Science of the Total Environment 773 (2021) 145580

Contents lists available at ScienceDirect

Science of the Total Environment

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

Optimal estimation of initial concentrations and emission sources with 4D-Var for air pollution prediction in a 2D transport model

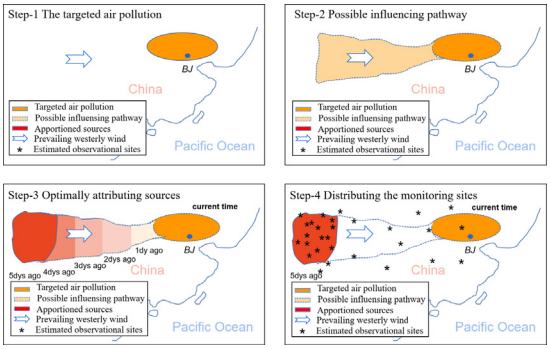
Caili Liu [a], Shaoqing Zhang [a,b,c,d,*], Yang Gao [e,c,**], Yuhang Wang [f], Lifang Sheng [a], Huiwang Gao [e], J.C.H. Fung [g]

[a] The College of Oceanic and Atmospheric Sciences, Ocean University of China, Qingdao, China
[b] Key Laboratory of Physical Oceanography, MOE, Institute for Advanced Ocean Study, Frontiers Science Center for Deep Ocean Multispheres and Earth System (FDOMES), Ocean University of China, China
[c] Ocean Dynamics and Climate Function Lab, Pilot National Laboratory for Marine Science and Technology (QNLM), Qingdao, China
[d] International Laboratory for High-Resolution Earth System Model and Prediction (iHESP), Qingdao, China
[e] Key Laboratory of Marine Environment and Ecology, and Frontiers Science Center for Deep Ocean Multispheres and Earth System (FDOMES), Ministry of Education, Ocean University of China, Qingdao 266100, China
[f] School of Earth and Atmospheric Sciences, Georgia Institute of Technology, Atlanta, GA 30332, United States of America
[g] Division of Environment and Sustainability, Hong Kong University of Science and Technology, China

HIGHLIGHTS

- Emission sources of air pollution are attributed by means of the 4-dimensional variational approach.
- Distributions of pollution sources are examined to be useful in distributing monitoring sites.
- Accurate air pollution forecast is available with effective deployment of the monitoring sites.
- Air pollution predictability is limited by the impacts of observational network.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 31 August 2020
Received in revised form 13 January 2021
Accepted 28 January 2021
Available online 5 February 2021

Editor: Pingqing Fu

Keywords:
Transport model
Optimal observational network
Adjoint sensitivity
Air pollution prediction
Optimization algorithm

ABSTRACT

Attributing sources of air pollution events by deploying an efficient observational network is an important and interesting problem in air quality control and forecast studies, but it is very challenging. In order to estimate the sensitivities of pollution events to emission sources, a comprehensive framework is built based on a horizontal 2-dimensional transport model and its adjoint in solving this problem. In an analysis of an idealized air pollution event of $PM_{2.5}$ over the region of North China, an objective function is defined to optimally estimate the initial concentrations and emission sources through a series of minimization procedures. Results by means of the 4-dimensional variational approach show that, with the optimal initial conditions and emission sources, the model can successfully forecast the pollution event in a few days. The optimal observing network based on sensitivity analysis takes only one third of the cost but greatly retains predictability skill compared to the full-grid observing system, while nearly no predictability skill is detectable if the same number of observational sites is randomly deployed. We evaluate air pollution predictability in the point of focusing on to what degree the root mean square errors between the modeled concentration and the targeted air pollution are limited by the optimal observational network. Results show that air pollution predictability in association with the optimal observational network is limited in the time scales about 6 days. With the high efficiency and in an economic

* Correspondence to: S. Zhang, The College of Oceanic and Atmospheric Sciences, Ocean University of China, Qingdao, China.
** Correspondence to: Y. Gao, Key Laboratory of Marine Environment and Ecology, and Frontiers Science Center for Deep Ocean Multispheres and Earth System (FDOMES), Ministry of Education, Ocean University of China, Qingdao 266100, China.
E-mail addresses: szhang@ouc.edu.cn (S. Zhang), yanggao@ouc.edu.cn (Y. Gao).

https://doi.org/10.1016/j.scitotenv.2021.145580

fashion, such a sensitivity-based optimal observing system holds promise for accurately predicting an air pollution event in the targeted area once the adjoint and variational procedure of a realistic atmosphere model including transport and chemical processes is performed.

1. Introduction

Since the industry revolution from 19th century, air pollution has been considered as a major threat to human health based on the World Health Organization (WHO) report (Venkatesan, 2016) and previous studies (Kampa and Castanas, 2008; Sweileh et al., 2018). Nowadays, with the development of numerical technique, air pollution forecast becomes the strong support of public decision making in air quality control. The accuracy of air pollution forecast is significantly affected by the initial conditions, which is also well known as the first-kind problem of predictability in weather forecast and related subjects (Mu and Wang, 2001). As a result, accurate air pollution forecast requires efficient information offered in advance, in which initial concentrations and emission sources play a key role (Henry, 2008; Sharma, 2009) and the effective deployment of monitoring sites is the first step (Joly and Peuch, 2012).

A number of studies have tried to track air pollution sources using different techniques. The trajectory clustering method (Wang et al., 2009) identifies potential emission sources based on the back trajectories of air mass with statistical analysis, but it may lose information in those areas without observations. Source attribution based on receptor-based models provides emission estimates, such as Positive Matrix Factorization (PMF) (Lee et al., 1999; Song et al., 2006) or Chemical Mass Balance (CMB) (Marmur et al., 2005), and three dimensional chemical transport models, i.e., tagged species method via the Ozone and Particle Source Apportionment Technology (OSAT/PSAT) within the Comprehensive Air Quality Model with Extensions (CAMx) (Karamchandani et al., 2017; Wagstrom et al., 2008) or the Integrated Source Apportionment Method (ISAM) within the Community Multiscale Air Quality Modeling (CMAQ) (Gao et al., 2020; Kwok et al., 2013), but source attribution using these methods is in local and only a single emission source or a limited number of predefined sources can be tracked.

In contrast, the adjoint technique is considered as an efficient tool to track multiple emission sources simultaneously with no need to set the predefined regions (Liu et al., 2007; Zhu and Zeng, 2002), and understand how modeled pollutant concentrations vary with emissions or reaction rates (Menut et al., 2000). It is in particular efficient and useful to tackle the backward problems with a limited number of outputs while the number of inputs is relatively large (Cacuci, 1981; Daescu and Carmichael, 2003; Hakami et al., 2006). Pudykiewicz (1998) firstly developed the adjoint model of a tracer transport equation to attribute the emission sources and then applied it to monitor the impacts of nuclear testing. An adaptive location method (Daescu and Carmichael, 2003) was proposed for the observational system in a general framework with adjoint sensitivity analyses. Though the adjoint method has been applied in the previous studies with a major focus on tackling the emission sources or improving the original emission inventory, i.e., GEOS-Chem (Zhang et al., 2016), WRF-Chem (Chen et al., 2019; Mizzi et al., 2016), CMAQ (Hakami et al., 2007; Park et al., 2018) and GRAPES–CUACE (An et al., 2016), apparently, there is a weak knowledge about the influences of attributed pollution sources along the pathway of pollutants on air pollution forecasts, which makes the distribution of monitoring sites still challenging.

The monitoring sites can be of different characteristics with regard to various air pollutants and the geographical area, and the observation stations are usually classified to make up observational networks for different objectives (Spangl et al., 2007). Flemming et al. (2005) illustrated that the classifications of air quality monitoring sites is linked either to an assessment of emissions or to the measurement of concentrations, and they showed the results of the concentration-based approach for observed multi-pollutants in Germany. Joly and Peuch (2012) overviewed the concentration dataset in studies before and then developed a method of linear discriminant analysis to classify monitoring stations in rural and urban areas. Most of the current studies (Flemming et al., 2005; Joly and Peuch, 2012; Kracht and Gerboles, 2019; Kracht et al., 2017; Spangl et al., 2007) focus on analysis of the concentration observations for the objective classification of monitoring stations, but the relevant evidence from the emission-based approach is rarely found in the literature.

Here, we use the 4-dimensional variational (4D-Var) approach to optimally attribute the sources of air pollution, aiming at providing a general framework with the effects of emission sources for distributing sites of an optimal observational network. Clearly illustrated in a 2-dimensional (2D) transport model framework, we show that monitoring sites of the observational network based on adjoint sensitivity analysis can offer optimal initial conditions to accurate prediction of air pollution with a minimum cost of monitoring sites. As an exploratory work, the study is not aiming to developing delicate classification schemes for the representativeness assessment of monitoring stations, which involves in too complex constraints beyond the ability of the current simple 2D framework. Therefore, the objective of the designed network only takes a simplest case, i.e., only considering the cost of network deployment.

In what follows, we first describe the methodology in Section 2, including the general description of optimally attributing sources in 4D-Var approach and four experimental designs. In Section 3, a 2D transport model and its adjoint are developed as a framework for illustration, and a minimization algorithm is adopted and examined. Sensitivity analysis is given in Section 4.1. The optimal initial concentrations and emission sources are analyzed in Section 4.2. In Section 4.3, the observational network based on adjoint sensitivity is designed, and the efficiency and cost are evaluated. We discuss the impacts of the optimal observational network on the air pollution predictability in Section 4.4. The summary and discussions are given in Section 5.

2. Methodology

2.1. Optimally attributing sources of air pollution in 4D-Var

Given a pollution event at the current time T (Fig. 1(a)), if the event is only resulted from the transport process of pollution sources, two questions have to be addressed for the prediction issue: (1) What is the possible pathway of the pollution sources in the past time period of (0, T) resulting in the event (Fig. 1(b))? (2) If an accurate prediction is pursued at the time 0, what kind of observing network (in terms of locations and coverage density) to measure initial concentrations and emissions (that could be persistent until T) is adequate (Fig. 1(cd))?

These two questions can be answered with 4D-Var approach. Given an atmosphere model including transport and chemical reaction processes $\frac{\partial c}{\partial t} = M(\boldsymbol{s})$, where the model state vector $\boldsymbol{s}$ includes the model dynamical state fields and the traceable pollutant being transported, c is the pollutant concentration. Define a cost function J to measure the strength of the pollution, i.e., the integral of the pollutants on the domain Ω: $J = \iiint f(c_T(x,y,z))dxdydz$, where c_T is the model pollutant concentration distribution at the current time.

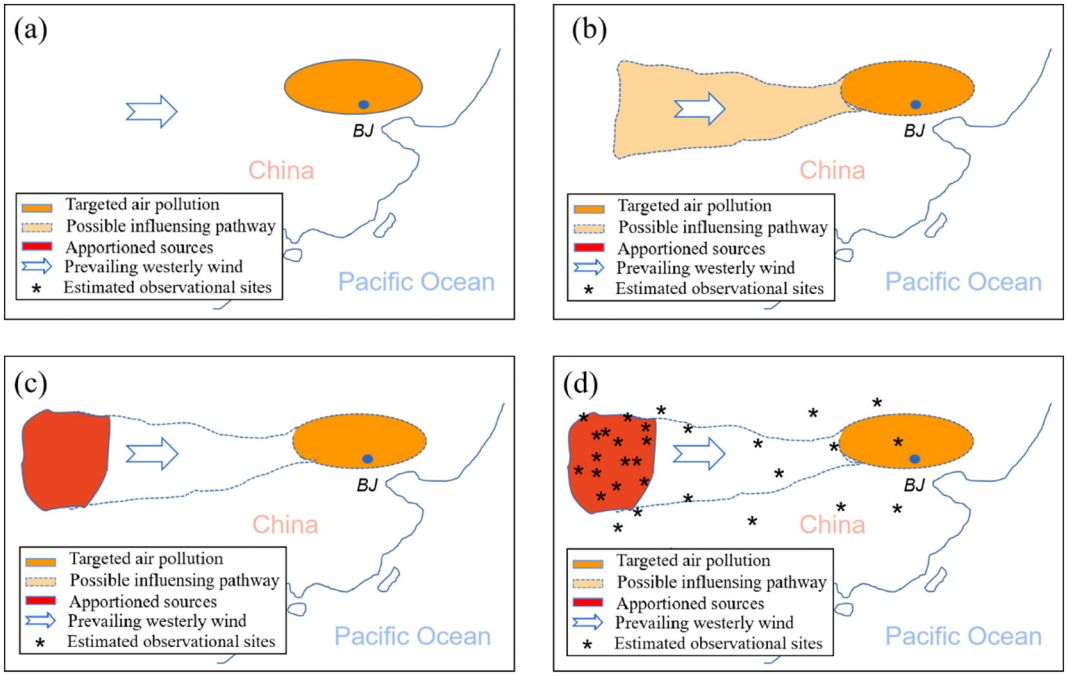


Fig. 1. Conceptual illustration of optimally attributing sources of air pollution (orange oval) under the background of prevailing westerly wind (arrow), around the target area over the North of China centered at Beijing (denoted by the blue dot), China. *a*) Occurrence of an air pollution event in the North of China. *b*) Possible influencing pathway of air pollution to the target area. *c*) The locations and strength of the sources (red-shaded) from the solution of the minimization problem in several days ago. *d*) Same as c), but added with the estimated observational sites (asterisks). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The answer to the first question is equivalent to what the contributive (sensitive) areas are for J of the pollutant sources in the past time periods. That is saying, we want to solve the first-order derivatives of J with respect to emission, i.e., $\nabla_{E_0} J$, where ∇ is the gradient notation. This can be solved by the adjoint model of the transport model integrated backward in time from T to 0, denoted as $-\frac{\partial \widehat{\delta c}}{\partial t} = \left(\frac{\partial M}{\partial \boldsymbol{s}}\right)^T \widehat{\delta \boldsymbol{s}}$, in (T, 0), as illustrated in the lower part of Fig. A in Supplementary materials marked as "Adjoint model, integrating backward in time."

The answer to the second question is a minimization problem: what are the exact distributions of initial concentrations and emissions leading to the occurrence of the pollution event at time T? In other words, we would like to estimate the optimal concentration c_0 and E_0 for the air pollution forecast so that the model concentration ${c_T}^M$ will be accurately approached to the air pollution c_T at time T. This optimal estimate can be solved through the minimization procedures shown in Fig. A in Supplementary materials, in which the following equation is minimized, i.e.,

$$\min_{c_0, E_0} \left\{ J = \iiint_{(x,y,z)}^{\Omega} \left[{c_T}^M(x,y,z) - c_T(x,y,z) \right]^2 dxdydz \right\} \quad (1)$$

with respect to initial concentration c_0 and emission E_0. The minimization is combining the adjoint model with a minimization algorithm.

The minimization is usually implemented by iterations (Fig. A in Supplementary materials) using the gradients solved from the adjoint model. In each iteration, an atmosphere model including transport and chemical reactions model is integrated forward in time to derive the model concentration c_t at time T and calculate the value of cost function J. Then the adjoint model is integrated backward in time to derive the gradient of J with respect to initial concentrations and emissions, i.e., $\nabla_{c_0} J$ and $\nabla_{E_0} J$. Finally, a minimization algorithm is adopted to estimate the optimally initial concentrations and emissions using the cost function and the associated gradients as the minimization converges when the norm of gradient of the cost function becomes small enough.

2.2. Experimental design

To gain a clear understanding of optimally attributing sources of a pollution event in 4D-Var in this study, we design the following 4 idealized experiments which simulate the pollution prediction problem from the simplest to the relatively realistic scenarios.

Exp1. An air pollution event with a uniform $PM_{2.5}$ concentration of 150 μg m^{-3}, which is normally considered as the level of severe pollution based on air quality index (AQI) (HJ 633–2012; (MEPPRC, 2012)), was assumed to occur over the region of North China (the area used to define objective function in Section 3.3). The initial concentrations with the value of zero are applied, and the pollution event is the consequence of the emission at the initial time only (i.e., 5 days ago). To clearly demonstrate the development of methodology, here we only consider idealized air pollution configuration. We will give discussions about the applications in observations and predictions of real pollution situations at the result analysis in Section 5.

Exp2. The same as **Exp1**, but the emission is assumed to be persistent for a while (in the first 24 h, for instance).

Exp3. The same as **Exp2**, but the emission is assumed to be persistent until the pollution event occurs.

Exp4. The same as **Exp3**, but the pollutant has an initial distribution, i.e., the pollution event is the consequence of a combination of the persistent emission and initial concentrations. This is the closest case to a real prediction problem. Using this experimental setting and 4D-Var procedure, we will construct and examine the optimal observing system in the pollution prediction, and complete the study of the optimally attributing sources of air pollution in 4D-Var.

Once the atmosphere model including pollution evolution processes (transport and chemical reactions, for instance) and its adjoint are set, with the approaches described in Section 2.1 (as illustrated in Fig. A in Supplementary materials) to **Exp1**, we can derive the distribution of optimal initial emission that causes the targeted pollution event ahead a

few days (5 days, for instance). Applying the approach to **Exp2** (**Exp3**), we can derive the optimal distribution of the pollution emission that is discharged persistently for a while in the beginning of the forecast period (the whole forecast period). When we apply the 4D-Var approach to **Exp4**, we can derive the optimal distributions and strength for both the persistent emission source and initial pollutant concentration.

Next, we develop a 2-dimensional transport model and its adjoint, and with the aid of a minimization algorithm to implement the methodology framework we described above and investigate the issue of optimizing pollution predictions.

3. A 2D transport model, adjoint and minimization

3.1. Development of a 2D transport model and its tangent linear

a. The 2D transport model

To implement the 4D-Var approach of optimally-attributing emission source for a pollution event described in Section 2.1, we develop a simple 2D transport model that delineates the processes including advection, diffusion, emission, and deposition. The chemical reactions are not considered in this study for simplicity. The concentration variations of air pollutants can be described by the nonlinear partial differential equation (Daescu and Carmichael, 2003) as Eq. (2).

$$\frac{\partial}{\partial t}c = -\nabla\cdot(\boldsymbol{u}c) + \nabla\cdot\left[\rho\boldsymbol{K}\cdot\nabla\left(\frac{c}{\rho}\right)\right] + E - D \tag{2}$$

In a spatial domain horizontally $\boldsymbol{x} = (x,y)$ belongs to domain Ω in Euclidean space R^2. With the time interval of $[t_0, t_n]$, the concentration $c(t_k,\boldsymbol{x})$ at time t_k (k = 0, …, n) is solved by wind field $\boldsymbol{u}$, air density ρ, emissions source E and deposition process D. Note that only dry deposition is considered in this study. The operator $\nabla = (\partial/\partial x, \partial/\partial y)$ represents the gradient in mathematics. Diffusion coefficient K is 1 $m^2\ s^{-1}$ in the 2D transport process which is reasonably less than the values of 3–6 $m^2\ s^{-1}$ usually used in urban areas (Pérez-Roa et al., 2006). We adopt a global atmosphere circulation model to drive the transport of pollutant, of which the west-eastward boundary condition is periodic along the latitude circles, and the south-northward boundary condition is the averaged values of the calculations at the latitudes nearest the poles. The initial condition used in Eq. (2) is set as Eq. (3)

$$c(t_0,\boldsymbol{x}) = c_0(\boldsymbol{x}) \tag{3}$$

The deposition flux can be estimated by the product of the pollutant concentration and deposition velocity rate, which is set as a relatively low value of 1×10^{-4} m/s based on Zhao et al. (2019). The wind field $\boldsymbol{u} = (u,v)$ is provided by a global barotropic spectral (GBS) model (Qiao et al., 2005), and the horizontal velocity u and v are the derivatives of the geostrophic stream function ψ which is applied with vertically 1 layer for the troposphere, shown in Eq. (4), whereas ψ is solved in the Eq. (5). The initial condition for GBS model is the stream function, which is calculated based on 500 hPa wind vector u and v from the reanalysis of ERA5 hourly reanalysis data (https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era5: last access, July 17, 2020) of January 2020.

$$u = \frac{\partial\psi}{\partial y}, v = -\frac{\partial\psi}{\partial x} \tag{4}$$

$$\frac{\partial}{\partial t}\left(\nabla^2-\gamma^2\right)\psi + J\left(\psi,\nabla^2\psi\right) + \beta\frac{\partial\psi}{\partial x} + J(\psi,\mathrm{h}') = f_c, \tag{5}$$

where β represents the change rate of the Coriolis parameter f_0 at a certain latitude, f_c is the vorticity forcing and $\mathrm{h}' = (f_0/H_0)\mathrm{h}_{terrain}$ reflects the effect of topography determined by f_0, the average atmosphere equivalent depth H_0 (about 12,000 m in depths of the troposphere), and topography $\mathrm{h}_{terrain}$. The Cressman parameter $\gamma^2 = f^2/(gH_0)$ is a corrector of the systematical error in the barotropic atmosphere model where f represents the planetary vorticity (Rinne and Järvinen, 1993).

A leap-frog scheme is used in coding the main iteration module of the 2D transport model, which is assisted by a normal forward scheme for the start module. Gaussian grids system is popularly used and it is with a spatial resolution of 5.625° × 3.333° in longitude and latitude for GBS model and transport model in this research. Temporal resolution is 30 min (0.5 h). It is applied with vertically 1 layer for the troposphere with about 12,000 m in depths.

b. The tangent linear model (TLM)

Before constructing an adjoint model, it is necessary to derive the corresponding tangent linear model (Errico, 1997). Let $\boldsymbol{p}$ denote the vector of parameters in the transport model which has no correlation with the model state c, TLM, as the Taylor expansion of the nonlinear equation (Eq. (2)) and the first order approximation of the concentration perturbation c', can be written as below (Eqs. (6) and (7)) based on Eqs. (18) and (19) in Daescu and Carmichael (2003).

$$\frac{dc'}{dt} = M'_{\boldsymbol{c}}(c,\boldsymbol{p})\cdot c' + M'_{\boldsymbol{p}}(c,\boldsymbol{p})\cdot\boldsymbol{p}' \tag{6}$$

$$c'(t_0,\boldsymbol{x}) = c'_0(\boldsymbol{x}) \tag{7}$$

where $M_{\boldsymbol{c}}'(c,\boldsymbol{p})$ and $M_{\boldsymbol{p}}'(c,\boldsymbol{p})$ are the Jacobian matrices of M with respect to c or $\boldsymbol{p}$, respectively, and the Lagrangian differential operator in Eq. (2) has been expressed in the form of Eulerian differential operator in Eq. (6).

c. Evaluation of TLM

To evaluate the consistency between the tangent linear model and nonlinear transport model in computing a small perturbation, the nonlinear transport model was run with two cases, one with the initial conditions and the other with the initial conditions plus a small perturbation. The difference between these two cases is used to be divided by the TLM results with the same initial perturbation $\alpha\delta\boldsymbol{s}$, referred to as $R_{\tan}(\alpha)$ with the magnitude of α at the same order as the perturbation. Based on the illustration of Zhang et al. (2001) for tangent linear test, Eq. (8) can be derived.

$$\lim_{\alpha\to0}\mathrm{R}_{\tan}(\alpha) = \lim_{\alpha\to0}\frac{\|M(\boldsymbol{s}+\alpha\delta\boldsymbol{s})-M(\boldsymbol{s})\|}{\|\alpha M'\delta\boldsymbol{s}\|} = 1 \tag{8}$$

Here $\boldsymbol{s} = (c^T,\boldsymbol{p}^T)^T$ is the vector of all parameters in the transport model and $M' = \partial M/\partial\boldsymbol{s}$ represents the tangent linear operator of the transport model, and $\|\cdot\|$ denotes a l_2 Euclidean norm. Setting a horizontal point emission which holds a constant value of 5000 μg $m^{-2}\ s^{-1}$, i.e., the emitted strength is 0.416 μg $m^{-3}\ s^{-1}$ when divided by the depth of 12,000 m vertically in one atmosphere column, we produce the solid line of Fig. B in Supplementary materials by performing integrations of the nonlinear model and tangent model described above.

With α tending to be infinitesimal or zero, the perturbations computed by nonlinear transport model and tangent linear model should be comparable, indicating a value of 1 might be achieved for $R_{\tan}(\alpha)$. The accuracy was quantified based on the metric of $log(R_{tan}(\alpha) - 1)$ (solid line in Fig. B in Supplementary materials), and it shows that when the magnitude's order α of a small perturbation decreases from 10^{-1} to 10^{-16}, $log(R_{tan}(\alpha) - 1)$ starts to decrease from −0.8 until −6.5, indicating that the tangent linear increment is approaching to the nonlinear increment. Then $log(R_{tan}(\alpha) - 1)$ drops very slowly as α ranges from 10^{-6} to 10^{-10}, while it slightly increases when α continues to decrease. The increase of the logarithm value says the model errors derived with the magnitude's order α become comparable to the increment in magnitude. The solid line of Fig. B in Supplementary materials ensures the correctness of Eq. (8) for the tangent linear model, which gets it ready for the development of the adjoint model.

3.2. Adjoint model and sensitivity

The adjoint model is needed to derive the sensitivity information of a pollution event governed by a nonlinear dynamical system. The sensitivity information is represented by the first-order derivatives (or called gradient) of a defined objective functional that measures the pollution with respect to some pollution-associated control variables. Since the gradient represents the descending direction of the objective functional, it serves as the key parameter in implementing minimization. Here we develop the adjoint of the 2D transport model described in last section to estimate the source of pollution (Pudykiewicz, 1998). Similar as the concept of an adjoint operator for pollutant's transport process (Daescu and Carmichael, 2003), we use inner product associated with the transport model (Eqs. (2) to (5)) to express the adjoint problem. Denote $\langle \cdot, \cdot \rangle$ as the inner product, a cost function J is of the form

$$J = \langle \boldsymbol{w}c, c \rangle \tag{9}$$

where $\boldsymbol{w}$ is a weights' vector independent on the model state c. Corresponding to a perturbation of input parameters, one can get a perturbation form of the cost function with respect to a perturbed c' (denoting the derivative of J with respect to c) *as*

$$J' = \langle \boldsymbol{w}, c' \rangle \tag{10}$$

An adjoint variable $\boldsymbol{\lambda}(t)$ may be introduced and multiplied by Eq. (6), then Eq. (11) can be derived through the integration on $[t_0, t_n]$.

$$\int_{t_0}^{t_n} \langle \boldsymbol{\lambda}, \frac{dc'}{dt} \rangle dt = \int_{t_0}^{t_n} \langle M_c'^{T}(c, \boldsymbol{p})\boldsymbol{\lambda}, c' \rangle + \langle M_{\boldsymbol{p}}'^{T}(c, \boldsymbol{p})\boldsymbol{\lambda}, \boldsymbol{p}' \rangle dt \tag{11}$$

Integrated by parts and arrange the terms in Eq. (11), it is referred that

$$\langle \boldsymbol{\lambda},\ c' \rangle|_{t_0}^{t_n} = \int_{t_0}^{t_n} \langle M_c'^{T}(c, \boldsymbol{p})\boldsymbol{\lambda} + \frac{d\boldsymbol{\lambda}}{dt}, c' \rangle + \langle M_{\boldsymbol{p}}'^{T}(c, \boldsymbol{p})\boldsymbol{\lambda}, \boldsymbol{p}' \rangle dt \tag{12}$$

The adjoint problem of transport model Eq. (2) is given by

$$\frac{d\boldsymbol{\lambda}}{dt} = -M_{\boldsymbol{c}}'^{T}(c, \boldsymbol{p})\boldsymbol{\lambda} \tag{13}$$

with $\boldsymbol{\lambda}(t_k) = \boldsymbol{w}$ at the moment of t_k in the time integration. As long as a vector $\boldsymbol{\lambda}(t_k)$ in mathematics satisfies Eq. (13), it can be inferred from Eqs. (10), (12) and (13),

$$J' = \langle \boldsymbol{\lambda}(t_0), c_0' \rangle + \int_{t_0}^{t_k} \langle M_{\boldsymbol{p}}'^{T}(c, \boldsymbol{p})\boldsymbol{\lambda}, \boldsymbol{p}' \rangle dt \tag{14}$$

Therefore, the sensitivities associated with the cost function can be displayed in Eqs. (15) and (16).

$$\nabla_{\boldsymbol{c_0}} J = \boldsymbol{\lambda}(t_0) \tag{15}$$

$$\nabla_{\boldsymbol{p}} J = M_{\boldsymbol{p}}'^{T}(c, \boldsymbol{p})\boldsymbol{\lambda} \tag{16}$$

where $\nabla_{\boldsymbol{c_0}}(\cdot)$ and $\nabla_{\boldsymbol{p}}(\cdot)$ are used to express the first-order derivative (i.e. the gradient) of J with respect to the control variables c_0 and $\boldsymbol{p}$, or as the sensitivity of J with respect to c_0 and $\boldsymbol{p}$, respectively. In addition to serving as the decent direction in a minimization algorithm, the gradient carrying sensitivity information is also quite useful to understand which area (sensitive area) is important for the target pollution event. More discussions of sensitivity analysis for transporting air pollution are presented in Section 4.

The adjoint model is coded adopting the method of adjoint of "finite difference" (Giering and Kaminski, 1998; Sirkes and Tziperman, 1997). The adjoint of the transport model is coded through transposing all subroutines line-by-line in their tangent linear model. In forward time integration, we perform $M = M_1 M_2 \ldots M_{tn-1} M_{tn}$to derive variables of TLM such as $c_{tn}' = Mc_0' = M_1 M_2 \ldots M_{tn-1} M_{tn} c_0'$, while its transpose becomes $M^* = M^*_{t_n} M^*_{t_n - 1} \ldots M^*_2 M^*_1$ and variables in adjoint model are integrated backward in time. Thus, the adjoint problem of transport model in an inner product form such as $\langle Mc', Mc' \rangle = \langle c', M^* Mc' \rangle$ will be checked by the ratio $R_{adj} = \langle c'(t_n), c'(t_n) \rangle / \langle \boldsymbol{\lambda}(t_0), c_0' \rangle$. Using the leap-frog scheme, the magnitude's order of the ratio R_{adj} agrees to 15 precision in 240 hours integration (Table A in Supplementary materials) so as to guarantee the accuracy of the adjoint model, but a gradient test is necessary before the gradient is applied to a minimization algorithm, which will be discussed more details next.

3.3. Gradient test and minimization algorithm

a. Gradient test

In order to ensure that the results derived from an adjoint model represent the sensitivity of the corresponding transport model, a gradient test is necessary (Giering and Kaminski, 1998; Janisková and Lopez, 2013). Based on Giering and Kaminski (1998), the cost function J described in Eq. (10) can be further written in Eq. (18) measuring the distance between the model forecast and observations, and the gradient of J can be written in Eq. (19).

$$J = 0.5 \cdot \sum_{\boldsymbol{x}=(x,y)}^{\Omega} [c(\boldsymbol{s}, \boldsymbol{x}, t) - c_{obs}(\boldsymbol{x})]^T [c(\boldsymbol{s}, \boldsymbol{x}, t) - c_{obs}(\boldsymbol{x})] \tag{18}$$

where $c(\boldsymbol{s}, \boldsymbol{x}, t)$ is the modeled concentration derived from variables $\boldsymbol{s}$ at the moment of t, and $c_{obs}(\boldsymbol{x})$ is the observed concentrations, and $\boldsymbol{x}$ belongs to grid points in the target domain Ω.

$$\nabla_{\boldsymbol{p}} J = M^* [c(\boldsymbol{s}, \boldsymbol{x}, t) - c_{obs}(\boldsymbol{x})] \tag{19}$$

For each time step, based on Eq. (19), the output $c(\boldsymbol{s}, \boldsymbol{x}, t)$ from the transport model is used as the input of adjoint model for the backward calculation.

Starting from a single point emission with an arbitrary value, we integrate the transport model (Eqs. (2) and (5)) forward in time. Applying Eq. (18) to the target area of the pollution event, the North of China centered at Beijing in this study as 32°N–52°N in latitude, 103.25°E–136°E in longitude, we compute the cost function J. Following the procedure shown in Fig. A in Supplementary materials, using the first-order derivative of J with respect to the pollutant concentration in the pollution event over the target domain as the input of the adjoint, we integrate the adjoint model backward in time until the initial time. The output of adjoint model is the gradient of J with respect to initial concentrations and emissions. A gradient test procedure is used to ensure that the gradient of cost function calculated from the adjoint model is correct.

Similar to the tangent linear test expressed by Eq. (8), a small perturbation $\alpha\delta\boldsymbol{s}$ (α is the magnitude controller of the perturbation) is applied to perform the gradient test based on Zhang et al. (2001). We compare the increments computed from integrations of nonlinear model and the adjoint-derived gradient combined with the small perturbation. The nonlinear transport model is integrated twice, one with the control initial conditions $\boldsymbol{s}$ and the other with the perturbed initial conditions $\boldsymbol{s} + \alpha\delta\boldsymbol{s}$. A ratio is defined as the norm of the difference between two nonlinear model integrations divided by the norm of adjoint-derived increments, referred to as $R_{\mathrm{grd}}(\alpha)$,

$$\lim_{\alpha \to 0} R_{grd}(\alpha) = \lim_{\alpha \to 0} \frac{\|J[\boldsymbol{s} + \alpha\delta\boldsymbol{s}] - J[\boldsymbol{s}]\|}{\|\nabla_{\boldsymbol{s}} J \cdot \alpha\delta\boldsymbol{s}\|} = 1, \tag{20}$$

where $J[\cdot]$ denotes that the operator of cost function from the nonlinear model. Especially, we evaluate the gradient of cost function to emission source E_0, and the perturbation series $\alpha\delta E_0$ to produce the dashed line of Fig. B in Supplementary materials. With α ranges from 10^{-1} to 10^{-15}, the value of $log(R_{grd}(\alpha) - 1)$ starts to decrease to -7, and then increase

when α continues to decrease. The dashed line of Fig. B in Supplementary materials proves the correctness of the adjoint-derived gradient so that the system is ready to do the minimization.

b. Minimization algorithm

Taking Eq. (18) as an objective function (i.e. cost function), the optimization with respect to the variable vector $\boldsymbol{s}$ can be reached through an efficient minimization algorithm, the limited memory quasi-Newton line search method, which is characterized by fast convergence and calculation in dealing with the linear equations (Liu and Nocedal, 1989). As shown in Fig. A in Supplementary materials, in each iteration of the minimization approach, the transport model is first used to calculate the objective function, followed by the adjoint model integration to obtain the gradient, and the last step is to retrieve the minimized results through a series of optimization search procedures. The gradient test performed in the last section gives us sufficient confidence to apply the transport model and its adjoint to the minimization procedure.

Applying the minimization algorithm to the four experiments **Exp1, Exp2, Exp3** and **Exp4** described in Section 2.2, we obtain the behaviors of the normalized (divided by the value of the first iteration) cost function and norm of the gradient to emission in the space of iteration number in Fig. C in Supplementary materials. Although there are a few differences in local curves of the cost function and corresponding gradient norm, the trends of cost function values appear generally decreasing and all start to converge after 30 iterations during the minimization process (Fig. C(a–d) in Supplementary materials). At the same time, the curve trend of gradient norm for **Exp4** converges (in 34 iterations) faster than **Exp3** (in 50 iterations), while it converges faster for **Exp3** than **Exp2** (in 80 iterations). The convergence of gradient norm might be controlled by the dynamics of the transport process, in which more complex and realistic emission conditions are considered in **Exp4** than **Exp1, 2,** and **3**. Besides, in **Exp2, Exp3,** and **Exp4** (Fig. C(b–d) in Supplementary materials), the first 30 iterations make the cost function (blue) and the gradient (red) reduced over 99% and 95% respectively, but in **Exp1,** the cost function and the gradient decrease by 95% and 75% (Fig. C(a) in Supplementary materials) due to a much more suitable scale in minimization not found in our experiments. The cost function and its gradient finally maintain a stable level when the minimization reaches the convergence.

4. Optimization of initial conditions, emissions, and observational network

4.1. Sensitivity analyses

Sensitivities (i.e., the first-order derivatives, or called gradient) carry the information of how the targeted air pollution responds to variations of some pollution-associated control variables (previous pollutant concentrations and emissions, for instance). In this section, in order to prepare for the optimization process of attributing emission source, we perform sensitivity analyses for **Exp1–4** as described in Section 2.2. With the 2D transport model and its adjoint developed in Section 3, the sensitivities to emission sources are calculated. The minimization algorithm has been tested with respect to the air pollution event targeted to the North China ahead a few days (we test 5 days in this study) in Section 3.3b. Here we show the derived sensitivity distributions with respect to emission source in the converged iteration (110th iteration for all cases) in the minimization process (Fig. D in Supplementary materials). In general, a positive sensitivity in space means that the increasing emission in such an area enhances the targeted air pollution, while a negative sensitivity represents the sensitive area that has a contribution to the increasing deviation of a model concentration from the concentration of the air pollution event over the targeted domain. The minimization starts from the first guess of initial concentrations and emissions being zero. As shown in Fig. D in Supplementary materials, when the minimization comes to converging, in all **Exp1–4**, the consistent distribution of the convergent sensitivities ($\nabla_{\boldsymbol{EO}}J|_{EO=Eopt} = M^{*}[c_{opt} - c_{obs}(\boldsymbol{x})]$) is observed as positive values in the upstream and nearby. This suggests that possible influencing pathways of emission source might be constrained by the same dynamical mechanism of transport to the targeted area of air pollution event over the North China domain for **Exp1–4**. It's worth to note that idealized air pollution events are used for adjoint sensitivity analysis in this part as well as the development and analyses in Sections 4.2, 4.3 and 4.4 with this 2D transport model which is driven by a simple barotropic circulation model at the 500hpa isobaric surface. It's clear that although our ultimate goal of developing this methodology is for real air pollution prediction and control, the current 2D transport model needs to be upgraded to handle 3D transport and complex processes such as diffusion and chemistry etc. Applications in the real pollution control will be discussed more in Section 4.3 when the optimal observational network is analyzed.

With the sensitivities calculated in the converged iteration, the optimization of emissions (and optimal initial concentrations for **Exp4**) can be obtained and then cause the targeted air pollution over the North China domain after transport. Next, we will further discuss the details of the impact of optimal emission and optimal initial concentrations.

4.2. Optimized estimate of initial conditions of pollutant concentrations and emissions

Optimal initial concentrations and emissions are estimated by the minimization applied to **Exp4,** while applying minimization to **Exp1–3** only optimal emission can be estimated. In this section, we focus on the results of **Exp4** and analyze the impact of optimal initial concentrations and emissions on the target pollution event. Based on the $PM_{2.5}$ concentration in the air pollution event, the cost function is calculated, and the optimal initial concentrations and emissions are derived through the minimization algorithm as implemented by the iterative procedure in the illustration of Fig. A in Supplementary materials and the convergence is shown in Fig. C(d) in Supplementary materials. The distributions of optimally estimated initial concentrations and emissions are shown in Fig. 2. Presumably in the real world the areas with high emission might suffer high concentration of air pollutants. On the one hand, the underlying mechanism of the similarity is primarily attributable to the same dynamics of transport. On the other hand, the much lower initial condition and emission over the pollution area compared to the upwind region occur because of the lack of chemical reactions during transport process. It is admitted that the spatial distributions of the emission may yield large differences once the chemical reactions are taken into consideration. Nevertheless, the concept of the adjoint sensitivity still holds.

Driven by the optimized initial conditions and emissions shown in Fig. 2, the forward transport model is then integrated continuously for five-day (120 h) (assuming persistent emission), and the time evolution of $PM_{2.5}$ concentration for day 0–5 is shown in Fig. 3(a–f). Clearly, starting from the initial condition depicted in Fig. 3(a), the westerly wind drives the initial air pollutants as well as the emission to transport and accumulate eastward. At the end of five day (120 h) shown in Fig. 3(f), the concentration over the North of China (inside the red square) is used to compare to the preassigned air pollution event with $PM_{2.5}$ concentrations > 150 μg m^{-3}. The mean $PM_{2.5}$ concentration is 147.4 μg m^{-3}, with the values of mean fractional bias (MFB) and mean fractional error (MFE) at −2% and 6.5%, respectively, which satisfied the benchmark ($-60\% \le MFB \le 60\%$, $MFE \le 75\%$) for the MFB and MFE proposed by Boylan and Russell (2006).

The time evolution of adjoint-derived sensitivity to emission (Fig. 4) in general displays comparable spatial patterns relative to the concentration derived from the forward transport model using optimal initial concentrations and emissions particularly during hour 72–120 (compare Figs. 4(d–f) to 3(d–f)). Consistency of the distributions between

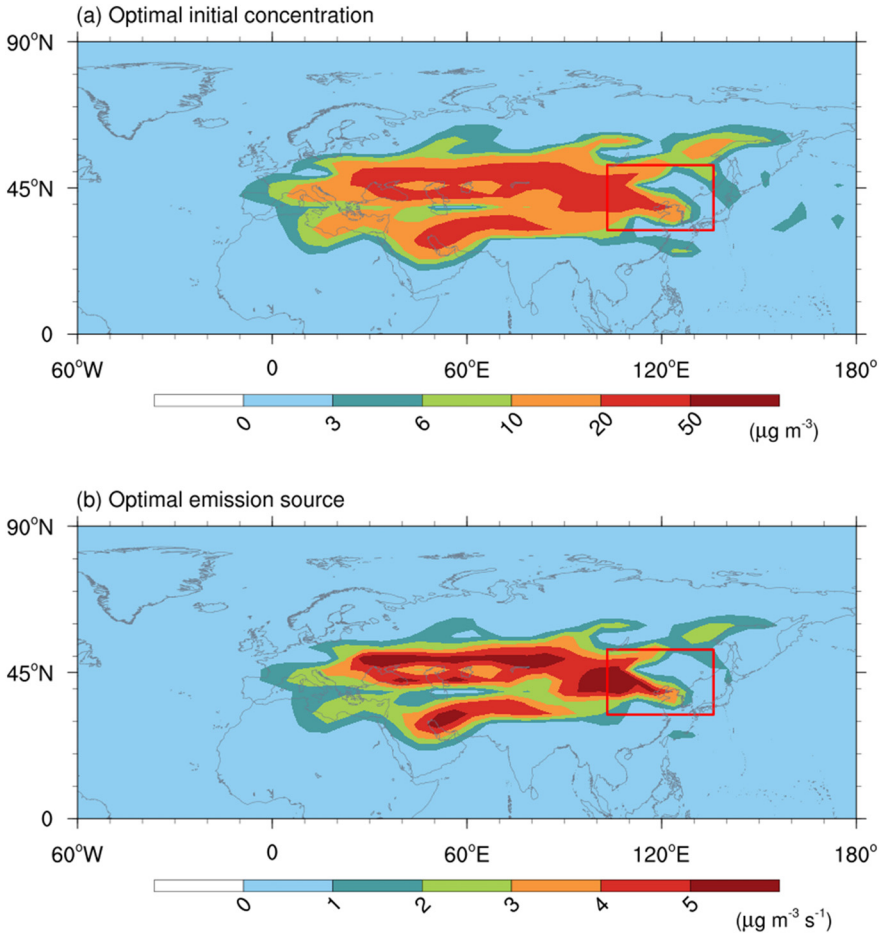


Fig. 2. Distributions of the optimal estimated a) initial $PM_{2.5}$ concentration and b) emission sources for the experiment that considering the emission source is persistent until the air pollution event occurs and the $PM_{2.5}$ pollutant has an initial concentration distribution (**Exp4**). The estimates are for the target air pollution event occurring over the North of China (red box) in 120 h after the initial time. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

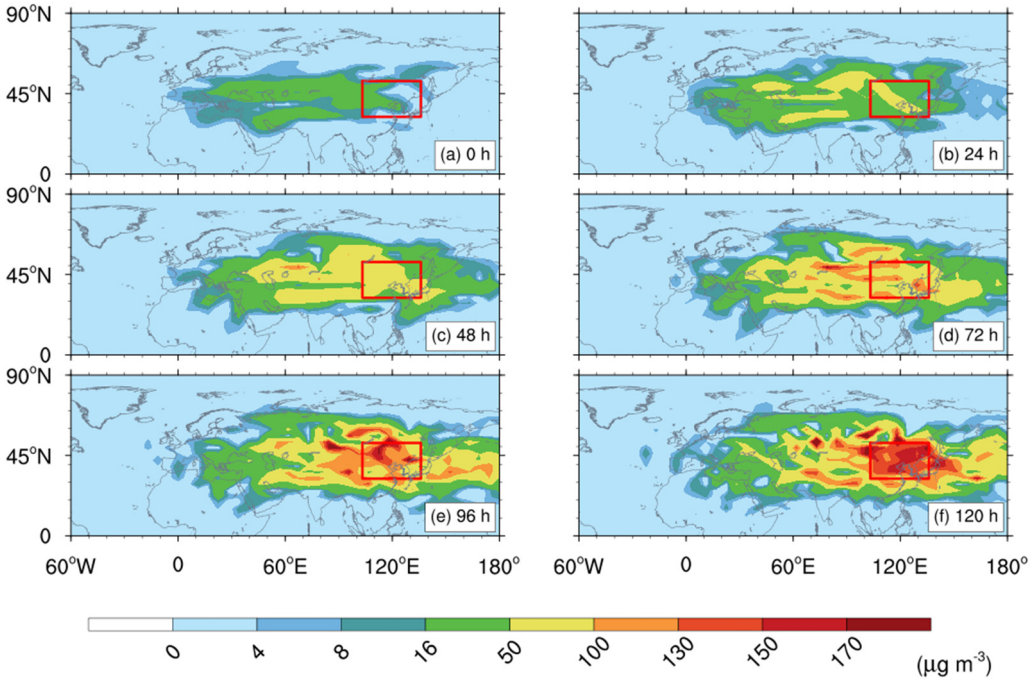


Fig. 3. Time evolution of the spatial distribution of $PM_{2.5}$ concentration (a-f) simulated by the forward transport model at a) initial time, b) 24-hour, c) 48-hour, d) 72-hour, e) 96-hour and f) 120-hour in **Exp4**.

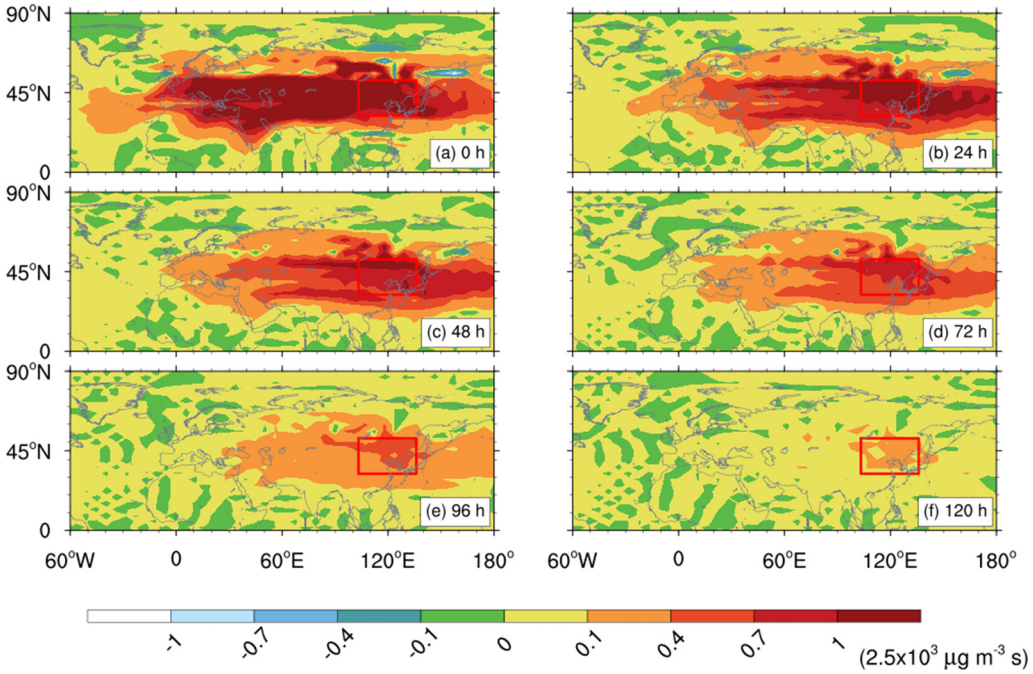


Fig. 4. The same as Fig. 3 but for the sensitivity with respect to the emission derived by integrating the adjoint backward.

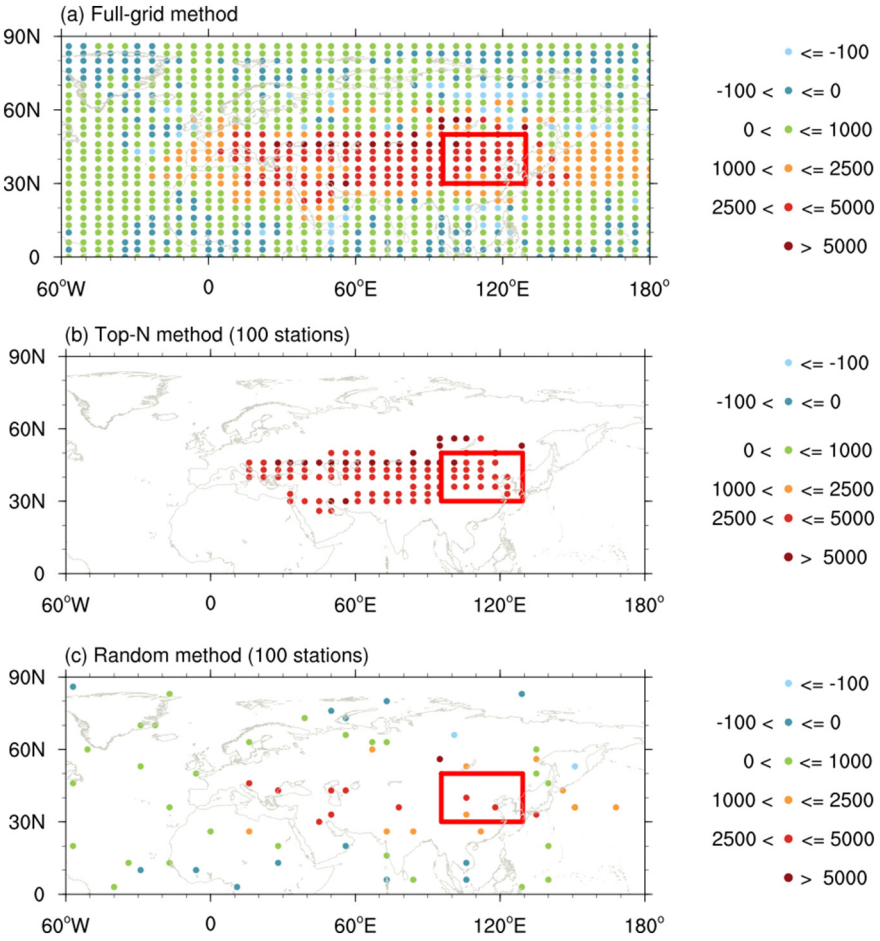


Fig. 5. Distributions of observing network based on a) full-grid method, b) top-N method (N set as 100), and c) randomly-sited 100 stations. Dots in each panel represent locations of observing sites, and colors represent the sensitivity (unit of μg m^{-3} s) to emission. That N equals to 100 is only for Fig. 5, but exactly, we conduct air pollution forecast using top-N method with N varying from 1 to 850 (and can be much more till the total number of model grids) for all other experiments.

sensitivity to emission and modeled concentration indicates that the adjoint model can efficiently distinguish the sensitive areas for the air pollution event over the North China, and it performs well ahead of 5 days. It is observed that spatial patterns of the sensitivity grow values at hours integrated from 120 to 0 backward, that is because the sensitivities between two adjacent integrations deliver values and added in the integrations of adjoint model.

4.3. Optimization of the observational network

Targeted to the air pollution event, we have estimated the optimal initial concentrations and emission sources in Section 4.2. In this section, we use the attributed source information to further design optimal deployment of observational stations in terms of two constraints, the cost and the effectiveness to form optimal observational network. Again, we only address an idealized situation in this study for the methodology development, which is readily implemented by a minimization algorithm. In the real world, an optimal observational network must consider more complex constraints such as such as monitoring areas of high population, measuring the maximum concentrations, or detecting the violations from ambient standards. Follow-up studies could implement such important constraints to address real pollution control issues.

Ideally, regardless of the cost, observations covering the entire targeted domain should yield the best prediction skills by providing the most accurate initial condition. In model simulations, this kind of observational network may include the total grid points of the model, and it is naturally named as the "full-grid method" as shown in Fig. 5(a). The sensitivities of the monitoring sites using the full-grid method distribute identically to the distribution depicted by Fig. 4(a). However, observational network designed in high cost is not sustainable in real application. The observational network cost is an important constraint in optimization. Here, the cost of observational network is defined as a linear function to the number of monitoring sites in the observational network, and we take the full-grid configuration as the reference for designing of the optimal observational network, in which the fraction of grids used over the full grids (850 in this case) is a measure of the cost. Therefore, the essential role of optimizing the observation sites is to improve the accuracy of air quality forecast with a reasonable cost in the construction of the observational network.

As in discussions on Fig. 4, the adjoint sensitivity to emission reflects sensitive areas in contribution to the targeted air pollution, and the locations with the higher sensitivity in general play larger roles in constraining the concentration over the designated receptor region. We then introduce the straight-forward way in distributing the monitoring sites about sensitivity, which is referred to as the "top-N method". The top-N method means that a number of N most sensitive observational sites is established to initialize the concentration and set the emission for the transport model run, and N can be any number from 1 to the maximum as the total number of domain grids. For the demonstration purpose, the sites (grids) with the highest 100 (Note that N equals to 100 is only for Fig. 5, but exactly, we conduct air pollution forecasts using top-N method with N varying from 1 to 850 for all other experiments) sensitivity values to emission are selected and shown in Fig. 5(b). Alternatively, a random method (Fig. 5(c)) was applied to compare with the top-N method using the same number of sites. Note that for the random method, the locations of the observation should be in the Northern Hemisphere as the targeted air pollution is under the influence of prevailing westerly wind around 35°N–65°N, and in Fig. 5(c) only a part of the region in Norther Hemisphere is shown and that's why the sites look like a little less than 100.

Based on the three methods discussed above, a series of numerical experiments are conducted to evaluate how efficient the initial concentrations and emissions observed from observational network is in affecting the air pollution forecast. First, the forecasts with the lead times of 5 days (120 h), 7 days (168 h), or 10 days (240 h) are simulated for three methods, respectively. For the full-grid method, the forecasts are conducted in the same way as that in Section 4.2 (**Exp4**) because of every grid point serving as a monitoring site. For the top-N method, a total of five ensemble members are used. The reason to use ensemble is the consideration of uncertainties in the initial conditions of the forecasts when modeled sensitivity is adopted. The ensemble members derive optimal initial concentrations and emissions via the drift forecast time of 2 or 4 h prior to or after the designated forecast hours. For example, the ensemble members of 5 days forecast use five groups of initial concentrations and emissions independently, which are optimally estimated ahead 116-hour, 118-hour, 120-hour, 122-hour and 124-hour of the air pollution event occurring at time T. Likewise, the drift forecast times are applied to the forecast of 7 days and 10 days in the same way. For the random method, the ensemble members are used too, but the members are designed by randomly distributing the sites five times for a selected number of sites.

RMSEs between the modeled concentrations and the air pollution event over the domain of North China are calculated. As shown in Fig. 6, the dashed lines indicate RMSE of 16.97 μg m^{-3} (black), 26.30 μg m^{-3} (purple) and 42.79 μg m^{-3} (yellow) for the forecast of

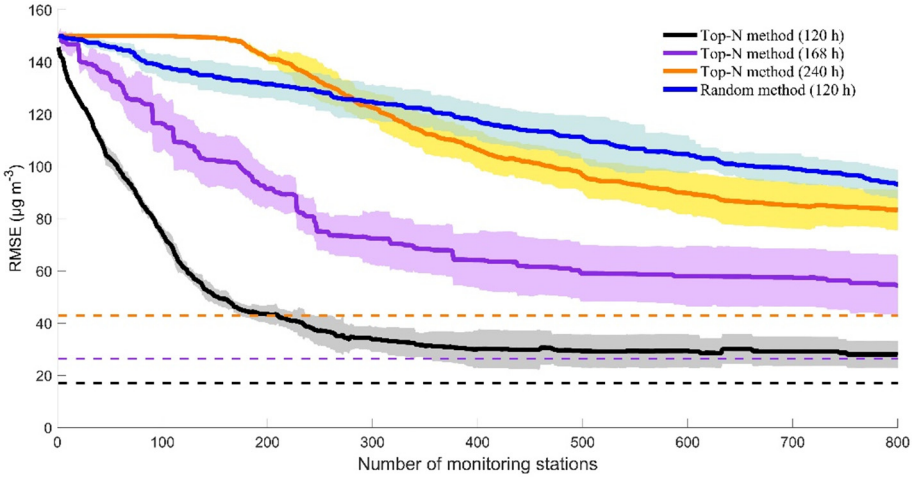


Fig. 6. The RMSE between the modeled and observed concentrations over North China along with number of observing stations representing different initial conditions, based on the top-N method with the forecast time of 120 h (black-solid), 168 h (purple), and 240 h (yellow), respectively, as well as a forecast of 120-hour with the random method (blue-dashed). For each method, the ensemble mean (wide-solid line) and standard deviation (shaded) of five members are shown. The black-, purple-, and yellow-dashed lines represent the RMSE of the forecast of 120-hour, 168-hour and 240-hour, respectively, driven by the initial conditions and persistent emissions using the full-grid observing system. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

5 days, 7 days and 10 days, respectively, which are driven by the initial conditions using full-grid method. The increase of RMSE along with the increment of lead time length in general indicates chaos in meteorology (Lorenz, 1963). The top-N method in general shows a decreasing trend by enlarging the number of monitoring stations, meaning that the more observational sites taken into optimization, the smaller forecast RMSE is expected. Taking the forecast of 120 h as an example, relatively sharp decreases of RMSE are observable at the beginning when the number of monitoring sites increases, and the slope becomes smaller once the number increases to a certain value, e.g., 300, with the RMSE reduced by 116.26 μg/m^3 (accounting for 87.35% in comparison with the full grid method). The random method, nevertheless, shows much higher RMSE compared to top-N method for the same forecast length (120 h) even when the number of monitoring sites increase to a relatively large value such as 500 or more, and the RMSE is reduced by only 25.14 μg/m^3 (accounting for 18.89% in comparison with the full gird) as 300 monitoring sites are used. For the forecast of 120 h, considering the number of monitoring sites as 300 as a transition point (one increases cost but not apparently improving efficiency) to a large degree, the time evolution of $PM_{2.5}$ spatial distribution of the concentration is shown in Fig. E in Supplementary materials. Compared to the baseline depicted in Fig. 2(a, b, d, f) for the hour of 0, 24, 72 and 120, the top-N method clearly generates the stronger eastward propagation (Fig. E (a–d) in Supplementary materials) compared to random method (Fig. E(e–h) in Supplementary materials), and at the hour of 120, the spatial distribution of $PM_{2.5}$ concentration based on top-N method is quite comparable to that based on the full-grid method (Fig. E(d) in Supplementary materials vs. Fig. 3(f)).

Initialized with the modeled data derived from the observational network based on the top-N method, the air pollution forecasting experiments are performed. We find that the observational network using the top-N method is with fewer number of monitoring sites, but it performs in high efficiency for accurate pollution forecast. This kind of observational network is referred to as the optimal observational network because it provides useful advanced information in the economic fashion which is due to reasonable distributions of monitoring sites.

4.4. Impacts of the optimal observational network on air pollution predictability

As the two questions illustrated in Section 2.1, we obtain the pathway of pollutant transport by solving the sensitivity to emissions from adjoint model (the answer to the first question) in Section 4.1. We deploy the monitoring sites with respect to sensitivity by minimizing the cost function of the targeted air pollution (answering the second question of Section 2.1) in Sections 4.2 and 4.3. Nevertheless, the impacts of the optimal observational network on air pollution predictability still remain unclear. This part is therefore inspired on the limit of air pollution predictability with observations from the optimal observational network using the top-N method based on optimal initial concentrations and emissions (**Exp4**).

We first plot variations of the averaged concentrations of the pollution forecast over the targeted domain during the forecast time [0, T] where T is 240 h, and three independent air pollution events described in Section 4.3 are shown in the panels of Fig. 7. The three air pollution events are assumed occurring over the North China at the time t_k ($\in[0,T]$) of 120 h, 168 h, and 240 h, respectively, and the optimal initial concentrations and emissions are adopted. Apparently in Fig. 7, the air pollution events are observed as scheduled on time as the peaks at 120 (Fig. 7 upper-thin), 168 (Fig. 7 medium-thin), and 240 h (Fig. 7 bottom-thin) indicating that the reference forecasts are credible based on optimal initial concentrations and emissions. However, the modeled concentrations for the forecast of the targeted pollution occurring at the time 120 h (Fig. 7 upper-thick), which starts with the observations of the observational network using the top-N method, displays more accurately to the reference concentrations (Fig. 7 upper-double-dashed) than that at the time 168 h (Fig. 7 medium-thick) and 240 h (Fig. 7 bottom-thick). We feel that air pollution predictability acts to be limited as the lead time length of the targeted pollution forecast increase, for which the optimal observational network provides the initial conditions for the forecasts.

In order to have a deep look at the impacts of the optimal observational network on the limit of air pollution predictability, we investigate RMSEs in measuring the difference between the modeled concentrations and the reference concentrations of the targeted air pollution. First, we estimate a series of optimal initial concentrations and emissions with respect to the targeted air pollution event occurring at the time t_k of 12 h, 24 h, …, 228 h, and 240 h during the forecast time [0, T], i.e., the air pollution events in Fig. 7 are three subcases of this series. Then, based on observations of the observational networks using the full-grid method and the top-N method, we use the optimal estimates of each case in the series, respectively, to derive the modeled concentrations. Therefore, several forecasts with respect to the targeted air pollution in different lead time lengths (say, 20 subcases) are obtained, and RMSE of each forecast is calculated at the lead time t_k for the targeted air pollution, respectively. For the demonstration purpose, the sites (grids) with the highest 300 (inspired by the discussions in Section 4.3) sensitivity values to emission are selected for the top-N method in this part. Thus, we have two plots of RMSEs for the targeted air pollution forecasts with the varying lead time lengths, one is based

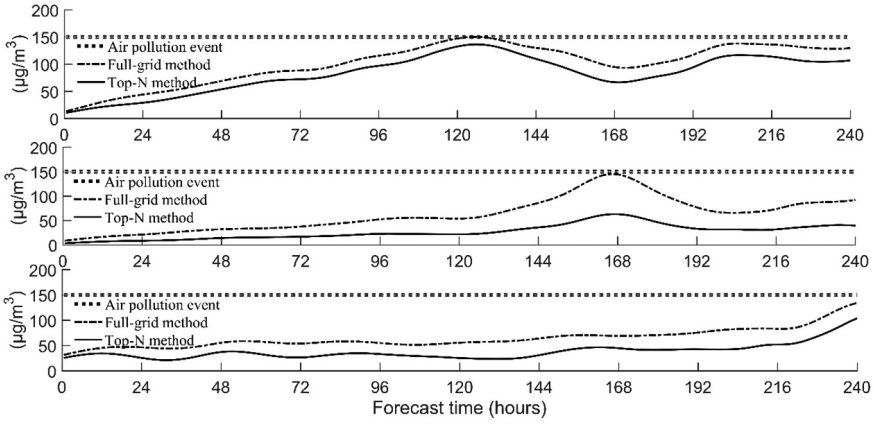


Fig. 7. Variations of the averaged concentration over the targeted domain for the 240 h forecasts, using the full-grid method (thin line), and top-N method (thick line) in each panel based on the optimal initial concentrations and emissions, respectively. The sites (grids) with the highest 300 sensitivity values to emission are selected for the top-N method, which is inspired by the discussions in Section 4.3. The upper panel uses the optimal initial concentrations and emissions with regard to the targeted air pollution event occurring at the time $t_k \in [0,T]$ where t_k is 120 h. The medium and bottom panels are the same as the upper panel but they are with respect to the pollution event occurring at the time where t_k equals to 168 h, and 240 h, respectively. Dotted lines are the reference concentrations of the $PM_{2.5}$ air pollution event with the value of 150 μg m^{-3}.

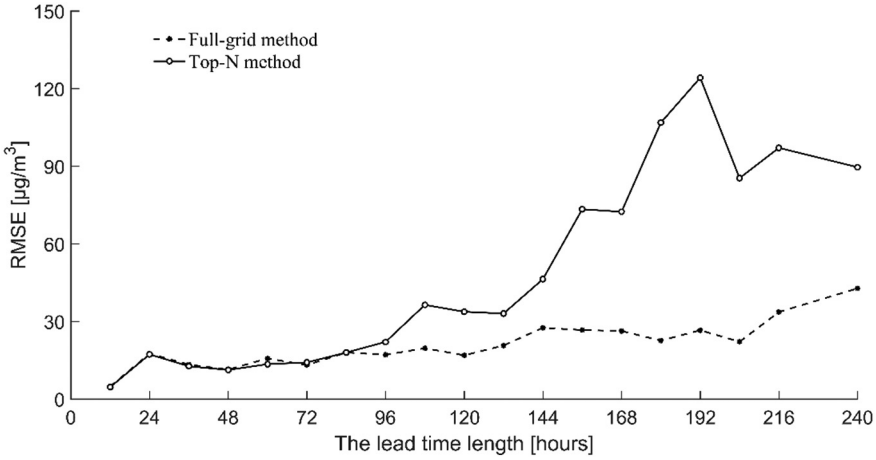


Fig. 8. RMSEs of air pollution forecasts with the lead time lengths in measuring the difference between the modeled concentrations and the reference concentrations of the targeted air pollution. Solid line plot represents the results based on the optimal network using the top-N method, and dashed line plot is based on the observational network using the full-grid method. The sites (grids) with the highest 300 sensitivity values to emission are selected for the top-N method, which is inspired by the discussions in Section 4.3. Note that the RMSEs are calculated at the scheduled time $t_k \in [0,T]$ when the targeted air pollution event occurs for the forecasts in each lead time length. For example, the RMSEs of dots at 120 h, 168 h, 240 h are for the peaks of the upper, medium and bottom panels in Fig. 7 at the time t_k in 120 h, 168 h, and 240 h, respectively.

on observations of the observational network using full-grid method (Fig. 8 dashed) and another using top-N method (Fig. 8 solid).

Though RMSEs described above using the full-grid method (which is equivalent to **Exp4**) could be in fluctuation for each case (as shown as the drifts of the dashed line in Fig. 8), which is subject to the different choices of scales in optimally estimating initial concentrations and emissions during the minimization process, the difference of RMSEs between the top-N method and the full-grid method (Fig. 8 solid vs. dashed) reflects the ability of the optimal observational network (which is based on sensitivity analysis) to impact the accuracy of prediction results, we refer it to as air pollution predictability here. We find that the two plots in Fig. 8 are close to each other in the lead time lengths from 12 h to 84 h indicating that the air pollution predictability is seized well, but then they slowly split up to a large degree, especially after 144 h (6 days). The limit of air pollution predictability in the forecasts based on the optimal observational network might be concluded to be in the time scale about 6 days as a result, in which the difference of RMSEs of the two plots is less than 20 μg m^{-3} (accounting for 14% in comparison with 150 μg m^{-3}) and is acceptable in accurate air pollution forecast.

5. Summary and discussion

Using the framework of 4-dimensional variational approach, the issue of air pollution predictability based on optimal observational network has been studied. In terms of optimal attribution of the air pollution sources, the first problem is to make it clear that the possible transport pathway of the pollution sources in the past time with respect to the targeted air pollution region. The answer is to calculate a cost function with the adjoint sensitivity analysis. The second problem is to design an observing network to measure initial concentrations and emissions, so as that accurate prediction will be realized. The answer is to solve a minimization problem and the general distribution of initial concentrations and emissions for an air pollution event are obtained. Four experiments are designed for optimally attributing the sources of pollution, in which pollution scenarios from the simplest to the more realistic are developed and examined. For the purpose of illustration, the framework of a 2D transport model and its adjoint is developed, and the transport model is driven by a global barotropic spectral model. From nonlinear equations to tangent linear derivation, and the adjoint version, the transport model is critically examined for the forecast lead time of 120 h. The gradient test ensures the representation of adjoint result consistent to the sensitivity of control variables in the model. Through an algorithm using the limited memory quasi-Newton method, the initial conditions of concentrations and emission sources are optimally estimated by the minimizations.

With the 4D-Var, we propose the optimized observational network with low cost and in high forecast skills. An idealized air pollution event of $PM_{2.5}$ occurring in the North of China is studied using the adjoint sensitivity analysis, and the optimal observational network is designed and discussed. Results show that the distribution of air pollution in forward time integration is well consistent to the adjoint sensitivity in backward time integration, and it indicates that the monitoring sites of observational network can be efficiently located according to the adjoint-derived sensitivity information. Results show that within some lead time scope, an excellent pollution forecast skill can be reached with the derived optimal observational network.

This study provides the relevant evidence from the emission-based approach to design optimal observational network in terms of cost-effectiveness which can inspire further efforts to resolve real pollution control problem, but it is preliminary and limited to an idealized framework.

The target pollutant in the 2D transport model is treated for only one pollutant, PM2.5, which is taken as an example in our current study. However, pollutants like PM10, CO and black carbons can also be adopted if the parameters of diffusion and deposition are applied in Eq. (2). In addition, an observing network is usually designed to observe multiple pollutants (Spangl et al., 2007), and the current method with the adjoint sensitivity analysis shall be modified to design the observing network for multiple pollutants. First, the cost function is calculated independently for each pollutant and the procedure will not be affected, but it is expected that a group of sensitivity fields are to be obtained for multiple pollutants. As the chemistry reactions are taken into account, the sensitivities calculation in adjoint model for multiple pollutants may become much more complex. Second, the optimization procedure of the current method needs to modify for identifying the area representativeness of monitoring sites and gaining a general observational network for all target pollutants. For example, we may develop the Combinatorial Optimization Problem (COP) in the optimization procedure, in which the Collaborative Optimization (CO) will be applied in the basis of the current method. The CO-approach could work for multi-objective problems and is widely used for multidisciplinary system designing (Tappeta and Renaud, 1997). In general, to design the observing network for multiple pollutants, the optimization could be implemented with some additional collaborative constraints and carried out in the follow-up studies.

Efforts on practical applications of the methodology in 3D numerical model are being carried using Community Multiscale Air Quality Modeling (CMAQ) and CMAQ-adjoint as well as Weather Research and Forecast (WRF) model, in which the missing chemical processes are included. Based on such a 3-dimensional atmosphere model that includes realistic transport dynamics and chemical reactions, with the development of minimization procedure and data assimilation scheme, an advanced pollution forecast system can be established.

CRediT authorship contribution statement

Liu Caili: Data curation, Modeling, Investigation, Roles/Writing - original draft. **Zhang Shaoqing:** Conceptualization, Methodology, Investigation, Funding acquisition, Writing - review & editing. **Gao Yang:** Methodology, Funding acquisition, Writing - review & editing. **Wang Yuhang:** Conceptualization, Investigation, Funding acquisition, Writing - review & editing. **Sheng Lifang:** Writing - review & editing. **Gao Huiwang:** Writing - review & editing. **J. C. H. Fung:** Writing - review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This research work is supported by the National Key Research and Development Program of China (2017YFC1404100 and 2017YFC1404104) and the National Natural Science Foundation of China (Grant No. 41775100, 41830964) as well as Shandong Province's "Taishan" Scholar Project and Qingdao "Creative and Initiative" frontier Scientist Program. Y. Wang was supported by the National Science Foundation Atmospheric Chemistry Program. This research is also supported by the Center for High Performance Computing and System Simulation, Pilot National Laboratory for Marine Science and Technology (Qingdao).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scitotenv.2021.145580.

References

An, X.Q., Zhai, S.X., Jin, M., Gong, S., Wang, Y., 2016. Development of an adjoint model of GRAPES–CUACE and its application in tracking influential haze source areas in north China. Geosci. Model Dev. 9 (6), 2153–2165. https://doi.org/10.5194/gmd-9-2153-2016.

Boylan, J.W., Russell, A.G., 2006. PM and light extinction model performance metrics, goals, and criteria for three-dimensional air quality models. Atmos. Environ. 40 (26), 4946–4959. https://doi.org/10.1016/j.atmosenv.2005.09.087.

Cacuci, D.G., 1981. Sensitivity theory for nonlinear systems. I. Nonlinear functional analysis approach. J. Math. Phys. 22, 2794–2802. https://doi.org/10.1063/1.525186.

Chen, D., Liu, Z., Ban, J., Chen, M., 2019. The 2015 and 2016 wintertime air pollution in China: SO2 emission changes derived from a WRF-Chem/EnKF coupled data assimilation system. Atmos. Chem. Phys. 19 (13), 8619–8650. https://doi.org/10.5194/acp-19-8619-2019.

Daescu, D.N., Carmichael, G.R., 2003. An adjoint sensitivity method for the adaptive location of the observations in air quality modeling. J. Atmos. Sci. 60(2), 434–450. https://doi.org/10.1175/1520-0469(2003)060<0434:aasmft>2.0.co;2.

Errico, R.M., 1997. What is an adjoint model? Bull.amer.meteor.soc. 78(11), 2577–2591. https://doi.org/10.1175/1520-0477(1997)078<2577:wiaam>2.0.co;2.

Flemming, J., Stern, R., Yamartino, R., 2005. A new air quality regime classification scheme for O, NO, SO and PM10 observations sites. Atmos. Environ. 39 (33), 6121–6129. https://doi.org/10.1016/j.atmosenv.2005.06.039.

Gao, Y., Shan, H., Zhang, S., Sheng, L., Li, J., Zhang, J., et al., 2020. Characteristics and sources of PM2.5 with focus on two severe pollution events in a coastal city of Qingdao, China. Chemosphere 247, 125861. https://doi.org/10.1016/j.chemosphere.2020.125861.

Giering, R., Kaminski, T., 1998. Recipes for adjoint code construction. ACM Trans. Math. Softw. 24 (4), 437–474. https://doi.org/10.1145/293686.293695.

Hakami, A., Seinfeld, J.H., Chai, T., Tang, Y., Carmichael, G.R., Sandu, A., 2006. Adjoint sensitivity analysis of ozone nonattainment over the continental United States. Environ. Sci. Technol. 40 (12), 3855–3864. https://doi.org/10.1021/es052135g.

Hakami, A., Henze, D.K., Seinfeld, J.H., Singh, K., Sandu, A., Kim, S., et al., 2007. The adjoint of CMAQ. Environ. Sci. Technol. 41 (22), 7807–7817. https://doi.org/10.1021/es070944p.

Henry, R.C., 2008. Locating and quantifying the impact of local sources of air pollution. Atmos. Environ. 42 (2), 358–363. https://doi.org/10.1016/j.atmosenv.2007.09.039.

Janisková, M., Lopez, P., 2013. *Linearized Physics for Data Assimilation at ECMWF* (Vol. II). Springer, Berlin Heidelberg.

Joly, M., Peuch, V.-H., 2012. Objective classification of air quality monitoring sites over Europe. Atmos. Environ. 47, 111–123. https://doi.org/10.1016/j.atmosenv.2011.11.025.

Kampa, M., Castanas, E., 2008. Human health effects of air pollution. Environ. Pollut. 151 (2), 362–367. https://doi.org/10.1016/j.envpol.2007.06.012.

Karamchandani, P., Long, Y., Pirovano, G., Balzarini, A., Yarwood, G., 2017. Source-sector contributions to European ozone and fine PM in 2010 using AQMEII modeling data. Atmos. Chem. Phys. 17 (9), 5643–5664. https://doi.org/10.5194/acp-17-5643-2017.

Kracht, O., and Gerboles, M. (2019). Spatial representativeness evaluation of air quality monitoring sites by point-centred variography. International Journal of Environment and Pollution, 65(1/2/3), 229-245. doi:10.1504/IJEP.2019.101843.

Kracht, O., Martin, F., Santiago, J., Piersanti, A., Vitali, L., Delaney, K., et al., 2017. Spatial representativeness of air quality monitoring sites: Outcomes of the FAIRMODE/AQUILA intercomparison exercise. EUR. 28987. EN, Publications Office of the European Union, Luxembourg. https://doi.org/10.2760/60611.

Kwok, R.H.F., Napelenok, S.L., Baker, K.R., 2013. Implementation and evaluation of PM2.5 source contribution analysis in a photochemical model. Atmos. Environ. 80, 398–407. https://doi.org/10.1016/j.atmosenv.2013.08.017.

Lee, E., Chan, C.K., Paatero, P., 1999. Application of positive matrix factorization in source apportionment of particulate pollutants in Hong Kong. Atmos. Environ. 33 (19), 3201–3212. https://doi.org/10.1016/s1352-2310(99)00113-2.

Liu, D.C., Nocedal, J., 1989. On the limited memory BFGS method for large scale optimization. Math. Program. 45 (1), 503–528. https://doi.org/10.1007/bf01589116.

Liu, F., Zhang, Y., Su, H., and Hu, J. (2007). Adjoint Model of Atmospheric Chemistry Transport Model CAMx:Construction and Application (in Chinese). *Acta Scientiarum Naturalium Universitatis Pekinensis, 43*(6), 764-770. doi:10.13209/j.0479-8023.2007.122.

Lorenz, 1963. Deterministic nonperiodic flow. J. Atmos. Sci. 20 (2), 130–141. https://doi.org/10.1175/1520-0469(1963)020<0130:dnf>2.0.co;2.

Marmur, A., Unal, A., Mulholland, J.A., Russell, A.G., 2005. Optimization-based source apportionment of PM2.5 incorporating gas-to-particle ratios. Environ. Sci. Technol. 39 (9), 3245–3254. https://doi.org/10.1021/es0490121.

Menut, L., Vautard, R., Beekmann, M., Honoré, C., 2000. Sensitivity of photochemical pollution using the adjoint of a simplified chemistry-transport model. J. Geophys. Res. Atmos. 105 (D12), 15379–15402. https://doi.org/10.1029/1999jd900953.

MEPPRC, 2012. Technical regulation on ambient air quality index (on trial): HJ 633-2012. Ministry of Environmental Protection of the People's Republic of China (MEPPRC). China Environmental Science Press, Beijing, China Retrieved from. http://www.mee.gov.cn/ywgz/fgbz/bz/bzwb/jcffbz/201203/W020120410332725219541.pdf.

Mizzi, A.P., Arellano Jr., A.F., Edwards, D.P., Anderson, J.L., Pfister, G.G., 2016. Assimilating compact phase space retrievals of atmospheric composition with WRFChem/DART: a regional chemical transport/ensemble Kalman filter data assimilation system. Geosci. Model Dev. 9, 965–978. https://doi.org/10.5194/gmd-9-965-2016.

Mu, M., Wang, J., 2001. Nonlinear fastest growing perturbation and the first kind of predictability. Sci. China Ser. D 44 (2), 1128–1139. https://doi.org/10.1007/BF02906869.

Park, S.-Y., Park, C., Yoo, J.-W., Lee, S.-H., Lee, H.W., 2018. Adjoint sensitivity of inland ozone to its precursors and meteorological and chemical influences. Atmos. Environ. 192, 104–115. https://doi.org/10.1016/j.atmosenv.2018.08.006.

Pérez-Roa, R., Castro, J., Jorquera, H., Pérez-Correa, J.R., Vesovic, V., 2006. Air-pollution modelling in an urban area: correlating turbulent diffusion coefficients by means of an artificial neural network approach. Atmos. Environ. 40 (1), 109–125. https://doi.org/10.1016/j.atmosenv.2005.09.032.

Pudykiewicz, J.A., 1998. Application of adjoint tracer transport equations for evaluating source parameters. Atmos. Environ. 32 (17), 3039–3050. https://doi.org/10.1016/S1352-2310(97)00480-9.

Qiao, F., Zhang, S., Yin, X., 2005. Study of initial vorticity forcing for block onset by a 4-dimensional variational approach. Adv. Atmos. Sci. 22 (2), 246–259. https://doi.org/10.1007/BF02918514.

Rinne, J., Järvinen, H., 1993. Estimation of the Cressman term for a barotropic model through optimization with use of the adjoint model. Mon. Weather Rev. 121 (3), 825–833. https://doi.org/10.1175/1520-0493(1993)121<0825:Eotctf>2.0.Co;2.

Sharma, V.K., 2009. Importance of source apportionment modeling in air pollution abatement policy. Clean Techn. Environ. Policy 12 (1), 3–4. https://doi.org/10.1007/s10098-009-0270-z.

Sirkes, Z., Tziperman, E., 1997. Finite difference of adjoint or adjoint of finite difference? Mon. Weather Rev. 125 (12), 3373–3378. https://doi.org/10.1175/1520-0493(1997)125<3373:fdoaoa>2.0.co;2.

Song, Y., Zhang, Y., Xie, S., Zeng, L., Zheng, M., Salmon, L.G., et al., 2006. Source apportionment of PM2.5 in Beijing by positive matrix factorization. Atmos. Environ. 40 (8), 1526–1537. https://doi.org/10.1016/j.atmosenv.2005.10.039.

Spangl, W., Schneider, J., Moosmann, L., Nagl, C., 2007. Representativeness and Classification of Air Quality Monitoring Stations, Umweltbundesamt Report. Retrieved from. https://ec.europa.eu/environment/air/quality/legislation/pdf/report_uba.pdf.

Sweileh, W.M., Al-Jabi, S.W., Zyoud, S.H., Sawalha, A.F., 2018. Outdoor air pollution and respiratory health: a bibliometric analysis of publications in peer-reviewed journals (1900 - 2017). Multidiscip. Respir. Med. 13, 15. https://doi.org/10.1186/s40248-018-0128-5.

Tappeta, R.V., Renaud, J.E., 1997. Multiobjective Collaborative Optimization. ASME. J. Mech. Des. 119 (3), 403–411. https://doi.org/10.1115/1.2826362.

Venkatesan, P., 2016. Ambien air pollution: air pollution is a major threat to health (WHO report). Retrieved from. https://doi.org/10.1016/S2213-2600(16)30014-5.

Wagstrom, K.M., Pandis, S.N., Yarwood, G., Wilson, G.M., Morris, R.E., 2008. Development and application of a computationally efficient particulate matter apportionment algorithm in a three-dimensional chemical transport model. Atmos. Environ. 42 (22), 5650–5659. https://doi.org/10.1016/j.atmosenv.2008.03.012.

Wang, Y.Q., Zhang, X.Y., Draxler, R.R., 2009. TrajStat: GIS-based software that uses various trajectory statistical analysis methods to identify potential sources from long-term air pollution measurement data. Environ. Model. Softw. 24 (8), 938–939. https://doi.org/10.1016/j.envsoft.2009.01.004.

Zhang, L., Shao, J., Lu, X., Zhao, Y., Hu, Y., Henze, D.K., et al., 2016. Sources and processes affecting fine particulate matter pollution over North China: an adjoint analysis of the Beijing APEC period. Environ. Sci. Technol. 50 (16), 8731–8740. https://doi.org/10.1021/acs.est.6b03010.

Zhang, S., Zou, X., Ahlquist, J.E., 2001. Examination of numerical results from tangent linear and adjoint of discontinuous nonlinear models. Mon. Weather Rev. 129, 2791–2804. https://doi.org/10.1175/1520-0493(2001)129<2791:EONRFT>2.0.CO;2.

Zhao, X.-f., Wang, T.-j., Huang, M.-t., Yang, F., Yang, D.-d., Jin, L.-s., 2019. Comparison of methods for dry deposition velocity and flux estimation of atmospheric pollutants—a case study in Xianlin District, Nanjing (in Chinese). Equip. Environ. Eng. 16 (6), 129–137. https://doi.org/10.7643/issn.1672-9242.2019.06.020.

Zhu, J., Zeng, Q., 2002. A mathematical theory framework for air pollution control (in Chinese). Sci. China Ser. D 32 (10), 864–870. https://doi.org/10.3969/j.issn.1674-7240.2002.10.011.