



# Retrieval of Unknown Number of Source Terms in Dispersion Events Involving Multiple Point Sources

Sarvesh Kumar Singh, Maithili Sharan

## ► To cite this version:

Sarvesh Kumar Singh, Maithili Sharan. Retrieval of Unknown Number of Source Terms in Dispersion Events Involving Multiple Point Sources. *Earth and Space Science*, 2019, 6, pp.1710-1729. 10.1029/2019EA000602 . insu-03721875

**HAL Id: insu-03721875**

**<https://insu.hal.science/insu-03721875>**

Submitted on 14 Jul 2022

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Copyright

# Earth and Space Science



## RESEARCH ARTICLE

10.1029/2019EA000602

### Key Points:

- New methodology is proposed for retrieving unknown number of point releases
- Criteria are proposed for discriminating the source terms
- Detection and false-alarm probability associated to each retrieved source is estimated

### Correspondence to:

S. K. Singh,  
Sarvesh-Kumar.Singh@lsce.ipsl.fr

### Citation:

Singh, S. K., & Sharan, M. (2019). Retrieval of unknown number of source terms in dispersion events involving multiple point sources. *Earth and Space Science*, 6, 1710–1729. <https://doi.org/10.1029/2019EA000602>

Received 19 FEB 2019

Accepted 7 JUL 2019

Accepted article online 16 JUL 2019

Published online 16 SEP 2019

Corrected 29 NOV 2019

This article was corrected on 29 NOV 2019. See the end of the full text for details.

## Retrieval of Unknown Number of Source Terms in Dispersion Events Involving Multiple Point Sources

Sarvesh Kumar Singh<sup>1</sup> and Maithili Sharan<sup>2</sup>

<sup>1</sup>Laboratoire des Sciences du Climat et de l'Environnement, CEA-CNRS-UVSQ, UMR8212, IPSL, Gif-sur-Yvette, France,

<sup>2</sup>Centre for Atmospheric Sciences, Indian Institute of Technology Delhi, New Delhi, India

**Abstract** In dispersion events involving multiple sources, estimation of correct number of sources is imperative prior to their localization or source term estimation. The study pertains to a dispersion scenario of multiple point releases simultaneously emitting a nonreactive tracer that results into a mixture of concentrations, as measurements, sampled at the receptors. The objective is to estimate correct number of sources along with their release parameters (mainly locations and strengths) using a limited merged set of concentration measurements. A new methodology, based on five criteria coupled with a multiple release identification algorithm, is proposed here. Criterion helps in discriminating the releases and accounts for detection and false-alarm probability associated with each retrieved source. The number of source terms are estimated based on sum of scores achieved by each predicted source under each criterion. The methodology is evaluated with synthetic, noisy synthetic, and pseudo-real data prepared from India Institute of Technology Delhi diffusion experiment conducted in 1991 at Delhi, India. With synthetic data, the number of point sources and their parameters are retrieved exactly as their true values. With pseudo-real data, true number of sources are retrieved. The source locations and strengths are retrieved within 30 m and a factor of 3, respectively. Synthetic simulations with controlled measurement noise shows that the proposed methodology is robust in presence of model uncertainties and can be applicable in real scenarios.

## 1. Introduction

Source term estimation (STE) has been an active topic of research in recent years due to increasing vigilance for the accidental/industrial hazards, fast detection, response management, and regular monitoring for the chemical, biological, and radiological (CBR) releases in the environment (Hutchinson et al., 2017; Rao, 2007; Singh, Sharan, et al., 2015). The objective is to utilize monitoring sensor based concentration measurements to retrieve unknown information of source terms in a dispersion event, that is, number of releases, time (or duration), origin, and strength (or mass) of the releases. This requires a methodology for an optimal integration of sensor measurements with an appropriate dispersion model in order to retrieve the source term information.

In STE, retrieval of point-type sources attain particular importance due to their similarity to several accidental/industrial source term and small size in comparison to the considerable spatial scales of a city/urban terrain. Assuming the nature of release (like, point-wise) and number of sources, often, a general complex STE problem is transformed to a relatively simplified problem, for instance, retrieval of a point emission source. Even though, challenges are posed mainly due to ill-posed nature of the inverse problem, imperfect dispersion model, and noise in sensor's measurements. Several contributions have been made for the retrieval of a point source which includes new developments in methods (Allen et al., 2007; Annunzio et al., 2012b; Bocquet, 2005; Ma et al., 2017; Pudykiewicz, 1998; Sharan et al., 2009, etc.), cost function formulations (Ma et al., 2018; Thomson et al., 2007; Wang et al., 2018, etc.), and applications (Chai et al., 2018; Kumar et al., 2015; Sharan, Issartel, et al., 2012; Singh & Rani, 2014; Singh, Turbelin, et al., 2015, etc.).

Comparing to a point source identification, retrieving unknown number of releases is rather a complex, difficult, and challenging problem. Complexity may occur in the presence of multiple sources emitting different or same tracer which could be reactive or nonreactive. In the absence of knowledge about nature of releases or number of sources, the estimation of nonparametric form of source term or source distribution is

©2019. The Authors.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

quite challenging. Moreover, it is not clear if concentrations sampled for a tracer at distant monitoring networks are resulted from single release or multiple releases. The difficulties may increase for discriminating the releases when two or more sources are located in the same direction and emitting the same tracer. A less intense source lying near to the monitoring stations may raise the same concentration levels as an intensive source located far away from the monitoring stations. The other challenges include noise in the measurements, meteorological variations, limited measurements, and ill-posed nature of the inverse problem. In addition, retrieval process may suffer due to wrong assumptions about the source term, noisy measurements, imperfections in the dispersion model, lack of knowledge of error statistics, and computational limitations.

In spite of the above difficulties, notable contributions are received addressing the estimation of multiple-point releases and estimation of involved number of source terms using the sensor-measured concentrations. The identification of multiple sources, assuming the number of sources, were addressed by Yee (2007) and Sharan, Singh, et al. (2012) using, respectively, a stochastic Bayesian approach and a least squares optimization approach that were similar to those applied previously for the recovery of a single source. In probabilistic framework, Yee (2008, 2012b) proposed a Reversible Jump Markov Chain Monte Carlo (RJM-CMC) method to identify the unknown number of releases that efficiently samples the number of point sources, their release parameters. The drawbacks of these probabilistic methods are requirement of a priori information and expensive computational efforts. Yee (2012a) relaxed the computational complexity of RJM-CMC method with a model selection approach and showed an evaluation using real measurements from Fusion Field Trials (Storwald, 2007). Albo et al. (2011) developed an Aerodyne Inverse Modeling System to characterize the number of releases along with their location and evaluated 84 trials of FFT07 experiment. Wade and Senocak (2013) proposed a composite ranking score-based approach with stochastic Bayesian method to reconstruct the multiple releases along with their number of releases. Annunzio et al. (2012a) proposed a Multi-Entity Field Approximation (MEFA) method for identifying the number of releases as well as locations of the multiple point releases. Sharan, Singh, et al. (2012) have proposed a two-step inversion algorithm which utilizes adjoint (also called, sensitivity) functions to identify the parameters of the multiple releases given the known number of releases. Using this two-step algorithm, Singh and Rani (2015) provided an evaluation of the proposed multiple release algorithm using real measurements from Fusion Field Trials. Singh et al. (2013) modified the two-step inversion algorithm by introducing weights given by Issartel et al. (2007) representing information from the monitoring network. Cantelli et al. (2015) developed a genetic algorithm based computational model to modify the search procedure in Sharan, Singh, et al. (2012) for the simultaneous estimation of locations and emission rates in steady state conditions.

In present study, a criteria-based methodology is proposed to estimate number of point releases in a multiple sources dispersion event. The methodology utilizes two-step multiple-point release identification algorithm proposed by Sharan, Singh, et al. (2012) to estimate the release parameters (mainly locations and strengths). The study is described in seven sections. Section 2 restates multiple-point release identification algorithm given by Sharan, Singh, et al. (2012) and describes five criteria to estimate the number of releases. The methodology is evaluated with pseudo-trials generated by combining single release trials, obtained in India Institute of Technology (IIT) Delhi diffusion experiment, in identical meteorological conditions. Section 3 describes preparation of these pseudo-trials from IIT Delhi diffusion experiment. Numerical computations and evaluation with the IIT Delhi pseudo-trials are given in sections 4 and 5, respectively. Finally, scope and limitations associated with the present study and conclusions are summarized in sections 6 and 7, respectively.

## 2. Inverse Modeling

The present study assumes a scenario where an unknown number of point releases are continuously and simultaneously emitting a nonreactive tracer that results into a mixture of concentrations sampled as a merged set of concentration measurements at the receptors. The objective here is to retrieve the number of point sources along with their release parameters using limited merged set of concentration measurements. For simplification, we assume that (i) measurements are linear mixture of concentrations that resulted from each point source, and (ii) releases are continuous and ground level (i.e., time and vertical dimensions are ignored in the formulations). Due to simplifications, the point sources can be localized by estimating their release locations and strengths. Note that bold symbol denotes a vector/matrix, italic symbol denotes a scalar/constant, and symbol “T” denotes “Transpose”.

The methodology is composed of a source-receptor formulation (section 2.1), multiple-point release identification algorithm (Sharan, Singh, et al., 2012; section 2.2), and criteria (section 2.3) to discriminate between the releases.

### 2.1. Source-Receptor Relationship

The unknown sources are mapped to the sampled measurements through the use of sensitivity coefficients in a discretized space (Pudykiewicz, 1998). The sensitivity coefficient describes sensitivity of unit measurement into the emission space with respect to each receptor. Let  $\mu \in \mathbb{R}^m$  is the vector of measured concentrations and  $\mathbf{s} \in \mathbb{R}^N$  is the vector of unknown emissions in a discretized space composed of  $N$  cells. The measurements  $\mu$  are related to the emissions  $\mathbf{s}$  as,

$$\mu = \mathbf{A}\mathbf{s} + \epsilon \quad (1)$$

where  $\mathbf{A} \in \mathbb{R}^{m \times N}$  (such that  $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_N]$ ) is the sensitivity matrix and  $\epsilon \in \mathbb{R}^m$  is the residual vector accounting for the instrumental noise and model errors. Each column vector  $\mathbf{a}_i \in \mathbb{R}^m$  (such that  $\mathbf{a}_i = \mathbf{a}(\mathbf{x}_i)$  where  $\mathbf{x} = (x, y)$  is a location vector) of matrix  $\mathbf{A}$  denotes sensitivity of a cell with respect to the  $m$  measurements. The matrix  $\mathbf{A}$  is obtained as solutions from the adjoint dispersion model with respect to each measurement (Sharan et al., 2009).

Let us assume that an unknown emission vector is composed of  $p$  different point releases located at cells  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_p$  emitting continuously a tracer with release rates  $q_1, q_2, \dots, q_p$ , respectively, such that

$$s_i = q_i \delta(\mathbf{x} - \mathbf{x}_i), i = 1, 2, \dots, p \quad (2)$$

By substituting equation (2) into equation (1), one obtains

$$\mu = \mathbf{H}\mathbf{q} + \epsilon \quad (3)$$

in which matrix  $\mathbf{H} \in \mathbb{R}^{m \times p}$  denotes sensitivity of  $p$  releases locations with respect to  $m$  measurements and  $\mathbf{q} \in \mathbb{R}^p$  is the vector of unknown release strength. Now, for a fixed number of  $p$  point releases, the problem is to estimate vectors  $\mathbf{q}, \mathbf{x}_1, \dots, \mathbf{x}_p$  such that the Euclidean norm  $\epsilon^T \epsilon$  is minimum.

### 2.2. Multiple-Point Release Identification

For estimation of release parameters (locations and strengths), an algorithm mentioned by Sharan, Singh, et al. (2012) is adopted here. Given  $m > p$ , identification of  $p$  point releases is considered as an overdetermined parametric estimation problem which involves estimation of  $p$  location vectors and  $p$  strengths. The estimation is based on minimization of a least squares cost function formulated as the sum of the squares of residuals (i.e.,  $\epsilon^T \epsilon$ ). This can be written as

$$J = \frac{1}{2} [\mu - \mathbf{H}\mathbf{q}]^T [\mu - \mathbf{H}\mathbf{q}] \quad (4)$$

In equation (4),  $\epsilon$  are assumed normally distributed with mean zero and constant variance. For a fixed set of  $\mathbf{x}_1, \dots, \mathbf{x}_p$ , an estimate of strength ( $\hat{\mathbf{q}}$ ) can be estimated by equating  $\frac{\partial J}{\partial \mathbf{q}} = 0$  as

$$\hat{\mathbf{q}} = [\mathbf{H}^T \mathbf{H}]^{-1} \mathbf{H}^T \mu \quad (5)$$

It is not difficult to show that the estimate  $\hat{\mathbf{q}}$  will be a local minima of  $J$  if  $\mathbf{H}^T \mathbf{H}$  is positive definite since  $\frac{\partial^2 J}{\partial \mathbf{q}^2} = \mathbf{H}^T \mathbf{H}$ . Using the equations (4) and (5), a minimum value of  $J$  for the chosen set  $\mathbf{x}_1, \dots, \mathbf{x}_p$  is derived as

$$\hat{J} = \frac{1}{2} \left[ \mu^T \mu - \mu^T \mathbf{H} (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \mu \right] \quad (6)$$

Using equations (5) and (6), a step-wise algorithm is constructed to retrieve the locations and strengths of  $p$ -point releases.

- Step 1: Choose a set of  $p$  location vectors  $\mathbf{x}_1, \dots, \mathbf{x}_p$  and compute the product  $\mathbf{H}^T \mathbf{H}$  for chosen pair of locations.
- Step 2: If  $\mathbf{H}^T \mathbf{H}$  is positive definite, proceed to step 3, otherwise go to step 1.
- Step 3: Estimate  $\hat{\mathbf{q}}$  using equation (5) and compute  $\hat{J}$ . Store  $\hat{J}$  along with the set of  $p$  locations.

- Step 4: Repeat steps 1 to 3 for all possible sets of  $p$  location vectors.
- Step 5: Find minimum of  $\hat{J}$  among the stored values.
- Step 6: The set of  $p$  location vectors corresponding to  $\min(\hat{J})$  will be the desired set of release locations.
- Step 7: The strengths for  $p$  locations can be estimated using equation (5).

### 2.3. Estimating Minimal Number of Releases

Estimation of unknown number of releases is performed in three steps. In the first step, methodology begins by assuming a priori maximum number of releases, say  $p = p^{\max}$ . The release parameters (locations and strengths) corresponding to  $p^{\max}$  releases are estimated using the multiple-point release identification algorithm. In the second step, proposed criteria are applied to obtain scores corresponding to  $p^{\max}$  releases. Based on achieved scores, an estimate for number of releases ( $p^{\text{est}}$ ) is determined. In the final step, the release parameters are retrieved for  $p = p^{\text{est}}$  number of releases by reapplying multiple-point release identification algorithm. Estimation of minimal number of releases is based on an analysis of retrieved strength  $\hat{\mathbf{q}}$  and matrix  $\hat{\mathbf{H}}$  containing the sensitivity vectors corresponding to the retrieved release locations. Here, we mentioned five criteria which analyze the proportion of strengths during the retrieval, significance of sensitivity vectors in explaining the measurements, collinearity and principal components among the sensitivity vectors, and detection and false-alarm probability associated with each retrieved source. Each criterion results to a score in the interval  $[0,1]$ .

#### 2.3.1. Criterion-1 ( $C^1$ )

The first criterion  $C^1$  is based on analyzing the retrieved strength  $\hat{q}_i$  at  $p$  release locations. Note that the measurement vector  $\boldsymbol{\mu}$  is a linear combination of the  $\hat{q}_i$  and  $\mathbf{a}_i$  for  $i = 1, 2, \dots, p$  (equation (3)). With noise free measurements and perfect model setup, it is expected that  $\hat{q}_i$  will be nonzero only for the linearly independent  $\mathbf{a}_i$  vectors. However, in presence of measurement noise and model errors,  $\hat{q}_i$  may become nonzero for all  $p$  release locations since  $\mathbf{a}_i$  may not be linearly independent. Thus, a score defined as  $\frac{\hat{q}_i}{\sum_{i=1}^p \hat{q}_i}$ , indicating the proportion of  $\hat{q}_i$ , is considered as a primary indicator to determine the number of potential sources. A disadvantage with this criterion is the chance of missing a source term with low emission strength but the other criterion can account for it.

#### 2.3.2. Criterion-2 ( $C^2$ )

This criterion analyzes relative contribution of the measurements in explaining the retrieved source terms. For this, a projection matrix  $\mathbf{D} \in \mathbb{R}^{p \times m}$  is derived, which describes the projection of measurements onto the sensitivity vectors corresponding to the retrieved locations. It also describes influence of measurement on the sensitivity vector. The projection of  $\boldsymbol{\mu}$  onto the sensitivity vectors  $\mathbf{a}_i$  is defined as

$$\mathbf{d}_i = \frac{\boldsymbol{\mu}^T \mathbf{a}_i}{\|\mathbf{a}_i\|} \mathbf{a}_i \quad i = 1, 2, \dots, p^{\max} \quad (7)$$

where  $\mathbf{d}_i$  are column vectors of projection matrix  $\mathbf{D}$ . The eigen decomposition of covariance matrix of  $\mathbf{D}$  helps to identify the components that have the largest spread (or variance) of this projection. The normalized variance of this projection can be interpreted as relative influence of the measurements on the sensitivity vectors. Further, covariance matrix of  $\mathbf{D}$  is computed whose eigenvalues will describe the relative contribution of measurements in explaining the individual releases since the sensitivity vectors corresponding to the estimated release locations are the basis vectors representative for releases in the projection matrix. Thus, a score indicating the relative proportion of the eigenvalues is considered an indicator to filter the minimal number of releases among  $p$  releases.

#### 2.3.3. Criterion-3 ( $C^3$ )

This criterion examines for the variance or spread of the sensitivity vectors associated to the retrieved locations. This is derived through the covariance matrix of  $\hat{\mathbf{H}}$ , that is,  $\mathbf{Q} = \text{cov}(\hat{\mathbf{H}})$  (see Appendix A), which shows the variance and covariances among the sensitivity vectors. An eigen analysis of  $\mathbf{Q}$  highlights principal components, spread in the sensitivity vectors and variance explained by the corresponding principal components. This also helps in determining collinearity among the sensitivity vectors. An occurrence of large eigenvalue can be related to the significance of a release as the number of significant eigenvalues indicate number of principle spread directions in the sensitivity vectors. A small eigenvalue may correspond to the artificial release location. Thus, the proportion of eigenvalues of  $\mathbf{Q}$  is considered as a score to determine the number of significant releases.

#### 2.3.4. Criterion-4 ( $C^4$ )

This criterion is a scaled version of  $C^3$  which examines the correlation matrix  $\mathbf{R} = \text{corr}(\hat{\mathbf{H}})$  (see Appendix A) for the correlations among sensitivity vectors in  $\hat{\mathbf{H}}$ . This criterion is efficient in filtering out collinear or linearly dependent sensitivity vectors which might correspond to the artificial release location. In this criterion, a score is derived as eigenvalues of correlation matrix  $\mathbf{R}$ . Similar to  $C^3$ , a large eigenvalue will be an indication of the potential release and a small eigenvalue will correspond to an artificial release.

#### 2.3.5. Criterion-5 ( $C^5$ ): Detection and False-Alarm Probability

This criterion originated from Neyman-Pearson detection theory-based (Neyman et al., 1933; Royall, 2017) eigen-thresholding method (Chang, 2013). This defines associated detection and false-alarm probability of the corresponding releases based on an eigenvalue threshold. Let  $\mathbf{Q} \in \mathbb{R}^{p \times p}$  and  $\mathbf{R} \in \mathbb{R}^{p \times p}$  are the covariance and correlation matrices of matrix  $\hat{\mathbf{H}}$ . Let  $\lambda_1^Q \geq \lambda_2^Q \geq \dots \geq \lambda_p^Q$  and  $\lambda_1^R \geq \lambda_2^R \geq \dots \geq \lambda_p^R$  are two sets of eigenvalues generated by  $\mathbf{Q}$  and  $\mathbf{R}$  called covariance and correlation eigenvalues, respectively. Since the component dimensionality is equal to the total number of eigenvalues, each eigenvalue specifies a component dimension and provides an indication of the significance of that particular component in terms of energy or variance.

If there is no source contained in a particular component, the corresponding correlation eigenvalue and covariance eigenvalue in this component should reflect only the noise variance, in which case, correlation eigenvalue and covariance eigenvalue are equal. This fact provides us with a base from which we can formulate the difference between the correlation eigenvalue and its corresponding covariance eigenvalue as a binary composite hypothesis testing problem (Chang & Du, 2004). We assume here that (i) sources are positive and associated with white noise (Gaussian distributed with zero mean), and (ii) the covariance between these two types of eigenvalues is asymptotically zero. Thus, we can expect that  $\lambda_i^R \geq \lambda_i^Q$  for distinct number of sources. In order to determine the number of distinct sources, a binary hypothesis can be formulated as

$$\begin{aligned} H_0 : z_i &= \lambda_i^R - \lambda_i^Q = 0, & i = 1, 2, \dots, p \\ H_1 : z_i &= \lambda_i^R - \lambda_i^Q > 0 \end{aligned} \quad (8)$$

where null hypothesis  $H_0$  and the alternative hypothesis  $H_1$  represent the case that the correlation eigenvalue is equal to its corresponding covariance eigenvalue and the case that the correlation eigenvalue is greater than its corresponding covariance eigenvalue, respectively. These pair of eigenvalues  $\lambda_i^R$  and  $\lambda_i^Q$  under hypotheses  $H_0$  and  $H_1$  are random variables whose asymptotic conditional probability densities are given by

$$\begin{aligned} p_0(z_i) &= p(z_i|H_0) \sim N(0, \sigma_z^2) \\ p_1(z_i) &= p(z_i|H_1) \sim N(\bar{z}, \sigma_z^2) \end{aligned} \quad (9)$$

where  $N(\bar{z}, \sigma_z^2)$  denotes Gaussian distribution with mean  $\bar{z}$  and variance  $\sigma_z^2$ . In view of equation (9), the false-alarm probability ( $p_F$ ) and detection power (i.e., detection probability,  $p_D$ ) can be defined as

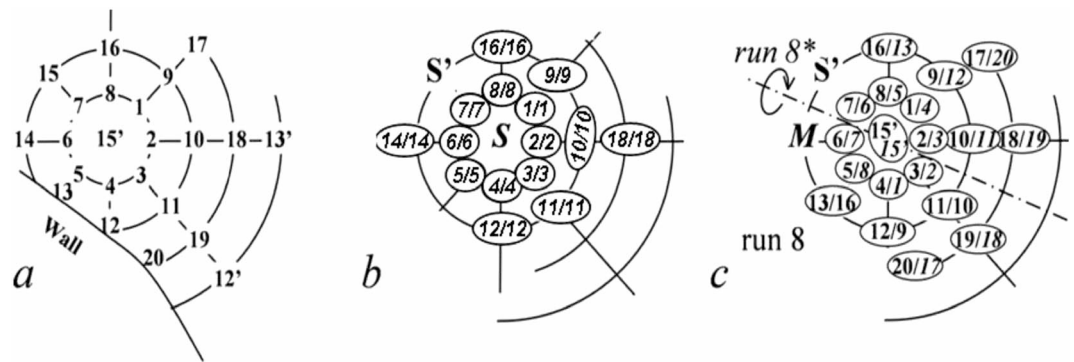
$$\begin{aligned} p_F &= \int_{\tau_i}^{\infty} p_0(z) dz \\ p_D &= \int_{\tau_i}^{\infty} p_1(z) dz \end{aligned} \quad (10)$$

Thus, for a fixed  $p_F$ , a threshold  $\tau_i$  is determined for each  $z_i$  such that  $z_i > \tau_i$  which indicates that there is an energy contributing to the eigenvalue  $\lambda_i^R$  in the  $p$  dimension (i.e., alternative hypothesis  $H_1$  is true) and corresponding detection probability can be computed. This criterion provides two important information related to  $p$  retrieved releases: (i) what are the false-alarm and detection probabilities with respect to each predicted source? and (ii) what is the probability of detecting a release under a given threshold allowing a limited false-alarm probability?. Thus, a score is defined as  $C_i^5 = p_D^i - p_F^i$ .

### 3. IIT Delhi Experiment: Pseudo-Runs

IIT Delhi experiment was a short-scale dispersion experiment, conducted in 1991, New Delhi, India, involving several runs (or trials) of tracer  $\text{SF}_6$  release from single source location in low wind convective conditions (Sharan et al., 1996; Singh et al., 1991). In this experiment, the source location varies in some runs of the release. Taking advantage of identical meteorological conditions and different source locations in some runs,





**Figure 1.** (a) Layout of the site for IIT Delhi tracer diffusion experiment. (b) Representation of pseudo-runs #11–13 and #8–11 (Sharan, Singh, et al., 2012). Within each ellipse the right and left italic numbers indicate combined receptors. (c) Schematic similar to panel (b) for pseudo-run #8–8\* in which run #8 is combined with itself after a mirror symmetry sending the source to M.

Sharan, Singh, et al. (2012) combined single release runs to generate pseudo-runs with pseudo-real measurements. The details have been given in Sharan, Singh, et al. (2012); however, for completeness, salient features are restated here.

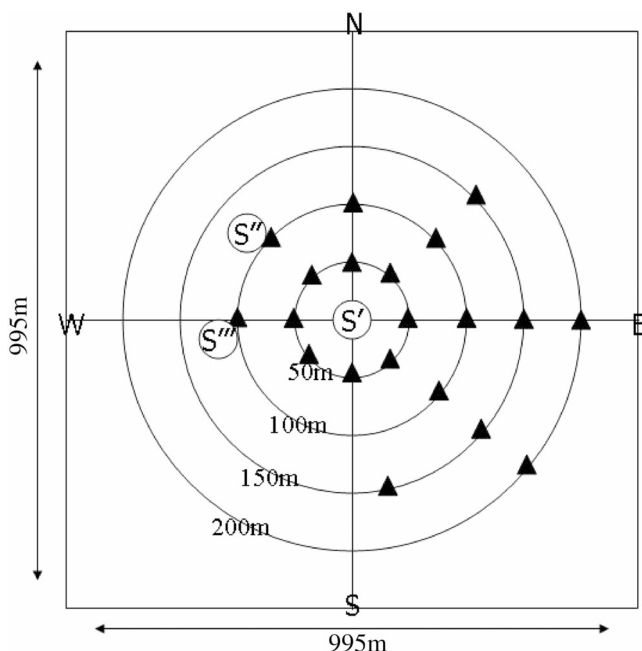
The monitoring network of IIT Delhi experiment consists of 20 receptors, arranged in circular arcs of radii 50, 100, 150, and 200 m at an angular distance of 45° (Figures 1 and 2). The height of release was 1 m and the concentrations were also sampled at approximately the same height. There were a total of 14 runs in which only three (run #8, run #11, and run #13) corresponding to steady state convective conditions are chosen under similar meteorological conditions (mean wind speeds of 1.1, 0.9, and 1.1 m/s, respectively, and mean wind directions of 125°, 121°, and 141°) for generating the pseudo-real measurements.

The tracer release experiments were divided into two categories: (i) the release point taken at the center (run #11 with release rate  $3 \times 10^{-3}$  g/s) and, (ii) shifted to 100 m northwest (runs #8 and #13 with release rate  $5 \times 10^{-3}$  g/s). To generate the pseudo-real measurements, both types of experiments are combined as if two releases occur simultaneously and the concentration measurements on the coinciding receptors are simply added. In this manner, two runs (#11–13 and #8–11) are constructed simply for two simultaneous

point releases. The wind direction and wind speed are taken as the average of two runs. Another pseudo-run #8–8\* corresponding to two simultaneous releases is prepared by superimposing the mirror image of receptors in run #8 along the symmetric axis for the wind direction of 112.50° (Figure 1). The coinciding receptor's measurements are added to generate pseudo-measurements. Similarly, a third pseudo-run corresponding to three simultaneous point releases is made by adding the concentration measurements of coinciding receptors of run #11 with the run #8–8\*. The wind direction is taken as 112.5° and the wind speed is averaged between runs #8 and #11.

#### 4. Numerical Computations

The inversion computations are implemented on a discretized domain of size 995 m × 995 m, discretized into 199 × 199 cells. The size of each cell is 5 m × 5 m. The cell size is chosen such that the measurement cells are sufficiently separated from each other. The center of the receptor arcs corresponds to the cell (100, 100). In pseudo-runs #11–13 and #8–11, two sources correspond to the cells (100, 100) and (86, 114) whereas in run #8–8\*, two sources lie on cells (86, 114) and (80, 100). Similarly, in run #8–8\*–11, three sources correspond to the cells (100, 100), (86, 114), and (80, 100). Further, in text, the true sources will be denoted as  $s_1^{\text{true}}$ ,  $s_2^{\text{true}}$ ,  $s_3^{\text{true}}$  representing cells (100,100), (86, 114), and (80, 100), respectively.



**Figure 2.** An illustration of computational domain for IIT Delhi dispersion experiment.

Overall, the computational steps for estimation of unknown releases can be divided into four steps: (i) computation of sensitivity vectors (or matrix  $\mathbf{A}$ ), (ii) prior estimation of location and strength for a priori chosen maximum number of sources ( $p^{\max}$ ) using multiple release identification algorithm (section 2.2), (iii) estimation of the number of sources  $p^{\text{est}}$  using the criteria (section 2.3) and (iv) estimation of location and strength corresponding to the  $p^{\text{est}}$  sources by reapplying multiple release identification algorithm. In this study, a maximum number of sources for prior estimation is chosen as  $p^{\max} = 5$ .

#### 4.1. Computation of Sensitivity Matrix

The sensitivity vector  $\mathbf{a}_i$  (column of sensitivity matrix  $\mathbf{A}$ ) describes the sensitivity of emission space with respect to the measurement. The sensitivity vectors are derived from the solutions of an adjoint dispersion model. In this study, an analytical dispersion model developed by Sharan et al. (1996) is considered for its use in adjoint mode. Due to simplicity of the model, an adjoint of the model can be implemented by reversing the wind speed and replacing the source terms by the measurement cells with unit release (Sharan et al., 2009). In computations, source and receptors are assumed at ground level. The sensitivities have peaks on the cells containing measurements which spread to their neighboring cells. This raises a numerical artifact in the inversion computations (Issartel et al., 2007). To resolve this, the measurement cells and neighboring cells are further subdivided into  $99 \times 99$  cells and an average value of the sensitivity coefficient is computed for the measurement cells.

#### 4.2. Score for Estimating Number of Releases

The estimated strengths corresponding to  $p^{\max}$  releases and matrix  $\hat{\mathbf{H}}$  (containing sensitivity vectors corresponding to the retrieved  $p^{\max}$  source locations) are analyzed under the criterion proposed in section 2.3. Each criterion analyzes different features associated with the sensitivity vectors and results to a value between 0 and 1 for each retrieved source. Combining all the criteria, a joint score  $\beta_i$  is built for each retrieved source by summing the values resulted from each criterion, that is,  $\beta_i = C_i^1 + C_i^2 + C_i^3 + C_i^4 + C_i^5$  for  $i = 1, 2, \dots, p^{\max}$ . An  $i$ th retrieved source is considered significant (or potential) if it achieves more than a threshold score  $\beta_i \geq 1$ . Thus,  $p^{\text{est}}$  is counted as number of sources which scores more than the threshold score.

#### 4.3. Reduction in Computational Time

The implementation of the algorithm presented in section 2.2 is computationally expensive. An exhaustive searching for  $p$  pair of source locations require to visit  $N(N-1)\dots(N-p+1)/p! \approx N^p$  cells, where  $N = 199 \times 199 = 39,601$  cells. The storage and computations of such a large number of pairs is tedious and expensive; for instance, the computational time taken in the estimation of source parameters for simultaneous three-point releases is approximately 90 hr on an Intel(R)Core™ i5 – 3427UCPU@1.80GHz laptop machine. This is expected to increase further with an increase in the number of sources and number of cells in the domain. Hence, in order to reduce the computational time, two strategies are proposed which reduces the search space for an application of the algorithm.

##### 4.3.1. Filtering of Well-Monitored Regions

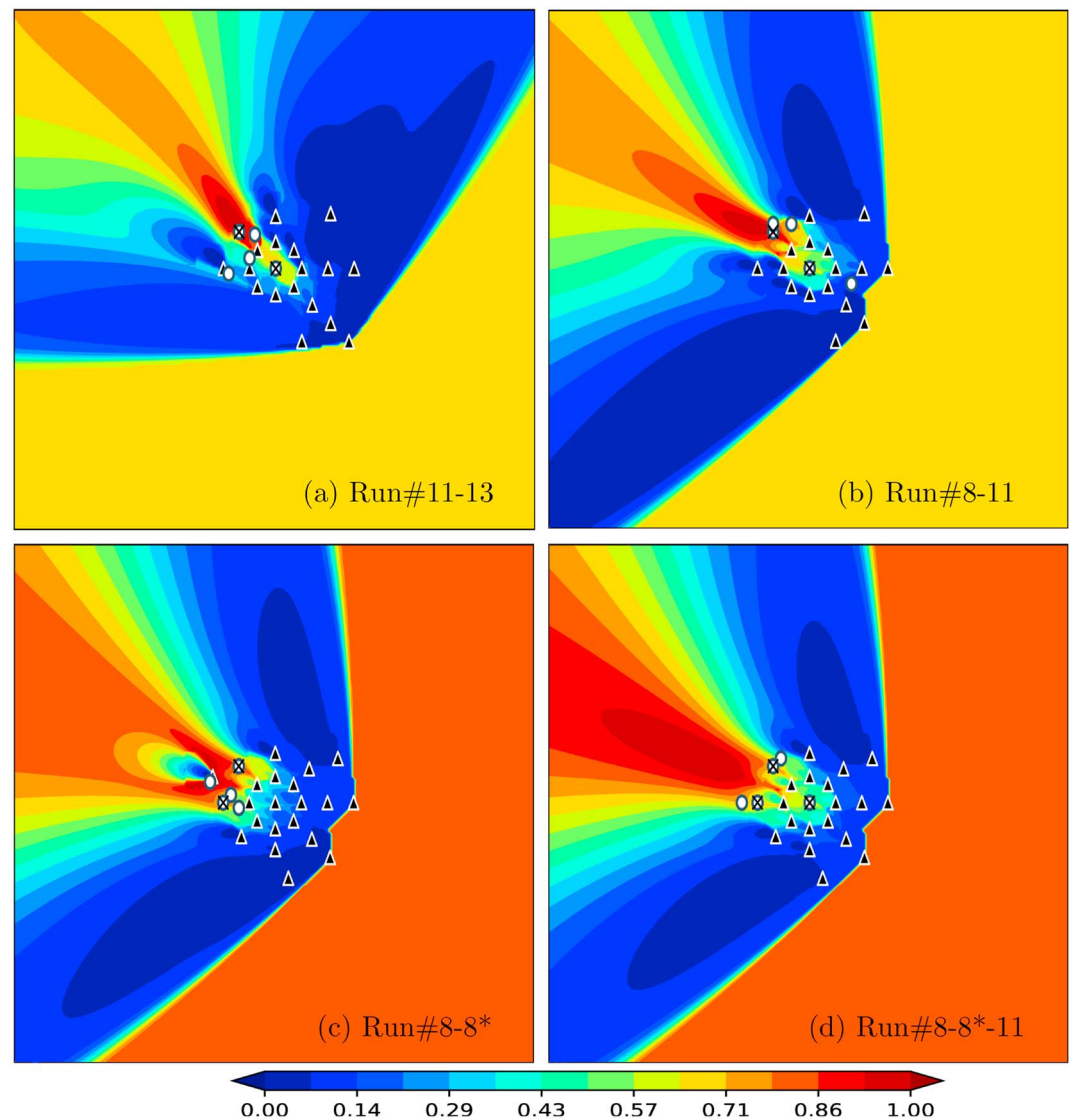
In order to obtain an information about possible source regions, we have adopted two strategies: (i) filtering regions based on the norm of sensitivity vectors and (ii) estimating a nonparametric or distributed form of the source term independent to its nature and type. For the first strategy, we choose a threshold on the norm of sensitivity vector as  $\sqrt{\mathbf{a}^T(\mathbf{x})\mathbf{a}(\mathbf{x})} > 10^{-7}$ . So, the regions having sensitivity vector norm lesser than  $10^{-7}$  are ignored for the source retrieval.

For the second strategy, an estimate for the source distribution is estimated by minimizing a least squares functional  $J(\mathbf{s}) = \frac{1}{2}(\boldsymbol{\mu} - \mathbf{A}\mathbf{s})^T(\boldsymbol{\mu} - \mathbf{A}\mathbf{s})$  following equation (1) as

$$\hat{\mathbf{s}} = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \boldsymbol{\mu} \quad (11)$$

The estimate  $\hat{\mathbf{s}}$  is similar to the maximum likelihood estimate of the source term under constant prior information and describes the conditional distribution of unknown releases based on the given set of measurements. The estimate is independent of assumptions about nature or number of releases. The estimate is utilized to filter the well-seen regions a priori to the application of the multiple-point release identification algorithm (section 2.2). To filter well-seen source regions, threshold criteria is defined as, a location  $\mathbf{x}$  is chosen only if  $\hat{s}(\mathbf{x}) / \max(\hat{\mathbf{s}}) > 0.05$ . Indeed, the first strategy with a threshold value of  $10^{-7}$  favors sources located close to the monitoring network. For the second strategy, a source estimate (known as the nonparametric, minimum norm estimate) is used with a threshold value of 5%. This estimate has the property to





**Figure 3.** Estimate  $\hat{s} / \max(\hat{s})$  for synthetic measurements in runs #11-13 (panel a), #8-11 (panel b), #8-8\* (panel c) and #8-8\*-11 (panel d). The symbol “x” and “o” denote true and retrieved source, respectively. The symbol “black triangle” denotes receptors.

have large values in the regions located close to the monitoring networks. Overall, these two strategies favor sources located directly upwind of the monitoring network. In the present study, this reduces the search space to approximately  $(N/4)^p$  cells instead of  $N^p$  cells and hence, reduces the computational cost.

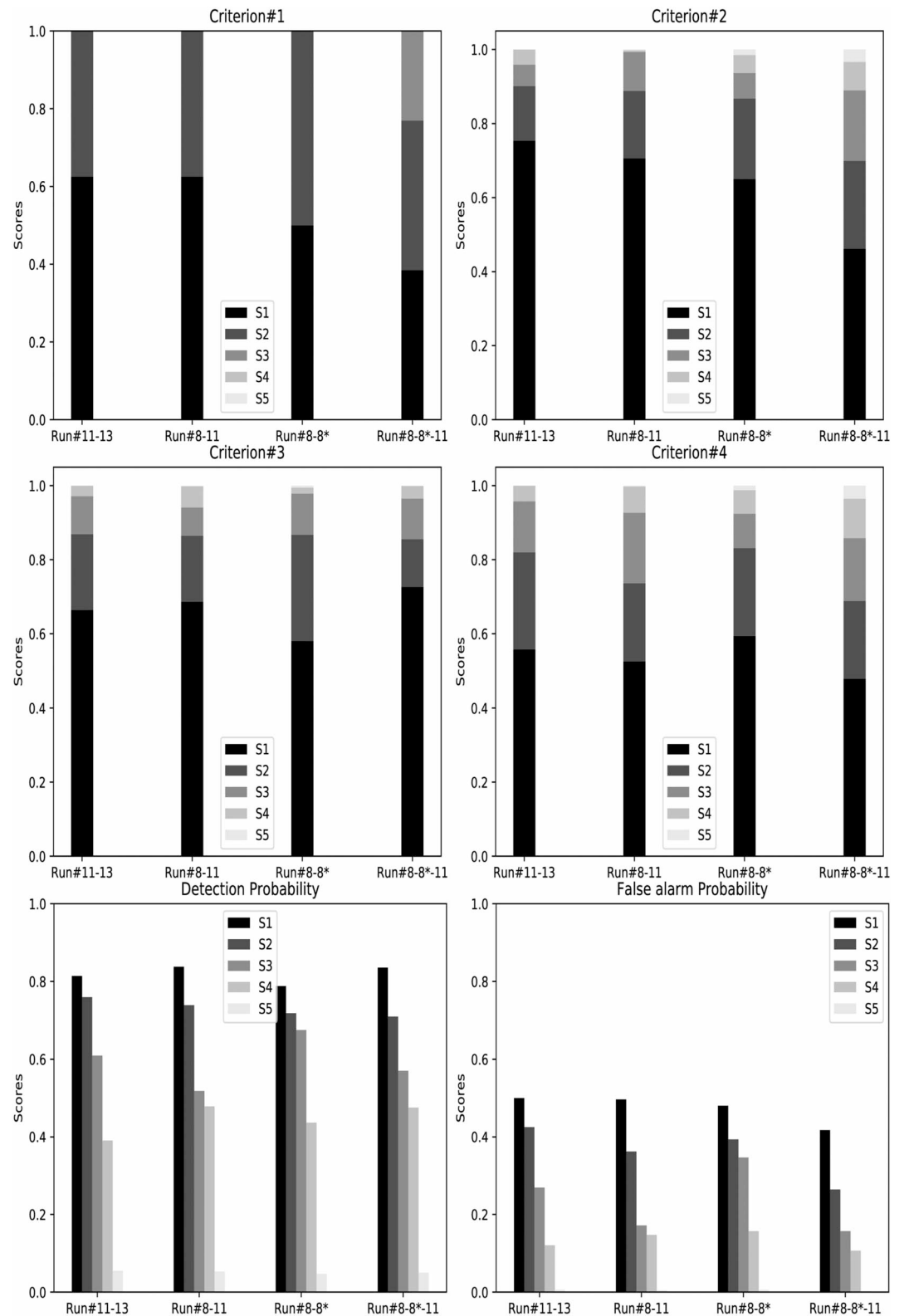
#### 4.3.2. Visiting Alternative Pair of Points

To further reduce the computational expense and time, a gross estimation of the source parameters is carried out in the well-monitored source regions by visiting one cell out of each three cells in each direction. This reduces the number of pairs by a factor of  $3^{2p}$  for the  $p$ -set. Finally, a gross estimation is further refined, for all possible sets of pairs associated with a locally minimum  $\hat{J}$ , by revisiting each pair in their seven neighboring cells.

Following both the strategies, computational time taken in estimation of two, three, and five simultaneous releases is 20 s, 3 hr, and 10 hr, respectively.

## 5. Results and Discussions

The estimation of unknown releases is performed for both synthetic and pseudo-real measurements. Synthetic measurement refers to the model concentrations generated using a forward analytical dispersion



**Figure 4.** Cumulative bar chart for the scores that resulted from each criterion using synthetic measurements. Top panels show criterion #1 and criterion #2. In the middle, criterion #3 and criterion #4 are shown. The bottom panels display Detection and False alarm probability.

model (Sharan et al., 1996) assuming the true values of the release parameters (locations and strengths). Synthetic measurements are free from any measurement or model errors and, thus, are utilized to verify consistency of the methodology and proposed criteria.

The source estimation results are discussed in terms of location errors and factor of retrieved to true release strength. Location error is derived as a Euclidean distance between the retrieved source and nearest lying true source. The features of source estimation are illustrated through the isopleths of estimate  $\hat{s}/\max(\hat{s})$  (Figures 3, 6, and 9). The maxima region of estimate  $\hat{s}$  highlights the regions with maximum source information, whereas small values of  $\hat{s}/\max(\hat{s})$  indicate regions with poor source information. In case of least squares identification of single point source, Singh and Rani (2014) highlight that the maximum of  $\hat{s}$  is associated with the location of the point release. However, this is not true in case of unknown multiple releases. Though, a sharp maximum may highlight presence of a source term while a flat maxima region may indicate large uncertainty in discriminating the sources.

The criterion scores for each source term are represented in the form of a cumulative bar chart for both synthetic (Figure 4) and pseudo-real data (Figure 7). The retrieved source terms are represented as "S1, S2, S3, S4, and S5". In criterion  $C^5$ , the detection and false-alarm probability are shown for each source term. The retrieved sources having a score  $\beta_i < 1$  are considered artificial, which appear due to limited measurement information and uncertainties. In the following, the source estimation results are presented for both synthetic and pseudo-real runs.

## 5.1. Synthetic Measurements

### 5.1.1. A Priori Source Retrieval

Here, we describe source retrieval results for a priori chosen  $p^{\max} = 5$  sources with synthetic measurements. In all the pseudo-runs, true source locations are exactly retrieved among the  $p^{\max}$  retrieved locations. We can say that the sources S1 and S2 coincide with two true sources in runs #11–13, 8–11, and 8–8\*. Similarly, S1, S2, and S3 coincide exactly with  $s_1^{\text{true}}$ ,  $s_2^{\text{true}}$ , and  $s_3^{\text{true}}$ , respectively, in run #8–8\*–11. The other source locations are retrieved within a distance of 85 m from the true sources. The estimate  $\hat{s}$  has relatively sharp maxima region in runs #11–13, 8–11, and 8–8\* whereas a large flat maxima region is observed in run #8–8\*–11 (Figure 3). We observed that  $s_2^{\text{true}}$  falls in the global maxima region in runs #11–13 and 8–11 whereas in other runs, true sources fall in the locally maxima regions which are relatively less resolved by the monitoring network.

In run #11–13, the other three sources (S3, S4, and S5) are obtained in the vicinity of receptors with a location error of 30, 53, and 82 m, respectively. In run #8–11, S3 and S4 are observed at a distance of 15 and 38 m from  $s_2^{\text{true}}$ , while S5 is observed in the far downwind (85 m from  $s_1^{\text{true}}$ ) of the monitoring network. In run #8–8\*, S3, S4, and S5 are observed within 50 m (21, 30, and 47 m) radii of the  $s_3^{\text{true}}$ . In run #8–8\*–11, S4 and S5 are observed close (21 and 30 m) to the  $s_2^{\text{true}}$  and  $s_3^{\text{true}}$ , respectively. Overall, we observed that the other sources (except those coinciding with true sources), are observed either in the vicinity of the receptors or downwind of the monitoring network.

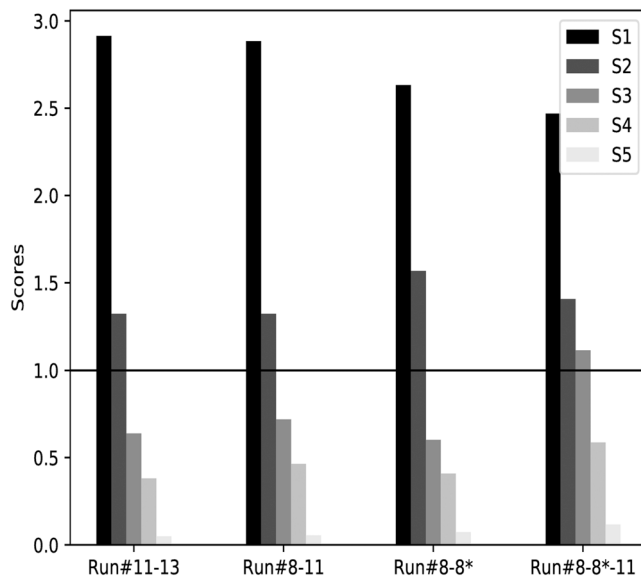
With synthetic data, the release strengths are estimated exactly to their true values for the retrieved locations coinciding with the true source locations. For other retrieved sources, strengths are estimated zero. This gives a hint that the other sources with zero strength are not real. However, we will further validate this hint with the evaluation of criterion scores.

### 5.1.2. Criteria Scores

The scores are determined based on the analysis of sensitivity vectors corresponding to the retrieved source locations and retrieved strengths under each criterion (Figure 4).

The first criterion ( $C^1$ ) clearly shows that there exist only two true point sources (S1 and S2) in runs # 11–13, 8–11, 8–8\*, and three point sources (S1, S2, and S3) in run #8–8\*–11. The strengths for other source terms in these runs tend to zero which indicates that the other source terms are nonexistent.

In second criterion ( $C^2$ ), it is observed that the S1 has large impact from the measurements ( $>0.6$  in runs # 11–13, 8–11, and 8–8\* and  $>0.4$  in run #8–8\*–11) in comparison to the other source terms. Cumulatively, S1 and S2 are associated with  $\approx 90\%$  of the measurement spread in two release runs. Similarly, S1, S2, and S3 cumulatively describe  $\approx 90\%$  of the measurement spread in run #8–8\*–11. The contribution of other source terms are observed comparatively very small ( $<10\%$ ) in comparison to the cumulative contribution of S1 and S2 in two release runs and S1, S2, and S3 in run #8–8\*–11.



**Figure 5.** Score  $\beta$  for synthetic measurements.

in all the runs. In run #8-8\*, S3 also has a significant probability to be detected. For other source terms (except for S3 in runs #11-13, #8-8\*, and #8-8\*-11), the detection probability is observed relatively small ( $<50\%$ ). This shows that S1 and S2 are clearly detectable in all the runs while S3 also has more than 50% probability to be detected in runs #11-13, #8-8\*, and #8-8\*-11. Thus, runs #11-13 and #8-8\* might include the possibility of containing three source terms. The analysis of false-alarm probability shows that all the source terms have less than 50% probability of false detection. Comparatively, false detection probability is observed relatively large ( $\approx 0.5$ ) for S1 in all the runs, whereas, for other source terms, false detection probability is significantly small.

The sum  $\beta$  of all the scores resulted from each criterion for each source term shows that there exist only two potential source terms (S1 and S2) in runs #11-13, #8-11, #8-8\*, and three source terms (S1, S2, and S3) in run#8-8\*-11 (Figure 5) which exactly matches with the true number of sources in these runs.

### 5.1.3. Source Term Estimation for $p^{\text{est}}$ Releases

The retrieval algorithm is reapplied for the estimated number of significant releases ( $p^{\text{est}} = 2$ ) in runs #11-13, #8-11, and #8-8\* and  $p^{\text{est}} = 3$  in run #8-8\*-11. In all the runs, source locations and strengths are retrieved exactly as given true parameters. The exact retrieval of the source terms equivalent to true sources is not surprising and rather reflect the mathematical consistency of the multiple release identification algorithm.

## 5.2. Pseudo-Real Measurements

Here, the source retrieval and criterion will be associated with an unknown structure of model errors and error due to artificiality in preparation of the real measurements.

### 5.2.1. A Priori Source Retrieval

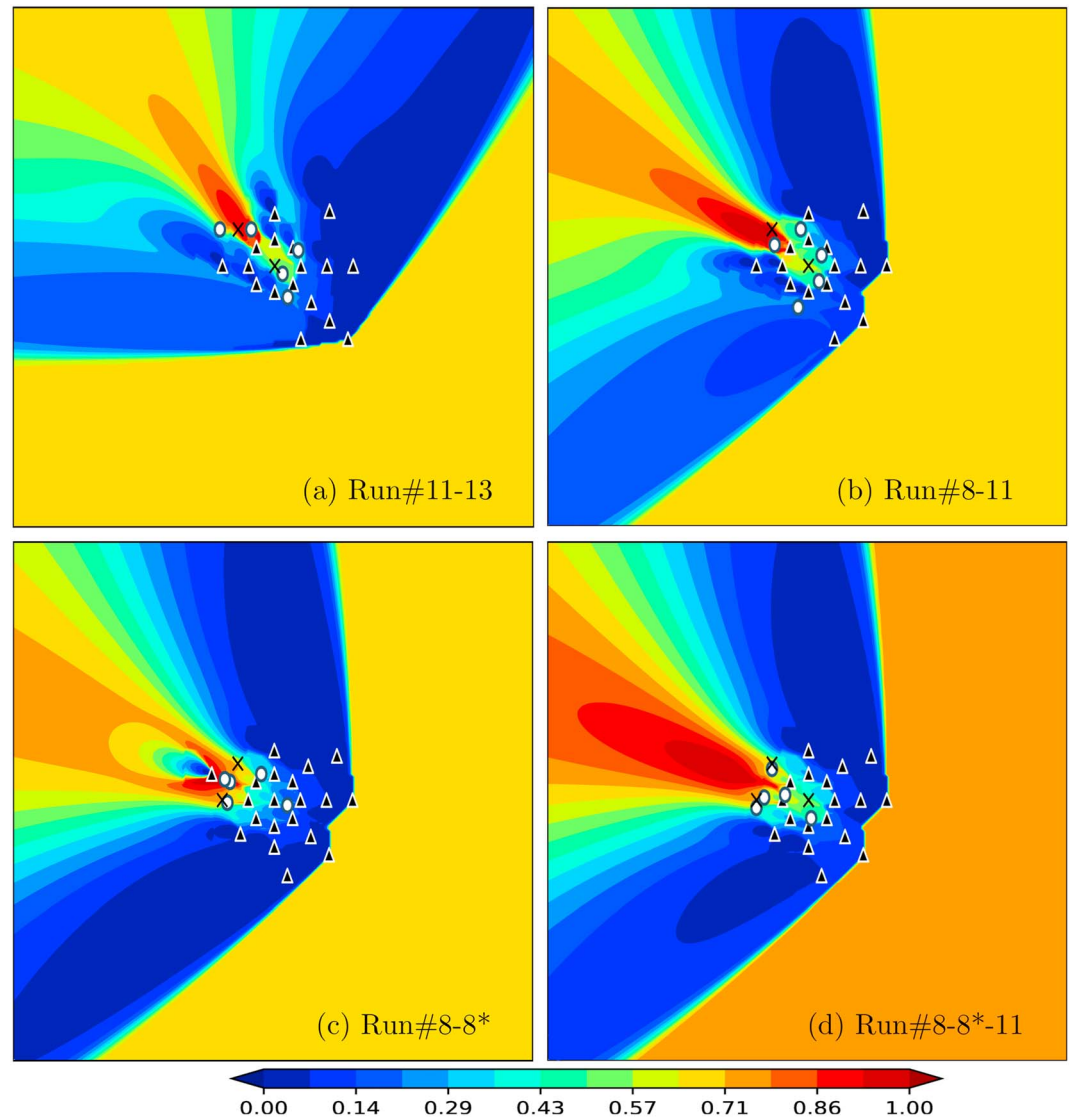
Among  $p^{\text{max}} = 5$  releases, the source locations are retrieved with a minimum and maximum location error of 7 and 120 m, respectively. We observed that, at least, one of the source terms is always obtained within a 30-m radii of the true sources. The estimate  $\hat{s}$  has sharp maximum in runs #11-13 and #8-8\* and a flat maxima region in runs #8-11 and #8-8\*-11. The true sources mostly fall in the locally maxima region.

In run#11-13, two sources S1 and S2 are retrieved with a distance of 10 and 17 m, respectively, to the true sources  $s_1^{\text{true}}$  and  $s_2^{\text{true}}$ . Among other sources, S3 is obtained at 30-m distance to the  $s_2^{\text{true}}$ , whereas S4 and S5 are retrieved within 60-m distance of  $s_2^{\text{true}}$ . Source term S1 lies in globally maxima region of the estimate  $\hat{s}$ , and S2 corresponds to the locally maxima region. Sources S4 and S5 are retrieved downwind of the monitoring network close to the receptors whereas S3 is retrieved upwind of the monitoring network but falls in the poorly resolved region.

Following third criterion ( $C^3$ ), it is observed that S1 describes a large proportion ( $\approx 0.6$ ) of total variance of the sensitivity vectors. In two release runs, S1 and S2 cumulatively describe more than 80% of the total variance. Similarly, in run #8-8\*-11, S1, S2, and S3 cumulatively describe  $\approx 95\%$  of the total variance. On comparing the proportions of S2 and S3, we observed that they contribute similarly in run #8-8\*-11 whereas the proportion of S3 is almost half of S2 in two release runs. Overall, criterion  $C^3$  indicates a small possibility of presence of S3 in two release runs.

The fourth criterion ( $C^4$ ) shows that S1 has relatively large proportion in comparison to the other source term, but the other source terms, especially S2 and S3 in two release runs and S2, S3, and S4 in run #8-8\*-11, are observed significant. The S1 and S2 in runs #11-13 and #8-8\*, and S1, S2, and S3 in run #8-8\*-11, cumulatively describe  $\approx 80\%$  of the total variance of the sensitivity vectors. In run #8-11, criterion  $C^4$  indicates the presence of three significant source terms as S3 has approximately similar impact as S2 and, S1 and S2 cumulatively describe only  $\approx 75\%$  of the total variance. Overall,  $C^4$  indicates a small possibility of presence of S3 in run #8-8\* but the impact of S3 is significant in runs #11-13 and #8-11. Similarly, there is a small possibility of the presence of S4 in run #8-8\*-11.

On comparing the detection probability for the source terms, it is clear that two source terms S1 and S2 has large probability ( $>0.7$ ) to be detected

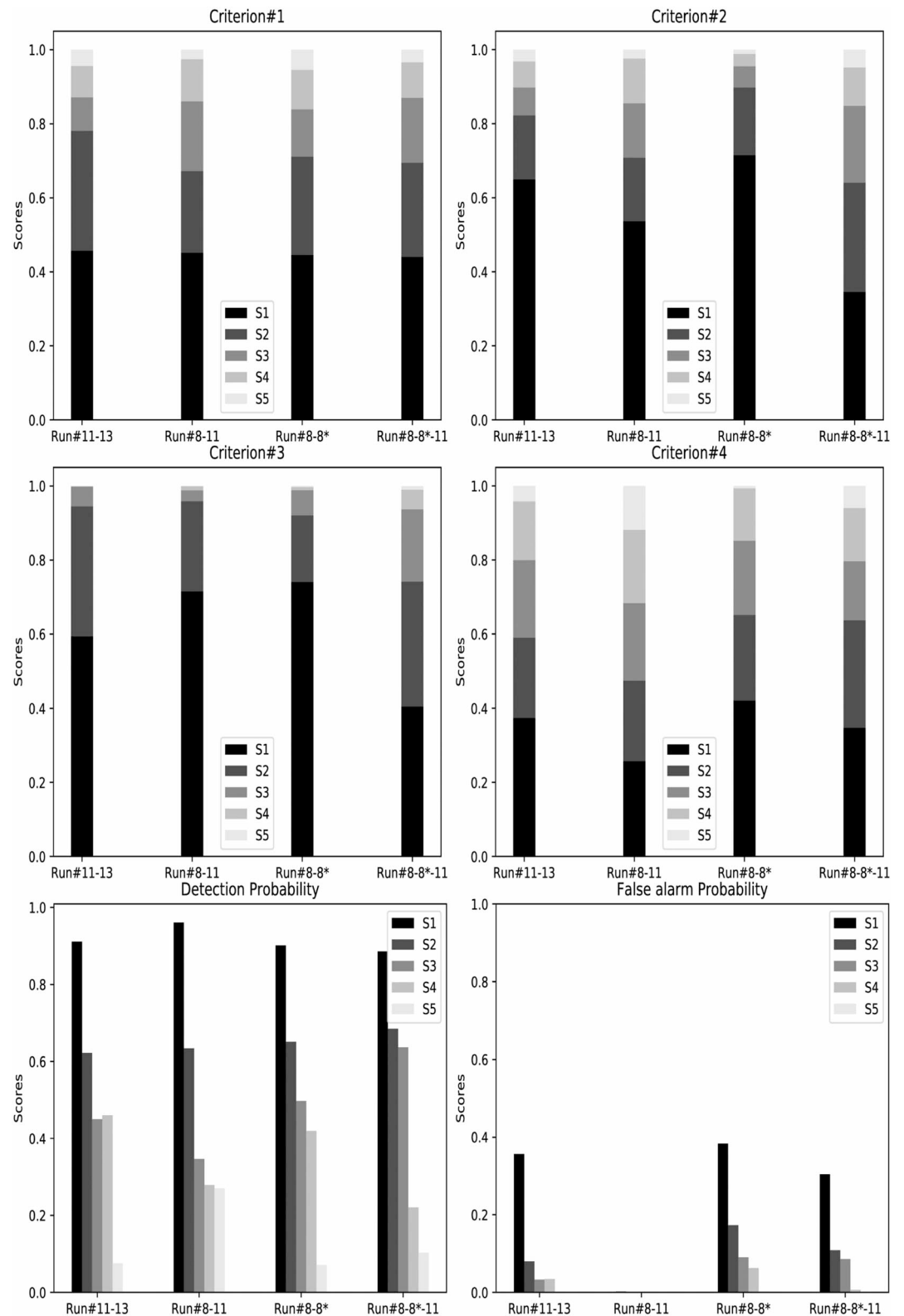


**Figure 6.** Estimate  $\hat{s} / \max(\hat{s})$  for pseudo-real measurements in runs #11-13 (panel a), #8-11 (panel b), #8-8\* (panel c) and #8-8\*-11 (panel d). The symbol “x” and “o” denote true and retrieved source, respectively. The symbol “black triangle” denotes receptors.

In run #8-11, source term S1 is retrieved within 15 m of the true source  $s_2^{\text{true}}$ , and two sources S2 and S3 are retrieved within 40 m from the true source  $s_1^{\text{true}}$ . The sources S4 and S5 are retrieved at a distance of 50 and 77 m from the true sources  $s_2^{\text{true}}$  and  $s_1^{\text{true}}$ , respectively. Source S1 is observed in the global maxima region of  $\hat{s}$  and S2, S3, and S4 lies around the locally maxima region. The source term S5 is retrieved downwind of the monitoring network in the poorly resolved region.

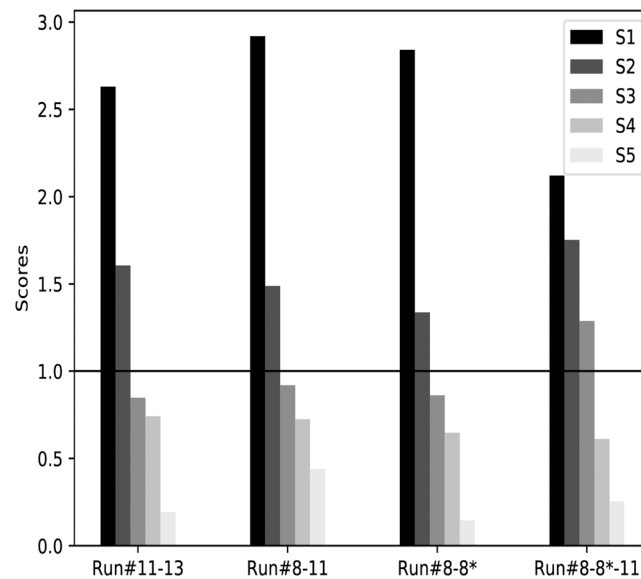
In run #8-8\*, source S1 is retrieved within 7 m of the  $s_3^{\text{true}}$  and sources S2 and S3 are retrieved within 35-m distance from the  $s_2^{\text{true}}$ . Sources S4 and S5 are retrieved at 46- and 120-m distance, respectively, from the  $s_2^{\text{true}}$ . The source S1 lies in the locally maxima region of  $\hat{s}$  and S2 and S3 lies in the global maxima regions of  $\hat{s}$  (Figure 6). However, source terms S4 and S5 are observed close to the receptors and downwind of the true source locations which is not well seen by the monitoring network.

In run #8-8\*-11, source S1 is retrieved at 25-m distance to  $s_1^{\text{true}}$ . The S2 is retrieved close (within 7 m) to  $s_2^{\text{true}}$  and S4 and S5 are retrieved within 7 m to the  $s_3^{\text{true}}$ . The source S5 is observed between the receptors, almost equidistant ( $\approx 50$  m) from all the true sources. All the sources lies in almost well-seen region of the monitoring network. The global maxima of  $\hat{s}$  is obtained flat along the upwind of the monitoring network.



**Figure 7.** Cumulative bar chart for the scores resulted from each criterion using pseudo-real measurements. Top panels show criterion #1 and criterion #2. In the middle, criterion #3 and criterion #4 are shown. The bottom panels display Detection and False alarm probability.





**Figure 8.** Score  $\beta$  for pseudo-real measurements.

### 5.2.2. Criteria Scores

With criterion  $C^1$ , we observed that the proportion of source strength for S1 and S2 is relatively higher than other source terms in all the runs and cumulatively describes  $\approx 70\%$  of the total score (Figure 7). In runs #11–13 and #8–8\*, only two source terms are observed significant, that is, scores for other source terms are comparatively small, whereas in runs #8–11 and #8–8\*–11, three sources cumulatively achieve  $>80\%$  score. Overall,  $C^1$  indicates clearly two significant source terms in runs #11–13 and #8–8\*, three source terms in runs #8–11 and #8–8\*–11.

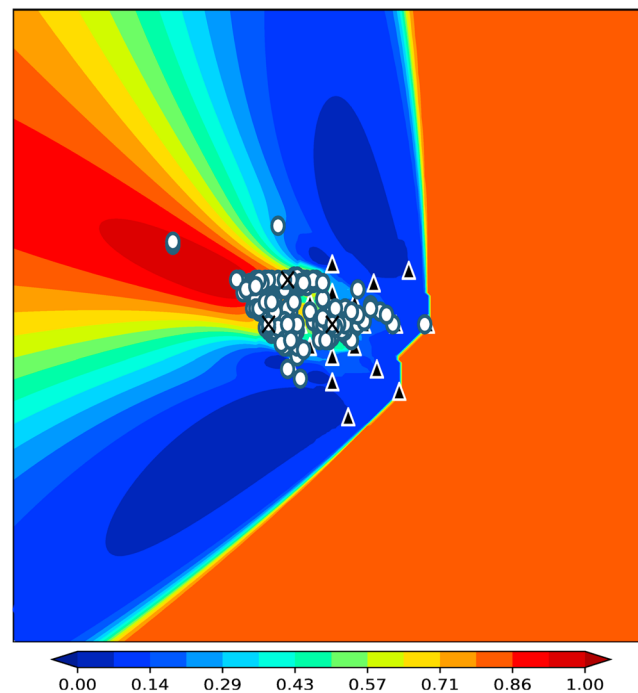
Criterion  $C^2$  shows that S1 and S2 are associated with  $\approx 80\%$  of the measurement spread in runs #11–13 and #8–8\*, while it takes three source terms in runs #8–11 and #8–8\*–11 to achieve 80% of the total. The relative score of S3 is observed relatively significant in runs #8–11 and #8–8\*–11 and proportions of S4 and S5 are observed small in all the runs.

In criterion  $C^3$ , S1 and S2 describe  $\approx 90\%$  of the total variance of the sensitivity vectors in runs #11–13, #8–11, and #8–8\*, whereas in run#8–8\*–11, same is being achieved by S1, S2, and S3. The scores corresponding to S4 and S5 are observed comparatively very small in all the runs. This indicates the presence of two releases in runs #11–13, #8–11, #8–8\*, and three releases in run#8–8\*–11.

In criterion  $C^4$ , S1 shows relatively large score in comparison to the other source terms in all the runs whereas S2 and S3 have similar scores in runs #11–13, #8–11, and #8–8\*. In all the runs (except run#8–11), S1, S2, and S3, cumulatively achieve  $\approx 80\%$  score, however, in run#8–11, it takes four source terms cumulatively to achieve 80% of the score. Thus,  $C^4$  indicates the presence of three sources in runs #11–13, #8–8\*, and #8–8\*–11 and four source terms in run#8–11.

With the scores of detection probability, it can be seen that S1 and S2 have more than 60% probability to be detected in all the runs. In run#8–8\*–11, S1, S2, and S3 have more than 60% detection probability. Comparatively, detection probability for other source terms (S4 and S5) are small ( $<50\%$ ) in all the runs. Analyzing false-alarm probability, we observed that S1 can have a small probability ( $<40\%$ ) of false alarm in runs #11–13, #8–8\*, and #8–8\*–11.

With the joint score  $\beta$  for pseudo-real data (Figure 8), we observed that only two source terms achieve the threshold score in runs #11–13, #8–11, and #8–8\*. Similarly, only three sources are higher than the threshold score in run #8–8\*–11. Thus, it shows that only two sources are present in runs #11–13, #8–11, and #8–8\* and three sources in run #8–8\*–11, which exactly matches with the true number of releases in these runs.



**Figure 9.** Estimate  $\hat{s}/\max(\hat{s})$  for noisy synthetic measurements in run #8-8\*-11. The symbol “x” and “o” denote true and retrieved source, respectively. The symbol “black triangle” denotes receptors. The retrieved source locations are shown for 50 simulations.

### 5.2.3. Source Term Estimation for $p^{\text{est}}$ Releases

The location and strength parameters are reestimated for the estimated number of releases ( $p^{\text{est}} = 2$  in runs #11-13, #8-11, #8-8\*, and  $p^{\text{est}} = 3$  in run #8-8\*-11). Overall in all the runs, source locations are retrieved within 30-m radius of the corresponding true release locations.

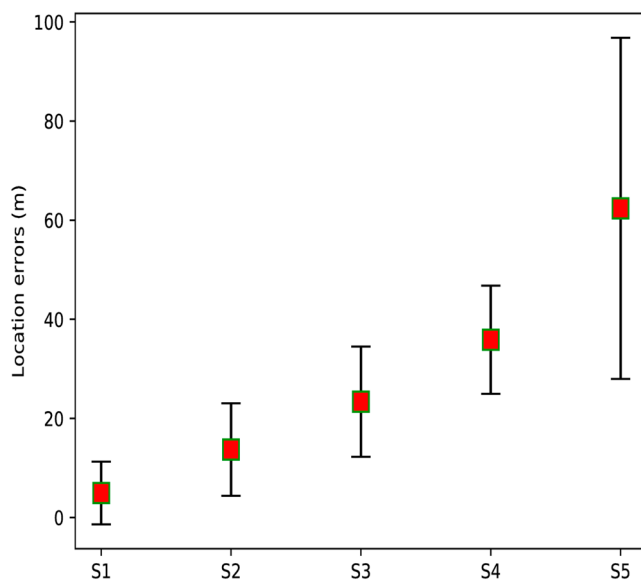
In run #11-13, first source is retrieved very close (7 m) to the  $s_2^{\text{true}}$  and the second source is retrieved at 30 m from the  $s_1^{\text{true}}$ . The strengths for both the releases are retrieved within a factor of 2 (i.e., ratio  $\frac{\text{retrieved}}{\text{true}}$  is 0.9 and 1.7) to the corresponding true values. In runs #8-11 and #8-8\*, the first source is retrieved at 15-m distance to the  $s_2^{\text{true}}$  and second source is retrieved at 25 m to the  $s_1^{\text{true}}$ . In both the runs, the release strength are overpredicted within a factor of 3. In run #8-8\*-11, two sources are retrieved at a distance of 5 and 15 m from the  $s_2^{\text{true}}$  and  $s_3^{\text{true}}$ , while the third source is retrieved at 30-m distance to the  $s_1^{\text{true}}$ . The release strength is overpredicted within a factor of 3 in this run.

### 5.3. Sensitivity Analysis: Noisy Synthetic Measurements

Noisy synthetic measurements refer to the measurements prepared by adding a Gaussian noise, in 30% proportion, to the synthetic measurements (i.e.,  $\mu^{\text{noisy}} = \mu^{\text{model}}(1 + 0.3\alpha)$ , where  $\alpha$  is a Gaussian distributed random number with mean zero and variance 1). A sensitivity analysis is required to analyze the impact of measurement errors (including model errors) on the estimation of releases and the criteria scores. Such simulations also help to analyze the uncertainty in the STE results due to model or measurement errors in the real scenarios. To analyze the results, 50 STE simulations are made using these noisy synthetic measurements for a priori source estimation (taking  $p^{\text{max}} = 5$ ).

The results are presented in the form of mean and standard deviations (sd) of the retrieved source parameters (Figure 9), location errors (Figure 10), criterion scores (Figure 11), and combined score (Figure 12).

To visualize the distribution of retrieved source locations, the retrieved locations are plotted for the 50 simulations. We observed that the sources are retrieved mostly within 60 m of the true source locations. Less than 10% of the retrieved sources are obtained too far (both downwind or upwind) from the true source locations. To analyze the retrieval, location error is computed for each retrieved source with respect to its nearest falling true source and location errors are averaged for all the simulations. The average location errors for



**Figure 10.** Errorbar plots for location errors corresponding to the S1–S5 sources in run #8–8\*–11 based on noisy synthetic simulations. The red square denotes the average location errors and bar denotes the standard deviations.

S1, S2, and S3 are observed within 25 m (Figure 10). The uncertainty (i.e., standard deviation) in the location errors is observed within 11 m for four source terms. The uncertainty is obtained large (35 m) for source S5.

With the criterion scores (Figure 11), we observed that source S1 achieves maximum score with all the criterion (Figure 11). In Figure 11,  $C^1$ ,  $C^2$ , and  $C^3$  show approximately three significant (score > 0.15) source terms while  $C^4$  shows only two significant source terms. In  $C^1$ ,  $C^2$ ,  $C^3$ , and  $C^4$ , sources S1, S2, and S3 cumulatively achieve >80% score.

In  $C^1$ , coefficient of variation (cv) ( $cv = \frac{sd}{mean}$ ) shows a variation of  $\approx 25$ –33% for sources S1, S2, and S3, while cv is observed relatively larger (45% and 80%) for S4 and S5 (Figure 11). In  $C^2$  and  $C^3$ , cv is observed within 20% for sources S1, S2, S3, and S4 and relatively large (>50%) for S5. The CV is observed much higher in  $C^4$ , as 30% for source S1 and 60–100% for S2, S3, S4, and S5. The detection probability is observed >50% for sources S1, S2, and S3 while for other sources, detection probability is low (<30%). False-alarm probability is also observed low (<20%) for all the sources. The cv is observed within 20% for sources S1, S2, and S3 whereas uncertainty is large for S4 and S5 (30% and 140%, respectively). The uncertainty in false-alarm probability is also observed large (cv between 70% and 300%) for all the sources but even with uncertainty, the probability of false alarm is always observed within 40%.

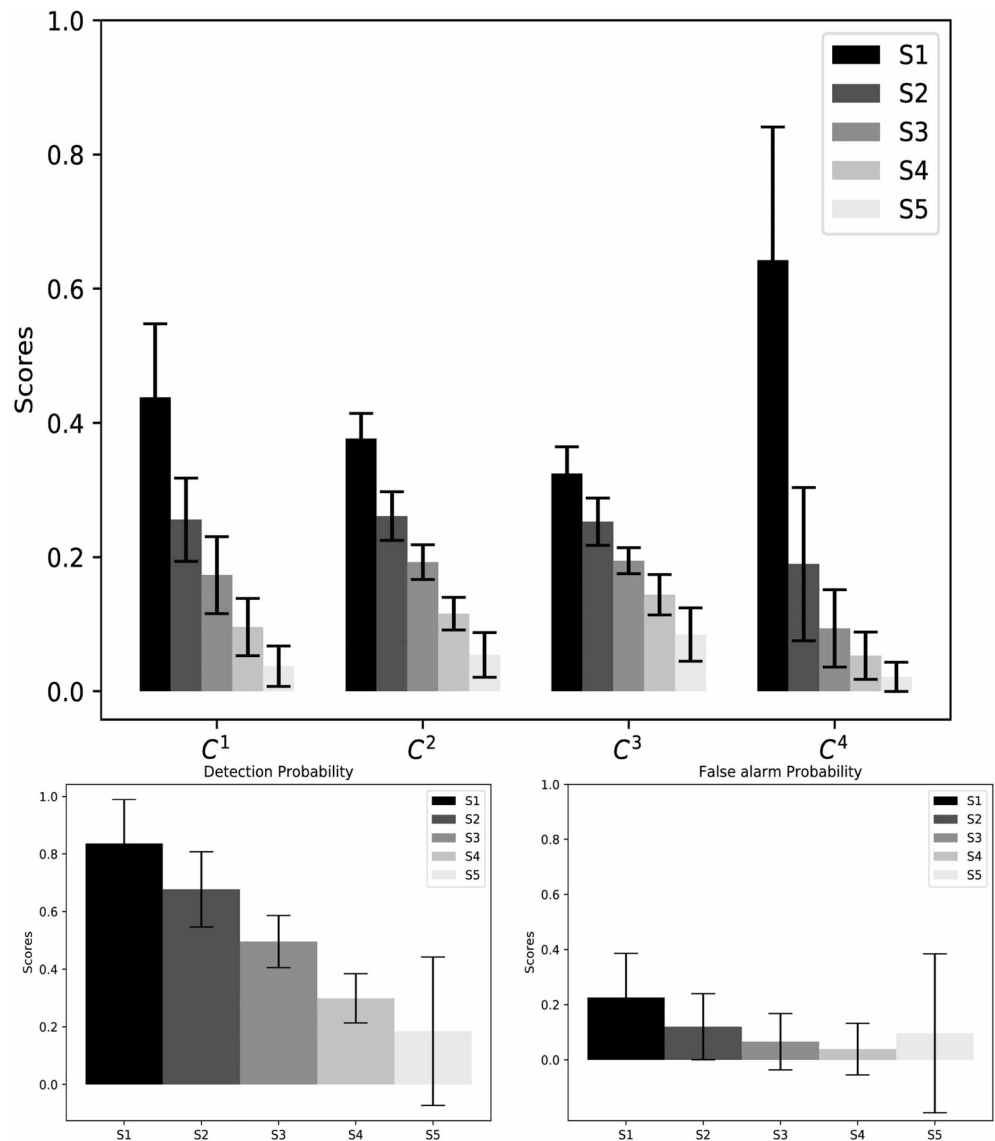
The average combined score  $\beta$  (Figure 12) shows that the sources S1, S2, and S3 achieve more than threshold score. The standard deviation in  $\beta$  is observed within 0.30 (i.e., 30% of the threshold score). The standard deviation is observed relatively large in magnitude for source S2 and comparatively small for source S5. However, in terms of the coefficient of variation ( $cv = sd/mean$ ), we could see that  $\beta$  varies within 10% for source S1,  $\approx 20\%$  for S2 and S3, and 25–30% for S4 and S5. The cv estimates for  $\beta$  shows robustness of the proposed  $\beta$  score in precisely identifying the number of releases. Although, in the presence of large model errors, it is quite possible that the methodology may miss the presence of a source term. For instance, with the standard deviation estimates, it can be analyzed that the uncertainties may force the score below threshold for source S3. This is an indication of model error's impact.

#### 5.4. Causes for Retrieval Errors

Overall, the proposed methodology works well in estimating the number of releases as well as corresponding release parameters. The deviations between true and retrieved source locations are mainly due to the meteorological limitations (use of constant wind speed, wind direction) and limited form of parameterizations. These can be attributed as model errors. An overprediction in the retrieval of emission strength is related to the underprediction behavior of model. However, the retrieval of the strength may be affected depending on the location of the retrieved source upwind or downwind of the network. The estimate  $\hat{s}$  helps in discriminating well-resolved or poorly resolved source informations. Thus, large errors are expected in the retrieval if source locations (both true and retrieved) fall to the poorly resolved regions. The number of measurements plays an important role in improving resolution of the source information. However, in this study, the number of measurements are very limited (lesser than 18) in all the runs. Another limitation is artificiality of the pseudo runs. In preparing pseudo trials, only common receptors corresponding to two real runs are combined which results into the relatively lesser number of measurements. In addition, averaging of the winds from two real runs is not fully consistent with the addition of corresponding concentration measurements.

## 6. Scope and Limitations

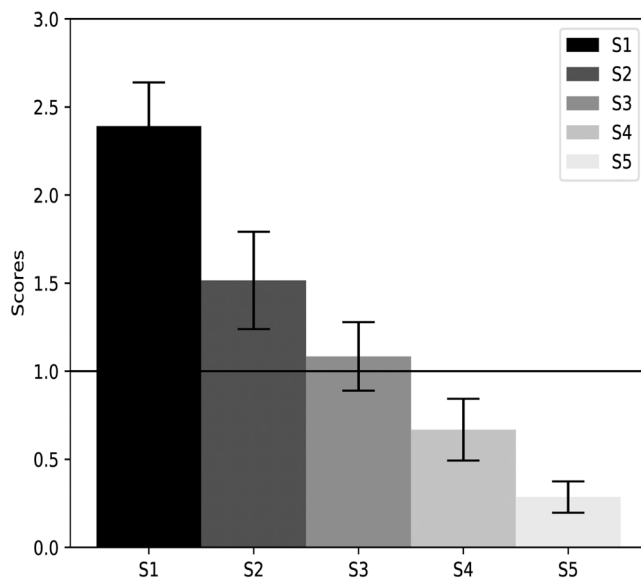
This section describes the key points associated with the methodology from both the technical and application points of view. The formulation of source-receptor relationship (section 2.1) is based on the assumption that the measurements are linear mixture of the concentrations that resulted from each source. However, this may not be true in real scenarios especially in the presence of reactive tracers, sink terms, or meteorological variability. The proposed methodology is based on the use of an adjoint dispersion model and minimization of a functional which is overdetermined in nature. The derivation of an exact adjoint



**Figure 11.** Cumulative bar chart for the mean scores resulted from each criterion using synthetic noisy measurements in run#8–8\*–11. The standard error bars are also shown with respect to each score. Top panel shows cumulative scores while the Detection and False alarm probability are shown in the bottom panel.

model is a difficult task for complex dispersion models accounting complex interactions between tracer and atmospheric turbulence. In this study, the formulated inverse problem was overdetermined (as  $m > 3p$ ) nevertheless, the number of measurements are still very limited and more number of measurements are required for better estimation. The numerical discretization of the computational domain also plays an important role in view of resolution of the source estimation and computational implementation of the algorithm. The cell size should be chosen such that the releases or receptors are sufficiently separated in the discretized domain. This is important to avoid the correlations between the source terms and among the receptors. Moreover, it is difficult to capture refined aspects of complicated releases or more sensitive localizations smaller than the discretized cell.

Computationally, the implementation of the proposed methodology is simple but time consuming and expensive. This can be improved by the use of parallel computation or hybrid search algorithms like simulated annealing, genetic algorithm, particle swarm optimization, etc. In source term estimation, the metaheuristic algorithms are utilized for the retrieval of single point source. This includes important contributions from Allen et al. (2007), Wang et al. (2018) for applying genetic algorithm, Thomson et al. (2007) for simulated annealing, Ma et al. (2017) for particle swarm optimization, etc. The implementation and



**Figure 12.** Score  $\beta$  for synthetic noisy measurements in run#8-8\*-11.

application of these algorithms were limited to the estimation of few parameters (i.e., mainly for the retrieval of a point source). The application of hybrid search algorithms for estimation of multiple release parameters is not straightforward. Nevertheless, estimating the global minimum of the cost function and simultaneously estimating all the release parameters for several number of sources will still be a challenge from the computational point of view. The accuracy and computational time are subjected to the number of releases, release parameters, number of cells in the discretized domain. The use of a priori source information via estimate  $\hat{s}$  helps to constrain the additional degree of freedom in the discretized model space. The source estimation methodology has two advantages: (i) it does not require any initial guesses for the unknown release parameters and (ii) it utilizes the resolution of the source information  $\hat{s}$  distributed in the domain, based on the given monitoring design and number of measurements. This helps to discriminate between the well-resolved or poorly resolved source regions and further, reduces the computational time required in the search procedure. The accuracy and uncertainties in the source retrieval are subjected to the design of the monitoring network, number of measurements, distance between the releases, model errors, and error statistics.

The present evaluation shows that the technique is able to characterize the releases well when they are sufficiently apart from each other,

and enough measurements are available to provide a well-resolved and distinct source regions. The technique provides a source retrieval close or upwind of the true releases when source information is distinctly resolved. The releases are retrieved relatively closer to each other or downwind of the true releases when source information is not distinctly resolved. The major deviations are observed due to lack of measurements and model representativeness errors. Thus, the accuracy of the source retrieval can be enhanced by improving the model characteristics toward the observations.

The estimation process of the release parameters is sensitive with respect to the error-statistics of measurement and model errors. In the present study, the observation error-covariance matrix is assumed identity matrix. The estimates of release parameters may vary for a different choice of observation error covariance matrix. With the present setup, a posteriori uncertainty of the source terms cannot be addressed directly. However, they can be approximated by using the Hessian of the cost function. Note that the posterior uncertainties in the release parameters will depend on the assumed prior statistical information about the measurement errors.

The present study utilizes constant wind speed and constant wind direction for the tracer dispersion. It is expected that the concentration measurements with variable meteorological scenarios may be more informative than those considered in the constant meteorological situations. In future, this will be further explored with availability of the data sets.

## 7. Conclusion

An inversion methodology is proposed here to estimate the number of point releases as well as corresponding release parameters in the dispersion events involving several simultaneous releases. The study proposes new criterion which analyzes release strengths, collinearity among the sensitivity vectors, spread of measurements described by the sensitivity vectors and spread of sensitivity vectors corresponding to the a priori retrieved releases. In addition, the criterion leads to the detection and false-alarm probability associated with the retrieved releases.

To evaluate the proposed methodology, four pseudo-runs, prepared from IIT Delhi diffusion experiment, are utilized. The methodology is evaluated with both synthetic and pseudo-real data. With synthetic data, we were able to estimate the true number of releases, their locations and strengths exactly as originally prescribed. With pseudo-real data, number of releases are estimated exactly as prescribed in all the four runs. The release locations are reconstructed within a minimum error of 7 m and maximum error of 30 m from their true locations. The strength is also retrieved within a factor of 3 in all the runs. The incurred

errors in retrieval with the pseudo-real data are mainly due to artificiality of the data set and model errors. The sensitivity analysis highlights robustness of the inversion methodology in the presence of model errors and impact of model errors on the criteria scores.

The proposed inversion technique is general for any number of sources and can perform well for elevated as well as instantaneous releases. This requires a further evaluation with existing or upcoming data sets for multiple contaminant releases at varying scales from local to global.

## Appendix A: Covariance and Correlation Matrix

The matrix  $\hat{\mathbf{H}} \in \mathbb{R}^{m \times p}$  containing the sensitivity vectors corresponding to the retrieved release locations is defined as

$$\hat{\mathbf{H}} = \begin{bmatrix} a_1(\hat{\mathbf{x}}_1) & a_1(\hat{\mathbf{x}}_2) & a_1(\hat{\mathbf{x}}_3) & \dots & a_1(\hat{\mathbf{x}}_p) \\ a_2(\hat{\mathbf{x}}_1) & a_2(\hat{\mathbf{x}}_2) & a_2(\hat{\mathbf{x}}_3) & \dots & a_2(\hat{\mathbf{x}}_p) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_m(\hat{\mathbf{x}}_1) & a_m(\hat{\mathbf{x}}_2) & a_m(\hat{\mathbf{x}}_3) & \dots & a_m(\hat{\mathbf{x}}_p) \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1p} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \dots & a_{mp} \end{bmatrix} \quad (\text{A1})$$

where  $\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \dots, \hat{\mathbf{x}}_p$  are the estimated source locations and  $a_i(\hat{\mathbf{x}}_j)$  can be interpreted as sensitivity element corresponding to the  $i$ th measurement at source location  $\hat{\mathbf{x}}_j$ . The column mean of the sensitivity elements is defined as

$$\bar{a}_j = \frac{1}{m} \sum_{i=1}^m a_{ij} = m^{-1} \gamma_m^T \mathbf{a}_j \quad (\text{A2})$$

where  $\mathbf{a}_j$  is the  $j$ th column of  $\hat{\mathbf{H}}$  and  $\gamma_p \in \mathbb{R}^{p \times 1}$  is a column vector containing identity elements (i.e., 1). The row mean of the sensitivity elements is defined as

$$\bar{a}_i = \frac{1}{p} \sum_{j=1}^p a_{ij} = p^{-1} \mathbf{a}_i' \gamma_p \quad (\text{A3})$$

where  $\mathbf{a}_i'$  is  $i$ th row of  $\hat{\mathbf{H}}$ . The variance of  $j$ th variable is defined as

$$s_j^2 = \frac{1}{m} \sum_{i=1}^m (a_{ij} - \bar{a}_j)^2 \quad (\text{A4})$$

Now the covariance matrix  $\mathbf{Q} = \text{cov}(\hat{\mathbf{H}})$  is defined as

$$\mathbf{Q} = \frac{1}{m} \hat{\mathbf{H}}_c^T \hat{\mathbf{H}}_c \quad (\text{A5})$$

where  $\hat{\mathbf{H}}_c = \hat{\mathbf{H}} - \gamma_m \bar{\mathbf{a}}^T = \mathbf{C} \hat{\mathbf{H}}$  is the centered matrix,  $\bar{\mathbf{a}}^T = [\bar{a}_1, \bar{a}_2, \dots, \bar{a}_p]$ , and  $\mathbf{C} = \mathbf{I}_m - m^{-1} \gamma_m \gamma_m^T$  is the centering matrix.  $\mathbf{I}_m \in \mathbb{R}^{m \times m}$  is identity matrix.

Now, the correlation matrix  $\mathbf{R}$  is defined as

$$\mathbf{R} = \frac{1}{m} \hat{\mathbf{H}}_s^T \hat{\mathbf{H}}_s \quad (\text{A6})$$

where  $\hat{\mathbf{H}}_s = \mathbf{C} \hat{\mathbf{H}} \mathbf{D}^{-1}$  and  $\mathbf{D} = \text{diag}(s_1, s_2, \dots, s_p)$  is a diagonal scaling matrix.

## Acknowledgments

The work is partially supported under the grant of JC Bose Fellowship to Maithili Sharan from Science and Engineering Research Board, Government of India. The IIT diffusion data are available in Singh et al. (1991).

## References

- Albo, S. E., Oluwale, O. O., & Miale-Lye, R. C. (2011). The aerodyne inverse modeling system (aims): Source estimation applied to the 07 experiment and to simulated mobile sensor data. *Atmospheric Environment*, 45(33), 6085–6092. <https://doi.org/10.1016/j.atmosenv.2011.07.033>
- Allen, C. T., Haupt, S. E., & Young, G. S. (2007). Source characterization with a genetic algorithm-coupled dispersion-backward model incorporating SCIPUFF. *Journal of Applied Meteorology and Climatology*, 46, 273–287. <https://doi.org/10.1175/JAM2459.1>
- Annunzio, A. J., Young, G. S., & Haupt, S. E. (2012a). A multi-entity field approximation to determine the source location of multiple atmospheric contaminant releases. *Atmospheric Environment*, 62, 593–604. <https://doi.org/10.1016/j.atmosenv.2012.08.032>



- Annunzio, A. J., Young, G. S., & Haupt, S. E. (2012b). Utilizing state estimation for the source term estimation of a contaminant. *Atmospheric Environment*, 46, 580–589. <https://doi.org/10.1016/j.atmosenv.2011.04.080>
- Bocquet, M. (2005). Reconstruction of an atmospheric tracer source using the principle of maximum entropy-I: Theory. *Quarterly Journal of Royal Meteorological Society*, 131, 2191–2208. <https://doi.org/10.1256/qj.04.67>
- Cantelli, A., D'Orta, F., Cattini, A., Sebastianelli, F., & Cedola, L. (2015). Application of genetic algorithm for the simultaneous identification of atmospheric pollution sources. *Atmospheric Environment*, 115, 36–46. <https://doi.org/10.1016/j.atmosenv.2015.05.030>
- Chai, T., Stein, A., & Ngan, F. (2018). Weak-constraint inverse modeling using HYSPLIT-4 lagrangian dispersion model and cross-appalachian tracer experiment (CAPTEX) observations: Effect of including model uncertainties on source term estimation. *Geoscientific Model Development*, 11(12), 5135–5148. <https://doi.org/10.5194/gmd-11-5135-2018>
- Chang, C. I. (2013). *Hyperspectral imaging: Techniques for spectral detection and classification*. US: Springer. <https://books.google.fr/books?id=ZFLSBwAAQBAJ>
- Chang, C.-I., & Du, Q. (2004). Estimation of number of spectrally distinct signal sources in hyperspectral imagery. *IEEE Transactions on geoscience and remote sensing*, 42(3), 608–619.
- Hutchinson, M., Oh, H., & Chen, W.-H. (2017). A review of source term estimation methods for atmospheric dispersion events using static or mobile sensors. *Information Fusion*, 36, 130–148. <https://doi.org/10.1016/j.inffus.2016.11.010>
- Issartel, J.-P., Sharan, M., & Modani, M. (2007). An inversion technique to retrieve the source of a tracer with an application to synthetic satellite measurements. *Proceedings of Royal Society A*, 463, 2863–2886. <https://doi.org/10.1098/rspa.2007.1877>
- Kumar, P., Feiz, A.-A., Singh, S. K., Ngae, P., & Turbelin, G. (2015). Reconstruction of an atmospheric tracer source in an urban-like environment. *Journal of Geophysical Research: Atmospheres*, 120, 12,589–12,604. <https://doi.org/10.1002/2015JD024110>
- Ma, D., Tan, W., Wang, Q., Zhang, Z., Gao, J., Wang, X., & Xia, F. (2018). Location of contaminant emission source in atmosphere based on optimal correlated matching of concentration distribution. *Process Safety and Environmental Protection*, 117, 498–510. <https://doi.org/10.1016/j.psep.2018.05.028>
- Ma, D., Tan, W., Zhang, Z., & Hu, J. (2017). Parameter identification for continuous point emission source based on tikhonov regularization method coupled with particle swarm optimization algorithm. *Journal of Hazardous Materials*, 325, 239–250. <https://doi.org/10.1016/j.jhazmat.2016.11.071>
- Neyman, J., Pearson, E. S., & Pearson, K. (1933). IX. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706), 289–337. <https://doi.org/10.1098/rsta.1933.0009>
- Pudykiewicz, J. A. (1998). Application of adjoint tracer transport equations for evaluating source parameters. *Atmospheric Environment*, 32, 3039–3050. [https://doi.org/10.1016/S1352-2310\(97\)00480-9](https://doi.org/10.1016/S1352-2310(97)00480-9)
- Rao, K. S. (2007). Source estimation methods for atmospheric dispersion. *Atmospheric Environment*, 41(33), 6964–6973. <https://doi.org/10.1016/j.atmosenv.2007.04.064>
- Royall, R. (2017). *Statistical evidence: A likelihood paradigm*. New York: CRC Press. <https://books.google.fr/books?id=wIzHDwAAQBAJ>
- Sharan, M., Issartel, J.-P., & Singh, S. K. (2012). A point-source reconstruction from concentration measurements in low-wind stable conditions. *Quarterly Journal of the Royal Meteorological Society*, 138, 1884–1894. <https://doi.org/10.1002/qj.1921>
- Sharan, M., Issartel, J.-P., Singh, S. K., & Kumar, P. (2009). An inversion technique for the retrieval of single-point emissions from atmospheric concentration measurements. *Proceedings of Royal Society A*, 465, 2069–2088. <https://doi.org/10.1098/rspa.2008.0402>
- Sharan, M., Singh, S. K., & Issartel, J. P. (2012). Least square data assimilation for identification of the point source emissions. *Pure and Applied Geophysics*, 169, 483–497. <https://doi.org/10.1007/s00024-011-0382-3>
- Sharan, M., Yadav, A. K., Singh, M. P., Agarwal, P., & Nigam, S. (1996). A mathematical model for dispersion of air pollutants in low wind conditions. *Atmospheric Environment*, 30, 1209–1220. [https://doi.org/10.1016/1352-2310\(95\)00442-4](https://doi.org/10.1016/1352-2310(95)00442-4)
- Singh, M. P., Agarwal, P., Nigam, S., & Gulati, A. (1991). Tracer experiments—A report (Technical Report). New Delhi, Delhi, India.
- Singh, S. K., & Rani, R. (2014). A least-squares inversion technique for identification of a point release: Application to fusion field trials 2007. *Atmospheric Environment*, 92, 104–117. <https://doi.org/10.1016/j.atmosenv.2014.04.012>
- Singh, S. K., & Rani, R. (2015). Assimilation of concentration measurements for retrieving multiple point releases in atmosphere: A least-squares approach to inverse modelling. *Atmospheric Environment*, 119, 402–414. <https://doi.org/10.1016/j.atmosenv.2015.08.063>
- Singh, S. K., Sharan, M., & Issartel, J.-P. (2013). Inverse modelling for identification of multiple-point releases from atmospheric concentration measurements. *Boundary-Layer Meteorology*, 146, 277–295. <https://doi.org/10.1007/s10546-012-9765-y>
- Singh, S. K., Sharan, M., & Issartel, J.-P. (2015). Inverse modelling methods for identifying unknown releases in emergency scenarios: An overview. *International Journal of Environment and Pollution*, 57, 68–91. <https://doi.org/10.1504/IJEP.2015.072121>
- Singh, S. K., Turbelin, G., Issartel, J.-P., Kumar, P., & Feiz, A. A. (2015). Reconstruction of an atmospheric tracer source in fusion field trials: Analyzing resolution features. *Journal of Geophysical Research: Atmospheres*, 120, 6192–6206. <https://doi.org/10.1002/2015JD023099>
- Storwald, D. P. (2007). Detailed test plan for the fusing sensor information from observing networks (fusion) field trial (FFT-07) (Document No. WDTC-TP-07-078). Dugway Proving Ground, Utah, USA: Meteorology Division, West Desert Test Center, U.S. Army Dugway Proving Ground WDTC.
- Thomson, L. C., Hirst, B., Gibson, G., Gillespie, S., Jonathan, P., Skeldon, K. D., & Padgett, M. J. (2007). An improved algorithm for locating a gas source using inverse methods. *Atmospheric Environment*, 41(6), 1128–1134. <https://doi.org/10.1016/j.atmosenv.2006.10.003>
- Wade, D., & Senocak, I. (2013). Stochastic reconstruction of multiple source atmospheric contaminant dispersion events. *Atmospheric Environment*, 74, 45–51. <https://doi.org/10.1016/j.atmosenv.2013.02.051>
- Wang, Y., Huang, H., Huang, L., & Zhang, X. (2018). Source term estimation of hazardous material releases using hybrid genetic algorithm with composite cost functions. *Engineering Applications of Artificial Intelligence*, 75, 102–113. <https://doi.org/10.1016/j.engappai.2018.08.005>
- Yee, E. (2007). Bayesian probabilistic approach for inverse source determination from limited and noisy chemical or biological sensor concentration measurements. *Proceedings SPIE* (Vol. 6554, pp. 65540W). Orlando, FL, USA: Chemical and Biological Sensing VIII. <https://doi.org/10.1117/12.721630>
- Yee, E. (2008). Theory for reconstruction of an unknown number of contaminant sources using probabilistic inference. *Boundary-Layer Meteorology*, 127, 359–394. <https://doi.org/10.1007/s10546-008-9270-5>
- Yee, E. (2012a). Inverse dispersion for an unknown number of sources: Model selection and uncertainty analysis. *International Scholarly Research Notices: Applied Mathematics*, 2012, 465320.
- Yee, E. (2012b). Probability theory as logic: Data assimilation for multiple source reconstruction. *Pure and Applied Geophysics*, 169(3), 499–517.

**Erratum**

In the originally published version of this article, several math symbols were incorrectly typeset. The symbols have since been updated and this version may be considered the authoritative version of record.