

# Potential Source Density Function: A New Tool for Identifying Air Pollution Sources

In Sun Kim<sup>1,2</sup>, Yong Pyo Kim<sup>3</sup>, Daehyun Wee<sup>1\*</sup>

<sup>1</sup> Department of Environmental Science and Engineering, Ewha Womans University, Seoul 03760, Korea

<sup>2</sup> Climate and Air Quality Research Department, National Institute of Environmental Research, Incheon 22689, Korea

<sup>3</sup> Department of Chemical Engineering and Materials Science, System Health & Engineering, Ewha Womans University, Seoul 03760, Korea

## ABSTRACT

Potential source density function (PSDF) is developed to identify, that is, locate and quantify, source areas of ambient trace species based on Gaussian process regression (GPR), a machine-learning technique. The PSDF model requires backward trajectories and sampling data at a receptor site in the calculation as in the conventional model to locate source areas of ambient trace species, such as the potential source contribution function (PSCF). The PSDF model can identify source areas quantitatively and provide information on the reliability of the estimation, while the PSCF model cannot. To verify and evaluate the capability of the PSDF model, tests are carried out using three scenarios based on ambient trajectory analysis data and simulated source distributions. The test results demonstrate that the PSDF model can identify the sources of ambient trace species more accurately than the PSCF model. The PSDF model can quantify the size of the source contaminating the air parcels passing through it, and the model can detect the variation of source intensity. Also, in the test, we evaluate reliability of the information provided by the PSDF model. In addition, future works are recommended to improve the model and increase its applicability.

**Keywords:** Gaussian process, Regression, Trajectory analysis, Air pollution, Source identification

## OPEN ACCESS

**Received:** September 7, 2021

**Revised:** December 22, 2021

**Accepted:** January 11, 2022

**\* Corresponding Author:**


dhwee@ewha.ac.kr

**Publisher:**

Taiwan Association for Aerosol  
Research

**ISSN:** 1680-8584 print

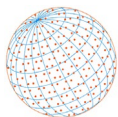
**ISSN:** 2071-1409 online

 **Copyright:** The Author(s).  
This is an open access article  
distributed under the terms of the  
[Creative Commons Attribution  
License \(CC BY 4.0\)](https://creativecommons.org/licenses/by/4.0/), which permits  
unrestricted use, distribution, and  
reproduction in any medium,  
provided the original author and  
source are cited.

## 1 INTRODUCTION

Combination of ambient measurements and models is an essential approach to understand the atmospheric environment. Numerous chemical species in the atmosphere are measured at many monitoring stations in Korea. A total of 552 monitoring stations are operated either by the Ministry of Environment of the Republic of Korea or by local governments. Each station has been producing hourly and daily average concentrations of several chemical species since 2016 (NIER, 2017). Many researchers also measure various ambient species across the country. Therefore, development and application of tools to interpret diverse and numerous ambient measurement data are essential in understanding the atmospheric environment.

To identify sources of air pollutants, a variety of models has been developed and applied to the atmospheric environment. Mathematical models based on fundamental atmospheric chemistry and physics can track emission from sources, their atmospheric transport and transformation, and contribution to the concentrations at a given location (receptor). However, several factors limit application of these models, including need for temporal and spatial emission inventory and meteorological field data. Most receptor models address the source identification problem based on a certain statistical theory. Some models attempt to relate measured concentrations at a given location to their sources based on statistical theory without reconstructing atmospheric transport of the material (Seinfeld and Pandis, 2016). The potential source contribution function (PSCF) is one of the receptor models to identify sources of ambient trace species and has been applied to



diverse research to identify the source area of ambient trace species (Ashbaugh *et al.*, 1985; Zeng and Hopke, 1989; Cheng *et al.*, 1993a, b; Hopke *et al.*, 1995; Peng *et al.*, 2007; Yu, 2013). This method offers users two major advantages. First, the PSCF model is attractive for users because it requires relatively few input data sets: only the measured concentrations of ambient trace species and the corresponding backward trajectories, which are easy to obtain. Second, the calculation is fast with small computing resources. However, the PSCF model also has some limitations. First, the PSCF value in a cell with a small number of trajectories can be more sensitive to certain high concentration events compared to a cell with a large number of trajectories, and this can cause unrealistically high PSCF values in cells with a small number of trajectories. Generally, to reduce this effect, an arbitrary weighting function is applied to downweigh the PSCF values in the cell in which the total number of trajectories is less than three times the average number of trajectories per cell (Hopke *et al.*, 1995; Polissar *et al.*, 2001a, b). Still, even with application of this arbitrary weighting function, the problem has not been resolved. Second, the PSCF model cannot perfectly quantify source areas. The PSCF values indicate the probability of potential sources located in the cells, but do not represent the intensity (or size, number) of sources. For example, two areas, Locations A and B, have the same PSCF value. This does not mean that these two locations equally influence the air quality of a receptor site. Furthermore, the PSCF model cannot provide users with information on reliability of estimated sources. Direct and quantitative evaluation of both the intensity and reliability of certain areas cannot be achieved using PSCF values. Finally, with change in the criteria applied in the calculation, the estimated values is also subject to change. Generally, mean or median of air pollutant concentrations is applied as a criterion. Mean or median can change with data set, so PSCF values with different criteria cannot be quantitatively compared. Several researches extend and modify the conventional models to reduce the limitation of PSCF and to enhance the application of trajectories (Stohl *et al.*, 2002; Lin *et al.*, 2003; Kim *et al.*, 2016; Kim *et al.*, 2019).

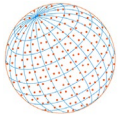
Machine learning is a branch of artificial intelligence concerned with construction of programs that learn from experience (Daintith and Wright, 2008). It is widely used to construct an accurate and useful approximation when the process cannot be identified completely (Alpaydin, 2010). For example, in a certain system, we do not know how to relate input to output due to lack of knowledge on the relationship between them. When we possess large amounts of input and output data, the goal is to “learn” to relate output to input. In other words, we build a computer (machine) to automatically extract a model for the system and let it estimate output from given input. Therefore, machine-learning techniques can be very effective in atmospheric study where physical and chemical processes are complicated and some processes are still unidentified. Albeit powerful, machine-learning techniques have only limited number of applications in the field of atmospheric research: for predicting spatiotemporal distributions of air pollutant concentrations (Yang *et al.*, 2018; Zhan *et al.*, 2017; Lary *et al.*, 2015; Petelin *et al.*, 2013) and for forecasting the concentrations of air pollutants (mainly concentration of particulate matter) (Shaban *et al.*, 2016).

In this study, we introduce a new model for identifying source areas of ambient trace species, called potential source density function (PSDF). This model can estimate the source distribution, that is, location and intensity, influencing the ambient concentrations at a receptor site based on Gaussian process regression (GPR), a machine-learning technique. The PSDF model requires known concentrations of ambient trace species and the corresponding backward trajectories, and the calculation is fast with small computing resources, like the conventional models. However, the PSDF model provides users with improved information about air pollution sources. Source distribution estimated by the PSDF model helps one understand the intensity and consistency of each area's influence on the concentrations of ambient trace species at a receptor site.

The paper is organized as follows. In Sect. 2, where the background theory is discussed, the concept of PSDFs is described by referring to the theory of Gaussian process regression (GPR). In Sect. 3, numerical examples are provided to demonstrate the possibility of using the developed method for studying the locations of contamination sources. A brief discussion follows in Sect. 4.

## 2 THEORY

In this section, we introduce the formulation of our PSDF method. First, we discuss the basics of the PSDF method (Sect. 2.1), and then continue on its implementation (Sect. 2.2), where the



structured kernel interpolation (SKI) method is introduced. Then, we explain how to specify hyperparameters in the model (Sect. 2.3).

## 2.1 Potential Source Density Function (PSDF)

Let there be a function  $f(\mathbf{x})$ , where  $\mathbf{x}$  represents the spatial coordinate variable typically in  $\mathbf{R}^2$ . We have  $N_s$  trajectories  $\xi_i(t)$  ( $1 \leq i \leq N_s$ ), where  $\xi_i(t) = \mathbf{x}_0$ . These are all backward trajectories with a final landing point  $\mathbf{x}_0$ , that is,  $t_i - T \leq t \leq t_i$  for all  $\xi_i$ , where  $T$  is a given time interval for backward tracking. The following integral is provided as the result of noisy measurement associated with

each trajectory:  $c_i = \int_{t=t_i-T}^{t_i} f(\xi_i(t))dt + \epsilon_i$ . Here,  $\epsilon_i$  is an additive independent measurement noise

with zero mean and variance  $\sigma_s^2$ . Thus, the set of all observed data consists of  $N_s$  pairs of  $[c_i, \xi_i]$ , whose set is denoted as  $C$ .

The formal setting mentioned above can be translated into the context of air pollution research as follows. One may consider  $c_i$  as the concentration of a pollutant measured at the sampling location  $\mathbf{x}_0$ . The sampled air parcel contains gradually accumulated pollutants collected while traveling along its trajectory  $\xi_i$ . This process of accumulation is modeled as an integral over a source density function, which is denoted here as  $f$ . By estimating  $f$  from the observed data, one can potentially locate and identify the contamination sources. When estimated,  $f$  is called the PSDF of the pollutant.

Many approaches can be used for estimation of  $f$ . In this study, we use Gaussian process regression (GPR) (Rasmussen and Williams, 2006). Let us assume that  $f$  is a Gaussian process:

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')). \quad (1)$$

Here,  $m(\mathbf{x})$  and  $k(\mathbf{x}, \mathbf{x}')$  are the mean and covariance functions of  $f(\mathbf{x})$ , respectively. Since we have very little prior knowledge on  $f$ , a Gaussian process with zero mean is usually taken as the prior:

$$f(\mathbf{x}) \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}')). \quad (2)$$

A simple square exponential kernel is employed here as the covariance function of choice:

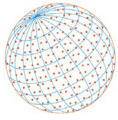
$$k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left(-\frac{|\mathbf{x} - \mathbf{x}'|^2}{2\ell^2}\right) \quad (3)$$

Here,  $\ell$  is the correlation length scale, and  $\sigma_f$  is a factor representing the strength of covariance.

It should be noted that the present problem is slightly different from typical cases of Gaussian process regression, in which the set of training outputs is composed of direct observations of pointwise values. The difference becomes obvious when covariances are computed. For example, the covariance of  $c_i$  and  $c_j$  is given as follows:

$$\begin{aligned} \text{cov}(c_i, c_j) &= \text{cov}\left(\int_{t=t_i-T}^{t_i} f(\xi_i(t))dt + \epsilon_i, \int_{t=t_j-T}^{t_j} f(\xi_j(t))dt + \epsilon_j\right) \\ &= \int_{t=t_i-T}^{t_i} \int_{t'=t_j-T}^{t_j} \text{cov}(f(\xi_i(t)), f(\xi_j(t'))dt'dt + \sigma_s^2 \delta_{ij} \\ &= \int_{t=t_i-T}^{t_i} \int_{t'=t_j-T}^{t_j} k(\xi_i(t), \xi_j(t'))dt'dt + \sigma_s^2 \delta_{ij}. \end{aligned} \quad (4)$$

For later use, we employ the following notation:



$$K_{c_i, c_j} = \int_{t=t_i-T}^{t_i} \int_{t'=t_j-T}^{t_j} k(\xi_i(t), \xi_j(t')) dt' dt, \quad (5)$$

which provides the element value in the  $i$ -th row and  $j$ -th column of the matrix  $K_{C,C}$ .

Since the set of observed data is composed of integrals over trajectories, evaluation of the related covariance naturally involves integrals over time. For a case with a large number of observed trajectories, a naively implemented integral may be time-consuming to evaluate. Thus, reducing the computational cost for evaluation of covariances is an essential task for making the concept practically viable. In the following, we propose an efficient method for evaluation of covariances, which is based on the concept of structured kernel interpolation (SKI) (Wilson and Nickisch, 2015).

## 2.2 Structured Kernel Interpolation for PSDFs

To reduce the computational cost for evaluation of covariances, it is proposed to use the SKI scheme (Wilson and Nickisch, 2015). The starting point is to apply the method of the subset of regressors (SoR) (Silverman, 1985) to Eq. (4):

$$\text{cov}(c_i, c_j) = K_{c_i, c_j} + \sigma_s^2 \delta_{ij} \approx \text{cov}_{\text{SoR}}(c_i, c_j) = K_{c_i, U} K_{U, U}^{-1} K_{U, c_j} + \sigma_s^2 \delta_{ij} \quad (6)$$

Here,  $U = [\mathbf{u}_l]$  is a set of  $N_U$  inducing points in  $\mathbf{R}^2$  ( $1 \leq l \leq N_U$ ).  $K_{c_i, U}$ ,  $K_{U, U}$ , and  $K_{U, c_j}$  are the  $1 \times N_U$ ,  $N_U \times N_U$ , and  $N_U \times 1$  covariance matrices generated from the exact kernel of Eq. (3), respectively. For example, the  $j$ -th element in  $K_{c_i, U}$ , that is,  $K_{c_i, \mathbf{u}_j}$  is given as follows:

$$K_{c_i, \mathbf{u}_j} = \int_{t=t_i-T}^{t_i} \text{cov}(f(\xi_i(t)), f(\mathbf{u}_j)) dt = \int_{t=t_i-T}^{t_i} k(\xi_i(t), \mathbf{u}_j) dt, \quad (7)$$

and the element value in the  $i$ -th row and  $j$ -th column of  $K_{U, U}$ , that is,  $K_{\mathbf{u}_i, \mathbf{u}_j}$  is given as follows:

$$K_{\mathbf{u}_i, \mathbf{u}_j} = \text{cov}(f(\mathbf{u}_i), f(\mathbf{u}_j)) = k(\mathbf{u}_i, \mathbf{u}_j). \quad (8)$$

If the inducing points  $U$  are on a regular grid with identical spacing between neighbouring points,  $K_{U, U}$  can be efficiently evaluated using the underlying Kronecker-Toeplitz structure, reducing the computational burden (Wilson and Nickisch, 2015). Eventually, we can approximately construct an  $N_S \times N_S$  matrix  $K_{C,C}$  as follows:

$$K_{C,C} \approx K_{C,U} K_{U,U}^{-1} K_{U,C}, \quad (9)$$

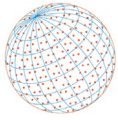
where each row of  $K_{C,U}$  is given by  $K_{c_i, U}$  and  $K_{U,C} = K_{C,U}^\top$ .

Next, we approximate the  $N_S \times N_U$  matrix  $K_{C,U}$  of cross-covariances between the trajectories and the inducing points by interpolating on the  $N_S \times N_U$  covariance matrix  $K_{U,U}$ . This is the central part of the SKI scheme, which yields

$$K_{C,U} \approx W_{C,U} W_{U,U}, \quad (10)$$

where  $W_{C,U}$  is an  $N_S \times N_U$  matrix of interpolation weights. Each row of  $W_{C,U}$  is computed by evaluating the following time integral:

$$W_{c_i, U} = \int_{t=t_i-T}^{t_i} W_{\xi_i(t), U} dt, \quad (11)$$



where  $W_{\xi_i(t),U}$  is the matrix of interpolation weights for  $\xi_i(t)$ . In this work, we use uniformly spaced two-dimensional grid points as  $U$  and a linear interpolation scheme, using only four of the nearest grid points of  $\mathbf{v}(t)$ . For example, considering the situation where  $\xi_i(t)$  is located within a rectangle whose vertices are denoted as  $\mathbf{u}_1$  (southwest),  $\mathbf{u}_2$  (southeast),  $\mathbf{u}_3$  (northwest), and  $\mathbf{u}_4$  (northeast), we write

$$W_{\xi_i(t),\mathbf{u}_1} = \frac{L_x - a}{L_x} \times \frac{L_y - b}{L_y}, \quad (12)$$

$$W_{\xi_i(t),\mathbf{u}_2} = \frac{a}{L_x} \times \frac{L_y - b}{L_y}, \quad (13)$$

$$W_{\xi_i(t),\mathbf{u}_3} = \frac{L_x - a}{L_x} \times \frac{b}{L_y}, \quad (14)$$

$$W_{\xi_i(t),\mathbf{u}_4} = \frac{a}{L_x} \times \frac{b}{L_y}, \quad (15)$$

and

$$W_{\xi_i(t),\mathbf{u}_j} = 0 \quad \text{for } j \notin \{1, 2, 3, 4\}. \quad (16)$$

Here,  $L_x$  and  $L_y$  are the length of longitudinal side and that of the latitudinal side, respectively.  $a$  is the distance from the west side, and  $b$  is that from the south side. Thus, all but four elements in  $W_{\xi_i(t),U}$  vanish, making  $W_{C,U}$  a sparse matrix.

Substituting Eq. (10) into Eq. (9), we get

$$K_{C,C} \approx K_{C,U} K_{U,U}^{-1} K_{U,C} \approx W_{C,U} K_{U,U} K_{U,U}^{-1} K_{U,U} W_{C,U}^T = W_{C,U} K_{U,U} W_{C,U}^T. \quad (17)$$

Though the computational advantage of using Eq. (17) to approximately evaluate  $K_{C,C}$  may not be obvious, it is significant. It can be summarized as follows:

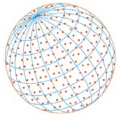
1. The most important gain is due to separation of two major operations in the procedure, that is, the time integrals and the covariance evaluations.  $W_{C,U}$  does not involve evaluation of the covariance kernel, while  $K_{U,U}$  does not involve any integral over trajectories. Thus, the time integrals over the trajectories can be evaluated first to construct  $W_{C,U}$ , which is stored for later uses, and then only the covariance matrix  $K_{U,U}$  can be evaluated multiple times without involving integrals.
2. The computational cost of evaluating  $W_{C,U}$  scales like  $O(N_s)$ , since it only involves a single integral over a trajectory instead of a double one. This is more affordable than directly evaluating the double integral in Eq. (4) for all  $1 \leq i \leq N_s$  and  $1 \leq j \leq N_s$ , which is  $O(N_s^2)$ , especially in cases with large  $N_s$ .
3. Further, as we have already pointed out, the underlying Kronecker-Toeplitz structure reduces the cost of computing  $K_{U,U}$ .

All these points indicate the method proposed here as efficient enough for practical applications.

The other covariance matrices in the Gaussian process regression can be evaluated essentially in the same fashion. Let us denote the set of test points at which the predictive distribution should be given as  $X_*$ . The cross-covariance matrix between the test points and the trajectories  $K_{C,X_*}$  and the covariance matrix between the test points  $K_{X_*,X_*}$  can be evaluated as follows:

$$K_{C,X_*} \approx W_{C,U} K_{U,U} W_{X_*,U}^T, \quad (18)$$

and



$$K_{X_*, X_*} \approx W_{X_*, U} K_{U, U} W_{X_*, U}^T, \quad (19)$$

where  $W_{X_*, U}$  is the  $N_T \times N_U$  matrix of interpolation weights for the test points  $X_*$ .

Now, we summarize the overall computational process for a concise reference. The objective is to predict the values of  $f$  at  $N_T$  test points  $\mathbf{x}_{*,m}$  ( $1 \leq m \leq N_T$ ). The training vector is constructed by combining the concentration measurement data, that is,  $\mathbf{c} = [c_1 \ c_2 \ c_3 \ \cdots \ c_{N_S}]^T$ . The set of trajectories associated with  $\mathbf{c}$  is defined as  $X = [\xi_1 \ \xi_2 \ \xi_3 \ \cdots \ \xi_{N_S}]^T$ . We define the test output vector as  $\mathbf{f}_* = [f_{*,1} \ f_{*,2} \ f_{*,3} \ \cdots \ f_{*,N_S}]^T$ , where  $f_{*,m}$  is the estimated value of  $f(\mathbf{x}_{*,m})$ . According to the prior, the joint distribution of the training vector and the test output vector is given as follows:

$$\begin{bmatrix} \mathbf{c} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N} \left( \mathbf{0}, \begin{bmatrix} \mathbf{A} & \mathbf{C} \\ \mathbf{C}^T & \mathbf{B} \end{bmatrix} \right), \quad (20)$$

where  $\mathbf{A} = K_{C,C} + \sigma_S^2 \mathbf{I}$ ,  $\mathbf{B} = K_{X_*, X_*}$ , and  $\mathbf{C} = K_{C, X_*}$ , which are all approximately evaluated by the SKI scheme discussed in the previous subsection.  $\mathbf{I}$  is an identity matrix of an appropriate size.

Applying a standard argument for multivariate Gaussian distributions to this distribution (Rasmussen and Williams, 2006), we can construct the conditional distribution to provide the key predictive equations for Gaussian process regression:

$$\mathbf{f}_* | X, \mathbf{c}, X_* \sim \mathcal{N}(\bar{\mathbf{f}}_*, \text{cov}(\mathbf{f}_*)), \quad (21)$$

where

$$\bar{\mathbf{f}}_* = E(\mathbf{f}_* | X, \mathbf{c}, X_*) = \mathbf{C}^T \mathbf{A}^{-1} \mathbf{c}, \quad (22)$$

and

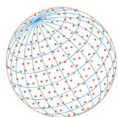
$$\text{cov}(\mathbf{f}_*) = \mathbf{B} - \mathbf{C}^T \mathbf{A}^{-1} \mathbf{C}. \quad (23)$$

The log marginal likelihood is given as follows:

$$\log p(\mathbf{c} | X) = -\frac{1}{2} \mathbf{c}^T \mathbf{A}^{-1} \mathbf{c} - \frac{1}{2} \log |\mathbf{A}| - \frac{N_S}{2} \log 2\pi. \quad (24)$$

## 2.3 Specification of Hyperparameters

To complete the specifications of the model, we need to determine hyperparameters. There are three hyperparameters in the PSDF model:  $\ell$ ,  $\sigma_f$ , and  $\sigma_s$ . In usual cases, these hyperparameters are determined by maximizing the likelihood of obtaining the training vector from the prior. That is, the hyperparameters are set to the optimal values with which the maximum of the log marginal likelihood, that is, Eq. (24), is attained. This approach is typically referred to as the type II maximum likelihood (ML-II) approximation (Rasmussen and Williams, 2006). However, the process is time-consuming because it involves multiple evaluation of costly predictive equations. Even worse, it does not always converge well. Especially, in typical cases of air pollution research, very close trajectories may exhibit very different measured values for pollutant concentrations. Such an anomaly is not unexpected since contamination sources are not always active. For example, a fossil-fuel power plant does not emit pollutants when it is idle. If two similar trajectories involve the area of a power plant, but one has visited it while the plant is idle, while the other has visited it while the plant is operating, these two similar trajectories may provide contradicting information, which can easily confuse the learning model. Thus, it is necessary to find reasonable values without the full ML-II approximation.



To reduce computational difficulty associated with the full ML-II approximation, we impose an additional condition on the hyperparameters. We first estimate the variance of the training data set, that is,  $\text{Var}[c_i]$ , by computing the variance of the measured concentrations. On the other hand, according to Eq. (4), this variance must be equal to  $\text{cov}(c_i, c_i) = \sigma_f^2 T^2 + \sigma_s^2$ . Imposing this condition onto  $\sigma_f$  and  $\sigma_s$ , we fix both  $\sigma_f$  and  $\sigma_s$  by identifying a single parameter  $r$  ( $0 < r < 1$ ):

$$\sigma_s^2 = r \text{Var}[c_i], \quad (25)$$

and

$$\sigma_f^2 = \frac{(1-r)\text{Var}[c_i]}{T^2}. \quad (26)$$

Thus, instead of performing the full ML-II approximation with three hyperparameters, we perform the ML-II approximation with two hyperparameters only:  $\ell$  and  $r$ . With the reduced dimension of the search space, convergence is more easily obtained, making the entire process more robust.

Separation of the variance, performed in Eqs. (25–26), also provides some physical interpretations. Namely, the total variance of the measured concentrations can have two subparts. One part comes from the temporal variation of the source activity, which is an effect unaccounted for in the model. As already mentioned above, a fossil-fuel power plant does not emit pollutants when it is idle. Such temporal variance, or any uncertainty caused by an unaccounted effect, is represented by  $\sigma_s^2$ . The other part comes from the spatial covariance, which is represented by  $\sigma_f^2 T^2$ . Thus, each of the divided parts in Eqs. (25–26) may indicate the physical nature of the uncertainty in the measured data set. The situation is similar to that of ANOVA, where one partitions the variance of the data into two parts, that is, one measuring the signal and the other the noise (Helsel *et al.*, 2020).

The hyperparameters for the PSDF model can be specified by the ML-II approximation with the two variables  $c$  and  $r$ . However, in most practical applications of the PSDF model, we can reduce the number of hyperparameters even further. Other models utilizing ambient data and backward trajectories divide the domain into a  $0.5^\circ$  by  $0.5^\circ$  grid (Zhang *et al.*, 2015; Kim *et al.*, 2016). Also, the typical spatial scale in atmospheric chemical transport models ranges from 20 km to 80 km for regional and continental scale (Seinfeld and Pandis, 2016). Considering these typical spatial scales, one can obtain an appropriate value for  $\ell$  without the full ML-II approximation.

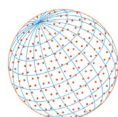
Once we fix  $\ell$  to a value corresponding to the typical spatial scale associated with the problem, then we only need to determine one hyperparameter, that is,  $r$ . Now, one may use the ML-II approximation and maximize Eq. (24) with respect to  $r$  only, which is much simpler than the original problem of full optimization over three hyperparameters. One may simply evaluate Eq. (24) with a few different values of  $r$  to see which gives the most reasonable result. For example, one may try three values:  $r = 0.1$  (small uncertainty due to unaccounted effects),  $r = 0.5$  (medium uncertainty; equipartition of the variance), and  $r = 0.9$  (high uncertainty). MATLAB (MATLAB and Statistics Toolbox Release 2021a, The MathWorks, Inc., Natick, Massachusetts, United States.) has been used in developing of the PSDF model, but the model can be executed on a free software platform like GNU Octave (Eaton *et al.*, 2019) without any major modification.

In Sect. 3, we validate our PSDF model by performing numerical experiments. We use real backward trajectories, but the ambient data set, that is, the pollutant concentrations, is constructed from assumed source distributions. Thus, in this numerical experiment, uncertainty due to unaccounted effects is naturally small. Hence,  $r$  is fixed at a low value, that is, 0.1, to minimize the influences of other factors in the variation of ambient data in application of PSDF and to evaluate the PSDF results. On the ground of resolution in other atmospheric models used for similar study (Zhang *et al.*, 2015; Kim *et al.*, 2016),  $\ell$  is fixed at  $0.5^\circ$ .

The algorithm of identification of the locations of contamination sources with a PSDF model can be summarized as follows:

1. Prepare a grid of  $N_T$  test points ( $X_*$ ) in the spatial domain.





2. Create an additional, regular grid for interpolation ( $U$ ), which can largely overlap with the grid for the test points. The total number of points on this new grid for interpolation is  $N_U$ .
3. Set up the interpolation matrices  $W_{C,U}$  and  $W_{X_*,U}$ . During evaluation of  $W_{C,U}$ , one needs to perform time integrals, i.e., Eq. (11).
4. Maximize Eq. (24) by adjusting  $\ell$  and  $r$ . This step fixes all the hyperparameters in the model. If there are already good estimates for  $\ell$  and  $r$  from earlier experiences of similar problems, one may skip this step and just use the acceptable values obtained previously.
5. Obtain  $\mathbf{A} = K_{C,C} + \sigma_C^2 \mathbf{I}$ ,  $\mathbf{B} = K_{X_*,X_*}$ , and  $\mathbf{C} = K_{C,X_*}$ .
6. Estimate the values of the PSDF, that is,  $\bar{\mathbf{f}}_*$ , at the test points, using Eq. (22).

### 3 RESULTS

We validate the capability of the PSDF model to locate source areas (Sect. 3.3) and to quantify intensity of source areas (Sect. 3.4). Numerical examples are provided to demonstrate the capability and characteristics of the PSDF model. In this study, we use only actual backward trajectories, without the measured concentration data at a receptor site. To evaluate the PSDF results with controlled interaction between ambient data and backward trajectories, the simulated ambient data are applied to the PSDF model instead of actual ambient data. The simulated ambient data are generated by a hypothetical source distribution and backward trajectories. The capability and characteristics of PSDF results are analysed using the simulated ambient data and the actual backward trajectories. Additionally, after applying the PSDF model with changing simulated ambient data, the results are evaluated compared to those computed from the conventional model, PSCF. These examples are useful not only to see whether the scheme possesses reasonable capability to identify the locations of contamination sources, but also to study the generic behaviour of the scheme. The overall process for validation of PSDF is illustrated in Fig. 1.

#### 3.1 Simulated Sampling Information and the Corresponding Backward Trajectories

We assume that the receptor site is located at Anmyeondo Global Atmosphere Watch Station (AMY), National Institute of Meteorological Sciences, Anmyeon Island, Korea (36.5386°N, 126.3299°E), where ambient trace species are collected from June 2015 to May 2017 (local time).

The PSDF model requires backward trajectories and ambient data in the calculation. Backward trajectory analysis was performed for the sampling days using the Hybrid Single-Particle Lagrangian Integrated Trajectory 4 (HYSPPLIT4) model with meteorological data of the Global Data Assimilation

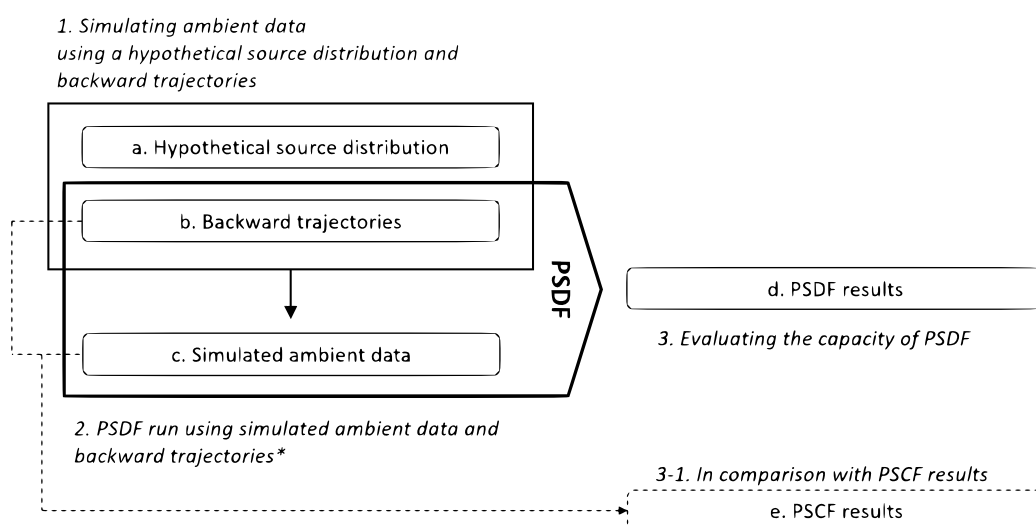
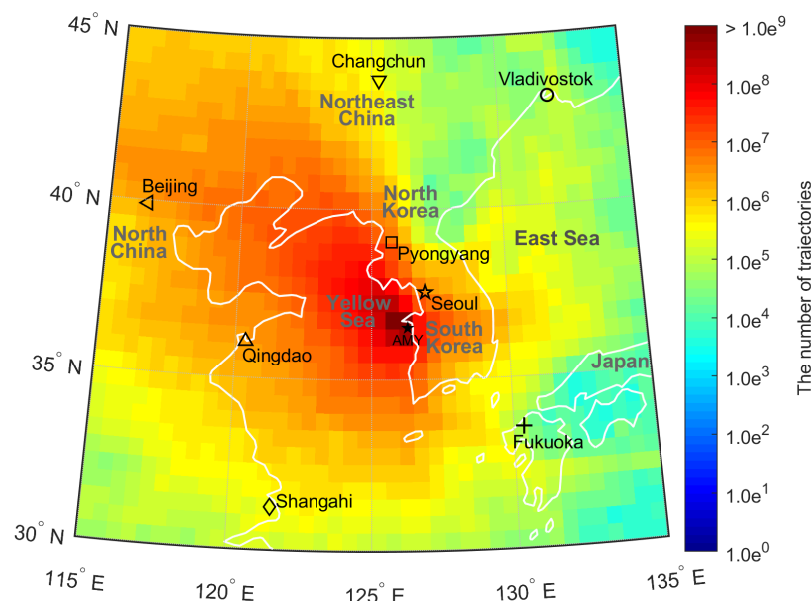
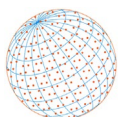


Fig. 1. Structure of validation.





**Fig. 2.** Backward trajectory frequency between 2015 and 2017 at Anmyeondo Global Atmosphere Watch Station (AMY).

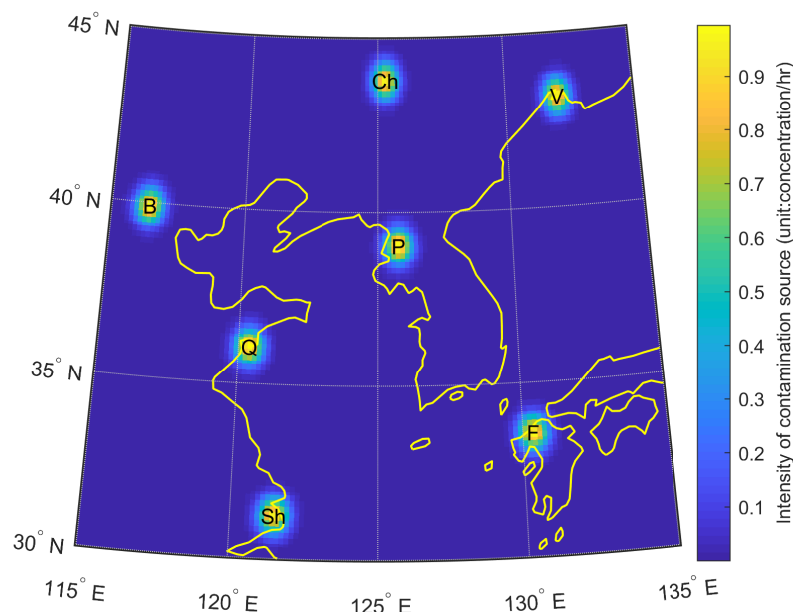
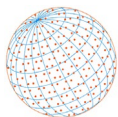
System (GDAS), which is operated by the National Centers for Environmental Prediction (NCEP) of the U.S. National Weather Service Organization (Stein *et al.*, 2015; Rolph *et al.*, 2017). The trajectory vertical motion method was to use the vertical velocity data fields supplied with the meteorological model data. Twenty-four trajectories, that is, one for each hour, were created during each day between 1<sup>st</sup> June 2015 and 31<sup>st</sup> May 2017; hence, the total number of trajectories available is 17544.

Air pollutants can be injected much higher than the mixed layer height or the planetary boundary layer (Lin *et al.*, 2003; Colarco *et al.*, 2004; Trentmann *et al.*, 2006). The starting height of backward trajectories in the PSCF model is chosen from 100 m to 3000 m (Heo *et al.*, 2009; Kim *et al.*, 2016, 2019). The starting height, that is, height at the arrival point, was set to 1500 meters, representatively in the middle of the range, in this study. In application of PSDF model in actual data, several heights have to be considered depending on air pollutant characteristics, PBL, etc. The time interval for backward tracking, that is,  $T$ , was 120 hours. Fig. 2 shows the spatial distribution of the backward trajectories used in this study.

### 3.2 Simulated Ambient Data

To generate the simulated ambient data, information of contamination sources is required. In this numerical experiment, the contamination sources are assumed to be located at seven major cities around the sampling location, that is, Beijing, Qingdao, Shanghai, Changchun, Pyongyang, Fukuoka, and Vladivostok, as shown in Fig. 3. These seven cities were selected because they exhibit a certain influence on the pollution level at the sampling site, based on our previous investigations and information from an emission inventory EDGAR v.4.3.2 (Crippa *et al.*, 2018). It was shown that the level of levoglucosan at Seoul could be influenced by emissions from Beijing, Qingdao, and Changchun (Kim *et al.*, 2019). In these areas, large amounts of organic carbons (OC) are generated from agricultural waste burning. Shanghai and Fukuoka are also known to emit large amounts of OC from agricultural waste burning, but these two cities exhibited relatively small effects on the level of levoglucosan at Seoul (Kim *et al.*, 2019). The influence from the area around Pyongyang can be significant due to its proximity to the sampling site at Seoul (Kim *et al.*, 2019). Vladivostok is also included as a source location because it is a relatively large city in the northeast of Seoul.

These seven cities exhibit wide variation in the number of trajectory visits; and hence, we can assess how the number of trajectory visits can affect the identification capability of the PSDF model by considering these seven cities. Pyongyang experiences a relatively large number of trajectory visits due to its proximity to the sampling site, while Fukuoka has relatively few visits, as shown in Fig. 2.



**Fig. 3.** Hypothetical source distribution according to Scenario 1 (B: Beijing, China, Q: Qingdao, China, Sh: Shanghai, China, Ch: Changchun, China, P: Pyongyang, North Korea, F: Fukuoka, Japan, and V: Vladivostok, Russia).

We assume that the source intensity follows a 2D-Gaussian distribution, and that source intensity at each location in the total distribution of source intensity including the seven sources is described as follows:

$$S(x, y) = \sum_i a_i \exp \left[ -\frac{(x - x_i)^2}{b_i^2} - \frac{(y - y_i)^2}{b_i^2} \right], \quad (27)$$

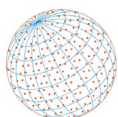
where  $a_i$  is the maximum intensity of the  $i$ -th contamination source, and  $b_i$  is the effective size of the  $i$ -th contamination source.  $x_i$  and  $y_i$  represent the longitude and latitude of the location of the  $i$ -th source center, respectively.

Two scenarios are being studied: Scenarios 1 and 2. Scenario 1 assumes that the maximum intensity of the contamination source,  $a_i$ , and the effective size of the contamination source,  $b_i$ , are set to 1 and 0.5 for all seven cities, respectively, as shown in Fig. 3. Thus, the influence of the number of visits can be identified by this choice of source locations. In Scenario 2, on the other hand, the contamination sources located in Qingdao and Changchun are modified to have different characteristics in maximum intensity and effective size, as shown in Table 1 and Fig. 4.

The simulation is carried out in the following fashion. The end point at  $t = t_i$  of the  $i$ -th backward trajectory is the sampling location, and the trajectory starts its journey from the corresponding starting location at  $t = t_i - T$ . The air parcel in each trajectory starts its travel along the corresponding trajectory with no contaminant. If the trajectory corresponding to the air parcel visits a location that possesses a contamination source, the intensity of the contamination source at the location is integrated according to the concentration of simulated measurement of that air parcel. That is, the simulated concentration of the air pollutant is obtained via the integral of

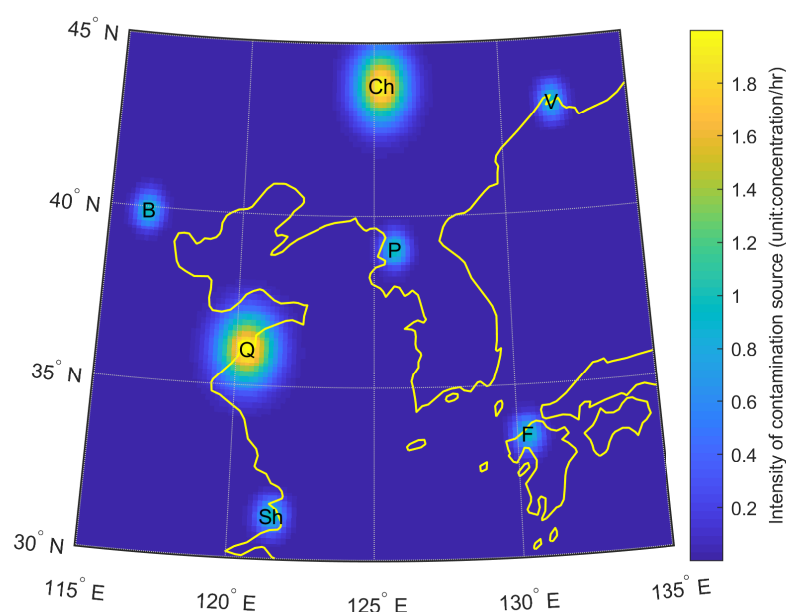
$$c_i = \int_{t=t_i-T}^{t_i} f(\xi_i(t)) dt, \text{ where } f \text{ represents the assumed source distribution for each case. According}$$

to the processes, three sets of ambient data are generated based on two scenarios and are applied to the PSDF and the PSCF models.



**Table 1.** Information on the seven hypothetical sources depending on scenario. Maximum intensity of the contamination source ( $a_i$ ) and effective size of the contamination source ( $b_i$ ) are given in units of concentration/hour and degrees, respectively.

Source Location	Scenario 1		Scenario 2	
	$a_i$	$b_i$	$a_i$	$b_i$
China				
Beijing	1	0.5	1	0.5
Qingdao	1	0.5	2	1
Shanghai	1	0.5	1	0.5
Changchun	1	0.5	2	1
North Korea				
Pyongyang	1	0.5	1	0.5
Japan				
Fukuoka	1	0.5	1	0.5
Russia				
Vladivostok	1	0.5	1	0.5

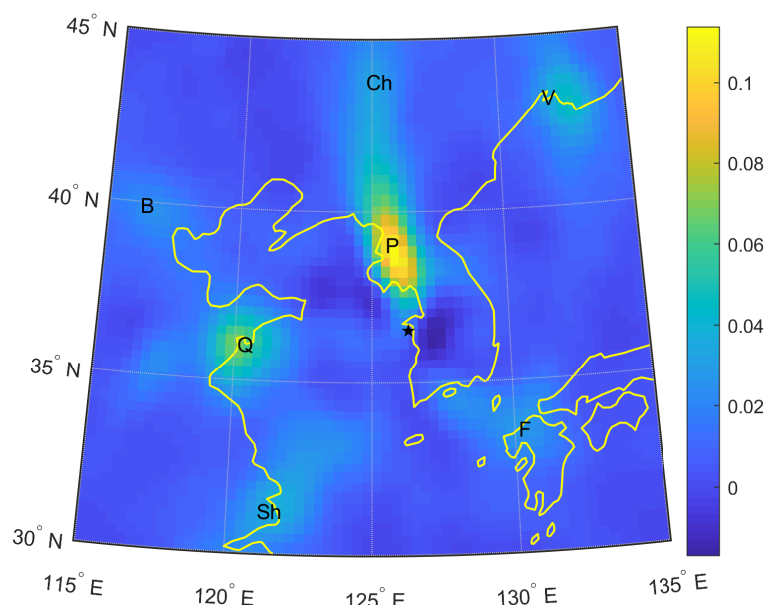
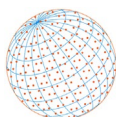


**Fig. 4.** Hypothetical source distribution according to Scenario 2. Symbol legends are the same as in Fig. 3.

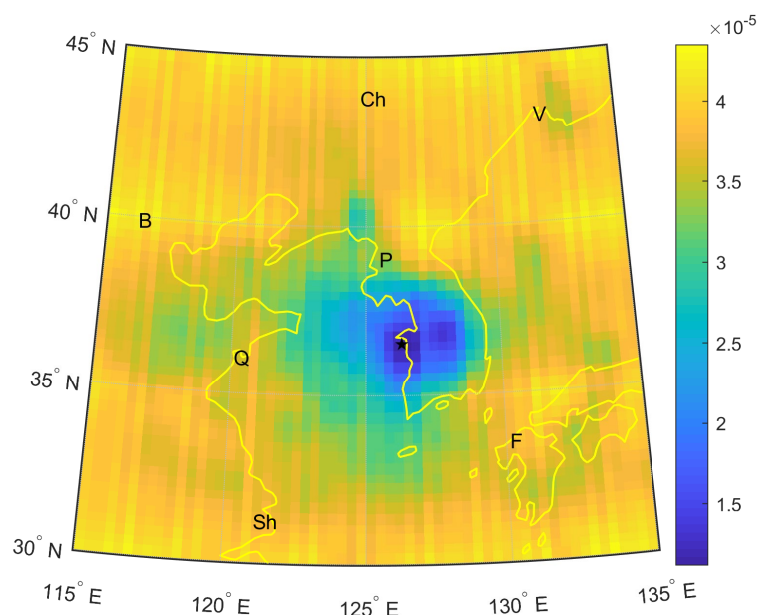
### 3.3 Results of Source Identification

The ambient data simulated using the source distribution of Scenario 1 are applied in the PSDF model to evaluate the capability to locate source areas. The simulation result based on Scenario 1 is presented in Fig. 5. PSDF can identify seven sources with the sources at Pyongyang and Qingdao well identified in particular. As shown in Fig. 2, these source locations comprise reasonably large numbers of trajectory visits, confirming that an adequate number of trajectory visits is crucial for appropriate identification of contamination sources. This is true in all the methods based on backward trajectories. This demonstrates that the PSDF model can reasonably identify the spatial length scale of the contamination source area, at least for regions with enough trajectory visits.

The PSDF model can also provide certain information on reliability of estimated PSDF values. Fig. 6 shows the variances of the estimated PSDF values. Fukuoka with small PSDF values exhibits large variance, suggesting large uncertainty associated with the estimated PSDF values for this city with a very small number of trajectory visits. The other sources like Pyongyang have smaller variances, which suggests that the estimated PSDF values for these sources are relatively reliable.



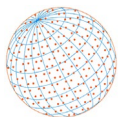
**Fig. 5.** Estimated PSDF,  $\bar{f}_*$ , constructed from the trajectories corresponding to the data shown in Fig. 2, with a generated  $c$  from the source distribution of Fig. 3.



**Fig. 6.** Standard deviation of the estimated PSDF,  $\sqrt{\text{Var}[\bar{f}_*]}$ , constructed from the trajectories corresponding to the data shown in Fig. 2, with a generated  $c$  from the source distribution of Fig. 3.

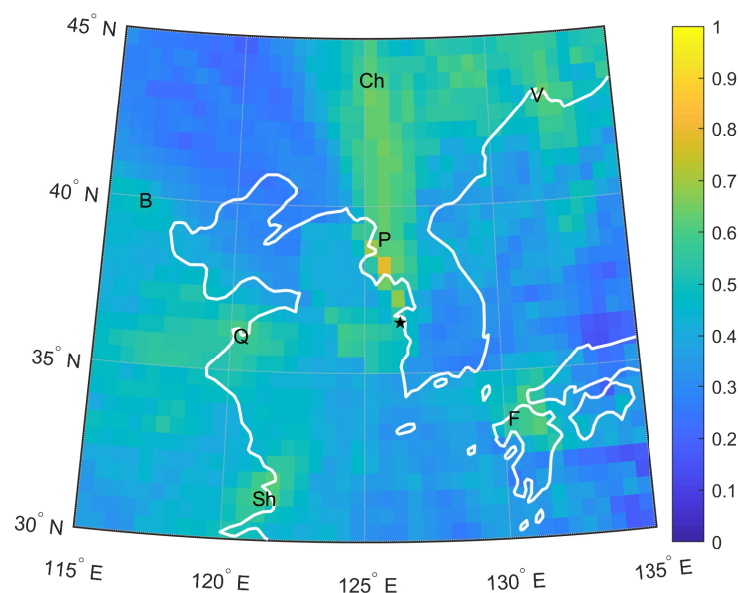
The spatial distribution of variances in the PSDF results, as shown in Fig. 6, exhibits similar patterns to the distribution of backward trajectory, as shown in Fig. 2.

The PSCF model, a conventional model utilizing ambient data and backward trajectories, also as applied in this study. The potential source contribution function (PSCF) is a simple tool to indicate potential source regions that contribute high air pollutant concentration based on the total number of trajectories over a given geographic region and the number of trajectories for high air pollutant concentration at the receptor (Ashbaugh *et al.*, 1985). The PSCF value in a cell with a small number of trajectories can be more sensitive to certain high-concentration events compared to a cell with a small number of trajectories, which can cause unrealistically high PSCF

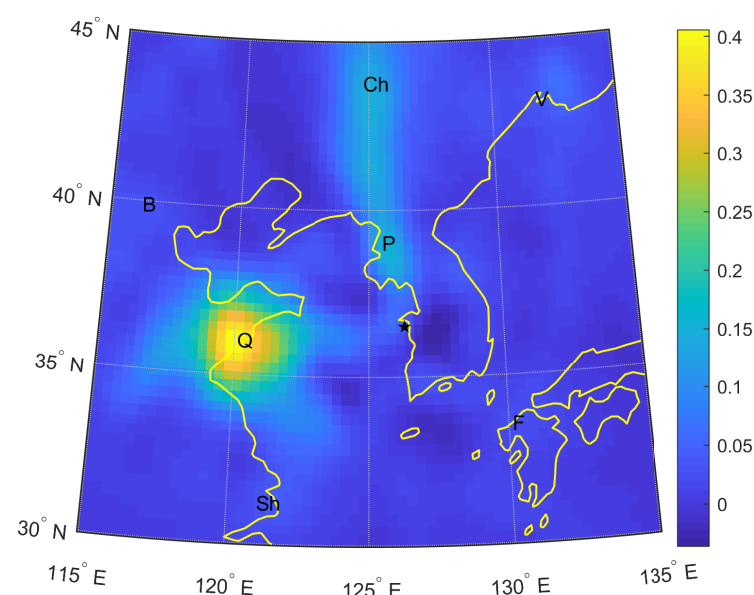


values in cells with a low number of trajectories. Generally, to reduce this effect, an arbitrary weighting function is applied to downweigh the PSCF values in the cell in which the total number of trajectories is less than three times the average number of trajectories per cell (Hopke *et al.*, 1995; Polissar *et al.*, 2001a, b). In this study, the arbitrary weighting function is used as the criterion for the weighting function widely used in other studies. The result of PSCF applied to Scenario 1 with the weighting function is presented in Fig. 7. In the PSCF result, cells with high values indicate a high probability of containing emission sources.

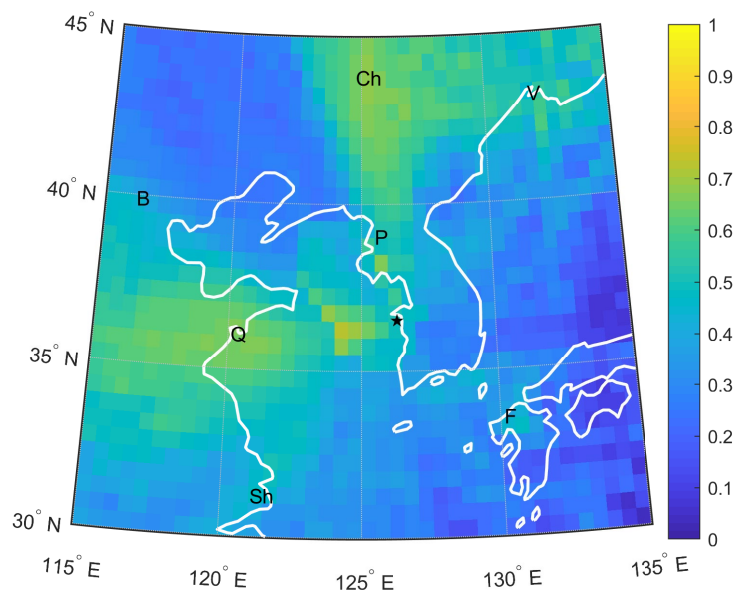
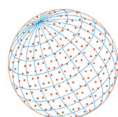
Scenario 2 is designed to evaluate the capability of the PSDF model to quantify sources with varying intensity and size. The contamination sources located in Qingdao and Changchun are stronger and broader than others in Scenario 2, as shown in Fig. 4 and Table 1. The simulated result of PSDF based on Scenario 2 is presented in Fig. 8. Five contamination sources excluding



**Fig. 7.** PSCF from the trajectories corresponding to the data shown in Fig. 2, with a generated  $c$  from the source distribution of Fig. 3.



**Fig. 8.** Estimated PSDF,  $\bar{f}_s$ , constructed from the trajectories corresponding to the data shown in Fig. 2, with a generated  $c$  from the source distribution of Fig. 4.



**Fig. 9.** PSCF from the trajectories corresponding to the data shown in Fig. 2, with a generated  $c$  from the source distribution of Fig. 4.

those at Qingdao and Changchun are identified with a similar pattern to the simulated results based on the other two scenarios. The maximum PSDF values in contamination sources at Qingdao and Changchun are much higher in Fig. 8 than those in Fig. 5. Especially, the influence of contamination source located in Changchun exceeds that level in Pyongyang, the most influential area in the PSDF results based on Scenario 1. The PSDF model also can detect the difference of effective source size (area) between Scenario 1 and Scenario 2. Because of the difference in number of trajectory visits, the PSDF model can more clearly recognize the variation of effective size in the source area of Qingdao than that of Changchun.

The PSCF model also is applied to Scenario 2. The simulated result of PSCF with this scenario is presented in Fig. 9. The PSCF model shows different changes depending on the source distribution. Though PSCF cannot quantify the variation of source intensity, the values in areas around Qingdao and Changchun, as in Fig. 9, are larger than those in Fig. 7. Also, the PSCF values are high in wide areas around Qingdao and Changchun. However, too many cells outside of the seven contamination source areas are also identified as potential source locations around Qingdao and Changchun.

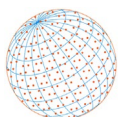
Analysing Scenario 1 and Scenario 2 with the PSDF and PSCF models, verified that the PSDF model can more sensitively and quantitatively represent the variation of source distributions and clearly detect the locations of sources.

## 4 DISCUSSION: RELAXATION OF ASSUMPTIONS AND EXTENSION OF THE METHOD

The PSDF model is developed to identify (locate and quantify) sources of ambient trace species based on Gaussian process regression (GPR). The PSDF model requires only backward trajectories and sampling data at a receptor site in the calculation, which is not significantly more time-consuming compared to the conventional model using the same input data such as the potential source contribution function (PSCF). Algorithms of PSDF are improved by structured kernel interpolation (SKI). The user can directly estimate the reliability of results because the PSDF model can provide the variance of estimated strengths.

The PSDF model is an effective tool to understand the characteristics of the atmospheric environment. Although the PSDF model is simple and easy to use, it can be applied to investigate sources of ambient trace species transported in a regional range; for example, from other countries in Northeast Asia to Korea. To use the PSDF model, there is need for concentration data of trace





species, that are stable and not reactive in the atmosphere; for example, levoglucosan and PAHs (Polycyclic aromatic hydrocarbons). As mentioned in Sect. 1, such data sets are widely available.

Here we propose a few possible extensions of the PSDF scheme presented in the main text. Many of them can be implemented without major efforts, but some of them should be pursued further with a certain amount of additional research efforts.

#### 4.1 Multiple Sampling Sites

In many cases, air pollutants are simultaneously sampled at several different locations. An approach to deal with multiple sampling sites in the PSCF scheme was once proposed (Peng *et al.*, 2007), but it was given without much justification.

On the other hand, the PSDF method described in this article can be extended without any modification in the formulation to such a case. Actually, there is no a priori need for all the trajectories to share the same final landing point  $\mathbf{x}_0$ . For example, let us suppose that there are two different sampling sites, i.e.,  $A$  and  $B$ . Each of them is supposed to have its own set of ten pairs of observed concentrations and backward trajectories:

$$C_A = \{[C_{A1}, \xi_{A1}], [C_{A2}, \xi_{A2}], \dots, [C_{A10}, \xi_{A10}]\} \quad (28)$$

and

$$C_B = \{[C_{B1}, \xi_{B1}], [C_{B2}, \xi_{B2}], \dots, [C_{B10}, \xi_{B10}]\}. \quad (29)$$

Then, one can simply make one set of data out of these two data sets by taking a union:  $C = C_A \cup C_B$ .

#### 4.2 Long Duration of Sampling; Assigning Multiple Trajectories to One Measurement

There may be cases where one measurement of pollutant concentration corresponds to many backward trajectories simultaneously. Such a case frequently occurs in practice, since a typical duration of continuous sampling is 1 day, i.e., 24 hours. A usual method to apply the PSCF scheme in such a case is to create 24 trajectories for each hour and assign the same concentration to all those 24 trajectories.

However, this is not really the best practice. The underlying implication behind such a practice would be that the air parcels of all the backward trajectories are equally carrying the same amount of the pollutant in them, which is clearly not true. Those air parcels carry various amounts of the pollutant, and they all make different contributions to the measured concentration on that day. Obviously, in such a case, a better interpretation is that the concentrations in the air parcels of all those 24 backward trajectories are ‘averaged’ to yield the measured concentration. This way of interpretation can be rigorously implemented in the PSDF method. The averaged concentration can be formally written as follows:

$$c_i = \frac{1}{24} \sum_{j=1}^{24} \left[ \int_{t=t_i-T}^{t_i} f(\xi_j(t)) dt \right] + \epsilon_i, \quad (30)$$

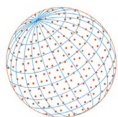
where  $j$  is the index of those 24 trajectories arriving the sampling site during that day,  $\xi_j$  presents the  $j$ -th trajectory, and  $t_j$  is the time of the arrival of the  $j$ -th trajectory.

To incorporate Eq. (30) into the present formulation based on the SKI scheme, the only place that should be modified is Eq. (11), which should be rewritten in the following form:

$$W_{c_i, U} = \frac{1}{24} \sum_{j=1}^{24} \left[ \int_{t=t_i-T}^{t_i} W_{\xi_j(t), U} dt \right]. \quad (31)$$

All the other part of the formulation remains essentially the same as before.





### 4.3 Implementation of Temporal Correlation

In the atmospheric environment, understanding temporal characteristics of ambient trace species is essential because the processes to emit, transport, transform, and remove trace species vary according to season, time of day, and events occurring at certain times. Temporal variation of ambient data has been extensively analysed, such as that of seasonal variation. However, such analysis can be limited for large amounts of data. Temporal correlation can be incorporated into the formulation of PSDF simply by assuming  $f$  not only as a Gaussian process over space, but as a Gaussian process over space and time:

$$f(\mathbf{x}, t) \sim \mathcal{GP}(0, k_x(\mathbf{x}, \mathbf{x}') \times k_t(t, t')), \quad (32)$$

where  $k_x$  and  $k_t$  represent the covariance function for space and for time, respectively. The present work assumes  $k_t(t, t') = 1$ .

### 4.4 Inclusion of a Temporal Profile

Airborne pollutants may experience temporal changes during transport. An airborne pollutant may decay, diffuse, or even be deposited. Such processes can be included in the PSDF formulation as a temporal profile, i.e.,

$$c_i = \int_{t=t_i-T}^{t_i} f(\xi_i(t)) g_i(t) dt + \epsilon_i, \quad (33)$$

where  $g_i$  is the temporal profile of the airborne pollutant corresponding to the  $i$ -th trajectory. If the pollutant is decaying via a first-order reaction with its time constant  $\tau$ , one can specify

$$c_i = \int_{t=t_i-T}^{t_i} f(\xi_i(t)) e^{-(t_i-t)/\tau} dt + \epsilon_i, \quad (34)$$

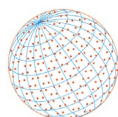
and the effect of the decay between the time of injection and the time of sampling can be properly considered.

Although the inclusion of decay would be the most promising application of temporal profiles in the PSDF formulation, there can be other important applications. A certain pollutant (e.g.,  $\text{O}_3$ ) can be created by a secondary formation process from a precursor emitted by a source (e.g.,  $\text{NO}_x$ ). In such a case, one may want to identify the source emitting the precursor, not the source directly emitting the measured pollutant itself. The corresponding secondary formation process may require a certain time to form the pollutant from its precursor, and the pollutant concentration may decay in time after exhibiting a peak. Such a behaviour can be implemented in the PSDF formulation simply by specifying  $g_i(t) = g(t_i - t)$ , where  $g$  is a bell-shaped function with an appropriate temporal time scale.

Another application can be the inclusion of the vertical information of the backward trajectories. An air parcel traveling through a trajectory passing too high may not catch pollutants emitted by sources on the ground. A previous study suggested that the vertical information in backward trajectories may be important in the PSCF model and developed a simple algorithm to account for the height of trajectories with high concentrations (Kim *et al.*, 2016). Vertical information in backward trajectories also can be easily included by specifying  $g_i$  that vanishes when the height of the trajectory is beyond a certain threshold, which may improve the capability to identify source areas.

## ACKNOWLEDGMENTS

This work was primarily supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (2019-R1F1A1062571). This research was also supported by Technology Development Program to Solve Climate Changes through the National Research

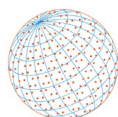


Foundation of Korea (NRF) funded by the Ministry of Science, ICT (2019M1A2A2103953). ISK was partly supported during the study by the Ewha Womans University scholarship of 2016.

The authors gratefully acknowledge the NOAA Air Resources Laboratory (ARL) for providing the HYSPLIT transport and dispersion model and/or READY website (<http://www.ready.noaa.gov>). The authors also gratefully acknowledge the Emissions of Atmospheric Compounds and Compilation of Ancillary Data (ECCAD, <http://eccad.aeris-data.fr/>) and the EU Joint Research Centre Emissions Database for Global Atmospheric Research (EDGAR) for providing emission data (<http://edgar.jrc.ec.europa.eu/>).

## REFERENCES

- Alpaydin, E. (2010). Introduction to Machine Learning, 2nd ed. Adaptive Computation and Machine Learning Series. The MIT Press.
- Ashbaugh, L.L., Malm, W.C., Sadeh, W.Z. (1985). A residence time probability analysis of sulfur concentrations at Grand Canyon National Park. *Atmos. Environ.* 19, 1263–1270. [https://doi.org/10.1016/0004-6981\(85\)90256-2](https://doi.org/10.1016/0004-6981(85)90256-2)
- Cheng, M.D., Hopke, P.K., Barrie, L., Rippe, A., Olson, M., Landsberger, S. (1993a). Qualitative determination of source regions of aerosol in Canadian high Arctic. *Environ. Sci. Technol.* 27, 2063–2071. <https://doi.org/10.1021/es00047a011>
- Cheng, M.D., Hopke, P.K., Zeng, Y. (1993b). A receptor-oriented methodology for determining source regions of particulate sulfate observed at Dorset, Ontario. *J. Geophys. Res.* 98, 16839–16849. <https://doi.org/10.1029/92JD02622>
- Colarco, P., Schoeberl, M., Doddridge, B., Marufu, L., Torres, O., Welton, E. (2004). Transport of smoke from Canadian forest fires to the surface near Washington, DC: Injection height, entrainment, and optical properties. *J. Geophys. Res.* 109, D06203. <https://doi.org/10.1029/2003JD004248>
- Crippa, M., Guizzardi, D., Muntean, M., Schaaf, E., Dentener, F., van Aardenne, J.A., Monni, S., Doering, U., Olivier, J.G.J., Pagliari, V., Janssens-Maenhout, G. (2018). Gridded emissions of air pollutants for the period 1970–2012 within EDGAR v4.3.2. *Earth Syst. Sci. Data* 10, 1987–2013. <https://doi.org/10.5194/essd-10-1987-2018>
- Daintith, J., Wright, E. (2008). A Dictionary of Computing, 6th ed. Oxford University Press.
- Eaton, J.W., Bateman, D., Hauberg, S., Wehbring, R. (2019). GNU Octave version 5.2.0 manual: A high-level interactive language for numerical computations. <https://www.gnu.org/software/octave/doc/v5.2.0/>
- Helsel, D.R., Hirsch, R.M., Ryberg, K.R., Archfield, S.A., Gilroy, E.J. (2020). Statistical Methods in Water Resources: U.S. Geological Survey Techniques and Methods, Book 4, Chapter A3. Tech. Rep., Reston, VA. <https://doi.org/10.3133/tm4a3>
- Heo, J.B., Hopke, P.K., Yi, S.M. (2009). Source apportionment of PM<sub>2.5</sub> in Seoul, Korea, *Atmos. Chem. Phys.* 9, 4957–4971. <https://doi.org/10.5194/acp-9-4957-2009>
- Hopke, P.K., Barrie, L.A., Li, S.M., Cheng, M.D., Li, C., Xie, Y. (1995). Possible sources and preferred pathways for biogenic and non-sea-salt sulfur for the high Arctic. *J. Geophys. Res.* 100, 16595–16603. <https://doi.org/10.1029/95JD01712>
- Kim, I.S., Lee, J.Y., Wee, D., Kim, Y.P. (2019). Estimation of the contribution of biomass fuel burning activities in North Korea to the air quality in Seoul, South Korea: Application of the 3D-PSCF method. *Atmos. Res.* 230, 104628. <https://doi.org/10.1016/j.atmosres.2019.104628>
- Kim, I.S., Wee, D., Kim, Y.P., Lee, J.Y. (2016). Development and application of three-dimensional potential source contribution function (3D-PSCF). *Environ. Sci. Pollut. Res.* 23, 16946–16954. <https://doi.org/10.1007/s11356-016-6787-x>
- Lary, D.J., Lary, T., Sattler, B. (2015). Using machine learning to estimate global PM<sub>2.5</sub> for environmental health studies. *Environ. Health Insights* 9, EHI.S15664. <https://doi.org/10.4137/EHI.S15664>
- Lin, J.C., Gerbig, C., Wofsy, S.C., Andrews, A.E., Daube, B.C., Davis, K.J., Grainger, C.A. (2003) A near-field tool for simulating the upstream influence of atmospheric observations: The Stochastic Time-Inverted Lagrangian Transport (STILT) model. *J. Geophys. Res.* 108, 4493. <https://doi.org/10.1029/2002JD003161>



- National Institute of Environmental Research (NIER) (2017). Annual Report of Air Quality in Korea, 2016. NIER, Ministry of Environment, Republic of Korea.
- Peng, X.L., Choi, M.P.K., Wong, M.H. (2007). Receptor modeling for analyzing PCDD/F and dioxin-like PCB sources in Hong Kong. *Environ. Model. Assess.* 12, 229–237. <https://doi.org/10.1007/s10666-006-9070-6>
- Petelin, D., Grancharova, A., Kocijan, J. (2013). Evolving Gaussian process models for prediction of ozone concentration in the air. *Simul. Model. Pract. Th.* 33, 68–80. <https://doi.org/10.1016/j.simpat.2012.04.005>
- Polissar, A.V., Hopke, P.K., Harris, J.M. (2001a). Source regions for atmospheric aerosol measured at Barrow, Alaska. *Environ. Sci. Technol.* 35, 4214–4226. <https://doi.org/10.1021/es0107529>
- Polissar, A.V., Hopke, P.K., Poirot, R.L. (2001b). Atmospheric aerosol over Vermont: Chemical composition and sources. *Environ. Sci. Technol.* 35, 4604–4621. <https://doi.org/10.1021/es0105865>
- Rasmussen, C.E., Williams, C.K.I. (2006). *Gaussian Processes for Machine Learning*. The MIT Press.
- Rolph, G., Stein, A., Stunder, B. (2017). Real-time environmental applications and display system: READY. *Environ. Modell. Softw.* 95, 210–228. <https://doi.org/10.1016/j.envsoft.2017.06.025>
- Seinfeld, J.H., Pandis, S.N. (2016). *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change*. John Wiley & Sons.
- Shaban, K.B., Kadri, A., Rezk, E. (2016). Urban air pollution monitoring system with forecasting models. *IEEE Sens. J.* 16, 2598–2606. <https://doi.org/10.1109/JSEN.2016.2514378>
- Silverman, B.W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *J. R. Stat. Soc.* 47, 1–21. <https://doi.org/10.1111/j.2517-6161.1985.tb01327.x>
- Stein, A., Draxler, R.R., Rolph, G.D., Stunder, B.J., Cohen, M., Ngan, F. (2015). NOAA's HYSPLIT atmospheric transport and dispersion modeling system. *Bull. Am. Meteorol. Soc.* 96, 2059–2077. <https://doi.org/10.1175/BAMS-D-14-00110.1>
- Stohl, A., Eckhardt, S., Forster, C., James, P., Spichtinger, N., Seibert, P. (2002) A replacement for simple back trajectory calculations in the interpretation of atmospheric trace substance measurements. *Atmos. Environ.* 36, 4635–4648. [https://doi.org/10.1016/S1352-2310\(02\)00416-8](https://doi.org/10.1016/S1352-2310(02)00416-8)
- Trentmann, J., Luderer, G., Winterrath, T., Fromm, M.D., Servranckx, R., Textor, C., Herzog, M., Graf, H.F., Andreae, M.O. (2006) Modeling of biomass smoke injection into the lower stratosphere by a large forest fire (Part I): Reference simulation. *Atmos. Chem. Phys.* 6, 5247–5260. <https://doi.org/10.5194/acp-6-5247-2006>
- Wilson, A., Nickisch, H. (2015). Kernel interpolation for scalable structured Gaussian processes (KISS-GP). In: *Proceedings of the 32<sup>nd</sup> International Conference on Machine Learning*, PMLR 37, pp. 1775–1784.
- Yang, W., Deng, M., Xu, F., Wang, H. (2018). Prediction of hourly PM<sub>2.5</sub> using a space-time support vector regression model. *Atmos. Environ.* 181, 12–19. <https://doi.org/10.1016/j.atmosenv.2018.03.015>
- Yu, T.Y. (2013). Identification of source regions of PM<sub>10</sub> with backward trajectory-based statistical models during PM<sub>10</sub> episodes. *Environ. Monit. Assess.* 185, 6465–6475. <https://doi.org/10.1007/s10661-012-3038-6>
- Zeng, Y., Hopke, P.K. (1989). A study of the sources of acid precipitation in Ontario, Canada. *Atmos. Environ.* 23, 1499–1509. [https://doi.org/10.1016/0004-6981\(89\)90409-5](https://doi.org/10.1016/0004-6981(89)90409-5)
- Zhan, Y., Luo, Y., Deng, X., Chen, H., Grieneisen, M.L., Shen, X., Zhu, L., Zhang, M. (2017). Spatiotemporal prediction of continuous daily PM<sub>2.5</sub> concentrations across China using a spatially explicit machine learning algorithm. *Atmos. Environ.* 155, 129–139. <https://doi.org/10.1016/j.atmosenv.2017.02.023>
- Zhang, Z.Y., Wong, M.S., Lee, K.H. (2015). Estimation of potential source regions of PM<sub>2.5</sub> in Beijing using backward trajectories. *Atmos. Pollut. Res.* 6, 173–177. <https://doi.org/10.5094/APR.2015.020>