Remote Sensing of Environment 206 (2018) 98–106

Contents lists available at ScienceDirect

Remote Sensing of Environment

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

Predicting the minimum height of forest fire smoke within the atmosphere using machine learning and data from the CALIPSO satellite

Jiayun Yao[a,*], Sean M. Raffuse[b], Michael Brauer[a], Grant J. Williamson[c], David M.J.S. Bowman[c], Fay H. Johnston[d], Sarah B. Henderson[a,e]

[a] School of Population and Public Health, University of British Columbia, Vancouver, Canada
[b] Air Quality Research Center, University of California, Davis, USA
[c] School of Biological Sciences, University of Tasmania, Hobart, Australia
[d] Menzies Institute for Medical Research, University of Tasmania, Hobart, Australia
[e] Environmental Health Services, British Columbia Centre for Disease Control, Vancouver, Canada

ARTICLE INFO

Keywords:
Forest fire smoke
CALIPSO
Vertical profile
Machine learning
Statistical model
Population exposure

ABSTRACT

Forest fire smoke is a growing public health concern as more intense and frequent fires are expected under climate change. Remote sensing is a promising tool for exposure assessment, but its utility for health studies is limited because most products measure pollutants in the total column of the atmosphere, and not the surface concentrations most relevant to population health. Information about the vertical distribution of smoke is vital for addressing this limitation. The CALIPSO satellite can provide such information but it cannot cover all smoke events due to its narrow ground track. In this study, we developed a random forests model to predict the minimum height of the smoke layer observed by CALIPSO at high temporal and spatial resolution, using information about fire activity in the vicinity, geographic location, and meteorological conditions. These pieces of information are typically available in near-real-time, ensuring that the resulting model can be easily operationalized. A total of 15,617 CALIPSO data blocks were identified as impacted by smoke within the province of British Columbia, Canada from 2006 to 2015, and 52.1% had smoke within the boundary layer, where the population might be exposed. The final model explained 82.1% of the variance in the observations with a root mean squared error of 560 m. The most important variables in the model were wind patterns, the month of smoke observation, and fire intensity within 500 km. Predictions from this model can be 1) directly applied to smoke detection from the existing remote sensing products to provide another dimension of information; 2) incorporated into statistical smoke models with inputs from remote sensing products; or 3) used to inform estimates of vertical dispersion in deterministic smoke models. These potential applications are expected to improve the assessment of ground-level population exposure to forest fire smoke.

1. Introduction

Smoke emitted from forest fires is a major contributor to poor air quality in many parts of the world. Fire smoke is a complex mixture of pollutants, among which particulate matter (PM) is the most consistently elevated and commonly studied (Durán et al., 2014; Naeher et al., 2007). Given that climate change favors more intense and frequent fires, smoke emissions are projected to increase over the coming decades (Spracklen et al., 2009). Under these conditions, exposure to forest fire smoke is becoming a pressing public health concern with short- and long-term air quality impacts affecting an increasing number of people (Liu et al., 2016). Forest fire smoke exposure has been consistently associated with a wide range of acute respiratory health endpoints, from increased medication dispensations to hospital admissions and mortality. Evidence for cardiovascular endpoints is also emerging (Liu et al., 2015; Reid et al., 2016a). However, better exposure assessment is needed to improve epidemiologic studies of these effects (Atkinson et al., 2014; Pope III and Dockery, 2006).

It is challenging to assess population exposure to forest fire smoke both retrospectively and in near-real-time because smoke is dynamic and variable in both space and time. Good exposure assessment requires tools with high spatial and temporal resolution, as well as sufficient spatial coverage. Conventional ground-based air quality monitoring can be limited with respect to these spatial considerations, but remote sensing technology has provided promising alternatives. Remotely sensed fire and/or smoke products have been used to directly or

* Corresponding author at: School of Population and Public Health, University of British Columbia, 2206 East Mall, Vancouver V6T 1Z3, Canada.
E-mail address: Jiayun.Yao@bccdc.ca (J. Yao).

https://doi.org/10.1016/j.rse.2017.12.027
Received 16 March 2017; Received in revised form 5 December 2017; Accepted 18 December 2017

indirectly assess smoke exposure in many recent studies (Faustini et al., 2015; Henderson et al., 2011; Kollanus et al., 2017; Reid et al., 2016b; Yao et al., 2016). However, most remote sensing products provide information integrated from the top to the bottom of the atmosphere, rather than information specifically at the surface where populations are exposed. Better information about the vertical distribution of smoke in the atmosphere at adequate temporal resolution would improve the utility of remote sensing products for exposure assessment in epidemiologic and public health surveillance applications.

There are two remote sensing platforms that can provide information about the vertical distribution of atmospheric aerosols. The Multi-angle Imaging Spectro-Radiometer (MISR) aboard the Terra satellite (Kahn et al., 2007) can measure the altitude of the layer of maximum contrast, which can be the land surface, cloud top, or smoke plume top. However, MISR does not always provide the complete vertical profile of the smoke plume, especially for thin smoke or smoke far from the source (Kahn et al., 2008). On the other hand, products derived from the Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observation (CALIPSO) satellite can provide both aerosol feature classifications (e.g. smoke, dust, etc.) and a complete picture of the vertical profile. However, data from CALIPSO are limited by its extremely narrow swath of observations. Given that CALIPSO passes over most locations only once every 16 days, it is unable to measure the vertical profile of most smoke plumes.

To achieve more spatially and temporally resolved information on the vertical distribution of smoke we need to turn to models rather than measurements. One approach is deterministic modelling, which relies on equations of the physical process of fire and smoke, as well as atmospheric transport (Paugam et al., 2016). In these models, the initial maximum smoke plume rise over the fire (injection height) and the subsequent vertical distribution of smoke (downwind height) are modelled separately, with the former as an important input for the latter. Such models are currently used for operational smoke forecasting systems, including the BlueSky (Larkin et al., 2010; Sakiyama, 2013) and FireWork (Pavlovic et al., 2016) frameworks in Canada. So far, most studies have focused on using CALIPSO and MISR data to evaluate and improve the injection height estimate, which describes the highest altitude of the smoke emission near the fire locations (Raffuse et al., 2012; Sessions et al., 2011; Sofiev et al., 2012; Val Martin et al., 2010). However, the more important estimate for public health is the downwind height, which describes smoke dispersion to the surface where human populations can be exposed.

Another approach is statistical modelling, which uses empirical data to describe the statistical relationships between relevant variables without considering the intermediate physical processes. We previously developed the near-real-time Optimized Statistical Smoke Exposure Model (OSSEM) to operationally estimate daily PM concentrations across the Canadian province of British Columbia (Yao et al., 2016; Yao and Henderson, 2014) using remote sensing data. However, the initial implementation of OSSEM did not account for the vertical profile of smoke in the atmosphere. This limitation may prevent improvement in the model performance, especially when we try to expand the model to finer temporal resolution. To address this limitation, we must develop a complementary near-real-time model that can operationally estimate the minimum height of smoke in the column. To the best of our knowledge there are no such models for injection height or downwind height, although some studies have recognized their feasibility and have identified potentially predictive variables related to fire intensity and meteorological conditions (Freitas et al., 2006; Freitas et al., 2007; Kahn et al., 2007; Mazzoni et al., 2007; Peterson et al., 2014; Val Martin et al., 2012). The process of building such models is further facilitated by the rapidly developing field of machine learning, where new methods can be used to model relationships that are far more sophisticated and complex than those that can be captured by conventional regression. The simplicity of statistical modelling with easily accessible data can provide timely estimates in the operational setting, especially in the situation where computational power or expertise to develop and maintain a complex deterministic model is lacking. However, most machine learning approaches come with a major limitation that the relationships between variables and their physical meanings are very difficult to properly assess.

In this study, we applied random forests to predict the minimum height of the smoke layer observed by CALIPSO using information about fire activity in the vicinity, geographic location, and meteorological conditions. These pieces of information are typically available in near-real-time, ensuring that the resulting model can be operationally applied. The predictions can be 1) directly applied to smoke detections from the existing remote sensing products; 2) incorporated into statistical models with inputs from remote sensing products; or 3) used to inform estimates of injection height and downwind dispersion in deterministic models. These potential applications are expected to improve the assessment of ground-level population exposure to forest fire smoke for both epidemiologic research and public health surveillance.

2. Methods

2.1. Study setting

The study area is the province of British Columbia (BC), on the west coast of Canada. With approximately 70% of its land area covered by forests, BC is prone to forest fires (B.C. Ministry of Forests, 2010). In recent years a severe pine beetle infestation has left approximately 18 million hectares even more susceptible, and several of the worst fire seasons on record have occurred in the past decade (Natural Resources Canada, 2016). Smoke from these fires has caused episodes of very poor air quality in many parts of the province during the fire season, which we have defined as April through September based on historical data. This study covers all forest fire seasons between 2006 and 2015, when data were available from all sources.

2.2. The CALIPSO data and the response variable

The CALIPSO satellite was launched in April 2006, as part of the US National Aeronautics and Space Administration (NASA) Afternoon Constellation (A-train), in a polar orbit with a 16-day repeat cycle (Fig. 1A). The primary instrument carried by CALIPSO is the Cloud-Aerosol Lidar with Orthogonal Polarization (CALIOP). The CALIOP instrument generates laser pulses at two wavelengths (532 nm and 1064 nm) every 333 metres (m) along the satellite ground track, and then collects the returned backscatter signals. Analysis of these signals allows for discrimination between cloud and aerosol, and provides insight into the vertical distribution, shape, size, and type of detected aerosols (Winker et al., 2006).

We obtained the CALIPSO Level 2 lidar vertical feature mask data (Version 3) from the CALIPSO Search and Subsetting Web Application for all overpasses in BC during the study period. These data were stored in a block for each 5-kilometre (km) segment of the footprint along the swath (Fig. 1A and B), with data cells at different vertical and horizontal resolutions depending on the altitude (Fig. 1B and C). The vertical resolution is 30 m for all altitudes below 8.2 km, which are the most relevant to human activity and exposure. A detailed description of the algorithm used to classify features of the CALIPSO data can be found elsewhere (Liu et al., 2005; Liu et al., 2009; Omar et al., 2009; Vaughan et al., 2005; Vaughan et al., 2009). In brief, each data cell is classified as *cloud*, *aerosol*, or *other*, with a quality assessment (QA) of *low*, *medium*, or *high*. The distinction between *cloud* and *aerosol* is primarily based on the scattering strength and the spectral dependence of backscattering. If a data cell is classified as *aerosol*, its type is then further classified into *clean marine*, *dust*, *polluted continental*, *clean continental*, *polluted dust*, *smoke*, or *other* (Fig. 1), with a binary QA of *confident* or *not confident*. Aerosol subtype classifications are primarily