International Journal of
Environmental Research and Public Health

MDPI

Article

A Bayesian Downscaler Model to Estimate Daily $PM_{2.5}$ Levels in the Conterminous US

Yikai Wang [1], Xuefei Hu [2], Howard H. Chang [1], Lance A. Waller [1], Jessica H. Belle [2] and Yang Liu [2,*]

[1] Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, GA 30322, USA; johnzon.wyk@gmail.com (Y.W.); howard.chang@emory.edu (H.H.C.); lwaller@emory.edu (L.A.W.)

[2] Department of Environmental Health, Rollins School of Public Health, Emory University, Atlanta, GA 30322, USA; xuefeihucn@hotmail.com (X.H.); jessicabelle4@gmail.com (J.H.B.)

[*] Correspondence: yang.liu@emory.edu

Received: 3 August 2018; Accepted: 10 September 2018; Published: 13 September 2018

Abstract: There has been growing interest in extending the coverage of ground particulate matter with aerodynamic diameter $\leq$ 2.5 µm ($PM_{2.5}$) monitoring networks based on satellite remote sensing data. With broad spatial and temporal coverage, a satellite-based monitoring network has a strong potential to complement the ground monitor system in terms of the spatiotemporal availability of the air quality data. However, most existing calibration models focus on a relatively small spatial domain and cannot be generalized to a national study. In this paper, we proposed a statistically reliable and interpretable national modeling framework based on Bayesian downscaling methods to be applied to the calibration of the daily ground $PM_{2.5}$ concentrations across the conterminous United States using satellite-retrieved aerosol optical depth (AOD) and other ancillary predictors in 2011. Our approach flexibly models the $PM_{2.5}$ versus AOD and the potential related geographical factors varying across the climate regions and yields spatial- and temporal-specific parameters to enhance model interpretability. Moreover, our model accurately predicted the national $PM_{2.5}$ with an R^2 at 70% and generated reliable annual and seasonal $PM_{2.5}$ concentration maps with its SD. Overall, this modeling framework can be applied to national-scale $PM_{2.5}$ exposure assessments and can also quantify the prediction errors.

Keywords: $PM_{2.5}$; Bayesian downscaler; exposure modeling; aerosol optical depth; MODIS

1. Introduction

Particulate air pollution has become a major environmental and public health concern worldwide in recent years. Particularly, particulate matter with aerodynamic diameter $\leq$2.5 µm ($PM_{2.5}$) is shown to have a strong association with various adverse health outcomes, such as increased mortality and morbidity and aggravated respiratory and cardiovascular symptoms [1]. Ambient $PM_{2.5}$ is either directly emitted from various anthropogenic and biogenic sources or generated in the atmosphere from complex photochemical reactions [2]. Consequently, $PM_{2.5}$ concentrations vary in space and time at sub-kilometer to continental scales [3]. Thus, it is important to accurately assess the population exposure of $PM_{2.5}$. However, in the interest of reducing cost, $PM_{2.5}$ monitors are usually sparsely distributed and tend to be concentrated among urban areas, and most $PM_{2.5}$ monitors operated by the US Environmental Protection Agency (EPA) and the IMPROVE network only operate on a one-in-three-day or one-in-six-day schedule, leaving significant temporal gaps. Due to these spatial and temporal limitations, it is difficult for current $PM_{2.5}$ networks to provide sufficient data to fully assess $PM_{2.5}$ for health effect studies and it could lead to biased results for some key scientific questions.

One emerging solution to these problems is spatial models driven by remotely sensed particle properties from the satellite platform as well as gridded meteorological and land use information. The most robust and widely used satellite parameter is the aerosol optical depth (AOD), which measures the overall particle light extinction caused by airborne particles in the atmospheric column. Many previous studies have shown that AOD has a strong positive association with $PM_{2.5}$. In addition to AOD, previous studies have shown that meteorological and land use information are important factors to predict the ground-level concentration of $PM_{2.5}$ and the relationship between AOD and $PM_{2.5}$ [4–6]. All these properties between AOD and $PM_{2.5}$ make it possible to develop statistical methods to calibrate $PM_{2.5}$ using AOD and other geographical factors. Over the past decade, various MODIS-driven $PM_{2.5}$ exposure models have been developed, from relatively simple linear regressions [7] to complex multilevel spatial models [8] and Bayesian hierarchical models [9]. Bayesian hierarchical models have more flexibility in modeling the complex temporal and spatial pattern of $PM_{2.5}$, and compared with other spatial models based on mixed-effects terms [7–10], one major advantage of the Bayesian model is its underlying nature to quantify the prediction uncertainty through the Markov chain Monte Carlo (MCMC) algorithm, which is crucial for scientific research. Therefore, in this study, we extended the Bayesian model proposed by Chang et al. [9] into a national Bayesian model to examine $PM_{2.5}$ under a national domain.

So far, most satellite-driven $PM_{2.5}$ exposure models have been developed at the urban to regional scales in order to support health effect studies in specific regions [5,9,11–15]. High-performance national scale $PM_{2.5}$ exposure models are still limited partially because of the high-computational demand in order to make national $PM_{2.5}$ prediction surfaces. A couple of national-scale studies involved machine learning methods [16,17]. Di et al. [16] developed a neural network approach, incorporated with convolutional layers to account for spatiotemporal autocorrelation, to predict $PM_{2.5}$ concentrations in the continental United States from 2000 to 2012. Hu et al. [17] developed a random forest model with ~40 predictors to predict $PM_{2.5}$ exposure in the conterminous United States in 2011. These emerging methods can provide relatively high predication accuracy but offer little insight into how different predictors behave across such large domains. For example, random forests only provide an importance value for each predictor to indicate which predictor is more important in the training process. Both neural networks and random forests do not provide quantification of uncertainties in prediction and parameter estimation. These methods also cannot provide straightforward estimates of the model prediction errors. On the other hand, statistical models provide a balance between model predication accuracy and the ability for interpretation and serve as the most reliable and commonly used approaches in calibrating the $PM_{2.5}$. For example, Lee et al. [13] proposed a mixed-effects model with random temporal intercept and slope on AOD to evaluate the time-varying effects. This type of model assumes that the temporal-random-effect-based model requires the independence assumption between different days, which is generally not practical and, thus, they have limited power to make predictions out of the temporal domain. They also fail to adjust the spatial variability in large spatial domains and can provide biased results. Similar for the hierarchical models [5,11,12], it is tricky to quantify the uncertainties in prediction or parameter estimation based on such models, which limits their power to be used in real applications. Thus, all these models are not directly applicable to the national domain.

Chang et al. [9] reported a Bayesian downscaling model which adopted the Gaussian spatial process to incorporate the spatial correlation into the model, which increases the power to borrow information across neighborhoods, through which the challenge of spatial misalignment between the point-referenced monitoring measurements and the gridded areal AOD data can be solved. It also models the conditional correlation between adjacent observed days, which allows us to estimate the random effects on the day without $PM_{2.5}$ measurements. In addition, this model adopts a full Bayesian approach, by which the model uncertainty can be obtained easily. However, this model is only applicable to small spatial domains for three reasons. First, it assumes that the temporal correlation structure is constant across the whole spatial domain, but based on our study, this is not realistic in a large spatial domain. Similarly, it assumes that the spatial correlation structure is constant across the whole year, which is not realistic. Second, the original model is not flexible enough to capture the huge spatial variability in the national domain. For example, it assumes a constant effect of land use across different states that fails to consider the localized difference, which is one of the goals of a national study. Finally, directly generalizing the original model is computationally expensive because the spatial correlation matrix is of high dimensions and is very sparse.

In this paper, we enhanced and expanded the original Bayesian downscaler to the entire continental United States We developed a regional- and temporal-specific Bayesian downscaling approach to gain more flexibility. Our model incorporated AOD data, meteorological fields, and land use variables to estimate daily ground-level $PM_{2.5}$ concentrations over the conterminous United States for the year 2011. The estimated regional- and temporal-specific parameters were scientifically meaningful and the prediction accuracy was evaluated through general and spatial cross-validation frameworks. Our model predicted the daily averaged $PM_{2.5}$ concentrations across the entire continental United States and also the prediction uncertainty maps.

2. Data and Methods

2.1. Data Collection

The 24-h averaged $PM_{2.5}$ measurements for 2011 were downloaded from the US EPA's Air Quality System Technology Transfer Network (http://www.epa.gov/ttn/airs/airsaqs/). Collection 6 level 2 Aqua MODIS retrievals at a nominal spatial resolution of 10 km were regridded to the 12 × 12 km^2 Community Multi-Scale Air Quality (CMAQ) modeling system (https://www.epa.gov/cmaq). Regridding to a fixed grid is necessary when modeling with level 2 MODIS data because MODIS pixels shift in location and size with each satellite overpass. In addition, given the 10-km nominal resolution of MODIS AOD pixels, regridding to the commonly used 12-km CMAQ grid does not compromise the AOD spatial resolution significantly. Doing so may also benefit future comparisons between CMAQ simulation results and our model predictions. We calculated AOD averages using AOD retrievals from the combined deep-blue and dark-target parameters. Meteorological fields were obtained from the North American Regional Reanalysis (NARR) (http://www.emc.ncep.noaa.gov/mmb/rreanl/), with a spatial resolution of ~32 km and a temporal resolution of 3 h, and the North American Land Data Assimilation System Phase 2 (NLDAS-2) (http://www.emc.ncep.noaa.gov/mmb/rreanl/), with a spatial resolution of ~13 km and a temporal resolution of 1 h. Elevation data at a spatial resolution of ~30 m were downloaded from the National Elevation Dataset (http://ned.usgs.gov). Road data were extracted from ESRI StreetMap USA. Percentage forest cover data at a spatial resolution of ~30 m were extracted from the 2011 Landsat-derived land cover map downloaded from the National Land Cover Database (NLCD) (http://www.mrlc.gov). Primary $PM_{2.5}$ emissions were obtained from the 2011 EPA National Emissions Inventory (NEI) facility emissions report (https://www.epa.gov/air-emissions-inventories/2011-national-emissions-inventory-nei-data).

2.2. Climate Regions and Temporal Domains

To improve computational efficiency, we divided the conterminous U.S. into nine NOAA-defined climate regions [18], which include Northeast, Southeast, South, Ohio Valley (Central), Upper Midwest (East North Central), Northern Rockies and Plains (West North Central), Southwest, Northwest, and West (Figure 1). After examining aerosol light extinction measurements in various regions of the world, Anderson et al. [3] reported that the typical mesoscale variability of lower-tropospheric particles ranges between 40 and 400 km Therefore, by dividing our national domain into nine multistate regions, we were still able to sufficiently capture the spatial and temporal correlations of ground-level $PM_{2.5}$. We added a 100-km buffer to each climate region and averaged overlapping predictions from neighboring regions to generate a smooth national $PM_{2.5}$ concentration surface. In addition, the spatial pattern varies significantly across the year. To reduce the high computational demand of our Bayesian model, we divided the year 2011 into three 4-month temporal periods and developed a Bayesian downscaling model in each period. Since the typical $PM_{2.5}$ residence time in the boundary layer ranges from a couple of days to two weeks, this treatment had minimal impact on model performance.

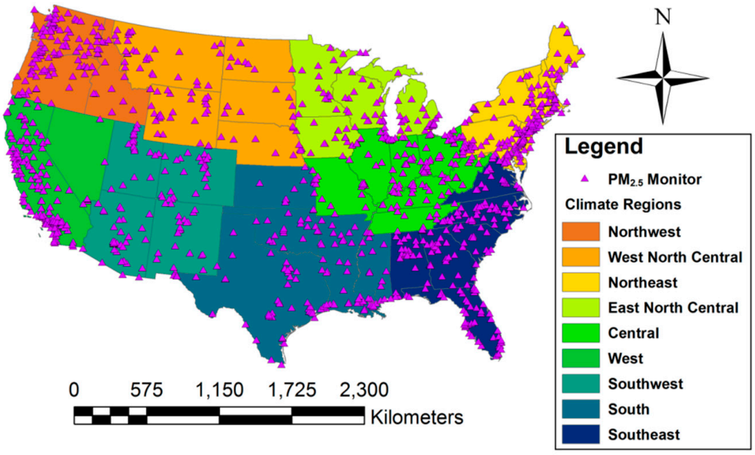

Figure 1. The nine climate regions and the spatial location of the monitors.

2.3. National Bayesian Downscaling Model

For each regional and temporal subdomain, we adopted the basic framework of the Bayesian downscaling model proposed by Chang et al. [9]. In this model, let *PM(s,t)* denote the $PM_{2.5}$ concentration at location s and day t, where s can be viewed as the unique spatial coordinates. Similarly, let *AOD(s,t)* denote the AOD measurement at the grid cell containing the monitor s and day t. For one specific climate region reg, a function of s and the temporal domain, the first level model between AOD and $PM_{2.5}$ is given as

$$PM(s,t) = \alpha_0(s,t) + \alpha_1(s,t)\ AOD(s,t) + \gamma_{reg,\ tem}(s,t)\ Z(s,t) + \varepsilon(s,t) \tag{1}$$

where $\alpha_0(s,t)$ and $\alpha_1(s,t)$ are the day-specific and location-specific random intercept and slope and the residual error ε(s,t) is assumed to be independently normal with mean zero and regional- and temporal-specific variance $\sigma_{\mathrm{reg,tem}}{}^2$. *Z(s,t)* represents for the covariates having a constant association with $PM_{2.5}$, where $\gamma_{\mathrm{reg,tem}}$ represents for the regional- and temporal-specific fixed effect between *Z(s,t)* and *PM(s,t)*. Here, *Z(s,t)* includes fire, forest coverage, emission, relative humidity (RH), temperature, wind speeds, major roadway length, boundary layer height (Hpbl), and the interaction between AOD and temperature.

The spatiotemporal random effects $\alpha_0(s,t)$ and $\alpha_1(s,t)$ are specified using additive setting. For clarity, we present the model setting for one specific region and temporal domain: $\alpha_i(s,t) = \beta_i(\mathrm{s}) + \beta_i(\mathrm{t})$, i = 0,1, where β_i(s) and β_i(t) are independent spatial and temporal effects. The spatial effects are modeled using a latent structure of two independent spatial Gaussian processes W_1(s) and W_2(s), where $\beta_0(\mathrm{s}) = c_1W_1(\mathrm{s})$ and $\beta_1(\mathrm{s}) = c_2W_1(\mathrm{s}) + c_3W_2(\mathrm{s})$ and the covariance function of W_i(s) for each region is given by the exponential function multiplied by a tapering function. The regional-specific temporal effects β_0(t) and β_1(t) are modeled as two independent daily time series using a first-order random walk, which can be defined through the conditional distribution of a particular day given all other days. More details about the model specifications can be found in the online supplementary materials.

2.4. Model Fitting and Prediction

Model fitting was carried out using Markov chain Monte Carlo (MCMC) techniques [9,19]. Details of the MCMC algorithms and the prior settings can be found in the online supplementary materials. Prediction performance was evaluated using two different cross-validation methods: fully random cross-validation (random CV) and spatial cross-validation (spatial CV) [9,20]. In random CV, we randomly split the data into 10 folds and fit the model using 9 folds and evaluated the fitted model using the remaining fold, which can be used to evaluate the overall prediction ability of our approach. The spatial CV was similar to random CV except that rather than randomly splitting the data, we split the data based on its spatial location. The spatial CV results were used to evaluate the ability in spatial extrapolation. In addition, through the MCMC approach, we quantified the prediction uncertainty, i.e., interval estimates. We also calculated the prediction statistics by comparing the predicted $PM_{2.5}$ measurements with the observations, which include root-mean-square error (RMSE), 90% posterior interval (PI) length, and its empirical coverage probability and linear coefficient of determination R^2 value. All analyses were carried out in R version 3.2.3 (https://www.r-project.org/).

3. Results

3.1. Data Description and Summary

The histograms of all variables are illustrated in Figure 2, which shows that all the variables are approximately unimodal and log-normal distributed. Log-transformation was conducted for fire, emission, and road for the following analysis. Z-transformation was conducted for all variables except for $PM_{2.5}$ and AOD to remove the collinearity between covariates and to make the scale comparable. The annual mean $PM_{2.5}$ concentration for all monitors was 9.88 μg/m^3, with an SD of 6.17 μg/m^3. The overall mean of AOD was 0.14, with an SD of 0.15. The region-specific descriptive statistics for $PM_{2.5}$ and AOD are summarized in Table 1. The number of records, monitors, days, and the percentage of data coverage for each region are summarized in Table 2. Among the nine regions, the Ohio Valley has the highest mean $PM_{2.5}$ concentration at 11.29 μg/m^3 and it has most records (18,642) and monitors (361). In terms of missing AOD data, the West has the best data coverage (30%), while the Northeast has the worst (12%) due to both cloud cover and snow cover in winter.

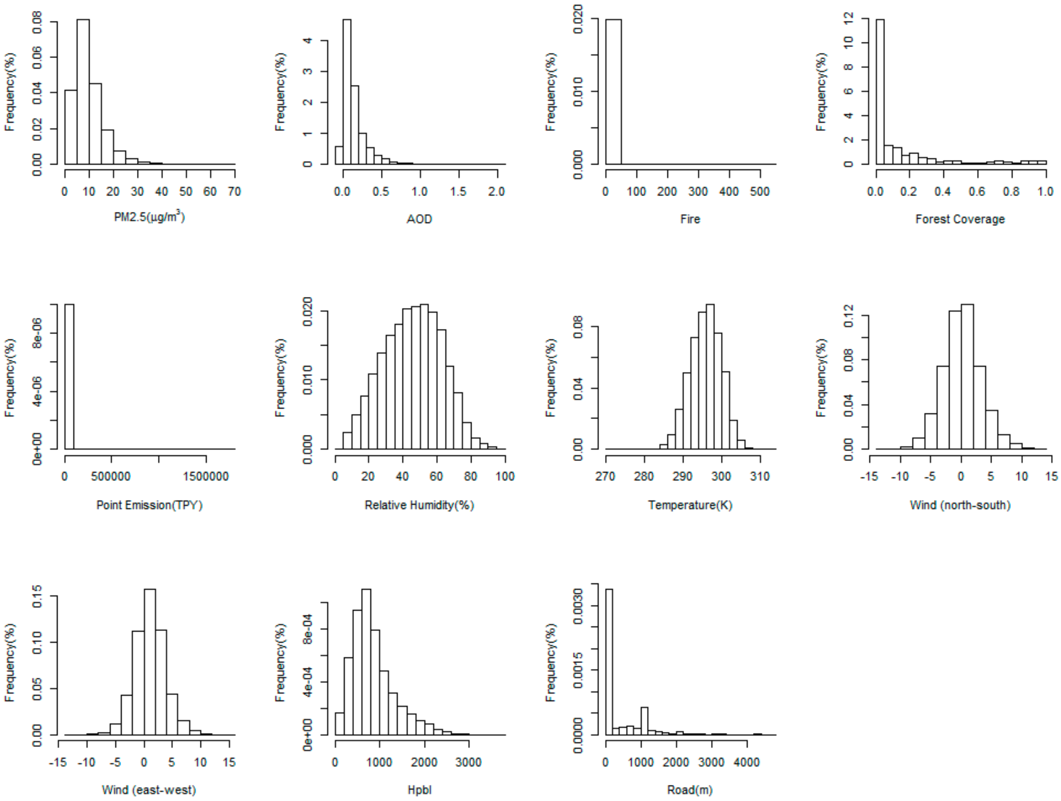

Figure 2. Histograms of the dependent and independent variables.

Table 1. Descriptive statistics for $PM_{2.5}$ and aerosol optical depth (AOD).

Regions	$PM_{2.5}$ (SD)	AOD (SD)
West	10.72 (7.17)	0.10 (0.12)
Northwest	6.23 (4.05)	0.12 (0.11)
Southwest	7.40 (4.75)	0.10 (0.11)
Northern Rockies and Plains	7.40 (4.11)	0.12 (0.13)
Upper Midwest	10.33 (5.87)	0.18 (0.17)
South	10.17 (5.09)	0.13 (0.15)
Southeast	10.83 (5.34)	0.15 (0.17)
Ohio Valley	11.29 (5.79)	0.17 (0.17)
Northeast	10.68 (6.10)	0.19 (0.19)

Table 2. Region-specific counts and data coverage.

Regions	Number of Records	Number of Days	Number of Monitors	Coverage
West	17,096	356	159	30%
Northwest	9486	295	170	19%
Southwest	9567	363	138	19%
Northern Rockies and Plains	7463	328	150	15%
Upper Midwest	6208	304	145	14%
South	15,899	364	189	23%
Southeast	17,525	361	257	19%
Ohio Valley	18,642	354	361	15%
Northeast	8913	302	238	12%

3.2. Regional and Temporal Varying Geographical Associations

In this section, we explore the different patterns across climate regions and temporal domains revealed by the significant parameters in our model. First of all, the national Bayesian downscaling model fits well across different regions and temporal domains in terms of model R^2 and slope, as shown in Table 3. The Northeast tends to have the best model fitting. The climate regions with the highest set of R^2 are the Upper Midwest (0.85) and Northeast (0.84) regions. The slopes in these two climate regions are consistently higher than 0.95 across all time domains, indicating minimal systematic biases in model fitting. On the other hand, the model tends to have a lower R^2 in the South and Northwest regions, where the annual R^2 for the South climate region is 0.64 and the annual R^2 for the Northwest climate region is 0.63.

Table 3 presents all the significant geographical and meteorological factors (p-value < 0.05), which exhibit substantial inter-region differences among the regional models. First of all, AOD is the most important covariate and is significant for most regions and temporal domains. We noticed that the effect of AOD on $PM_{2.5}$ is weaker in May through August than other months. This pattern is commonly observed in all climate regions after we condition the temperature to be the average level in the specific spatial and temporal domain. Furthermore, fire, RH, temperature (TMP), and the interaction between AOD and TMP are significant across all regions. Other covariates, including forest coverage, emission, wind speed, Hpbl, and road length, vary across regions and temporal domains. Forest coverage is a significant factor in explaining the pattern of $PM_{2.5}$ in the West, Northwest, and Southwest climate regions across the entire year but is not significant in the Northern Rockies and Plains regions. Emission is not significant in most regions except the West, where it significantly explains the variability of $PM_{2.5}$. On the other hand, Hpbl has a significantly negative association with $PM_{2.5}$ in the West region but is not significant in the Northwest and Northeast regions.

Table 3. Statistically significant geographical and meteorological predictors.

Region	Temporal	AOD	Fire	Forest	Emission	RH	TMP	Vgrd	Ugrd	Hpbl	Road	AOD * TMP	R^2	Slope
West	1	21.2 (5.9)		−0.7 (0.3)	0.6 (0.2)	2.5 (0.2)	2.4 (0.4)	1.2 (0.2)		−0.2 (0.1)		9.8 (3.4)	0.65	0.88
	2	4.1 (1)		−0.8 (0.3)	0.4 (0.1)	0.7 (0.1)	2.7 (0.2)			−0.1 (0.1)			0.77	0.94
	3	31.2 (5.2)	0.2 (0.1)	−1.9 (0.3)	0.8 (0.3)	0.6 (0.1)		1.4 (0.1)	−0.4 (0.1)	−0.7 (0.1)		−8.5 (2.5)	0.72	0.91
Northwest	1			−1.5 (0.5)	−0.7 (0.3)	−1.9 (0.5)							0.57	0.84
	2	5.4 (1.1)	0.1 (0)	−0.2 (0.1)		0.4 (0.1)	1.5 (0.2)					4.4 (1)	0.62	0.92
	3	25.4 (3.8)	0.4 (0.1)	−0.4 (0.2)			1.2 (0.4)	−0.4 (0.1)				14.7 (2.1)	0.69	0.9
Southwest	1	10.6 (5)	0.3 (0.1)	−0.7 (0.2)		0.5 (0.2)	2.4 (0.3)	0.5 (0.1)				−11 (2.8)	0.69	0.89
	2	5.5 (1.8)		−0.3 (0.1)		0.4 (0.2)	3.5 (0.3)	0.3 (0.1)	0.6 (0.1)				0.6	0.88
	3	18.8 (4.5)		−0.5 (0.2)			0.7 (0.2)		−0.2 (0.1)	−0.3 (0.1)			0.68	0.9
Northern Rockies and Plains	1		0.4 (0.1)			1.2 (0.3)		0.7 (0.2)					0.82	0.95
	2	4.4 (1.5)	0.3 (0.1)			0.3 (0.1)	2.4 (0.2)	0.4 (0.1)				3.1 (1)	0.67	0.92
	3	11.1 (2.1)	0.3 (0.1)				2.1 (0.2)		−0.6 (0.1)	−0.4 (0.1)			0.73	0.92
Upper Midwest	1					0.5 (0.3)		1.3 (0.2)		−0.4 (0.2)			0.79	0.95
	2	4.4 (1.7)	0.3 (0.1)	−0.6 (0.2)		0.9 (0.1)	2.7 (0.3)	1 (0.1)	−0.3 (0.1)			3.7 (1)	0.82	0.95
	3	9.5 (3)	0.4 (0.1)	−0.6 (0.2)	0.3 (0.1)		2.5 (0.2)	0.4 (0.1)	−0.3 (0.1)	−0.2 (0.1)			0.85	0.96
South	1	13.1 (2.2)	0.5 (0)	−0.3 (0.1)			1.4 (0.2)	0.3 (0.1)	−0.2 (0.1)				0.59	0.91
	2		0.2 (0.1)			0.5 (0.1)	4.2 (0.3)	−0.2 (0.1)	0.2 (0.1)		0.3 (0.1)	4.5 (1.1)	0.67	0.94
	3	14.7 (1.9)	0.3 (0)	−0.7 (0.1)		−0.2 (0.1)	1 (0.2)	0.3 (0.1)	−0.4 (0.1)	−0.4 (0.1)	0.3 (0.1)		0.65	0.93
Southeast	1	15.1 (1.9)	0.3 (0)	−0.3 (0.1)		−0.3 (0.1)	0.8 (0.2)	0.5 (0.1)				4.6 (0.9)	0.68	0.94
	2	4.6 (1.6)	0.1 (0)			1.7 (0.2)	7 (0.4)	−0.6 (0.1)	0.2 (0.1)			6.1 (1.1)	0.74	0.95
	3	11.1 (1.4)	0.3 (0)	−0.6 (0.1)		−0.7 (0.1)	0.8 (0.2)	0.7 (0.1)		−0.3 (0.1)		6.9 (1.1)	0.69	0.94
Ohio Valley	1	21.2 (2.9)	0.7 (0)	−0.5 (0.1)		0.4 (0.1)	0.7 (0.2)	0.7 (0.1)	−0.3 (0.1)			5.5 (1)	0.68	0.94
	2	5.7 (1.3)				2.2 (0.1)	5.5 (0.3)	0.3 (0.1)			0.2 (0.1)	2.9 (0.7)	0.74	0.95
	3	14.4 (5.2)	0.3 (0.1)	−0.8 (0.1)			1.7(0.2)	0.5 (0)	−0.3 (0)	−0.3 (0.1)		3.3(1.3)	0.77	0.95
Northeast	1	10.6 (2.8)	0.4 (0.1)					0.9 (0.1)	−0.4 (0.1)				0.8	0.95
	2		−0.2 (0.1)			1.2 (0.2)	6.4 (0.4)	−0.4 (0.1)				8.5 (1.1)	0.84	0.96
	3	31 (2.5)	1.5 (0.2)	−0.8 (0.2)		1.8 (0.2)	1.4 (0.4)					27.5 (2)	0.8	0.95

* All predictors are significant at $\alpha = 0.05$ level.

3.3. Model Cross-Validation

The overall CV R^2 for the entire study area and study period is 0.70 and the slope between predicted $PM_{2.5}$ and the observed $PM_{2.5}$ is 0.98, indicating good agreement between CV estimates and observations. Regional results of random and spatial 10-fold CV including R^2 and slope are listed in Tables 4 and 5. Results show that the CV-based performance of our model varies across regions. For example, our model achieves the highest R^2 under both CV settings (random R^2 = 0.78, spatial R^2 = 0.70) in the Northwest as well as in the Upper Midwest and Ohio Valley regions. On the other hand, the Southwest region has the lowest R^2 of 0.54 for random CV. For spatial CV, our model does not perform as well as for random CV, where in the Northwest region, the random CV R^2 is 0.60 and the spatial CV R^2 is only 0.39. Specifically, Figures 3 and 4 show the scatterplots of CV estimates and observed $PM_{2.5}$ concentration levels across nine climate regions. The CV-based $PM_{2.5}$ estimates have good linear agreement with the observations in the West, South, Upper Midwest, Southwest, Northwest, and Ohio Valley regions. However, in the Southwest, North, and Northwest regions, the model tends to underestimate at higher $PM_{2.5}$ concentrations.

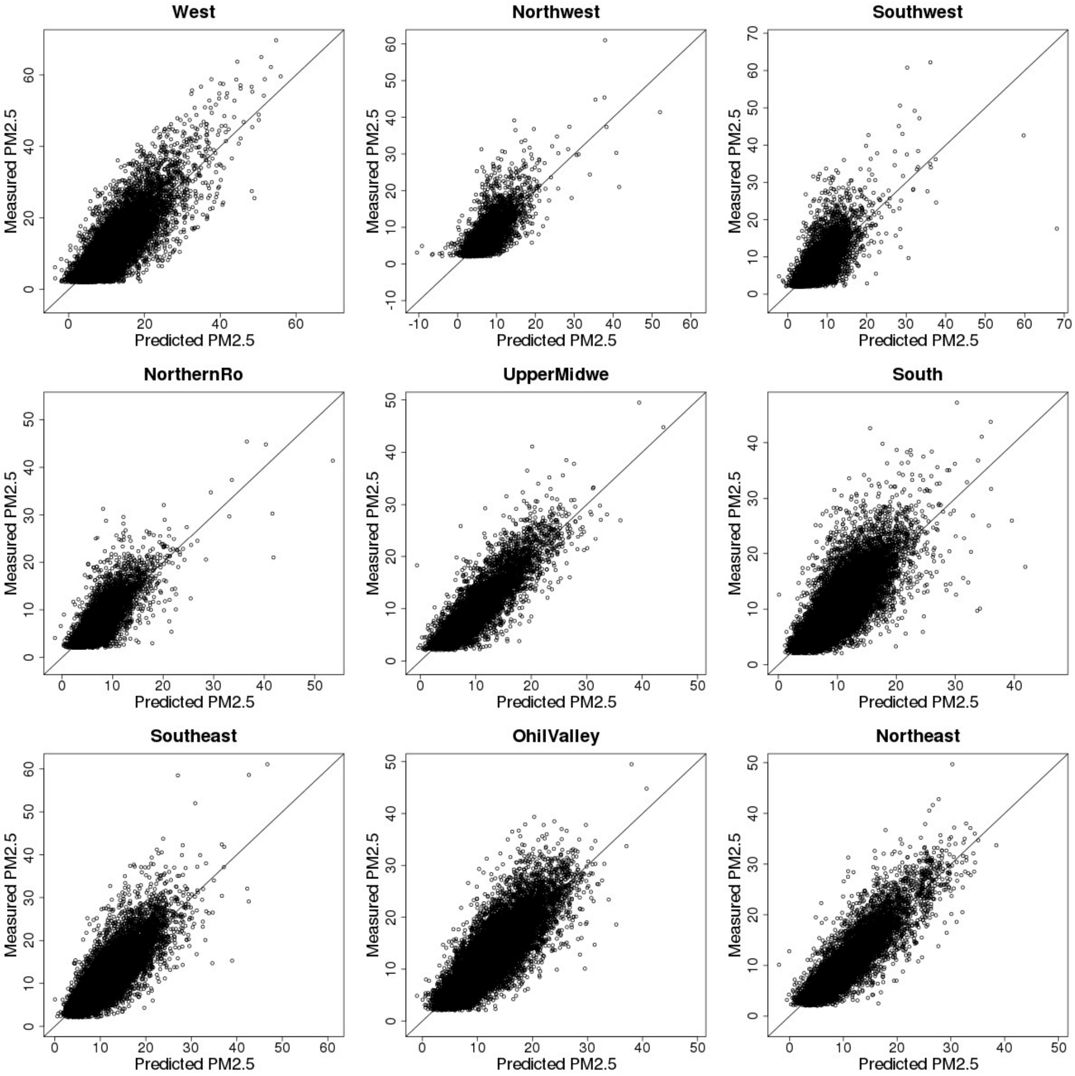

Figure 3. Tenfold cross validation results.

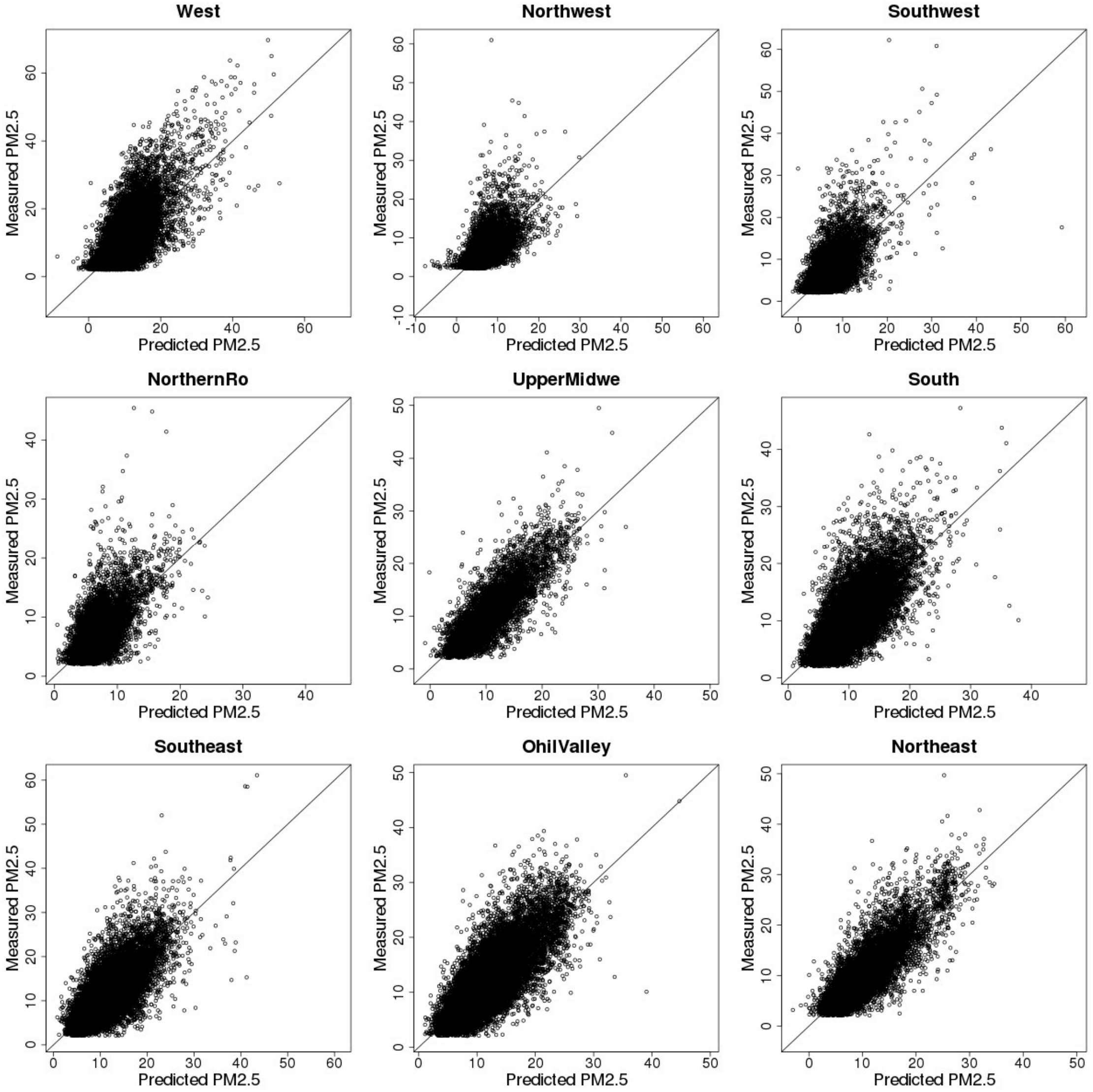

Figure 4. Spatial 10-fold cross validation results.

Table 4. Tenfold cross validation results.

Regions	R^2	Intercept	Slope
West	0.69	0.04	0.99
Northwest	0.60	0.35	0.95
Southwest	0.54	0.40	0.94
Northern Rockies and Plains	0.60	0.29	0.95
Upper Midwest	0.76	−0.04	0.99
South	0.59	0.27	0.97
Southeast	0.69	0.19	0.98
Ohio Valley	0.71	0.07	0.99
Northeast	0.78	0.07	0.99

Table 5. Spatial 10-fold cross validation results.

Regions	R^2	Intercept	Slope
West	0.46	0.36	1.02
Northwest	0.39	1.01	0.83
Southwest	0.40	0.96	0.87
Northern Rockies and Plains	0.37	0.94	0.90
Upper Midwest	0.69	−0.01	0.99
South	0.50	0.38	0.96
Southeast	0.58	0.77	0.92
Ohio Valley	0.65	0.18	0.97
Northeast	0.70	0.33	0.97

3.4. Model Prediction

The predicted annual average $PM_{2.5}$ concentrations and their model-based SDs are visualized in Figures 5 and 6. As shown in Figure 5, the predicted annual mean of $PM_{2.5}$ concentration is smoothed across all the spatial domains, even among the climate buffer regions, indicating that the national Bayesian model fits the data well. Furthermore, a strong spatial differential pattern exists in the annual $PM_{2.5}$ spread, where the $PM_{2.5}$ concentration is higher in eastern regions than western regions. California, the Great Lakes regions, and the east coast regions, including New York and Washington, have especially high annual average $PM_{2.5}$ concentrations. On the other hand, the lowest annual $PM_{2.5}$ concentration is found in Midwestern states, such as Utah, Colorado, Wyoming, and Idaho, with extensive forest coverage and sparse human activities. These indicate that our model can capture large-scale spatial patterns of $PM_{2.5}$ well. Our model is also able to discover the small features of the predicted $PM_{2.5}$ concentration surface, where we can observe high $PM_{2.5}$ concentration levels in urban centers such as Atlanta, Dallas, Houston, Miami, and Salt Lake City.

Regarding prediction uncertainty, Figure 6 shows the spatial spread of the standard deviation of the annual average $PM_{2.5}$ concentrations. The West region, including California and Nevada, has a higher SD compared with other regions. The South and Southeast regions have the lowest SD on average. More specifically, from the spatial distribution of SD, we observed a higher peak at the Miami, Houston, and Dallas areas and Colorado state. Similarly, we visualized the predicted seasonal average $PM_{2.5}$ concentrations and their SDs and the results are in Supplemental Figures S1 and S2.

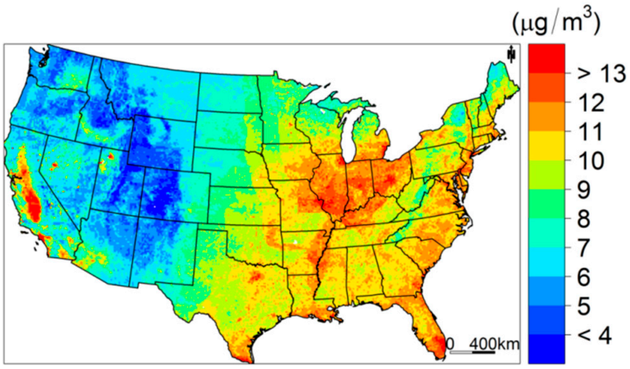

Figure 5. Predicted annual $PM_{2.5}$ concentration across the continental United States.

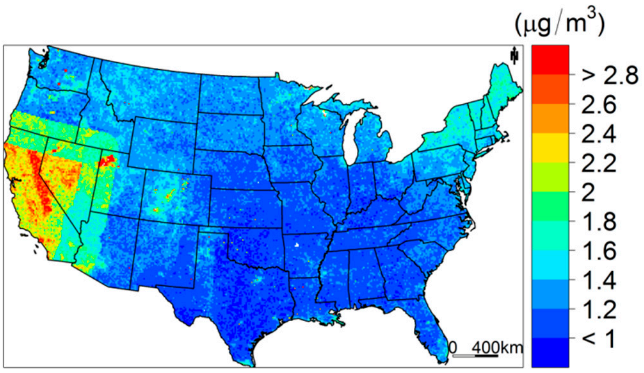

Figure 6. The uncertainty (standard deviation) of the predicted annual $PM_{2.5}$ concentration across the continental United States.

4. Discussion

In our national Bayesian downscaling approach, we first adopted nine climate regions and three temporal regions to separate the data into subregions, which provided more flexible model fitting. Then, we utilized the Bayesian downscaling approaches in each sub-block to quantify the geographical patterns and the association between AOD and $PM_{2.5}$. Compared with regional models, including regional hierarchical models, mixed-effects models, and regional Bayesian downscaling methods, our approach provides nationally cohesive predictions and quantifies the model prediction errors. Compared with machine learning models (e.g., neural networks and random forests), our approach incorporates the core of the statistical approaches, providing insights into the physical and geographical information of the problem. The model uncertainty provided by the Bayesian approach is much more informative than that generated from machine learning models.

Our approach has several strengths. First, it uses a latent spatial process to incorporate spatial correlation, which can borrow information across neighborhoods and is able to make more reliable predictions compared with mixed-effects models. Second, based on the climate region and temporal separation, our model is much more flexible in terms of model fitting and therefore fits the data better than the traditional Bayesian models. As shown in the Table 1, there is a significant difference in the geographical patterns across regions and temporal domains, which is an important sign that in different climate regions, the association between AOD and $PM_{2.5}$ is complex and region specific. This further confirms that using a single model for a whole national domain is not realistic, as it cannot reveal the real physical mechanism in which researchers are interested. Moreover, our proposed approach can perform parallel setting and is much faster than traditional approaches.

Finally, compared with the machine learning method for national calibration, our approach has slightly weaker prediction ability, probably because of the fewer predictors used in our models than in theirs. For example, Di et al. [16] included more than 50 predictors in the neural network model, and the random forest model in Hu et al. [17] contained ~40 predictors. In addition, both approaches used convolutional layers for nearby $PM_{2.5}$ measurements and land use terms in their models, and both studies point out that convolutional layers can help to improve prediction accuracy. Although we could have included these predictors in our models, it would have required additional computing resources and consumed additional computing time. We will address this issue in future research.

5. Conclusions

We presented a national Bayesian downscaler model to estimate daily $PM_{2.5}$ concentrations in the continental United States using satellite aerosol remote sensing data and meteorological and land use parameters. Overall, our national Bayesian downscaling model performs well at the national scale. It has the advantage of explicitly displaying the important predictors of $PM_{2.5}$ in different geographical regions, which allows model simplification and further improvements of model performance. It should be noted that the goal of this approach is to study the geographical patterns across different regions and seasons and our approach successfully provides great insights into these and provides much more informative results than machine learning methods. As we mentioned, the limited prediction ability of our approach in some specific regions, i.e., the South region, is one limitation. The reason for this is that even though we separated the national domain into sub-blocks, each region is still very large, which makes it difficult for a single model to fit across such a large domain. Thus, one future direction of this model is to provide a more flexible approach.

Supplementary Materials: The following are available online at http://www.mdpi.com/1660-4601/15/9/1999/s1, Figure S1: Seasonal Predicted $PM_{2.5}$ Concentration, Figure S2. Standard Deviation of Seasonal Predicted $PM_{2.5}$ Concentration, Text S1: Detailed Model Specification and Estimation Procedure.

Author Contributions: Y.L. conceived this study. H.H.C. and L.A.W. contributed to study design. Y.W. and X.H. performed model development and prepared the manuscript. J.B. processed model input data. All authors commented on the manuscript.

Funding: This work was partially supported by the NASA Applied Sciences Program (grant no. NNX16AQ28G; PI: Liu). This publication was developed under Assistance Agreement No. 83586901 awarded by the US Environmental Protection Agency to Emory University (PI: Liu). It has not been formally reviewed by the EPA. The views expressed in this document are solely those of [name of recipient or names of authors] and do not necessarily reflect those of the Agency. The EPA does not endorse any products or commercial services mentioned in this publication.

Conflicts of Interest: The authors declare no conflict of interest. The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Pope, C.A., III; Dockery, D.W. Health effects of fine particulate air pollution: Lines that connect. *J. Air Waste Manag. Assoc.* **2006**, *56*, 709–742. [CrossRef] [PubMed]
2. Seinfeld, J.H.; Pandis, S.N. *Atmospheric Chemistry and Physics: From air Pollution to Global Change*; John Wiley & Sons: New York, NY, USA, 1998.
3. Anderson, T.; Charlson, R.J.; Winker, D.M.; Ogren, J.A.; Holmen, K. Mesoscale variations of tropospheric aerosols. *J. Atmos. Sci.* **2003**, *60*, 119–136. [CrossRef]
4. Liu, Y.; Sarnat, J.A.; Kilaru, A.; Jacob, D.J.; Koutrakis, P. Estimating ground-level $PM_{2.5}$ in the eastern united states using satellite remote sensing. *Environ. Sci. Technol.* **2005**, *39*, 3269–3278. [CrossRef] [PubMed]
5. Hu, X.; Waller, L.A.; Lyapustin, A.; Wang, Y.; Al-Hamdan, M.Z.; Crosson, W.L.; Estes Jr, M.G.; Estes, S.M.; Quattrochi, D.A.; Puttaswamy, S.J.; et al. Estimating ground-level $PM_{2.5}$ concentrations in the Southeastern United States using MAIAC AOD retrievals and a two-stage model. *Remote Sens. Environ.* **2014**, *140*, 220–232. [CrossRef]
6. Hu, X.; Waller, L.A.; Lyapustin, A.; Wang, Y.; Liu, Y. 10-year spatial and temporal trends of $PM_{2.5}$ concentrations in the southeastern US estimated using high-resolution satellite data. *Atmos. Chem. Phys.* **2014**, *14*, 6301–6314. [CrossRef] [PubMed]
7. Gupta, P.; Christopher, S.A.; Wang, J.; Gehrig, R.; Lee, Y.; Kumar, N. Satellite remote sensing of particulate matter and air quality assessment over global cities. *Atmos. Environ.* **2006**, *40*, 5880–5892. [CrossRef]
8. Ma, Z.; Hu, X.; Sayer, A.M.; Levy, R.; Zhang, Q.; Xue, Y.; Tong, S.; Bi, J.; Huang, L.; Liu, Y. Satellite-based spatiotemporal trends in $PM_{2.5}$ concentrations: China, 2004–2013. *Environ. Health Perspect.* **2015**, *124*, 184–192. [CrossRef] [PubMed]
9. Chang, H.H.; Hu, X.; Liu, Y. Calibrating MODIS aerosol optical depth for predicting daily $PM_{2.5}$ concentrations via statistical downscaling. *J. expo. Sci. Environ. Epidemiol.* **2014**, *24*, 398–404. [CrossRef] [PubMed]
10. Cheng, Q.; Gao, X.; Martin, R. Exact prior-free probabilistic inference on the heritability coefficient in a linear mixed model. *Electron. J. Stat.* **2014**, *8*, 3062–3076. [CrossRef]
11. Hu, X.; Waller, L.A.; Al-Hamdan, M.Z.; Crosson, W.L.; Estes Jr, M.G.; Estes, S.M.; Quattrochi, D.A.; Sarnat, J.A.; Liu, Y. Estimating ground-level $PM_{2.5}$ concentrations in the southeastern U.S. using geographically weighted regression. *Environ. Res.* **2013**, *121*, 1–10. [CrossRef] [PubMed]

12. Hu, X.; Waller, L.A.; Lyapustin, A.; Wang, Y.; Liu, Y. Improving satellite-driven $PM_{2.5}$ models with Moderate Resolution Imaging Spectroradiometer fire counts in the southeastern U.S. *J. Geophys. Res. Atmos.* **2014**, *119*, 11375–11386. [CrossRef] [PubMed]
13. Lee, H.; Liu, Y.; Coull, B.; Schwartz, J.; Koutrakis, P. A novel calibration approach of MODIS AOD data to predict $PM_{2.5}$ concentrations. *Atmos. Chem. Phys.* **2011**, *11*, 7991–8002. [CrossRef]
14. Lee, H.J.; Coull, B.A.; Bell, M.L.; Koutrakis, P. Use of satellite-based aerosol optical depth and spatial clustering to predict ambient $PM_{2.5}$ concentrations. *Environ. Res.* **2012**, *118*, 8–15. [CrossRef] [PubMed]
15. Kloog, I.; Koutrakis, P.; Coull, B.A.; Lee, H.J.; Schwartz, J. Assessing temporally and spatially resolved $PM_{2.5}$ exposures for epidemiological studies using satellite aerosol optical depth measurements. *Atmos. Environ.* **2011**, *45*, 6267–6275. [CrossRef]
16. Di, Q.; Koutrakis, P.; Schwartz, J. A hybrid prediction model for $PM_{2.5}$ mass and components using a chemical transport model and land use regression. *Atmos. Environ.* **2016**, *131*, 390–399. [CrossRef]
17. Hu, X.; Belle, J.H.; Meng, X.; Wildani, A.; Waller, L.A.; Strickland, M.J.; Liu, Y. Estimating $PM_{2.5}$ concentrations in the conterminous United States using the Random Forest Approach. *Environ. Sci. Technol.* **2017**, *51*, 6936–6944. [CrossRef] [PubMed]
18. Thomas, K.; Walter, J.K. *Regional and National Monthly, Seasonal, and Annual Temperature Weighted by Area, 1895–1983*; National Climatic Data Center: Asheville, NC, USA, 1984.
19. Andrieu, C.; De Freitas, N.; Doucet, A.; Jordan, M.I. An introduction to MCMC for machine learning. *Mach. Learn.* **2003**, *50*, 5–43. [CrossRef]
20. Wang, Y.; Zhao, Y.; Zhang, L.; Liang, J.; Zeng, M.; Liu, X. Graph construction based on re-weighted sparse representation for semi-supervised learning. *J. Inf. Comput. Sci.* **2013**, *10*, 375–383.