Environment International 121 (2018) 550–560

Contents lists available at ScienceDirect

Environment International

journal homepage: www.elsevier.com/locate/envint

The sensitivity of satellite-based $PM_{2.5}$ estimates to its inputs: Implications to model development in data-poor regions

Guannan Geng[a], Nancy L. Murray[b], Howard H. Chang[b], Yang Liu[a,*]

[a] Department of Environmental Health, Rollins School of Public Health, Emory University, Atlanta, GA 30322, USA
[b] Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, GA 30322, USA

ARTICLE INFO

Handling Editor: Xavier Querol

Keywords:
Bayesian ensemble model
Fine particulate matter
Exposure assessment
Statistical model
Satellite remote sensing

ABSTRACT

Exposure to fine particulate matter ($PM_{2.5}$) has been associated with a wide range of negative health outcomes. The overwhelming majority of the epidemiological studies that helped establish such associations was conducted in regions with sufficient ground observations and other supporting data, i.e., the data-rich regions. However, air pollution health effects research in the data-poor regions, where pollution levels are often the highest, is still very limited due to the lack of high-quality exposure estimates. To improve our understanding of the desired input datasets for the application of satellite-based $PM_{2.5}$ exposure models in data-poor areas, we applied a Bayesian ensemble model in the southeast U.S. that was selected as a representative data-rich region. We designed four groups of sensitivity tests to simulate various data-poor scenarios. The factors considered that would influence the model performance included the temporal sampling frequency of the monitors, the number of ground monitors, the accuracy of the chemical transport model simulation of $PM_{2.5}$ concentrations, and different combinations of the additional predictors. While our full model achieved a 10-fold cross-validated (CV) R^2 of 0.82, we found that when reducing the sampling frequency from the current 1-in-3 day to 1-in-9 day, the CV R^2 decreased to 0.58, and the predictions could not capture the daily variations of $PM_{2.5}$. Half of the current stations (i.e., 30 monitors) could still support a robust model with a CV R^2 of 0.79. With 20 monitors, the CV R^2 decreased from 0.71 to 0.55 when 100% additional random errors were added to the original CMAQ simulations. However, with a sufficient number of ground monitors (e.g., 30 monitors), our Bayesian ensemble model had the ability to tolerate CMAQ errors with only a slight decrease in CV R^2 (from 0.79 to 0.75). With fewer than 15 monitors, our full model collapsed and failed to fit any covariates, while the models with only time-varying variables could still converge even with only five monitors left. A model without the land use parameters lacked fine spatial details in the prediction maps, but could still capture the daily variability of $PM_{2.5}$ (CV $R^2 \geq 0.67$) and might support a study of the acute health effects of $PM_{2.5}$ exposure.

1. Introduction

Exposure to fine particulate matter with aerodynamic diameters of 2.5 μm or less ($PM_{2.5}$) has been associated with various adverse health outcomes, including cardiovascular and respiratory diseases, lung cancer, and premature death in numerous epidemiologic studies worldwide (Pope and Dockery, 2006; Brook et al., 2010; Turner et al., 2011; Raaschou-Nielsen et al., 2013). However, the overwhelming majority of these studies were conducted in the developed countries with sufficient ground $PM_{2.5}$ monitors and high-quality supporting information to provide exposure estimates, i.e., the data-rich regions. According to the Global Burden of Disease study, nearly 87% of the world's population lived in areas exceeding the World Health Organization Air Quality Guideline of 10 $\mu g/m^3$ $PM_{2.5}$ at the annual level, and high $PM_{2.5}$ levels were commonly found in developing countries (Brauer et al., 2016). Assessing the disease burden of $PM_{2.5}$ is still difficult in these polluted regions in the world (Tonne, 2017), mainly due to the lack of high-resolution $PM_{2.5}$ exposure estimates (i.e., the data-poor regions).

In the past decade, satellite-based aerosol optical depth (AOD) has become a valuable information source to extend the spatial and temporal coverage of $PM_{2.5}$ ground monitoring networks (Wang and Christopher, 2003; Engel-Cox et al., 2004; Liu et al., 2004; Van Donkelaar et al., 2010; Geng et al., 2015; Lee et al., 2016). Various empirical statistical models have been proposed to estimate daily to annual mean $PM_{2.5}$ concentrations at various spatial resolutions (Hu et al., 2014a; Kloog et al., 2014; Di et al., 2016; Zheng et al., 2016; Hu et al., 2017). Because of the complex relationship between satellite-

* Corresponding author.
E-mail address: yang.liu@emory.edu (Y. Liu).

https://doi.org/10.1016/j.envint.2018.09.051
Received 27 July 2018; Received in revised form 26 September 2018; Accepted 26 September 2018

retrieved AOD and ground $PM_{2.5}$ measurements (Hoff and Christopher, 2009), additional factors such as meteorological fields, land use variables and other satellite data were incorporated in the statistical models to better resolve the AOD-$PM_{2.5}$ relationship (Kloog et al., 2012; Hu et al., 2014b; Just et al., 2015). For example, in a national-scale geographically weighted regression model over China (Ma et al., 2014), the prediction accuracy was improved when including meteorological and land use data in the model (cross-validated R^2 from 0.52 to 0.64). Recently, statistical models also attempted to integrate chemical transport model (CTM)-simulated $PM_{2.5}$ concentrations to fill the data gaps left by missing satellite AOD, despite their higher absolute prediction errors (Friberg et al., 2016; Xiao et al., 2017; Geng et al., 2018). However, CTM-simulated $PM_{2.5}$ can have different uncertainty levels in different regions. This is because the limited activity and emission factor data and empirical choices of spatial proxies could contribute more biases in the estimation of the gridded emissions in developing countries (Zhang et al., 2009; Geng et al., 2017). How the errors in the CTM $PM_{2.5}$ affect the performance of the statistical models is an important yet poorly studied issue.

Since the development of statistical models depends on ground $PM_{2.5}$ observations, many studies (Liu et al., 2009; Chang et al., 2014; Hu et al., 2014a) have estimated $PM_{2.5}$ exposure data in the U.S. where ~1600 monitoring stations routinely make $PM_{2.5}$ measurements daily, 1-in-3 day, or 1-in-6 day since 2000. There is an increasing body of literatures in China (Ma et al., 2016; Zheng et al., 2016) on the estimation of satellite-based $PM_{2.5}$ concentrations after 2013, when China established its national $PM_{2.5}$ monitoring network. There are also studies in southern Ontario (Tian and Chen, 2010) and the Mexico City (Just et al., 2015), etc. These studies have shown good performance in their statistical models and provided useful datasets for the following epidemiological studies (Liu et al., 2016; Di et al., 2017). Despite the rapid development of this emerging research field and strong desire to apply these new exposure modeling techniques in data-poor regions, developing a high-performance $PM_{2.5}$ statistical model for air quality management and health impact assessment requires a thorough understanding about the impacts of the availability of ground $PM_{2.5}$ observations and other supporting information, which may not be readily available in a data-poor region.

To date, an important question remains to be answered: does a set of minimum data requirements exist for developing a high-performance $PM_{2.5}$ model in an area with limited data and resources? For example, how important is $PM_{2.5}$ ground sampling frequency to the performance of a satellite-driven $PM_{2.5}$ model? Is the inclusion of meteorological or land use parameter essential to model prediction power? How valuable are CTM simulations to the predicted $PM_{2.5}$ concentration surface? In this study, we addressed this issue by examining the sensitivity of a flexible satellite-based statistical model to its key input variables including the spatial and temporal availability of ground $PM_{2.5}$ observations, meteorological and land use parameters, and the quality of CTM simulations. In addition, we considered the complexity of model structure as well as the joint impacts of these factors. We used the southeastern U.S. as a representative data-rich region and designed multiple sets of sensitivity tests to simulate various data-poor scenarios. It is our hope that the process demonstrated in this study could provide a framework for evaluating the feasibility of building a high-performance $PM_{2.5}$ statistical model in a data-poor region.

2. Materials and methods

2.1. Datasets

The study domain is approximately 600 × 550 km^2 in the southeastern U.S., which covers part of Tennessee, North Carolina, South Carolina, Alabama and Georgia (Fig. 1). This area has a changing terrain from the Appalachian Mountains in the northeast to the Piedmont in the middle, then the coastal plain in the south. The sizes of cities range from the Atlanta Metropolitan area with over 5.5 million people, to medium to small size cities and rural towns. More importantly, this region has a relatively dense ground air quality monitoring network with 60 stations, high-quality meteorological and land use data, and well-developed CTM simulations, which could be used to simulate various levels of data accessibility and quality in a wide range of data-poor regions in the world.

Daily mean ground-level $PM_{2.5}$ measurements for 2003–2005 using federal reference method were obtained from the U.S. Environmental Protection Agency's Air Quality System (https://www.epa.gov/outdoor-air-quality-data/). The numbers of observations per year for each monitor are shown in Fig. 1. These monitors had three different sampling schemes: daily, 1-in-3 day, and 1-in-6 day. On average, there were 115 observation days per year for each monitor, which was a typical 1-in-3 day sampling schedule.

We utilized the satellite-based AOD data retrieved by the Multi-angle Implementation of Atmospheric Correction (MAIAC) algorithm at 1 km spatial resolution (Lyapustin et al., 2011a; Lyapustin et al., 2011b) based on the Moderate Resolution Imaging Spectroradiometer (MODIS). MAIAC AOD from both Terra (overpass time at 10:30 am) and Aqua (overpass time at 1:30 pm) satellite were merged to improve the spatial coverage of AOD data.

$PM_{2.5}$ simulations from the USEPA Models-3/Community Multiscale Air Quality (CMAQ) model version 4.5 at a 12 km spatial resolution (Byun and Schere, 2006) were also used in this study. Other variables compiled in this study included: elevation at 30 m spatial resolution from the National Elevation Data set (NED, http://ned.usgs.gov), forest cover at 30 m spatial resolution from the 2001 National Land Cover Database (NLCD, http://www.mrlc.gov), road lengths of limited access highway extracted from ESRI StreetMap USA (Environmental Systems Research Institute, Inc., Redland, CA), relative humidity (RH) and wind speed data at ~13 km spatial resolution from the North American Land Data Assimilation Systems, and the primary $PM_{2.5}$ emissions from point sources provided by the 2002 USEPA National Emissions Inventory.

All data were integrated into the 1 km MAIAC grid. CMAQ $PM_{2.5}$ data and meteorological fields were matched to the centroid of each grid using the nearest neighbor approach. Elevation and forest cover data were averaged and road lengths were summed within the 1 km MAIAC grid. Overall, we had 8722 records of paired data for 2003–2005 over the study domain.

2.2. Bayesian ensemble approach

In this work, we utilized a two-stage Bayesian ensemble approach to estimate daily full-coverage $PM_{2.5}$ concentrations and conducted sensitivity tests to study the impact of the input dataset on the model's performance. The Bayesian ensemble model is a modeling framework that takes advantages of the satellite remote sensing, the ground measurements and the CTM simulations. It had better performance than the commonly used multi-stage statistical models (Murray et al., 2018), especially in the mountainous regions (Geng et al., 2018). The details of the Bayesian ensemble model are provided in Murray et al. (2018) and a brief summary is presented below.

In the first stage, two statistical downscalers were involved. One was to calibrate the spatially and temporally varying $PM_{2.5}$-AOD relationship, which was the AOD downscaler. The other was to calibrate the CMAQ simulated $PM_{2.5}$, which was the CMAQ downscaler. The downscaler models could be written as below (Chang et al., 2014):

$$Y_{st} = \alpha_{st} + \beta_{st} X_{st} + \gamma Z_{st} + \varepsilon_{st} \quad (1)$$

where Y_{st} was the $PM_{2.5}$ measurement at monitor s on day t. X_{st} represented the main predictor of the downscaler at monitor s on day t, which were MAIAC AOD and CMAQ $PM_{2.5}$ for the AOD and CMAQ downscaler respectively. Z_{st} were additional predictors including meteorological fields and land use variables. In our AOD downscaler, the Z vector included RH, wind speed, elevation, forest cover, limited