Defence Research and Development Canada
Recherche et développement pour la défense Canada

Validation of a Sensor-Driven Modeling Paradigm for Multiple Source Reconstruction with FFT-07 Data

E. Yee

Defence R&D Canada

Technical Report

DRDC Suffield TR 2009-040

May 2009

Canada

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
MAY 2009		

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Validation of a Sensor-Driven Modeling Paradigm for Multiple Source Reconstruction with FFT-07 Data	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Defence R&D Canada - Suffield,PO Box 4000, Station Main,Medicine Hat, AB,Canada T1A 8K6, ,**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

13. SUPPLEMENTARY NOTES
The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**		**64**	

Validation of a Sensor-Driven Modeling Paradigm for Multiple Source Reconstruction with FFT-07 Data

E. Yee
Defence R&D Canada – Suffield

Defence R&D Canada – Suffield

Technical Report

DRDC Suffield TR 2009-040

May 2009

Principal Author

Original signed by E. Yee

E. Yee

Approved by

Original signed by Mr C. Laforce

Mr C. Laforce
Head, Personal Protection Sector

Approved for release by

Original signed by Dr P. D'Agostino

Dr P. D'Agostino
Chair, DRDC Suffield DRP

This work is partially supported by the Chemical Biological Radiological-Nuclear and Explosives Research and Technology Initiative (CRTI) Program under project number CRTI-07-0196TD.

Abstract

A Bayesian probabilistic inferential framework provides a natural and logically consistent method for source reconstruction that fully utilizes the information provided by a limited number of noisy concentration data obtained from a network (or, array) of detectors. This report addresses the application of this framework to the difficult problem of estimating the parameters of an *a priori* unknown number of sources, using an array of detectors. To this purpose, Bayesian probability theory is used to formulate the full joint posterior probability density function for the number of sources and the parameters (e.g., location, emission rate, activation and deactivation times) that describe each source. A simulated annealing algorithm, applied in conjunction with a reversible-jump Markov chain Monte Carlo technique, is used to draw random samples from the posterior probability density function. By calculating the marginal posterior probability distribution of the number of sources from these samples, a *maximum a posteriori* estimate $\widehat{N}_s$ for the number of sources can be obtained, and all samples of source distribution models with exactly $\widehat{N}_s$ discrete sources can be used to provide best estimates for the source parameters (along with their associated uncertainties). The method is validated against a real dispersion experiment involving various combinations of multiple source releases conducted under a multinational cooperative **FU**sing **S**ensor **I**nformation from **O**bserving **N**etworks (FUSION) Field Trial 2007 (FFT-07) undertaken at US Army Dugway Proving Ground (DPG) in September 2007.

Résumé

Le schéma probabiliste inférentiel bayésien est une méthode naturelle et logiquement constante de reconstruction des sources qui utilise au maximum l'information procurée par un nombre limité de données de concentration obtenues d'un réseau de détecteurs. Ce rapport traite d'appliquer ce schéma au problème ardu de l'estimation des paramètres d'un nombre à priori inconnu de sources à l'aide d'un réseau de détecteurs. C'est dans ce but qu'on utilise la théorie bayésienne des probabilités pour formuler la fonction de densité conjointe à postériori pour le nombre de sources et les paramètres (ex. : endroit, taux d'émission, heures d'activation et de désactivation) qui décrivent chaque source. On utilise un algorithme recuit simulé appliqué en conjonction avec une technique Monte Carlo par chaîne de Markov à sauts réversibles pour tirer des échantillons au hasard au moyen de la fonction de probabilité à postériori. En calculant la distribution de probabilité à postériori du nombre de sources à partir de ces échantillons, on peut obtenir *l'estimation du maximum à postériori* $\widehat{N}_s$ du nombre de sources et on peut utiliser tous les échantillons provenant des modèles de distribution de sources avec exactement $\widehat{N}_s$ sources discrètes pour procurer les meilleures estimations concernant les paramètres des sources (ainsi que les fluctuations qui y sont reliées). La méthode a été validée par une expérience consistant en une dispersion réelle comprenant plusieurs combinaisons d'émissions de sources multiples ; cette expérience a été conduite au moyens d'essais pratiques en 2007 (FFT-07) par les réseaux d'observation de capteurs d'information FUSION (**FU***sing* **S***ensor* **I***nformation from* **O***bserving* **N***etworks*) et effectués sur le polygone d'essais de l'arme des États-Unis, Army Dugway Proving Ground (DPG), en septembre 2007.

Executive summary

Validation of a Sensor-Driven Modeling Paradigm for Multiple Source Reconstruction with FFT-07 Data

E. Yee; DRDC Suffield TR 2009-040; Defence R&D Canada – Suffield; May 2009.

Background: A critical capability gap in current emergency and retrospective management efforts directed at terrorist incidents involving the covert release of chemical, biological and radiological (CBR) agents into the atmosphere is the ability to determine the number of sources and, for each of these sources, the location, amount of agent released, and the time of release, following the detection of an event by a network of CBR sensors. This is the so-called source reconstruction problem (also referred to as the source characterization or source inversion problem in various studies). In order to address this capability gap, a multinational collaborative program for the development of methodologies for source reconstruction, involving the fusion of (usually noisy) CBR concentration measurements from remote and deployable networks of sensors with model concentration data obtained from advanced atmospheric dispersion models, has been included as a task under The Technical Cooperation Program (TTCP), CBR Defence Group Technical Panel 9 (TP-9) on Hazard Assessment.

Principal results: In this report, we address the problem of source reconstruction for the difficult case of multiple sources when even the number of sources is unknown *a priori*. A new methodology is developed to solve this problem. This methodology provides simultaneous estimates for the number of sources and for the parameters (e.g., release location, emission rate) that characterize each source. The method is successfully validated against a real dispersion experiment involving various combinations of multiple source releases. This field experiment was conducted under a multinational cooperative **FU**sing **S**ensor **I**nformation from **O**bserving **N**etworks (FUSION) Field Trial 2007 (FFT-07), which was designed and sponsored by TTCP CBR Defence Group TP-9 and conducted at US Army Dugway Proving Ground (DPG) in September 2007.

Significance of results: The algorithm described in this report can be used to interpret agent concentration measurements obtained from an array of CBR detectors in order to estimate unknown source characteristics and quantify uncertainties in this estimation. The technique can be used to estimate the location, emission rate and duration of clandestine CBR agent release(s). Once the sources have been characterized in terms of their various descriptive parameters (e.g., release location, emission rate), this information can be subsequently used with dispersion models to predict the future agent transport in the atmospheric environment and to construct toxic corridors with specified confidence levels in order to support various aspects of the decision-making process for emergency managers and first responders (civilian and military). The comprehensive source reconstruction tool developed here provides a general methodology for the fusion of CBR detector data with model predictions of agent dispersal in the atmosphere. The application of this sensor-driven

modeling paradigm will undoubtedly result in a more complete situational awareness in the operational CBR environment. Indeed, the sensor-driven modeling paradigm developed herein can be integrated potentially into operational warning and reporting (information) systems that combine automated data acquisition, analysis, source reconstruction, display and distribution of CBR hazard prediction and associated decision-support products. This will undoubtedly lead to a greatly improved Common Operating Picture (COP) in the CBR battlespace.

Future work: The next step is to develop a fully operational implementation of the source reconstruction capability described in this report and to incorporate this operational capability into the integrative multiscale urban modeling system implemented in the computational infrastructure at a government operations facility (Environmental Emergency Response Section at Canadian Meteorological Centre). This multiscale urban modeling system was developed under a previous Chemical Biological Radiological-Nuclear and Explosives Research and Technology Initiative (CRTI) Project 02-0093RD and provided an advanced, high-fidelity, validated, state-of-the-science modeling system for the prediction of urban flows and the dispersion of CBR agents released into these flows. The successful completion of the future work proposed here will give a government operations facility the necessary tools to provide a 'whole-of-government' (comprehensive) single authoritative source for expert quality-assured sensor-driven CBR hazard predictions and concomitant decision-support aids, which will form the basis for making decisions for responding to and mitigating hazardous release incidents. These products can be used by emergency managers, planners and first responders (civil and military) in various federal, provincial and municipal agencies for informed response decision making in both domestic and international operations, as well as for support to major events of national and international significance [e.g., Vancouver Winter Olympics, Group of Eight (G8) Summit]. Furthermore, this development is in direct alignment with Defence R&D Canada's Science and Technology Strategy for a Secure Canada and contributes to the solution of a hard problem in the defence and security domain to build a reusable major events security capability.

Sommaire

Validation of a Sensor-Driven Modeling Paradigm for Multiple Source Reconstruction with FFT-07 Data

E. Yee ; DRDC Suffield TR 2009-040 ; R & D pour la défense Canada – Suffield ; mai 2009.

Contexte : Il existe actuellement une lacune grave en termes de capacités en gestion des urgences et efforts rétrospectifs contre les incidents terroristes d'émission invisible d'agents chimiques, biologiques et radiologiques (CBR) dans l'atmosphère. Cette lacune consiste en l'incapacité à déterminer le nombre de sources et pour chacune de ces sources, le lieu, la quantité d'ágent émis et l'heure de l'émission, après qu'un réseau de capteurs CBR ait effectué la détection d'un tel événement. Il s'agit d'un problème de reconstruction des sources (appelé aussi dans certaines études la caractérisation des sources ou problème d'inversion des sources). Pour combler cette lacune, on a inclus un programme de collaboration multinational à la tâche du Programme de coopération technique, Panel 9 (TP-9) du groupe technique de défense CBR d'évaluation des risques. Ce programme consiste en la fusion de mesures de concentrations CBR (habituellement bruyantes) provenant de réseaux de capteurs éloignés et pouvant être déployés dont les modèles de données de concentration sont obtenus à partir de modèles perfectionnés de dispersion atmosphérique.

Résultats principaux : Nous traitons, dans ce rapport, de ce problème de reconstruction des sources dans les cas difficiles de sources multiples dont le nombre est inconnu *a priori*. On a développé une nouvelle méthodologie pour résoudre ce problème ; cette méthodologie procure des estimations simultanées du nombre de sources et de ses paramètres (ex. : le lieu d'émission, le taux d'émission) qui caractérisent chaque source. On a réussi à valider cette méthode en expérimentant sur une dispersion réelle ayant des combinaisons variées de sources multiples d'émission. Cet expérience sur le terrain a été conduite au moyen d'essais pratiques, en 2007, (FFT-07) par les réseaux d'observation de capteurs d'information FUSION (**FU***sing* **S***ensor* **I***nformation from* **O***bserving* **N***etworks*). Ces essais ont été conçus et parrainés par le Groupe TP-9 de la défense CBR du Programme de coopération technique. Ils ont été conduits sur le polygone d'essais de l'arme des États-Unis, Army Dugway Proving Ground (DPG), en septembre 2007.

Portée des résultats : On peut utiliser l'algorithme décrit dans ce rapport pour interpréter les mesures de concentration des agents obtenus d'un réseau de détecteurs CBR en vue d'estimer les caractéristiques inconnues de la source et quantifier les fluctuations de cette estimation. On peut utiliser cette technique pour estimer le lieu, le taux et la durée d'émission d'un agent CBR clandestin ayant été émis. Une fois les sources caractérisées en termes de paramètres descriptifs variés (ex. : lieux et taux d'émission), l'information peut être subséquemment utilisée au moyen de modèles de dispersion visant à prédire le transport futur d'agents dans le milieu atmosphérique et visant à construire des corridors toxiques ayant des coefficients de confiance en mesure de soutenir certains aspects du processus de

prise de décisions des gestionnaires des mesures d'urgences et premiers intervenants (civils et militaires). L'outil de reconstruction complète des sources, mis au point ici, procure une méthodologie générale pour la fusion des données de détecteurs CBR au moyen de modèles de prédiction de la dispersion des agents dans l'atmosphère. L'application de ce paradigme de modèle à base de capteurs résultera sans aucun doute en une reconnaissance situationnelle plus complète lors d'opérations dans un environnement CBR. En réalité, ce paradigme de modèle à base de capteurs, mis au point ici, a le potentiel d'être intégré dans des systèmes opérationnels d'avertissement et de transmission (de l'information) qui combinent l'acquisition automatisée des données avec les analyses, la reconstruction des sources, l'affichage et la distribution des prédictions des risques CBR et des produits d'aide à la décision. Ceci aboutira sans aucun doute à une grande amélioration de l'image commune de la situation opérationnelle (ICSO) dans l'espace de combat CBR.

Perspectives d'avenir : La prochaine étape est de développer l'implémentation totale de la capacité de reconstruction des sources, décrite dans ce rapport et d'incorporer cette capacité opérationnelle au système intégratif de modélisation urbaine à échelles multiples, implémenté dans l'infrastructure computationnelle des installations opérationnelles gouvernementales (Section de réponse aux urgences environnementales du Centre météorologique canadien). Ce système de modélisation urbain à échelles multiples a déjà été mis au point par le Projet 02-0093RD de l'Initiative de recherche et de technologie (CRTI) chimique, biologique, radiologique, nucléaire et explosive ; il s'agit d'un système de modélisation perfectionné, de haute fidélité, validé et d'une science d'avant garde, visant à prédire les flots urbains et la dispersion des agents CBR émis dans ces flots. La réalisation des travaux futurs proposés ici, procurera aux installations opérationnelles gouvernementales, les outils nécessaires facilitant une source gouvernementale autorisée, unique est globale (compréhensive) qui fera des prédictions d'une qualité assurée des risques CBR provenant des capteurs. Ces outils procureront aussi les aides concomitantes qui seront à la base des prises de décisions ayant trait à la réponse et à l'atténuation des incidents d'émission. Ces produits pourront être utilisés par les gestionnaires des mesures d'urgences, les planificateurs et les premiers intervenants (civils et militaires) des agences fédérales, provinciales et municipales qui seront alors en mesure de prendre des décisions éclairées en matière d'opérations domestiques et internationales ainsi qu'en soutien à des événements importants de portée nationale et internationale (ex. : les jeux olympiques d'hiver de Vancouver, le sommet du Groupe des huit (G8)). De plus, ce développement s'aligne directement avec la Stratégie scientifique et technologique pour un Canada sécuritaire de R&D pour la défense Canada et contribue à résoudre le problème difficile en matière de défense et de sécurité consistant à construire une capacité en matière de sécurité des événements importants qui soit réutilisable.

Table of contents

List of figures

List of tables

This page intentionally left blank.

1 Introduction

The development of increasingly more sophisticated sensing technologies for the monitoring of the concentration of hazardous contaminants [e.g., chemical, biological, or radiological (CBR) agents] released into the turbulent atmosphere has generated interest in utilizing this information for the reconstruction of the contaminant source responsible for the observed concentration pattern. More specifically, in public security applications for countering terrorist incidents involving the covert release of a CBR agent in a densely populated urban centre, a critical requirement is the characterization of the unknown source(s) following event detection by a network (array) of CBR sensors. These sensors are placed at different points in space within a designated region in order to function as detectors/monitors to provide quantitative measurements of the concentration of various air admixtures of contaminants.

For example, the Department of Homeland Security (DHS) has deployed (albeit sparse) arrays of biological agent sensors in 31 (with plans to expand to 120) cities across the United States as part of the BioWatch program [1] in order to provide detection and warning of a covert bioterrorism event. In the context of homeland security, the BioWatch program has provided the impetus for recent research efforts directed towards the source reconstruction problem for determination of the location, emission rate, and other characteristics of unknown source(s) of contamination. In consequence, a multinational collaborative program for the development of methodologies for source reconstruction has been included as a task under The Technical Cooperation Program (TTCP), CBR Group Technical Panel 9 on Hazard Assessment.

Mathematically, the source reconstruction problem is an inverse problem. Let $c(\mathbf{x}, t)$ denote the instantaneous concentration at location $\mathbf{x}$ and time t. The ensemble-mean concentration $C(\mathbf{x}, t) \equiv \langle c(\mathbf{x}, t) \rangle$ ($\langle \cdot \rangle$ denotes an ensemble-averaging operation) is related to the source density function $S(\mathbf{x}, t)$ through a Volterra integral equation of the first kind (exact source-receptor relation):

$$C(\mathbf{x}, t) = \int_{t_0}^{t} \int_{\mathbb{R}^3} p(\mathbf{x}, t|\mathbf{x}', t') S(\mathbf{x}', t') \, d\mathbf{x}' \, dt', \tag{1}$$

where for simplicity it is assumed implicitly that at the (arbitrary) initial time t_0, $C(\mathbf{x}, t_0) = 0$. In Eq. (1), $p(\mathbf{x}, t|\mathbf{x}', t') \equiv \langle R_\kappa(\mathbf{x}, t|\mathbf{x}', t') \rangle$ is the kernel function with R_κ determined as the fundamental solution of an advection-diffusion equation:

$$\frac{\partial R_\kappa}{\partial t} + \mathbf{u} \cdot \nabla R_\kappa = \kappa \nabla^2 R_\kappa, \tag{2}$$

with initial condition $R_\kappa(\mathbf{x}, t_0|\mathbf{x}', t_0) = \delta(\mathbf{x} - \mathbf{x}')$ (where $\delta(\mathbf{x})$ is the Dirac delta function). In Eq. (2), $\mathbf{u}$ is the instantaneous velocity field which is assumed to be turbulent (and, hence, random and unpredictable) and κ is the molecular diffusivity which is taken to be constant. The source reconstruction problem is as follows: given $C(\mathbf{x}, t)$ and a specific $p(\mathbf{x}, t|\mathbf{x}', t')$, what is the source density function $S(\mathbf{x}', t')$?

The right-hand-side (RHS) of Eq. (1) defines a formal operator G as follows:

$$(GS)(\mathbf{x},t) = \int_{t_0}^{t} \int_{\mathbb{R}^3} p(\mathbf{x},t|\mathbf{x}',t')S(\mathbf{x}',t')\,d\mathbf{x}'\,dt', \tag{3}$$

which allows us to write Eq. (1) in the symbolic form

$$GS = C. \tag{4}$$

Formally, the source reconstruction problem can be solved by constructing the inverse operator G^{-1}. Unfortunately, this direct mathematical inversion is not possible for a number of reasons. Firstly, there is no comprehensive theory of turbulence so an exact form of $p(\mathbf{x},t|\mathbf{x}',t')$ (and of the operator G) is not available. Secondly, as a pure integral equation the problem may have no solutions (singular operator G), and when it does the solutions may not be uniquely determined. Thirdly, $C(\mathbf{x},t)$ cannot be specified completely and precisely — measurements of C are available only at a finite number of space-time points and these measurements are usually noisy suggesting that even if G^{-1} was known exactly, the presence of noise in the concentration data may introduce instabilities into the solution. For these reasons, the source reconstruction problem using incomplete and noisy concentration data is an ill-posed problem.

A mathematical tool that has been developed for treating instabilities that occur in inverse problems is regularization [2]. In this approach, the class of admissible solutions for S is restricted by imposing further constraints on the solution (although frequently these constraints are somewhat arbitrary and ad-hoc). More specifically, regularization of the source reconstruction problem involves finding the "best" result from among all those source distributions that agree with the incomplete and noisy concentration data, by minimizing a cost functional of the form:

$$J(S) = ||C - GS||^2 + \lambda\Phi(S), \tag{5}$$

where the first term on the RHS of Eq. (5) is a measure of the misfit between the measured (noisy) concentration C and the model concentration GS (where $||\ \cdot\ ||$ is some appropriately defined norm), $\Phi(S)$ is the regularization functional that is used to impose some constraint on the solution, and λ is a regularizing parameter that imposes a relative weight (or, importance) between the data and the constraint.

Robertson and Persson [3] and Robertson and Langner [4] minimized Eq. (5) with $\Phi(S) = 0$, using a four-dimensional (space and time) variational data assimilation method with an adjoint transport equation, to recover emission rate profiles for an experiment with synthetic data and for the European Tracer Experiment (ETEX) [5], respectively. Seibert and Stohl [6] and Seibert [7] use a regularization functional of the form $\Phi(S) = ||S||^2$ (which imposes an upper bound on the "energy" of the source distribution S) to reconstruct the distribution of emission rates for application to the radionuclide monitoring system implemented under the Comprehensive Test Ban Treaty (CTBT) [8]. Thomson et al [9] investigated the use of three different regularization functionals $\Phi(S)$ for source reconstruction, with the minimum of the cost function $J(S)$ obtained using a random search algorithm with simulated annealing.

The three regularization functionals were an entropy functional, a smoothing functional in which the source strength in a grid cell is compared to the average of the source strengths in the neighboring grid cells, and a functional that imposes an upper bound on the total emission flux. Bocquet [10] used entropy as a regularizer (viz., applied the principle of maximum entropy on the mean) and solved the resulting constrained minimization of the cost function by using duality theory to convert the problem to a simpler unconstrained minimization problem for the Lagrangian multipliers (which were introduced to impose constraints on the mean to ensure that the model concentration is consistent with the measured concentration). Finally, Allen et al [11] minimized a misfit measure defined as the squared differenced between the logarithmic measured and modeled concentration with $\Phi(S) = 0$. In effect, these investigators assumed that it was known *a priori* that the unknown source is a continuous point source with unknown source location (x, y) and source strength Q, and used a genetic algorithm to select the source location and strength so as to minimize the misfit measure.

The regularization approach selects a particular source distribution S by minimizing, maximizing, or optimizing some form of cost functional. Unfortunately, the reliability of the inferred S cannot be obtained from any single selection, no matter what optimal properties in terms of quality and utility this single selection of S purportedly embodies. To deal with uncertainty (which reflects also the non-uniqueness of solutions for S using incomplete and noisy concentration data), it is necessary to apply Bayesian probability theory to the problem, rather than simply apply regularization. To this purpose, a probabilistic approach using a Bayesian inferential scheme that allowed the uncertainty in the inference for S to be determined was developed by Yee [12] and demonstrated using Project Prairie Grass data for dispersion over open terrain. This methodology was further developed, refined, and generalized in subsequent work: (1) application of the methodology to complex environments (dispersion in built-up environments) by Yee [13], Keats et al [14] and Chow et al [15]; (2) generalization of the methodology to deal with a non-conservative scalar by Keats et al [16]; and, (3) application of the methodology to source reconstruction for long-range dispersion on continental scales by Yee et al [17]. Yee [18] generalized the methodology to the reconstruction of multiple sources when the number of sources was known *a priori*. Finally, Yee [19], Yee [20] and Yee [21] developed the theory underlying the application of a Bayesian probabilistic inferential framework for addressing the problem of source reconstruction for the difficult case of multiple sources when the number of sources is unknown *a priori*. For a general review of source estimation methods for atmospheric dispersion, the reader is referred to Rao [22].

The aim of this paper is to use some new concentration data, measured by a sensor array consisting of 100 detectors for releases involving multiple sources, to test the procedure proposed by Yee [21] for multiple source reconstruction for the case when the number of sources is unknown *a priori*.

2 Bayesian inference for source reconstruction

In this report, we focus on a source distribution S associated with N_s transient point sources with the k-th source located at a vector position $\mathbf{x}_{s,k}$ and with source activation and deactivation times T_b^k and T_e^k, respectively, between which the source is releasing contaminant at a constant emission rate Q_k ($k = 1, 2, \ldots, N_s$). This source distribution has the following explicit form:

$$S(\mathbf{x}, t) = \sum_{k=1}^{N_s} Q_k \delta(\mathbf{x} - \mathbf{x}_{s,k}) \big[\mathcal{H}(t - T_b^k) - \mathcal{H}(t - T_e^k)\big], \tag{6}$$

where $\mathcal{H}(\,\cdot\,)$ is the Heaviside unit step function. Now, let us assemble the parameters for this particular source distribution (consisting of N_s transient point sources) into the following source parameter vector:

$$\theta_{N_s} \equiv \big(\mathbf{x}_{s,1}, T_b^1, T_e^1, Q_1, \ldots, \mathbf{x}_{s,N_s}, T_b^{N_s}, T_e^{N_s}, Q_{N_s}\big) \in \mathbb{R}^{6N_s}. \tag{7}$$

Furthermore, let $\Theta \equiv (N_s, \theta_{N_s})$.

The measured mean concentration $d_{i,j^{(i)}}$ obtained by a detector at receptor location $\mathbf{x}_{d_i}$ and time $t_{d_j}^{(i)}$ is assumed be the sum of a modeled mean concentration signal $\overline{C}(\mathbf{x}_{d_i}, t_{d_j}^{(i)}; \Theta)$ and noise $e_{i,j^{(i)}}$:

$$d_{i,j^{(i)}} = \overline{C}(\mathbf{x}_{d_i}, t_{d_j}^{(i)}; \Theta) + e_{i,j^{(i)}}, \qquad i = 1, 2, \ldots, N_d \;\text{ and } j = 1, 2, \ldots, N_t^{(i)}, \tag{8}$$

where N_d is the number of detectors, $N_t^{(i)}$ is the number of concentration time samples measured at the i-th detector and

$$\overline{C}(\mathbf{x}_{d_i}, t_{d_j}^{(i)}; \Theta) = \int_{t_0}^{T} dt \int_{\mathcal{D} \subset \mathbb{R}^3} d\mathbf{x}\; C(\mathbf{x}, t) h(\mathbf{x}, t | \mathbf{x}_{d_i}, t_{d_j}^{(i)}) \equiv \langle\!\langle C | h \rangle\!\rangle (\mathbf{x}_{d_i}, t_{d_j}^{(i)}), \tag{9}$$

is the expected (mean) concentration "seen" by the detector at location $\mathbf{x}_{d_i}$ and time $t_{d_j}^{(i)}$, $\mathcal{D} \times [t_0, T]$ is the space-time volume enclosing the source distribution S and the detectors, and $C(\mathbf{x}, t)$ is the ensemble-mean concentration determined in accordance to Eq. (1) for the source distribution S given by Eq. (6). In Eq. (9), $h(\mathbf{x}, t | \mathbf{x}_d, t_d)$ is the spatial-temporal filtering function (of $\mathbf{x}$ and t) for the mean concentration measurement made by the detector at location $\mathbf{x}_d$ and time t_d with

$$\int_{t_0}^{T} dt \int_{\mathcal{D}} d\mathbf{x}\; h(\mathbf{x}, t | \mathbf{x}_d, t_d) = 1. \tag{10}$$

Note from Eq. (9) that $\overline{C}$ can be determined as the inner (or scalar) product $\langle\!\langle C | h \rangle\!\rangle$ of the mean concentration C and the detector spatial-temporal filtering function h.

To simplify the notation, we rewrite the measurement model of Eq. (8) as follows:

$$d_J = \overline{C}_J(\Theta) + e_J, \qquad J = 1, 2, \ldots, N, \tag{11}$$

where $N \equiv \sum_{i=1}^{N_d} N_t^{(i)}$ is the total number of measured concentration data and $\overline{C}_J(\Theta) \equiv \overline{C}(\mathbf{x}_{d_i}, t_{d_j}^{(i)}; \Theta)$. In Eq. (11), the index J is used to denote the label $(i, j^{(i)})$ (ordered in some regular or convenient manner). With this background, the problem of source reconstruction reduces to the following problem: estimate N_s (number of sources) and θ_{N_s} (parameters for each source) or, equivalently, estimate Θ given the concentration data $\mathbf{D} \equiv (d_1, d_2, \ldots, d_N)$.

Yee [21] developed a new method based on Bayesian probability theory for the inference of Θ. The components of the Bayesian inference scheme for source reconstruction are shown in Figure 1. Bayesian probability theory can be derived from more fundamental principles starting with the formulation of a small number of requirements that any theory of plausibility (or inference) ought to verify. These requirements (desiderata) were first provided by Cox [23], with an eloquent description of the complete development described by Jaynes [24] in his definitive treatise. Within the context of the source reconstruction problem formulated above, Bayes' theorem yields the following result:

$$p(\Theta|\mathbf{D}, I) = \frac{p(\Theta|I)p(\mathbf{D}|\Theta, I)}{p(\mathbf{D}|I)}, \tag{12}$$

where I is the background (contextual) information available in the problem (e.g., model that defines the mapping from a source distribution S to the concentration C, background meteorology). The various factors that appear in Eq. (12) have the following interpretation. Firstly, $p(\Theta|I)$ is the prior probability density function (PDF) for a proposition Θ about the source, predicated on the contextual information specified by I, with "|" denoting "conditional upon". The prior PDF encodes all the prior information about the source before receipt of the concentration data $\mathbf{D}$. Secondly, $p(\mathbf{D}|\Theta, I)$ is the likelihood function and is the probability that we observe the concentration data $\mathbf{D}$, when Θ is known exactly (viz., the source distribution is known). Thirdly, $p(\mathbf{D}|I)$ is referred to as the evidence and, in our case here, is simply a normalization constant. Finally, $p(\Theta|\mathbf{D}, I)$ is the posterior PDF for the proposition Θ about the source, in light of the new information introduced through the newly acquired concentration data $\mathbf{D}$.

We are seeking the posterior PDF $p(\Theta|\mathbf{D}, I)$, which encodes our inferences about Θ. Because $p(\mathbf{D}|I)$ is simply a normalization constant, the posterior PDF of interest in Eq. (12) can be specified within a normalization constant as

$$p(\Theta|\mathbf{D}, I) \propto p(\Theta|I)p(\mathbf{D}|\Theta, I). \tag{13}$$

The problem now reduces to the assignment of $p(\Theta|I)$ (prior distribution) and $p(\mathbf{D}|\Theta, I)$ (likelihood function).

For the prior distribution, the logical independence of the source parameters is assumed, in which case $p(\Theta|I)$ factorizes as follows:

$$p(\Theta|I) \equiv p\big(N_s, \theta_{N_s}|I\big) = p(N_s|I) \prod_{k=1}^{N_s} p(Q_k|I)p(\mathbf{x}_{s,k}|I)p(T_b^k|I)p(T_e^k|T_b^k, I). \tag{14}$$

The prior on the number of sources $p(N_s|I)$ is chosen to be a binomial distribution with parameter $p^* \in [0, 1]$ (binomial rate), and with a domain of definition between $N_{s,\min}$

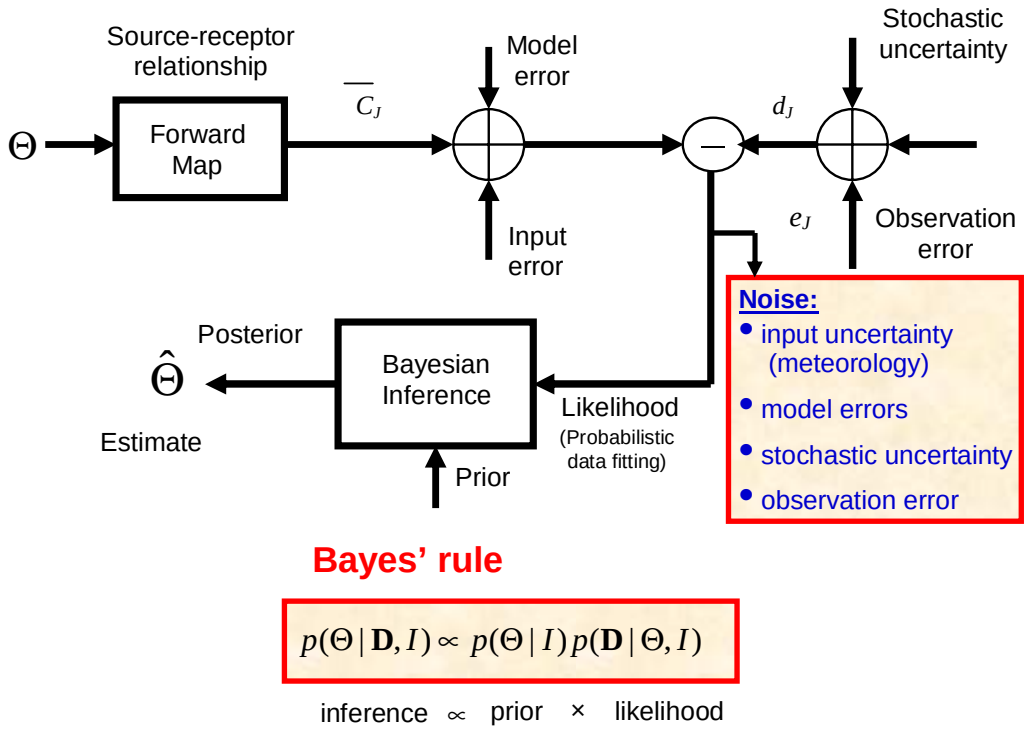


Figure 1: *The components of the Bayesian inference scheme for source reconstruction.*

(minimum number of sources) and $N_{s,\max}$ (maximum number of sources); viz., $p(N_s|I) \sim B(p^*; N_{s,\min}, N_{s,\max})$[1] where the symbol '$\sim$' denotes "is distributed as". The prior on emission rate $p(Q_k|I)$ for $k = 1, 2 \ldots, N_s$ is chosen to be a Bernoulli-uniform mixture: $p(Q_k|I) \sim \mathcal{BU}(\gamma; Q_{\max})$ where γ is the probability that the source is turned on (viz., $\Pr\{Q_k > 0\} = \gamma$) and $Q_{\max}$ is an *a priori* upper bound on the expected emission rate.[2] The prior on the source location $p(\mathbf{x}_{s,k}|I)$ for $k = 1, 2 \ldots, N_s$ is chosen to be uniform (flat) over the some spatial region $\mathcal{D} \subset \mathbb{R}^3$, so $p(\mathbf{x}_{s,k}|I) \sim \mathcal{U}(\mathcal{D})$. Finally, the priors on the source activation (on) and deactivation (off) times for the k-th source are chosen to be uniform over $[t_0, T_{\max}]$ and $[T_b^k, T_{\max}]$, respectively, where $T_{\max}$ is an upper bound on the time at which the source was turned on or off (viz., $p(T_b^k|I) \sim \mathcal{U}([t_0, T_{\max}])$ and $p(T_e^k|T_b^k, I) \sim \mathcal{U}([T_b^k, T_{\max}])$ for $k = 1, 2, \ldots, N_s$). Note that the domain of definition for

[1] The binomial distribution $B(p^*; N_{s,\min}, N_{s,\max})$ has a probability distribution function defined as follows:

$$p(N_s|I) = \frac{(N_{s,\max} - N_{s,\min})!}{(N_s - N_{s,,\min})!(N_{s,\max} - N_s)!}\, p^{*(N_s - N_{s,\min})}(1 - p^*)^{N_{s,\max} - N_s},$$

for $N_s = N_{\mathrm{s,min}}, N_{\mathrm{s,min}} + 1, \ldots, N_{\mathrm{s,max}}$. Note that in this definition, the standard form of the binomial distribution has been offset by the minimum number of sources $N_{\mathrm{s,min}}$.

[2] More specifically, a Bernoulli-uniform mixture model for Q_k has the following form:

$$p(Q_k|I) = (1 - \gamma)\delta(Q_k) + \gamma \mathbb{I}_{(0,Q^*)}(Q_k)/Q^*,$$

$(k = 1, 2, \ldots, N_s)$ and $\mathbb{I}_A(x)$ is the indicator function for set A, with $\mathbb{I}_A(x) = 1$ if $x \in A$ and $\mathbb{I}_A(x) = 0$ if $x \notin A$.

the uniform distribution assigned to $p(T_e^k|T_b^k, I)$ explicitly encodes the fact that the time the k-th source is turned off must occur after it has been turned on.

The likelihood function $p(\mathbf{D}|\Theta, I)$ encodes all the information provided by the concentration data about the unknown source distribution S. The "noise" component e_J in Eq. (11) represents the difference (residual) between the measured and modeled (or, predicted) mean concentration. It is assumed that the only information we have about the noise component e_J is that it has a known noise power (or variance) σ_J^2. The variance σ_J^2 of this noise can be decomposed into four basic contributions as discussed by Rao [25] (see Figure 1): (1) model errors arising from uncertainties in the representation of various physical processes (e.g., turbulent diffusion) in the dispersion model used to predict $C(\mathbf{x}, t)$; (2) input error arising from uncertainties in the values of model parameters and/or in the specification of the meteorology used to "drive" the dispersion model; (3) stochastic uncertainty arising from the turbulent nature of the atmosphere; and, (4) measurement noise inherent in the concentration detector. In spite of the complexity of the noise structure, application of the principle of maximum entropy (see Jaynes [24]) to our state of knowledge concerning the noise, results in the following Gaussian form for the likelihood function:

$$p(\mathbf{D}|\Theta, I) = \frac{1}{\prod_{J=1}^{N} \sqrt{2\pi}\sigma_J} \exp\left(-\frac{1}{2}\chi^2(\Theta)\right), \tag{15}$$

where

$$\chi^2(\Theta) \equiv \sum_{J=1}^{N} \left(\frac{d_J - \overline{C}_J(\Theta)}{\sigma_J}\right)^2. \tag{16}$$

Using the forms for the prior distribution and the likelihood function assigned above, the posterior distribution of Eq. (13) can be written as follows:

$$\begin{aligned} p(\Theta|\mathbf{D}, I) &\equiv p(N_s, \theta_{N_s}|\mathbf{D}, I) \\ &\propto \frac{1}{\prod_{J=1}^{N} \sqrt{2\pi}\sigma_J} \exp\left(-\frac{1}{2}\sum_{J=1}^{N}\left(\frac{d_J - \overline{C}_J(\Theta)}{\sigma_J}\right)^2\right) \\ &\quad \times \frac{(N_{s,\max} - N_{s,\min})!}{(N_s - N_{s,,\min})!(N_{s,\max} - N_s)!}\, p^{*(N_s - N_{s,\min})}(1-p^*)^{N_{s,\max} - N_s} \\ &\quad \times \prod_{k=1}^{N_s}\Big[(1-\gamma)\delta(Q_k) + \gamma \mathbb{I}_{(0,Q_{\max})}(Q_k)/Q_{\max}\Big] \\ &\quad \times \mathbb{I}_{\mathcal{D}}(\mathbf{x}_{s,k})\mathbb{I}_{(t_0,T_{\max})}(T_b^k)\frac{\mathbb{I}_{(T_b^k,T_{\max})}(T_e^k)}{(T_{\max} - T_b^k)}. \end{aligned} \tag{17}$$

The posterior PDF $p(\Theta|\mathbf{D}, I)$ embodies the state of knowledge about the source parameters given the prior information encoded in $p(\Theta|I)$ and the newly acquired concentration data $\mathbf{D}$, the latter of which modulates our prior belief about Θ through the likelihood function $p(\mathbf{D}|\Theta, I)$. In consequence, $p(\Theta|\mathbf{D}, I)$ allows us to estimate all the interesting statistics about the source parameter Θ. More specifically, the posterior distribution $p(\Theta|\mathbf{D}, I)$ allows

one to summarize its information through a few statistics such as location and dispersion measures which provide quantitative information on the central (or, "best") value of a source parameter and the uncertainty associated with this value.

For example, the maximum a posteriori estimate for the number of (discrete) sources can be obtained from

$$\widehat{N}_s = \underset{N_s}{\operatorname{argmax}}\ p(N_s|\mathbf{D}, I), \tag{18}$$

where $p(N_s|\mathbf{D}, I)$ is the (marginal) posterior probability for the number of sources, conditioned on the concentration data and the background information. Given an estimate $\widehat{N}_s$ for the number of sources, posterior quantities of interest about $\theta_{\widehat{N}_s}$ involve expectation values of the following form:

$$\langle \psi(\theta_{\widehat{N}_s}) \rangle = \int \psi(\theta_{\widehat{N}_s}) p(\widehat{N}_s, \theta_{\widehat{N}_s}|\mathbf{D}, I)\, d\theta_{\widehat{N}_s}\,. \tag{19}$$

In particular, the choice $\psi(\theta_{\widehat{N}_s}) = \theta_{\widehat{N}_s}$ gives the posterior mean of the source parameters, which can be used as an estimate for these parameters; viz., $\widehat{\theta}_{\widehat{N}_s} = \langle \theta_{\widehat{N}_s} \rangle$.

Similarly, the choice $\psi(\theta_{\widehat{N}_s}) = \left(\theta_{\widehat{N}_s} - \langle \theta_{\widehat{N}_s} \rangle\right)^2$, gives the posterior variance of the source parameters, whose square root (or, posterior standard deviation) can be used as a measure of uncertainty in the estimate for these parameters; viz., $\sigma(\theta_{\widehat{N}_s}) = \left\langle \left(\theta_{\widehat{N}_s} - \langle \theta_{\widehat{N}_s} \rangle\right)^2 \right\rangle^{1/2}$. Alternatively, a p% highest posterior distribution (HPD) interval that encloses a source parameter with p% probability, and constructed so that the lower and upper bounds of the specified interval are such that the probability density function within the interval is everywhere larger than outside it, can be used as a uncertainty specification for a source parameter.

3 Computational framework

This section describes the computational procedures which are used for extracting the source parameter estimates required for event reconstruction. In this report, the background treatment on the computational methodology is necessarily brief. The reader is referred to Yee et al [17] and Yee [21] for a more complete description of this topic. There are two major issues in the computational framework applied to Bayesian inference for source reconstruction that need to be addressed; namely, (1) a computationally efficient methodology for the computation of the source-receptor relationship required in the determination of the likelihood function, and (2) a methodology for sampling from the posterior distribution for the source parameters.

3.1 Fast computation of source-receptor relationship

To apply the Bayesian inference methodology to source reconstruction, we need to relate the hypotheses of interest about the unknown source distribution (encoded in Θ) to the available concentration data d_J measured by the network (or array) of detectors. This requires the

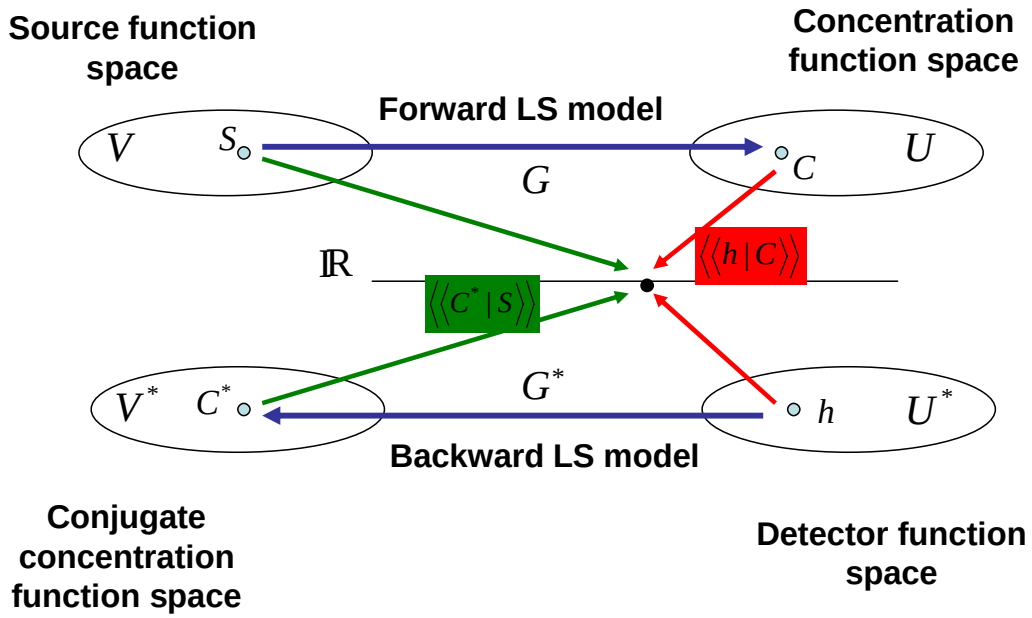


Figure 2: *Commutative diagram illustrating two mathematically equivalent representations for the source-receptor relationship.*

computation of a modeled (or predicted) concentration $\overline{C}_J$ as prescribed by Eqs (9) and (1). This is the source-receptor relationship (realized in terms of an atmospheric dispersion model) that describes the mapping $\mathcal{M}_{\mathrm{SR}}$ from the hypothesis space $\mathcal{H}$ of source distributions to the concentration sample space S^N of concentration data, so $\mathcal{M}_{\mathrm{SR}} : \Theta \in \mathcal{H} \rightarrow \mathbf{D} \in S^N$ (N is the number of concentration data).

The likelihood function given in Eqs (15) and (16) is not a closed-form expression and its evaluation is computationally expensive owing to the fact that $\overline{C}_J$ ($J = 1, 2, \ldots, N$) needs to be determined for a given source distribution Θ [using Eqs (9) and (1)]. Moreover, a simulation-based posterior inference using Markov chain Monte Carlo sampling requires a large number of computations of the source-receptor relationship to be undertaken. In consequence, a fast and efficient technique for performing computations of the source-receptor relationship (for any given source distribution Θ) is required for the rapid sampling of the posterior distribution. To this purpose, Keats et al [14] and Yee et al [17] described a computationally efficient methodology for determination of the source-receptor relationship using an adjoint representation for this relationship.

Figure 2 illustrates two mathematically equivalent (dual) representations for the source-receptor relationship $\mathcal{M}_{\mathrm{SR}}$. If we interpret the source distribution S (which we encode as Θ) as a vector in V (vector space of source functions), then we can construct a vector (function) C^* which is dual (or conjugate) to S belonging to the dual vector space V^* (or, conjugate concentration function space). This vector space is identified to be the space of all linear functionals $C^* : V \rightarrow \mathbb{R}$. There is a one-to-one correspondence between the vector

$S \in V$ and the dual vector $C^* \in V^*$, and this correspondence can be defined through the scalar product $\langle\langle C^*|S\rangle\rangle$ that pairs S with C^*.[3]

Similarly, C and h [see Eq. (9)] can be interpreted as dual (or conjugate) vectors lying in the concentration function space U and detector function space U^*, respectively, with U^* being the dual vector space to U with the one-to-one correspondence between elements of these two spaces defined through the scalar product $\langle\langle h|C\rangle\rangle$ with $C \in U$ and $h \in U^*$. More interestingly, C can be paired with its dual h and C^* can be paired with its dual S such that the duality relationship $\langle\langle C|h\rangle\rangle = \langle\langle h|C\rangle\rangle = \langle\langle C^*|S\rangle\rangle$ is exactly satisfied. Now, S can be related to C through the mapping G (viz., $C = G(S) \equiv GS$) which corresponds to the direct (source-oriented) representation of the source-receptor relationship. However, a mathematically equivalent representation of the source-receptor relationship can be formulated by relating h to C^* through the adjoint mapping G^* (viz., $C^* = G^*(h) \equiv G^*h$)[4] with G^* explicitly constructed so that the duality relation is exactly satisfied: $\langle\langle C^*|S\rangle\rangle = \langle\langle G^*h|S\rangle\rangle = \langle\langle h|GS\rangle\rangle = \langle\langle h|C\rangle\rangle = \langle\langle C|h\rangle\rangle$, for any source $S \in V$ and any receptor $h \in U^*$. This is the dual (receptor-oriented) representation for the source-receptor relationship.

In more concrete terms, the dual (adjoint) representation of the source-receptor relationship given in Eq. (9) can be expressed explicitly as follows:

$$\begin{aligned}
\overline{C}(\mathbf{x}_{d_i}, t_{d_j}^{(i)}; \Theta) &\equiv \int_{t_0}^{T} dt \int_{\mathcal{D}} d\mathbf{x}\; C(\mathbf{x},t) h(\mathbf{x},t|\mathbf{x}_{d_i}, t_{d_j}^{(i)}) \equiv \langle\langle C|h\rangle\rangle(\mathbf{x}_{d_i}, t_{d_j}^{(i)}) \\
&= \int_{-\infty}^{T} dt' \int_{\mathcal{D}} d\mathbf{x}'\, C^*\big(\mathbf{x}', t'|\mathbf{x}_{d_i}, t_{d_j}^{(i)}\big) S(\mathbf{x}', t') \\
&\equiv \langle\langle C^*|S\rangle\rangle(\mathbf{x}_{d_i}, t_{d_j}^{(i)}),
\end{aligned} \tag{20}$$

where $C^*(\mathbf{x}', t'|\mathbf{x}_{d_i}, t_{d_j}^{(i)})$ is an adjunct (dual) "concentration" at space-time point $(\mathbf{x}', t')$ associated with the sensor concentration data at location $\mathbf{x}_{d_i}$ and time $t_{d_j}^{(i)}$. In the source-oriented approach, the computation of $\overline{C}_J$ using Eqs (9) and (1) requires the determination of $C(\mathbf{x}, t)$. The latter quantity requires the specification of the kernel function $p(\mathbf{x}, t|\mathbf{x}', t')$. Unfortunately, we have nothing approaching a comprehensive theory of turbulent diffusion and, as a consequence, there are no known exact solutions for $p(\mathbf{x}, t|\mathbf{x}', t')$ for the case of complex turbulent flows (e.g., atmospheric flows). Approximations for the kernel function $p(\mathbf{x}, t|\mathbf{x}', t')$ can be obtained using either an Eulerian or Lagrangian description of atmospheric diffusion.

In this report, we will use a first-order Lagrangian stochastic (LS) trajectory simulation method to approximate $p(\mathbf{x}, t|\mathbf{x}', t')$ (and thence, $C(\mathbf{x}, t)$). To this purpose, we consider

[3] In fact, the scalar product $\langle\langle C^*|S\rangle\rangle$ can be interpreted as an explicit mathematical representation for collection of all linear functionals $C^* : V \to \mathbb{R}$ that can be defined on V.

[4] In functional analysis, a linear operator $G^* : U^* \to V^*$ is called the dual or pull-back of the linear operator $G : V \to U$ if $\langle\langle G^*h|S\rangle\rangle = \langle\langle h|GS\rangle\rangle,\ \forall S \in V, h \in U^*$.

Thomson's [26] model for inhomogeneous Gaussian turbulence, where the forward increments of the velocity $\mathbf{U}(t) \equiv (U_i(t))$ and position $\mathbf{X}(t) \equiv (X_i(t))$ $(i = 1, 2, 3)$ of a marked fluid element (particle) are given by the following stochastic differential equation:[5]

$$\begin{aligned} d\mathbf{X}(t) &= \mathbf{U}(t)\, dt, \\ d\mathbf{U}(t) &= \mathbf{a}(\mathbf{X}(t), \mathbf{U}(t), t)\, dt + \big(C_0\epsilon(\mathbf{X}(t), t)\big)^{1/2}\, d\mathbf{W}(t), \end{aligned} \tag{21}$$

where ϵ is the turbulence kinetic energy (TKE) dissipation rate, C_0 is the Kolmogorov "universal" constant, $d\mathbf{W}(t) \equiv (dW_i(t))$ are the increments of a vector-valued (three-dimensional) Wiener process, and $\mathbf{a} \equiv (a_i)$ is the drift coefficient vector which assumes the following form:

$$a_i = -\frac{1}{2}\big(C_0\epsilon\big)\Gamma_{ik}^{-1}\big(u_k - \overline{U}_k\big) + \frac{\phi_i}{P_E}, \tag{22}$$

where

$$\begin{aligned} \frac{\phi_i}{P_E} \equiv\ & \frac{1}{2}\frac{\partial \Gamma_{il}}{\partial x_l} + \frac{\partial \overline{U}_i}{\partial t} + \overline{U}_l \frac{\partial \overline{U}_i}{\partial x_l} \\ & + \left(\frac{1}{2}\Gamma_{lj}^{-1}\left[\frac{\partial \Gamma_{il}}{\partial t} + \overline{U}_m \frac{\partial \Gamma_{il}}{\partial x_m}\right] + \frac{\partial \overline{U}_i}{\partial x_j}\right)\big(u_j - \overline{U}_j\big) \\ & + \frac{1}{2}\Gamma_{lj}^{-1}\frac{\partial \Gamma_{il}}{\partial x_k}\big(u_j - \overline{U}_j\big)\big(u_k - \overline{U}_k\big). \end{aligned} \tag{23}$$

Here, $\overline{U}_i$ is the mean Eulerian velocity, $\Gamma_{ij} \equiv \overline{(u_i - \overline{U}_i)(u_j - \overline{U}_j)}$ is the Reynolds stress tensor (where an overline is used to denote an ensemble average), and P_E is the background (Eulerian) velocity PDF (which is implicitly assumed here to possess a Gaussian form). In this source-oriented approach, marked particles with initial space-time coordinates $(\mathbf{x}_i, t_i)$ are sampled from a space-time density function that is proportional to the source distribution $S(\mathbf{x}, t)$, and the forward-time trajectories of these tagged fluid particles are computed using Eqs (21) to (23) for $t > t_i$ (where t_i is the initial time at which a tagged particle was released from $S(\mathbf{x}, t)$). The displacement statistics of these particles are used to determine $C(\mathbf{x}, t)$.

For source reconstruction, we use the receptor-oriented approach described by Eq. (20) for the efficient calculation of $\overline{C}_J$, which requires the determination of the adjunct (or dual) concentration $C^*\big(\mathbf{x}', t' | \mathbf{x}_{d_J}, t_{d_J}\big)$. To this purpose, it was shown by Thomson [26] and Flesch et al [27] that the following backward-time Lagrangian trajectory simulation model is the dual to the forward-time Lagrangian trajectory simulation model given by Eqs (21) to (23):

$$\begin{aligned} d\mathbf{X}^b(t') &= \mathbf{U}^b(t')\, dt', \\ d\mathbf{U}^b(t') &= \mathbf{a}^b(\mathbf{X}^b(t'), \mathbf{U}^b(t'), t')\, dt' + \big(C_0\epsilon(\mathbf{X}^b(t'), t')\big)^{1/2}\, d\mathbf{W}(t'), \end{aligned} \tag{24}$$

with

$$a_i^b = \frac{1}{2}\big(C_0\epsilon\big)\Gamma_{ik}^{-1}\big(u_k - \overline{U}_k\big) + \frac{\phi_i}{P_E}. \tag{25}$$

[5] The forward increments of the marked particle position and velocity are defined as $d\mathbf{X}(t) \equiv \mathbf{X}(t + dt) - \mathbf{X}(t)$ and $d\mathbf{U}(t) \equiv \mathbf{U}(t + dt) - \mathbf{U}(t)$, respectively, with $dt > 0$.

In Eq. (24), $d\mathbf{X}^b(t') \equiv \mathbf{X}^b(t') - \mathbf{X}^b(t'-dt')$ and $d\mathbf{U}^b(t') \equiv \mathbf{U}^b(t') - \mathbf{U}^b(t'-dt')$ $(dt' > 0)$ are the backward increments of the position $\mathbf{X}^b(t')$ and velocity $\mathbf{U}^b(t')$ of a marked particle at time t' along a backward-time trajectory. In this receptor-oriented (dual) approach, marked particles with final space-time coordinates are sampled from a space-time density function that is proportional to the sensor response function $h(\mathbf{x}', t'|\mathbf{x}_d, t_d)$ [at the detector space-time location $(\mathbf{x}_d, t_d)$], and the backward-time trajectories of these tagged fluid particles are computed using Eqs (24) and (25) for $t' < t_f$ (where t_f is the final time at which a tagged particle was released from $h(\mathbf{x}', t'|\mathbf{x}_d, t_d)$). The displacement statistics of these particles are used to determine $C^*(\mathbf{x}', t'|\mathbf{x}_d, t_d)$.

Substituting Eq. (6) into Eq. (9), the model concentration "seen" by the detector at space-time point $(\mathbf{x}_{d_i}, t_{d_j}^{(i)})$ is given explicitly as follows:

$$\overline{C}_J(\Theta) \equiv \overline{C}(\mathbf{x}_{d_i}, t_{d_j}^{(i)}; \Theta) = \sum_{k=1}^{N_s} Q_k \int_{T_b^k}^{\min(T, T_e^k)} C^*(\mathbf{x}_{s,k}, t_s|\mathbf{x}_{d_i}, t_{d_j}^{(i)})\, dt_s. \tag{26}$$

It should be noted that the computation of $\overline{C}_J(\Theta)$ can be obtained for any source distribution S (encoded as the parameter vector Θ), without having to re-compute C^*. Indeed, it should be emphasized that because C^* does not depend on the source distribution, it can be pre-calculated using the backward-time LS trajectory simulation model for each available detector space-time location, and this pre-calculated C^* can be used in Eq. (26) for a computationally efficient determination of $\overline{C}_J(\Theta)$ $(J = 1, 2, \ldots, N)$ required for the rapid evaluation of the likelihood function $p(\mathbf{D}|\Theta, I)$.

3.2 Markov chain Monte Carlo sampling

All the information arising from the application of Bayesian probability theory to the problem of source reconstruction is embodied in the posterior PDF of the parameters $\Theta \equiv (N_s, \theta_{N_s}) \in \mathbb{R}^{6N_s+1}$ [see Eq. (17)] that define the source distribution S. The posterior quantities of interest are expectation values with a generic form given by Eq. (19), which involves an integration of the product of an arbitrary function $\psi(\theta_{N_s})$ and $p(N_s, \theta_{N_s}|\mathbf{D}, I)$ (for a fixed value of N_s) over a subset of the parameter space involving $\theta_{N_s} \in \mathbb{R}^{6N_s}$. Unfortunately, in potentially high-dimensional spaces ($N_s \gg 1$), a numerical integration of Eq. (19) involving the evaluation of the integrand on some grid in the θ_{N_s}-space would be prohibitively expensive (and indeed, even for the case $N_s = 1$ implying $\theta_{N_s} \in \mathbb{R}^6$, the "curse of dimensionality" renders the quadrature practically infeasible).

One method for overcoming this "curse of dimensionality" is given by the application of Markov chain Monte Carlo (MCMC) algorithms for posterior sampling (see Gilks et al [28] and Gelman et al [29]). To this purpose, Yee [21] described the formulation of a reversible-jump MCMC (RJMCMC) algorithm applied with parallel tempering for generating samples from the posterior distribution $p(\Theta|\mathbf{D}, I) \equiv p(N_s, \theta_{N_s}|\mathbf{D}, I)$ given by Eq. (17). As a consequence, only the important details of the algorithm that are required to understand the following Bayesian analysis of field trial data will be presented here.

The objective of MCMC sampling is to construct an auxiliary Markov chain whose stationary (or, invariant) distribution is the posterior distribution $p(\Theta|\mathbf{D}, I)$ of the source parameters $\Theta \equiv (N_s, \theta_{N_s})$. The difficulty in the construction of the Markov chain in the current application resides in the fact that number of sources N_s in unknown *a priori*, so the dimension of the hypothesis (or, parameter) space is unknown (viz., $\theta_{N_s} \in \mathbb{R}^{6N_s}$ with N_s unknown). More specifically, with reference to the posterior distribution $p(\Theta|\mathbf{D}, I)$ given by Eq. (17), it is necessary to consider $(N_{s,\max} - N_{s,\min} + 1)$ candidate models for the source distribution S and associated with each of these candidate models is a posterior distribution $p(N_s, \theta_{N_s}|\mathbf{D}, I)$ depending upon an unknown parameter vector $\theta_{N_s} \in \mathbb{R}^{N_s}$ where $N_s \in \{N_{s,\min}, N_{s,\min}+1, \ldots, N_{s,\max}\}$ is a model indicator that defines the dimension $(6N_s)$ of the hypothesis (parameter) space. In order to allow changes in the dimensionality of the model, a reversible-jump MCMC algorithm is used to construct the Markov chain for Θ. The formalization of RJMCMC algorithms for dealing with variable dimension models has been described by Green [30].

Consider a Markov chain with "state" vector $\{\Theta^{(t)}\} \equiv \{N_s^{(t)}, \theta_{N_s^{(t)}}\}$ $(t = 0, 1, 2, \ldots)$ that is constructed so that its stationary distribution coincides with $p(\Theta|\mathbf{D}, I)$ given by Eq. (17). After convergence of the Markov chain to its stationary distribution, the samples drawn from this chain can be used to estimate any posterior statistic of interest. The construction of the Markov chain uses a RJMCMC algorithm in which the MCMC moves are separated into two categories; namely, propagation moves which do not change the dimensionality of the source distribution and trans-dimensional jump moves which change the source distribution by ± 1 discrete source (source atom). In the current application, the trans-dimensional jump move changes the dimensionality of the hypothesis space by ± 6 (as each source atom in Eq. (6) involves six degrees of freedom).

For the dimension-conserving moves, we partition the parameter vector θ_{N_s} (N_s fixed) as follows: $\theta_{N_s} = (\theta^1, \theta^2)$ where $\theta^1 \equiv (Q_1, Q_2, \ldots, Q_{N_s}) \in \mathbb{R}^{N_s}$ and $\theta^2 = (\mathbf{x}_{s,1}, T_b^1, T_e^1, \ldots, \mathbf{x}_{s,N_s}, T_b^{N_s}, T_e^{N_s}) \in \mathbb{R}^{5N_s}$. The parameters associated with θ^1 are *linearly* related to the model concentration data, as is evident from Eq. (26). For these parameters, a Gibbs sampler is used for the update (propagation) move. More specifically, the Gibbs sampler updates Q_k as a direct draw from the univariate full conditional posterior distribution $p(Q_k|\theta^1_{-k}, \theta^2, \mathbf{D}, I)$ where θ^1_{-k} is the vector θ^1 with its k-th component removed $(k = 1, 2, \ldots, N_s)$. From Eq. (17), it can be shown that the full conditional posterior distribution for Q_k assumes a simple Bernoulli-Gaussian (truncated) distribution which can be sampled from directly (see Yee [21]).

The remaining parameters in θ^2 (e.g., source location, source on time, source off time) are related *non-linearly* to the model concentration. In consequence, the Gibbs sampler cannot be used to update these parameters owing to the fact that the full conditional posterior distribution for these parameters cannot be determined analytically. Furthermore, even if this distribution can be determined analytically, it would correspond to a non-standard probability distribution from which it is not easy (straightforward) to draw a random sample. For these reasons, the Metropolis-Hastings (M-H) sampler is used to update the parameters in θ^2. To this purpose, we update $\theta^2_l \in \theta^2$ $(l = 1, 2 \ldots, 5N_s)$ by randomly

sampling a new candidate $\theta_l^{2'}$ from a proposal distribution that is taken to be a mixture of Gaussian distributions, each with mean θ_l^2 and different variances β_l^2. The variances in this mixture of Gaussian distributions are chosen typically to cover several orders of magnitude. For the M-H steps here, we used a mixture consisting of seven Gaussian distributions with the base standard deviation β_l for a particular parameter θ_l^2 taken to be equal to 0.01 times the 'length' of the domain of definition for that parameter (assigned using the prior distribution), with the remaining six standard deviations chosen to be equal to 0.1, 0.2, 0.5, 2, 5, and 10 times the base standard deviation. The M-H acceptance probability for this update move for θ_l^2 is evaluated. If the move is accepted, the new candidate $\theta_l^{2'}$ is taken to be the updated value for this parameter; otherwise, the original value θ_l^2 becomes the updated value for this parameter (viz., the Markov chain does not move in the θ_l^2-subspace).

The dimension-changing moves that modify the source distribution by ± 1 source atom (and the dimensionality of the hypothesis space by ± 6) are provided by: (1) a creation move $\mathcal{C}$ that results in the addition of a single source atom, so $\Theta' = (N_s + 1, \theta_{N_s+1}) \equiv \mathcal{C}(\Theta) = \mathcal{C}((N_s, \theta_{N_s}))$ where $\Theta' \in \mathbb{R}^{1+6N_s+6}$; and, (2) an annihilation move $\mathcal{C}^\dagger$ that involves the removal of a single existing source atom from the current source distribution, so $\Theta' = (N_s - 1, \theta_{N_s-1}) \equiv \mathcal{C}^\dagger(\Theta) = \mathcal{C}^\dagger((N_s, \theta_{N_s}))$ where $\Theta' \in \mathbb{R}^{1+6N_s-6}$. If a creation move $\mathcal{C}$ is selected, the "coordinates" of the new source atom are obtained by drawing random samples from a proposal density that is chosen to be the prior density for each coordinate. On the other hand, if an annihilation (reverse) move $\mathcal{C}^\dagger$ is selected, a source atom in the current source distribution is randomly picked and removed. The acceptance probability for the creation move $\mathcal{C}$ or the annihilation move $\mathcal{C}^\dagger$ is similar to that for a M-H algorithm involving only dimension-conserving moves, with the exception of a term involving the Jacobian of the transformation $\mathcal{C}$ or $\mathcal{C}^\dagger$. From this viewpoint, the RJMCMC is simply a generalized M-H sampler for the hypothesis space that includes the dimensionality of the space as an additional parameter to be sampled. Finally, the probabilities for creation and annihilation moves are chosen as follows: $p_{\mathcal{C}}^{N_s} = 0$ for $N_s = N_{s,\max}$ and $p_{\mathcal{C}^\dagger}^{N_s} = 0$ for $N_s = N_{s,\min}$; otherwise,

$$\begin{aligned} p_{\mathcal{C}}^{N_s} &= \frac{1}{2} \min \left\{ 1, \frac{p(N_s + 1|I)}{p(N_s|I)} \right\}, \\ p_{\mathcal{C}^\dagger}^{N_s+1} &= \frac{1}{2} \min \left\{ 1, \frac{p(N_s|I)}{p(N_s + 1|I)} \right\}. \end{aligned} \tag{27}$$

To summarize, the Markov chain consists of a sequence of states $\Theta^{(t)}$ $(t = 0, 1, 2, \ldots)$ resulting from individual updates consisting of three basic moves: (1) $\mathcal{M}_{\mathcal{C},\mathcal{C}^\dagger}$ involving the creation of a source atom at a random location, or annihilation of an existing source atom with probabilities $p_{\mathcal{C}}$ and $p_{\mathcal{C}^\dagger}$, respectively; (2) $\mathcal{M}_1$ involving updates of the emission rates of the source atoms using Gibbs sampling; and, (3) $\mathcal{M}_2$ involving updates of the location, source on time and source off time of the source atoms using M-H sampling. The state vector $\Theta^{(t-1)}$ of the Markov chain at iteration $t-1$ is updated to the state vector $\Theta^{(t)}$ at iteration t using the following procedure:

1. Specify values for $(N_{s,\min}, N_{s,\max}, Q_{\max}, T_{\max}, t_0, \gamma, p^*)$ which define $p(\Theta|I)$.

2. Choose an initial state $\Theta^{(0)}$ for the Markov chain by sampling from $p(\Theta|I)$.

3. For $t \in \{1, 2, \ldots, t_{\text{upper}}\}$, conduct the following sequence of moves:

$$\Theta^{(t-1)} \quad \xrightarrow{\mathcal{M}_{c,c^\dagger}} \quad \Theta_\star \quad \xrightarrow{\mathcal{M}_1} \quad \Theta_{\star\star} \quad \xrightarrow{\mathcal{M}_2} \quad \Theta^{(t)}, \tag{28}$$

where $\Theta_\star$ and $\Theta_{\star\star}$ denote some intermediate transition states between iterations $t-1$ and t.

To improve the "speed" with which a Markov chain traverses the hypothesis space (or, to increase the mixing rate of the chain in the hypothesis space), Yee [21] implemented a form of parallel tempering based on a Metropolis-coupled MCMC algorithm described by Geyer [31]. In this approach, r Markov chains are run in parallel, each with a different stationary distribution. These chains are run simultaneously, but occasionally a proposal is made to swap the states of two randomly selected chains. In consequence, the states in the "ladder" of Markov chains can swap positions with a certain acceptance probability as each chain equilibrates. The family of interrelated stationary distributions, corresponding to this "ladder" of r Markov chains, is chosen to have the following form:

$$p_i(\Theta|\mathbf{D}, I) \propto p(\Theta|I)p^{\lambda_i}(\mathbf{D}|\Theta, I), \qquad i = 1, 2, \ldots, r, \tag{29}$$

where $\lambda_i \in [0, 1]$ $(i = 1, 2, \ldots, r)$ is an increasing sequence (viz., $\lambda_i < \lambda_j$ for $i < j$) with $\lambda_1 \equiv 0$ and $\lambda_r \equiv 1$. Note that each distribution in Eq. (29) involves a tempering parameter λ which is used to raise the likelihood to the λ power to give a modified posterior proportional to $p(\Theta|I)p^\lambda(\mathbf{D}|\Theta, I)$. When $\lambda = 0$, the likelihood function is switched off and the modified posterior distribution reduces exactly to the prior distribution. On the other hand, when $\lambda = 1$ the modified posterior distribution is exactly the posterior that we wish to sample from. In between these two extremes, with $\lambda \in (0, 1)$, the effects of the concentration data $\mathbf{D}$ are introduced gradually through the modified (or "softened") likelihood function $p^\lambda(\mathbf{D}|\Theta, I)$.

In this study, rather than use a parallel tempering scheme, we employ a related (and simpler) simulated annealing scheme to facilitate chain mobility in the hypothesis space. In this scheme, we consider an ensemble of N_{mem} (typically between 50 and 200) of source distributions (or, source molecules) that have been randomly drawn from the modified posterior $p_\lambda(\Theta|\mathbf{D}, I) \propto p(\Theta|I)p^\lambda(\mathbf{D}|\Theta, I)$. These samples will be labelled $\Theta_k(\lambda)$, with $\lambda \in [0, 1]$ $(k = 1, 2, \ldots, N_{\text{mem}})$. Note that $p_0(\Theta|\mathbf{D}, I) = p(\Theta|I)$ (prior distribution of Θ) and $p_1(\Theta|\mathbf{D}, I) = p(\Theta|\mathbf{D}, I)$ (posterior distribution of Θ). In this framework, it is useful to interpret the tempering parameter λ as an inverse temperature parameter T (so, $\lambda = 1/T$), with $\lambda \in [0, 1]$ implying $T \in [1, \infty]$. The posterior distribution $p_1(\Theta|\mathbf{D}, I)$ corresponds to the temperature $T = 1$, whereas the modified $p_\lambda(\Theta|\mathbf{D}, I)$ $(\lambda \in [0, 1))$ corresponds to "heating" the posterior distribution to a temperature $T = 1/\lambda > 1$ which results in a flattening of the distribution.

When the stochastic sampling scheme begins and $\lambda = 0$ (infinite temperature), we randomly draw N_{mem} source molecules $\Theta_k(0)$ $(k = 1, 2 \ldots, N_{\text{mem}})$ from $p_0(\Theta|\mathrm{D}, I)$ (prior distribution);

viz., $\Theta_k(0) \sim p(\Theta|I)$.[6] Given an ensemble of $N_{\rm mem}$ source molecules $\Theta_k(\lambda)$ that has achieved equilibrium (at temperature $T = 1/\lambda$) with respect to the modified posterior $p_\lambda(\Theta|\mathbf{D}, I)$,[7] an ensemble of $N_{\rm mem}$ source molecules $\Theta_k(\lambda+\delta\lambda)$ that is consistent with $p_{\lambda+\delta\lambda}(\Theta|\mathbf{D}, I)$ (at the reduced temperature $T = 1/(\lambda + \delta\lambda)$, $\delta\lambda > 0$) can be obtained by using the weighted resampling method (see Gamerman and Lopes [32]) applied to $\Theta_k(\lambda)$ ($k = 1, 2, \ldots, N_{\rm mem}$). To this purpose, each source molecule $\Theta_k(\lambda)$ is assigned an importance weight w_k as follows:

$$\begin{aligned} w_k &= \frac{p_{\lambda+\delta\lambda}(\Theta_k(\lambda)|\mathbf{D}, I)}{p_\lambda(\Theta_k(\lambda)|\mathbf{D}, I)} \times \left(\sum_{j=1}^{N_{\rm mem}} w_j \right)^{-1} \\ &= p^{\delta\lambda}(\mathbf{D}|\Theta_k(\lambda), I) \times \left(\sum_{j=1}^{N_{\rm mem}} w_j \right)^{-1}, \end{aligned} \tag{30}$$

for $k = 1, 2, \ldots, N_{\rm mem}$. Next, a resample is drawn from the discrete distribution concentrated at the sample of source molecules $\{\Theta_k(\lambda)\}_{k=1}^{N_{\rm mem}}$ from $p_\lambda(\Theta|\mathbf{D}, I)$ with respective weights $\{w_k\}_{k=1}^{N_{\rm mem}}$. In other words, the resampling step here involves generating a new set $\{\Theta_{k^*}\}_{k^*=1}^{N_{\rm mem}}$ by resampling (with replacement) $N_{\rm mem}$ times from the set $\{\Theta_k(\lambda)\}_{k=1}^{N_{\rm mem}}$ so that $\Pr\{\Theta_{k^*} = \Theta_k(\lambda)\} = w_k$. This resample can be interpreted as an ensemble of source molecules Θ_{k^*} ($k^* = 1, 2, \ldots, N_{\rm mem}$) that is drawn from the modified posterior $p_{\lambda+\delta\lambda}(\Theta|\mathbf{D}, I)$.

There are many ways to implement this resampling (with replacement) from $\{\Theta_k(\lambda)\}_{k=1}^{N_{\rm mem}}$, but in this report we use the systematic resampling scheme described by Kitagawa [33]. This scheme requires $O(N_{\rm mem})$ time to execute and minimizes the Monte Carlo variation in the resample. The basic operations for this scheme can be described briefly as follows. A vector $(n_1, n_2, \ldots, n_{N_{\rm mem}})$ of copies of the source molecules $\Theta_k(\lambda)$ ($k = 1, 2 \ldots, N_{\rm mem}$) is obtained by computing a vector $(\rho_1, \rho_2, \ldots, \rho_{N_{\rm mem}})$ of the cumulative sums of $N_{\rm mem}(w_1, w_2, \ldots, w_{N_{\rm mem}})$, generating a random draw $u \sim \mathcal{U}([0, 1])$, and determining n_k ($k = 1, 2, \ldots, N_{\rm mem}$) from

$$\begin{aligned} n_k &= \lfloor \rho_k + u \rfloor - \lfloor \rho_{k-1} + u \rfloor, \quad k = 2, 3, \ldots, N_{\rm mem} - 1, \\ n_1 &= \lfloor \rho_1 + u \rfloor, \quad n_{N_{\rm mem}} = \lfloor \rho_{N_{\rm mem}-1} + u \rfloor, \end{aligned} \tag{31}$$

where $\lfloor \cdot \rfloor$ denotes the "integer part of". After resampling, the weights for each member of the resample are reset to $w_k = 1/N_{\rm mem}$ (viz., an equal weight is assigned to each member of the resample). After each resampling step to give an ensemble of $N_{\rm mem}$ source molecules $\Theta_k(\lambda + \delta\lambda)$ that is in equilibrium at temperature $T = 1/(\lambda + \delta\lambda)$ (approximately or better) with respect to $p_{\lambda+\delta\lambda}(\Theta|\mathbf{D}, I)$, we apply N_r (typically 25) Markov chain transitions $T_{\lambda+\delta\lambda}$ to each member of the ensemble obtained from the resampling operation. These Markov chain transitions utilize the update scheme given by Eq. (28) and leave $p_{\lambda+\delta\lambda}(\Theta|\mathbf{D}, I)$ invariant.

[6] The prior distribution $p(\Theta|I)$ is composed of standard probability distributions for which independent sampling is easy.

[7] This simply implies that $\{\Theta_k(\lambda)\}_{k=1}^{N_{\rm mem}}$ can be interpreted as an ensemble of source molecules drawn from $p_\lambda(\Theta|\mathbf{D}, I)$.

An annealing schedule for $\lambda \in [0,1]$ is required for the simulated annealing. In this report, we applied simulated annealing with 200 values of λ uniformly spaced in the interval $[0, 0.05]$ and 400 values of λ geometrically spaced in the interval $(0.05, 1]$. This gentle annealing schedule allows the ensemble of N_{mem} source molecules to transition slowly through a series of quasi-equilibrium states from the prior distribution ($\lambda = 0$, or infinite temperature) at one end of the annealing schedule to the posterior distribution ($\lambda = 1$, or unit temperature) at the other end of the schedule. When $\lambda = 1$, the annealing phase is complete and probabilistic exploration of the hypothesis space proceeds (for each of the N_{mem} source molecules in the ensemble) in accordance to the scheme summarized in Eq. (28). The annealing phase of the scheme, corresponding to values of $\lambda \in [0, 1)$, is associated with the burn-in phase of the algorithm. When $\lambda = 1$, the MCMC algorithm has reached an equilibrium, at which point the probabilistic exploration corresponding to the sampling from the posterior distribution $p(\Theta|\mathbf{D}, I)$ begins. These samples drawn from the posterior distribution can be used to make inferences about all characteristics of the source parameters (e.g., posterior means, variances, HPD intervals).

Interestingly, the simulated annealing phase of the proposed scheme can be used to estimate the normalization constant (or, evidence) $Z \equiv p(\mathbf{D}|I)$ for the posterior distribution $p(\Theta|\mathbf{D}, I)$ [see Eq. (12)]. To accomplish this objective, we use an approach referred to as thermodynamic integration (see, von der Linden [34]), which has its origins in the problem of the evaluation of partition functions in statistical mechanics familiar to physicists. Briefly, thermodynamic integration focusses on the modified evidence $Z(\lambda)$ defined as follows:

$$Z(\lambda) \equiv \int p(\Theta|I) p^{\lambda}(\mathbf{D}|\Theta, I)\, d\Theta, \tag{32}$$

which can be seen to be simply the normalization constant (evidence) corresponding to the modified posterior $p_{\lambda}(\Theta|\mathbf{D}, I)$. Obviously, $Z(0) = 1$ because the prior $p(\Theta|I)$ is normalized and $Z(1) = Z$ is the desired normalization constant (evidence) for the posterior $p(\Theta|\mathbf{D}, I)$. Taking the logarithmic derivative of Eq. (32) with respect to λ, and re-arranging yields

$$\ln[Z(1)] - \ln[Z(0)] = \ln[Z(1)] = \int_0^1 \big\langle \ln[p(\mathbf{D}|\Theta, I)] \big\rangle_{\lambda}\, d\lambda, \tag{33}$$

where $\langle\, \cdot\, \rangle_{\lambda}$ denotes the mathematical expectation operation with respect to the modified posterior $p_{\lambda}(\Theta|\mathbf{D}, I)$.

Additionally, the computation of $Z(\lambda)$ allows the determination of the information (gain) corresponding to the receipt of the concentration data $\mathbf{D}$ and the updating of our state of knowledge concerning the (unknown) source S (encoded as Θ) from the prior distribution $p(\Theta|I)$ to the posterior distribution $p(\Theta|\mathbf{D}, I)$. This information gain (amount of useful information about Θ embodied in $\mathbf{D}$) is given by the Kullback-Leibler divergence $D_{\text{KL}}(\lambda)$ at $\lambda = 1$ [viz., by $D_{\text{KL}}(1)$] where $D_{\text{KL}}(\lambda)$ is defined as follows (Cover and Thomas [35]):

$$\begin{aligned} D_{\text{KL}}(\lambda) &\equiv \int \ln\left(\frac{p_{\lambda}(\Theta|\mathbf{D}, I)}{p(\Theta|I)}\right) p_{\lambda}(\Theta|\mathbf{D}, I)\, d\Theta \\ &= \lambda \int \ln\big[p(\mathbf{D}|\Theta, I)\big]\, p_{\lambda}(\Theta|\mathbf{D}, I)\, d\Theta - \ln[Z(\lambda)] \int p_{\lambda}(\Theta|\mathbf{D}, I)\, d\Theta \\ &= \lambda \big\langle \ln[p(\mathbf{D}|\Theta, I)] \big\rangle_{\lambda} - \ln[Z(\lambda)], \end{aligned} \tag{34}$$

on using the fact that $p_\lambda(\Theta|\mathbf{D},I) = p(\Theta|I)p^\lambda(\mathbf{D}|\Theta,I)/Z(\lambda)$. The Kullback-Leibler divergence defined in Eq. (34) for $\lambda = 1$ is simply the negative of the entropy (negentropy) of the posterior relative to the prior and, as such, is the information gain provided by the receipt of the concentration data $\mathbf{D}$. More specifically, the information gain "compresses" the posterior relative to the prior so that $D_{\mathrm{KL}}(1)$ can simply be interpreted as the logarithm of the volumetric factor by which the prior has been compressed to become the posterior (the greater this compression, the greater is the information gain provided by the concentration data).

4 FUSION Field Trial 2007 (FFT-07)

The **FU**sing **S**ensor **I**nformation from **O**bserving **N**etworks (FUSION) Field Trial 2007 (FFT-07) was conducted to obtain research-grade concentration data from a network (or array) of fast-response detectors resulting from releases of a passive tracer from single and multiple sources. The scientific objective of this field campaign was to acquire a comprehensive meteorological and dispersion dataset that can be used to validate methodologies developed for source reconstruction. Details of the instrumentation deployed and the experiments conducted in FFT-07 are given in Storwold [36], so only a brief summary of FFT-07 will be presented here. In particular, only the relevant details of the experiments that are required for the interpretation of the results in this report are emphasized.

The experiments in FFT-07 were carried out in September 2007 at Tower Grid on US Army Dugway Proving Ground, Utah about 2 km west of Camel Back Ridge on the Great Salt Lake Desert. The terrain was flat, uniform, and homogeneous with short grass interspersed with low shrubs that are between 0.25 to 0.75 m in height, providing an upwind fetch that is uniform and unobstructed for 5 km or more in a wide sector. The test elevation site was 1330 m above mean sea level, and the terrain gradually rises towards the southeast by about 1 m over horizontal distances of about 2000 m. The easterly through southerly drainage flows that predominate during the early morning hours at the site originate on the higher terrain to the southeast and are channeled by Camel Back Ridge.

In all the experiments, the tracer gas used was propylene (C_3H_6) which was chosen because of its low toxicity, relatively high vapor pressure (938 kPa at 21°C), and low ionization potential (9.73 eV), hence giving the concentration detectors good sensitivity. The concentration detectors used were fast-response digital photo-ionization (dPID) detectors manufactured commercially by Aurora Scientific Inc. (Aurora, Ontario, Canada). These detectors give a frequency response of 50 Hz with a sensitivity of about 0.025 parts per million (ppm) by volume of propylene. The detectors were calibrated regularly over their entire operating range using a specially designed calibration unit mounted in a small rack. This unit allowed the user to mix gas from propylene calibration gas cylinders with air to generate a user-selectable gas concentration. A least-squares regression fit of the calibration data to a second-order polynomial, $c = a_0 + a_1 V + a_2 V^2$, provided an overall calibration for each detector, where V is the analog-to-digital (A/D) output of the detector, c is the absolute propylene concentration (ppm), and a_i $(i = 0, 1, 2)$ are the fitted calibration constants.

In the experiments, a plume was formed in the atmospheric surface layer by releasing propylene from one or more (up to a maximum of four) purpose-designed gas dissemination systems (see Chandler [37]). The system consisted of up to four propylene cylinders connected in parallel and immersed in a warm-water bath. A regulator was connected to the outlet of the cylinders to ensure a constant downstream pressure, and a flexible hose was used to connect the regulator to the inlet fitting of the mass flow controller. This controller was used to set, control, and measure the flow through the dissemination system. The controller allowed a constant gas flow rate to be maintained at a user-selected reference level between 10 and 250 l min^{-1}. A quick-release connector mated the outlet of the mass flow controller to the dissemination hose, which was connected to the base of the disseminator, a section of 40 PVC (polyvinyl chloride) pipe 1 m in length and 0.05 m in diameter.

The network (or array) of concentration detectors used in FFT-07 is shown in Figure 3. A total of 100 dPIDs (indicated by the filled squares in Figure 3) was arranged in a staggered configuration consisting of 10 rows of 10 detectors. The rows of detectors were spaced 50 m apart. The spacing between detectors along each row was 50 m. The (x, y) local coordinate system that will be used here is shown in Figure 3. The direction corresponding to the negative x-axis (referred to as grid north) of the array is oriented 25° west of north to take advantage of the prevailing wind direction at the test site in the sector from south-south-west to south-east. Winds incident on the array of detectors from this sector resulted in a flat and homogeneous upwind fetch of more than 10 km. Note that x is the lengthwise distance from the windward edge (grid south edge) of the array and y is the spanwise distance from the grid west edge of the array. The overall length (along the x-direction) and width (along the y-direction) of the detector array were 450 m and 475 m, respectively. Finally, the z (vertical) coordinate is defined so $z = 0$ is the ground surface.

For ease of reference, the rows of the detector array will be numbered from 1 to 10, with row 1 forming the grid north edge of the array (at $x = -450$ m) and row 10 forming the grid south edge of the array (at $x = 0$ m). Along row i of the array ($i = 1, 2, \ldots, 10$), the dPID detectors are numbered as $10(i - 1) + j$ ($j = 1, 2, \ldots, 10$), where the j index of the detector increases from grid west to grid east. Consequently, in this numbering scheme for detectors, detector number 91 [first detector ($j = 1$) in row 10 ($i = 10$)] is located at the origin of the local (x, y) coordinate system. For geo-referencing purposes, it is noted that this location corresponds to the following geodetic coordinates: 40.0923° N latitude and 112.9757° W longitude. The concentration detectors along the ten sampling lines in the array were placed at a height, z_d, of 2.0 m.

In the experiments used for the current analysis, propylene gas was released continuously over a period of 10 min from one or more source locations (up to four) at a height, z_s, of 2.0 m. The four source locations (labelled 1 to 4 in Figure 3) are as follows: (1) source 1 is at $(x_s, y_s) = (33.0, 171.0)$ m; source 2 is at $(x_s, y_s) = (33.8, 240.7)$ m; source 3 is at $(x_s, y_s) = (30.0, 312.9)$ m; and, source 4 is at $(x_s, y_s) = (26.0, 384.4)$ m. Unfortunately, for these experiments, only the mass flow controller for source 3 functioned properly. The mass flow controllers for sources 1, 2 and 4 failed to properly regulate the flow owing to the fact that the impurities in the low-grade of propylene used for these experiments contaminated the sensor and electronic circuitry in these controllers. Consequently, the control signals

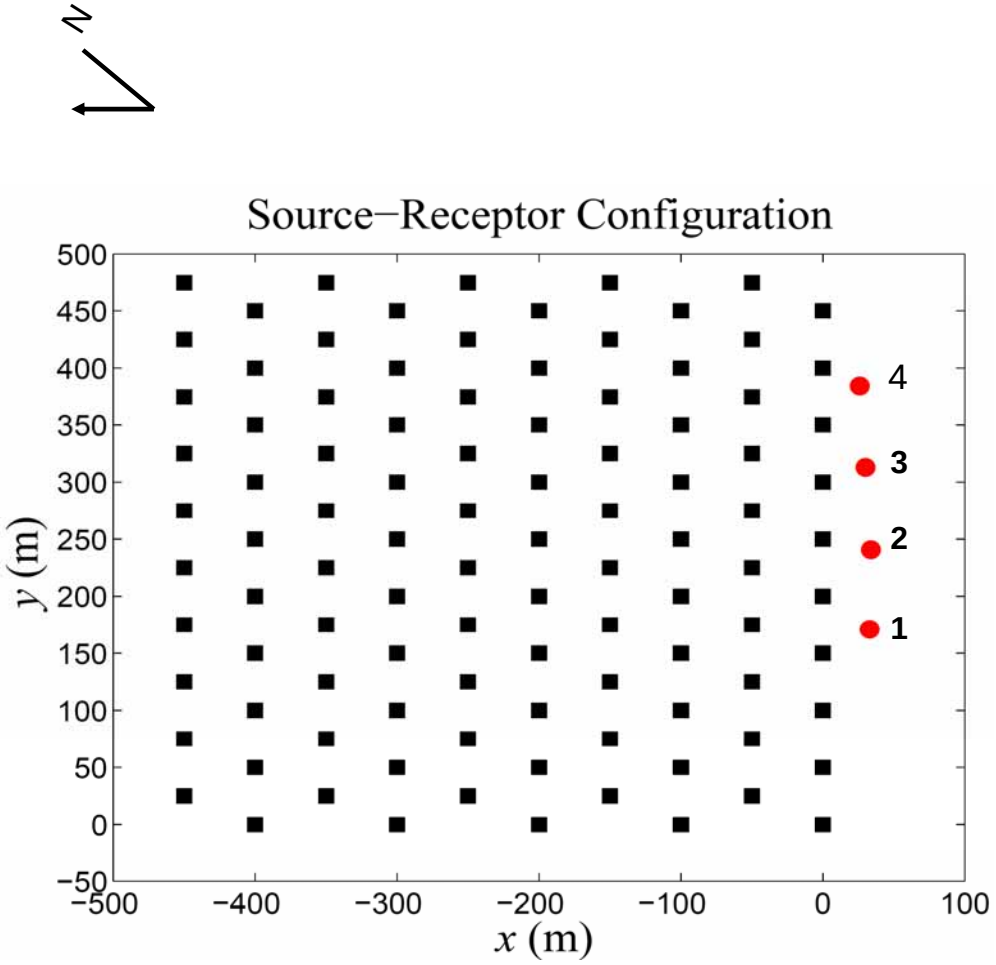


Figure 3: *A schematic diagram of the geometry of the network (or, array) of concentration detectors (dPIDs) used in FFT-07. The locations of the 100 detectors (filled squares) and four sources (filled circles) are shown.*

that were used to set the valve to regulate the flow rate in these controllers failed to function correctly.

Three-dimensional (3-D) sonic anemometers (R. M. Young, Model 81000) were arranged on three 32-m lattice towers along a transect parallel to the x-axis and midline of the concentration detector array. These three towers were located at the center of the detector array at $(x, y) = (-225.0, 237.5)$ m (or, grid center) and at positions 750.0 m upwind and downwind of grid center at $(x, y) = (525.0, 237.5)$ m and $(x, y) = (-975.0, 237.5)$ m, respectively. Each of these towers was instrumented with five 3-D sonic anemometers at the 2-, 4-, 8-, 16-, and 32-m levels. The sonic anemometer data were recorded at 10 Hz.

An estimate of the momentum roughness length, z_0, was determined from mean wind profiles measured under near-neutral stratification (for which the mean wind speed variation

with height can be represented by a semi-logarithmic relation). An estimate of $z_0 = 1.3 \pm 0.2$ cm was obtained from the 3-D sonic anemometry data on the upwind tower at $(x, y) = (525.0, 237.5)$ m, as an average over 35 periods (each with a 10-minute sampling time) with the magnitude of the Obukhov length L from the 3-D sonic anemometer at 2-m height exceeding 500 m (viz., $|L| > 500$ m at $z = 2$ m).

Table 1: *Summary of turbulence statistics measured with a 3-D sonic anemometer at the 2-m level on the upwind tower. Here, S_2 is the horizontal mean wind speed at the 2-m level, α is the mean wind direction, u_* is the friction velocity, L is the Obukhov length, and σ_i ($i \equiv u, v, w$) are the standard deviations of the fluctuating wind velocity in the three coordinate directions (alongwind, crosswind, and vertical, respectively).*

Trial	S_2 (m s^{-1})	α (deg)	u_* (m s^{-1})	L (m)	σ_u/u_* (-)	σ_v/u_* (-)	σ_w/u_* (-)
I	2.60	142.3	0.418	10.7	2.18	1.62	1.17
II	3.00	138.9	0.183	15.4	2.36	1.72	1.23
III	3.61	155.1	0.282	−27.3	2.33	1.86	1.10

The mean wind and turbulence statistics data for three experiments used to test the source reconstruction methodology are given in Table 1. Here, the wind components are calculated in a rotated coordinate system with the alongwind (u), crosswind (v), and vertical (w) components, aligned along the x'-, y'-, and z-directions, respectively. The standard deviations of the fluctuating wind in these three coordinate directions are denoted σ_u, σ_v, and σ_w, respectively. Measurements of the mean horizontal wind speed $S_2 \equiv \left(\bar{u}^2 + \bar{v}^2\right)^{1/2}$ and mean wind direction α, obtained from the 3-D sonic anemometer at the 2-m level on the upwind tower, are summarized in Table 1 for these three experiments. The mean wind direction corresponding to a wind from the $+x$-axis direction (see Figure 3) will be denoted in the usual compass convention as $\alpha = 155^\circ$ (and is associated with a wind direction from grid south). This mean wind direction corresponds to normal incidence on the detector array (viz., the direction is perpendicular to the sampling lines of detectors along the y-axis). For simplicity, α' will denote the deviation of the mean wind direction from $\alpha = 155^\circ$ (or, normal incidence). The Obukhov length L was estimated as

$$L = -\frac{u_*^3 \overline{T}}{k_v g \overline{w'T'}}, \tag{35}$$

where $T = \overline{T} + T'$ is the sonic temperature (which can be decomposed into a mean value $\overline{T}$ and fluctuation therefrom), u_* is the friction velocity at the surface, $k_v \approx 0.4$ is the von Karman constant, and g is the acceleration due to gravity. The friction velocity was estimated as follows:

$$u_* = \left[(\overline{u'w'})^2 + (\overline{v'w'})^2\right]^{1/4}. \tag{36}$$

5 Application to FFT-07 data

In this section, we apply the source reconstruction algorithm to the three dispersion data sets identified in Table 1. Trials I, II and III correspond to one- three- and four-source examples.

In all three examples, the backward-time LS model given by Eqs (24) and (25) was used to determine $\overline{C}_J(\Theta)$ from Eq. (26). This model was applied to short-range dispersion in the atmospheric surface layer over a level and unobstructed terrain. The mean wind and turbulence statistics for the three examples were assumed to be horizontally homogeneous and stationary, with the relevant flow statistics for these examples summarized in Table 1. The backward-time LS model was applied with the Kolmogorov constant $C_0 = 4.8$, a value which was recommended by Wilson et al [38] from a calibration of the model against concentration data obtained from Project Prairie Grass.

The wind flow and turbulence statistics parameterization for the LS model are prescribed based on standard surface-layer relationships from Monin-Obukhov theory and summarized in the Appendix of Yee [21]. These relationships have been modified slightly so as to be consistent with the measured values of σ_u/u_*, σ_v/u_* and σ_w/u_* (at the surface) compiled in Table 1 for the three trials. Finally, the three surface-layer parameters z_0, u_* and L required for the wind statistics parameterization have been either reported in main text of Section 4 or summarized in Table 1.

All the examples in this report involve continuously emitting sources and, as a consequence, $T_b^k \to -\infty$ and $T_e^k \to \infty$ in Eqs (6) and (26) $[k = 1, 2, \ldots, N_s]$. As a consequence, the relevant source parameters are as follows: $\Theta = (N_s, \mathbf{x}_{s1}, Q_1, \ldots, \mathbf{x}_{sN_s}, Q_{N_s})$. Furthermore, it is assumed that the height of the sources above ground level ($z_s = 2.0$ m) is known *a priori*, so the only unknown location parameters are (x_s, y_s) of the sources in the horizontal plane. In view of these assumptions, the adjunct concentration $C^*(\mathbf{x}_s|\mathbf{x}_d)$ in Eq. (26) can be pre-calculated for one detector position $\mathbf{x}_d$ (for the known source height z_s) in any given trial, with the adjunct concentration C^* (considered as a function of $\mathbf{x}_s$) at all other detector positions obtained by a simple translation of $C^*(\mathbf{x}_s|\mathbf{x}_d)$ in the horizontal (x, y) plane.

5.1 Trial I: one-source example

For this example, the source corresponding to source 3 (see Figure 3) was turned on. The (constant) emission rate from this source was $Q = 2.8$ g s^{-1}. The source-detector configuration for this example is depicted in Figure 4. Only six detectors (shown by the filled blue squares in Figure 4) in the array (viz., detector numbers 25, 36, 46, 56, 66, and 77) were used for the source reconstruction. For this example, the mean wind direction was $\alpha' = 155^\circ - \alpha = 12.7^\circ$ away from normal incidence to the detector array.

It is assumed that the number of sources is known *a priori* (viz., this knowledge formed part of the background information I) for this example. To this end, $N_{\mathrm{s,min}} = N_{\mathrm{s,max}} = 1$.[8]

[8] Note that the value of the hyperparameter p^* that characterizes the binomial prior for N_s in this example

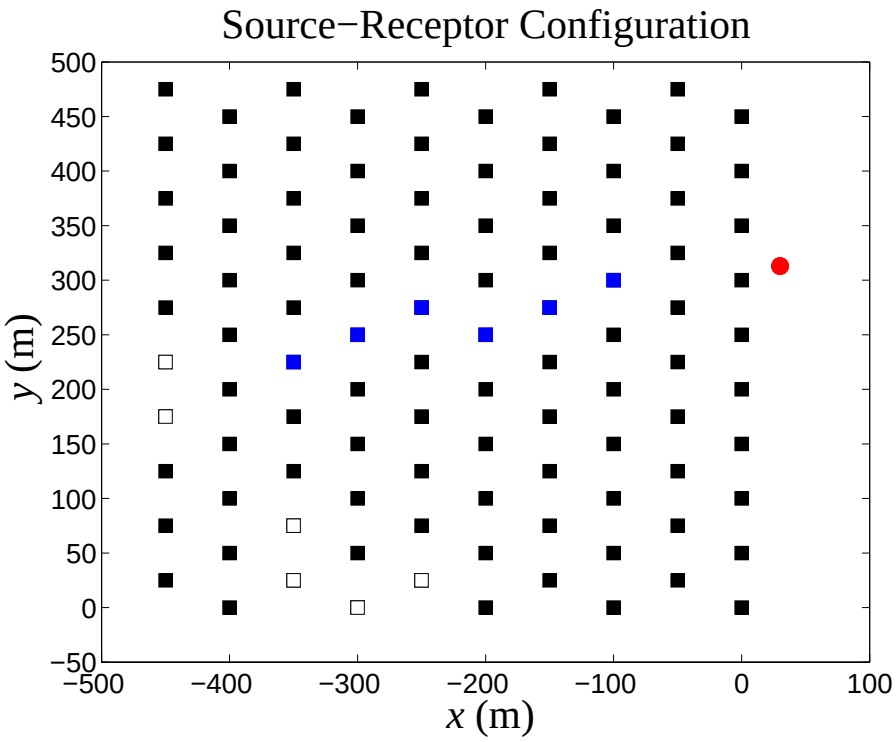


Figure 4: *In Trial I, one source located upwind of the array of detectors was turned on. The solid dot shows the location of this source. Squares show the location of the detectors in the array: open and filled squares indicate that the detector at the given location is missing and present, respectively, in the array. A filled blue square marks the detectors that were used for the source reconstruction.*

The other parameters needed to define the prior distribution $p(\Theta|I)$ for this example are chosen as follows: $Q_{\max} = 100$ g s^{-1}; $\gamma = 1.0$; and $\mathcal{D} = [0, 100] \times [0, 500]$ providing the prior bounds on the source location in the (x, y)-plane. Recall that the (x, y) coordinate system used here is chosen as shown in Figure 4.

After the initial simulated annealing phase was completed using $N_{\text{mem}} = 50$ members of an ensemble of source distributions (originally, randomly drawn from the prior distribution $p(\Theta|I)$ for $\lambda = 0$), the MCMC algorithm was run for an additional 1,000 iterations for each member of the ensemble, giving a total of 50,000 samples of source distributions (encoded as Θ) obtained from the posterior distribution $p(\Theta|\mathbf{D}, I)$ during the probabilistic exploration phase ($\lambda = 1$) of the algorithm. The marginal distributions for the source location (x_s, y_s) and the emission rate $Q \equiv q_s$ obtained from these 50,000 samples are shown in Figure 5. The actual source location was $(x_s, y_s) = (30.0, 312.9)$ m and the emission rate was $Q = 2.8$ g s^{-1}. The analysis of the samples gave the following estimates for the source parameters expressed as a posterior mean value and standard deviation (s.d.) and the boundary (lower

is irrelevant, owing to the fact that $p(N_s|I) \equiv 1$ ($N_s \in [N_{\text{s,min}}, N_{\text{s,max}}]$) for $N_{\text{s,min}} = N_{\text{s,max}} = 1$.

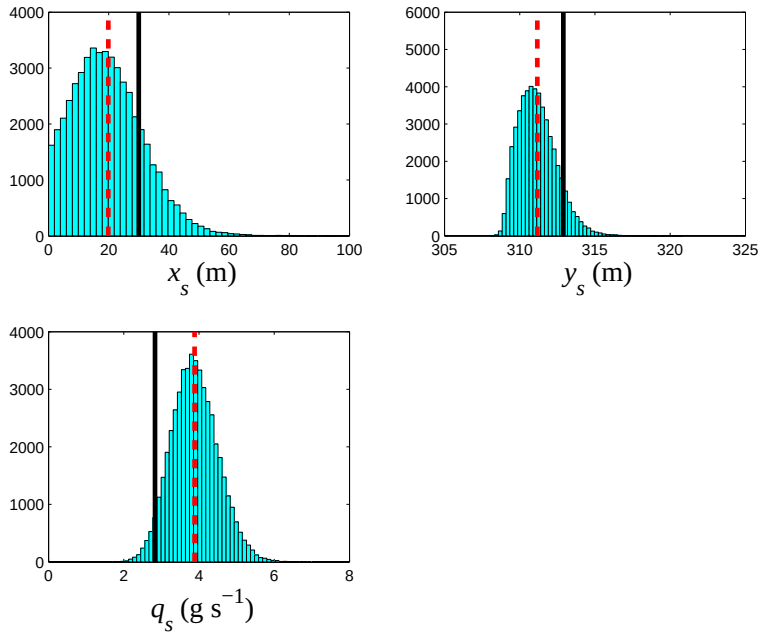


Figure 5: *The marginal posterior histograms for the source location* (x_s, y_s) *and for the emission rate* q_s *obtained for Trial I. The vertical solid lines mark the true values of the source parameters and the vertical dashed lines mark the best estimates of the source parameters obtained from the posterior mean.*

and upper) of the 95% HPD interval: $\widehat{x}_s = 19.9 \pm 11.6$ m $(0.0, 40.5)$ m; $\widehat{y}_s = 311.2 \pm 1.3$ m $(309.0, 313.6)$ m; and, $\widehat{q}_s = 3.9 \pm 0.6$ g s^{-1} $(2.7, 5.1)$ g s^{-1}. Although the information embodied in the six concentration detectors allowed the crosswind position y_s of the source to be well determined, this information only weakly constrained the alongwind position x_s of the source which is estimated with a large uncertainty. As a consequence, there is also a large uncertainty in the determination of the source emission rate. Indeed, the information gain (provided by the concentration data) for this example was determined to be only $D_{\mathrm{KL}}(1) = 10.2$ natural units (nits) [see Eq. (34)].

5.2 Trial II: three-source example

For Trial II, three sources were turned on upwind of the detector array as shown in Figure 6. These sources correspond to sources 2, 3 and 4 shown in Figure 3. The emission rate Q from source 3 was known to be 3.8 g s^{-1}, whereas the emission rates for sources 2 and 4 were not known in this example owing to the fact that the mass flow controllers for these two sources malfunctioned because of their contamination by the impurities in the propylene cylinders connected to them. The source reconstruction for this example will be attempted using 37 detectors in the array, which have been marked using a filled blue square in Figure 6. In this trial, the mean wind direction had an obliquity angle with respect to the normal to the detector array of $\alpha' = 155^\circ - \alpha = 16.1^\circ$.

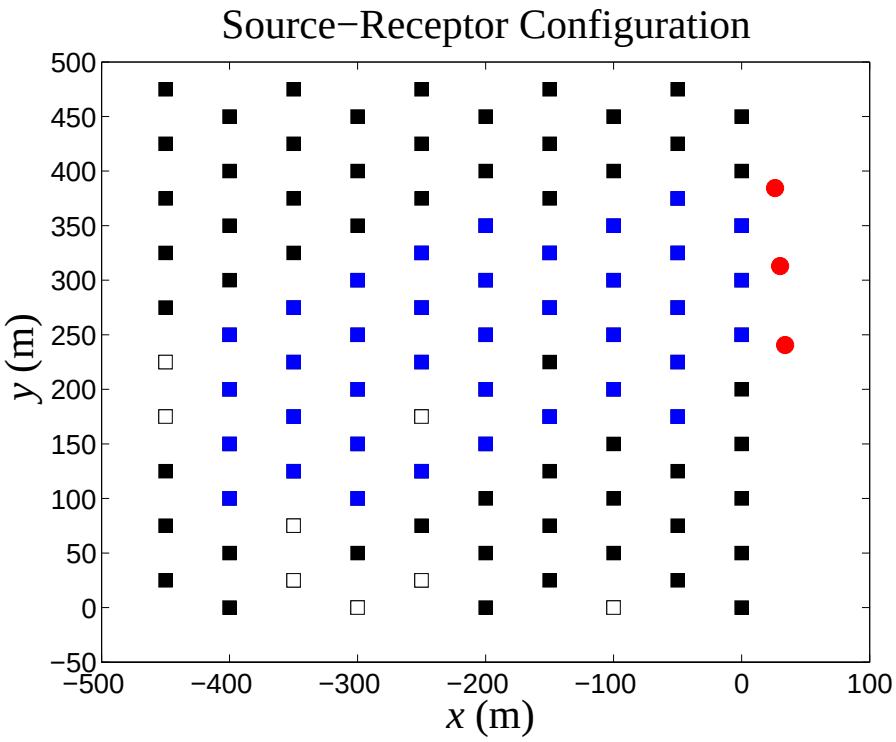


Figure 6: *In Trial II, three sources located upwind of the array of detectors were turned on. The solid dots show the locations of the three sources. Squares show the location of the detectors in the array: open and filled squares indicate that the detector at the given location is missing and present, respectively, in the array. A filled blue square marks the detectors that were used for the source reconstruction.*

For this example, we assume that the number of sources N_s is unknown *a priori*. Furthermore, it is assumed that an exact upper bound on the maximum number of possible sources is available. In consequence, for our specification of the binomial prior for N_s, we choose $N_{\mathrm{s,min}} = 1$ and $N_{\mathrm{s,max}} = 4$, with $p^* = 1/3$.[9] As in the first example, the hyperparameters Q_{max} and γ that define the Bernoulli-uniform prior for the emission rate Q, are initialized to 100.0 g s^{-1} and 0.25, respectively. The prior bounds on the (x_s, y_s) location of each of the unknown sources are given by $\mathcal{D} = [0, 100] \times [0, 500]$.

We applied our source reconstruction algorithm to this example. The simulated annealing phase of the algorithm was initiated (at $\lambda = 0$) with $N_{\mathrm{mem}} = 50$ different source distributions S (encoded as Θ) drawn randomly from the prior distribution $p(\Theta|I)$. A sequence of modified posterior distributions with $\lambda \in (0, 1]$ and the associated annealing schedule as described in Section 3.2 was used in the simulated annealing phase. After the termination of this phase with $\lambda = 1$, a further 1000 iterations of the RJMCMC procedure were applied

[9] This choice for $p(N_s|I)$ implies that the expected number of sources in the domain is $\langle N_s \rangle = N_{\mathrm{s,min}} + (N_{\mathrm{s,max}} - N_{\mathrm{s,min}})p^* = 2$. In consequence, the prior distribution for N_s favors the wrong choice for the number of sources in this example.

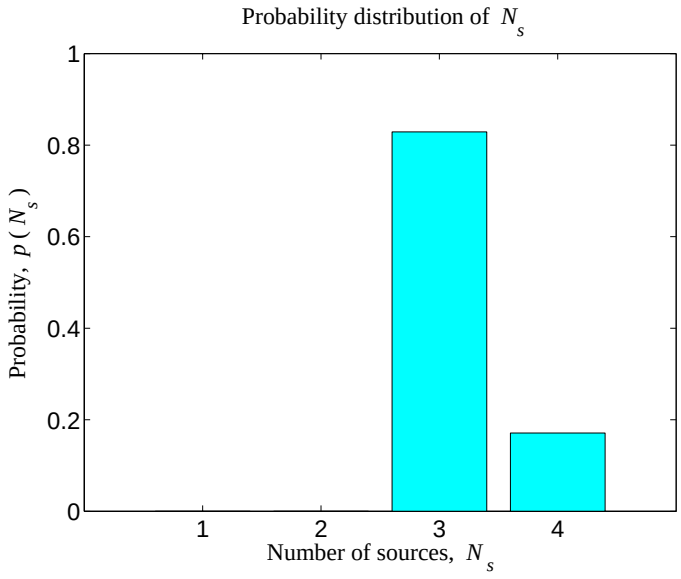


Figure 7: *The posterior distribution for the number of sources, $p(N_s) \equiv p(N_s|\mathbf{D}, I)$, for Trial II estimated using 50,000 samples obtained from the probabilistic exploration phase of the stochastic sampling algorithm.*

to each of the $N_{\rm mem} = 50$ members of the ensemble of source distributions to give 50,000 samples of source distributions drawn from the posterior distribution $p(\Theta|\mathbf{D}, I)$ during the probabilistic exploration phase of the algorithm.

Figure 7 displays the posterior probability distribution for the number of sources $p(N_s) \equiv p(N_s|\mathbf{D}, I)$ for this example. The most probable number of sources is $N_s = 3$, and this coincides with the correct number of sources. In consequence, our *maximum a posteriori* estimate for the number of discrete sources present in Trial II is $\widehat{N}_s = 3$ [cf. Eq. (18)], and this choice is favoured with a probability of about 0.82. It is important to emphasize that samples with $N_s = 4$ discrete sources were also drawn from the posterior distribution during the probabilistic exploration phase of the algorithm. However, it turns out that with the prior specification $\gamma \equiv \Pr\{Q_k > 0\} = 0.25$ ($k = N_{\rm s,min}, N_{\rm s,min} + 1, \ldots, N_{\rm s,max}$), many of the samples obtained with $N_s = 4$ had one of the discrete sources turned off (so, the emission rate $Q = 0$ for this source). For the case of a four-source sample, but with one of the sources turned off (with the result that this source does not contribute to the model concentration "seen" by the detectors in the array), we classify the source distribution sample here as having three discrete sources rather than four. This convention for the determination of the number of discrete sources, associated with a sample of a source distribution model, was used in the construction of Figure 7.

Given the fact that our best estimate for the number of sources is $\widehat{N}_s = 3$, we can now estimate the source parameters (x_s, y_s, q_s) corresponding to each of these three discrete sources. Towards this objective, we first extract all samples of source distributions drawn from the posterior distribution $p(\Theta|\mathbf{D}, I)$ having exactly three sources (viz., we choose

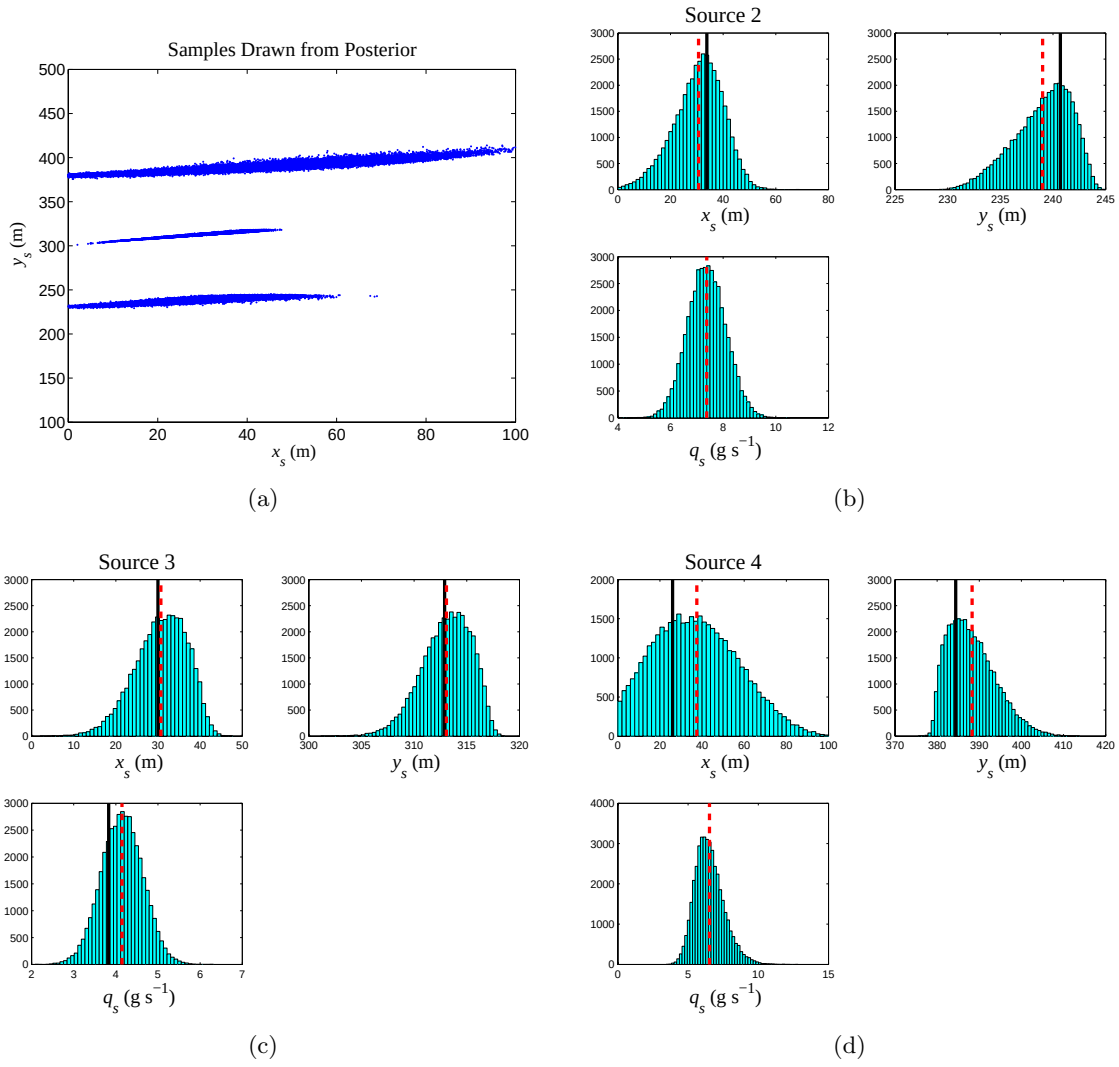


Figure 8: *Inference of the discrete source parameters obtained from samples drawn from the posterior distribution $p(\Theta|\mathbf{D}, I)$ having exactly three discrete sources. (a) Samples from the posterior distribution having three discrete sources projected onto the (x_s, y_s) subspace. (b,c,d) Histograms for the three parameters, namely alongwind location x_s, crosswind location y_s, and emission rate q_s that characterize sources 2, 3, and 4 (cf. Figure 3). In each frame, the solid vertical line indicates the true value of the parameter (if known) and the dashed vertical line corresponds to the best estimate of the parameter obtained as the posterior mean of the marginal posterior distribution for the parameter.*

all sample indices t such that $\Theta^{(t)}$ has a first component that satisfies $N_s = 3$ for $t = 1, 2, \ldots, 50000$). Figure 8(a) displays samples corresponding to $N_s = 3$, projected onto the (x_s, y_s) subspace. From this figure, we can identify three clusters of points that determine the locations of the three discrete sources. Observe that the concentration data $\mathbf{D}$ constrain the crosswind positions y_s of the three sources, but there is considerable uncertainty in the

determination of the alongwind positions x_s of the sources (which is especially acute for the source located at the largest value of y_s).

Table 2: *The posterior mean, posterior standard deviation, and lower and upper bounds of the 95% HPD interval of the parameters $x_{s,k}$ (m), $y_{s,k}$ (m), and $q_{s,k}$ (g s^{-1}) for k = 1, 2, and 3 calculated from samples of source distribution models with $N_s = 3$ (the latter corresponding to the most probable number of sources in the domain as inferred from Figure 7).*

	Parameter	Mean	Standard Deviation	95% HPD	Actual
$k = 1$					
	x_s (m)	30.7	9.3	$(11.0, 47.5)$	33.8
	y_s (m)	239.0	2.7	$(233.5, 243.5)$	240.7
	q_s (g s^{-1})	7.4	0.7	$(5.96, 8.81)$	–
$k = 2$					
	x_s (m)	30.7	6.2	$(18.6, 41.6)$	30.0
	y_s (m)	313.1	2.3	$(308.5, 317.2)$	312.9
	q_s (g s^{-1})	4.1	0.5	$(3.1, 5.1)$	3.8
$k = 3$					
	x_s (m)	37.5	19.8	$(0.8, 73.0)$	26.0
	y_s (m)	388.3	5.7	$(379.0, 399.3)$	384.4
	q_s (g s^{-1})	6.5	1.0	$(4.8, 8.9)$	–

An important issue in the interpretation of the results is the identifiability problem that arises owing to the fact that the posterior distribution of Θ is invariant under a reordering (or, relabelling) of the identifiers used for each discrete source in the source distribution model [cf. Eq. (17)]. More specifically, it can be seen that $p(\Theta|\mathbf{D}, I)$ is invariant under a permutation of the discrete source identifier k. However, from Figure 8(a), it is evident that we can uniquely identify the discrete sources in a source distribution model if we impose an ordering constraint on the $y_{s,k}$-locations of the three sources so that $y_{s,1} < y_{s,2} < y_{s,3}$. Furthermore, comparing Figures 3 and 8(a), it can be seen that after relabelling the discrete sources of each source distribution model in accordance to this ordering constraint, the sources at $y_{s,1}$, $y_{s,2}$ and $y_{s,3}$ are associated with sources 2, 3 and 4, respectively. Figures 8(b), (c) and (d) display the histograms of the alongwind position x_s, crosswind position y_s and emission rate q_s of the three identified sources, using the same label for the sources as provided in Figure 3. The posterior mean and standard deviation, as well as the 95% HPD interval of the discrete source parameters for each of the three identified sources are summarized in Table 2. Note that sources 2 ($k = 1$) and 4 ($k = 3$), whose emission rates were not regulated owing to the failure of the mass flow controllers connected to these sources, appear to have higher emission rates than source 3 ($k = 2$) whose emission rate was properly regulated. Indeed, the measured emission rate from source 3 for Trial II was $Q = 3.8$ g s^{-1}, which compares well with the posterior mean estimate for this emission rate given in Table 2 (for $k = 2$). Finally, the information gain provided by the concentration

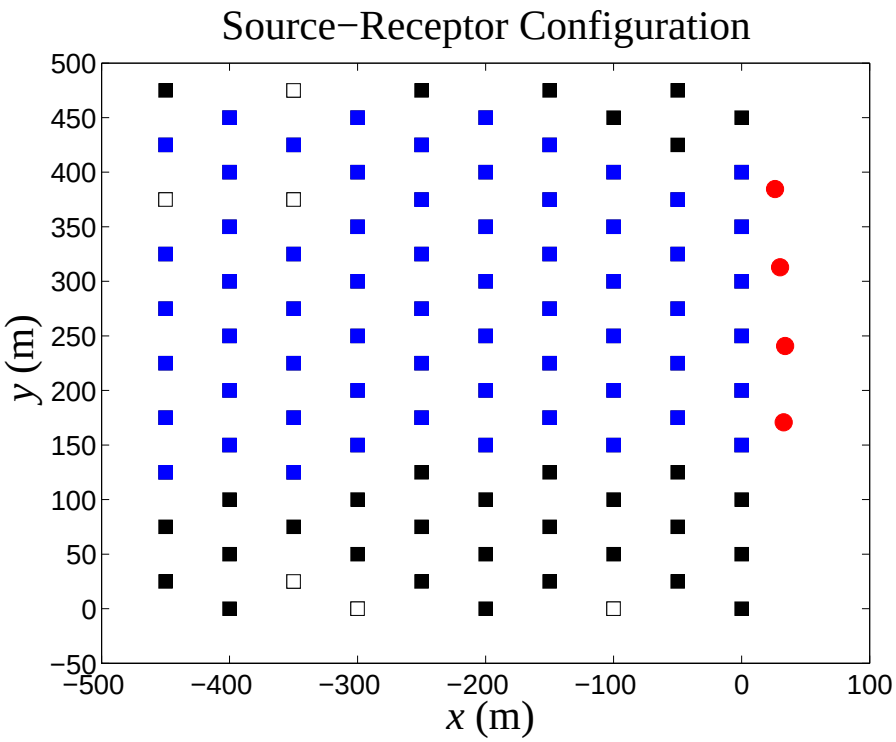


Figure 9: *In Trial III, four sources located upwind of the array of detectors were turned on. The solid dots show the locations of the four sources. Squares show the location of the detectors in the array: open and filled squares indicate that the detector at the given location is missing and present, respectively, in the array. A filled blue square marks the detectors that were used for the source reconstruction (Case 1).*

data for this example was found to be $D_{KL}(1) = 35.2$ nits.

5.3 Trial III: four-source example

5.3.1 Case 1: 62 detectors

This example involves four continuously emitting sources. In case 1, the detectors used for the source reconstruction algorithm are shown in Figure 9. As can be seen from this figure, this case involves the use of 62 detectors in the array for source inversion. All the detectors in the array that measured a significantly non-zero mean concentration were used for the reconstruction, as well as a number of detectors for which the measured mean concentration was nominally zero (viz., over the sampling time for the trial, the instantaneous concentration did not exceed the detection threshold concentration). In this example, the mean wind direction was normally incident to the detector array; more specifically, $\alpha' = 155^\circ - \alpha = -0.1^\circ$ away from normal incidence to the detector array (cf. Table 1).

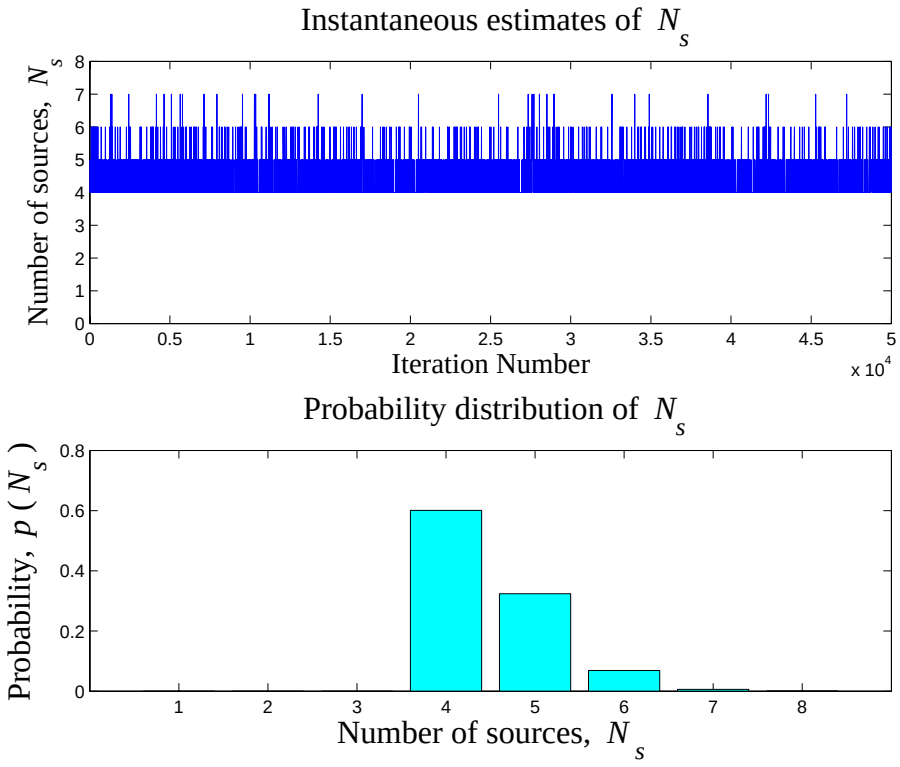


Figure 10: *Trace plot (top) of the number of discrete sources N_s in the source distribution model samples drawn from $p(\Theta|\mathbf{D}, I)$ during the probabilistic exploration phase of the stochastic sampling algorithm for Trial III (62 detectors used for source reconstruction), and the corresponding posterior distribution for the number of sources, $p(N_s) \equiv p(N_s|\mathbf{D}, I)$. In the trace plot, the number of samples displayed has been decimated by a factor of 10.*

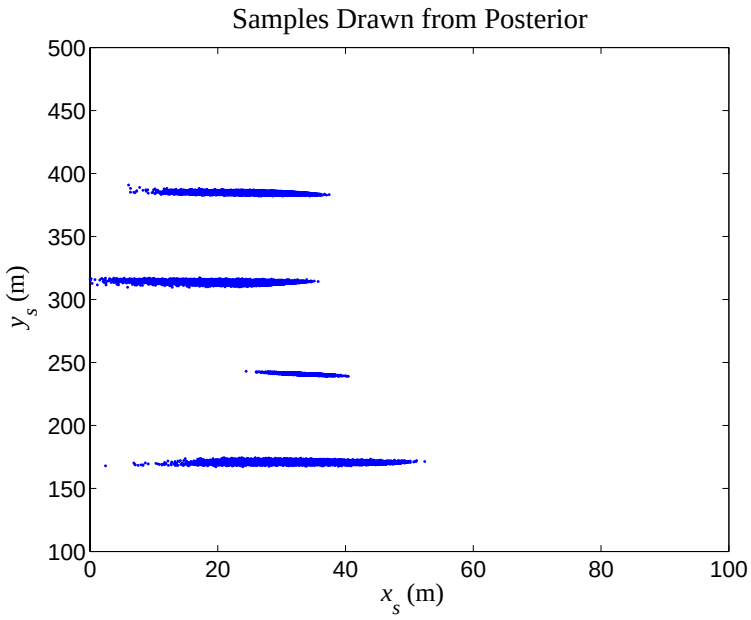
The proposed stochastic sampling algorithm randomly initializes all unknown source parameters in accordance to the prior distribution $p(\Theta|\mathbf{D})$. In this example (unlike the case dealt with in Trial II above), it is assumed that we do not have a prior knowledge of the maximum number of available sources used in FFT-07. As a consequence, we choose $N_{s,\min} = 1$ and $N_{s,\max} = 8$, with $p^* = 1/7$ in the specification of $p(N_s|I)$ implying that the expected number of sources is $\langle N_s \rangle = 2$. In consequence, our initial specification for the prior distribution of N_s favors the wrong choice for the actual number of sources. The remaining hyperparameters defining $p(\Theta|\mathbf{D})$ are chosen as follows: $\gamma = 0.25$ and $Q_{\max} = 100.0$ g s^{-1} for specification of the prior for the emission rate Q; and, $\mathcal{D} = [0, 100] \times [0, 500]$ m which is used to define the prior bounds for the location (x_s, y_s) of any source. An ensemble of $N_{\text{mem}} = 50$ members of source distribution models Θ were drawn from $p(\Theta|\mathbf{D})$ and used for the simulated annealing phase of the MCMC algorithm. After $\lambda = 1$ was achieved, 1000 further iterations of the RJMCMC algorithm were applied to each of these source distribution model members during the probabilistic exploration phase of the algorithm to give 50,000 samples of source distribution models drawn from the posterior distribution $p(\Theta|\mathbf{D}, I)$.

Figure 10 (top) shows a trace plot for the number of discrete sources in a source distribution model against the sample (or, iteration) number. From this plot, it can be seen that after the initial burn-in phase of the stochastic sampling algorithm (simulated annealing phase), the samples of source distribution models drawn from $p(\Theta|\mathbf{D}, I)$ during the probabilistic exploration phase generally mixes well over N_s. In particular, annihilation moves for models from $N_s = 4$ to 3 do not occur. However, dimension-changing moves involving transitions from $N_s = 4$ to 5 (and, vice-versa), as well as higher-order transitions (e.g., from $N_s = 6$ to 7 and its reverse) are seen to occur. Even so, it is seen that the transitions to large values of N_s (e.g., $N_s \approx N_{s,\max}$) are rare and short-lived. Figure 10 (bottom) displays the marginal posterior distribution $p(N_s) \equiv p(N_s|\mathbf{D}, I)$ for the number of sources. Note that the most probable number of sources for Trial III is 4 ($\widehat{N}_s = 4$), which is favoured with a probability of about 0.6. The most probable value for N_s in this case coincides with the correct number of sources $N_s = 4$. This result is obtained in spite of the fact that the prior distribution for N_s was initialized with $p^* = 1/7$ (with an a priori expected number of sources $\langle N_s \rangle = 2$), implying that the algorithm is not sensitive to this hyperparameter. The information embodied in the concentration data was sufficient to move the stochastic simulations towards the more complex model with $N_s = 4$ source atoms.

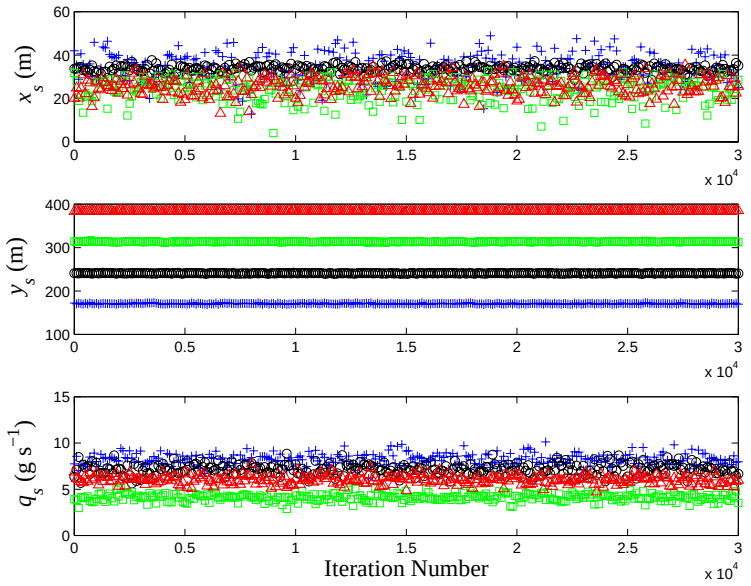
Figure 11(a) displays samples of all source distribution models drawn from the posterior distribution $p(\Theta|\mathbf{D}, I)$ with $N_s = 4$. Note that there are four clusters of points, with the centroids of each of these clusters coinciding (approximately or better) with the true location of the four sources (see Figure 3). It is evident that the concentration data $\mathbf{D}$ constrain the y_s-locations of the four sources, but the x_s-locations of these sources are subject to greater uncertainty. An examination of Figure 11(a) suggests that identifiability of the sources can be obtained by ordering the discrete sources of each source distribution model on the y_s-coordinate. More specifically, the label k for each discrete source in a source distribution model sample are reordered so that the $y_{s,k}$-locations of the discrete sources verify $y_{s,1} < y_{s,2} < y_{s,3} < y_{s,4}$. With this relabelling of discrete sources, the index k for a source follows the labelling of sources given by Figure 3. Figure 11(b) shows traces of the source parameters associated with each discrete source of the source distribution model samples (with $N_s = 4$), after the relabelling of the discrete sources. As can be seen from this figure, the y_s-locations of the discrete sources are well separated, but the x_s-locations and emission rates q_s of the discrete sources are very similar and significantly overlap one another.

Figure 12 shows the marginal posterior distribution (histogram) of the parameters x_s, y_s and q_s for each of the four discrete sources identified in Figure 11. The posterior mean and standard deviation, as well as the lower and upper bounds for the 95% HPD interval, for each of these four identified discrete sources are summarized in Table 3. For this example, it is seen that generally the estimates for the source parameters are quite good and, certainly, the true values of the parameters (when these are known) lie within the stated errors.

An examination of Table 3 shows that the emission rates for the three unregulated sources (viz., sources 1, 2, and 4) generally are larger than that for the regulated source (viz., source 3). More specifically, measurements of the emission rate from source 3 for Trial III yielded $Q = 3.8$ g s^{-1}, which agrees with the estimated value given in Table 3 for this source of



(a) Sample density plot



(b) Trace plots of source parameters

Figure 11: *(a) Density plot consisting of samples of source distribution models obtained for $N_s = 4$ projected onto the (x_s, y_s) subspace. (b) Trace plots of source parameter estimates against sample (or, iteration) number, after relabelling of discrete sources (+ (blue): $k = 1$; ∘ (black): $k = 2$; □ (green): $k = 3$; △ (red): $k = 4$). In the trace plots, the number of samples has been decimated by a factor of 100.*

$\widehat{q}_s = 4.1 \pm 0.4$ g s^{-1}. As in the previous examples, the localization of the source in the x_s-direction (alongwind) is generally poorer than that in the y_s-direction (crosswind). It can be seen that the crosswind locations of the sources can be determined with an accuracy of about ±1 m (standard deviation), but the alongwind locations of the sources can only

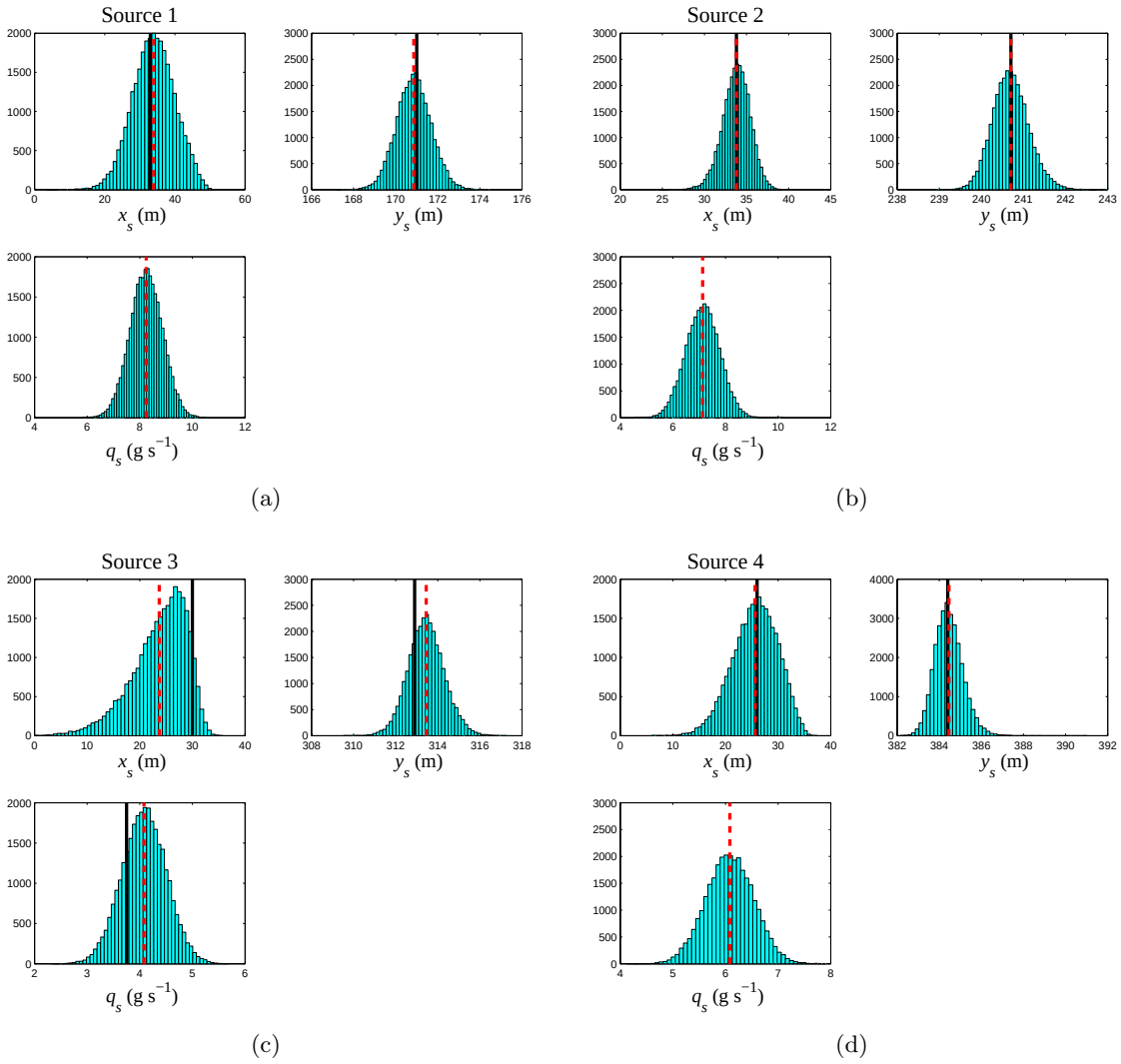


Figure 12: *Inference of the discrete source parameters obtained from samples drawn from the posterior distribution $p(\Theta|\mathbf{D}, I)$ having exactly four discrete sources. (a,b,c,d) Histograms for the three parameters, namely alongwind location x_s, crosswind location y_s, and emission rate q_s that characterize sources 1, 2, 3, and 4 (cf. Figure 3). In each frame, the solid vertical line indicates the true value of the parameter (if known) and the dashed vertical line corresponds to the best estimate of the parameter obtained as the posterior mean of the marginal posterior distribution for the parameter.*

be determined with an accuracy of about ±5 m (standard deviation). This appears to be true in this example, except for source 2 whose alongwind location is determined to an accuracy of about ±2 m (standard deviation). The reason for the increased accuracy in the determination of the alongwind location of source 2 stems from the fact that only detector 96 (along line 10) – see Figure 9 – measured a significantly non-zero mean concentration and this concentration was due solely to the emission from source 2. The information content embodied in this detector allowed the alongwind location of source 2 to be inferred with

Table 3: *The posterior mean, posterior standard deviation, and lower and upper bounds of the 95% HPD interval of the parameters $x_{s,k}$ (m), $y_{s,k}$ (m), and $q_{s,k}$ (g s^{-1}) for k = 1, 2, 3, and 4 calculated from samples of source distribution models with $N_s = 4$ for Case 1 (the latter corresponding to the most probable number of sources in the domain as inferred from Figure 10).*

	Parameter	Mean	Standard Deviation	95% HPD	Actual
$k = 1$					
	x_s (m)	34.0	6.0	$(22.8, 45.9)$	33.0
	y_s (m)	170.9	0.8	$(169.3, 172.5)$	171.0
	q_s (g s^{-1})	8.2	0.6	$(7.0, 9.4)$	–
$k = 2$					
	x_s (m)	33.8	1.6	$(30.5, 37.0)$	33.8
	y_s (m)	240.5	0.4	$(239.8, 241.5)$	240.7
	q_s (g s^{-1})	7.1	0.6	$(5.9, 8.4)$	–
$k = 3$					
	x_s (m)	23.7	5.2	$(13.1, 32.1)$	30.0
	y_s (m)	313.4	0.9	$(311.8, 315.2)$	312.9
	q_s (g s^{-1})	4.1	0.4	$(3.3, 4.9)$	3.8
$k = 4$					
	x_s (m)	25.6	4.3	$(17.5, 33.8)$	26.0
	y_s (m)	384.5	0.7	$(383.2, 385.8)$	384.4
	q_s (g s^{-1})	6.1	0.5	$(5.2, 7.0)$	–

greater accuracy than the other three sources. Finally, for this example, the information gain obtained from the concentration data $\mathbf{D}$ was found to be $D_{KL}(1) = 52.5$ nits, implying that the information contained in the concentration data allowed the "posterior volume" of the hypothesis space (volume of hypothesis space of reasonably large plausibility after receipt of the concentration data) to decrease by a factor of $\exp(D_{KL}(1)) \approx 6.3 \times 10^{22}$ relative to the "prior volume" of the hypothesis space (volume of hypothesis space of reasonably large plausibility before the receipt of the concentration data).

5.3.2 Case 2: 45 detectors

In case 2 of the example of four continuously emitting sources (Trial III), we consider the use of 45 detectors (shown in Figure 13) in the array for source reconstruction. In contrast to Case 1, all the detectors on sampling lines 8, 9 and 10 (viz., the three sampling lines closest to the emitting sources) have been removed. The stochastic sampling algorithm was applied to the concentration data obtained from the remaining 45 detectors, using exactly the same hyperparameters for the prior distribution $p(\Theta|I)$ as described previously for Case 1. The simulated annealing phase was applied to an ensemble of $N_{\text{mem}} = 50$ members of source distribution models Θ drawn from $p(\Theta|I)$. After the termination of the simulated annealing

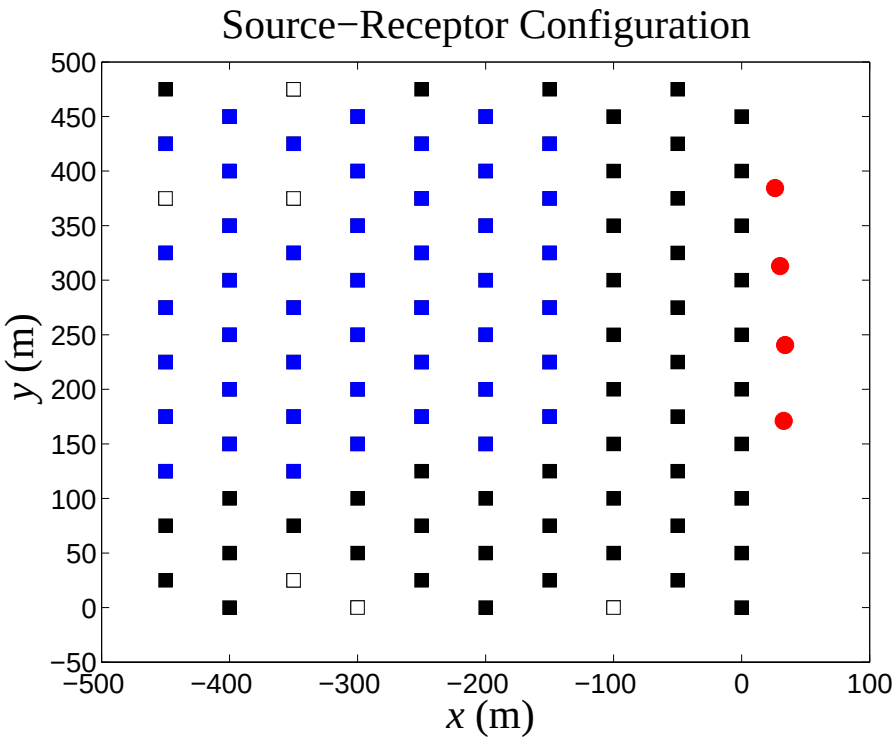


Figure 13: *In Trial III, four sources located upwind of the array of detectors were turned on. The solid dots show the locations of the four sources. Squares show the location of the detectors in the array: open and filled squares indicate that the detector at the given location is missing and present, respectively, in the array. A filled blue square marks the detectors that were used for the source reconstruction (Case 2).*

phase, the RJMCMC algorithm was run for 1000 iterations on the $N_{\text{mem}} = 50$ members of the ensemble (associated with the inverse temperature $\lambda = 1$) to generate 50,000 samples of source distribution models drawn from the posterior distribution $p(\Theta|\mathbf{D}, I)$ (during the probabilistic exploration phase).

Figure 14 exhibits a histogram of the number of sources N_s for this case, obtained from the 50,000 samples of source distribution models drawn from $p(\Theta|\mathbf{D}, I)$. The RJMCMC simulations settle in a distribution which slightly favors $N_s = 5$ sources. However, the hypotheses $N_s = 4$ and 6 are equally (approximately or better) probable, albeit with a smaller probability than the hypothesis $N_s = 5$. The results of Figure 14 would suggest that the most probable number of source in Trial III is $\widehat{N}_s = 5$ (which turns out to be incorrect).

Next, let us extract all samples of source distribution models having exactly $N_s = 5$ discrete sources. Figure 15(a) displays these samples projected onto the (x_s, y_s) plane. There are four prominent "streaks" roughly parallel to the x_s-axis, and the narrowness of these streaks usefully constrain the locations of four discrete sources in the y_s-direction. The length of

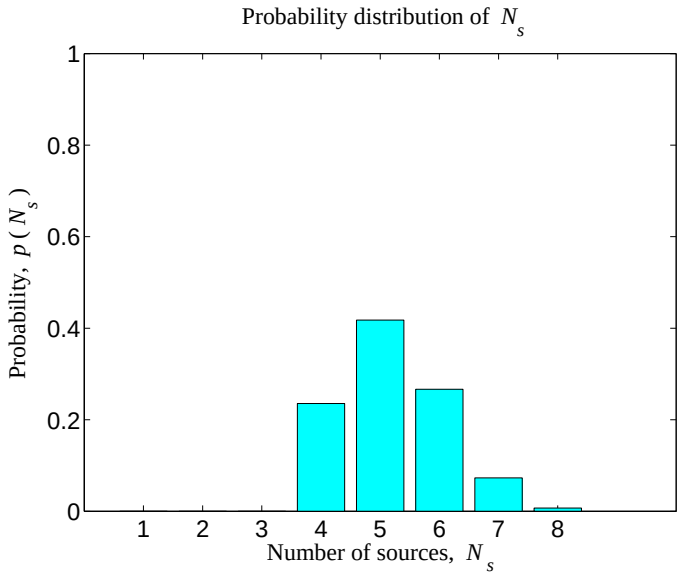
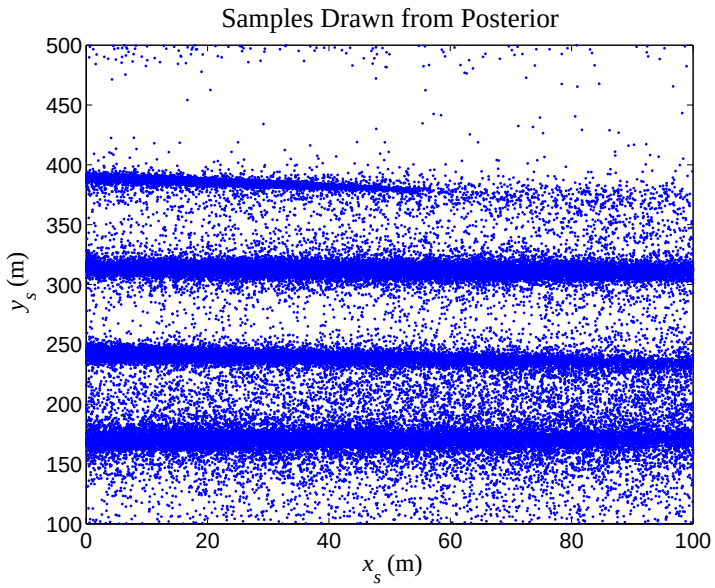


Figure 14: *The posterior distribution for the number of sources, $p(N_s) \equiv p(N_s|\mathbf{D}, I)$, for Trial III (Case 2) estimated using 50,000 samples obtained from the probabilistic exploration phase of the stochastic sampling algorithm.*

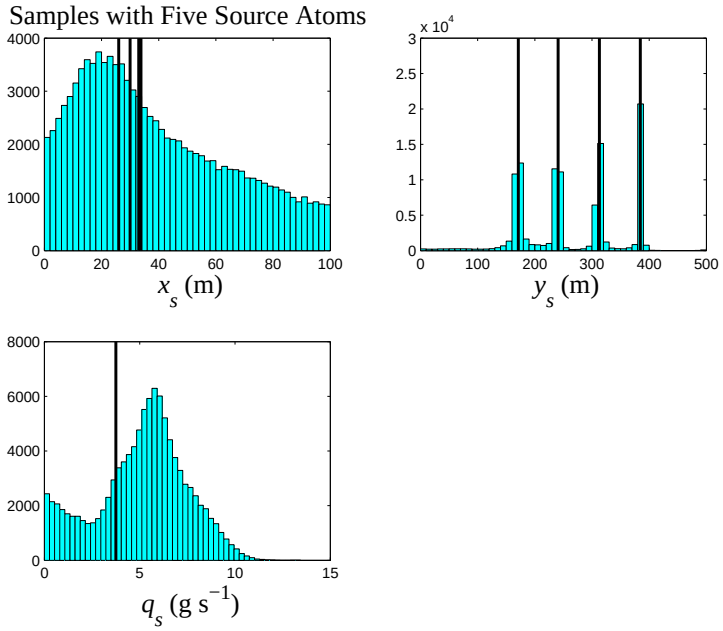
the streaks imply that the x_s-locations of these sources are only weakly constrained by the concentration data $\mathbf{D}$ (viz., there is significant uncertainty in the x_s-locations of the four identified sources). In addition to the four streaks of points, there is a diffuse cloud of points in the (x_s, y_s)-plane associated with the location of the fifth discrete source in the source distribution model samples. Note that these points appear to be randomly scattered in the (x_s, y_s)-plane, implying that the concentration data $\mathbf{D}$ do not contain information that can usefully constrain the location of this fifth discrete source.

Figure 15(b) exhibits the marginal posterior distributions for the source parameters corresponding to the samples of source distribution models with $N_s = 5$ discrete sources. Note that the histogram for y_s exhibits four modes whose locations coincide (approximately or better) with the y_s-locations of the four sources present in Trial III. The histogram for x_s has a single mode at $x_s \approx 25$ m which corresponds roughly to the x_s-locations of the four sources, but this marginal posterior distribution is very broad indicating a large uncertainty in the determination of the alongwind locations of the sources. Finally, the histogram for q_s is bimodal. There are broad modes at $q_s \approx 0$ g s^{-1} and at ≈ 7 g s^{-1}. The latter mode is associated with the four emitting sources identified in the histogram for y_s. The mode with emission rate near zero is associated with the fifth discrete source, and it is seen that the emission rate from this source is small. Indeed, the concentration contributed by this fifth source is generally smaller than the uncertainty in the specification of the concentration data $\mathbf{D}$. This simply implies that this source is not usefully constrained by the available concentration data for this case, and may simply correspond to an extraneous source.

Even though Figure 14 suggests that the most probable number of sources is $\widehat{N}_s = 5$, the



(a) Sample density plot



(b) Histogram of source parameters

Figure 15: *(a) Density plot consisting of samples of source distribution models obtained for $N_s = 5$ projected onto the (x_s, y_s) subspace. (b) Histogram of the alongwind location x_s, crosswind location y_s, and emission rate q_s of the discrete sources obtained from samples of source distribution models with $N_s = 5$. The vertical line(s) in each panel marks the true value of the source parameter (if known).*

analysis of the samples of source distribution models with $N_s = 5$ discrete sources (see Figure 15) suggests that only four of these sources are usefully constrained by the concentration data $\mathbf{D}$. Of course, there may have been one or more additional sources than those identified in Figure 15, but if the concentration data do not constrain these additional

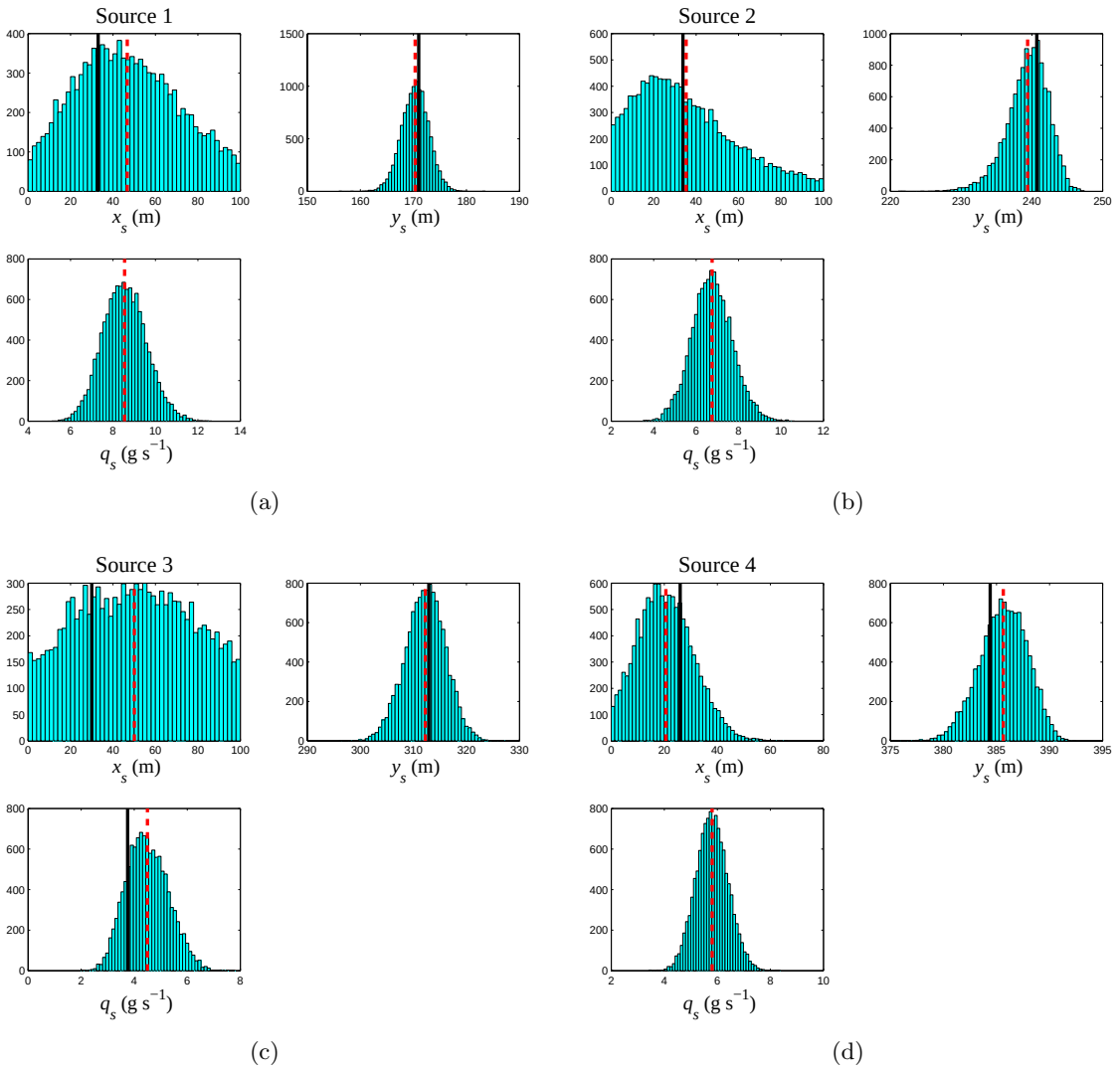


Figure 16: *Inference of the discrete source parameters obtained from samples drawn from the posterior distribution $p(\Theta|\mathbf{D}, I)$ having exactly four discrete sources. (a,b,c,d) Histograms for the three parameters, namely alongwind location x_s, crosswind location y_s, and emission rate q_s that characterize sources 1, 2, 3, and 4 (cf. Figure 3). In each frame, the solid vertical line indicates the true value of the parameter (if known) and the dashed vertical line corresponds to the best estimate of the parameter obtained as the posterior mean of the marginal posterior distribution for the parameter.*

sources, then no useful inferences can be made about them. As a consequence, it is useful to extract all samples of source distribution models with $N_s = 4$ discrete sources, and use this information to infer the source parameters for the four sources identified in Figure 15. Towards this objective, Figure 16 displays the histograms of the source parameters $\{x_s, y_s, q_s\}$ constructed from all samples with $N_s = 4$ [after relabelling the discrete sources for each sample so that their $y_{s,k}$ locations are ordered as $y_{s,1} < y_{s,2} < y_{s,3} < y_{s,4}$ — an ordering constraint that is obvious to impose after a perusal of Figure 15(b)].

Table 4: *The posterior mean, posterior standard deviation, and lower and upper bounds of the 95% HPD interval of the parameters $x_{s,k}$ (m), $y_{s,k}$ (m), and $q_{s,k}$ (g s^{-1}) for k = 1, 2, 3, and 4 calculated from samples of source distribution models with $N_s = 4$ for Case 2.*

	Parameter	Mean	Standard Deviation	95% HPD	Actual
$k = 1$					
	x_s (m)	46.7	23.7	$(3.0, 90.2)$	33.0
	y_s (m)	170.3	2.6	$(165.2, 175.3)$	171.0
	q_s (g s^{-1})	8.5	1.0	$(6.6, 10.6)$	–
$k = 2$					
	x_s (m)	35.3	22.9	$(0.0, 80.7)$	33.8
	y_s (m)	239.4	2.9	$(233.6, 244.8)$	240.7
	q_s (g s^{-1})	6.8	0.9	$(4.0, 8.6)$	–
$k = 3$					
	x_s (m)	50.1	26.4	$(6.4, 98.8)$	30.0
	y_s (m)	312.2	3.8	$(304.5, 319.2)$	312.9
	q_s (g s^{-1})	4.5	0.8	$(3.1, 6.1)$	3.8
$k = 4$					
	x_s (m)	20.6	10.3	$(0.4, 38.8)$	26.0
	y_s (m)	385.6	2.2	$(381.1, 389.8)$	384.4
	q_s (g s^{-1})	5.8	0.6	$(4.6, 7.0)$	–

Table 4 summarizes the posterior mean and standard deviation, as well as the lower and upper bounds for the 95% HPD interval of the source parameters, for each of the four discrete sources. The estimated values for the crosswind locations of the four sources agree well with the true crosswind locations, albeit the uncertainties in these inferred locations are now about three times as large as those for Case 1. The estimates of the alongwind locations of three of the four sources are quite good, although the uncertainties in these estimates are large. An examination of Figure 16(c) and Table 4 for $k = 3$ indicates that the available concentration data $\mathbf{D}$ for Case 2 does not constrain the x_s-location for Source 3. Indeed, from Figure 16(c), it can be seen that the marginal posterior distribution of x_s for Source 3 is almost a uniform distribution over its domain of definition (and, hence, almost identical to the marginal prior distribution of x_s for Source 3). The posterior mean value of q_s for Source 3 overestimates the true value of the emission rate ($Q = 3.8$ g s^{-1}). Even so, the true emission rate here is contained within the one standard deviation uncertainty interval for this parameter. Finally, for Case 2, the information gain provided by the concentration data $\mathbf{D}$ was found to be $D_{KL}(1) = 39.2$ nits for Case 2 (which is 13 nits less information gain than provided by the concentration data for Case 1). Viewed in another way, the concentration data available on the last three sampling lines (which were used for the source reconstruction for Case 1, but not for Case 2) provided 13 nits of additional information, and this information allowed the posterior "volume" in the hypothesis space for Case 1 to be reduced by a factor of $\exp(13) \approx 4.4 \times 10^5$ relative to the posterior "volume"

in the hypothesis space for Case 2.

6 Conclusions

In the report, we have developed and tested an innovative Bayesian method for source reconstruction for the difficult case when the number of sources is unknown *a priori*. More specifically, Bayesian probability theory was applied to formulate a posterior distribution for the number of sources and for the parameters that characterize each of these sources (e.g., location, emission rate, time of release, etc.). The evaluation of this posterior distribution in order to extract its features of interest is realized using an efficient computational technology described in Yee [21]. The computational algorithm uses simulated annealing for the burn-in phase and a reversible-jump MCMC method for the probabilistic exploration phase.

The source reconstruction methodology has been successfully validated against a real dispersion experiment involving various combinations of multiple source releases (FFT-07), with measurements of the resulting concentration field obtained from an array of 100 detectors. In particular, three different trials were used to test the methodology: namely, Trial I was used to test the source reconstruction methodology for the case in which the number of sources is known *a priori*; Trial II was used to test the source reconstruction methodology for the case in which the number of sources is unknown *a priori*, but the maximum possible number of sources is known (four in this case); and, Trial III was used to test the source reconstruction methodology for the case in which the number of sources is unknown *a priori*, and the maximum number of possible sources is not available. For Trial III, the effect of reducing the number of detectors on the quality of the inference was investigated. These various examples illustrate the effectiveness of the proposed source reconstruction methodology and demonstrate the reliable determination of the number of sources and estimation of the source parameters (along with the associated uncertainties) corresponding to each of the identified sources.

There are several possible extensions to this work. A number of the field experiments conducted in FFT-07 involved the presence of a spurious additive background in the concentration measured on the detector array. It would be useful to extend the source reconstruction methodology to treat an unrecognized spurious source of signal. This may require the development of a rational procedure within the Bayesian framework that would allow the proper separation of the true concentration signal from an underlying background (whose source may be unknown). The various contributions to the noise term in our model of the mean concentration observations were simply lumped together and it was assumed that a noise scale parameter (or standard deviation) associated with the aggregate error was known. Of course, apparent inconsistencies in the inference may arise from an incorrect estimate of the uncertainty of this aggregate error. It would be useful to generalize the methodology to treat the possible uncertainty in the specification of the standard deviation for the aggregate error.

References

[1] Shea, D. A. and Lister, S. A. (2003), The BioWatch Program: Detection of Bioterrorism (online), `http://www.fas.org/sgp/crs/terror/RL32152.html` (Access Date: September 2008). Congressional Research Service Report No. RL 35152.

[2] Tikhonov, A. N. and Arsenin, V. Y. (1977), *Solutions of Ill-Posed Problems*, Wiley, New York. (F. John, Translation Editor).

[3] Robertson, L. and Persson, C. (1993), Attempts to apply four dimensional data assimilation of radiological data using the adjoint technique, *Radiation Protection Dosimetry*, 50, 333–337.

[4] Robertson, L. and Langner, J. (1998), Source function estimate by means of variational data assimilation applied to the ETEX-I tracer experiment, *Atmospheric Environment*, 32, 4219–4225.

[5] Van Dop, H., Addis, R., Fraser, G., Girardi, F., Graziani, G., Inoue, Y., Kelly, N., Klug, W., Kulmala, A., Nodop, K., and Pretel, J. (1998), ETEX: A European tracer experiment; observations, dispersion modeling and emergency response, *Atmospheric Environment*, 32, 4089–4094.

[6] Seibert, P. and Stohl, A. (2000), Inverse modeling of the ETEX-1 release with a Lagrangian particle model, In Barone, G., Builtjes, P., and Giunta, G., (Eds.), *Proceedings of the Third GLOREAM Workshop, September 1999, Ischia, Italy*, pp. 95–105.

[7] Seibert, P. (2000), Methods for source determination in the context of the CTBT radionuclide monitoring system, In *Proceedings Informal Workshop on Meteorological Modeling in Support of CTBT Verification*. 6 pp.

[8] Hourdin, F. and Issartel, J.-P. (2000), Sub-surface nuclear tests monitoring through the CTBT xenon network, *Geophysical Research Letters*, 27, 2245–2248.

[9] Thomson, L. C., Hirst, B., Gibson, G., Gillespie, S., Jonathan, P., Skeldon, K. D., and Padgett, M. J. (2007), An improved algorithm for locating a gas source using inverse methods, *Atmospheric Environment*, 41, 1128–1134.

[10] Bocquet, M. (2005), Reconstruction of an atmospheric tracer source using the principle of maximum entropy. I: Theory, *Quarterly Journal of the Royal Meteorological Society*, 131, 2191–2208.

[11] Allen, C. T., Young, G., and Haupt, S. E. (2007), Improving pollutant source characterization by better estimating wind direction with a genetic algorithm, *Atmospheric Environment*, 41, 2283–2289.

[12] Yee, E. (2005), Probabilistic inference: an application to the inverse problem of source function estimation, In *The Technical Cooperation Program (TTCP) Chemical and Biological Defence (CBD) Group Technical Panel 9 (TP-9) Annual Meeting*. Defence Science and Technology Organization, Melbourne, Australia.

[13] Yee, E. (2006), A Bayesian approach for reconstruction of the characteristics of a localized pollutant source from a small number of concentration measurements obtained by spatially distributed "electronic noses", In *Russian-Canadian Workshop on Modeling of Atmospheric Dispersion of Weapon Agents*. Karpov Institute of Physical Chemistry, Moscow, Russia.

[14] Keats, A., Yee, E., and Lien, F.-S. (2007), Bayesian inference for source determination with applications to a complex urban environment, *Atmospheric Environment*, 41, 465–479.

[15] Chow, F. K., Kosović, B., and Chan, S. (2008), Source inversion for contaminant plume dispersion in urban environments using building-resolving simulations, *Journal of Applied Meteorology and Climatology*, 47, 1553–1572.

[16] Keats, A., Yee, E., and Lien, F.-S. (2007), Efficiently characterizing the origin and decay rate of a nonconservative scalar using probability theory, *Ecological Modeling*, 205, 437–452.

[17] Yee, E., Lien, F.-S., Keats, A., and D'Amours, R. (2008), Bayesian inversion of concentration data: Source reconstruction in the adjoint representation of atmospheric diffusion, *Journal of Wind Engineering and Industrial Aerodynamics*, 96, 1805–1816.

[18] Yee, E. (2007), Bayesian probabilistic approach for inverse source determination from limited and noisy chemical or biological sensor concentration measurements, In Fountain III, Augustus W., (Ed.), *Proceedings of SPIE, Chemical and Biological Sensing VIII*, Vol. 6554, 65540W, doi:10.1117/12.721630. 12 pp.

[19] Yee, E. (2007), Bayesian inversion of concentration data for an unknown number of contaminant sources, (DRDC Suffield TR 2007-085) Defence R&D Canada – Suffield.

[20] Yee, E. (2008), Inverse dispersion of an unknown number of contaminant sources, In *Proceedings of 15th Joint Conference on the Applications of Air Pollution Meteorology with the A&WMA*. New Orleans, LA, Paper 7.1, 17 pp.

[21] Yee, E. (2008), Theory for reconstruction of an unknown number of contaminant sources using probabilistic inference, *Boundary-Layer Meteorology*, 127, 359–394.

[22] Rao, K. S. (2007), Source estimation methods for atmospheric dispersion, *Atmospheric Environment*, 41, 6964–6973.

[23] Cox, R. T. (1946), Probability, frequency, and reasonable expectation, *American Journal of Physics*, 14, 1–13.

[24] Jaynes, E. T. (2003), *Probability Theory: The Logic of Science*, Cambridge University Press, Cambridge, UK.

[25] Rao, K. S. (2005), Uncertainty analysis in atmospheric dispersion modeling, *Pure and Applied Geophysics*, 162, 1893–1917.

[26] Thomson, D. J. (1987), Criteria for the selection of stochastic models of particle trajectories in turbulent flows, *Journal of Fluid Mechanics*, 180, 529–556.

[27] Flesch, T., Wilson, J. D., and Yee, E. (1995), Backward-time Lagrangian stochastic dispersion models and their application to estimate gaseous emissions, *Journal of Applied Meteorology*, 34, 1320–1332.

[28] Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1996), *Markov Chain Monte Carlo in Practice* , CRC Press, Chapman and Hall, Boca Raton, Florida.

[29] Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2003), *Bayesian Data Analysis*, Second Edition (Texts in Statistical Science), CRC Press, Chapman and Hall, Boca Raton, Florida.

[30] Green, P. J. (1995), Reversible jump MCMC computation and Bayesian model determination, *Biometrika*, 82, 711–732.

[31] Geyer, C. J. (1991), Markov chain Monte Carlo maximum likelihood, In *Computing science and Statistics: Proceedings of the 23rd symposium on the interface.* Keramidas, E. M. (ed).

[32] Gamerman, D. and Lopes, H. F. (2006), *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*, Second Edition (Texts in Statistical Science), CRC Press, Chapman and Hall, Boca Raton, Florida.

[33] Kitagawa, G. (1996), Monte Carlo filter and smoother for non-Gaussian nonlinear state space models, *Journal of Computational Graphics and Statistics*, 5, 1–25.

[34] von der Linden, W., Fischer, R., and Dose, V. (1996), Evidence integrals, In Hanson, K. M. and Silver, R. N., (Eds.), *Maximum Entropy and Bayesian Methods*, pp. 443–448, Kluwer Academic Publishers, Dordrecht, The Netherlands.

[35] Cover, T. M. and Thomas, J. A. (1991), *Elements of Information Theory*, John Wiley & Sons, Inc., New York, New York.

[36] Storwold, D. P. (2007), Detailed test plan for the Fusing Sensor Information from Observing Networks (FUSION) Field Trial 2007 (FFT-07), (WDTC-TP-07-078) West Desert Test Center, US Army Dugway Proving Ground.

[37] Chandler, G. M. (2007), Propylene dissemination system for FUSION field trials, (DRDC Suffield CR-2007-129) Defence R&D Canada – Suffield.

[38] Wilson, J. D., Flesch, T. K., and Harper, L. A. (2001), Micro-meteorological methods for estimating surface exchange with a disturbed wind flow, *Agricultural and Forest Meteorology*, 107, 207–225.

This page intentionally left blank.

DOCUMENT CONTROL DATA

(Security classification of title, body of abstract and indexing annotation must be entered when document is classified)

1. ORIGINATOR (The name and address of the organization preparing the document. Organizations for whom the document was prepared, e.g. Centre sponsoring a contractor's report, or tasking agency, are entered in section 8.)

   Defence R&D Canada – Suffield
   Box 4000, Station Main, Medicine Hat, Alberta,
   Canada T1A 8K6

2. SECURITY CLASSIFICATION (Overall security classification of the document including special warning terms if applicable.)

   UNCLASSIFIED

3. TITLE (The complete document title as indicated on the title page. Its classification should be indicated by the appropriate abbreviation (S, C or U) in parentheses after the title.)

   Validation of a Sensor-Driven Modeling Paradigm for Multiple Source Reconstruction with FFT-07 Data

4. AUTHORS (Last name, followed by initials – ranks, titles, etc. not to be used.)

   Yee, E.

5. DATE OF PUBLICATION (Month and year of publication of document.)	6a. NO. OF PAGES (Total containing information. Include Annexes, Appendices, etc.)	6b. NO. OF REFS (Total cited in document.)
May 2009	60	38

7. DESCRIPTIVE NOTES (The category of the document, e.g. technical report, technical note or memorandum. If appropriate, enter the type of report, e.g. interim, progress, summary, annual or final. Give the inclusive dates when a specific reporting period is covered.)

   Technical Report

8. SPONSORING ACTIVITY (The name of the department project office or laboratory sponsoring the research and development – include address.)

   Defence R&D Canada – Suffield
   Box 4000, Station Main, Medicine Hat, Alberta, Canada T1A 8K6

9a. PROJECT NO. (The applicable research and development project number under which the document was written. Please specify whether project or grant.)	9b. GRANT OR CONTRACT NO. (If appropriate, the applicable number under which the document was written.)
10te05	

10a. ORIGINATOR'S DOCUMENT NUMBER (The official document number by which the document is identified by the originating activity. This number must be unique to this document.)	10b. OTHER DOCUMENT NO(s). (Any other numbers which may be assigned this document either by the originator or by the sponsor.)
DRDC Suffield TR 2009-040	

11. DOCUMENT AVAILABILITY (Any limitations on further dissemination of the document, other than those imposed by security classification.)

   (X) Unlimited distribution
   ( ) Defence departments and defence contractors; further distribution only as approved
   ( ) Defence departments and Canadian defence contractors; further distribution only as approved
   ( ) Government departments and agencies; further distribution only as approved
   ( ) Defence departments; further distribution only as approved
   ( ) Other (please specify):

12. DOCUMENT ANNOUNCEMENT (Any limitation to the bibliographic announcement of this document. This will normally correspond to the Document Availability (11). However, where further distribution (beyond the audience specified in (11)) is possible, a wider announcement audience may be selected.)

13. ABSTRACT (A brief and factual summary of the document. It may also appear elsewhere in the body of the document itself. It is highly desirable that the abstract of classified documents be unclassified. Each paragraph of the abstract shall begin with an indication of the security classification of the information in the paragraph (unless the document itself is unclassified) represented as (S), (C), (R), or (U). It is not necessary to include here abstracts in both official languages unless the text is bilingual.)

A Bayesian probabilistic inferential framework provides a natural and logically consistent method for source reconstruction that fully utilizes the information provided by a limited number of noisy concentration data obtained from a network (or, array) of detectors. This report addresses the application of this framework to the difficult problem of estimating the parameters of an *a priori* unknown number of sources, using an array of detectors. To this purpose, Bayesian probability theory is used to formulate the full joint posterior probability density function for the number of sources and the parameters (e.g., location, emission rate, activation and deactivation times) that describe each source. A simulated annealing algorithm, applied in conjunction with a reversible-jump Markov chain Monte Carlo technique, is used to draw random samples from the posterior probability density function. By calculating the marginal posterior probability distribution of the number of sources from these samples, a *maximum a posteriori* estimate $\widehat{N}_s$ for the number of sources can be obtained, and all samples of source distribution models with exactly $\widehat{N}_s$ discrete sources can be used to provide best estimates for the source parameters (along with their associated uncertainties). The method is validated against a real dispersion experiment involving various combinations of multiple source releases conducted under a multinational cooperative **FU**sing **S**ensor **I**nformation from **O**bserving **N**etworks (FUSION) Field Trial 2007 (FFT-07) undertaken at US Army Dugway Proving Ground (DPG) in September 2007.

14. KEYWORDS, DESCRIPTORS or IDENTIFIERS (Technically meaningful terms or short phrases that characterize a document and could be helpful in cataloguing the document. They should be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location may also be included. If possible keywords should be selected from a published thesaurus. e.g. Thesaurus of Engineering and Scientific Terms (TEST) and that thesaurus identified. If it is not possible to select indexing terms which are Unclassified, the classification of each should be indicated as with the title.)

atmospheric dispersion
Bayesian inference
Markov chain Monte Carlo
sensor arrays
sensor/model data fusion
source reconstruction

Defence R&D Canada

Canada's Leader in Defence
and National Security
Science and Technology

R & D pour la défense Canada

Chef de file au Canada en matière
de science et de technologie pour
la défense et la sécurité nationale

www.drdc-rddc.gc.ca