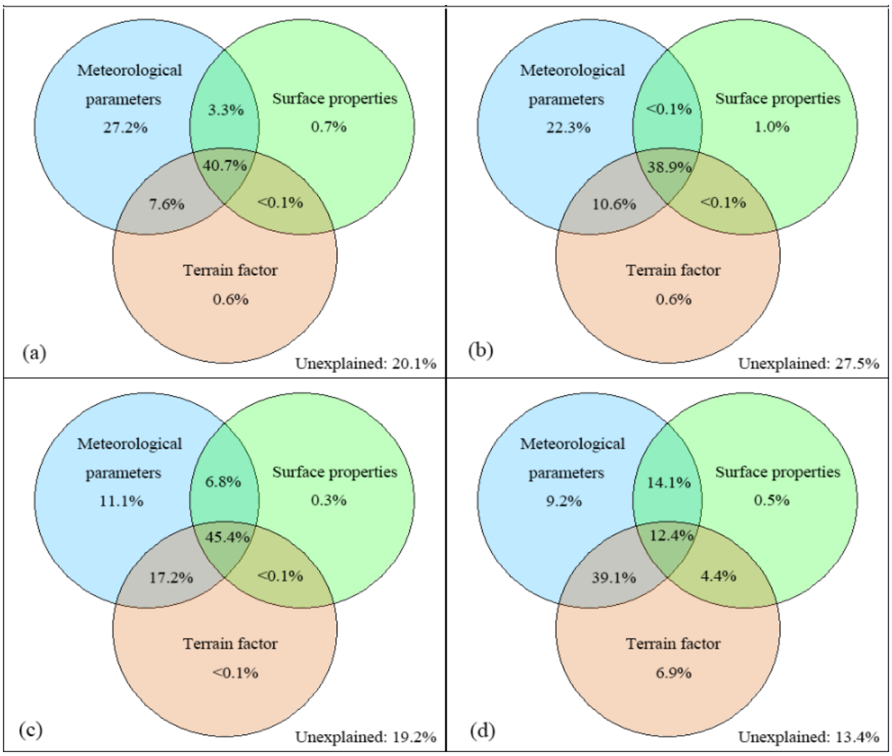
To examine the contribution of environmental covariates to the FEC AOD dynamic, the redundancy analysis (RDA) was used to explore the association between different seasons of FEC AOD and the environmental covariates. The twelve environmental covariates were divided into three groups, meteorology, surface properties, and terrain. The variance proportion driving the variation of FEC AOD on different temporal scales was tested from the environmental covariates 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 Tab.2 and Fig.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.

https://doi.org/10.5194/essd-2021-426
Preprint. Discussion started: 7 April 2022

457 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%

458



459

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

462

463

464

465

https://doi.org/10.5194/essd-2021-426
Preprint. Discussion started: 7 April 2022

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 a 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 of 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. 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 Fig.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 (Fig.13). The uncertainty is lower than 0.5, accounting for 77.3% of the region, and the lowest uncertainty (0.3)

https://doi.org/10.5194/essd-2021-426
Preprint. Discussion started: 7 April 2022

corresponds to the largest proportion of FEC AOD (0.1 - 0.2). With the AOD increasing, the uncertainty also remains on rise, that is to say, the high AOD areas often feature high uncertainty.

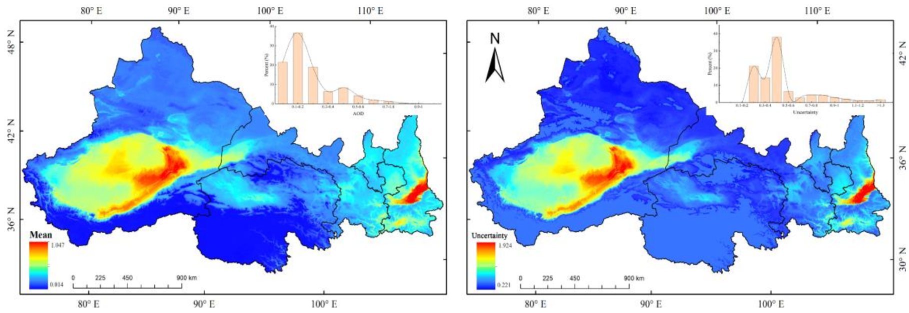


Figure 12 Distribution of mean and uncertainty in the prediction model of AOD.

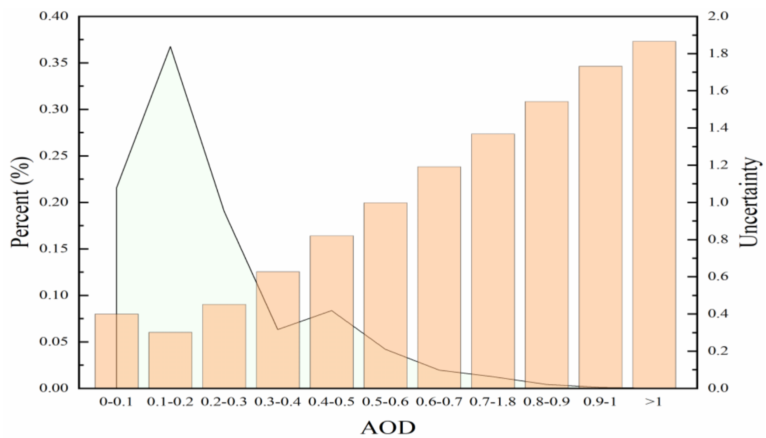


Figure 13 The average uncertainty corresponding to different levels of AOD. The light-colored area surrounded by black lines is the AOD percentage, and the histogram is the uncertainty.

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). Despite our selection of 12 environmental covariates that can explain most AOD variation, there are

https://doi.org/10.5194/essd-2021-426
Preprint. Discussion started: 7 April 2022

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 have 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).

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

https://doi.org/10.5194/essd-2021-426
Preprint. Discussion started: 7 April 2022

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 accuracy 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 accuracy and ability to capture local information, even superior to MAIAC and MxD08 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. Jingzhe Wang supported and supervised the study and reviewed the paper. Chuanmei Zhu and Zhe Zhang made conceptualization and investigation.

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), Guangdong Basic and Applied Basic Research Foundation (No. 2020A1515111142) and China Postdoctoral Science Foundation (No. 2020M672776). 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.

https://doi.org/10.5194/essd-2021-426
Preprint. Discussion started: 7 April 2022

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.

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.

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.
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.
Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., and Moore, R.: Google Earth Engine: Planetary-scale geospatial analysis for everyone, Remote Sensing of Environment, 202, 18-27, https://doi.org/10.1016/j.rse.2017.06.031, 2017.
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.
Hu, T., Myers Toman, E., Chen, G., Shao, G., Zhou, Y., Li, Y., Zhao, K., and Feng, Y.: Mapping fine-scale human disturbances in a working landscape with Landsat time series on Google Earth Engine, ISPRS Journal of Photogrammetry and Remote Sensing, 176, 250-261, https://doi.org/10.1016/j.isprsjprs.2021.04.008, 2021.
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, 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, ps://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.
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,

https://doi.org/10.5194/essd-2021-426
Preprint. Discussion started: 7 April 2022

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.

Xiao, Q., Wang, Y., Chang, H. H., Meng, X., Geng, G., Lyapustin, A., and Liu, Y.: Full-coverage high-resolution daily PM2.5 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.

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.

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.

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.