


## Article

# Bayesian Aerosol Retrieval-Based PM<sub>2.5</sub> Estimation through Hierarchical Gaussian Process Models

Junbo Zhang <sup>1</sup>, Daoji Li <sup>2</sup>, Yingzhi Xia <sup>1</sup> and Qifeng Liao <sup>1,\*</sup> <sup>1</sup> School of Information Science and Technology, ShanghaiTech University, Shanghai 201210, China<sup>2</sup> Department of Information Systems and Decision Sciences, California State University, Fullerton, CA 92831, USA

\* Correspondence: liaoqf@shanghaitech.edu.cn

**Abstract:** Satellite-based aerosol optical depth (AOD) data are widely used to estimate land surface PM<sub>2.5</sub> concentrations in areas not covered by ground PM<sub>2.5</sub> monitoring stations. However, AOD data obtained from satellites are typically at coarse spatial resolutions, limiting their applications on small or medium scales. In this paper, we propose a new two-step approach to estimate 1-km-resolution PM<sub>2.5</sub> concentrations in Shanghai using high spatial resolution AOD retrievals from MODIS. In the first step, AOD data are refined to a 1 × 1 km<sup>2</sup> resolution via a Bayesian AOD retrieval method. In the second step, a hierarchical Gaussian process model is used to estimate PM<sub>2.5</sub> concentrations. We evaluate our approach by model fitting and out-of-sample cross-validation. Our results show that the proposed approach enjoys accurate predictive performance in estimating PM<sub>2.5</sub> concentrations.

**Keywords:** Bayesian retrieval algorithm; PM<sub>2.5</sub>; hierarchical Gaussian process model; MAIAC

**MSC:** 62M20; 62J05; 62F15



**Citation:** Zhang, J.; Li, D.; Xia, Y.; Liao, Q. Bayesian Aerosol Retrieval-Based PM<sub>2.5</sub> Estimation through Hierarchical Gaussian Process Models. *Mathematics* **2022**, *10*, 2878. <https://doi.org/10.3390/math10162878>

Academic Editor: Jiancang Zhuang

Received: 4 July 2022

Accepted: 10 August 2022

Published: 11 August 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Particulate matter with an aerodynamic diameter of less than 2.5 μm (PM<sub>2.5</sub>) can affect human health (see [1] for detailed discussions). Exposure to PM<sub>2.5</sub> over a few hours to weeks can trigger cardiovascular disease-related mortality; longer-term exposure (e.g., a few years) increases the risk for cardiovascular mortality and ischaemic heart disease and even reduces life expectancy [2–5]. In recent years, the public has regarded PM<sub>2.5</sub> as one of the primary air pollutants, especially after the Chinese government proposed the National Ambient Air Quality Standard for PM<sub>2.5</sub> and began to establish ground-based PM<sub>2.5</sub> monitor networks in 2012 [6]. Although the ground-based PM<sub>2.5</sub> monitor networks can accurately measure the land surface PM<sub>2.5</sub> concentrations, the number of monitoring sites is limited.

To assess PM<sub>2.5</sub> concentrations in large areas, satellite-based products such as the aerosol optical depth (AOD) are commonly used due to their broad coverage. Various statistical models have been developed to predict PM<sub>2.5</sub> concentrations using AOD data in the literature, including the geographically weighted regression model and the linear mixed-effects model [7,8]. In addition, based on a moderate resolution imaging spectroradiometer (MODIS), Gaussian process modeling in a Bayesian hierarchical setting has been used to estimate PM<sub>2.5</sub> concentrations [9]. However, AOD data obtained from satellites are typically at coarse spatial resolutions, limiting their applications on small or medium scales. To address this issue, many methods have been introduced to refine AOD data. For example, Lipponen et al. [10] proposed a Bayesian aerosol retrieval algorithm for MODIS AOD retrieval over land. Wang et al. [11] suggested a hierarchical Bayesian approach for multi-angle imaging spectroradiometer (MISR) data such that the resolution of AOD can be improved from 17.6 to 4.4 km. Based on the multi-angle implementation of the atmospheric correction (MAIAC) retrieval algorithm in [12], Wei et al. [13] developed a new space-time

random forest model to predict  $PM_{2.5}$  concentrations using MODIS MAIAC AOD and other pertinent variables related to meteorological conditions, land use, and human activities.

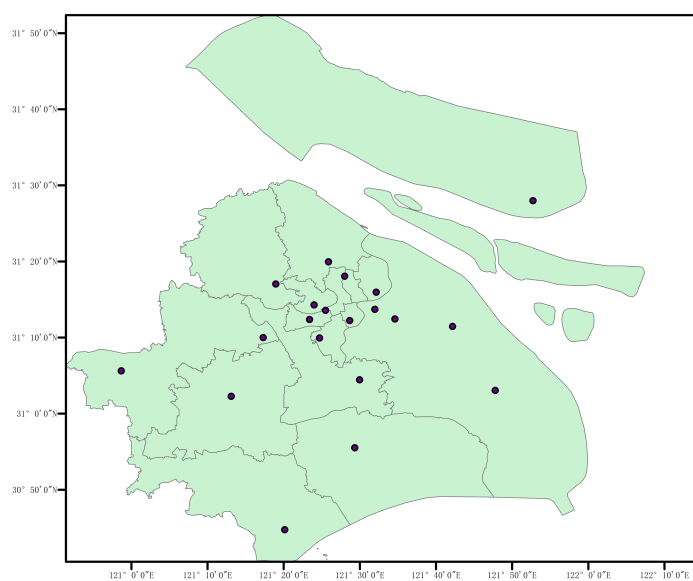
Although the space-time random forest model has good performance, one main limitation of the random forest model is that a large number of trees can make the algorithm too slow and ineffective for real-time predictions. There is a tradeoff between the training time (and space) and increased number of trees. In addition, random forest may not obtain good results for small data. When estimating  $PM_{2.5}$  1-km-resolution, it can be expected that the results from the random forest model are not good if only a few observations are available at some location. To address these challenges, in this paper, we propose a new two-step approach to predict  $PM_{2.5}$  concentrations using high spatial resolution AOD retrievals from MODIS. In the first step, AOD data are refined to a  $1 \times 1$  km<sup>2</sup> resolution via a Bayesian AOD retrieval method. In the second step, a hierarchical Gaussian process model is used to estimate  $PM_{2.5}$  concentrations. Our approach combines the strengths of recent advances in [9,10]. In addition, the studies on  $PM_{2.5}$  concentration estimates in Shanghai are limited in the literature, although Shanghai is one of the largest cities in China and a popular tourist destination. Our work tries to fill this gap.

The rest of the paper is organized as follows. Section 2 describes the study region and the datasets used in this work. Section 3 introduces the Bayesian algorithm to retrieve AOD and the Hierarchical Gaussian process model to estimate  $PM_{2.5}$  concentrations. Section 4 provides the experimental results for our approach. We discuss some limitations and extensions of our work in Section 5.

## 2. Datasets

In this study, we are interested in estimating  $PM_{2.5}$  concentrations in Shanghai. Shanghai is one of the largest cities in China, with intensive human activities and high amounts of complex aerosols in the air. However, the studies on  $PM_{2.5}$  concentration estimates in Shanghai are limited in the literature. We would like to remark that the proposed procedure in this paper can be used to estimate  $PM_{2.5}$  concentrations in any other city.

The  $PM_{2.5}$  concentration monitoring network in Shanghai consists of 20 monitoring stations across all districts of Shanghai, and these are shown as dots in Figure 1. The solid lines are the city border and district borders, while the green filling is the mainland of Shanghai. We consider the hourly ground-level  $PM_{2.5}$  concentrations in Shanghai from 1 January 2021 to 31 December 2021, which are provided by the China Environmental Monitoring Center (<http://www.cnemc.cn/>, accessed 1 February 2022).



**Figure 1.** Spatial distributions of  $PM_{2.5}$  concentrations monitoring stations in Shanghai. Solid lines: city border and district borders. Black dots:  $PM_{2.5}$  concentrations monitoring stations.

The MODIS instruments on the Earth Observation System's (EOS) Terra and Aqua satellites provide measurements of the atmosphere, land, and ocean in the visible, near-infrared, and infrared from 0.4 to 14.4  $\mu\text{m}$  of the 36 spectral bands. The instrument collects the data once or twice a day at a given location and views the entire globe every one to two days. The MODIS product *MOD04\_L2* provides initial AOD data in Shanghai with a spatial resolution of about  $1 \text{ km} \times 1 \text{ km}$  at nadir and other variables needed in the Bayesian retrieval algorithm of our procedure, while *MCD19A2* offers MAIAC AOD products. Both can be downloaded from Level-1 and the Atmosphere Archive and Distribution System Distributed Active Archive Center (<https://ladsweb.modaps.eosdis.nasa.gov/>, accessed 1 February 2022).

### 3. Methods

Now we are ready to introduce our procedure for  $\text{PM}_{2.5}$  concentrations estimation. As mentioned earlier, our procedure includes two steps. In the first step, we use the Bayesian aerosol retrieval algorithm [10] to refine the AOD data, while in the second step, the proposed hierarchical Gaussian process model estimates land surface  $\text{PM}_{2.5}$  concentrations using the refined AOD data from the first step.

#### 3.1. Bayesian Aerosol Retrieval Algorithm

The Bayesian aerosol retrieval algorithm was firstly proposed by Lipponen et al. [10] for the retrieval of AOD over land in Finland using MODIS aerosol products, including the retrieved aerosol properties and measurement data with a spatial resolution of about  $10 \times 10 \text{ km}^2$  at nadir. In the first step of our procedure, we use the Bayesian aerosol retrieval algorithm proposed to refine the AOD data in Shanghai with a spatial resolution of about  $1 \times 1 \text{ km}^2$  at nadir. The Bayesian aerosol retrieval algorithm is composed of three main parts. The first part is building a radiative transfer model called the forward model. The second part is the AOD retrieval based on the forward model, which is called the inversion part. The last part is combining the retrieved AOD with MAIAC AOD products.

##### 3.1.1. Establishing Forward Models

We first construct a radiative transfer model using MODIS products to simulate top of Atmosphere (TOA) reflectance. Then we use the simulated TOA reflectance and the TOA reflectance measured by MODIS to do the retrieval. The AOD values are the outputs of the retrieval. The radiative transfer model we used here is the same as the Dark Target (DT) algorithm [14], except that we perform the retrieval using a Bayesian algorithm. There are two versions for the DT algorithm. One is for retrieval over land, and the other one is for retrieval over the ocean. We will use the DT algorithm over land. The DT algorithm uses bright aerosols against a dark target because it can reflect so much solar radiation to the space that the MODIS sensor can capture it. The main idea of this DT algorithm is to find the aerosol properties that minimize the difference between the TOA reflectance measured by the MODIS sensor and the TOA reflectance calculated by radiative transfer simulations using the aerosol properties. We reformulate it as a Bayesian form, and more details will be given later.

In this study, the TOA reflectance is calculated as follows:

$$f(\tau, \eta, \rho^s; \gamma) = \eta \times f_{fine}(\tau, \rho^s; \gamma) + (1 - \eta) \times f_{coarse}(\tau, \rho^s; \gamma), \quad (1)$$

where  $f(\tau, \eta, \rho^s; \gamma)$  denotes the total simulated reflectance,  $f_{fine}$  represents the simulated reflectance corresponding to fine aerosol models,  $f_{coarse}$  is the simulated reflectance corresponding to coarse aerosol models,  $\tau$  denotes AOD,  $\eta$  is the fine mode fraction (FMF),  $\rho^s$  denotes the surface reflectance, and  $\gamma$  denotes the vector of all additional parameters corresponding to auxiliary variables, such as wavelength, aerosol models, solar zenith angle, view zenith angle, and relative azimuth angle in Table 1. There are five main aerosol models related to this work, including three different fine aerosol models, one coarse aerosol model, and one continental aerosol model. In the DT retrieval, the fine aerosol

model is taken from a predefined database that contains aerosol model information based on the location and the season. The TOA reflectances and other radiative transfer-related variables corresponding to each aerosol model are precomputed and stored in lookup tables (LUT) to make the algorithm computationally more efficient. We use the same LUT as the Collection 6 MODIS aerosol products in [15]. The variables in the LUT and corresponding input parameters are listed in Table 1.

**Table 1.** Input variables and parameters in LUT.

Variables	Input Parameters
AOD	From 0 to 6, step 1
Wavelength	0.466, 0.554, 0.645, 2.113
Aerosol type	0, 1, 2, 3, 4
Solar zenith angle	From 0 to 84, step 12
View zenith angle	From 0 to 84, step 12
Relative azimuth angle	From 0 to 180, step 12
Topographic altitude	From −0.2 to 9, step 0.1

We use the same aerosol models and data processing procedures (such as cloud screening) as those in the DT algorithm [14]. Therefore, we retrieve the same pixels as the DT algorithm. For each pixel, the forward model at different wavelength  $\lambda$  is as follows:

$$f_{\lambda}(\tau, \rho_{\lambda}^s; \gamma) = \rho_{\lambda}^a(\tau; \theta_0, \theta, \phi) + \frac{T_{\lambda}(\tau; \theta_0)T_{\lambda}(\tau; \theta)\rho_{\lambda}^s(\tau; \theta_0, \theta, \phi)}{1 - s_{\lambda}(\tau)\rho_{\lambda}^s(\tau; \theta_0, \theta, \phi)}, \tag{2}$$

where  $\rho_{\lambda}^a$  denotes the atmospheric path reflectance depending on AOD  $\tau$ , wavelength  $\lambda$ , solar zenith angle  $\theta_0$ , view zenith angle  $\theta$ , and relative azimuth angle  $\phi$ ,  $T_{\lambda}(\tau; \theta_0)$  and  $T_{\lambda}(\tau; \theta)$  are the downward and upward atmospheric transmissions at wavelength  $\lambda$ , respectively,  $s = s_{\lambda}$  is the atmospheric backscattering ratio at wavelength  $\lambda$ , and  $\rho_{\lambda}^s$  denotes the surface reflectance at wavelength  $\lambda$ . The simulated reflectances  $f_{fine}$  and  $f_{coarse}$  in (1) are calculated separately using forward model (2) based on the LUT.

### 3.1.2. Retrieve AOD Using Bayesian Methods

The inversion part of the Bayesian aerosol retrieval algorithm can be formulated as an optimization problem. The retrieval problem is to find the parameters that maximize the following posterior distribution

$$\pi(\tau, \eta, \rho^s; \gamma | \rho^{TOA}), \tag{3}$$

where  $\rho^{TOA}$  denotes TOA reflectances measured by the MODIS instrument. We apply the Bayes theorem to the posterior distribution, and the posterior becomes

$$\pi(\tau, \eta, \rho^s; \gamma | \rho^{TOA}) \propto \pi(\rho^{TOA} | \tau, \eta, \rho^s; \gamma)\pi(\tau, \eta, \rho^s). \tag{4}$$

The function that describes the relationship between TOA reflectance measured by the MODIS instrument and TOA reflectance simulated by radiative transfer models is

$$\rho^{TOA} = f(\tau, \eta, \rho^s; \gamma) + e, \tag{5}$$

where  $f(\tau, \eta, \rho^s; \gamma)$  is the TOA reflectance simulated by radiative transfer models in (1),  $e$  denotes the total error, including the observation noise and the approximation error in TOA reflectances due to aerosol and radiative transfer models. In our study, we assume that the random error  $e$  follows a Gaussian distribution, that is,  $e \sim \mathcal{N}(\mathbb{E}_e, \Gamma_e)$ . In this retrieval algorithm, errors mainly come from two sources, which are observation noise  $n$  and approximation error  $u$ . The observation noise is caused by the instrument, while the approximation error is caused by the variables such as geometry and aerosol models

used in the radiative transfer model. In this study, the observation noise is modeled as Gaussian zero-mean random variable, and its variances are based on MODIS aerosol product *STD\_Reflectance\_Land*. The approximation error is modeled as an additive Gaussian random variable. We use the approximation error established in [10], which runs N simulations using AOD, FMF values from the AERONET, and surface reflectance values from the MODIS MCD43C3 product to construct a database of  $(\rho^{MODIS} - \rho^{simulation})$  for all regions (AERONET stations all over the world) and month and estimates the expected value and covariance matrix using sample median and sample covariance of the approximation error model, respectively. Then, we can obtain

$$\pi(\rho^{TOA} | \tau, \eta, \rho^s; \gamma) = \pi_e(\rho^{TOA} - f(\tau, \eta, \rho^s; \gamma)). \tag{6}$$

In this study, we retrieve total AOD  $\tau$  at 0.55  $\mu\text{m}$ , fine mode fraction  $\eta$ , and land surface reflectance  $\rho^s$  at four different wavelengths: 0.47, 0.55, 0.67, and 2.1  $\mu\text{m}$ . We model AOD, FMF, and surface reflectances as mutually uncorrelated variables at all bands. We depict all unknown parameters in a granule using multivariate Gaussian prior models. The prior probability density models encode prior knowledge such as spatial correlation information, seasonal variability, or positivity constraints into the retrieval. The prior models are fully described by their expected value vectors and covariance matrices:

- $\tau \sim \mathcal{N}(\mathbb{E}_\tau, \Gamma_\tau)$ , where  $\mathbb{E}_\tau$  is the expected value vector of AOD, and  $\Gamma_\tau$  is the covariance matrix of AOD. Following [16], the nearest value from the MAC-V2 climatology is taken as the prior expectation for each pixel to be retrieved. The MAC-V2 climatology is a  $12 \times 180 \times 360$  tensor (denotes month, latitude, and longitude), and we use the nearest value to the pixel we retrieve as the prior expectation. We define the  $(i, j)$  element of the prior covariance matrix  $\Gamma_\tau$  of AOD as

$$\Gamma_\tau(i, j) = \sigma_{\tau, \text{nugget}}^2 \delta_{i,j} + \sigma_{\tau, \text{sill}}^2 \exp \left\{ -3 \left\| \frac{\mathbf{x}_i - \mathbf{x}_j}{r_{\tau, \text{range}}} \right\|^p \right\}, \tag{7}$$

where  $\delta_{i,j} = 1$  if  $i = j$  and 0 otherwise,  $\sigma_{\tau, \text{nugget}}$  is the so-called nugget, representing the local variance,  $\sigma_{\tau, \text{sill}}$  is the spatial correlation,  $\|\mathbf{x}_i - \mathbf{x}_j\|$  measures the distance between the pixel  $i$  and  $j$ ,  $r_{\tau, \text{range}}$  describes the spatial correlation length, and  $p$  is the spatial smoothness of the AOD fields. The values of these covariance matrix parameters used in (7) are listed in Table 2.

- $\eta \sim \mathcal{N}(\mathbb{E}_\eta, \Gamma_\eta)$ , where  $\mathbb{E}_\eta$  is the expected value vector of FMF, and  $\Gamma_\eta$  is the covariance matrix of FMF. Similar to the AOD, the prior expectation value for FMF is also computed from the MAC-V2 climatology. The prior covariance matrix  $\Gamma_\eta$  of FMF is the same as the covariance matrix  $\Gamma_\tau$  of AOD, except for the values of those covariance matrix parameters. See Table 2 for the values of covariance matrix parameters used in the prior model for the FMF.
- $\rho^s \sim \mathcal{N}(\mathbb{E}_{\rho^s}, \Gamma_{\rho^s})$ , where  $\mathbb{E}_{\rho^s}$  and  $\Gamma_{\rho^s}$  are the expected value vector and covariance matrix of land surface reflectance, respectively. We also use Gaussian prior models for the surface reflectances. We use the blue-sky albedos computed with the weighting coefficient of 0.5 (50% of the white-sky albedo and 50% of the black-sky albedo) in the MODIS MCD43C3 albedo product as the expected values for the surface reflectances. For the Bayesian aerosol retrieval algorithm, the monthly surface reflectance is computed as the temporal average of surface reflectances  $\pm 45$  days around the middle day of the month. The expected values for the surface reflectances in the retrieval are computed as an average of the three closest pixels in the monthly surface reflectance. Both the temporal variance in the original surface albedo product and the variance due to averaging are taken into account in the construction of the surface reflectance variance.

**Table 2.** The covariance matrix parameters used in AOD and FMF prior models.

Parameter	AOD	FMF
$r_{\tau, range}$	50 km	50 km
$\sigma_{\tau, nugget}^2$	0.0025	0.01
$\sigma_{\tau sill}^2$	0.10	0.25
$p$	1.5	1.5

Then, with the prior models above, we can rewrite (4) as

$$\begin{aligned} & \pi(\tau, \eta, \rho^s; \gamma | \rho^{TOA}) \\ & \propto \pi_e(\rho^{TOA} - f(\tau, \eta, \rho^s; \gamma)) \pi(\tau) \pi(\eta) \pi(\rho^s) \\ & \propto \exp \left\{ -\frac{1}{2} (\rho^{TOA} - f(\tau, \eta, \rho^s; \gamma) - \mathbb{E}_e)^T \Gamma_e^{-1} (\rho^{TOA} - f(\tau, \eta, \rho^s; \gamma) - \mathbb{E}_e) \right. \\ & \quad - \frac{1}{2} (\tau - \mathbb{E}_\tau)^T \Gamma_\tau^{-1} (\tau - \mathbb{E}_\tau) - \frac{1}{2} (\eta - \mathbb{E}_\eta)^T \Gamma_\eta^{-1} (\eta - \mathbb{E}_\eta) \\ & \quad \left. - \frac{1}{2} (\rho^s - \mathbb{E}_{\rho^s})^T \Gamma_{\rho^s}^{-1} (\rho^s - \mathbb{E}_{\rho^s}) \right\}. \end{aligned}$$

We simultaneously retrieve all dark land pixels in a granule. The inversion part of the Bayesian aerosol retrieval algorithm looks for the maximum a posteriori (MAP) estimate for the unknown parameters. This is equivalent to the following optimization problem

$$\begin{aligned} (\tau, \eta, \rho^s)_{MAP} = \arg \min_{\tau, \eta, \rho^s} \left\{ \left\| \mathbf{L}_e \left[ \rho^{TOA} - f(\tau, \eta, \rho^s; \gamma) - \mathbb{E}_e \right] \right\|^2 + \left\| \mathbf{L}_\tau (\tau - \mathbb{E}_\tau) \right\|^2 \right. \\ \left. + \left\| \mathbf{L}_\eta (\eta - \mathbb{E}_\eta) \right\|^2 + \left\| \mathbf{L}_{\rho^s} (\rho^s - \mathbb{E}_{\rho^s}) \right\|^2 \right\}, \end{aligned} \tag{8}$$

where  $\mathbf{L}_\tau$ ,  $\mathbf{L}_\eta$ , and  $\mathbf{L}_{\rho^s}$  are the Cholesky factors of  $\Gamma_\tau^{-1}$ ,  $\Gamma_\eta^{-1}$ , and  $\Gamma_{\rho^s}^{-1}$ , respectively. We use AOD, FMF, and the surface reflectances at all bands in MODIS products MMOD04\_L2 as the initial values for  $\tau$ ,  $\eta$ , and  $\rho^s$ , and use the L-BFGS-B optimization algorithm in [17] to find  $(\tau, \eta, \rho^s)_{MAP}$ .

### 3.1.3. Combine AOD Retrievals with MAIAC AOD Products

We create a  $0.01^\circ \times 0.01^\circ$  grid with a total of  $118 \times 137$  grid cells that cover the whole region of Shanghai. If there is an AOD retrieval in a grid cell, we fill it with the AOD retrieval. After that, we fill the empty grid cells with MAIAC AOD products. The final grid cells are the products to be used in our hierarchical Gaussian process model for estimating PM<sub>2.5</sub> concentrations in Shanghai.

### 3.2. Hierarchical Gaussian Process Model

Next, with the refined data obtained in the first step by the Bayesian aerosol retrieval algorithm above, we use the hierarchical Gaussian process model introduced in [9] to estimate PM<sub>2.5</sub> concentrations in Shanghai. To be more specific, the hierarchical Gaussian process model is given by

$$\begin{aligned} p_i &= \beta_0 + \beta_1 \tau_i + \omega_i + \epsilon_i, \\ \omega_i &\sim \mathcal{N}(0, \mathcal{K}(h; \sigma^2, \phi)), \quad \epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2), \end{aligned} \tag{9}$$

where  $p_i$  and  $\tau_i$  denote the PM<sub>2.5</sub> and AOD at location  $i$ , respectively,  $\beta_0$  and  $\beta_1$  are, respectively, the intercept and the coefficient for AOD,  $\omega_i$  is a spatial random effect for the location  $i$ ,  $\epsilon_i$  is the random error,  $\mathcal{K}$  is the covariance function,  $\sigma^2$  is a variance parameter,  $h$  is the Euclidean distance between any two spatial locations, and  $\phi$  represents the spatial decay. Here, we assume that each random effect  $\omega_i$  follows a multivariate Gaussian process with

a mean zero and covariance function of  $\mathcal{K}(h; \sigma^2, \phi)$ , and the random error  $\delta_i$  also follows a Gaussian distribution with mean zero and variance  $\sigma_\epsilon^2$ . Following [9], the covariance function  $\mathcal{K}(h; \sigma^2, \phi)$  is given by

$$\mathcal{K}(h; \sigma^2, \phi) = \begin{cases} \sigma^2 [1 - 1.5h\phi + 0.5(h\phi)^3] & \text{if } 0 < h < \phi^{-1}, \\ 0 & \text{otherwise.} \end{cases} \quad (10)$$

We also set prior distributions for parameters  $\beta_0, \beta_1, \sigma^2, \phi$ , and  $\sigma_\epsilon^2$ . To be more precise, following [9,18], we use the following prior distributions:

- $\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} 1000 & 0 \\ 0 & 1000 \end{bmatrix}\right)$
- $\sigma^2$  follows the inverse gamma distribution with shape parameter 2 and scale parameter 2.
- $\phi$  follows the uniform distribution  $\mathcal{U}[3, 100]$ .
- $\sigma_\epsilon^2$  follows the inverse gamma distribution with shape parameter 2 and scale parameter 0.1.

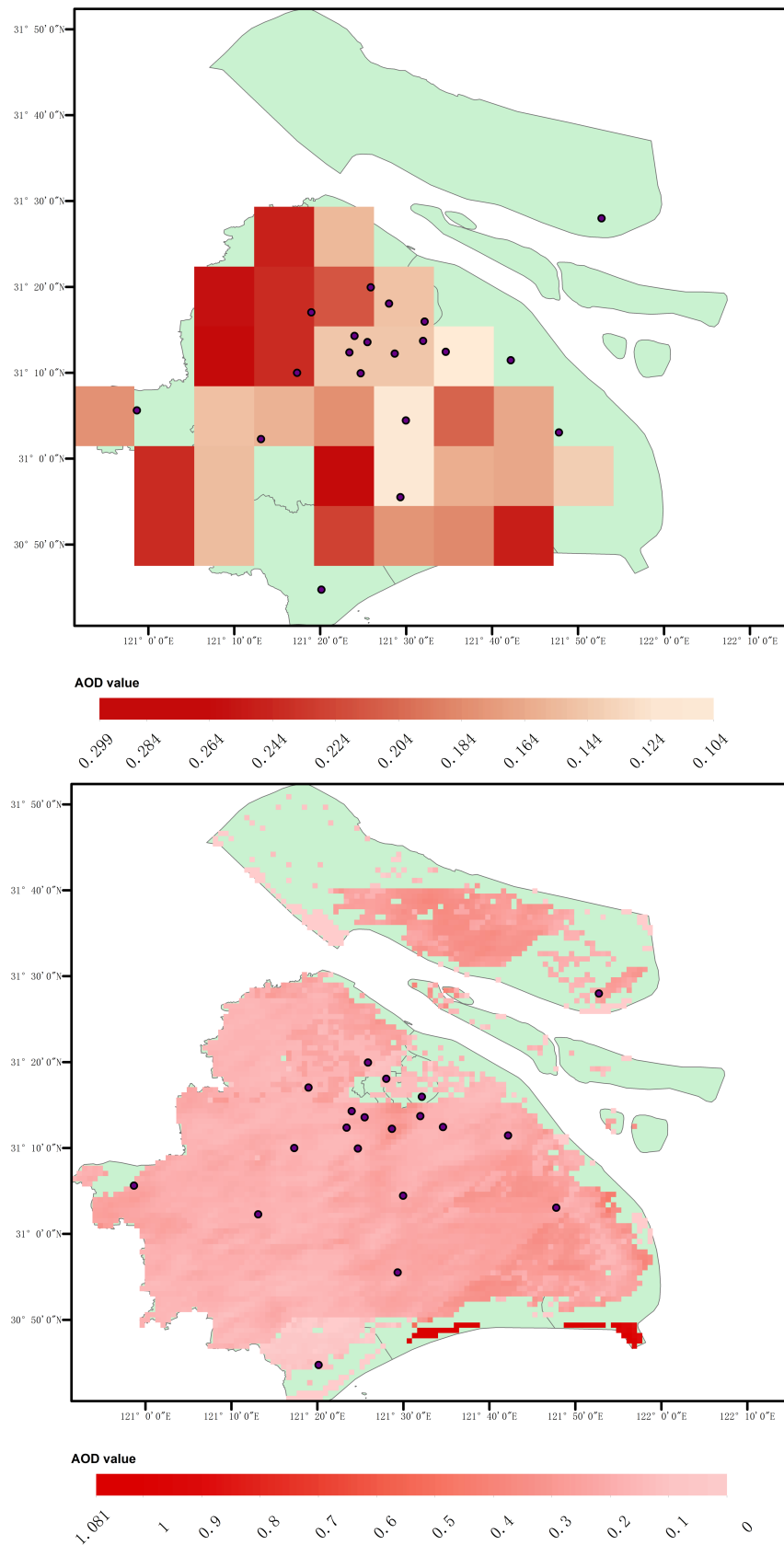
These parameters are updated by the Metropolis–Hastings algorithm with 5000 iterations for each parameter. Similar to [19–21], we use the first 2000 iterations as burn in and the last 3000 iterations to recover the PM<sub>2.5</sub> concentrations at location  $i$ . Let  $\beta_{0j}, \beta_{1j}, \sigma_j^2, \phi_j$ , and  $\sigma_{j\epsilon}^2$  be the values of these parameters  $\beta_0, \beta_1, \sigma^2, \phi$ , and  $\sigma_\epsilon^2$  from the prior distributions above for the  $j$ th iteration,  $j = 2001, \dots, 5000$ . Then we can obtain the random effect  $\omega_{ij}$  from  $\mathcal{N}(0, \mathcal{K}(h; \sigma_j^2, \phi_j))$  and then generate  $p_{ij}$  from  $\mathcal{N}(\beta_{0j} + \beta_{1j}\tau_i + \omega_{ij}, \sigma_{j\epsilon}^2)$ . The sample mean of these 3000  $p_{ij}$  is defined as the PM<sub>2.5</sub> concentrations at location  $i$ , which is the sample mean of all 3000  $p_{ij}$  at location  $i$ .

We evaluate the performance of our approach by out-of-sample validation using the cross-validation (CV). In this study, we use 5-fold CV on the day when there are more than five working PM<sub>2.5</sub> concentration monitoring stations. The training data were randomly split into five equal subsets based on the PM<sub>2.5</sub> concentration monitoring stations in 5-fold CV. Four subsets were used for the model fitting, and the remaining one was used for model validation. This process was repeated five times until all the subsets were tested. Three commonly used statistical performance metrics, including the coefficient of determination ( $R^2$ ), the mean absolute error (MAE), and the root-mean-square error (RMSE), are used to evaluate the model performance quantitatively. MAE represents the difference between model-estimated PM<sub>2.5</sub> concentrations and actual PM<sub>2.5</sub> concentrations measured at monitoring stations, while RMSE is the square root of the mean of the squares of the differences between model-estimated PM<sub>2.5</sub> concentrations and actual PM<sub>2.5</sub> concentrations measured at monitoring stations. Finally, we apply the hierarchical Gaussian process model to estimate 1-km-resolution PM<sub>2.5</sub> concentrations. We use Python and R in this study.

## 4. Results and Discussion

### 4.1. Evaluation of the Model Performance

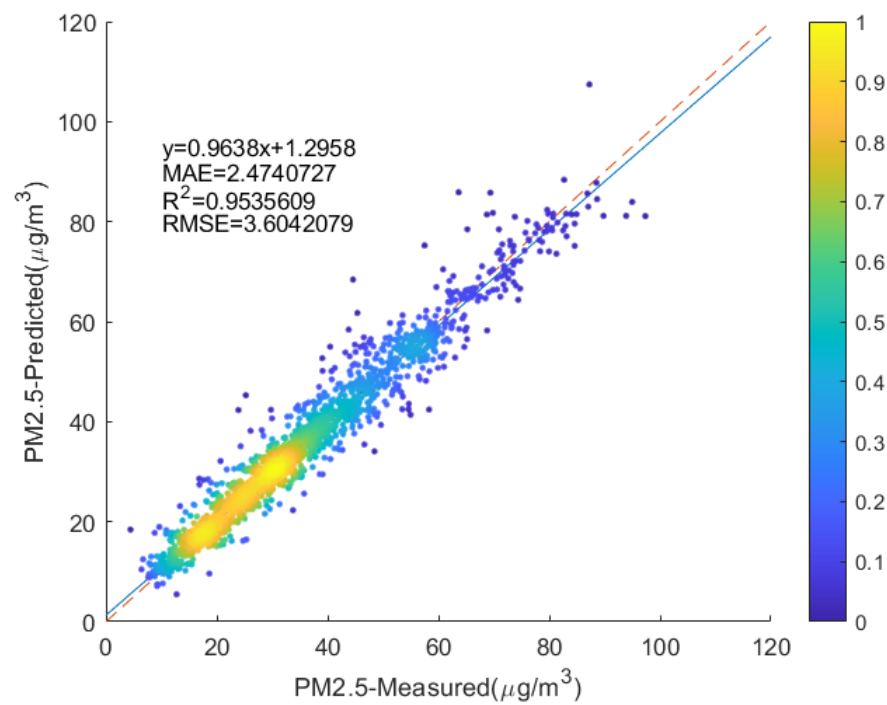
In this study, we improve the accuracy and coverage of the traditional MODIS AOD products using the Bayesian aerosol retrieval algorithm. Compared with existing results in [10,12], we replace the  $10 \times 10 \text{ km}^2$  grid cells with  $1 \times 1 \text{ km}^2$  grid cells and improve the accuracy by about 10–20%. Figure 2 shows the traditional MODIS AOD grid cells and AOD grid cells established by our approach in Shanghai. Note that we use different color scales for the two colorbars in this figure for better visibility. It can be seen from Figure 2 that the traditional MODIS grid cells are too coarse, and thus, many identical AOD values are used to estimate PM<sub>2.5</sub> concentrations at different geolocations. This can decrease the estimation accuracy of PM<sub>2.5</sub> concentrations. However, the grid cells in our approach are fine enough to capture the variation of AOD values at different geolocations.



**Figure 2.** Top panel: The traditional MODIS AOD grid cells. Bottom panel: AOD grid cells by our approach. Solid lines: city border and district borders. Black dots: PM<sub>2.5</sub> concentrations monitoring stations in Shanghai. Colored rectangles: the traditional MODIS AOD grid cells. For better visibility, different color scales are used in the two colorbars.



Figure 3 plots estimated PM<sub>2.5</sub> values in Shanghai obtained by our approach vs. actual PM<sub>2.5</sub> values for the whole year in Figure 3. It shows that our approach can estimate PM<sub>2.5</sub> concentrations in Shanghai very well. Three commonly used statistical performance metrics, MAE, R<sup>2</sup>, and RMSE, are also reported in the same figure. The value of R<sup>2</sup> is about 0.9536. The mean absolute error (MAE) is 3.6042079 μg/m<sup>3</sup> and the root-mean-square error (RMSE) is 2.47042 μg/m<sup>3</sup>. A large value of R<sup>2</sup> and small values of the RMSE and MAE indicate the predictive ability of our approach.



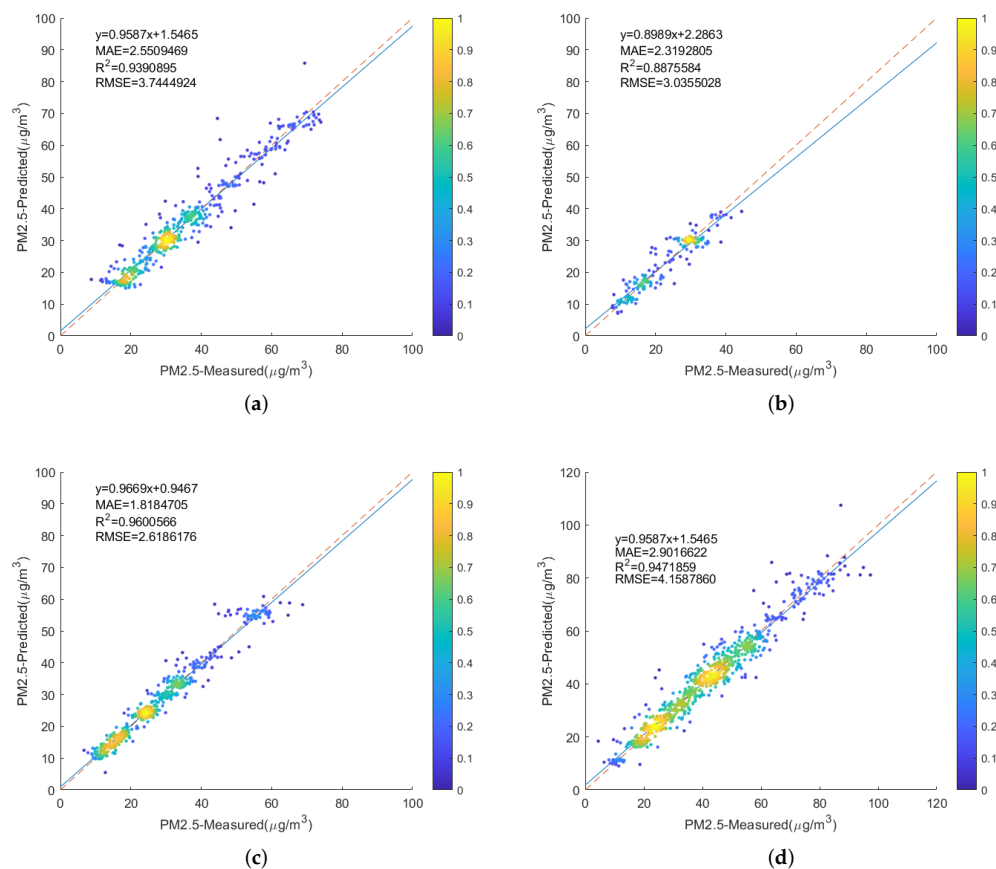
**Figure 3.** Scatter plot of estimated PM<sub>2.5</sub> values by our approach and actual PM<sub>2.5</sub> values for the whole year in Shanghai. The dashed line is the 1:1 reference line.

To consider the seasonal variations in the PM<sub>2.5</sub> concentrations, we also plot estimated PM<sub>2.5</sub> values in Shanghai obtained by our approach vs. actual PM<sub>2.5</sub> values for different seasons in Figure 4. It is clear from Figure 4 that our approach performs very well, especially for spring, autumn, and winter seasons. The performance of our approach for the summer season is not as good as that for other seasons. One of the possible reasons is that the cloud condition during the summer season in Shanghai hinders the accuracy of our Bayesian AOD retrieval and thus affects the estimation accuracy of PM<sub>2.5</sub> concentrations in the summer season.

Table 3 records the average, minimum and maximum of estimated PM<sub>2.5</sub> concentrations. It is clear that the PM<sub>2.5</sub> concentrations in Shanghai during the winter and spring seasons are high, while PM<sub>2.5</sub> concentrations during the summer and autumn seasons are relatively low. This finding is not surprising. One of the most important reasons is that there is more coal burning for heat in winter and early spring.

**Table 3.** The average, minimum, and maximum of estimated PM<sub>2.5</sub> concentrations in Shanghai.

	Average	Minimum	Maximum
Spring	34.16 μg/m <sup>3</sup>	16.53 μg/m <sup>3</sup>	68.90 μg/m <sup>3</sup>
Summer	18.69 μg/m <sup>3</sup>	8.07 μg/m <sup>3</sup>	38.20 μg/m <sup>3</sup>
Autumn	25.10 μg/m <sup>3</sup>	10.86 μg/m <sup>3</sup>	55.57 μg/m <sup>3</sup>
Winter	39.19 μg/m <sup>3</sup>	17.28 μg/m <sup>3</sup>	82.15 μg/m <sup>3</sup>
Whole year	29.62 μg/m <sup>3</sup>	10.86 μg/m <sup>3</sup>	74.41 μg/m <sup>3</sup>

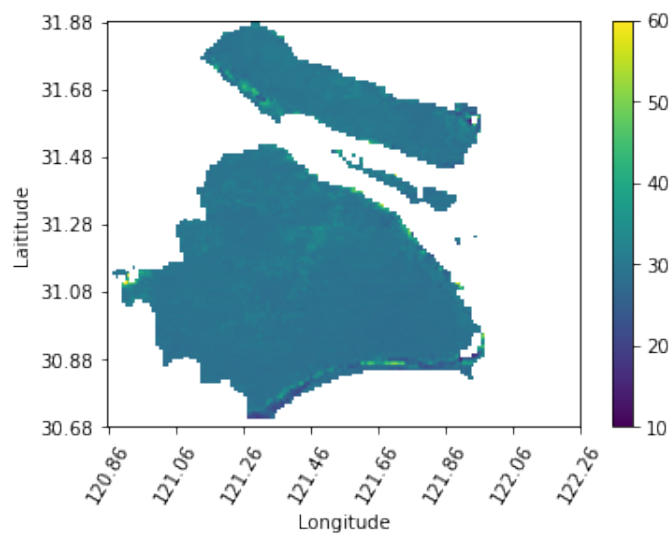


**Figure 4.** Scatter plot of estimated PM<sub>2.5</sub> values by our approach and actual PM<sub>2.5</sub> values for four seasons in Shanghai. The dashed line is the 1:1 reference line. (a) Spring; (b) Summer; (c) Autumn; (d) Winter.

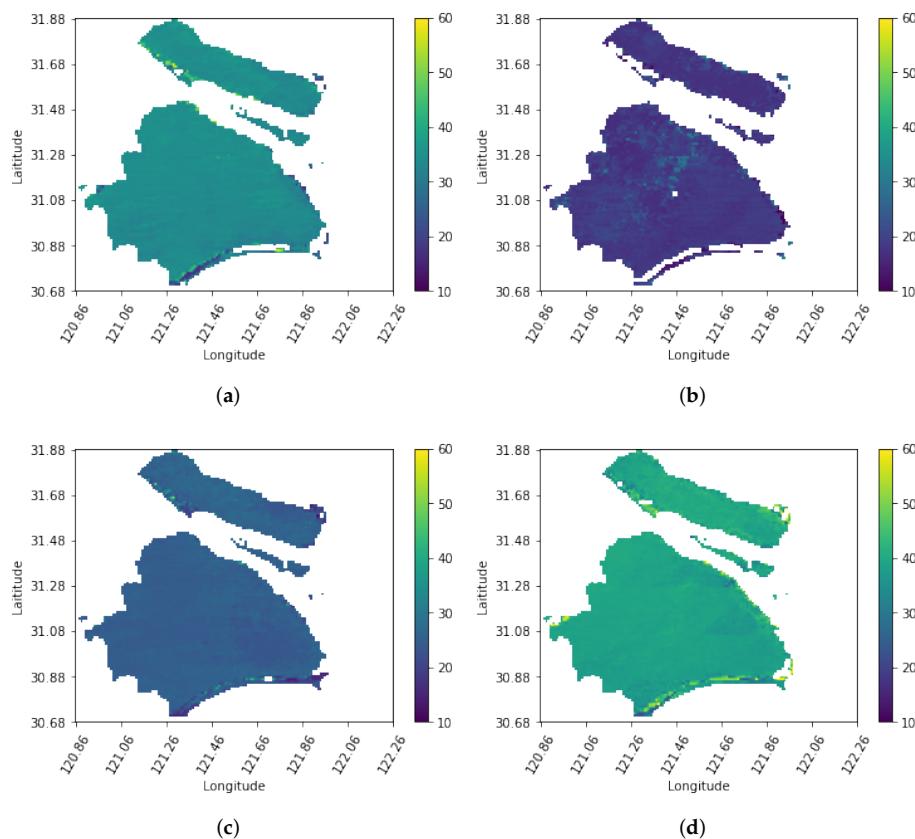
#### 4.2. Spatial and Seasonal Variations of PM<sub>2.5</sub> Concentrations in Shanghai

To further explore the spatial and seasonal variations of the estimated PM<sub>2.5</sub> concentrations in Shanghai by our approach, we show the annual and seasonal distributions of PM<sub>2.5</sub> concentrations in Shanghai estimated by our approach in Figures 5 and 6. It can be seen that the PM<sub>2.5</sub> concentrations in the central region of Shanghai are higher than those in surrounding suburban areas, especially during the summer season. This is not surprising because Shanghai is one of the largest cities in China and a global financial hub, and there are more human activities in the central region of Shanghai than the surrounding suburban areas. This causes more air pollutants such as vehicle exhaust, which accounts for the high PM<sub>2.5</sub> concentrations. This finding is also consistent with the fact that there are more visitors from other places to Shanghai during the summer season because Shanghai is a popular tourist destination renowned for its historical landmarks and a great mix of the traditional and modern.

We also find that the estimated PM<sub>2.5</sub> concentrations on the southwest side of Chongming island (which is the island in the upper right corner in Figure 5) are higher than in other regions of Shanghai. As we can see from Figure 1, there is only one PM<sub>2.5</sub> concentration monitoring station in Chongming island. The number of observations of PM<sub>2.5</sub> concentrations on the southwest side of the island is limited. Therefore, the PM<sub>2.5</sub> concentrations estimated by our approach in this particular region may be misleading.



**Figure 5.** The mean values of estimated  $PM_{2.5}$  concentrations for the year 2021 using our approach.



**Figure 6.** The mean values of estimated  $PM_{2.5}$  concentrations for four seasons using our model. (a) Spring; (b) Summer; (c) Autumn; (d) Winter.

### 5. Conclusions

In this paper, we propose a new two-step approach to estimate 1-km-resolution  $PM_{2.5}$  concentrations in Shanghai using high spatial resolution AOD retrievals from MODIS. It enjoys accurate predictive performance in model fitting and cross-validation. Our model only needs AOD products to estimate ground  $PM_{2.5}$  concentrations without sophisticated auxiliary variables.

We use the random cross-validation when evaluating the predictive performance of our approach. As pointed out by one anonymous reviewer, the random cross-validation

can be misleading in the presence of spatial correlations. There are many other approaches for cross-validation available in the literature; for example, see [22–26]. One can use one of these cross-validation techniques to evaluate the predictive performance of the proposed approach.

In addition, only two variables, PM<sub>2.5</sub> and AOD, are used in our hierarchical Gaussian process model. It is possible to include more additional variables, such as population, wind, and rivers, in our hierarchical Gaussian process model. In this case, one may need to set non-Gaussian priors for these additional variables [27]. This is beyond the scope of our current paper, and we leave it for future research.

**Author Contributions:** Conceptualization, J.Z. and Q.L.; methodology, Q.L. and D.L.; software, J.Z. and Y.X.; validation, J.Z. and Y.X.; formal analysis, J.Z. and Y.X.; investigation, J.Z. and Y.X.; resources, J.Z. and Y.X.; data curation, J.Z. and Y.X.; writing—original draft preparation, J.Z.; writing—review and editing, J.Z., D.L. and Q.L.; visualization, J.Z. and D.L.; supervision, Q.L.; project administration, Q.L. and D.L.; funding acquisition, Q.L. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work is supported by the Science and Technology Commission of Shanghai Municipality (No. 20JC1414300) and the Natural Science Foundation of Shanghai (No. 20ZR1436200).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Not applicable.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

AOD	Aerosol Optical Depth
FMF	Fine Mode Fraction
MODIS	Moderate Resolution Imaging Spectroradiometer
MAIMC	Multi-Angle Implementation of Atmospheric Correction
CV	Cross Validation
PM <sub>2.5</sub>	Particulate matter with an aerodynamic diameter less than 2.5 μm

## References

1. Yan, X.D.; Wang, Q.M.; Tie, C.; Jin, H.T.; Han, Y.X.; Zhang, J.L.; Yu, X.M.; Hou, Q.; Zhang, P.P.; Wang, A.P. Polydatin protects the respiratory system from PM<sub>2.5</sub> exposure. *Sci. Rep.* **2017**, *7*, 40030. [[CrossRef](#)] [[PubMed](#)]
2. Brook, R.D.; Rajagopalan, S.; Pope, C.A., III; Brook, J.R.; Bhatnagar, A.; Diez-Roux, A.V.; Holguin, F.; Hong, Y.; Luepker, R.V.; Mittleman, M.A. Particulate matter air pollution and cardiovascular disease: An update to the scientific statement from the American Heart Association. *Circulation* **2010**, *121*, 2331–2378. [[CrossRef](#)] [[PubMed](#)]
3. Hayes, R.B.; Lim, C.; Zhang, Y.; Cromar, K.; Shao, Y.; Reynolds, H.R.; Silverman, D.T.; Jones, R.R.; Park, Y.; Jerrett, M. PM<sub>2.5</sub> air pollution and cause-specific cardiovascular disease mortality. *Int. J. Epidemiol.* **2020**, *49*, 25–35. [[CrossRef](#)] [[PubMed](#)]
4. Wu, J.Z.; Ge, D.D.; Zhou, L.F.; Hou, L.Y.; Zhou, Y.; Li, Q.Y. Effects of particulate matter on allergic respiratory diseases. *Chronic Dis. Transl. Med.* **2018**, *4*, 95–102. [[CrossRef](#)] [[PubMed](#)]
5. Dominici, F.; Peng, R.D.; Bell, M.L.; Pham, L.; McDermott, A.; Zeger, S.L.; Samet, J.M. Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases. *JAMA* **2006**, *295*, 1127–1134. [[CrossRef](#)] [[PubMed](#)]
6. Yuan, Y.; Liu, S.; Castro, R.; Pan, X. PM<sub>2.5</sub> monitoring and mitigation in the cities of China. *Environ. Sci. Technol.* **2012**, *46*, 3627–3628. [[CrossRef](#)] [[PubMed](#)]
7. Song, W.; Jia, H.; Huang, J.; Zhang, Y. A satellite-based geographically weighted regression model for regional PM<sub>2.5</sub> estimation over the Pearl River Delta region in China. *Remote Sens. Environ.* **2014**, *154*, 1–7. [[CrossRef](#)]
8. Ma, Z.; Liu, Y.; Zhao, Q.; Liu, M.; Zhou, Y.; Bi, J. Satellite-derived high resolution PM<sub>2.5</sub> concentrations in Yangtze River Delta Region of China using improved linear mixed effects model. *Atmos. Environ.* **2016**, *133*, 156–164. [[CrossRef](#)]
9. Yu, W.; Liu, Y.; Ma, Z.; Bi, J. Improving satellite-based PM 2.5 estimates in China using Gaussian processes modeling in a Bayesian hierarchical setting. *Sci. Rep.* **2017**, *7*, 7048. [[CrossRef](#)] [[PubMed](#)]
10. Lipponen, A.; Mielonen, T.; Pitkänen, M.R.; Levy, R.C.; Sawyer, V.R.; Romakkaniemi, S.; Kolehmainen, V.; Arola, A. Bayesian aerosol retrieval algorithm for MODIS AOD retrieval over land. *Atmos. Meas. Tech.* **2018**, *11*, 1529–1547. [[CrossRef](#)]

11. Wang, Y.; Jiang, X.; Yu, B.; Jiang, M. A hierarchical Bayesian approach for aerosol retrieval using MISR data. *J. Am. Stat. Assoc.* **2013**, *108*, 483–493. [[CrossRef](#)]
12. Lyapustin, A.; Wang, Y.; Korkin, S.; Huang, D. MODIS collection 6 MAIAC algorithm. *Atmos. Meas. Tech.* **2018**, *11*, 5741–5765. [[CrossRef](#)]
13. Wei, J.; Huang, W.; Li, Z.; Xue, W.; Peng, Y.; Sun, L.; Cribb, M. Estimating 1-km-resolution PM<sub>2.5</sub> concentrations across China using the space-time random forest approach. *Remote Sens. Environ.* **2019**, *231*, 111221. [[CrossRef](#)]
14. Kaufman, Y.; Tanré, D.; Remer, L.A.; Vermote, E.; Chu, A.; Holben, B. Operational remote sensing of tropospheric aerosol over land from EOS moderate resolution imaging spectroradiometer. *J. Geophys. Res. Atmos.* **1997**, *102*, 17051–17067. [[CrossRef](#)]
15. Levy, R.; Mattoo, S.; Munchak, L.; Remer, L.; Sayer, A.; Patadia, F.; Hsu, N. The Collection 6 MODIS aerosol products over land and ocean. *Atmos. Meas. Tech.* **2013**, *6*, 2989–3034. [[CrossRef](#)]
16. Kinne, S.; O'Donnel, D.; Stier, P.; Kloster, S.; Zhang, K.; Schmidt, H.; Rast, S.; Giorgetta, M.; Eck, T.F.; Stevens, B. MAC-v1: A new global aerosol climatology for climate studies. *J. Adv. Model. Earth Syst.* **2013**, *5*, 704–740. [[CrossRef](#)]
17. Byrd, R.H.; Lu, P.; Nocedal, J.; Zhu, C. A limited memory algorithm for bound constrained optimization. *SIAM J. Sci. Comput.* **1995**, *16*, 1190–1208. [[CrossRef](#)]
18. Williams, C.K.; Rasmussen, C.E. *Gaussian Processes for Machine Learning*; MIT Press: Cambridge, MA, USA, 2006; Volume 2.
19. Finley, A.O.; Banerjee, S.; Carlin, B.P. spBayes: An R package for univariate and multivariate hierarchical point-referenced spatial models. *J. Stat. Softw.* **2007**, *19*, 1–24. [[CrossRef](#)] [[PubMed](#)]
20. Finley, A.O.; Banerjee, S.; Gelfand, A.E. spBayes for large univariate and multivariate point-referenced spatio-temporal data models. *J. Stat. Softw.* **2015**, *63*, 1–28. [[CrossRef](#)]
21. Bakar, K.S.; Kocic, P. Bayesian Gaussian models for point referenced spatial and spatio-temporal data. *J. Stat. Res.* **2017**, *51*, 17–40. [[CrossRef](#)]
22. Brenning, A. Spatial cross-validation and bootstrap for the assessment of prediction rules in remote sensing: The R package sperrorest. In Proceedings of the 2012 IEEE International Geoscience and Remote Sensing Symposium, Munich, Germany, 22–27 July 2012; pp. 5372–5375. doi: [[CrossRef](#)]
23. Roberts, D.R.; Bahn, V.; Ciuti, S.; Boyce, M.S.; Elith, J.; Guillera-Arroita, G.; Hauenstein, S.; Lahoz-Monfort, J.J.; Schröder, B.; Thuiller, W. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* **2017**, *40*, 913–929. [[CrossRef](#)]
24. Valavi, R.; Elith, J.; Lahoz-Monfort, J.J.; Guillera-Arroita, G. block CV: An r package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. *Methods Ecol. Evol.* **2019**, *10*, 225–232. [[CrossRef](#)]
25. Ploton, P.; Mortier, F.; Réjou-Méchain, M.; Barbier, N.; Picard, N.; Rossi, V.; Dormann, C.; Cornu, G.; Viennois, G.; Bayol, N. Spatial validation reveals poor predictive performance of large-scale ecological mapping models. *Nat. Commun.* **2020**, *11*, 4540. [[CrossRef](#)] [[PubMed](#)]
26. Brenning, A. Spatial machine-learning model diagnostics: A model-agnostic distance-based approach. *arXiv* **2021**, arXiv:2111.08478.
27. Ren, M.; Zhang, Q.; Zhang, J. An introductory survey of probability density function control. *Syst. Sci. Control Eng.* **2019**, *7*, 158–170. [[CrossRef](#)]