The Annals of Applied Statistics
2016, Vol. 10, No. 2, 756–785
DOI: 10.1214/15-AOAS901

A STATISTICAL MODELING APPROACH FOR AIR QUALITY DATA BASED ON PHYSICAL DISPERSION PROCESSES AND ITS APPLICATION TO OZONE MODELING

BY XIAO LIU*, KYONGMIN YEO*, YOUNGDEOK HWANG*, JITENDRA SINGH† AND JAYANT KALAGNANAM*

*IBM T.J. Watson Research Center** *and IBM Research Collaboratory Singapore*†

For many complex environmental processes such as air pollution, the underlying physical mechanism usually provides valuable insights into the statistical modeling. In this paper, we propose a statistical air quality model motivated by a commonly used physical dispersion model, called the scalar transport equation. The emission of a pollutant is modeled by covariates such as land use, traffic pattern and meteorological conditions, while the transport and decay of a pollutant are modeled through a convolution approach which takes into account the dynamic wind field. This approach naturally establishes a nonstationary random field with a space–time nonseparable and anisotropic covariance structure. Note that, due to the extremely complex interactions between the pollutant and environmental conditions, the space–time covariance structure of pollutant concentration data is often dynamic and can hardly be specified or envisioned directly. The relationship between the proposed spatial-temporal model and the physics model is also shown, and the approach is applied to model the hourly ozone concentration data in Singapore.

1. Introduction. Ground-level ozone is one of the major air pollutants regulated by many governmental agencies, for example, the U.S. Clean Air Act [United States Environmental Protection Agency (2012)] and World Health Organization air quality guidelines [World Health Organization (2005)]. When ozone is inhaled, it irritates the respiratory system, inflames the lining of lungs and increases the susceptibility to respiratory infections. In particular, the 8-hour concentration exceeding 240 μg/m^3 causes both healthy adults and asthmatics to experience significant reductions in lung function [United States Environmental Protection Agency (2003)]. Several studies have also shown that ozone is correlated with various toxic photochemical oxidants arising from similar sources, including the peroxyacyl nitrates, nitric acid and hydrogen peroxide [Han et al. (2011)]. In recent years, various statistical methods have been used for estimating ozone exposure [Shaddick et al. (2008), Wilson et al. (2014)], evaluating ozone control strategies [Reich et al. (2013)] and assessing health impact [Fuentes (2009)].

Received December 2014; revised January 2016.
Key words and phrases. Spatial-temporal modeling, air quality model, partial differential equation, space–time nonseparable and anisotropic random field.

Both physics-based and statistics-based methods have been proposed for modeling ozone concentration data. Physical models, which are usually computationally intensive, simulate the emission, transport, diffusion and removal processes of ozone. A typical physics-based air quality modeling system consists of a set of submodels: a meteorological model that forecasts meteorological conditions (e.g., the Weather Research and Forecasting Model [Skamarock et al. (2008)]), an emission model that calculates emission processes from a set of known sources (e.g., the Sparse Matrix Operator Kernel Emissions modeling system [Coats (1996)]), and an air quality model that simulates the dispersion and deposition of ozone (e.g., the Community Multi-scale Air Quality modeling system [United States Environmental Protection Agency (1998)]). Since physical air quality models are built on fundamental physics and chemistry, such models are extremely useful in long-term air quality planning; for example, the impact assessment of population growth and emission controls on future air quality. However, physical models are less effective in handling the uncertainties associated with various model inputs including source locations, emission rates and meteorological conditions. For example, in many practical situations, it is impossible to pinpoint the exact locations of all emission sources, and the errors of the estimated emission rates can reach the order of 300% [Christakos and Vyas (1998)]. Also, physical models rarely incorporate the observed pollutant concentration data, and the measured concentrations are often used for bias correction purposes only [Smith et al. (2013)].

Statistical air quality models, on the other hand, have advantages in handling the uncertainties associated with both pollutant concentration and meteorological conditions. Carroll et al. (1997) modeled the ozone distribution by a second-order stationary Gaussian random field, and investigated the relationship between ozone exposure and population density in Texas. Christakos and Vyas (1998) proposed a composite space–time model with a polynomial covariance function to model the ozone distribution over the Eastern United States. Sahu, Gelfand and Holland (2007) proposed an autoregressive space–time model for the daily 8-hour maximum ozone level in the state of Ohio. Stein (2007) investigated the spatial variation of total column ozone through modeling the axially symmetric processes on the sphere using expansions in spherical harmonics. Dou, Le and Zidek (2010) proposed a dynamic linear state-space model as well as a Bayesian spatial predictor for modeling hourly ozone concentrations over the eastern United States. Cameletti et al. (2013) proposed a space–time air quality model by representing a Gaussian field with the Matérn covariance function through stochastic partial differential equations. Often, to keep the model mathematically tractable, assumptions such as stationary, space–time separable and isotropic correlation structure are imposed. These assumptions are appropriate when air quality measurements are aggregated over a relatively large spatial area and time period (e.g., at daily, weekly or monthly levels). By carefully constructing the mathematical structure of the chosen spatial-temporal process (especially the covariance structure), the variation of pollutant

concentration, including the spatial heterogeneity and the temporal nonstationarity, can be approximated. However, for high-frequency air quality measurements such as the hourly level data considered in our paper, assumptions above rarely hold due to the high variability associated with pollutant concentrations and meteorological conditions. For example, hourly ozone concentrations measured at two sites can be correlated in various ways depending on the combinations of hourly wind direction and speed. In fact, there have been prolonged interests in the literature to provide flexible and effective ways to construct covariance functions that take into account such phenomena. Some key results are reported in Cressie and Huang (1999), Gneiting (2002), Banerjee, Carlin and Gelfand (2014), Fuentes et al. (2005) and Reich et al. (2011). To capture the dynamics of the correlation structure of daily ozone data, Reich et al. (2011) investigated an important class of covariate-dependent space–time covariance functions. For air quality data at the hourly level, the space–time covariance structures can be extremely complicated given the interactions between pollutant concentration and dynamic meteorological conditions (especially the wind conditions). It is always challenging to specify appropriate space–time covariance functions that adequately model such complicated covariance structures [Calder (2007), Ghosh et al. (2010)].

To leverage the strengths of both physics models and statistical modeling, different physical-statistical modeling approaches have been proposed in the literature. One important approach is based on the Bayesian Hierarchical Models (BHM) which have been well studied to utilize knowledge about complex systems to tackle environmental problems surrounding uncertainty. For example, Wikle et al. (2001) and Berliner (2003) proposed an important BHM framework to include physical knowledge into statistical analysis. Such a framework consists of three components: a data model, a process model and a parameter model. Wilson et al. (2014) employed this framework and studied the effect of temperature on ozone-related mortality. Malmberg et al. (2008) investigated the interpolation of the carbon monoxide field using a statistical-physical model, and a Bayesian hierarchical model was developed to integrate the deterministic physics model and remote sensing observations. In general, BHM is a powerful statistical tool to fuse the physical model output and observations.

The model developed in this paper is motivated by the fundamental physical principle of pollutant dispersion known as convection–diffusion. Convection-diffusion equations, which are given in the form of Partially Differential Equations (PDE), are the building blocks of almost all physical pollutant dispersion models and many other physical models for natural processes. Such equations describe the transport of air-borne pollutants inside a physical system through three basic processes: emission, convection and diffusion. Analogous to the physics model, our statistical modeling approach expresses the ozone concentration at a particular location and time as the sum of three components: ozone generated at the time, ozone generated in the past but transported to the location of interest, and a noise term. Since our approach does not require running an actual physics model as in BHM,

the total computational cost is significantly reduced. In particular, the emission of ozone is modeled by covariates such as land use, traffic pattern and meteorological conditions, while the convection–diffusion (i.e., the transport of pollutant) is modeled by a convolution approach which takes into account the dynamic wind field. It can be shown that the modeling approach naturally leads to a random field with a space–time nonseparable covariance structure. In spatial statistics, convolution was initially introduced as a novel way to construct the covariance function for complex nonstationary spatial processes [Higdon (2002, 2007)]. In the same spirit, Calder (2007) proposed a dynamic space–time model, in which the pollutant concentration is expressed as the convolution of a latent process. Such a model fits into the general dynamic modeling framework for spatial-temporal data introduced in Stroud, Müller and Sansó (2001). In particular, Brown et al. (2000) advocated the use of convolution to approximate the transport of a pollutant under a constant wind field. Huang and Hsu (2004) also proposed a nonstationary space–time model for the transport of ozone but without considering the interaction between emission and convection.

Recently, Sigrist, Künsch and Stahel (2015) presented a spatio-temporal Gaussian process directly derived from the solution of a Stochastic Partial Differential Equation (SPDE) describing the convection–diffusion process, and proposed to use a Fourier spectral method for efficient computation. However, it is difficult for the spectral method to deal with more complex practical scenarios with a nonstationary random field. In fact, the key results in Sigrist, Künsch and Stahel (2015) are developed under the assumption that the field is stationary. In this paper, we extend the idea of Brown et al. (2000) to account for spatially-varying, dynamic wind and pollution emission fields, and establish the relationship between the proposed convolution approach and the physical pollutant transport process by investigating the analytical solutions of the convection–diffusion equation under special conditions. In Section 3.2.1, we show the connection between our model and Sigrist, Künsch and Stahel (2015) under the special condition of uniform and steady wind in a homogeneous medium. The proposed model provides a general statistical modeling framework to explicitly model both the emission of pollutant and the interactions between emission and convection–diffusion motivated by fundamental physics. While Sigrist, Künsch and Stahel (2015) suggest that the pollutant generation process may be modeled by the temporally white and spatially colored random forcing term of the SPDE, we propose to model the generation of pollutant by a linear function with covariates, which leads to, together with the convolution approach, a linear air quality model in Section 4. We demonstrate the advantages of the proposed statistical air quality model through a case study with real ozone data. The case study also shows that the proposed model provides useful insights about the effects of the complex environmental conditions on ozone concentration.

The remainder of the article is organized as follows. In Section 2, we describe the study area and the data used in this paper. Section 3 provides the general modeling framework and presents some numerical illustrations of the model. In Section 4, the modeling details of the ozone data in Singapore are given. Discussions

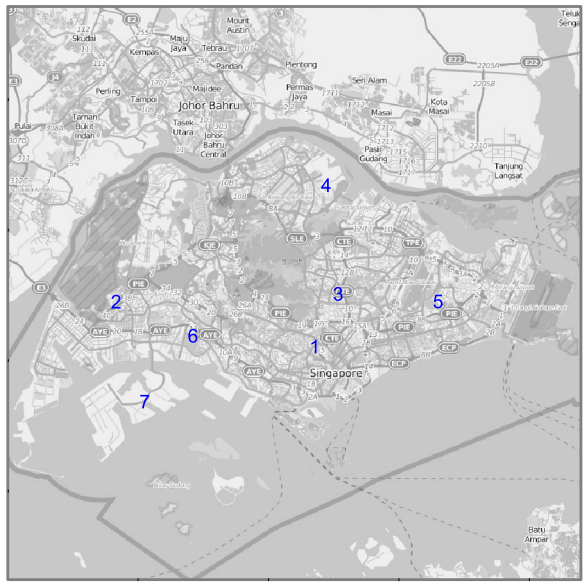


FIG. 1. *Locations of the* 7 *air quality monitoring stations with station IDs indicated by the numbers on the figure* (*the map is downloaded from www.openstreetmap.org*).

on parameter estimation and numerical results are also provided in this section. Finally, we give some concluding remarks in Section 5.

2. The study area and data. This section provides an introduction to the study area and the data provided by the National Environmental Agency (NEA) of Singapore. Section 2.1 describes the observed air quality data, while Section 2.2 describes the data which are used as covariates in our model. Details on how different datasets are preprocessed are also discussed.

2.1. *The air quality data*. As a tropical island country located 137 km north of the equator, Singapore has a total land area of 716.1 km^2 and a population close to 5.4 million. The population density is about 7669/km^2, ranked the third highest in the world. The motor vehicle population had also reached 974,170 as of 2013. The high population and vehicle densities, together with the warm tropical weather (the annual mean temperature ranges from 24° to 32°C), create ideal conditions for ozone to be generated.

Hourly ozone concentrations (in $\mu g/m^3$) are measured at air monitoring stations. In this paper, we use the data observed from seven stations. Locations of these stations are available from the website of NEA and are marked in Figure 1. The hourly ozone concentrations exhibit both temporal and spatial variation, and are subject to a high degree of uncertainty. As an illustration, Figure 2 displays the hourly ozone measurements from July 1, 2013 8AM to August 30, 2013 7AM. Due

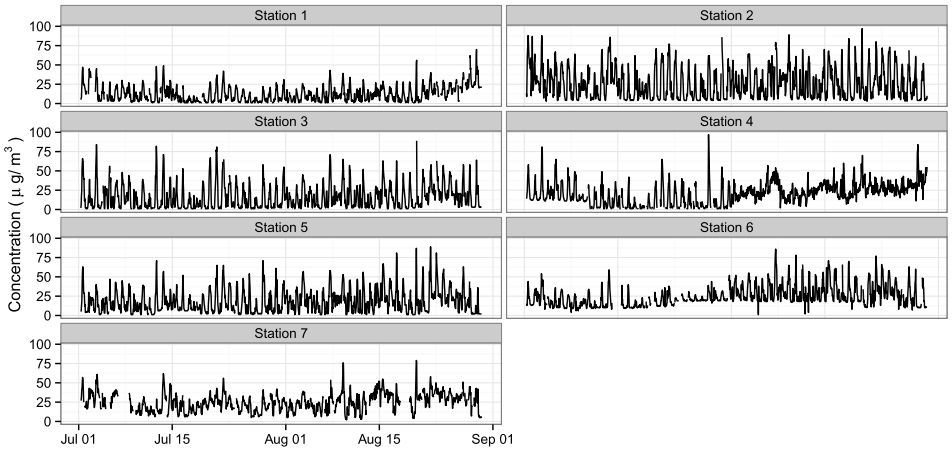
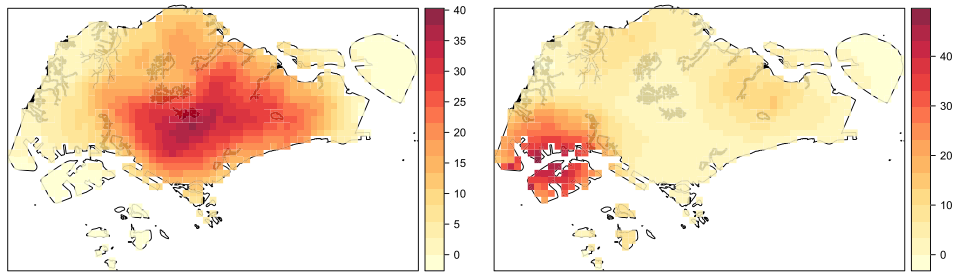


FIG. 2. *Hourly ozone concentration* (*the vertical axis*) *measured at the* 7 *air quality monitoring stations from July* 1, 2013 8*AM to Aug* 30, 2013 7*AM* (*the horizontal axis*). *The station ID is shown on the top of each subplot.*

to the confidentiality agreement, the data is masked by multiplying a constant to all original readings. The diurnal ozone variation is observed. As seen from the data at stations 1 and 2, both the maximum hourly ozone concentration of a day and the amplitude of the diurnal ozone variation vary substantially at different locations. The data from stations 4 and 6 show that the maximum daily ozone concentration and the amplitude of the diurnal ozone variation change over time. The temporal variation is largely due to the fact that ground-level ozone is formed by photochemical reactions in the presence of sunlight and precursor pollutants, such as the oxides of nitrogen (NO_x) and volatile organic compounds (VOCs). The spatial variation, on the other hand, is much more complex because it depends on factors such as land use, traffic and wind.

2.2. *Other datasets and data preprocessing*. In our model, the land use information, traffic data and meteorological data are used as covariates to explain the variation of hourly ozone concentration. Because these datasets are from different sources with a different format, we briefly explain in this subsection how the data is preprocessed.

High-resolution land use information is of great importance in explaining the spatial variation of ozone concentration. The land use type is highly correlated with the VOCs emission, which affects how much ozone can be generated at a given area [Xu, Vizuete and Serre (2012)]. The raw dataset of land use is similar to a map in which Singapore is divided into 110,830 areas with irregular boundaries. Each area is referred to as a spatial polygon in the Geographic Information System (GIS) and the land use type of each area is uniquely specified. In total, there are 32 land use categories. To model the spatial and temporal variation of ozone concentration



(a) (b)

FIG. 3. *Subplot* (a) *is an illustration of the calculated land use index for residential areas. For a grid point in figure* (a), *the figure shows the total residential areas* (*in* km^2) *within a* 5 *km radius of that point. Subplot* (b) *is an illustration of the computed land use index for industrial areas. For a grid point in figure* (b), *the figure shows the total industrial areas* (*in* km^2) *within a* 5 *km radius of that point. For both subplots, the land use index is computed for equal-spaced grid points and the spacing is approximately* 0.7 *km*.

in downtown and suburban areas, industrial and residential areas, as well as urban and nature reserve areas, we regroup the 32 categories into four main land use types: residential, nature reserve, commercial and industrial. Then, for any location $\mathbf{s}$ and land use type p ($p = 1, 2, 3, 4$), we define the land use index, denoted by $l^{(p)}(\mathbf{s})$, for that location as the total spatial area of land use of type p within a 5 km radius of location $\mathbf{s}$. The radius, 5 km, is chosen empirically. In our exploratory analysis, we vary the radius from 1 km to 10 km, and find that the model fitting error is the smallest when the radius is close to 5 km.

As an illustration, Figure 3(a) and 3(b) respectively show the land use index for residential areas (i.e., $\{l^{(1)}; \mathbf{s} \in \mathbb{S}\}$) and industrial areas (i.e., $\{l^{(4)}; \mathbf{s} \in \mathbb{S}\}$) of Singapore, where $\mathbb{S} \subseteq \mathbb{R}^2$ is the spatial domain of interest. In this figure, we create approximately 10,000 grid points within the spatial domain, and compute the land use index for each grid point. The size of a grid is roughly 0.7 km $\times$ 0.7 km.

Vehicles are the main sources of ozone precursor pollutants such as NO_x and VOCs. Hence, the traffic pattern of Singapore is more essential information that explains the spatial and temporal variations of ozone concentrations. Vehicle speed data are available from applications such as Google Map and OneMap. Figure 4 shows the median traffic speed (km/hour) at 8AM for major road links. Similar to how the land use index is defined, for any location $\mathbf{s}$ and traffic characteristic p ($p = 1$ if only traffic speed is used), we define the traffic index, $r^{(p)}(\mathbf{s})$, by the median value of that traffic characteristic p within a 2 km radius of location $\mathbf{s}$ at a given hour of day. Again, the radius is empirically chosen after exploratory analysis. In this paper, the traffic index on each grid point is computed based on the same grid setting described above.

Hourly measurements of meteorological variables, such as temperature, wind speed and direction, are recorded at meteorological stations. Given the size of

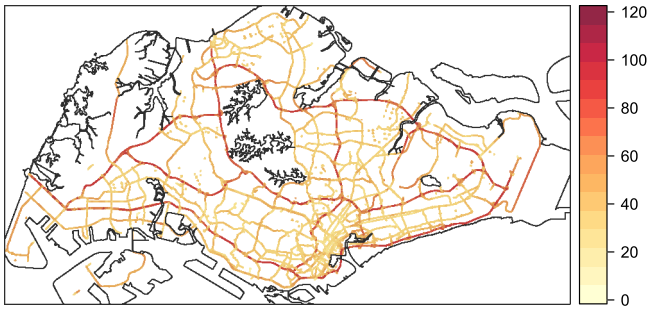


FIG. 4. *An illustration of the median traffic speed on road links at* 8*AM* (km/hour), *where the speed is shown by the color on road links.*

Singapore and its proximity to the equator, the spatial variation of temperature is small, and the temperature field for the entire island can be well constructed from station observations using statistical techniques such as Kriging. The wind field, however, could vary over space and time dramatically because of the complex dynamics due to the dense urban canopy and the small-scale coastal influences. Instead of generating the wind field using station observations using spatial-temporal statistics [Haslett and Raftery (1989), Reich and Fuentes (2007), Stein (2009), Wikle et al. (2001)], we obtain the wind field from the Weather Research and Forecasting (WRF) model, which is one of the mainstream numerical weather prediction models [Skamarock et al. (2008)]. Our WRF model incorporates both the high-resolution land use data and the sea surface temperature data respectively obtained from the NASA Moderate Resolution Imaging Spectroradiometer project and the Jet Propulsion Laboratory. The model generates a 1 km by 1 km high-resolution surface wind field with 30-min temporal resolution, which is one of the key inputs to the air quality model proposed in this paper. Figure 5 shows both the observed wind vectors at meteorological monitoring stations and the surface wind field obtained from our own WRF model. It is seen from Figure 5 that the WRF model is particularly useful in capturing the small-scale dynamics of the wind field and the phenomena of wind convergence.

Table 1 provides a summary of the datasets described above.

3. The model. In this section we first present the statistical spatio-temporal model in Section 3.1. Then, we show in Section 3.2 the connection between the statistical model and a physical dispersion model which is widely used.

3.1. *The spatio-temporal model.* Our goal is to build a statistical model motivated by the fundamental ideas behind physical dispersion models. A typical physical dispersion model consists of four major components, including the emission, convection, diffusion and decay of pollutants. For example, consider a widely used

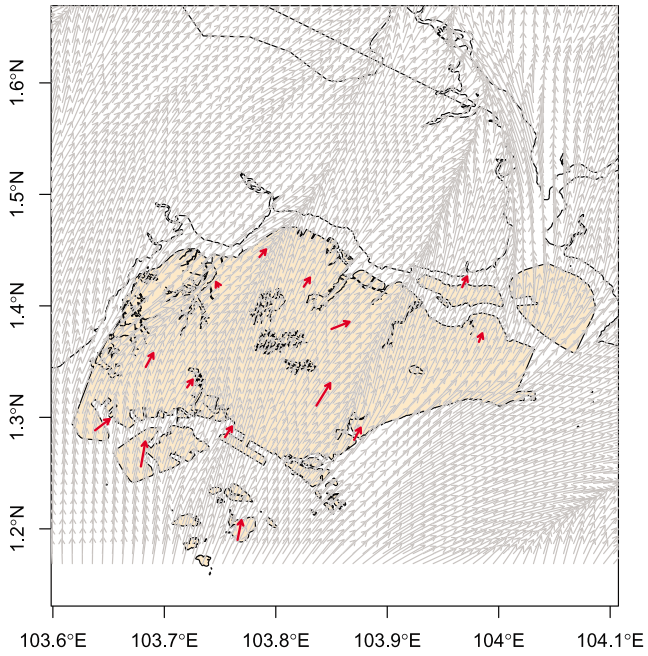


FIG. 5. *A snapshot of the observed wind vectors and the wind field generated by the WRF model, where red and grey arrows represent the observed wind vectors and the wind vectors generated by the WRF model, respectively.*

scalar transport equation for the dispersion of air-borne pollutants:

$$\frac{\partial \varphi(\mathbf{s},t)}{\partial t} = Q(\mathbf{s},t) - \nabla \cdot \left[\mathbf{v}(\mathbf{s},t)\varphi(\mathbf{s},t)\right] + \nabla \cdot \left[\mathbf{K}(\mathbf{s},t) \cdot \nabla \varphi(\mathbf{s},t)\right] - \frac{1}{\tau}\varphi(\mathbf{s},t), \tag{3.1}$$

TABLE 1
A summary of the settings of different datasets

Data	Raw data	Processed data
Air quality	Hourly observations at 7 stations	N.A.
Land use	Available on 110,830 spatial polygons	Land use index computed at equally spaced grids for 4 major land use categories
Traffic	Speed on road links	Traffic index computed at equally spaced grids
Weather	WRF output with 1 km spatial resolution and 30-min temporal resolution	N.A.

where φ is the pollutant concentration at some location $\mathbf{s}$ and time t, Q is the pollution emission rate from a pollution source, $\mathbf{v}$ is the wind vector, $\mathbf{K}$ is a second-order tensor of the eddy diffusivity, and τ is the relaxation timescale of pollution removal. Note that the second and the third terms on the right-hand side of (3.1) respectively represent the pollutant convection and diffusion processes. In physics, convection is used to describe the transport mechanism of a pollutant due to the flow of air (i.e., wind), while the diffusion process describes the mixing of the pollutant by local turbulent flow. The last term in (3.1) models either the physical (e.g., deposition on a surface or absorption in water) or chemical (e.g., photochemical decomposition of ozone) removal process of the airborne pollution.

Let $\{Y(\mathbf{s},t);\mathbf{s}\in\mathbb{R}^2, t\in\mathbb{N}^+\}$ be a spatial-temporal random field that represents the pollution concentration over continuous space and discrete time. Following the idea of (3.1), we model the pollutant concentration at location $\mathbf{s}$ and time t, $Y(\mathbf{s},t)$, by

$$Y(\mathbf{s},t) = g(\mathbf{s},t) + Z(\mathbf{s},t) + \varepsilon(\mathbf{s},t). \tag{3.2}$$

The first term $g(\mathbf{s},t)$ is the pollutant emission over the time interval $[t-\Delta,t]$ at location $\mathbf{s}$. Depending on the pollutant of interest, $g(\mathbf{s},t)$ is determined by various factors such as land use, traffic density, meteorological conditions and so on. The second term $Z(\mathbf{s},t)$ corresponds to the physical convection–diffusion and decay process, and is used to model the pollutant generated in the past but transported to location $\mathbf{s}$ at time t. The last term ε corresponds to an error process that accounts for the noise over the time interval $[t-\Delta,t]$. For example, the noise occurs during the pollutant emission and transport processes. We further assume that ε is a white-in-time isotropic Gaussian random field with spatial covariance function $c(\cdot;\boldsymbol{\theta})$. Here, general spatial covariance functions can be used, such as Exponential, Matern, Gaussian, etc.

In (3.2), the stochastic transport process of pollutant, $Z(\mathbf{s},t)$, largely determines the spatial and temporal correlation of the pollutant concentration data. We use a convolution model to approximate such a process:

$$\begin{aligned} Z(\mathbf{s},t) &= \zeta_\Delta\{\omega_\Delta * Y(\mathbf{s},t-\Delta)\} \\ &= \zeta_\Delta\left\{\int_{\mathbb{R}^2}\omega_\Delta(\mathbf{x})Y(\mathbf{s}-\mathbf{x},t-\Delta)\,d\mathbf{x}\right\} \\ &= \zeta_\Delta\{\omega_\Delta * Z(\mathbf{s},t-\Delta) + \omega_\Delta * g(\mathbf{s},t-\Delta) + \omega_\Delta * \varepsilon(\mathbf{s},t-\Delta)\}, \end{aligned} \tag{3.3}$$

where $*$ denotes the convolution operation and ζ_Δ is a scaling (decay) factor. As shown in Section 3.2, it is possible to establish the link between the statistical model and the physical dispersion model (3.1) if we respectively define the scaling factor ζ_Δ and the convolution kernel ω_Δ as

$$\zeta_\Delta = \exp(-\lambda\Delta), \tag{3.4}$$

$$\begin{aligned}\omega_\Delta(\mathbf{x}) &= \frac{1}{2\pi|\boldsymbol{\Sigma}_\Delta|^{1/2}}\exp\left\{-\frac{(\mathbf{x}-\boldsymbol{\mu}_\Delta)^\top\boldsymbol{\Sigma}_\Delta^{-1}(\mathbf{x}-\boldsymbol{\mu}_\Delta)}{2}\right\}\\ &= \phi(\mathbf{x};\boldsymbol{\mu}_\Delta,\boldsymbol{\Sigma}_\Delta),\end{aligned} \tag{3.5}$$

where ϕ denotes the probability density function of a bivariate Gaussian distribution, and $\boldsymbol{\Sigma}_\Delta$ is the covariance matrix of the Gaussian convolution kernel. Because the transport of pollutant is affected by wind speed and direction, we let both $\boldsymbol{\mu}_\Delta$ and $\boldsymbol{\Sigma}_\Delta$ depend on the wind vector, $\mathbf{v}(\mathbf{s},t)$, at location $\mathbf{s}$ and time t as follows:

$$\boldsymbol{\mu}_\Delta(\mathbf{s},t) = \mathbf{v}(\mathbf{s},t)\Delta \tag{3.6}$$

and

$$\boldsymbol{\Sigma}_\Delta\big(v(\mathbf{s},t)\big) = \mathbf{R}^{-1}\begin{pmatrix}\rho_1\Delta & 0\\ 0 & \rho_2\Delta\end{pmatrix}(\mathbf{R}^\top)^{-1}, \tag{3.7}$$

$$\mathbf{R} = \begin{pmatrix}\cos(\alpha_\mathbf{v}) & -\sin(\alpha_\mathbf{v})\\ \sin(\alpha_\mathbf{v}) & \cos(\alpha_\mathbf{v})\end{pmatrix}. \tag{3.8}$$

Here, $\alpha_\mathbf{v}\in[0,2\pi)$ is the counterclockwise rotation angle of the wind vector from the horizontal axis, and $\mathbf{R}$ is the rotation matrix. The parameters, ρ_1 and ρ_2, respectively control the standard deviations of the convolution kernel ω_Δ in directions which are parallel and perpendicular to the wind direction. In Section 3.2, we show that the parameters ρ_1 and ρ_2 can be interpreted in terms of the eddy diffusivities in the physical dispersion model (3.1).

Further, discretizing Δ in (3.3) into a number of n small time intervals ($\delta = \Delta/n$) gives

$$\begin{aligned}Z(\mathbf{s},t) &= \sum_{i=1}^{n}\exp(-i\lambda\delta)\omega_\delta^{*i} * g(\mathbf{s},t-i\delta)\\ &\quad+\exp(-n\lambda\delta)\omega_\delta^{*n} * Z(\mathbf{s},t-\Delta)\\ &\quad+\sum_{i=1}^{n}\exp(-i\lambda\delta)\omega_\delta^{*i} * \varepsilon(\mathbf{s},t-i\delta),\end{aligned} \tag{3.9}$$

with $*n$ denoting the n-fold convolution operation. Substituting (3.9) into (3.2), we have

$$\begin{aligned}Y(\mathbf{s},t) &= g(\mathbf{s},t)+\sum_{i=1}^{n}\exp(-i\lambda\delta)\omega_\delta^{*i} * g(\mathbf{s},t-i\delta)\\ &\quad+\exp(-\lambda\Delta)\omega_\Delta * Z(\mathbf{s},t-\delta)\\ &\quad+\sum_{i=1}^{n}\exp(-i\lambda\delta)\omega_\delta^{*i} * \varepsilon(\mathbf{s},t-i\delta)+\varepsilon(\mathbf{s},t).\end{aligned} \tag{3.10}$$

Since the convolution of Gaussians is still a Gaussian, we rewrite (3.10) to obtain the expression of the pollutant concentration over continuous space and discrete time:

$$
\begin{aligned}
Y(\mathbf{s},t) &= g(\mathbf{s},t) \\
&\quad + \sum_{i=1}^{n}\{\exp(-i\lambda\delta)\phi(\mathbf{s}) * g(\mathbf{s},t-i\delta)\} \\
&\quad + \exp(-n\lambda\delta)\phi(\mathbf{s}) * Z(\mathbf{s},t-\Delta) \\
&\quad + \sum_{i=1}^{n}\{\exp(-i\lambda\delta)\phi(\mathbf{s}) * \varepsilon(\mathbf{s},t-i\delta)\} + \varepsilon(\mathbf{s},t),
\end{aligned} \tag{3.11}
$$

where

$$
\phi(\mathbf{s}) = \phi\Bigg(\mathbf{s}; \sum_{j=0}^{i-1}\boldsymbol{\mu}_{\delta}\big(\mathbf{v}(\mathbf{s},t-j\delta)\big), \sum_{j=0}^{i-1}\boldsymbol{\Sigma}_{\delta}\big(\mathbf{v}(\mathbf{s},t-j\delta)\big)\Bigg). \tag{3.12}
$$

Let

$$
\Psi_{i,t}(\mathbf{s}) = \begin{cases} \exp(-i\lambda\delta)\phi(\mathbf{s}), & i>0, \\ 1, & i=0, \end{cases} \tag{3.13}
$$

then the covariance between the pollutant concentration at location $\mathbf{s}_1$ and time t_1 and the concentration at location $\mathbf{s}_2$ and time t_2 (assuming $t_1 \le t_2$ and $t_2 - t_1 = \jmath\delta$ for some $\jmath = 0, 1, 2, \ldots$) can be obtained from (3.11) as follows (the derivation is given in the Appendix):

$$
\begin{aligned}
&\operatorname{cov}\big(Y(\mathbf{s}_1,t_1), Y(\mathbf{s}_2,t_2)\big) \\
&\quad = I_{\{\jmath=0\}}c(\mathbf{d}) \\
&\qquad + \sum_{i=0}^{n}(\tilde{\Psi}_{i,t_1} * \Psi_{\jmath+i,t_2} * c)(\mathbf{d}),
\end{aligned} \tag{3.14}
$$

where $\mathbf{d} = \mathbf{s}_2 - \mathbf{s}_1$, $\tilde{\Psi}(\mathbf{s}) = \Psi(-\mathbf{s})$, and the function $I_{\{\jmath=0\}} = 1$ if $\jmath = 0$, otherwise, $I_{\{\jmath=0\}} = 0$. Note that, since the random field $\varepsilon(\mathbf{s},t)$ is isotropic, $c(\mathbf{d}) = c(\|\mathbf{d}\|)$ with $\|\mathbf{d}\|$ representing the distance between $\mathbf{s}_1$ and $\mathbf{s}_2$. It is seen from (3.14) that the covariance, $\operatorname{cov}(Y(\mathbf{s}_1,t_1), Y(\mathbf{s}_2,t_2))$, is determined by not only the separation of time and space, but also the dynamic wind conditions between times t_1 and t_2. This is due to the fact that wind conditions determine the transport of pollutant through convolution in our model. Since both wind speed and direction change over time, the resultant covariance structure is no longer isotropic and space–time separable. In a special case when the wind field is uniform, that is, the wind speed and direction are the same everywhere at

a given time, Ψ does not depend on location $\mathbf{s}$ any longer. Hence, (3.14) becomes

$$\begin{aligned}(3.15)\qquad &\operatorname{cov}\big(Y(\mathbf{s}_1,t_1),Y(\mathbf{s}_2,t_2)\big)\\ &\quad= I_{\{J=0\}}c(\mathbf{d})+\sum_{i=0}^{n}(\Psi_{i,t_1}*\Psi_{J+i,t_2}*c)(\mathbf{d}).\end{aligned}$$

As an extension, we could obtain the random field, $\{Y(\mathbf{s},t);\mathbf{s}\in\mathbb{R}^2,t\in\mathbb{R}\}$, which is continuous in both space and time from equation (3.11), by letting $n\to\infty$ and $\delta\to 0$. The expression is given by

$$\begin{aligned}(3.16)\qquad Y(\mathbf{s},t)&=\int_0^\infty \exp(-\lambda u)\phi(\mathbf{x};\bar{\boldsymbol{\mu}}_u,\bar{\boldsymbol{\Sigma}}_u)*g(\mathbf{s},t-u)\,du\\ &\quad+\int_0^\infty \exp(-\lambda u)\phi(\mathbf{x};\bar{\boldsymbol{\mu}}_u,\bar{\boldsymbol{\Sigma}}_u)*dB(\mathbf{x},t-u)\,du,\end{aligned}$$

where $\bar{\boldsymbol{\mu}}_u=\int_0^u \mathbf{R}(\mathbf{s},t-x)\mathbf{v}(\mathbf{s},t-x)\,dx$, $\bar{\boldsymbol{\Sigma}}_u=\int_0^u \mathbf{R}(\mathbf{s},t-x)\boldsymbol{\Sigma}(\mathbf{s},t-x)\,dx$, and $B(\mathbf{s},t)$ is a spatially correlated Brownian motion. In reality, since air quality measurements are sampled at discrete times, we only focus on the discrete case in this paper.

3.2. *The connection between the statistical model and physics.*

3.2.1. *The transport of pollutant.* It is possible to show that the convolution model above is an approximation to the physical convection–diffusion process by establishing the relationship between equations (3.3), (3.5) and the scalar transport equation (3.1). Note that, under a uniform and steady wind field in a homogeneous medium without any source, the scalar transport equation (3.1) can be simplified as

$$(3.17)\qquad \frac{\partial\varphi(\mathbf{s},t)}{\partial t}=-\mathbf{v}\cdot\nabla\varphi(\mathbf{s},t)+K\nabla^2\varphi(\mathbf{s},t)-\frac{1}{\tau}\varphi(\mathbf{s},t).$$

Then, applying the Fourier transform, equation (3.17) becomes

$$(3.18)\qquad \frac{d\tilde{\varphi}_\eta(t)}{dt}=-i(\boldsymbol{\eta}\cdot\mathbf{v})\tilde{\varphi}_\eta(t)-\eta^2K\tilde{\varphi}_\eta(t)-\frac{1}{\tau}\tilde{\varphi}_\eta(t),$$

in which η is the wave number and $\tilde{\varphi}_\eta$ is the Fourier coefficient of φ. It is not difficult to show that the solution of equation (3.18) is given by

$$(3.19)\qquad \tilde{\varphi}_\eta(t+\Delta)=\exp\left\{-\frac{1}{\tau}\Delta-\big[i(\boldsymbol{\eta}\cdot\mathbf{v})+K\eta^2\big]\Delta\right\}\tilde{\varphi}_\eta(t).$$

Finally, the backward Fourier transform leads to

$$(3.20)\qquad \varphi(\mathbf{s},t)=\exp\left\{-\frac{1}{\tau}\Delta\right\}\int_{\mathbb{R}^2}\Omega_\Delta(\mathbf{x})\varphi(\mathbf{s}-\mathbf{x},t-\Delta)\,d\mathbf{x},$$

in which the kernel $\Omega_\Delta(\mathbf{x})$ is

$$\Omega_\Delta(\mathbf{x}) = \frac{1}{2\pi|\mathbf{\Sigma}_\Delta|^{1/2}} \exp\left\{-\frac{(\mathbf{x}-\mathbf{v}\Delta)^\top \mathbf{\Sigma}_\Delta^{-1}(\mathbf{x}-\mathbf{v}\Delta)}{2}\right\}, \tag{3.21}$$

with the covariance matrix given by

$$\mathbf{\Sigma}_\Delta = \begin{pmatrix} 2K\Delta & 0 \\ 0 & 2K\Delta \end{pmatrix}. \tag{3.22}$$

The covariance matrix (3.22) is known as the eddy diffusivity tensor in physics. The comparison between (3.3) and (3.20) justifies why the decay term and the convolution kernel can be defined in ways shown in equations (3.4) and (3.5) with $\lambda = \tau^{-1}$. Equation (3.22) also suggests that

$$\rho_1 = 2K_\parallel(s,t), \qquad \rho_2 = 2K_\perp(s,t), \tag{3.23}$$

where $K_\parallel$ and $K_\perp$ are the eddy diffusivities respectively parallel and perpendicular to the wind direction, and can be obtained from the wind field through a physics parameterization of atmospheric turbulent mixing. In a special case when an isotropic diffusion model is used (i.e., $K = K_\parallel = K_\perp$), the covariance matrix reduces to equation (3.22). The connection shown above shows that the proposed statistical model is well motivated when the wind direction and speed do not change dramatically within a certain spatial region.

3.2.2. *The emission of pollutant.* To see the relationship between $g(\mathbf{s},t)$ in (3.2) and the source term $Q(\mathbf{s},t)$ in (3.1), we now consider the limit situation with zero diffusion and no decay, that is, $\mathbf{K} \to 0$ and $\tau \to \infty$. Then, equation (3.1) reduces to

$$\frac{\partial\varphi(\mathbf{s},t)}{\partial t} = -\nabla\cdot\big[\mathbf{v}(\mathbf{s},t)\varphi(\mathbf{s},t)\big] + Q(\mathbf{s},t). \tag{3.24}$$

At the same time, it is obvious that $\zeta_\Delta = 1$ and the convolution kernel ω_Δ in (3.5) in this limit becomes a Dirac delta function,

$$\omega_\Delta(\mathbf{x}) = \delta(\mathbf{x}-\mathbf{v}\Delta).$$

Hence, substituting (3.3) into (3.2) and omitting the error process yields

$$\begin{aligned} Y(\mathbf{s},t) &= \big\{\omega_\Delta * Y(\mathbf{s},t-\Delta)\big\} + g(\mathbf{s},t) \\ &= Y(\mathbf{s}-\mathbf{v}\Delta, t-\Delta) + g(\mathbf{s},t). \end{aligned} \tag{3.25}$$

Since $Y(\cdot,t)$ in the statistical model corresponds to the pollution concentration φ at t in the physical model, equation (3.25) can be rewritten as

$$\varphi(\mathbf{s},t) = \varphi(\mathbf{s}-\mathbf{v}\Delta, t-\Delta) + g(\mathbf{s},t). \tag{3.26}$$

Applying the Taylor expansion and assuming the mass conservation [i.e., $\nabla \cdot \mathbf{v}(\mathbf{s}, t) = 0$], equation (3.26) yields

$$\Delta\left(\frac{\partial \varphi(\mathbf{s}, t)}{\partial t} + \nabla \cdot \left[\mathbf{v}(\mathbf{s}, t)\varphi(\mathbf{s}, t)\right] - \left[g(\mathbf{s}, t)/\Delta\right]\right) = O(\Delta^2). \tag{3.27}$$

By comparing (3.24) and (3.27), it is clear that the convolution model (3.25) is a first-order approximation to the scalar transport equation (3.24), that is, (3.27) converges to (3.24) as $\Delta \to 0$. Hence, the source term, $g(\mathbf{s}, t)$, in the convolution model (3.2), is related to the pollution emission source in the physics model (3.1) through the following relationship: $g(\mathbf{s}, t) = Q(\mathbf{s}, t)\Delta$.

3.3. *Illustration.* Before a case study is presented in Section 4, two illustrative examples are presented here to help the readers develop a better understanding of our model through the visualization of model output. In both examples, we compute the pollutant concentration based on equation (3.11) with arbitrarily chosen values of model parameters, for illustrative purposes.

Figure 6 shows the computed pollutant concentration given a fixed-location point emission source. We see that the output of the proposed statistical model mimics the pollutant dispersion process. The direction of the pollutant dispersion changes as the wind direction gradually changes from west to south. Similarly, Figure 7 shows a hypothetical pollutant dispersion from a moving point emission source under the same wind field. The solid line in Figure 7 represents a major highway in Singapore.

The two examples illustrate how the proposed statistical air quality model mimics the physical phenomenon of pollutant dispersion. However, unlike a pure physical model with model parameters directly specified, we describe in the next section a procedure for estimating the unknown model parameters based on data. Animations based on the two illustrative examples are also available from the online supplementary materials [Liu et al. (2016)].

4. Application: The modeling of hourly ozone data.

4.1. *The modeling.* A real case study is presented to demonstrate how the modeling framework is applied to model the hourly ozone concentration data in Singapore. Because the model is motivated by the basic physics behind the pollution process, the case study also shows that the model provides useful insights into the effects of the complex environmental conditions on ozone concentration.

First, based on the basic physics and chemistry of how ground-level ozone is generated in an urban area, we model the emission of ozone, $g(\mathbf{s}, t)$, by a linear function of land use types and weather conditions. In particular,

$$g(\mathbf{s}, t) = \mathbf{l}_0(\mathbf{s}, t)\mathbf{A}_{\tau(t)}^{\top} + \mathbf{r}_0(\mathbf{s}, t)\mathbf{B}_{\tau(t)}^{\top} + \mathbf{m}_0(\mathbf{s}, t)\mathbf{C}_{\tau(t)}^{\top}, \tag{4.1}$$

where $\tau(t) : t \mapsto \tau \in [1, \ldots, 24]$ returns the hour of a calendar time t at which the measurements are taken. The covariates in (4.1) are described as follows:

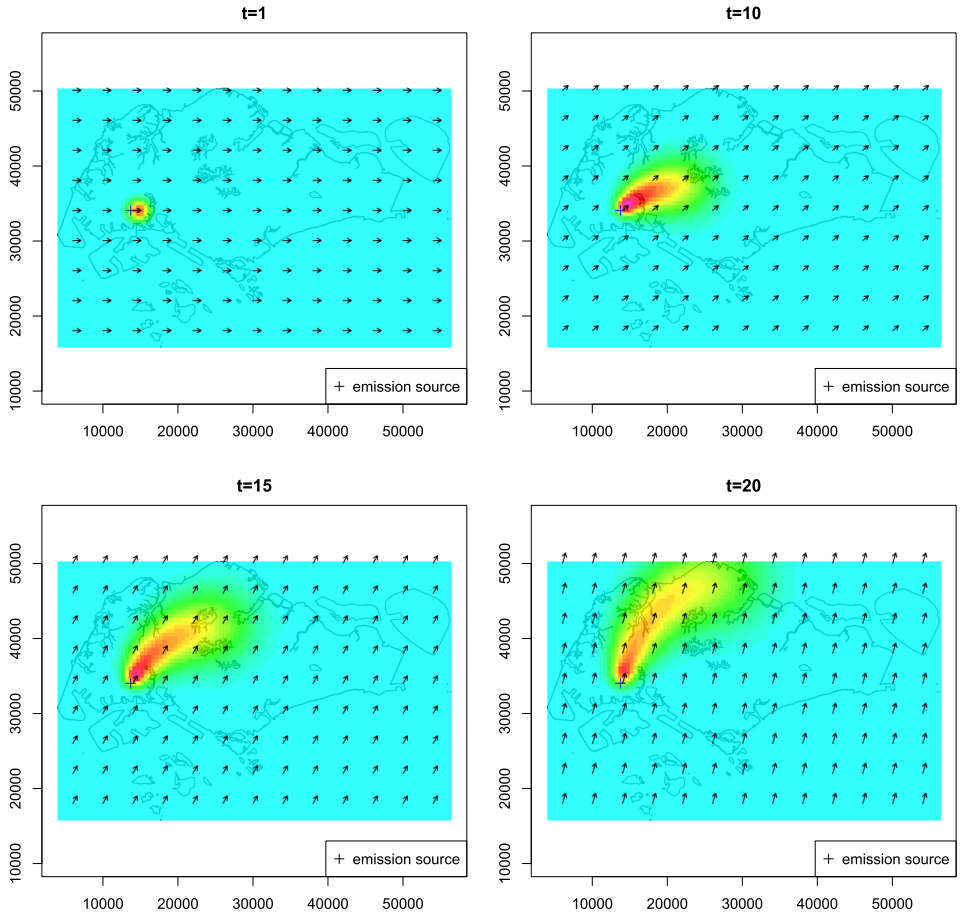


FIG. 6. *Illustration of the computed pollutant dispersion from a point source at times* $t = 1, 10, 15, 20$. *The wind field is shown by the arrows.*

- The vector $\mathbf{l}_0(\mathbf{s}, t) = (l^{(1)}(\mathbf{s}, t), \ldots, l^{(k_1)}(\mathbf{s}, t))$ represents the land use index, as defined in Section 2, at location $\mathbf{s}$ for the four land use types (i.e., $k_1 = 4$). Then, $\mathbf{A}_{\tau(t)} = (a^{(1)}_{\tau(t)}, \ldots, a^{(k_1)}_{\tau(t)})$ are the effects of different land use types on ozone emission at different hours of day.
- The vector $\mathbf{r}_0(\mathbf{s}, t) = (r^{(1)}(\mathbf{s}, t), \ldots, r^{(k_2)}(\mathbf{s}, t))$ represents the traffic characteristics at location $\mathbf{s}$ and time t as defined in Section 2, and $\mathbf{B}_{\tau(t)} = (b^{(1)}_{\tau(t)}, \ldots, b^{(k_2)}_{\tau(t)})$ are the effects of each traffic characteristic on ozone concentrations at different hours of day. In the case study, only the traffic speed is used and $k_2 = 1$.
- The vector $\mathbf{m}_0(\mathbf{s}, t) = (m^{(1)}(\mathbf{s}, t), \ldots, m^{(k_3)}(\mathbf{s}, t))$ represents a number of k_3 types of meteorological conditions at location $\mathbf{s}$ and time t, and $\mathbf{C}_{\tau(t)} = (c^{(1)}_{\tau(t)}, \ldots, c^{(q)}_{\tau(t)})$ are the effects of each type of meteorological condition on ozone concentrations at different hours of day. We only retain temperature,

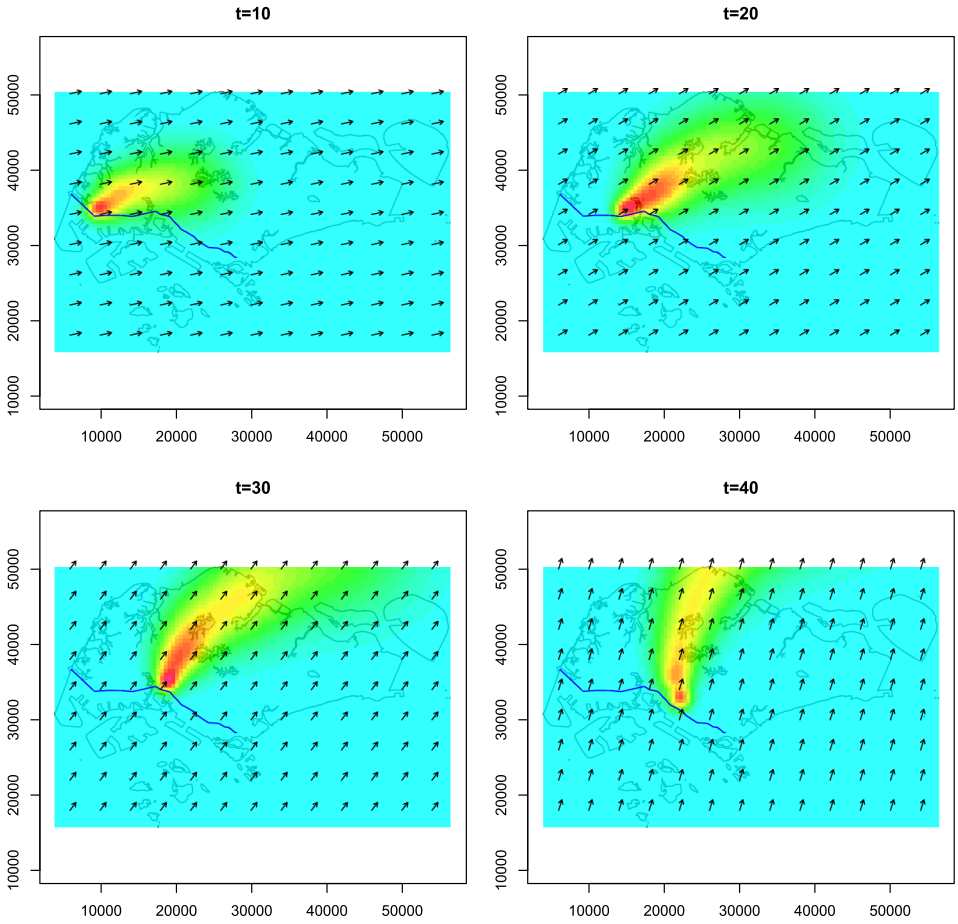


FIG. 7. *Illustration of the computed pollutant dispersion from a moving source along a trajectory at times* $t = 10, 20, 30, 40$. *The wind field is shown by the arrows.*

which turns out to be the most important meteorological condition for the generation of ozone and, hence, $k_3 = 1$. Although wind affects the transport of the pollutant, it does not affect the emission of ozone.

It is worth mentioning that the actual ozone generation process is a complicated photo-chemical process which mainly involves solar ultraviolet radiation, volatile organic compounds (VOCs) and oxides of nitrogen (NO_x). Hence, there might be a potential bias or lack of fitting if the covariates in (4.1) are not properly chosen. In (4.1), the emission term $g(\mathbf{s}, t)$ is modeled by a pure statistical approach with all covariates carefully chosen: temperature serves as a proxy of solar radiation when the solar radiation data is not available, while the land use and traffic data are mainly used to capture the emission of VOCs and NO_x.

Substituting (4.1) into (3.11), we obtain the expected ozone concentration at location $\mathbf{s}$ and time t:

$$\begin{aligned}
\mathbb{E}\big(Y(\mathbf{s},t)\big) &\approx g(\mathbf{s},t) + \sum_{i=1}^{n}\big\{\Psi_{i,t}(\mathbf{s}) * \big(\mathbf{l}_0(\mathbf{s},t-i\delta)\mathbf{A}^{\top}_{(\tau(t-i\delta))}\big)\big\} \\
&\quad + \sum_{i=1}^{n}\big\{\Psi_{i,t}(\mathbf{s}) * \big(\mathbf{r}_0(\mathbf{s},t-i\delta)\mathbf{B}^{\top}_{(\tau(t-i\delta))}\big)\big\} \\
&\quad + \sum_{i=1}^{n}\big\{\Psi_{i,t}(\mathbf{s}) * \big(\mathbf{m}_0(\mathbf{s},t-i\delta)\mathbf{C}^{\top}_{(\tau(t-i\delta))}\big)\big\} \\
&= \mathbf{l}_0(s,t)\mathbf{A}^{\top}_{\tau(t)} + \mathbf{r}_0(\mathbf{s},t)\mathbf{B}^{\top}_{\tau(t)} + \mathbf{m}_0(\mathbf{s},t)\mathbf{C}^{\top}_{\tau(t)} \\
&\quad + \sum_{i=1}^{n}\left\{\sum_{p=1}^{k_1} a^{(p)}_{\tau(t-i\delta)}\big[\Psi_{i,t}(\mathbf{s}) * l^{(p)}(\mathbf{s},t)\big]\right\} \\
&\quad + \sum_{i=1}^{n}\left\{\sum_{p=1}^{k_2} b^{(p)}_{\tau(t-i\delta)}\big[\Psi_{i,t}(\mathbf{s}) * r^{(p)}(\mathbf{s},t)\big]\right\} \\
&\quad + \sum_{i=1}^{n}\left\{\sum_{p=1}^{k_3} c^{(p)}_{\tau(t-i\delta)}\big[\Psi_{i,t}(\mathbf{s}) * m^{(p)}(\mathbf{s},t)\big]\right\}.
\end{aligned} \tag{4.2}$$

Since the hourly-level ozone concentration data are considered in this case study, we let $\delta = 1$, and $t - i\delta$ refers to the time which is a number of i hours ahead of the time t.

In fact, if we let

$$\begin{aligned}
\mathbf{l}_i(\mathbf{s},t) &= \big(\Psi_{i,t}(\mathbf{s}) * l^{(1)}(\mathbf{s},t), \ldots, \Psi_{i,t}(\mathbf{s}) * l^{(k_1)}(\mathbf{s},t)\big), \\
\mathbf{r}_i(\mathbf{s},t) &= \big(\Psi_{i,t}(\mathbf{s}) * r^{(1)}(\mathbf{s},t), \ldots, \Psi_{i,t}(\mathbf{s}) * r^{(k_2)}(\mathbf{s},t)\big), \\
\mathbf{m}_i(\mathbf{s},t) &= \big(\Psi_{i,t}(\mathbf{s}) * m^{(1)}(\mathbf{s},t), \ldots, \Psi_{i,t}(\mathbf{s}) * m^{(k_3)}(\mathbf{s},t)\big),
\end{aligned}$$

for $i = 0, \ldots, n$, the expression of (4.2) is further simplified to a linear form as follows,

$$\mathbb{E}\big(Y(\mathbf{s},t)\big) = \mathbf{x}(\mathbf{s},t)\boldsymbol{\beta}, \tag{4.3}$$

where $\boldsymbol{\beta} = (\mathbf{A}_1, \mathbf{B}_1, \mathbf{C}_1, \ldots, \mathbf{A}_{24}, \mathbf{B}_{24}, \mathbf{C}_{24})^{\top}$ is a $24k \times 1$ column vector ($k = 6$), and $\mathbf{x}(\mathbf{s},t)$ is a $1 \times 24k$ row vector given by

$$\mathbf{x}(\mathbf{s},t) = \big(\mathbf{x}^{(1)}(s,t), \ldots, \mathbf{x}^{(24)}(s,t)\big),$$

where

$$\mathbf{x}^{(j)}(\mathbf{s},t) = \sum_{i \in I_j} \mathbf{x}_i(\mathbf{s},t) \tag{4.4}$$

with $I_j = \{i; \tau(t - i\delta) = j\}$ and $\mathbf{x}_i(\mathbf{s}, t) = (\mathbf{l}_i(\mathbf{s}, t), \mathbf{r}_i(\mathbf{s}, t), \mathbf{m}_i(\mathbf{s}, t))$ for $i = 0, \ldots, n$ and $j = 1, \ldots, 24$. Equation (4.3) implies that the covariates in (4.1) are first transformed by convolution under the wind field, and the expected ozone concentration is given by a linear combination of the transformed covariates. Note that each $\mathbf{x}_i(\mathbf{s}, t)$ contains the following: four transformed land use index, the transformed traffic speed and the transformed temperature.

Finally, supposing that ozone concentrations are monitored at N_s stations over N_t hours, we obtain from (4.3) a linear model as follows:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \tag{4.5}$$

where $\mathbf{Y} = (Y(\mathbf{s}_1, t_1), \ldots, Y(\mathbf{s}_{N_s}, t_1), \ldots, Y(\mathbf{s}_{N_s}, t_{N_t}))^\top$ is a column vector of length $N_s \times N_t$ and $\mathbf{X} = (\mathbf{x}(\mathbf{s}_1, t_1), \mathbf{x}(\mathbf{s}_2, t_1), \ldots, \mathbf{x}(\mathbf{s}_{N_s}, t_1), \ldots, \mathbf{x}(\mathbf{s}_{N_s}, t_{N_t}))^\top$ is a $N_s N_t \times (kn + k)$ matrix, and $\mathbf{e}$ is the error process with mean zero and variance $\boldsymbol{\Sigma}_\mathbf{Y}$ given by (3.14).

4.2. *Parameter estimation.* The model in (4.5) contains the following parameters: the decay parameter λ, the eddy diffusivity $\mathbf{K}$, the parameter $\boldsymbol{\theta}$ in the spatial covariance function $c(\cdot; \boldsymbol{\theta})$, and the parameters $\boldsymbol{\beta}$. The physical parameters, including λ and $\mathbf{K}$, can be determined according to their physical meanings described in Section 3. In particular, the eddy diffusivity $\mathbf{K}$ represents the mixing of pollutants due to turbulent wind conditions. Hence, we choose a widely used parameterization [Byun and Schere (2006)]:

$$\mathbf{K} = 0.28(\delta_1\delta_2)\sqrt{\left(\frac{\partial v_1}{\partial s_1} - \frac{\partial v_2}{\partial s_2}\right)^2 + \left(\frac{\partial v_1}{\partial s_2} + \frac{\partial v_2}{\partial s_1}\right)^2},$$

in which v_1 and v_2 are the horizontal and vertical components of wind [$\mathbf{v} = (v_1, v_2)$]; s_1 and s_2 are x- and y-coordinates of a location $\mathbf{s}$; and δ_1 and δ_2 are the computational resolution of the wind field. To model the decay of ozone, we set the relaxation timescale $\tau = 2$ hours or, equivalently, $\lambda = 0.5$ (hour^{-1}), corresponding to about 50% decay after one and a half hours, and 95% decay after 6 hours.

The remaining parameters $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ can be estimated using the Iteratively Re-Weighted Generalized Least Squares (IRWGLS) which consists of the following steps:

Step 0 : Set the initial $\hat{\boldsymbol{\Sigma}}_\mathbf{Y}$ to an identify matrix of size $N_s \times N_t$.

Step 1 : Estimate $\boldsymbol{\beta}$ using the Feasible General Least Squares (FGLS):

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \hat{\boldsymbol{\Sigma}}_\mathbf{Y}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \hat{\boldsymbol{\Sigma}}_\mathbf{Y} \mathbf{Y}.$$

Step 2 : Based on $\mathbf{r} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} = (r(s_1, t_1), \ldots, r(s_{N_s}, t_{N_t}))^\top$, estimate $\hat{\boldsymbol{\theta}}$ of the parameter $\boldsymbol{\theta}$ associated with the covariance function $c(\cdot; \boldsymbol{\theta})$, and obtain the estimate of the covariance matrix, $\hat{\boldsymbol{\Sigma}}_\mathbf{Y}$, from equation (3.14).

Step 3 : Iterate steps 1 and 2 until the relative changes of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\theta}}$ are small.

Schabenberger and Gotway (2005) provide some useful discussions on both the advantages and disadvantages of the procedure above. To obtain $\hat{\boldsymbol{\theta}}$ in step 2, we use the fast cross-validation-type method [Carroll et al. (1997)]. In this case study, an Exponential model is chosen for $c(\cdot;\boldsymbol{\theta})$, and $c(\mathbf{d};\boldsymbol{\theta}) = \theta_1^2 \exp(-\theta_2\|\mathbf{d}\|)$ where $\|\mathbf{d}\|$ represents the spatial distance (in kilometer) between two locations. Specifically, let $\mathbf{r}_{(-\mathbf{s}_0,-t_0)}$ be a column vector of residuals with the residual at location $\mathbf{s}_0$ and time t_0 removed, and let $\eta(\mathbf{s}_0,t_0)$ be the leaving-one-station-out prediction error, that is, the error in predicting the residual $r(\mathbf{s}_0,t_0)$ using only $\mathbf{r}_{(-\mathbf{s}_0,-t_0)}$ given by

$$\eta(\mathbf{s}_0,t_0) = r(\mathbf{s}_0,t_0) - \hat{r}(\mathbf{s}_0,t_0),$$

where $\hat{r}(\mathbf{s}_0,t_0)$, given below, is the well-defined Simple Kriging predictor which is known to be the best linear estimator of $r(\mathbf{s}_0,t_0)$ under squared-error loss,

$$\hat{r}(\mathbf{s}_0,t_0) = \boldsymbol{\gamma}\boldsymbol{\Sigma}^{-1}_{\mathbf{r}_{(-\mathbf{s}_0,-t_0)}}\mathbf{r}_{(-\mathbf{s}_0,-t_0)}.$$

Here, $\boldsymbol{\gamma} = \operatorname{cov}(\mathbf{r}^{\top}_{(-\mathbf{s}_0,-t_0)}, r(\mathbf{s}_0,t_0))$, and the covariance matrix of $\mathbf{r}_{(-\mathbf{s}_0,-t_0)}$, $\boldsymbol{\Sigma}_{\mathbf{r}_{(-\mathbf{s}_0,-t_0)}}$, can be calculated using equation (3.14).

The optimum value of $\boldsymbol{\theta}$ is found by minimizing the sum of squared leaving-one-station-out prediction errors, that is,

$$\min_{\boldsymbol{\theta}} \sum_{i}^{N_s}\sum_{j}^{N_t} \eta^2(\mathbf{s}_i,t_j;\boldsymbol{\theta}). \tag{4.6}$$

A few remarks can be made as follows:

- In the first iteration, since $\hat{\boldsymbol{\Sigma}}_{\mathbf{Y}}$ is an identify matrix, $\hat{\boldsymbol{\beta}}$ in step 1 is the Ordinary Least Squares (OLS) estimator and is unbiased. In subsequent iterations, the finite-sample properties of the FGLS estimator, $\hat{\boldsymbol{\beta}}$, are usually unknown and can be studied case-by-case via Monte Carlo experiments. Asymptotically, the FGLS estimator possesses the asymptotic properties of the Maximum Likelihood estimator, and is equivalent to the Generalized Least Squares (GLS) estimator under regularity conditions.
- The cross-validation-type method is adopted for practical reasons. As noted by Carroll et al. (1997), due to the decay of pollutant concentration, the temporal correlation of pollutant concentration becomes extremely weak over a certain time lag. Hence, when solving the simple Kriging problem above, it is possible to only retain the data within certain time units of the time of interest (i.e., local Kriging). In that case, the dimension of the covariance involved in the simple Kriging problem can be significantly reduced. For example, because ozone is only formed during daylight and destroyed over night, we could only use the data within 24 hours of the time of interest, t_0, when solving the simple Kriging problem. This leads to a significant reduction of both the dimension of the

covariance matrix and the computational time compared with MLE which requires repeated inversion of a large covariance matrix. The cross-validation-type method has also been studied by Zhang and Wang (2010).

- Simulation studies can be used to study the finite-sample property of the estimators. The readers may refer to the online supplementary materials [Liu et al. (2016)]. In general, the simulation results do not suggest any systematic bias associated with $\hat{\beta}$, while the estimators, $\hat{\theta}_1$ and $\hat{\theta}_2$, seem to be negatively and positively biased, respectively.
- It is worth noting that the values of unknown model parameters are obtained using a hybrid approach, in which some parameters are computed from physical knowledge while the others are estimated using statistical methods. Recall that one of the main goals of this paper is to build a statistical model which is mathematically tractable and also reflects the fundamental physics of the air pollution process. In reality, the decay parameter λ is dynamic and depends on factors such as land use type, solar radiation, rainfall, pressure, temperature, concentration of other pollutants, etc. From the modeling point of view, if we allow λ to be dependent on other factors, then the linear form of (4.5) does not hold and the model becomes much less tractable. If the decay parameter λ is assumed to be an unknown constant, then it is still possible to estimate the value of λ using the same procedure above. However, the computational cost will be much higher because the design matrix $\mathbf{X}$ in (4.5) needs to be repeatedly computed for different values of λ. Note that, when λ is specified using the physical knowledge, we only need to compute $\mathbf{X}$ once.

4.3. *Results and interpretation*. Figure 8, as an illustration of the results, shows both the observed and fitted hourly ozone concentrations from July 20, 2013 9AM to July 30, 2013 8AM. Results from other time periods are similar. It is seen from Figure 8 that the model well captures the diurnal variation of hourly ozone concentration at different sites. Note that the temporal variation of hourly ozone concentrations at different sites can be very different. One might also note some lack of fit of the ozone concentration at station 6, which shows the difficulty of short-term ozone data modeling. One key challenge in modeling the urban air quality data for a metropolitan area like Singapore is that urban structures (say, buildings) affect the air quality, as urban structures greatly change the wind field. The wind field generated by our Numerical Weather Prediction model, or many other numerical weather models, does not take into account the effects of urban structures. As a result, there often exists some systematic bias between the numerically simulated and the actual wind field at some particular locations, for example, at the location of station 6. Another important reason is that the ozone generation and decay is a complicated photo-chemical process that involves many other chemicals in the air. Suppose that there is an unknown emission source of oxides of nitrogen (NO_x) near station 6, and, as a result, the behavior of ozone concentration can be very different at that location. Unlike most of the existing

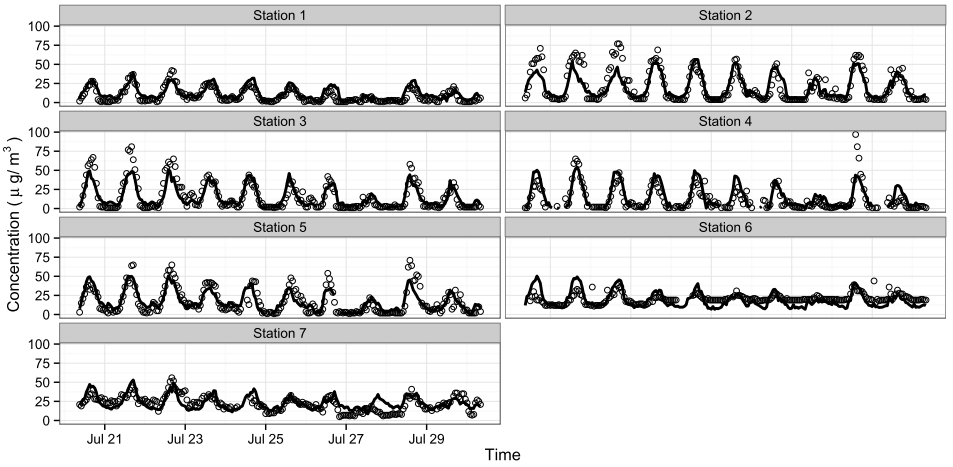


FIG. 8. *Fitted* (*solid line*) *and observed* (*circle*) *hourly ozone concentrations at the* 7 *stations from* 20 *July* 2013 9*AM to* 30 *July* 2013 8*AM*, *where the circles and lines represent the observed and fitted values respectively. The station ID is shown on the top of each subplot*.

literature that focuses on daily or 8-hour average ozone modeling, we consider in this paper the modeling of hourly ozone concentration which is subject to a much higher variability. Incorporating an additional dataset, if available, will certainly be helpful in explaining more uncertainties.

It is highly desirable that the estimated parameters can be interpreted associated with the physics behind the pollution generation process. The left panel of Figure 9 shows the estimated effects of temperature on ozone concentration at different hours of day. Because ozone can only be generated under solar radiation and is destroyed after sunset, we clearly see that the estimated effects of temperature on ozone concentration are positive during daylight (8AM–8PM), and are negative during night time (9PM–7AM). In particular, the effects are the largest around 3PM in the afternoon when the ozone concentration usually reaches the peak value of a day. The right panel of Figure 9 shows the estimated effects of traffic speed on ozone concentrations at each hour of day. An exactly opposite pattern is observed. During daytime, the estimated effects of traffic speed are negative, which indicates that higher traffic speed leads to lower ozone concentration. This is because higher traffic speed usually implies lower traffic volume. During nighttime, on the other hand, traffic speed has positive effects on ozone concentration. Recall that ozone is destroyed by nitric oxide (NO) at night and higher traffic speed usually implies a lower emission rate of NO. As a result, the ozone concentration remains relatively high when traffic speed is high.

Similarly, Figure 10 shows the estimated effects of different types of land use on ozone concentrations at different hours of day. The effects of residential area, nature reserve (including woods, forests, gardens and parks), commercial area and industrial area have similar effects on ozone concentrations in the sense that ozone

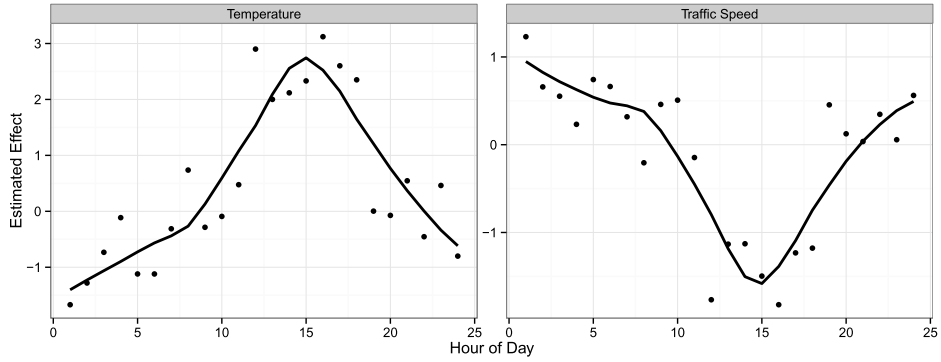


FIG. 9. *Estimated hourly effects of temperature* (*shown on the left*) *and traffic speed* (*shown on the right*) *on ozone concentrations*

is generated during daylight and destroyed at night. Particularly, it is clearly seen that the commercial area has the largest effects on the ozone concentration. Since most of the commercial areas are in downtown Singapore, such a conclusion is consistent with the well-known fact that ozone concentration in an urban area is strongly correlated with human activities.

We smooth the effects of different factors over time using the nonparametric Locally Weighted Scatterplot Smoothing (LOESS), and the smoothed curves are also shown in Figures 9 and 10. Alternatively, one might choose to refit the model by imposing some parametric relationships that describe the time-dependent effects of different factors, although there is no relevant physical or chemical knowledge that may help in choosing appropriate parametric forms to our knowledge.

The estimates of the parameters of the covariance function $c(\cdot;\boldsymbol{\theta})$ are given by $\hat{\theta}_1 = 0.5$ and $\hat{\theta}_2 = 0.2$. Based on (3.2), θ_1 captures the uncertainty associated with the pollutant emission, transport and decay processes at a particular location and

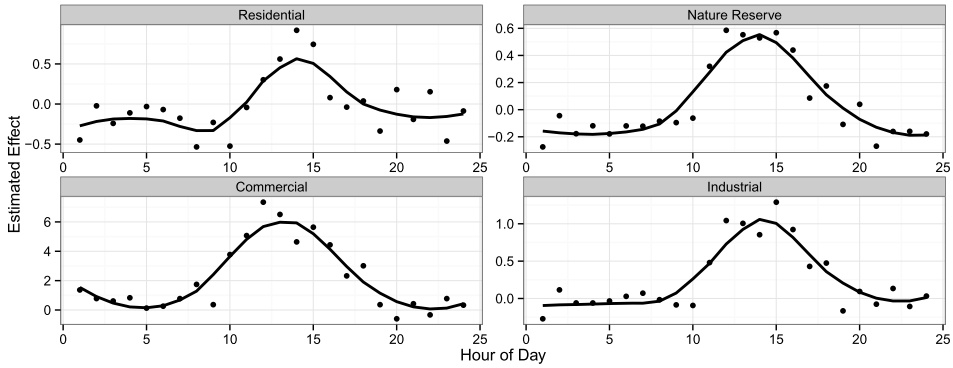


FIG. 10. *Estimated hourly effects of different land use type on ozone concentrations. The land use type is shown on the top of each subplot.*

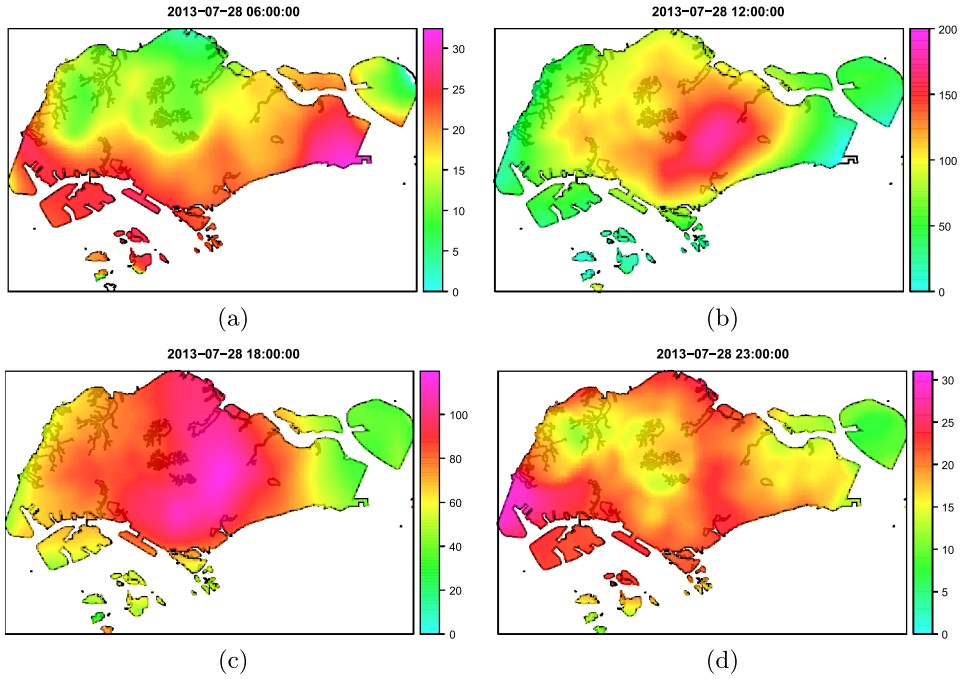


FIG. 11. *Ozone map at different hours on July*, 28 2013: (a) 6*AM*, (b) 12*PM*, (c) 6*PM*, (d) 11*PM*. *The concentration is represented by the color.*

within a one-hour time interval. From the estimate $\hat{\theta}_2 = 0.2$, the practical range of the semivariogram is $3\hat{\theta}_2^{-1} = 15$ kilometers. Of course, it is the stochastic transport process, $Z(\mathbf{s}, t)$, that largely determines the spatial and temporal correlation of the pollutant concentration data.

Once the model has been constructed, the ozone concentration at any location can be predicted and the ozone map of an entire Singapore is constructed. As an example, Figure 11 shows the ozone map at 6AM, 12PM, 6PM and 11PM of a particular day obtained by Universal Kriging. The daily change of the ozone pattern of that day is clearly seen. In the early morning before sunrise, the ozone concentration remains at a low level, and the concentration is slightly higher in the southern region. During the day, as temperature rises, ozone is generated and starts to build up. Figure 11(b) shows that ozone is mainly generated from the center of the island (i.e., downtown Singapore), and is gradually transported to the outskirts of Singapore due to pollutant dispersion. At 6PM of the day, ozone has been fully dispersed and the concentration over most parts of the island is almost uniform. At 11PM, the ozone concentration drops back to a low level; and it is interesting to see that the ozone removal rate is higher in some areas. It turns out those areas are either the nature reserves or the low-population density areas of Singapore, which are mainly located at the northwest part of the island.

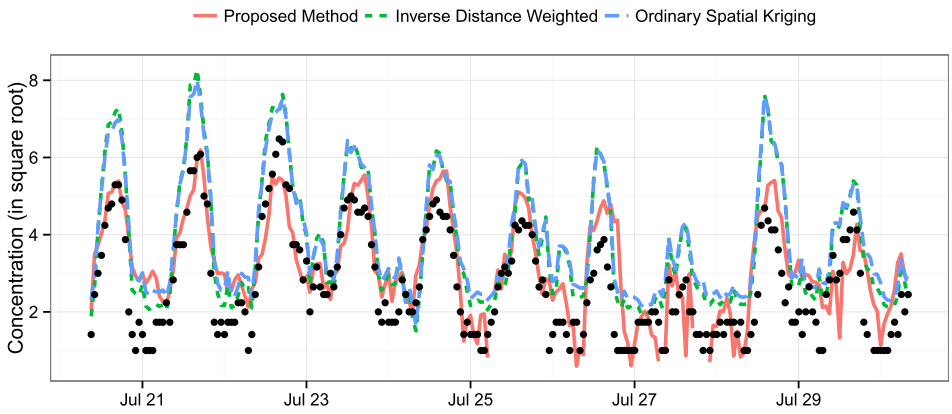


FIG. 12. *Cross-validation-based comparison between the proposed approach and other spatial interpolation methods.*

To further demonstrate the utility of the proposed model, a cross-validation-based comparison is performed between the proposed method and those commonly used spatial interpolation methods. For illustrative purposes, Figure 12 shows the cross-validation results at station one (see Figure 1 for the location of this station). In this figure, the observed hourly ozone concentrations (in square root) are shown by dots. The solid line, obtained from the proposed method, shows the leave-one-station-out predictions of the hourly ozone using only the data from the remaining 6 stations. The long- and short-dashed lines show the leave-one-station-out predictions, respectively obtained from the ordinary spatial Kriging with Matérn covariance function and the Inverse Distance Weighted interpolation (IDW). The proposed method successfully reduces the prediction error by approximately 40% compared with the ordinary spatial Kriging and IDW. This is mainly because the proposed method incorporates the dynamic environmental conditions and basic physics into the statistical modeling.

5. Conclusions. This paper proposed a spatial-temporal air quality model, and demonstrated the application of the model through a real case study. In the modeling of many complex environmental processes such as hourly level air quality, physical knowledge and principles can be extremely useful in model selection, construction and interpretation, although a pure physical model may not perform well in data fitting and prediction due to the uncertainty associated with many physical parameters. The model presented in this paper successfully established the link between the statistical air quality model and physical pollutant dispersion models, and can be seen as an example of how physics and statistics can be integrated in solving real-world problems. Although the paper focuses on the modeling of hourly ozone, the general framework can be applied for other pollutants or environmental processes.

APPENDIX

In the appendix, we show how equation (3.14) is derived. It follows from (3.11) and (3.13) that the covariance, $\operatorname{cov}(Y(\mathbf{s}_1,t_1),Y(\mathbf{s}_2,t_2))$ for $t_1 \le t_2$, can be written as

$$
\begin{aligned}
&\operatorname{cov}\big(Y(\mathbf{s}_1,t_1),Y(\mathbf{s}_2,t_2)\big)\\
&\quad=\operatorname{cov}\Bigg(\sum_{i=1}^{n}\big(\Psi_{i,t_1}(\mathbf{s}_1)*\varepsilon(\mathbf{s}_1,t_1-i\delta)\big),\varepsilon(\mathbf{s}_2,t_2)\Bigg)\\
&\qquad+\operatorname{cov}\big(\varepsilon(\mathbf{s}_1,t_1),\varepsilon(\mathbf{s}_2,t_2)\big)\\
&\qquad+\operatorname{cov}\Bigg(\sum_{i=1}^{n}\big(\Psi_{i,t_2}(\mathbf{s}_2)*\varepsilon(\mathbf{s}_2,t_2-i\delta)\big),\varepsilon(\mathbf{s}_1,t_1)\Bigg)\\
&\qquad+\operatorname{cov}\Bigg(\sum_{i=1}^{n}\big(\Psi_{i,t_1}(\mathbf{s}_1)*\varepsilon(\mathbf{s}_1,t_1-i\delta)\big),\\
&\qquad\sum_{i=1}^{n}\big(\Psi_{i,t_2}(\mathbf{s}_2)*\varepsilon(\mathbf{s}_2,t_2-i\delta)\big)\Bigg).
\end{aligned}
\tag{A.1}
$$

In what follows, we obtain the expression for each term on the right-hand side (RHS) of (A.1). Since $\varepsilon(\mathbf{s},t)$ is a white-in-time isotropic random field with spatial covariance function $c(\cdot)$, we immediately obtain the expressions of the first two terms on the RHS of (A.1):

$$
\operatorname{cov}\Bigg(\sum_{i=1}^{n}\big(\Psi_{i,t_1}(\mathbf{s}_1)*\varepsilon(\mathbf{s}_1,t_1-i\delta)\big),\varepsilon(\mathbf{s}_2,t_2)\Bigg)=0,\qquad t_1\le t_2 \tag{A.2}
$$

and

$$
\operatorname{cov}\big(\varepsilon(\mathbf{s}_1,t_1),\varepsilon(\mathbf{s}_2,t_2)\big)=\begin{cases}c(\mathbf{d}), & \text{if } t_1=t_2,\\ 0, & \text{otherwise,}\end{cases} \tag{A.3}
$$

where $\mathbf{d}=\mathbf{s}_2-\mathbf{s}_1$ is a vector. Note that, since the random field $\varepsilon(\mathbf{s},t)$ is isotropic, $c(\mathbf{d})=c(\|\mathbf{d}\|)$ with $\|\mathbf{d}\|$ representing the distance between $\mathbf{s}_1$ and $\mathbf{s}_2$.

Letting $t_1=t_2-\jmath\delta$ for some $\jmath\ge 0$, the expression of the third term on the RHS of (A.1) is derived as follows:

$$
\begin{aligned}
&\operatorname{cov}\Bigg(\sum_{i=1}^{n}\big(\Psi_{i,t_2}(\mathbf{s}_2)*\varepsilon(\mathbf{s}_2,t_2-i\delta)\big),\varepsilon(\mathbf{s}_1,t_1)\Bigg)\\
&\quad=\operatorname{cov}\big(\Psi_{\jmath,t_2}(\mathbf{s}_2)*\varepsilon(\mathbf{s}_2,t_1),\varepsilon(\mathbf{s}_1,t_1)\big)\\
&\quad=\operatorname{cov}\Bigg(\int_{\mathbb{R}^2}\Psi_{\jmath,t_2}(\mathbf{x})\varepsilon(\mathbf{s}_2-\mathbf{x},t_1)\,d\mathbf{x},\varepsilon(\mathbf{s}_1,t_1)\Bigg)\\
&\quad=(\Psi_{\jmath,t_2}*c)(\mathbf{d}).
\end{aligned}
\tag{A.4}
$$

The expression of the fourth term on the RHS of (A.1) can be derived in a similar way as follows:

$$\begin{aligned}
&\operatorname{cov}\left(\sum_{i=1}^{n}\big(\Psi_{i,t_1}(\mathbf{s}_1) * \varepsilon(\mathbf{s}_1, t_1 - i\delta)\big), \sum_{i=1}^{n}\big(\Psi_{i,t_2}(\mathbf{s}_2) * \varepsilon(\mathbf{s}_2, t_2 - i\delta)\big)\right) \\
&\quad = \sum_{i=1}^{n}\operatorname{cov}\big(\Psi_{i,t_1}(\mathbf{s}_1) * \varepsilon(\mathbf{s}_1, t_1 - i\delta), \Psi_{J+i,t_2}(\mathbf{s}_2) * \varepsilon(\mathbf{s}_2, t_2 - J\delta - i\delta)\big) \qquad \text{(A.5)} \\
&\quad = \sum_{i=1}^{n}(\tilde{\Psi}_{i,t_1} * \Psi_{J+i,t_2} * c)(\mathbf{d}),
\end{aligned}$$

where $\tilde{\Psi}_{i,t_1}(\mathbf{s}) = \Psi_{i,t_1}(-\mathbf{s})$.

Since $\tilde{\Psi}_{i,t} = \Psi_{i,t} = 1$ when $i = 0$, we have

$$\begin{aligned}
&\operatorname{cov}\big(Y(\mathbf{s}_1, t_1), Y(\mathbf{s}_2, t_2)\big) \\
&\quad = \sum_{i=0}^{n}(\tilde{\Psi}_{i,t_1} * \Psi_{J+i,t_2} * c)(\mathbf{d}) + I_{\{J=0\}}c(\mathbf{d})
\end{aligned}$$

where $I_{\{J=0\}} = 1$ when $J = 0$, otherwise, $I_{\{J=0\}} = 0$.

Acknowledgments. The authors would like to thank the National Environment Agency of Singapore for providing the environmental monitoring data used in this paper. The authors would also like to thank Dr. Nicoleta Serban (the Editor), the Associate Editor and two referees for their constructive comments that greatly helped to improve the quality of the paper.

SUPPLEMENTARY MATERIAL

A simulation study and some useful animations (DOI: 10.1214/15-AOAS901SUPP; .zip). Because the finite-sample properties of the estimators presented in Section 4.2 are usually unknown, a Monte Carlo simulation study is performed to investigate the statistical properties, such as unbiasedness, of the estimators. In addition, some useful animations are also provided to illustrate the proposed modeling approach.

REFERENCES

BANERJEE, S., CARLIN, B. P. and GELFAND, A. E. (2014). *Hierarchical Modeling and Analysis for Spatial Data*, 2nd ed. *Monographs on Statistics and Applied Probability* **135**. CRC Press, Boca Raton, FL. MR3362184

BERLINER, L. M. (2003). Physical-statistical modeling in geophysics. *Journal of Geophysical Research-Atmospheres* **108** STS 3-1–STS 3-10.

BROWN, P. E., KÅRESEN, K. F., ROBERTS, G. O. and TONELLATO, S. (2000). Blur-generated nonseparable space–time models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **62** 847–860. MR1796297

BYUN, D. and SCHERE, K. L. (2006). Review of the governing equations, computational algorithms, and other components of the models-3 community multiscale air quality (CMAQ) modeling system. *Applied Mechanics Reviews* **59** 51–77.

CALDER, C. A. (2007). Dynamic factor process convolution models for multivariate space–time data with application to air quality assessment. *Environ. Ecol. Stat.* **14** 229–247. MR2405328

CAMELETTI, M., LINDGREN, F., SIMPSON, D. and RUE, H. (2013). Spatio-temporal modeling of particulate matter concentration through the SPDE approach. *AStA Adv. Stat. Anal.* **97** 109–131. MR3045763

CARROLL, R., CHEN, E., LI, T., NEWTON, H., SCHMIEDICHE, H. and WANG, N. (1997). Ozone exposure and population density in Harris county. *Texas. Journal of the American Statistical Association* **92** 392–404.

CHRISTAKOS, G. and VYAS, V. (1998). A composite space–time approach to studying ozone distribution over eastern United States. *Atmospheric Environment* **32** 2845–2857.

COATS, C. (1996). High performance algorithms in the sparse matrix operator kernel emissions modelling system. In *Proceedings of the Ninth Joint Conference on Applications of Air Pollution Meteorology of the American Meteorological Society and the Air and Waste Management Association*. Atlanta, GA.

CRESSIE, N. and HUANG, H.-C. (1999). Classes of nonseparable, spatio-temporal stationary covariance functions. *J. Amer. Statist. Assoc.* **94** 1330–1340. MR1731494

DOU, Y., LE, N. D. and ZIDEK, J. V. (2010). Modeling hourly ozone concentration fields. *Ann. Appl. Stat.* **4** 1183–1213. MR2751338

FUENTES, M. (2009). Statistical issues in health impact assessment at the state and local levels. *Air Quality, Atmosphere and Health* **2** 47–55.

FUENTES, M., CHEN, L., DAVIS, J. M. and LACKMANN, G. M. (2005). Modeling and predicting complex space–time structures and patterns of coastal wind fields. *Environmetrics* **16** 449–464. MR2147536

GHOSH, S. K., BHAVE, P. V., DAVIS, J. M. and LEE, H. (2010). Spatio-temporal analysis of total nitrate concentrations using dynamic statistical models. *J. Amer. Statist. Assoc.* **105** 538–551. MR2759930

GNEITING, T. (2002). Nonseparable, stationary covariance functions for space–time data. *J. Amer. Statist. Assoc.* **97** 590–600. MR1941475

HAN, S., BIAN, H., FENG, Y., LIU, A., LI, X., ZENG, F. and ZHANG, X. (2011). Analysis of the relationship between O3, NO and NO2 in tianjin. *China. Aerosol and Air Quality Research* **11** 128–139.

HASLETT, J. and RAFTERY, A. (1989). Space–time modelling with long-memory dependence: Assessing Ireland's wind power resource (with discussion). *Applied Statistics* **38** 1–50.

HIGDON, D. (2002). Space and space–time modeling using process convolutions. In *Quantitative Methods for Current Environmental Issues* (C. Anderson, V. Barnett, P. Chatwind and A. El-Shaarawi, eds.) 37–56. Springer, London. MR2059819

HIGDON, D. (2007). A process-convolution approach to modeling temperatures in the North Atlantic Ocean. *Environ. Ecol. Stat.* **5** 173–190.

HUANG, H. and HSU, N. (2004). Modeling transport effects on ground-level ozone using a nonstationary space–time model. *Environmetrics* **15** 251–268.

LIU, X., YEO, K., HWANG, Y., SINGH, J. and KALAGNANAM, J. (2016). Supplement to "A statistical modeling approach for air quality data based on physical dispersion processes and its application to ozone modeling." DOI:10.1214/15-AOAS901SUPP.

MALMBERG, A., ARELLANO, A., EDWARDS, D. P., FLYER, N., NYCHKA, D. and WIKLE, C. (2008). Interpolating fields of carbon monoxide data using a hybrid statistical-physical model. *Ann. Appl. Stat.* **2** 1231–1248. MR2655657

REICH, B. J. and FUENTES, M. (2007). A multivariate semiparametric Bayesian spatial modeling framework for hurricane surface wind fields. *Ann. Appl. Stat.* **1** 249–264. MR2393850

REICH, B. J., EIDSVIK, J., GUINDANI, M., NAIL, A. J. and SCHMIDT, A. M. (2011). A class of covariate-dependent spatiotemporal covariance functions for the analysis of daily ozone concentration. *Ann. Appl. Stat.* **5** 2425–2447. MR2907121

REICH, B., COOLEY, D., FOLEY, K., NAPELENOK, S. and SHABY, B. (2013). Extreme value analysis for evaluating ozone control strategies. *Ann. Appl. Stat.* **7** 739–762. MR3112916

SAHU, S. K., GELFAND, A. E. and HOLLAND, D. M. (2007). High-resolution space–time ozone modeling for assessing trends. *J. Amer. Statist. Assoc.* **102** 1221–1234. MR2412545

SCHABENBERGER, O. and GOTWAY, C. A. (2005). *Statistical Methods for Spatial Data Analysis*. Chapman & Hall/CRC, Boca Raton, FL. MR2134116

SHADDICK, G., LEE, D., ZIDEK, J. V. and SALWAY, R. (2008). Estimating exposure response functions using ambient pollution concentrations. *Ann. Appl. Stat.* **2** 1249–1270. MR2655658

SIGRIST, F., KÜNSCH, H. R. and STAHEL, W. A. (2015). Stochastic partial differential equation based modelling of large space–time data sets. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **77** 3–33. MR3299397

SKAMAROCK, W. C., KLEMP, J. B., DUDHIA, J., GILL, D. O., BARKER, D. M., DUDA, M. G., HUANG, X. Y., WANG, W. and POWERS, J. G. (2008). A description of the advanced research WRF version 3, Boulder, Colorado, USA. Ncar Technical Note: NCAR/TN–475+STR.

SMITH, L., FUENTES, M., REICH, B. and EDER, B. (2013). Prediction of speciated particulate matter and bias assessment of numerical output data. *International Journal of Environmental Science and Engineering Research* **4** 8–17.

STEIN, M. L. (2007). Spatial variation of total column ozone on a global scale. *Ann. Appl. Stat.* **1** 191–210. MR2393847

STEIN, M. L. (2009). Spatial interpolation of high-frequency monitoring data. *Ann. Appl. Stat.* **3** 272–291. MR2668708

STROUD, J. R., MÜLLER, P. and SANSÓ, B. (2001). Dynamic models for spatiotemporal data. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **63** 673–689. MR1872059

UNITED STATES ENVIRONMENTAL PROTECTION AGENCY (1998). EPA third-generation air quality modeling system, models-3 (EPA-600/R-98/069a). U.S. Environmental Protection Agency, Research Triangle Park, NC.

UNITED STATES ENVIRONMENTAL PROTECTION AGENCY (2003). Air Quality index—A guide to air quality and your health, EPA-454/K-03-002. U.S. Environmental Protection Agency, Research Triangle Park, NC.

UNITED STATES ENVIRONMENTAL PROTECTION AGENCY (2012). Clean air act: Title I—Air pollution prevention and control. Available at http://epa.gov/oar/caa/title1.html.

WIKLE, C. K., MILLIFF, R. F., NYCHKA, D. and BERLINER, L. M. (2001). Spatiotemporal hierarchical Bayesian modeling: Tropical ocean surface winds. *J. Amer. Statist. Assoc.* **96** 382–397. MR1939342

WILSON, A., RAPPOLD, A. G., NEAS, L. M. and REICH, B. J. (2014). Modeling the effect of temperature on ozone-related mortality. *Ann. Appl. Stat.* **8** 1728–1749. MR3271351

WORLD HEALTH ORGANIZATION (2005). WHO Air Quality Guidelines for Paticularte Matter, Ozone, Nitrogen Dioxide and Sulfur Dioxide–Global Update (WHO/SDE/PHE/OEH/06.02), World Health Organization.

XU, Y., VIZUETE, W. and SERRE, M. (2012). Characterization of air quality ozone model performance using land use regression model: An application in exposure assessment for epidemology studies. In *The* 11*th Annual CMAS Conference*, Chapel Hill, NC.

ZHANG, H. and WANG, Y. (2010). Kriging and cross-validation for massive spatial data. *Environmetrics* **21** 290–304. MR2842244

X. LIU
K. YEO
Y. HWANG
J. KALAGNANAM
IBM T.J. WATSON RESEARCH CENTER
29-252, 1101 KITCHAWAN ROAD
YORKTOWN HEIGHTS, NEW YORK, 10598
USA
E-MAIL: liuxiao314923@gmail.com
kyeo@us.ibm.com
yhwang@us.ibm.com
jayant@us.ibm.com

J. SINGH
9 CHANGI BUSINESS PARK CENTRAL 1
THE IBM PLACE
IBM SINGAPORE
E-MAIL: jsingh@sg.ibm.com