A semi-empirical model for predicting hourly ground-level fine particulate matter (PM$_{2.5}$) concentration in southern Ontario from satellite remote sensing and ground-based meteorological measurements

Jie Tian*, Dongmei Chen

Department of Geography, Queen’s University, Kingston, ON, Canada K7L 2N6

**ARTICLE INFO**

Article history:
Received 16 January 2009
Received in revised form 7 September 2009
Accepted 7 September 2009

Keywords:
MODIS
Aerosol optical depth
PM$_{2.5}$
Meteorological fields
Modelling

**ABSTRACT**

A semi-empirical model is developed to predict the hourly concentration of ground-level fine particulate matter (PM$_{2.5}$) coincident to satellite overpass, at a regional scale. The model corrects the aerosol optical depth (AOD) data from the Moderate Resolution Imaging Spectroradiometer (MODIS) by the assimilated parameters characterizing the boundary layer and further adjusts the corrected value according to meteorological conditions near the ground. The model was built and validated using the data collected for southern Ontario, Canada for 2004. Overall, the model is able to explain 65% of the variability in ground-level PM$_{2.5}$ concentration. The model-predicted values of PM$_{2.5}$ mass concentration are highly correlated with the actual observations. The root-mean-square error of the model is 6.1 μg/m³. The incorporation of ground-level temperature and relative humidity is found to be significant in improving the model predictability. The coarse resolution of the assimilated meteorological fields limits their value in the AOD correction. Although MODIS AOD data is acquired on a daily basis and the valid data coverage can sometimes be very limited due to unfavourable weather conditions, the model provides a cost-effective approach for obtaining supplemental PM$_{2.5}$ concentration information in addition to the ground-based monitoring station measurement.

© 2009 Elsevier Inc. All rights reserved.

1. Introduction

Exposure to fine particulate matter with aerodynamic diameters less than 2.5 μm (PM$_{2.5}$) has negative effects on human health and may induce respiratory problems, cardiovascular and lung diseases, and additional health problems (Pope III et al., 2002; Kappos et al., 2004). Both short-term and long-term exposures to PM$_{2.5}$ have been linked to increased morbidity (Brunekreef & Holgate, 2002; Miller et al., 2007). The measurement of ground-level PM$_{2.5}$ concentration on a regular basis is therefore of great importance to epidemiological studies; it also provides valuable information for an effective management and forecasting of air quality (Al-Saadi et al., 2005). Air quality monitoring networks have been established in many industrialized countries to take measurement of pollutant concentrations at different locations, on a daily or hourly basis. The temporal resolution and accuracy of air quality data vary depending on the measurement method (Cyrys et al., 2001).

Although ground-based measurements are generally considered to be accurate, they are representative for only relatively small areas around point stations. Often, the limited spatial coverage and irregular distribution of ground-based monitoring stations largely restrict the study on space–time dynamics of air pollution and its impacts on human health and the environment. Alternatively, complex process-based air pollution models, which estimate pollutant concentrations by considering pollutant generation, transportation, and removal, are hampered in a lot of cases by the incomplete information of anthropogenic emission inventories and natural sources (Koelemeijer et al., 2006).

Satellite remote sensing has been employed to supplement the prediction of ground-level PM$_{2.5}$ concentration (GL-[PM$_{2.5}$]). Satellites are able to cover vast spaces at a relatively low cost. For aerosol studies, the launch of the Moderate Resolution Imaging Spectroradiometer (MODIS) has enabled the retrieval of aerosol optical depth (AOD) data globally from the satellite's spectral observation. MODIS AOD is a measure of light extinction by aerosols in an atmospheric column during the satellite overpass. With the evolution of the retrieval algorithm, MODIS AOD has become increasingly important in the role of producing more accurate estimation for the GL-[PM$_{2.5}$]. The Multiple Imaging Spectroradiometer (MISR) has the capability of providing aerosol data globally as well. MODIS and MISR AOD data differ in their supplemental PM$_{2.5}$ concentration information in addition to the ground-based monitoring station measurement.

A number of researches have been devoted to exploiting the quantitative relationship between satellite-measured AOD and GL-
[PM\textsubscript{2.5}]. Efforts have been made to investigate the correlation between these two variables and its variation across space and time (Engel-Cox et al., 2004; Hutchison et al., 2004; Hutchison et al., 2005; Gupta et al., 2006). Although the strength of correlation between satellite-measured AOD and GL-[PM\textsubscript{2.5}] was reported to be different for various study areas, satellite remote sensing has shown a great potential in the prediction of ground-level conditions. Attempts have been made to model the AOD–PM\textsubscript{2.5} relationship by incorporating certain environmental factors. For example, empirical models have been developed in a number of regions ranging from simple linear relationships to more complex nonlinear relationships involving meteorological and geographical data (Liu et al., 2005; Koelemeijer et al., 2006; Kumar et al., 2007). However, the predictive power of the models in the literature is not very satisfactory, even taking into account those models that require a relatively large number of inputs. Although certain columnar atmospheric characteristics (e.g. model-approximated boundary layer height) have been accounted, the meteorological condition at the ground level has not been explicitly considered in the process of modelling GL-[PM\textsubscript{2.5}].

In this paper, we intend to develop a new model that can be used to predict GL-[PM\textsubscript{2.5}] at a regional scale. The model is designed to synthetically utilize satellite observations, assimilated meteorological fields, and ground-based meteorological measurements, and is regarded as a semi-empirical model as the constitution of the model draws insights from the related theories and is statistically driven by data.

2. MODIS AOD retrieval and its relation to ground-level PM\textsubscript{2.5} concentration (GL-[PM\textsubscript{2.5}])

Aerosols are generally recognized as solid or liquid particles suspended in the air. AOD (\(\tau\)) is defined as the integral of aerosol extinction along an atmospheric column from the ground to the top of the atmosphere (Chylek et al., 2005). Theoretically, \(\tau\) can be calculated by:

\[
\tau = \int \alpha_{ext}(h)dh
\]

where \(\alpha_{ext}(h)\) is the aerosol extinction coefficient at the vertical elevation \(h\) above the ground. AOD is parameterized as the log of the ratio of irradiance at the top of atmosphere to irradiance at the surface. With the aid of a MODIS equipped satellite, AOD can be computed using the radiative transfer model based on radiances recorded by the sensor (Kaufman et al., 1997).

In comparison, GL-[PM\textsubscript{2.5}] measures the mass concentration of fine particulate matter (\(\leq 2.5 \, \mu\text{m}\) in diameter) in the atmosphere near the ground. The original AOD can be corrected (AOD\*) to better correlate with GL-[PM\textsubscript{2.5}] by using the following equation (Koelemeijer et al., 2006):

\[
\text{AOD}^{*} = \frac{\text{AOD}}{f(\text{Humidity})/\text{BLH}}
\]

where BLH denotes the planet boundary layer height and \(f(\text{Humidity})\) is a function of columnar humidity. The correction is to remove the effects of water vapor and atmospheric column height on the optical extinction efficiency.

Kaufman et al. (2003) suggest that the majority of aerosols reside in the lower troposphere, particularly in the planetary boundary layer (PBL) where the particles are more evenly distributed due to the active mixing. In the PBL, the properties of aerosols normally do not change substantially across different altitudes, especially when convection is salient. The impact of water vapor on AOD can be quantified as a function of the boundary layer humidity. A number of researchers (Chin et al., 2002; Liu et al., 2005; Donkelaar et al., 2006; Koelemeijer et al., 2006) have suggested that air humidity accounts for part of the optical extinction. For example, when air humidity is high, hygroscopic particles can grow exponentially in size, resulting in a significant increase of their extinction efficiencies and the overestimation of aerosol mass. On the other hand, a thicker PBL usually corresponds to a lower aerosol density for a given AOD value. BLH is therefore regarded as a denominator of MODIS AOD in its relationship to GL-[PM\textsubscript{2.5}]. Moreover, there are a number of other factors that link AOD and GL-[PM\textsubscript{2.5}], including particle composition, size distribution, and vertical profile. The information on these factors may be used to further improve the modelling of the AOD–PM\textsubscript{2.5} relationship (Hutchison et al., 2008).

3. Data collection and processing

A number of data from various sources were collected for this research, including the historical air quality data sets, MODIS aerosol imagery, assimilated meteorological fields characterizing columnar atmospheric properties, and ground-based meteorological measurements.

3.1. Ground-based PM\textsubscript{2.5} concentration data

The Ontario Ministry of Environment (OME) operates a monitoring network measuring ground-level PM\textsubscript{2.5} concentration (GL-[PM\textsubscript{2.5}]) on an hourly basis. The data is used to check compliance with acceptable values and report risks to the public. Hourly GL-[PM\textsubscript{2.5}] measurements (in \(\mu\text{g}/\text{m}^3\)) throughout year 2004 were obtained from the OME's electronic archive of historical air quality data. These measurements were taken by the monitoring stations across southern Ontario. According to the OME (1998), continuous measurements of hourly GL-[PM\textsubscript{2.5}] at the ground level are obtained by the Tapered Element Oscillating Microbalance (TEOM) method. This method measures the accumulation of mass on a filter that is attached to the tip of a hollow, tapered, and oscillating glass rod. Direct measurement of mass accumulation on the filter is obtained based on the change in the oscillating frequency over time. It should be noted that this method measures the dry mass of the aerosols under investigation and may undervalue the aerosol concentration due to aerosol evaporation. Invalid data (values are documented in the metadata) due to extreme weather conditions or equipment disorder was removed to only keep those valid data that represent the real concentrations of fine particulate matter around the station locations. Fifteen stations were selected to provide the data for modelling, while another 15 stations among the rest were preserved for validation (Fig. 1). Both the modelling stations and the validation stations were selected to have a reasonable spatial coverage and distribution. Such a selection was done on the intention of minimizing, if not eliminating, the possible spatial autocorrelation effects in the modelling or validation. Neither group of stations showed noticeable spatial distribution patterns (e.g. clusters).

3.2. MODIS satellite data

Aerosol Optical Depth (AOD), representing columnar aerosol loading of the atmosphere, is retrieved as a level 2 product (5-minute swath granules) from the MODIS observation, at a typical spatial resolution of 10 km. Two separate algorithms are applied for the retrieval of aerosols over land (Kaufman et al., 1997) and ocean (Tanré et al., 1997). The strategy for retrieving aerosol over land from MODIS was first introduced by Kaufman et al. (1997). The AOD over land can be retrieved at three wavelengths: 0.47 \(\mu\text{m}, 0.55 \mu\text{m}, \) and 0.66 \(\mu\text{m}.\) Although the algorithm for data retrieval has continued to evolve to achieve better accuracy over the past few years, the theoretical basis of the algorithm has not changed since its inception. The strategy of the algorithm is to examine a lookup table to determine the conditions that can best mimic the MODIS-observed spectral reflectance, and retrieve the associated aerosol properties including AOD. Although a
number of factors (e.g., the assumed aerosol model, the empirical relationship between the surface reflectance at different wavelengths) may induce uncertainty, MODIS AOD has been proven to be quite accurate by comparison to the sunphotometer measurements.

The AOD values (the field of Corrected_Optical_Depth_Land) used in this study were derived by the MODIS aerosol data team, based on their collection 5 algorithm (C005) of aerosol property retrieval. It was found from our previous studies that the AOD at 0.47 μm has the strongest correlation with GL-[PM2.5] (Tian, 2008), in comparison to the AOD at the other wavelengths (0.55 μm and 0.66 μm). Therefore, the AOD at 0.47 μm was used in the following modelling process. In total, there were 807 and 844 MODIS AOD images collected from Terra and Aqua, respectively, each of which included all of, or a portion of southern Ontario depending on acquisition time. Closer examination of the images showed that most of the valid AOD values were taken from April to November.

### 3.3. GEOS-4 meteorological fields

Meteorological fields including boundary layer height (BLH), specific humidity (SH), air pressure (AP), and air temperature (AT) were obtained from the NASA Global Modelling Assimilation Office. These meteorological fields are assimilated based on the various observations of the Goddard Earth Observing System (GEOS). The GEOS-4 meteorological fields were provided at a resolution of 1° latitude x 1.25° longitude. While the BLH data is two-dimensional, the SH and AT data are provided in three dimensions, including a stack of layers assimilated at 55 vertical layers. SH, AT, and AP were used to calculate the relative humidity (RH) at each of the six bottom layers, which together correspond to the lower troposphere (approximately the lower 2–3 km of the atmosphere). A well-known equation for calculating saturation water vapor (Buck, 1981) was adopted so that RH could be derived as the ratio of water vapor pressure (a function of SH and AP) over saturation water vapor pressure (a function of AT and AP). The GEOS-4 columnar relative humidity (GRH) of the lower troposphere was subsequently estimated by averaging the obtained RH values for the bottom atmospheric layers. The collected GEOS-4 data covers the time period of June to October instead of the full year due to its availability.

### 3.4. Ground-based meteorological measurements

The ground-based meteorological measurements of year 2004, including surface wind speed (SWS), surface temperature (ST), and surface relative humidity (SRH), were collected from the weather monitoring stations distributed across southern Ontario. The data was originally provided by Environment Canada (EC) and organized by watersheds. Only the data from the stations that provide hourly measurements were included in this study.

### 3.5. Data pre-processing and integration

The data from the above four sources were co-located in both space and time to establish a complete dataset that serves as the basis for the following analyses and modelling. The point-based station data (including GL-[PM2.5], ST, SRH, and SWS) and the area-based image data (including MODIS AOD, GEOS-4 BLH, and GEOS-4 RH) were associated in space based on their proximity. For instance, a GL-[PM2.5] reading from an air quality station was associated with the value of the MODIS AOD pixel covering the station. The BLH and GRH values were interpolated, using the inverse distance weighting method, for each station location from its four closest pixel centres. The meteorological data from the closest weather station was used to represent the meteorological condition for an air quality monitoring station. In time, a nearly instant MODIS AOD measurement was associated with the coincident hourly measurements of GL-[PM2.5], ST, SRH, and SWS, and multi-hour averages of GEOS-4 BLH (3-hour average) and GRH (6-hour average). Only the data that were spatially co-located at the selected 30 ground stations and temporally matched at the MODIS overpasses were included in the following analyses. More detailed characteristics of the data variables involved in the modelling process are summarized in Table 1. The entire pre-processed...
dataset was further divided into the modelling set and the validation set according to the station that measured the PM$_{2.5}$ data (refer to Section 3.1). The data in the modelling set was subsequently sampled to have longer time lag (more than 90 h) between the observations. This is intended to minimize the impact of temporal autocorrelation during the modelling.

4. Model development

Defining the form of a model for ground-level PM$_{2.5}$ concentration (GL-[PM$_{2.5}$]) prediction is critical in the modelling process. Deterministic models that explicitly describe the pollution process often have intensive data demands. For instance, one necessary input data required by many deterministic models is the inventory of pollution sources, which can be difficult to map out or quantify for their emissions. This is particularly the case for regional-scale studies, in which cities or towns are usually regarded as individual pollution sources. Instead, a statistical model can be developed to more effectively describe the quantitative or numerical relationship existing between the predictors and GL-[PM$_{2.5}$]. Two models were developed and compared in this research: (1) a simple linear regression model (termed as Model-I) that considers a linear relationship between MODIS AOD and GL-[PM$_{2.5}$], and (2) a semi-empirical model (Model-II) that considers the coexistence of the values for all the included variables. Since Model-I incorporates fewer variables than Model-II, there were more valid cases in the modelling dataset that can be used in Model-I than in Model-II.

It was therefore decided to develop Model-I based on two groups of modelling data. Model-IA used all valid data pairs of AOD and GL-[PM$_{2.5}$] from the modelling dataset, while Model-IB only used the pairs involved in the parameterization of Model-II. The model coefficients (slopes and intercepts) were determined by regressing GL-[PM$_{2.5}$] on AOD for these two groups of data.

In essence, Model-II consists of two major components: corrected MODIS AOD by the assimilated GEOS-4 meteorological fields, and the ground-based meteorological measurements. While the former characterizes the columnar aerosol density, the latter can be understood as the sensitive impact factors that significantly influence the GL-[PM$_{2.5}$]. Based on its relation to GL-[PM$_{2.5}$] described in Section 2, MODIS AOD was corrected in the manner shown by Eq. (2). GEOS-4 BLH and the calculated GEOS-4 RH (or GRH) were used as the representative variables of mixing height and boundary layer humidity, respectively. As highlighted in Section 1, the response of GL-[PM$_{2.5}$] to meteorological condition changes at the ground level has not been well addressed and still remains unclear. To perceive some understanding of the co-variation between GL-[PM$_{2.5}$] and the possible predictors prior to the modelling, GL-[PM$_{2.5}$] was plotted against ST, SRH, and SWS. A LOESS (locally weighted polynomial regression) trend line was added on each scatter-plot in Fig. 2. LOESS was used because it combines much of the flexibility of nonlinear regression. It fits simple models to localized subsets of the data and therefore does not require to fit a global model to the entire data. For more details, please refer to Cleveland and Devlin (1988).

As can be seen from Fig. 2A, despite the random variations, there seems to exist a semi-exponential relationship between GL-[PM$_{2.5}$] and ST. GL-[PM$_{2.5}$] does not change substantially when the ST is below around 18 °C, but increases superlinearly at higher temperatures and demonstrates an exponential trend. In comparison, SRH seems to be associated with GL-[PM$_{2.5}$] in a nearly linear fashion with a positive slope (see Fig. 2B). No noticeable co-variation is found between SWS and GL-[PM$_{2.5}$] (see Fig. 2C) as the points appear to be randomly distributed. SWS

<table>
<thead>
<tr>
<th>Variable name</th>
<th>Unit</th>
<th>Frequency</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ground-level concentration of fine particulate matter (GL-[PM$_{2.5}$])</td>
<td>μg/m$^3$</td>
<td>Hourly</td>
<td>OME air quality monitoring stations</td>
</tr>
<tr>
<td>MODIS AOD at 0.47 μm (AOD)</td>
<td>Unitless</td>
<td>Daily</td>
<td>MODIS satellite</td>
</tr>
<tr>
<td>GEOS-4 boundary layer height (BLH)</td>
<td>m</td>
<td>Every 6 h</td>
<td>GEOS-4 model assimilation</td>
</tr>
<tr>
<td>GEOS-4 relative humidity (GRH)</td>
<td>%</td>
<td>Every 6 h</td>
<td>GEOS-4 model assimilation</td>
</tr>
<tr>
<td>Surface temperature (ST)</td>
<td>°C</td>
<td>Hourly</td>
<td>EC weather monitoring stations</td>
</tr>
<tr>
<td>Surface RH (SRH)</td>
<td>%</td>
<td>Hourly</td>
<td>EC weather monitoring stations</td>
</tr>
<tr>
<td>Surface wind speed (SWS)</td>
<td>km/h</td>
<td>Hourly</td>
<td>EC weather monitoring stations</td>
</tr>
</tbody>
</table>

Fig. 2. Scatter plots of the ground-based meteorological measurements (A. surface temperature (ST); B. surface relative humidity (SRH); and C. surface wind speed (SWS)) against the dependant variable of ground-level PM$_{2.5}$ concentration (GL-[PM$_{2.5}$]).
was therefore decided not to be included in the modelling. It should also be noted that SRH and GRH were found not to be correlated. Therefore, both SRH and GRH were incorporated in our modelling process.

Based on the above pre-modelling analyses, the form of the semi-empirical model was decided and can be expressed as:

$$GL-\text{[PM}_{2.5}] = \exp(\beta_0) \cdot AOD^{\beta_{AOD}} \cdot BLH^{\beta_{BLH}} \cdot \exp(\beta_{GRH} \cdot GRH) \cdot \exp(\beta_{ST} \cdot ST) \cdot SRH^{\beta_{SRH}}$$

(3)

where GL-\text{[PM}_{2.5}], the dependent variable on the left-hand-side, is the hourly average of ground-level PM\textsubscript{2.5} concentration. The independent variables (or predictors) on the right-hand side include remotely sensed MODIS AOD, assimilated GEOS-4 boundary layer height (BLH) and relative humidity (RH), and ground-based meteorological measurements of surface temperature (ST) and surface relative humidity (SRH). The parameters \(\beta\) with different subscripts denote the respective regression coefficients. An exponential function of ST was used due to its superlinear impact to GL-\text{[PM}_{2.5}] shown by Fig. 2A. The AOD correction, using BLH and GRH, assumes a relatively smooth vertical profile of particulate matter so that its concentrations at various altitudes are correlated to the ground-level concentration. However, because the vertical profile is likely to be skewed to some degree near the ground, the surface meteorological condition should also be taken into consideration. ST and SRH were therefore employed to further adjust the corrected AOD values to approach the real GL-\text{[PM}_{2.5}] values.

As Model-II incorporates the closely related meteorological variables from both GEOS-4 model assimilation and ground-based measurement, it has a stronger physical meaning than Model-I. It should also be emphasized that the model predicts the hourly GL-\text{[PM}_{2.5}] coincident to the MODIS satellite overpass, at a regional scale. To facilitate the model parameterization, Eq. (3) was log-transformed into a linear regression form for the coefficients to be more easily resolved.

$$\ln(GL-\text{[PM}_{2.5}]) = \beta_0 + \beta_{AOD} \ln(AOD) + \beta_{BLH} \ln(BLH) + \beta_{GRH} \ln(GRH) + \beta_{ST} \ln(ST) + \beta_{SRH} \ln(SRH)$$

(4)

5. Results and discussion

5.1. Descriptive statistics

Descriptive statistics of the data used for the parameterization of Model-I/A/B and Model-II are summarized in Figs. 3 and 4, respectively. The MODIS AOD data used in Model-I/A has a mean value of 0.27 and a standard deviation of 0.25. The corresponding hourly ground-level PM\textsubscript{2.5} concentration (GL-\text{[PM}_{2.5}]) ranges from 0 \(\mu\text{g/m}^3\) to 78 \(\mu\text{g/m}^3\), with a mean value of 6.8 \(\mu\text{g/m}^3\). Not surprisingly, these two data exhibit a similar frequency distribution towards the lower bound over their value ranges, as these two variables have been proven to be well correlated. Model-I/B used the same MODIS AOD and GL-\text{[PM}_{2.5}] data as Model-II.

The data used to develop Model-II fall in a time frame of June to October due to the unavailability of the GEOS-4 data other than that time period in this research. The involved GL-\text{[PM}_{2.5}] data has a mean value of 11.7 \(\mu\text{g/m}^3\) and a standard deviation of 11.0 \(\mu\text{g/m}^3\). The maximum hourly GL-\text{[PM}_{2.5}] reaches a level of 48 \(\mu\text{g/m}^3\). Meanwhile, MODIS AOD is found to have a mean value of 0.36 and a standard deviation of 0.30 (Fig. 4A and B). The MODIS AOD and GL-\text{[PM}_{2.5}] data used for Model-II exhibit a similar frequency distribution as well. BLH (993 m \pm 410 m) and GRH (60% \pm 15.2%) are found to be less variable when compared with MODIS AOD and GL-\text{[PM}_{2.5}]; they both show a nearly normal distribution (Fig. 4C and D). For the ground-based meteorological measurements, ST varies dramatically over the seasons but is primarily within the range of 14 °C to 31 °C in our dataset. A relatively high mean value of SRH (61%) implies the humid atmospheric environment due to the existence of sufficient water bodies (e.g. the Great Lakes) within and around southern Ontario. The SRH data also exhibits a nearly normal distribution, with a minimum and maximum value of 30% and 89%, respectively (Fig. 4E and F).

5.2. Model parameterization

5.2.1. Model-I

Hourly GL-\text{[PM}_{2.5}] was first regressed on all the MODIS AOD data (558 of them) in the modelling dataset. The resultant model (Model-I) was found to be significant (\(p<0.001\)) and could explain 38% of the variability in the corresponding hourly GL-\text{[PM}_{2.5}] (Table 2). The slope of 22.2 is very close to the value (22.6) presented in Engel-Cox et al. (2004), which correlates MODIS AOD and GL-\text{[PM}_{2.5}] across the continental U.S. The difference between the estimated constant, 7.5 in Engel-Cox et al. (2004) and 1.0 here, may reflect the different background levels of GL-\text{[PM}_{2.5}] in the U.S. and in southern Ontario, Canada.

Hourly GL-\text{[PM}_{2.5}] was then regressed on the MODIS AOD data (195 of them) that was included in the parameterization of Model-II. The resultant model (Model-II) was also found to be significant (\(p<0.001\)). A larger slope and higher intercept were obtained, implying an increased sensitivity and elevated background levels, respectively. In comparison to Model-I, Model-II represents the linear AOD-PM\textsubscript{2.5}
relationship in the time period of summer and fall (June to October). Overall, Model-I provides a simple but effective way to estimating GL-[PM$_{2.5}$], especially when MODIS AOD is the only data available for the area of interest.

5.2.2. Model-II

The model described by Eq. (4) was fitted using the modelling dataset to estimate the model coefficients. Overall, the resultant model is found to be highly significant ($p<0.001$) and is able to explain 65% of the variability in the hourly GL-[PM$_{2.5}$] data (Table 3). This model surpasses Model-IB by an $r^2$ increase of 14%, which is substantial in the context of modelling ability. Closer examination of the model reveals that MODIS AOD is a highly significant predictor ($p<0.001$). The estimated power of MODIS AOD (0.328±0.052) was positive and smaller than 1, meaning GL-[PM$_{2.5}$] varies sublinearly with MODIS AOD measurements when the other predictors are controlled.

Although a negative coefficient ($-0.054$) was found for BLH, this predictor did not appear to be significant. It is understandable that aerosol density should increase when the boundary layer shrinks, and vice versa, given the fact that aerosols are mainly trapped within this layer of the atmosphere. The lack of significance may be attributed to the indirect method of measurement employed since GEOS-4 BLH values were obtained by model assimilation; furthermore, GEOS-4 BLH are provided in a very coarse resolution of 1°×1.25° and cannot capture the detailed variations over shorter distances (e.g. less than 100 km). GEOS-4 BLH may become an ineffective predictor if such variations are significant in the study area.

Table 2
Estimated coefficients of the simple linear regression model (Model-I) for the prediction of hourly ground-level PM$_{2.5}$ concentration (GL-[PM$_{2.5}$], μg/m$^3$) using MODIS AOD.

<table>
<thead>
<tr>
<th>Model predictors</th>
<th>Model-IA</th>
<th>Model-IB</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
<td>Std error</td>
</tr>
<tr>
<td>AOD</td>
<td>22.212</td>
<td>1.203</td>
</tr>
<tr>
<td>Intercept</td>
<td>1.036</td>
<td>0.437</td>
</tr>
</tbody>
</table>

Goodness of fit $r^2$ 0.38 0.51

Table 3
Regression coefficients of the semi-empirical model (Model-II) for the prediction of hourly ground-level PM$_{2.5}$ concentration (GL-[PM$_{2.5}$], μg/m$^3$) using MODIS AOD, GEOS-4 BLH, GEOS-4 RH, ST, and SRH.

<table>
<thead>
<tr>
<th>Model predictors</th>
<th>Estimate ($β$)</th>
<th>Std error</th>
<th>$p$ value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ln(AOD)</td>
<td>0.328</td>
<td>0.052</td>
<td>$&lt;0.001$</td>
</tr>
<tr>
<td>Ln(BLH)</td>
<td>$-0.054$</td>
<td>0.146</td>
<td>0.712</td>
</tr>
<tr>
<td>GRH</td>
<td>$-0.008$</td>
<td>0.004</td>
<td>0.064</td>
</tr>
<tr>
<td>ST</td>
<td>0.140</td>
<td>0.018</td>
<td>$&lt;0.001$</td>
</tr>
<tr>
<td>Ln(SRH)</td>
<td>1.678</td>
<td>0.236</td>
<td>$&lt;0.001$</td>
</tr>
<tr>
<td>Constant</td>
<td>$-6.700$</td>
<td>1.665</td>
<td>$&lt;0.001$</td>
</tr>
</tbody>
</table>

Fig. 4. Histograms and descriptive statistics of the variables (A. GL-[PM$_{2.5}$], B. MODIS AOD, C. GEOS-4 Boundary Layer Height (BLH), D. GEOS-4 Relative Humidity (GRH), E. Surface Relative Humidity (SRH), and F. Surface Temperature (ST)) involved in the development of Model-II.
The average GRH of the bottom atmospheric layers was found to be significant to the GL-PM2.5 prediction. The negative coefficient \((-0.008 \pm 0.004)\) indicates that the presence of water vapor accounts for part of the optical extinction sensed by MODIS. Given an AOD value, the higher the GRH is, the lesser the actual columnar aerosols are likely to be. The reason for this is that hygroscopic particles such as ammonium sulfates and ammonium nitrates can grow exponentially in size and dramatically enhance their light extinction efficiencies (Malm et al., 2000). The model therefore corrects the humidification effect in the optical depth estimation by MODIS. It was found that the incorporation of surface temperature (ST) and surface relative humidity (SRH) can significantly enhance the predictive power of the model. In particular, ST was found to be highly significant \((p<0.001)\). The positive sign of the estimated coefficient for surface temperature indicated that, a corrected AOD (AOD*) by PBL and GRH should correspond to a higher GL-PM2.5 near the ground as a response to a higher ST. Holben et al. (2001) suggests that higher insolation may enhance photochemical conversion of gases to aerosols. According to Dawson et al. (2007), the response of GL-PM2.5 to temperature was mainly the result of competing changes in sulfate and nitrate concentrations with a minor role played by organics. More specifically, PM2.5 sulfate concentrations tend to increase as temperature increases, while PM2.5 nitrate concentrations decrease. The concentration of sulfates increases faster than the corresponding decrease in nitrates resulting in an overall increase in PM2.5 (Dawson et al., 2007). As can be seen from Fig. 2A, when the surface temperature increases, the net effect of the competing changes is an elevated GL-PM2.5 in southern Ontario. On the other hand, higher temperatures are associated with summer months, in which the prevailing winds may affect the relative contributions of different aerosol sources to the PM2.5 concentration in the study area. SRH was also found to be significant \((p<0.001)\) in predicting GL-PM2.5. Given a certain AOD*, the actual GL-PM2.5 tends to have a higher value when the SRH increases. A positive correlation between GL-PM2.5 and SRH was also found in Barman et al. (2008). It has been suggested by Seinfeld and Pandis (2006) that, increases in humidity shift the equilibrium of the ammonia-nitric acid system toward the aerosol phase, and therefore results in elevated concentrations of ammonium nitrate aerosol. This may be able to explain the increase of GL-PM2.5 as a response to the increase of SRH.

Since BLH was found not to be significant in the modelling, the regression was subsequently performed again based only on the four independent predictors, excluding BLH. The resultant model (see Table 4) is highly comparable with the previous one in terms of goodness of fit \((r^2 =0.65)\) and the estimated coefficients. This verifies that the GEOS-4 BLH data contributes little, if nothing, to the GL-PM2.5 prediction of our model.

5.3. Model validation

The models developed in Section 5.2 were applied to the validation dataset in order to evaluate their performance in predicting hourly ground-level PM2.5 concentrations (GL-PM2.5). Model-IA was applied to all the MODIS AOD data (1582 of them) available in the validation dataset. The predicted values yielded a correlation coefficient of 0.6 \((r^2 =0.36)\) with the observed values. The root-mean-square error (RMSE) of this model was found to be 7.3 \(\mu g/m^3\) for a mean GL-PM2.5 of 7.8 \(\mu g/m^3\). As can be seen from Fig. 5A, the points appear to be rather scattered and mainly fall in a wide belt covering the regression line. Applying Model-IB to the MODIS AOD data (569 of them) that was used in the validation of Model-II resulted in a comparatively higher agreement \((r^2 =0.49)\) between the predicted values and the observed values (Fig. 5B). A RMSE of 9.7 \(\mu g/m^3\) was found for the validation data with a mean of 10.3 \(\mu g/m^3\). Although the predictions are not very accurate, a simple linear model is valuable to attempt the estimation of GL-PM2.5 merely from MODIS AOD (especially in summer and fall time). For some rural or remote areas, satellite observations like MODIS AOD may be the only data source available for the local GL-PM2.5 estimation.

Model-II, the proposed semi-empirical model, provided relatively better prediction results for most of the cases used for validation. In comparison to Model-IB, the predicted values by Model-II are more strongly correlated with the observed ones \((r^2 =0.64)\). It was noticed that the model tends to underestimate GL-PM2.5 when their actual values are relatively high (Fig. 6A). The regression line is very close to \(y=x\) when the values larger than 25 \(\mu g/m^3\) are excluded (Fig. 6B). This is mainly because that about 98% and 88% of the GL-PM2.5 values in the modelling dataset were below 40 \(\mu g/m^3\) and 25 \(\mu g/m^3\), respectively. The model was regressed to fit these data and hence does not represent the relationship between the predictors and higher GL-PM2.5 values very well. A RMSE of 6.1 \(\mu g/m^3\) was obtained for the entire validation dataset with a mean GL-PM2.5 of 10.3 \(\mu g/m^3\). For the observed GL-PM2.5 values of \(>20 \mu g/m^3\) (mean =29.1 \(\mu g/m^3\)), the model produced a RMSE of 15.7 \(\mu g/m^3\). Very high GL-PM2.5 values (e.g., \(>60 \mu g/m^3\)) are much less frequent when compared to the lower ones and they are often induced by some pollution events affecting only the nearest station and its relatively small surrounding area. The instant observations of MODIS AOD can hardly represent these elevated levels of GL-PM2.5 unless they last fairly long and impose impact on a considerably large area.

When dividing the validation dataset by land use, it was found that the model yielded slightly better predictions at agricultural/rural stations \((r^2 =0.70)\) than urban stations \((r^2 =0.64)\). There were 5 agricultural/rural stations and 10 urban stations in our validation dataset. Particulate matter is likely to be more evenly distributed over agricultural areas due to the relatively homogeneous background of vegetation and the absence of anthropogenic sources of PM2.5. Therefore, the models making use of MODIS AOD that characterize a relatively low area can predict GL-PM2.5 more precisely in agricultural/rural areas.

There are several factors accounting for the models’ errors, including the relatively coarse spatial resolution of AOD, the uncertainty in the measurement or estimation of the meteorological variables, and the incomplete understanding of the relationship between GL-PM2.5 and the predictors. More information (e.g. land use and vertical profile of aerosol concentration) may be incorporated to further improve the model accuracy. The data samples for the modelings and validations are believed to be representative for the situation in southern Ontario, but may not be ideal for testing the models’ performance for predicting higher levels of GL-PM2.5. It is needed to apply the proposed model to the data collected for more heavily polluted regions by PM2.5 in the future research.

6. Concluding remarks

This paper has proposed an effective semi-empirical model for the prediction of ground-level PM2.5 concentration (GL-PM2.5). Remotely sensed MODIS AOD, GEOS-4 assimilated relative humidity, and ground-based measurements of surface temperature and surface relative humidity have been found to be highly significant in the prediction. The model was able to explain around 65% of the variability

<table>
<thead>
<tr>
<th>Table 4</th>
<th>Regression coefficients of the semi-empirical model (Model-II) for the prediction of hourly ground-level PM2.5 concentration (GL-PM2.5, (\mu g/m^3)) using MODIS AOD, GEOS-4 RH (GRH), ST, and SRH.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model predictors</td>
<td>Estimate ((\hat{b}))</td>
</tr>
<tr>
<td>Ln(AOD)</td>
<td>0.316</td>
</tr>
<tr>
<td>GRH</td>
<td>-0.007</td>
</tr>
<tr>
<td>ST</td>
<td>0.141</td>
</tr>
<tr>
<td>Ln(SRH)</td>
<td>1.694</td>
</tr>
<tr>
<td>Constant</td>
<td>-7.238</td>
</tr>
</tbody>
</table>
in the hourly GL-[PM\textsubscript{2.5}] collected for this study, and predicted the realizations with a RMSE of 6.1 μg/m\textsuperscript{3}. Closer examination revealed that the model tends to underestimate the relatively larger GL-[PM\textsubscript{2.5}] values. Moreover, the model was able to provide slightly better prediction results for agricultural/rural areas than urban areas. In essence, the model corrects MODIS AOD by considering the optical extinction caused by humidi- 
fication and uses surface temperature and surface relative humidity to further calibrate the corrected AOD value towards the PM\textsubscript{2.5} concentration at the ground level. The form of the model was decided by analyzing the numerical relationship between the possible predictors and the dependent variable of GL-[PM\textsubscript{2.5}]. The observed response of GL-[PM\textsubscript{2.5}] to surface temperature and relative humidity can be used to compare with the results from the studies of particulate matter composition in southern Ontario. The proposed model is expected to be easily applied to other regions with some degree of coefficient modification. More studies are needed to compare the simple linear model with the proposed model for other regions and time scales before a general conclusion about its performance can be made.

AOD data with a finer spatial resolution and more detailed information about boundary layer height may be necessary to further improve the model predictability. Metrics characterizing land use structure should also be developed and incorporated in the model in future research. In addition, it should also be noted that the proposed model cannot be used to replace the monitoring stations or accomplish the monitoring task consistently throughout a year, simply because the data available is not guaranteed due to cloud contamination or snow cover. The model is developed to help predict the GL-[PM\textsubscript{2.5}] for the hours of MODIS overpasses, from which the GL-[PM\textsubscript{2.5}] at other hours may be interpolated. Epidemiological studies should more or less benefit from the accuracy improvement of the proposed new model over the simple linear model. Although there is still a considerable uncertainty associated with the model prediction, PM\textsubscript{2.5} air quality levels (e.g. good, moderate, sensitive, or bad) can be more confidently estimated. The simple linear regression model is still rather valuable for its simplicity and reliance on fewer data sources. These satellite data-based models enable the GL-[PM\textsubscript{2.5}] estimation over large areas, and allow the attempt to analyze the spatial–temporal response of human health to PM\textsubscript{2.5} concentration distribution.

**Fig. 5.** Scatter plots of the observed vs. the predicted hourly ground-level PM\textsubscript{2.5} concentrations (GL-[PM\textsubscript{2.5}], μg/m\textsuperscript{3}) by applying Model-IA (A) to all the MODIS AOD available in the validation dataset; and by applying Model-IB (B) to the MODIS AOD data that was used to validate Model-II.

**Fig. 6.** Scatter plots of the Model-II-predicted hourly ground-level PM\textsubscript{2.5} concentration (GL-[PM\textsubscript{2.5}], μg/m\textsuperscript{3}) vs. the observed GL-[PM\textsubscript{2.5}] that are less than 50 μg/m\textsuperscript{3} (A) and less than 25 μg/m\textsuperscript{3} (B), respectively.

References


Chin, M., Ginoux, P., Kinne, S., Torres, O., Holben, B. N., Duncan, B. N., et al. (2002). Tropospheric aerosol optical thickness from the GOCART Model and comparisons


