Environmental Pollution 233 (2018) 464–473

Contents lists available at ScienceDirect

Environmental Pollution

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

Spatiotemporal prediction of daily ambient ozone levels across China using random forest for human exposure assessment

Yu Zhan [a], Yuzhou Luo [b], Xunfei Deng [c], Michael L. Grieneisen [b], Minghua Zhang [b], Baofeng Di [a, *]

[a] Department of Environmental Science and Engineering, Sichuan University, Chengdu, Sichuan 610065, China
[b] Department of Land, Air, and Water Resources, University of California, Davis, CA 95616, USA
[c] Institute of Digital Agriculture, Zhejiang Academy of Agricultural Sciences, Hangzhou, Zhejiang 310021, China

ARTICLE INFO

Article history:
Received 10 July 2017
Received in revised form
11 September 2017
Accepted 8 October 2017

Keywords:
Ozone pollution
Spatiotemporal distributions
China
Human exposure
Machine learning

ABSTRACT

In China, ozone pollution shows an increasing trend and becomes the primary air pollutant in warm seasons. Leveraging the air quality monitoring network, a random forest model is developed to predict the daily maximum 8-h average ozone concentrations ($[O_3]_{MDA8}$) across China in 2015 for human exposure assessment. This model captures the observed spatiotemporal variations of $[O_3]_{MDA8}$ by using the data of meteorology, elevation, and recent-year emission inventories (cross-validation $R^2 = 0.69$ and RMSE = 26 μg/m^3). Compared with chemical transport models that require a plenty of variables and expensive computation, the random forest model shows comparable or higher predictive performance based on only a handful of readily-available variables at much lower computational cost. The nationwide population-weighted $[O_3]_{MDA8}$ is predicted to be 84 ± 23 μg/m^3 annually, with the highest seasonal mean in the summer (103 ± 8 μg/m^3). The summer $[O_3]_{MDA8}$ is predicted to be the highest in North China (125 ± 17 μg/m^3). Approximately 58% of the population lives in areas with more than 100 nonattainment days ($[O_3]_{MDA8}$>100 μg/m^3), and 12% of the population are exposed to $[O_3]_{MDA8}$>160 μg/m^3 (WHO Interim Target 1) for more than 30 days. As the most populous zones in China, the Beijing-Tianjin Metro, Yangtze River Delta, Pearl River Delta, and Sichuan Basin are predicted to be at 154, 141, 124, and 98 nonattainment days, respectively. Effective controls of O_3 pollution are urgently needed for the highly-populated zones, especially the Beijing-Tianjin Metro with seasonal $[O_3]_{MDA8}$ of 140 ± 29 μg/m^3 in summer. To the best of the authors' knowledge, this study is the first statistical modeling work of ambient O_3 for China at the national level. This timely and extensively validated $[O_3]_{MDA8}$ dataset is valuable for refining epidemiological analyses on O_3 pollution in China.

1. Introduction

Ambient ozone (O_3) is a worldwide air pollutant harmful to human health (Brunekreef and Holgate, 2002; WHO, 2006). Ambient O_3 is mainly formed through photochemical reactions between volatile organic compounds (VOCs) and nitrogen oxides (NO_x) in the presence of sunlight (USEPA, 2013). Exposure to ambient O_3 has been associated with human health problems such as respiratory and cardiovascular diseases (Kan et al., 2008; Shang et al., 2013; USEPA, 2013; WHO, 2006; Wong et al., 2008). To protect public health, the World Health Organization (WHO) proposes the O_3 air quality guideline of a daily maximum 8-h mean ($[O_3]_{MDA8}$) of 100 μg/m^3, which is also implemented in China (MEPC, 2012; WHO, 2006). The O_3 pollution level in China is predicted to be comparable to the global average (Brauer et al., 2016). In contrast to the decreasing trend of "visible" fine particulate matter ($PM_{2.5}$) pollution in China, "invisible" O_3 pollution shows an increasing trend and becomes the primary air pollutant in warm seasons (Anger et al., 2015; Brauer et al., 2016; Zhao et al., 2016). While site-based ambient O_3 concentrations have been regularly monitored since 2013 via the national air quality monitoring network for China (MEPC, 2015), nationwide spatiotemporal distributions of ambient O_3 levels are required for human exposure assessment.

Chemical transport models (CTMs), such as the Community Multiscale Air Quality model (CMAQ) (Byun and Schere, 2006), have been employed to predict the spatiotemporal distributions of

* Corresponding author.
E-mail address: dibaofeng@scu.edu.cn (B. Di).

https://doi.org/10.1016/j.envpol.2017.10.029
0269-7491/

ambient O_3 worldwide (Bell, 2006; Brauer et al., 2016; Hu et al., 2016; Wang et al., 2011; Zhang et al., 2011). By utilizing meteorological conditions predicted by climate models, CTMs simulate environmental processes of O_3 and its precursors, providing guidance on source control (Jacob, 1999). CTMs have been extensively validated against O_3 monitoring data in the United States, indicating that CTMs are capable of capturing monthly or seasonal O_3 concentrations at large spatial scales (USEPA, 2013; Zhang et al., 2011). However, fine-scale predictions by CTMs may deviate widely from observations due to imperfect input data and limited knowledge (USEPA, 2013). In addition, CTMs generally require high computing resources and a wealth of input data (Beelen et al., 2009). The predictive performance of CTMs for China tends to degrade due to the high uncertainty associated with the emission inventories, especially on a fine spatiotemporal scale (Streets et al., 2003; Zhang et al., 2008). While many efforts have been made to refine and update the emission inventories for China (Li et al., 2017; Wu and Xie, 2017), this process requires much extra labor and time. The publicly available emission inventories for China are only released for certain years, e.g., 2008 and 2010 (Li et al., 2017), which may also be used in CTMs simulations for more recent years. The predictive performance of such simulations for more recent years tend to be lower, and post hoc adjustment to the emission inventories are generally required. Moreover, long-term and regional/national monitoring data is critical for validating the CTM predictions (USEPA, 2009). However, validations of CTMs for China are quite limited due to the scarcity of ambient O_3 monitoring data before 2013, and the exposure assessment results based on these CTMs predictions may be encumbered (Brauer et al., 2016; Guo et al., 2016; Lou et al., 2015; Madaniyazi et al., 2016). In order to provide more recent and reliable O_3 distributions for human exposure assessment, therefore, statistical models based on the extensive O_3 monitoring network for China are more favored.

Statistical models such as geostatistical models and land use regressions (LUR) are commonly used to predict spatiotemporal distributions of ambient O_3 concentrations (Adam-Poupart et al., 2014; Beelen et al., 2009; Brauer et al., 2008; Wang et al., 2015). On the basis of O_3 observations, these statistical models are parameterized with the spatiotemporal autocorrelations of O_3 levels and/or their relationships with the predictor variables such as land use types and meteorological conditions. The statistical models that incorporate CTMs predictions as their predictor variables are also referred to as hybrid models, which tend to improve the predictive performance with increased model complexity (Di et al., 2017; Wang et al., 2016). Although statistical models do not explicitly simulate the environmental processes of O_3, they generally exhibit higher predictive performance than CTMs on fine spatiotemporal scales in the presence of extensive monitoring data (Adam-Poupart et al., 2014; Hoek et al., 2008; Marshall et al., 2008). While statistically modeling ambient O_3 distributions is common in the western countries (Adam-Poupart et al., 2014; Beelen et al., 2009; Brauer et al., 2008; Wang et al., 2015), the number of statistical studies for China is rather limited due to the paucity of monitoring data before 2013. Fortunately, the national air quality monitoring network for China has been established and expanded since 2013 (MEPC, 2015). In 2015, hourly ambient O_3 concentrations were measured at more than 1500 sites located in more than 300 cities. However, such a large dataset is still not sufficient to assess the nationwide human exposure levels, since many other areas still lack monitoring data. Statistical models trained with the monitoring data are needed to predict the O_3 concentrations in the unmonitored areas.

Machine learning algorithms, which are statistical models from the algorithmic modelling culture (Breiman, 2001b), generally show predictive performance that is comparable or superior to traditional statistical models such as general linear regression and kriging (Hastie et al., 2009; Hu et al., 2017; Kanevski, 2008; Li et al., 2011). Random forests, as a popular machine learning algorithm, make statistical predictions by averaging an ensemble of decorrelated classification or regression trees, and they are capable of handling nonlinear relationships and interaction effects (Breiman, 2001a). An algorithm comparison study found that random forests and gradient boosting machine (GBM) exhibit the best performance in predicting $PM_{2.5}$ concentrations among ten machine learning algorithms (Reid et al., 2015). The geographically-weighted GBM shows even higher prediction performance than GBM but is much more computationally expensive (Zhan et al., 2017). Random forests are thus one of the optimal choices when striking a balance between prediction accuracy and computing cost. Moreover, unlike some other machine learning algorithms (e.g., neural network), random forests can show the contribution of each predictor variable by the variable importance measure and the partial dependence plot (Hastie et al., 2009). The method for evaluating the variances of prediction made by random forests is also available (Wager et al., 2014). In a recent study, the random forest approach has shown good performance in predicting the $PM_{2.5}$ concentrations for the conterminous United States, but the prediction uncertainties are not evaluated (Hu et al., 2017). It is important to show the prediction uncertainties and demonstrate the applicability of random forests to different chemical species and geographical regions. Therefore, random forests are proposed to predict the nationwide spatiotemporal distributions of $[O_3]_{MDA8}$ in China for human exposure assessment by utilizing the valuable monitoring dataset.

This study aims to predict the spatiotemporal distributions of daily $[O_3]_{MDA8}$ across China in 2015. A $0.1^\circ \times 0.1^\circ$ grid is used to be consistent with the previous global or national exposure assessments (Brauer et al., 2016; Guo et al., 2016). Random forest is implemented with variable selection to statistically model the spatiotemporal variations of $[O_3]_{MDA8}$ based on the publicly available datasets including the meteorology, elevation, and emission inventories. On the basis of the predicted $[O_3]_{MDA8}$, we assess the exposure intensities and durations for the populations living in different regions of China. To the best of the authors' knowledge, this study is the first statistical modeling work of ambient O_3 for China at the national level. This timely and extensively validated $[O_3]_{MDA8}$ dataset are valuable for refining epidemiological analyses on O_3 pollution in China.

2. Materials and methods

2.1. Data preparation

The hourly O_3 monitoring data for 2015 were obtained from the national air quality monitoring network for mainland China and Hainan Island (MEPC, 2015), the Environmental Protection Department of Hong Kong for Hong Kong (EPDHK, 2015), and the Environmental Protection Administration of Taiwan for Taiwan (EPAROC, 2015). The O_3 concentrations were measured with the ultraviolet-spectrophotometry method. The highest 8-h moving average for each day was calculated as $[O_3]_{MDA8}$ after data cleaning at each monitoring site by following the procedure adopted by the California Air Resources Board (CARB, 2006). For a given day to be included, this procedure requires more than six hourly measurements in each third of the day (i.e., 0:00am-7:00am, 8:00am-15:00pm, and 16:00pm-23:00pm; Beijing Standard Time, UTC+8), and no more than two consecutive hourly measurements missing within that day. Around half a million $[O_3]_{MDA8}$ records were obtained from 1608 monitoring sites across China in 2015 (Fig. 1), which were mainly located in East, Central, North, and South China.