



Published in final edited form as:

*J Geophys Res Atmos.* 2014 October 16; 119(19): 11375–11386. doi:10.1002/2014JD021920.

## Improving satellite-driven $PM_{2.5}$ models with Moderate Resolution Imaging Spectroradiometer fire counts in the southeastern U.S

Xuefei Hu<sup>1</sup>, Lance A. Waller<sup>2</sup>, Alexei Lyapustin<sup>3</sup>, Yujie Wang<sup>3,4</sup>, and Yang Liu<sup>1</sup>

<sup>1</sup>Department of Environmental Health, Rollins School of Public Health, Emory University, Atlanta, Georgia, USA

<sup>2</sup>Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, Georgia, USA

<sup>3</sup>NASA Goddard Space Flight Center, Greenbelt, Maryland, USA

<sup>4</sup>Goddard Earth Sciences and Technology Center, University of Maryland, Baltimore County, Catonsville, Maryland, USA

### Abstract

Multiple studies have developed surface  $PM_{2.5}$  (particle size less than 2.5  $\mu\text{m}$  in aerodynamic diameter) prediction models using satellite-derived aerosol optical depth as the primary predictor and meteorological and land use variables as secondary variables. To our knowledge, satellite-retrieved fire information has not been used for  $PM_{2.5}$  concentration prediction in statistical models. Fire data could be a useful predictor since fires are significant contributors of  $PM_{2.5}$ . In this paper, we examined whether remotely sensed fire count data could improve  $PM_{2.5}$  prediction accuracy in the southeastern U.S. in a spatial statistical model setting. A sensitivity analysis showed that when the radius of the buffer zone centered at each  $PM_{2.5}$  monitoring site reached 75 km, fire count data generally have the greatest predictive power of  $PM_{2.5}$  across the models considered. Cross validation (CV) generated an  $R^2$  of 0.69, a mean prediction error of 2.75  $\mu\text{g}/\text{m}^3$ , and root-mean-square prediction errors (RMSPEs) of 4.29  $\mu\text{g}/\text{m}^3$ , indicating a good fit between the dependent and predictor variables. A comparison showed that the prediction accuracy was improved more substantially from the nonfire model to the fire model at sites with higher fire counts. With increasing fire counts, CV RMSPE decreased by values up to 1.5  $\mu\text{g}/\text{m}^3$ , exhibiting a maximum improvement of 13.4% in prediction accuracy. Fire count data were shown to have better performance in southern Georgia and in the spring season due to higher fire occurrence. Our findings indicate that fire count data provide a measurable improvement in  $PM_{2.5}$  concentration estimation, especially in areas and seasons prone to fire events.

### 1. Introduction

Previous studies have found strong associations between  $PM_{2.5}$  (particle size less than 2.5  $\mu\text{m}$  in aerodynamic diameter) and various adverse health outcomes including respiratory and

cardiovascular diseases [Dominici et al., 2006]. Thus, obtaining accurate spatiotemporally resolved PM<sub>2.5</sub> concentrations is crucial to assess air pollution levels and address public health concerns.

Although stationary ambient monitors have been established to measure ground-level PM<sub>2.5</sub> exposure, satellite remote sensing estimated concentrations have the potential to expand the coverage of ground networks and the ability to improve estimates of population exposure to PM<sub>2.5</sub> [van Donkelaar et al., 2010]. Many satellite aerosol optical depth (AOD) products have been used to estimate PM<sub>2.5</sub> concentrations, such as those from the Moderate Resolution Imaging Spectroradiometer (MODIS), the Multiangle Imaging Spectroradiometer (MISR), and the Geostationary Operational Environmental Satellite Aerosol/Smoke Product [Hu et al., 2013; Liu et al., 2007, 2005; Paciorek et al., 2008]. Recently, a new AOD product with 1 km spatial resolution derived by a multiangle implementation of atmospheric correction (MAIAC) algorithm has been reported [Lyapustin et al., 2011b]. Hu et al. [2014b] successfully estimated PM<sub>2.5</sub> exposure at 1 km resolution in the southeastern U.S. using the MAIAC AOD. Hu et al. [2014a] further used the MAIAC AOD estimated PM<sub>2.5</sub> concentrations to investigate the 10 year spatial and temporal trends of PM<sub>2.5</sub> in the southeastern U.S.

The quantitative relationships between PM<sub>2.5</sub> and AOD have been established by multiple studies for various regions. Wang and Christopher [2003] developed empirical relationships between remotely sensed AOD and surface PM<sub>2.5</sub> in the southeastern U.S. However, using AOD as the only PM<sub>2.5</sub> concentration predictor may be insufficient to fully account for the complex PM<sub>2.5</sub>-AOD relationship. Liu et al. [2005] found that meteorological, geographical, and seasonal conditions can strongly affect the association between PM<sub>2.5</sub> and MISR AOD. As a result, these factors need to be incorporated when establishing the PM<sub>2.5</sub>-AOD relationship. To accommodate for this, previous studies have developed statistical models to predict PM<sub>2.5</sub> concentrations using satellite-derived AOD as the primary predictor and meteorological fields and land use variables as secondary predictors. For instance, Liu et al. [2007] incorporated surface temperature, surface wind speed, and mixing height in a predictive model and found that they are all significant predictors of PM<sub>2.5</sub> concentrations. Liu et al. [2009] further pointed out that road length and population density within areas are effective predictors of PM<sub>2.5</sub>. Kloog et al. [2011] introduced a number of useful land use variables (e.g., elevation, percent of open spaces, PM<sub>2.5</sub> point emissions, and area source PM<sub>2.5</sub> emissions) that can potentially be incorporated into statistical models for PM<sub>2.5</sub> concentration prediction. In addition, Hu et al. [2013] reported that forest cover is significantly associated with PM<sub>2.5</sub>. However, to our knowledge, to date, fire data have not been used in statistical models for PM<sub>2.5</sub> concentration estimation. Fires, including wildfires and prescribed burning, are important emission sources of primary air pollutants and precursors of secondary pollutants. Wildfires can be a natural disaster threatening human life and property, while prescribed burning is a vegetation management technique to control vegetation, enhance biotic production and diversity, control diseases and insects, and reduce dead fuel accumulation. Prescribed burning is widely used in the United States, especially in the southeast [Hardy et al., 2001]. Zeng et al. [2008] pointed out that prescribed burning emissions can lead to a maximum increase of 25 µg/m<sup>3</sup> in PM<sub>2.5</sub> concentrations within a day. Tian et al. [2009] indicated that 55% and 80% of PM<sub>2.5</sub> mass were attributed to

prescribed burning in Georgia in January and March 2002, respectively. *Zhang et al.* [2010] reported that 13% of  $PM_{2.5}$  mass concentrations were contributed by biomass burning annually in the southeastern U.S., indicating that fires are significant contributors to  $PM_{2.5}$  levels. In addition, *Hu et al.* [2014b] predicted  $PM_{2.5}$  concentrations in the southeastern U.S. and hypothesized that abnormally high concentrations in areas such as southern Georgia were partially caused by fires. As a result, fire data could be potentially useful predictors of  $PM_{2.5}$ , particularly in areas where fire events frequently occur. Establishing accurate biomass burning emissions inventories is rather difficult, and another approach is to use observations of active fires [*Al-Saadi et al.*, 2008]. Satellite remote sensing has provided a consistent means of detecting active burning through various sensors such as MODIS and advanced very high resolution radiometer [*Cardoso et al.*, 2005; *Justice et al.*, 2002]. The primary objective of this study is to investigate how remotely sensed fire count data can be incorporated in a spatial statistical model setting to predict surface  $PM_{2.5}$  concentrations and whether this parameter can significantly improve  $PM_{2.5}$  prediction accuracy, especially in areas where fires frequently occur. In addition, sensitivity analyses of fire counts on prediction accuracy of surface  $PM_{2.5}$  concentrations were conducted to preliminarily examine how much predictive power fire count data have at various levels of fire occurrence.

## 2. Materials and Methods

### 2.1. Study Area

The study area is approximately  $1800 \times 1200$  km<sup>2</sup> in the southeastern U.S. (Figure 1). This region contains various land cover in which wildfires and prescribed burning are prone to occur, including forestland, agricultural land, pasture/hay, and swamp (e.g., peat fires in the Okefenokee Swamp).

### 2.2. $PM_{2.5}$ Measurements

Twenty-four hour averaged (daily mean)  $PM_{2.5}$  concentrations in the study domain from 2007 were collected from the U.S. Environmental Protection Agency (EPA) federal reference method samplers and downloaded from the EPA's Air Quality System Technology Transfer Network (<http://www.epa.gov/ttn/airs/airsaqs/>).  $PM_{2.5}$  concentrations less than  $2 \mu\text{g}/\text{m}^3$  (~2.5% of total records) are below the established limit of detection and were filtered [*Environmental Protection Agency*, 2008].

### 2.3. Remote Sensing Data

**2.3.1. AOD Data**—We obtained 2007 MAIAC AOD data from the National Aeronautics Space Administration (NASA) Goddard Space Flight Center. MAIAC used the time series of MODIS measurements and simultaneous processing of a group of pixels in fixed  $25 \times 25$  km<sup>2</sup> blocks to retrieve aerosol parameters over land at 1 km resolution [*Lyapustin et al.*, 2011a, 2011b, 2012]. Validation showed that the accuracy of MAIAC is similar to that of operational Collection 5 MODIS Dark Target AOD over dark and vegetated surfaces and better over most urban regions [*Lyapustin et al.*, 2011b]. The improvement is mostly related to the fact that MAIAC retrieves the spectral regression coefficient to assess surface reflectance at visible wavelengths from the  $2.1 \mu\text{m}$  measurements, while the Dark Target algorithm relies instead on empirical parameterization. In this paper, we combined and took

the average of Terra and Aqua AOD values in order to increase the spatial coverage. The averaged measurements from Aqua and Terra are suitable for predicting daily  $PM_{2.5}$  concentrations, because they are a good approximation of the daily mean AOD values, as pointed out by *Zhang et al.* [2012]. We adopted the approach of *Puttaswamy et al.* [2013] and *Hu et al.* [2014b] when one of the AOD products is missing. In the instance of missing values, a linear regression was fitted to establish the association between daily mean Terra-MAIAC and Aqua-MAIAC AOD values. The missing AOD value was estimated by this regression equation, and daily AOD values were obtained by averaging both the observed and estimated values. Finally, in addition to the internal cloud mask provided by MAIAC [*Lyapustin et al.*, 2008], an upper bound of 2.0 was established for the combined AOD values (~0.1% of data were filtered) to reduce potential cloud contamination. We employed the same method in our previous studies [*Hu et al.*, 2014a, 2014b].

**2.3.2. Fire Data**—We obtained the fire count data for the study area from the U.S. Department of Agriculture Forest Service's Remote Sensing Applications Center for 2007 (<http://activefiremaps.fs.fed.us>). These data include Terra and Aqua MODIS fires and thermal anomalies from the NASA MCD14ML product, collection 5.1. These data were downloaded in the Environmental Systems Research Institute (ESRI) shapefile format and provided the centroids of the 1 km fire detections. *Hawbaker et al.* [2008] reported that the MODIS active fire product may underrepresent fires in regions with frequent cloud cover or rapid burning, as well as small and low-intensity fires. *Hantson et al.* [2013] also found that the detection rate of the MODIS fire product is low for small fires. Although MODIS fire data also provide fire radiative power (FRP), we found that fire counts yielded slightly more accurate predictions than FRP, and thus, we used fire counts in this analysis. To link fire count data to each  $PM_{2.5}$  monitoring site, buffer zones centered at each  $PM_{2.5}$  monitoring site with different radii were generated to implement the sensitivity analysis of the impact of radius on prediction accuracy of  $PM_{2.5}$  concentrations. The radius that exhibited the greatest improvement in prediction accuracy was selected for this analysis. The fire counts within each buffer were assigned to the corresponding  $PM_{2.5}$  monitoring site. Because our objective was to add the fire predictor to improve the overall prediction accuracy of  $PM_{2.5}$  across the entire study area, we used all  $PM_{2.5}$  monitoring sites for radius selection. It should be noted that the optimum radius may vary for different study areas, including other subregions in the southeastern U.S. and regions outside of the southeastern U.S.

## 2.4. Meteorological Fields

The meteorological fields were downloaded at a spatial resolution of ~13 km from the North American Land Data Assimilation System (NLDAS) (<http://ldas.gsfc.nasa.gov/nldas/>) website, including relative humidity and wind speed. The NLDAS provided quality-controlled, spatiotemporally consistent, real-time, and retrospective forcing data sets [*Cosgrove et al.*, 2003]. We adopted the method used by *Hu et al.* [2014b] that employs hourly averaged NLDAS measurements from 10 A.M. to 4 P.M. local standard time as daytime meteorological parameters to match the satellite overpass times.

## 2.5. Land Use Variables

We downloaded elevation data at a spatial resolution of 1 arc sec (~30 m) from the national elevation data set (<http://ned.usgs.gov>). The A1 road data (limited access highway) were obtained from ESRI StreetMap USA (Environmental Systems Research Institute, Inc., Redland, CA). A summed value of road length within a  $1 \times 1$  km<sup>2</sup> MAIAC grid cell was calculated for that grid cell, and a value of zero was assigned to grid cells that had no roads. 2006 Landsat-derived land cover maps encompassing the study region at a spatial resolution of 30 m were downloaded from the National Land Cover Database (<http://www.mrlc.gov>). To generate forest cover maps, the value of one was assigned to forest pixels and a value of zero was assigned to others. We obtained primary PM<sub>2.5</sub> emissions (t/yr) from the 2008 EPA National Emissions Inventory facility emissions reports. A summed value of point emissions within a  $1 \times 1$  km<sup>2</sup> MAIAC grid cell was calculated for that grid cell, and a value of zero was assigned to grid cells with no emissions.

## 2.6. Data Integration

We reprojected the data to the USA Contiguous Albers Equal Area Conic U.S. Geological Survey coordinate system. For model fitting, a nearest neighbor approach was used to link meteorological fields and AOD values to PM<sub>2.5</sub> monitoring sites. Averaged forest cover and elevation and summed road length and point emissions over the  $1 \times 1$  km<sup>2</sup> square buffer centered at each PM<sub>2.5</sub> monitoring site were assigned to the corresponding PM<sub>2.5</sub> monitoring site. For concentration prediction, we performed the same procedure for each  $1 \times 1$  km<sup>2</sup> MAIAC grid cell. Fire data were linked to each monitoring site and MAIAC pixel centroid by calculating the fire counts within a buffer centered at each monitoring site and MAIAC pixel centroid with a selected radius.

## 2.7. Model Structure and Validation

To estimate PM<sub>2.5</sub> concentrations using satellite AOD, *Lee et al.* [2011] and *Kloog et al.* [2011] developed linear mixed effects (LMEs) models to account for the day-to-day variability within the relationship between PM<sub>2.5</sub> and AOD. *Hu et al.* [2014b] introduced a second-stage model using geographically weighted regression (GWR) to explain potential spatial variability in the relationship. This method achieved reasonable prediction accuracy and was adopted for use in this study. In addition, some of the predictors are local variables and some could be universal. Whether a predictor should be incorporated needs to be tested in the model. *Hu et al.* [2014a] found that for various years, the best prediction accuracies were obtained by using different combinations of predictors. We tested a number of different predictors in the model, including air temperature, mixing height, and area source emissions, to find the most suitable combination of predictors that would yield the best prediction accuracy. Since the primary purpose of this paper was to investigate whether fire data can be used to improve prediction accuracy of PM<sub>2.5</sub> concentrations, particularly in areas with high fire occurrence, fire parameters were toggled on and off in the two-stage model to compare performances of models with and without fire predictors (hereinafter referred to as the “fire model” and the “nonfire model”). The model structure can be expressed as follows:

$$PM_{2.5,st} = (b_0 + b_{0,t}) + (b_1 + b_{1,t})AOD_{st} + (b_2 + b_{2,t})RelativeHumidity_{st} + (b_3 + b_{3,t})WindSpeed_{st} + b_4Elevation_s + b_5MajorRoads_s + b_6ForestCover_s + b_7PointEmissions_s + b_8FireCountData_s + \varepsilon_{st}$$

$$(b_{0,t}, b_{1,t}, b_{2,t}, b_{3,t}) \sim N[(0, 0, 0, 0), \Psi] \quad (1)$$

$$PM_{2.5-resi_{st}} = \beta_{0,s} + \beta_{1,s}AOD_{st} + \varepsilon_{st}$$

where equation (1) is the model structure for the fire model. The nonfire model has the same model structure to equation (1) but does not include fire count data as a predictor. For the first-stage LME model,  $b_0$  and  $b_{0,t}$  are the fixed and day-specific random intercepts, respectively, and  $b_i$  ( $i = 1, 2, 3, \dots, 8$ ) and  $b_{i,t}$  ( $i = 1, 2, 3$ ) are the fixed and day-specific random slopes for predictors, respectively;  $PM_{2.5,st}$ ,  $AOD_{st}$ , relative humidity $_{st}$ , and wind speed $_{st}$  are the  $PM_{2.5}$  measurements ( $\mu\text{g}/\text{m}^3$ ), the MAIAC AOD value (unitless), the relative humidity value (%), and the wind speed value (m/s) at site  $s$  in day  $t$ , respectively; elevation $_s$ , major roads $_s$ , forest cover $_s$ , point emissions $_s$ , and fire count data $_s$  are the elevation value (m), the road length value (m), the forest cover value, point emissions (t/yr), and the fire count data at site  $s$ , respectively; and  $\Psi$  is an unstructured variance-covariance matrix for the random effects. Element of  $\Psi$  in position  $i$  and  $j$  is the covariance between the  $i$ th and  $j$ th elements of a vector of random variables. There are several possible covariance structures for  $\Psi$ . We selected unstructured because unstructured covariance matrix has no a priori constraint and each variance and covariance is estimated uniquely from the data using a likelihood-based method. As a result, it can achieve the best possible model fit. For the second-stage GWR model,  $PM_{2.5-resi_{st}}$  and  $AOD_{st}$  denote the residuals from the stage one model and the MAIAC AOD value (unitless) at site  $s$  in month  $t$ , respectively.  $\beta_{0,s}$  and  $\beta_{1,s}$  are the location-specific intercept and slope, respectively. The first stage of the model can explain the day-to-day variability within the relationship between  $PM_{2.5}$  and AOD, while the second stage is able to account for the spatial variability in the association. Model fitting estimates the fixed and day-specific random intercepts and slopes for the first-stage LME model and location-specific intercepts and slopes for the second-stage GWR model. After we obtained those intercepts and slopes, we calculated the two components of the  $PM_{2.5}$  prediction from the first- and second-stage models separately and then added them together. Validation was conducted by calculating a variety of statistical indicators between fitted values from the model and observations such as the coefficient of determination ( $R^2$ ), mean prediction error (MPE), and root-mean-square prediction errors (RMSPEs). The MPE and RMSPE are defined as

$$MPE = \frac{\sum_{i=1}^n |\hat{y}_i - y_i|}{n} \quad (2)$$



$$\text{RMSPE} = \sqrt{\frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{n}} \quad (3)$$

where  $\hat{y}_i$  is the predicted  $\text{PM}_{2.5}$  concentrations,  $y_i$  is the observed values, and  $n$  is the total number of the observations. Additionally, we adopted a tenfold cross-validation (CV) technique to assess potential model overfitting and randomly split the entire data set into 10 subsets with each subset containing approximately 10% of the data [Kohavi, 1995]. In each round of cross validation, one subset was used for model testing, and the remaining nine subsets were used for model fitting. Predictions were made for the held-out subset. We repeated the process 10 times to test every subset and calculated  $R^2$ , MPE, and RMSPE between CV-predicted values and observations to assess potential model overfitting.

### 3. Results

#### 3.1. Radius Selection

To assign fire counts to each  $\text{PM}_{2.5}$  monitor and MAIAC grid centroid, a buffer zone centered at each monitor and centroid was generated, and the buffer zones radii were determined. The radius is believed to be related to the distance that fire smoke plumes could transport, on average, during the study period and within the study region. For radius selection, fire counts calculated in buffers with different radii were used in cross validation, and CV RMSPE was then calculated for each radius. The radius that generated the lowest CV RMSPE value (representing the best prediction accuracy) was adopted in this analysis. The result of radius selection is illustrated in Figure 2, and it showed that the lowest CV RMSPE was achieved when the radius reached 75 km. As a result, we selected 75 km as the buffer radius in this paper, regardless of the fact that the differences of CV RMSPE generated from various radii were actually small and inconsequential. The RMSPE is insensitive to the buffer radii considered, because fire counts within various radii are highly correlated. For all monitoring sites, the correlation coefficient ( $r$ ) ranged from 0.05 to 0.997 with a mean of 0.63.

#### 3.2. Descriptive Statistics

Table 1 summarizes the descriptive statistics of dependent and predictor variables used in model fitting. The annual mean  $\text{PM}_{2.5}$  concentrations in 2007 ranged from 2 to  $145 \mu\text{g}/\text{m}^3$ , the annual mean AOD values extended from 0 to 1.95, and the fire counts (calculated within a 75 km buffer) ranged from 0 to 91, indicating frequent fire occurrence near some  $\text{PM}_{2.5}$  monitoring sites. Figure 3 displays the seasonal distribution of fire counts in the study domain in 2007. Fire occurrence was the highest in spring and lowest in winter.

#### 3.3. The Impact of Fires on $\text{PM}_{2.5}$ Predictions

Our first-stage linear mixed effects model generated a slope of 0.20 ( $p < 0.0001$ ) for the fire predictor, indicating a statistically significant and positive relationship between  $\text{PM}_{2.5}$  concentrations and local fire counts. The accuracy assessment results showed that our fire

model performed well. Model fitting generated an  $R^2$  of 0.75, MPE of  $2.47 \mu\text{g}/\text{m}^3$ , and RMSPE of  $3.83 \mu\text{g}/\text{m}^3$ , while cross validation (CV) generated an  $R^2$  of 0.69, MPE of  $2.75 \mu\text{g}/\text{m}^3$ , and RMSPE of  $4.29 \mu\text{g}/\text{m}^3$ . These results indicate a reasonable fit compared to previous studies [Hu et al., 2014a]. To further assess the extent of predictive power that fire data can add to our model, the performances of the fire and nonfire models were compared. Table 2 lists the CV statistics of the two models, and the results showed that by adding the fire predictor to the model, the overall prediction accuracy was improved. The CV RMSPE decreased by  $0.025 \mu\text{g}/\text{m}^3$ , representing a 0.58% reduction. However, the improvement was small because 62.3% of  $\text{PM}_{2.5}$  monitoring sites were not affected by fires (zero fire spots within their buffers), and the mean fire count for all monitoring sites was low (1.41 in this study). For those sites unaffected by fires, prediction accuracy was still improved with the overall CV RMSPE reduced by  $0.021 \mu\text{g}/\text{m}^3$ . Both fire and nonfire models overestimated the  $\text{PM}_{2.5}$  concentrations, while the fire model produced better model fit and yielded less bias than the nonfire model. The reason for the observed behavior of this model needs to be further explored. On the other hand, the CV RMSPE for three  $\text{PM}_{2.5}$  monitoring sites located in southern Georgia, including Albany, Brunswick, and Valdosta, decreased  $0.65 \mu\text{g}/\text{m}^3$ , an 11.8% drop, and the mean fire count for the three monitoring sites was 5.26, as compared to a mean of 1.41 for all monitoring sites. This result could suggest that the fire predictor performed better in areas with higher fire occurrence. With fire counts as a predictor, the estimated  $\text{PM}_{2.5}$  concentrations are also more accurate in spring. That is, CV RMSPE was improved by 1.07% for the entire study area and by 25.6% for southern Georgia in spring, while the improvement was 0.25% and 1.64% for the study area and southern Georgia, respectively, in other seasons. Likewise, Figure 4 illustrates that the performances of the fire and nonfire models were largely similar for the entire study area and period, and the slope indicates the overall bias of our predicted concentrations away from the observations. For the study area, when predictions from both models were regressed against observations, the same slopes of 0.94 were obtained, and a slope of 1 was reached when they were regressed against each other, indicating similar performances of two models. However, in southern Georgia, a slope of 0.84 was obtained when predictions from the fire model were regressed against observations, while the slope dropped to 0.8 when those from the nonfire model were regressed against observations. Moreover, the slope increased from 0.65 to 0.73 in spring when the fire predictor was incorporated, indicating reduced biases. These results suggest that fire count data are an effective predictor of  $\text{PM}_{2.5}$  in regions and periods with high fire occurrence. To further evaluate the performance of the fire predictor in the model for areas with various levels of fire occurrence, we established a number of thresholds for fire counts. We then calculated the differences of CV RMSPE between the fire and nonfire models separately for the  $\text{PM}_{2.5}$  monitors with fire counts above the thresholds (Figures 5a and 5b) and found that the higher the fire counts in buffers centered at monitoring sites, the greater the improvement the fire model achieved over the nonfire model at those sites. The decrease of CV RMSPE exhibits a general inclining trend as the threshold of fire counts become larger with decrease up to  $1.5 \mu\text{g}/\text{m}^3$ ; a maximum drop of 13.4% in CV RMSPE. We further calculated the differences of CV RMSPE between the fire and nonfire models for the three  $\text{PM}_{2.5}$  monitoring sites located in southern Georgia (Figures 5c–5f), since differences between  $\text{PM}_{2.5}$  concentrations estimated from the fire and nonfire models primarily occurred in this area. The results showed that the reduction of CV



RMSPE increased as the mean fire count increased, indicating greater improvement of prediction accuracy from the fire model at sites with higher fire counts. In Valdosta, the annual reduction was  $0.9 \mu\text{g}/\text{m}^3$  when the mean fire count reached seven; a 14.33% improvement in CV RMSPE. Furthermore, a  $2.69 \mu\text{g}/\text{m}^3$  reduction in spring was detected when the mean fire count reached 25 in Valdosta; a 29.06% decrease in CV RMSPE.

### 3.4. Annual and Seasonal Mean $\text{PM}_{2.5}$ Concentrations

The annual and seasonal mean  $\text{PM}_{2.5}$  concentrations estimated from both the fire and nonfire models are shown in Figure 6. The results showed that the patterns of  $\text{PM}_{2.5}$  estimates from both models were largely similar. For instance, high concentrations occurred in large cities and along major highways, while low concentrations appeared in mountainous and rural areas. Figure 6 also shows that the differences between  $\text{PM}_{2.5}$  concentrations estimated from the fire and nonfire models match well with the spatial distribution of the fire density. That is, large differences occurred in areas with high fire density, while small differences appeared in areas with low fire density. These differences primarily occurred in the southeastern part of our domain, particularly in southern Georgia where fires frequently occur. In this region,  $\text{PM}_{2.5}$  estimates generated from the fire model were higher than those estimated from the nonfire model. In conjunction with our accuracy assessment results for the three  $\text{PM}_{2.5}$  monitoring sites in southern Georgia, the results suggested that high  $\text{PM}_{2.5}$  concentrations in this region were partially a result of fires, and these contributions could be effectively captured by adding fire count data as a predictor to our model, especially for the areas with high fire occurrence. The comparison between seasonal mean  $\text{PM}_{2.5}$  estimates generated from the fire and nonfire models showed that the differences were larger in spring than those in the rest of the year, which was consistent with the seasonal distribution of fire counts. The mean relative differences in the study region were 1.2%, 2.3%, 0.1%, and 1.3% for winter, spring, summer, and fall, respectively. The fires mainly occurred in spring, and as a result, fire contributions to  $\text{PM}_{2.5}$  concentrations were also high in spring. Other seasons had lower fire contributions due to fewer observed fire occurrences. There were also some noticeable differences along the Mississippi River and in southern Alabama in fall. Investigating the  $\text{PM}_{2.5}$  monitoring sites in these regions, we found a  $0.01 \mu\text{g}/\text{m}^3$  (0.2%) improvement in CV RMSPE in the Mississippi River region in fall and a  $0.08 \mu\text{g}/\text{m}^3$  (2%) improvement in CV RMSPE in southern Alabama when the fire predictor was incorporated. This improvement was small, which might be due to lower fire occurrence in fall as compared to that in spring.

## 4. Discussion

A strength of our analysis was that we used high spatial resolution MAIAC AOD data to estimate high spatial resolution daily average  $\text{PM}_{2.5}$  concentrations. This method made it possible to examine whether fire counts can provide additional predictive power for  $\text{PM}_{2.5}$  concentration prediction at a much finer scale than using MODIS or MISR. Higher resolution can help reveal more local details of the relationship and thus is more desirable. Our results showed that the differences between  $\text{PM}_{2.5}$  concentrations estimated from the fire and nonfire models matched well with the fire density. Furthermore, our results showed that incorporating fire data as a predictor can improve  $\text{PM}_{2.5}$  prediction accuracy, and fire

data may partially account for  $PM_{2.5}$  variability in the model, which suggests that elevated  $PM_{2.5}$  levels cannot be fully explained by AOD. This may be because some fire-generated smoke plumes were missed by Terra and Aqua and therefore were not captured by MAIAC AOD, while fire spots can remain detectable at the thermal infrared wavelengths for a significantly long period of time. Thus, at some locations, fire count data may introduce additional  $PM_{2.5}$  emissions sources missed by MODIS AOD into the model and lead to more accurate predictions of  $PM_{2.5}$  at those locations. However, fire count data can only partially supplement AOD in  $PM_{2.5}$  predictive models since MODIS fire data miss many small and short-lived fires. In addition, the improvement of accuracy of  $PM_{2.5}$  estimates over the entire study area was rather small because the distribution of fire spots was spatiotemporally unbalanced. Since fires occurred more frequently in southern Georgia and in spring than other areas and seasons, the improvement of prediction accuracy is greater in this area and season.

One of the difficulties of this analysis was linking fire count data to each  $PM_{2.5}$  monitoring site and MAIAC pixel centroid. The dispersion and transport of fire smoke plumes are affected by meteorological conditions such as boundary layer height, precipitation, wind direction, and speed. As a result, it is a challenge to determine which fire incident contributes fine particles to a particular monitoring site or MAIAC pixel centroid. In this paper, we defined a buffer centered at each  $PM_{2.5}$  monitoring site and MAIAC pixel centroid, and the buffer radius was determined by comparing the performances of the models with fire counts calculated using various buffer radii. The radius that can reduce the CV RMSPE between predicted values and observations the most was selected in this analysis. Although this simplification could bias our estimation, this method was effective in our two-stage model setting. Another potential problem regarding linking fire counts to  $PM_{2.5}$  monitoring sites and MAIAC pixel centroids was lag time. Transport of fire smoke to those locations took time and mainly depended on meteorological conditions, such as wind speed. However, the mean wind speed in our domain was 3.85 m/s, which translated to an air mass movement of ~332 km within a day. Given that this movement is much farther than the 75 km buffer radius, we used same day fire counts to estimate the  $PM_{2.5}$  concentrations.

One limitation of the study is that only fire count data were used in this analysis, not including fire size and intensity data. We might expect that fire contributions to  $PM_{2.5}$  concentrations between fire incidents would vary. Our statistical model, however, assigned the same weight to every fire incident; this may be a source of bias. Although we do not expect that the resulting prediction errors were large because the fires in the study region were generally small, further analyses using more comprehensive fire data, including fire size and intensity, are needed for more robust modeling, especially for applying this model structure to studies conducted in areas with large wildfires, such as the western U.S.

## 5. Conclusions

This paper evaluated the performance of MODIS fire count data as a predictor to estimate surface  $PM_{2.5}$  concentrations using a two-stage model in the southeastern U.S. in 2007. Our findings showed that fire counts were an effective predictor of  $PM_{2.5}$  concentrations, especially in southern Georgia and the spring season. Our analysis further exhibited that the

predictive power of fire counts generally increased with the increasing number of surrounding fire spots. The differences between PM<sub>2.5</sub> estimates generated from the fire and nonfire models corresponded well with the spatial distribution of fire density. This study provides insight into the utility of enhancing predictions of surface PM<sub>2.5</sub> levels by incorporating fire count data, and future research should utilize more comprehensive fire size and intensity data.

## Acknowledgments

This work was partially supported by NASA Applied Sciences Program (grants NNX09AT52G and NNX11AI53G, PI: Liu). In addition, this publication was made possible by USEPA grant R834799. Its contents are solely the responsibility of the grantee and do not necessarily represent the official views of the USEPA. Further, USEPA does not endorse the purchase of any commercial products or services mentioned in the publication. PM<sub>2.5</sub>, meteorological, land use, and fire data used in this paper are available free through the links provided in section 2 of the paper. Inquiries regarding the MAIAC data should be directed to Alexei Lyapustin (alexei.i.lyapustin@nasa.gov) at the NASA Goddard Space Flight Center.

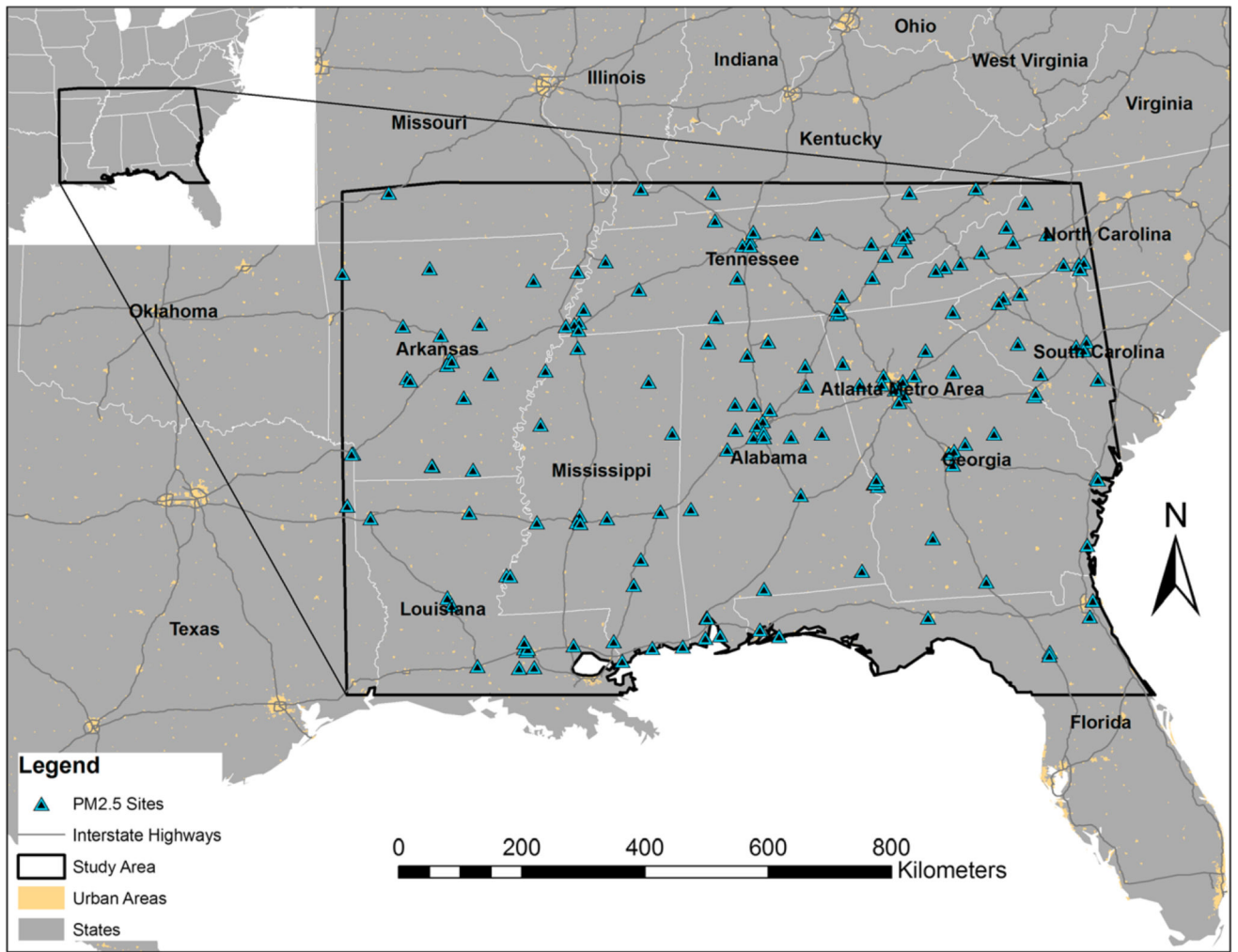
## References

- Al-Saadi J, et al. Intercomparison of near-real-time biomass burning emissions estimates constrained by satellite fire data. *J. Appl. Remote Sens.* 2008; doi: 10.1117/1.2948785
- Cardoso MF, Hurr G, Moore Iii B, Nobre CA, Bain H. Field work and statistical analyses for enhanced interpretation of satellite fire data. *Remote Sens. Environ.* 2005; 96(2):212–227.
- Cosgrove BA, et al. Real-time and retrospective forcing in the North American Land Data Assimilation System (NLDA) project. *J. Geophys. Res.* 2003; 108(D22):8842. doi: 10.1029/2002JD003118
- Dominici F, Peng RD, Bell ML, Pham L, McDermott A, Zeger SL, Samet JM. Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases. *JAMA-J. Am. Med. Assoc.* 2006; 295(10):1127–1134.
- Environmental Protection Agency. Ambient Air Qual. Monit. Program, vol. II, U.S. Environmental Protection Agency, Office of Air Quality Planning and Standards, Air Quality Assessment Division, RTP, N.C., 27711, II(Ambient Air Quality Monitoring Program). 2008. Quality Assurance Handbook for Air Pollution Measurement Systems.
- Hanson S, Padilla M, Corti D, Chuvieco E. Strengths and weaknesses of MODIS hotspots to characterize global fire occurrence. *Remote Sens. Environ.* 2013; 131(0):152–159.
- Hardy, CC., Hermann, SM., Mutch, RE. The wildland fire imperative. In: Hardy, CC., et al., editors. *Smoke Management Guide for Prescribed and Wildland Fire: 2001 Edition*. Natl. Interagency Fire Cent.; Boise, Idaho: 2001.
- Hawbaker TJ, Radeloff VC, Syphard AD, Zhu Z, Stewart SI. Detection rates of the MODIS active fire product in the United States. *Remote Sens. Environ.* 2008; 112(5):2656–2664.
- Hu X, Waller LA, Al-Hamdan MZ, Crosson WL, Estes MG Jr, Estes SM, Quattrochi DA, Sarnat JA, Liu Y. Estimating ground-level PM<sub>2.5</sub> concentrations in the southeastern U.S. using geographically weighted regression. *Environ. Res.* 2013; 121(0):1–10. [PubMed: 23219612]
- Hu X, Waller LA, Lyapustin A, Wang Y, Liu Y. 10-year spatial and temporal trends of PM<sub>2.5</sub> concentrations in the southeastern US estimated using high-resolution satellite data. *Atmos. Chem. Phys.* 2014a; 14(12):6301–6314.
- Hu X, et al. Estimating ground-level PM<sub>2.5</sub> concentrations in the southeastern United States using MAIAC AOD retrievals and a two-stage model. *Remote Sens. Environ.* 2014b; 140(0):220–232.
- Justice CO, Giglio L, Korontzi S, Owens J, Morissette JT, Roy D, Descloitres J, Alleaume S, Petitcolin F, Kaufman Y. The MODIS fire products. *Remote Sens. Environ.* 2002; 83(1–2):244–262.
- Kloog I, Koutrakis P, Coull BA, Lee HJ, Schwartz J. Assessing temporally and spatially resolved PM<sub>2.5</sub> exposures for epidemiological studies using satellite aerosol optical depth measurements. *Atmos. Environ.* 2011; 45(35):6267–6275.

- Kohavi, R. Proceedings of the 14th International Joint Conference on Artificial Intelligence. Vol. 2. Morgan Kaufmann; San Francisco, Calif: 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection; p. 1137-1143.
- Lee HJ, Liu Y, Coull BA, Schwartz J, Koutrakis P. A novel calibration approach of MODIS AOD data to predict PM<sub>2.5</sub> concentrations. *Atmos. Chem. Phys.* 2011; 11(15):7991–8002.
- Liu Y, Sarnat JA, Kilaru A, Jacob DJ, Koutrakis P. Estimating ground-level PM<sub>2.5</sub> in the eastern United States using satellite remote sensing. *Environ. Sci. Technol.* 2005; 39(9):3269–3278. [PubMed: 15926578]
- Liu Y, Franklin M, Kahn R, Koutrakis P. Using aerosol optical thickness to predict ground-level PM<sub>2.5</sub> concentrations in the St. Louis area: A comparison between MISR and MODIS. *Remote Sens. Environ.* 2007; 107(1–2):33–44.
- Liu Y, Paciorek CJ, Koutrakis P. Estimating regional spatial and temporal variability of PM<sub>2.5</sub> concentrations using satellite data, meteorology, and land use information. *Environ. Health Perspect.* 2009; 117(6):886–892. [PubMed: 19590678]
- Lyapustin A, Wang Y, Frey R. An automatic cloud mask algorithm based on time series of MODIS measurements. *J. Geophys. Res.* 2008; 113:D16207.doi: 10.1029/2007JD009641
- Lyapustin A, Martonchik J, Wang YJ, Laszlo I, Korkin S. Multiangle implementation of atmospheric correction (MAIAC): 1. Radiative transfer basis and look-up tables. *J. Geophys. Res.* 2011a; 116:D03210.doi: 10.1029/2010JD014985
- Lyapustin A, Wang Y, Laszlo I, Kahn R, Korkin S, Remer L, Levy R, Reid JS. Multiangle implementation of atmospheric correction (MAIAC): 2. Aerosol algorithm. *J. Geophys. Res.* 2011b; 116:D03211.doi: 10.1029/2010JD014986
- Lyapustin A, Wang Y, Laszlo I, Hilker T, Hall FG, Sellers PJ, Tucker CJ, Korkin SV. Multi-angle implementation of atmospheric correction for MODIS (MAIAC): 3. Atmospheric correction. *Remote Sens. Environ.* 2012; 127(0):385–393.
- Paciorek CJ, Liu Y, Moreno-Macias H, Kondragunta S. Spatiotemporal associations between GOES aerosol optical depth retrievals and ground-level PM<sub>2.5</sub>. *Environ. Sci. Technol.* 2008; 42(15): 5800–5806. [PubMed: 18754512]
- Puttaswamy SJ, Nguyen HM, Braverman A, Hu X, Liu Y. Statistical data fusion of multi-sensor AOD over the Continental United States. *Geocarto Int.* 2013; doi: 10.1080/10106049.2013.827750
- Tian D, Hu Y, Wang Y, Boylan JW, Zheng M, Russell AG. Assessment of biomass burning emissions and their impacts on urban and regional PM<sub>2.5</sub>: A Georgia case study. *Environ. Sci. Technol.* 2009; 43(2):299–305. [PubMed: 19238955]
- van Donkelaar A, Martin RV, Brauer M, Kahn R, Levy R, Verduzco C, Villeneuve PJ. Global estimates of ambient fine particulate matter concentrations from satellite-based aerosol optical depth: Development and application. *Environ. Health Perspect.* 2010; 118(6):847–855. [PubMed: 20519161]
- Wang J, Christopher SA. Intercomparison between satellite-derived aerosol optical thickness and PM<sub>2.5</sub> mass: Implications for air quality studies. *Geophys. Res. Lett.* 2003; 30(21):2095.doi: 10.1029/2003GL018174
- Zeng T, Wang Y, Yoshida Y, Tian D, Russell AG, Barnard WR. Impacts of prescribed fires on air quality over the southeastern United States in spring based on modeling and ground/satellite measurements. *Environ. Sci. Technol.* 2008; 42(22):8401–8406. [PubMed: 19068824]
- Zhang X, Hecobian A, Zheng M, Frank NH, Weber RJ. Biomass burning impact on PM<sub>2.5</sub> over the southeastern US during 2007: Integrating chemically speciated FRM filter measurements, MODIS fire counts and PMF analysis. *Atmos. Chem. Phys.* 2010; 10(14):6839–6853.
- Zhang Y, et al. Aerosol daytime variations over North and South America derived from multiyear AERONET measurements. *J. Geophys. Res.* 2012; 117:D05211.doi: 10.1029/2011JD017242

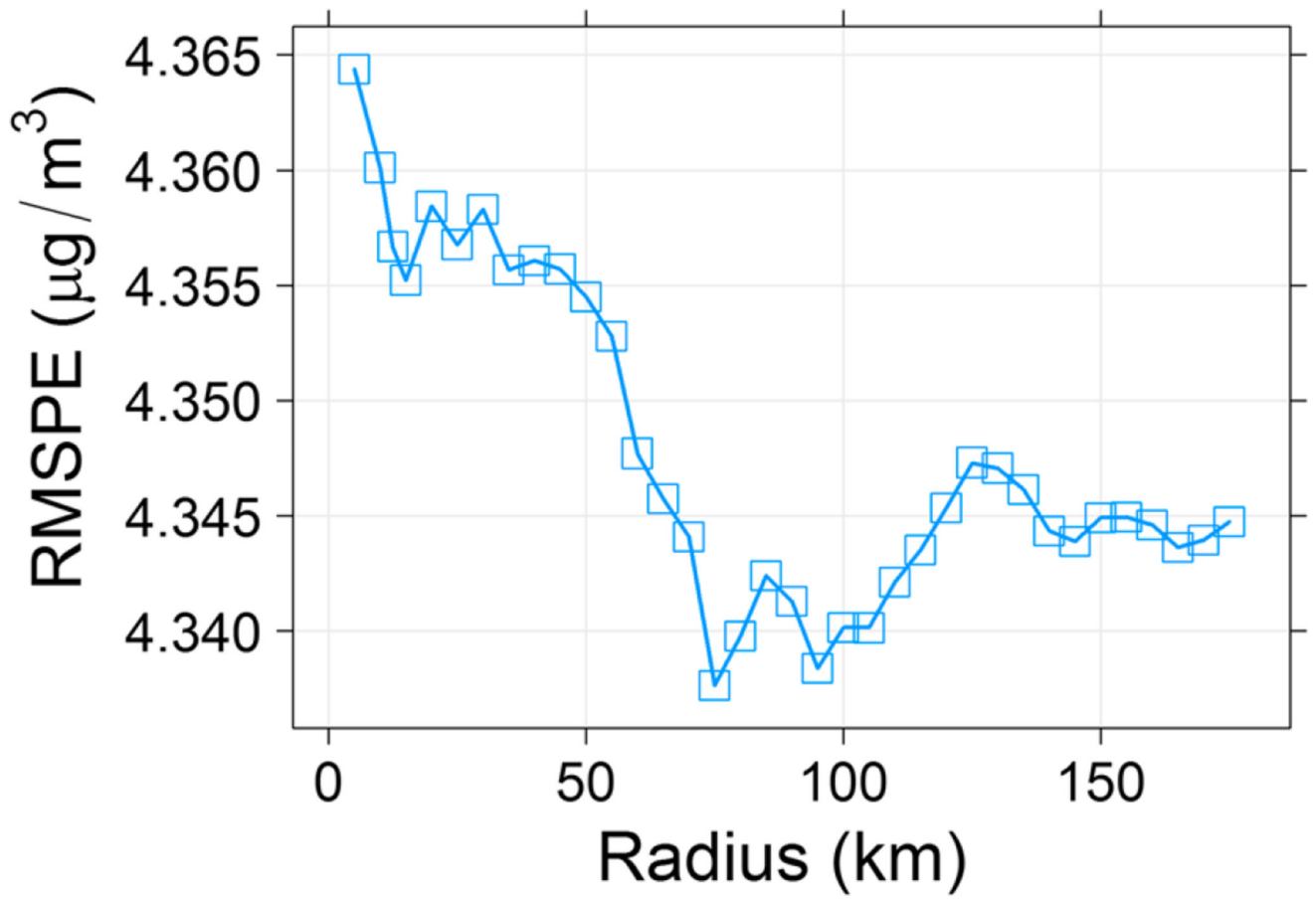
**Key Points**

- MODIS fire count data were utilized to enhance predictions of PM<sub>2.5</sub> levels
- Fire count data provide a measurable improvement in PM<sub>2.5</sub> prediction
- Fire counts have better performance in areas and seasons prone to fire events

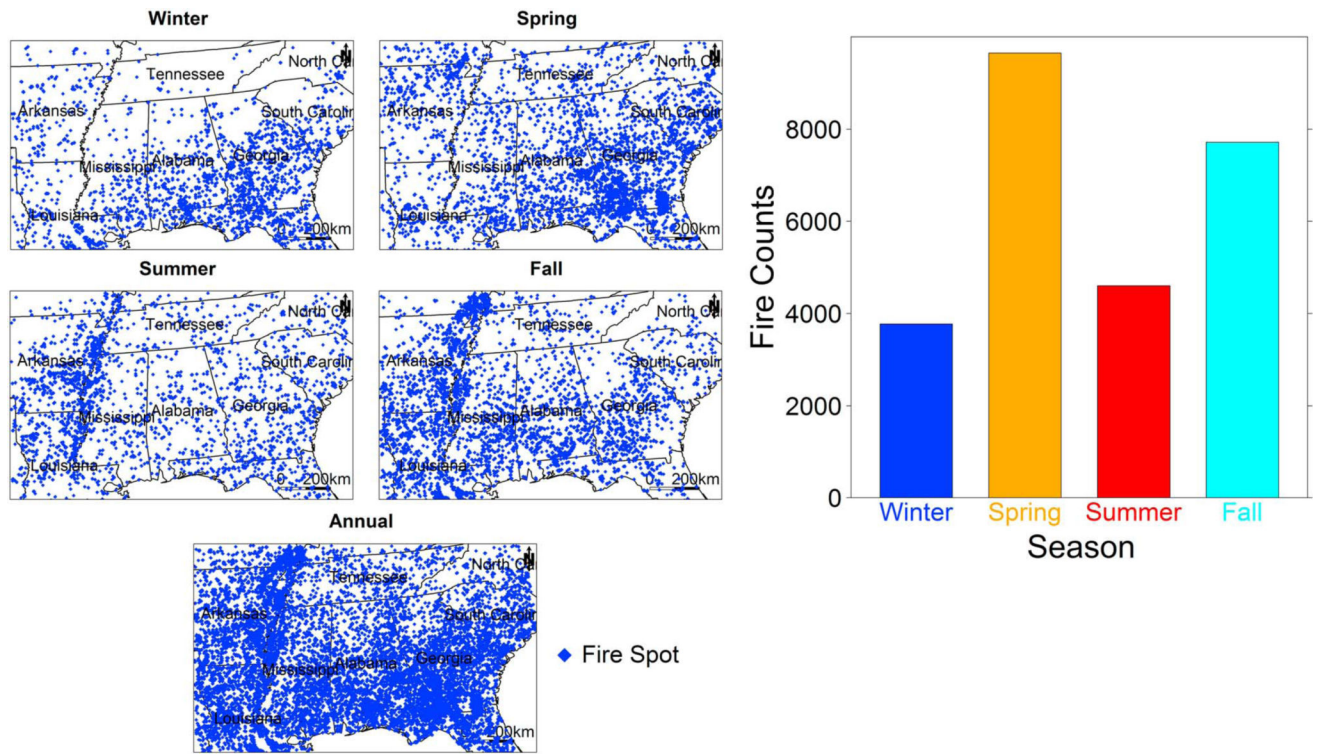


**Figure 1.** Study area indicating location of PM<sub>2.5</sub> monitoring stations for year 2007.

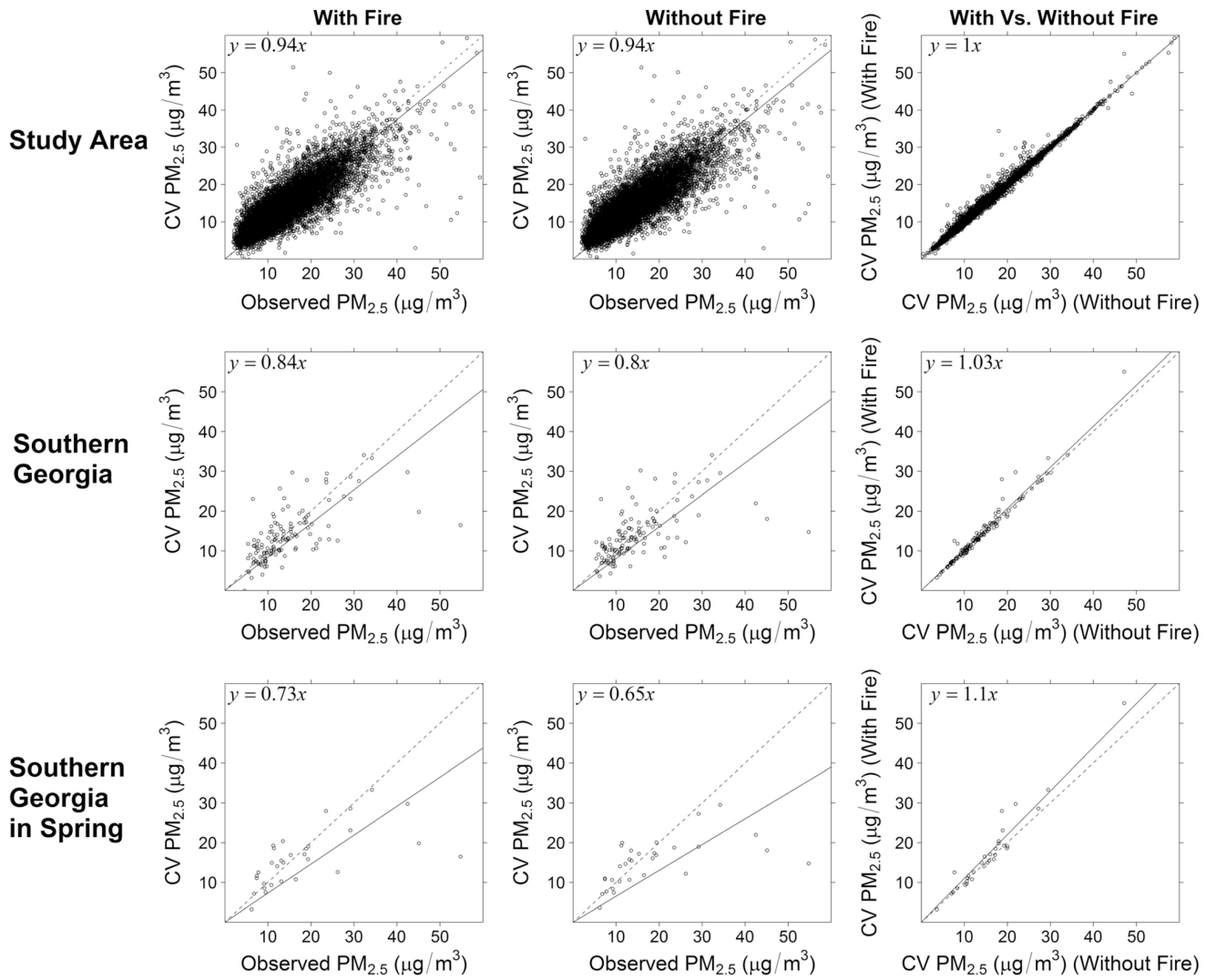




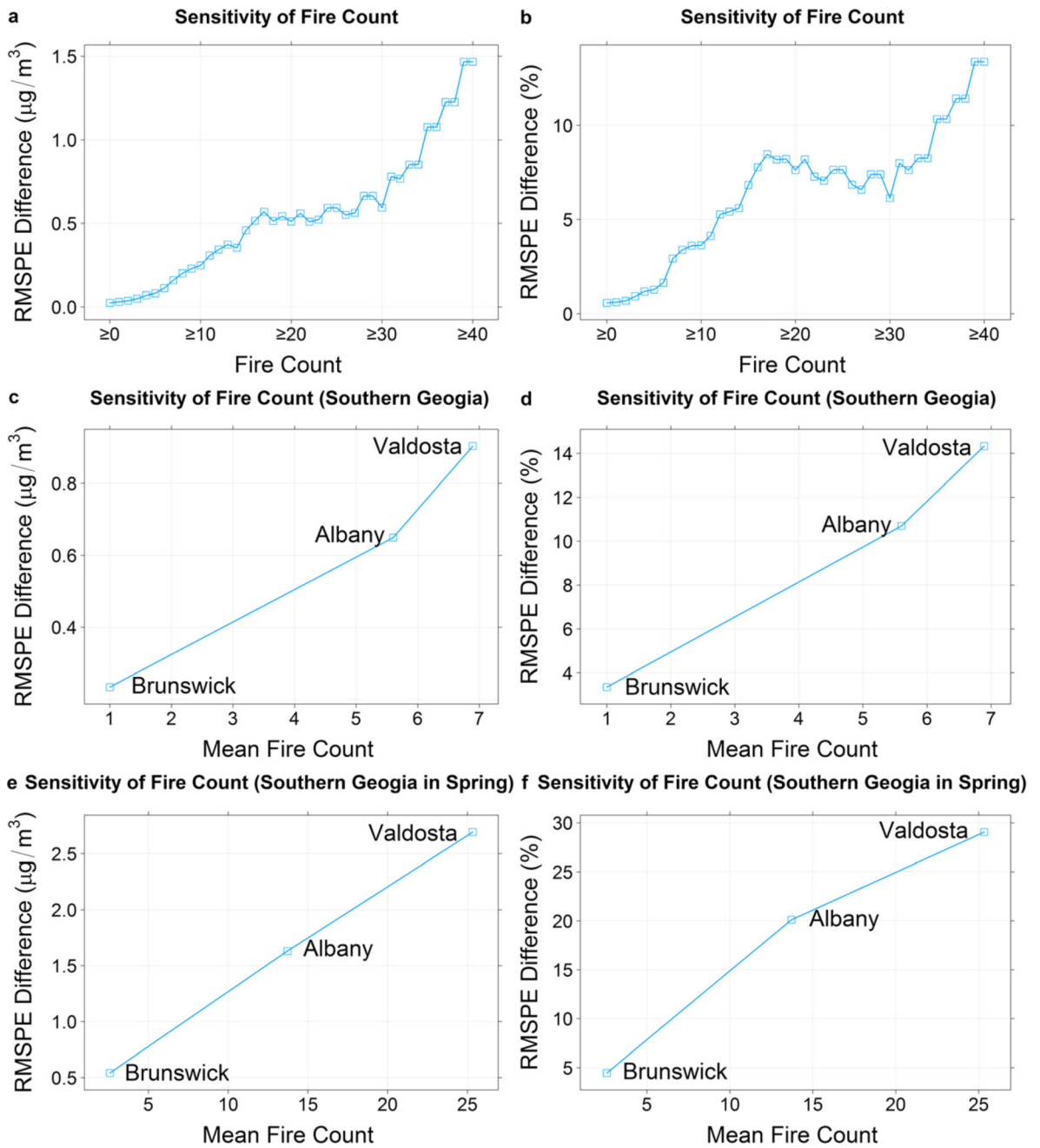
**Figure 2.**  
The impact of buffer radius on root-mean-square prediction errors (RMSPEs).



**Figure 3.** The seasonal distribution of fire counts within the study area in 2007.

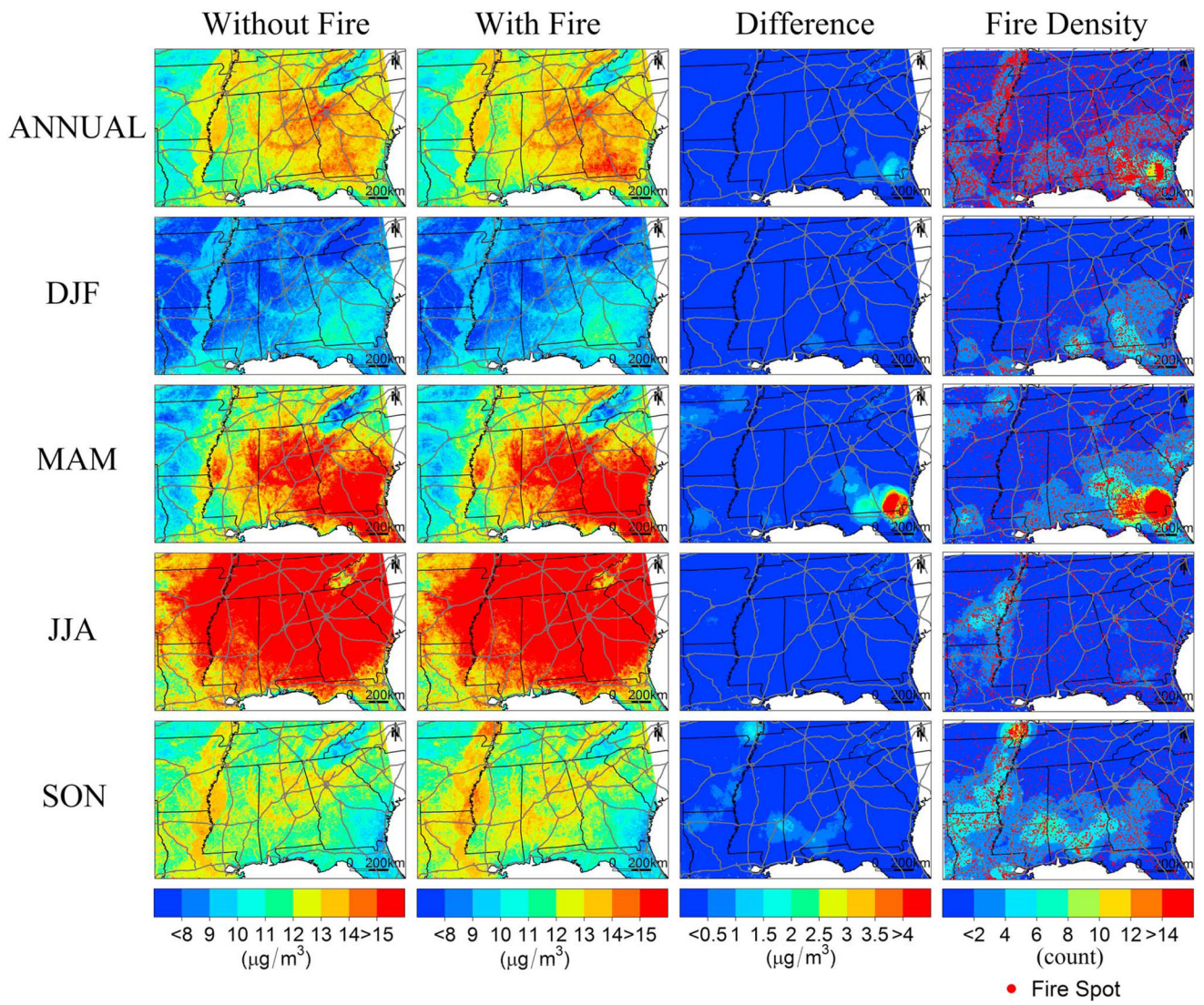


**Figure 4.** Model validation and comparison between observed and predicted  $PM_{2.5}$  values.



**Figure 5.** The impact of fire counts on model performance as measured by RMSPE difference (RMSPE difference = RMSPE of the nonfire model-RMSPE of the fire model).





**Figure 6.** Annual and seasonal mean PM<sub>2.5</sub> concentration predictions, differences between PM<sub>2.5</sub> concentrations predicted from the fire and nonfire models, and the density of fire counts in the study area in 2007.

**Table 1**Descriptive Statistics Summarizing Observations for PM<sub>2.5</sub> Monitoring Sites for Year 2007

Variable	Mean	SD	Min	Max
PM <sub>2.5</sub> (µg/m <sup>3</sup> )	13.67	7.70	2.00	145.00
Relative humidity (%) <sup>a</sup>	49.81	11.75	18.74	86.83
Wind speed (m/s) <sup>a</sup>	3.85	2.02	0.05	15.11
Forest cover 2006 <sup>b</sup>	0.11	0.14	0.00	0.79
Road length (m) <sup>b</sup>	2072.00	375.69	0.00	137.00
Elevation (m) <sup>b</sup>	165.97	150.89	-1.67	981.25
Point emissions 2008 (t/yr) <sup>b</sup>	2.75	15.81	0.00	138.21
AOD <sup>a</sup>	0.28	0.25	0.00	1.95
Fire count <sup>c</sup>	1.41	3.36	0.00	91.00

<sup>a</sup>Nearest neighbor.<sup>b</sup>Within a 1 × 1 km<sup>2</sup> square buffer.<sup>c</sup>Within a buffer with a radius of 75 km.



Model Comparisons Across Season and Extent of the Study Area for Observations From 2007

**Table 2**

	Study Area			Southern Georgia (Albany, Brunswick, and Valdosta)		
	Mean Fire Count	CV RMSPE Improvement ( $\mu\text{g}/\text{m}^3$ )	CV RMSPE Improvement (%)	Mean Fire Count	CV RMSPE Improvement ( $\mu\text{g}/\text{m}^3$ )	CV RMSPE Improvement (%)
Spring	1.77	0.06	1.07	12.74	1.72	25.60
Nonspring	1.28	0.01	0.25	2.56	0.08	1.64
All	1.41	0.02	0.58	5.26	0.65	11.8