H13-214

VALIDATION OF A BAYESIAN INFERENTIAL FRAMEWORK FOR MULTIPLE SOURCE RECONSTRUCTION USING FFT-07 DATA

Eugene Yee

Defence R&D Canada – Suffield, Medicine Hat, Alberta, Canada

Abstract: This paper applies a Bayesian probabilistic inferential framework to the difficult problem of the estimation of the parameters of an *a priori* unknown number of sources, using a limited number of noisy concentration data obtained from an array (or, network) of sensors. To this purpose, Bayesian probability theory is used to formulate the full joint posterior probability density function for the number of (unknown) sources and for the parameters (e.g., location, emission rate, source on and off 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. The method is validated using a real dispersion experiment involving a release of a propylene tracer from four discrete sources. This experiment was conducted under a multinational cooperative **FU**sing **S**ensor **I**nformation from **O**bserving **N**etworks (FUSION) Field Trial 2007 (FFT-07).

Key words: *Bayesian inference, inverse dispersion, Markov chain Monte Carlo, sensor-model data fusion, sensor arrays.*

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, toxic industrial materials] released into the turbulent atmosphere has generated interest in utilizing this information for the reconstruction of the contaminant sources 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 (or, array) of CBR sensors. The 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 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. Further motivation is provided by a network of 40 radiological detectors that has been set up as a verification tool for the Comprehensive Test Ban Treaty (CTBT) in order to provide world wide monitoring of radioactive noble gases that can be used potentially for source localization and characterization of a clandestine nuclear test.

To address the problem of source reconstruction, a probabilistic approach using a Bayesian inferential scheme has been developed, refined and generalized over the past several years by the author and colleagues: (1) application of the methodology to complex environments (inverse dispersion in built-up environments) has been developed by Yee E. (2006) and Keats A. *et al* (2007a); (2) generalization of the methodology to deal with a non-conservative scalar has been described by Keats A. *et al* (2007b); and, (3) application of the methodology to source reconstruction for long-range dispersion on continental scales has been demonstrated successfully in Yee E. *et al* (2008). Yee E. (2007) generalized the methodology to the reconstruction of multiple sources when the number of sources was **known** *a priori*. Finally, Yee E. (2008) 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*.

The objective 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 E. (2008) for multiple source reconstruction for the case when the number of sources is unknown *a priori*. Furthermore, the computational framework used by Yee E. (2008) for sampling from the posterior distribution of the source parameters is significantly improved in the current paper.

BAYESIAN INFERENCE FOR SOURCE RECONSTRUCTION

In this paper, we apply Bayesian probability theory to address the problem of source reconstruction. Within the context of this problem, Bayes' theorem yields the following result:

$$p(\Theta|\mathbf{D},I) = \frac{p(\Theta|I)p(\mathbf{D}|\Theta,I)}{p(\mathbf{D}|I)}, \tag{1}$$

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 equation (1) have the following interpretation. Firstly, $p(\Theta|I)$ is the prior probability density function (PDF) for a proposition (or, hypothesis) Θ about the source, predicated on the contextual information specified by I, with "|" denoting "conditional upon". Secondly, $p(\mathbf{D}|\Theta,I)$ is the likelihood function and is the probability that we would have observed the concentration data $\mathbf{D}$, if Θ were 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 $\mathbf{D}$. Because $p(\mathbf{D}|I)$ is simply a normalization

constant, the problem for the specification of the posterior PDF of Θ reduces to the assignment of $p(\Theta|I)$ (prior distribution) and $p(\mathbf{D}|\Theta,I)$ (likelihood function).

Before we can proceed further, we need to be explicit about Θ. In this paper, we focus on a source distribution *S* associated with N_s transient point sources with the *k*-th source located at 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 emitting at a constant release rate Q_k (k = 1, 2, …, N_s). The source distribution has the following explicit form:

$$S(\mathbf{x},t) = \sum_{k=1}^{N_s} Q_k \delta(\mathbf{x}-\mathbf{x}_{s,k})[H(t-T_b^k) - H(t-T_e^k)], \tag{2}$$

where $H(s)$ and $\delta(s)$ denote the Heaviside step and Dirac delta functions, respectively. Now, we can assemble the parameters for this particular source distribution into the following source parameter vector:

$$\Theta \equiv \left(N_s, \mathbf{x}_{s,1}, T_b^1, T_e^1, Q_1, \ldots, \mathbf{x}_{s,N_s}, T_b^{N_s}, T_e^{N_s}, Q_{N_s}\right) \in \mathbb{R}^{6N_s+1}. \tag{3}$$

With this background, the problem of source reconstruction reduces to the following: estimate Θ given the concentration data $\mathbf{D} \equiv (d_1, d_2, \ldots, d_N)$ where *N* is the number of concentration data.

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 **D**, the latter of which modulates our prior belief about Θ through the likelihood function $p(\mathbf{D}|\Theta,I)$. In this paper, the posterior PDF is specified as follows (to within a normalization constant):

$$\begin{aligned} p(\Theta|\mathbf{D},I) &\equiv p(N_s,\theta_{N_s}|\mathbf{D},I) \\ &\propto \frac{1}{\prod_{J=1}^{N}\sqrt{2\pi}\sigma_J}\exp\left(-\frac{1}{2}\sum_{J=1}^{N}\left(\frac{d_J - C_J(\Theta)}{\sigma_J}\right)^2\right) \\ &\quad \times \frac{(N_{s,\max}-N_{s,\min})!}{(N_s-N_{s,,\min})!(N_{s,\max}-N_s)!}\, p^{*(N_s-N_{s,\min})}(1-p^*)^{N_{s,\max}-N_s} \\ &\quad \times \prod_{k=1}^{N_s}\left[(1-\gamma)\delta(Q_k) + \gamma \mathbf{I}_{(0,Q_{\max})}(Q_k)/Q_{\max}\right] \\ &\quad \times \mathbf{I}_{\mathcal{D}}(\mathbf{x}_{s,k})\mathbf{I}_{(t_0,T_{\max})}(T_b^k)\frac{\mathbf{I}_{(T_b^k,T_{\max})}(T_e^k)}{(T_{\max}-T_b^k)}. \end{aligned} \tag{4}$$

Here, **I** denotes the indicator function, C_J is the *J*-th model concentration (and is a function of the source distribution encoded in the parameter vector Θ), and σ_J is the noise standard deviation corresponding to the *J*-th datum (and incorporates the effects arising from model error, measurement noise, and stochastic uncertainty associated with either d_J or C_J).

In equation (4), the prior on the number of sources N_s is chosen to be a binomial distribution with parameter p^* (binomial rate) where $p^* \in [0, 1]$ and with a domain of definition between $N_{s,\min}$ and $N_{s,\max}$ (minimum and maximum number of sources, respectively). The prior on the emission rate is chosen to be a Bernoulli-uniform mixture, with γ defined as the probability that the source is turned on and $Q_{\max}$ defined as the *a priori* upper bound on the expected emission rate. The prior on the source location is chosen to be uniform (flat) over some spatial region *D* that is assumed to contain the contaminant sources. 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_0 is a lower bound on the time at which the source was turned on and $T_{\max}$ is an upper bound on the time at which the source was turned on or off. Note that the prior for the source off time explicitly encodes the fact that the time the *k*-th source is turned off must occur after it has been turned on. Finally, a Gaussian form has been used for the likelihood function in equation (4).

COMPUTATIONAL FRAMEWORK

This section describes briefly the computational procedures that were used for extracting the source parameter estimates required for event reconstruction. The reader is referred to Yee E. *et al* (2008) and Yee E. (2008) for a more complete description of the computational methodology. 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.

Fast computation of the source-receptor relationship

The likelihood function is not a closed-form expression and its evaluation is computationally expensive owing to the fact that C_J (J = 1, 2, … , N) needs to be determined for a given source distribution Θ. 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 a given source distribution Θ) is required to facilitate the rapid sampling from the posterior distribution. To this purpose, Keats A. *et al* (2007a) and Yee E. *et al* (2008) described a computationally efficient methodology for determination of the source-receptor relationship using an adjoint representation for this relationship.

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 Θ. The posterior quantities of interest are expectation values of $p(\Theta|\mathbf{D},I)$, which necessarily involves an integration in a potentially high-dimensional hypothesis space. One method for overcoming this "curse of dimensionality" is given by the application of Markov chain Monte Carlo (MCMC) algorithms for posterior sampling. To this purpose, Yee E. (2008) described the formulation of a reversible-jump MCMC (RJMCMC) algorithm applied with parallel tempering for generating samples from the posterior distribution given in equation (4).

The objective of MCMC sampling is to construct an auxiliary Markov chain whose stationary (or, invariant) distribution is the posterior distribution of Θ. To summarize, the Markov chain consists of a sequence of states $\Theta^{(t)}$ (t = 0, 1, 2, ...) resulting from individual updates consisting of three basic moves: (1) dimension-changing moves M_0 involving the creation of a source atom at a random location, or annihilation of an existing source atom; (2) fixed-dimension moves M_1 involving updates of the emission rates of the source atoms using Gibbs sampling; and, (3) fixed-dimension moves M_2 involving updates of the location, source on and off times of the source atoms using Metropolis-Hastings (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 time t using the following procedure:

1. Specify the values ($N_{s,\min}$, $N_{s,\max}$, $Q_{\max}$, $T_{\max}$, t_0, γ, 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:

$$\Theta^{(t-1)} \xrightarrow{M_0} \Theta_{\star} \xrightarrow{M_1} \Theta_{\star\star} \xrightarrow{M_2} \Theta^{(t)}, \tag{5}$$

where Θ_* and Θ_{**} 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 E. (2008) implemented a form of parallel tempering based on a Metropolis-coupled MCMC algorithm. 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. 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) source distributions (or, source molecules) that have been randomly drawn from the following modified distribution:

$$p_\lambda(\Theta|\mathbf{D},I) \propto p(\Theta|I)p^\lambda(\mathbf{D}|\Theta,I). \tag{6}$$

The samples will be labelled $\Theta_k(\lambda)$, with $\lambda \in [0,1]$ (k = 1, 2, ... , N_{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 parameter λ as an inverse temperature T (so, $\lambda = 1/T$), with $\lambda \in [0,1]$ implying $T \in [1, \infty]$. The posterior distribution corresponds to the temperature T = 1, whereas the modified $p_\lambda(\Theta|\mathbf{D},I)$ 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 λ = 0 (infinite temperature), we randomly draw N_{mem} source molecules $\Theta_k(0)$ (k = 1, 2, ... , N_{mem}) from $p_0(\Theta|\mathbf{D},I)$ (prior distribution). Given an ensemble of N_{mem} source molecules $\Theta_k(\lambda)$ that has achieved equilibrium (at temperature $T = 1/\lambda$) with respect to the modified posterior $p_\lambda(\Theta|\mathbf{D},I)$, an ensemble of N_{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 D. and H. F. Lopes, 2000) applied to $\Theta_k(\lambda)$ (k = 1, 2, ... , N_{mem}). An annealing schedule for $\lambda \in [0,1]$ is required for the simulated annealing. In this paper, 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 (λ = 0, or infinite temperature) at one end of the annealing schedule to the posterior distribution (λ = 1, or unit temperature) at the other end of the schedule. When λ = 1, the annealing phase is complete and probabilistic exploration of the hypothesis space proceeds (for each N_{mem} source molecules in the ensemble) in accordance to the scheme summarized in equation (5). The annealing phase of the scheme, corresponding to $\lambda \in [0,1)$, is associated with the burn-in phase of the algorithm. When λ = 1, the MCMC algorithm has reached an equilibrium, at which point the probabilistic exploration corresponding to the sampling from the posterior distribution 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, and highest posterior distribution (HPD) intervals).

EXAMPLE: APPLICATION OF METHODOLOGY

In this section, we apply the source reconstruction methodology to a real dispersion data set; namely, the **FU**sing **S**ensor **I**nformation from **O**bserving **N**etworks (FUSION) Field Trial 2007 (FFT-07). The experiments in FFT-07 were carried out in September 2007 at Tower Grid on US Army Dugway Proving Ground. In these experiments, the tracer gas used was propylene (C_3H_6). The concentration detectors used were fast-response digital photo-ionization (dPID) detectors. These

detectors give a frequency response of 50 Hz with a sensitivity of about 0.025 parts per million (ppm) by volume of propylene. In FFT-07, a network of concentration detectors was used, consisting of a total of 100 dPIDs 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 overall (alongwind) length and (crosswind) width of the detector array were 450 m and 475 m, respectively. Three-dimensional (3-D) sonic anemometers were arranged on three 32-m lattice towers along a transect parallel to the (alongwind) length dimension and midline of the concentration detector array.

This example involves four continuously emitting sources. We used 62 detectors in the array for the 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. In this example, the mean wind direction was normally incident to the detector array. The proposed stochastic sampling algorithm was applied with $N_{s,\min} = 1$, $N_{s,\max} = 8$, $p^* = 1/7$, $\gamma = 0.25$, $Q_{\max} = 100$ g s^{-1}, and prior bounds for location of any source was contained in the domain $D \equiv [0,100] \times [0,500]$ m (constraining the x_s and y_s locations of the sources). An ensemble of $N_{\text{mem}} = 50$ members of source distribution models Θ were drawn from the prior distribution and used for the simulated annealing phase of the 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 50000 samples of source distribution models drawn from the posterior distribution $p(\Theta|\mathbf{D},I)$.

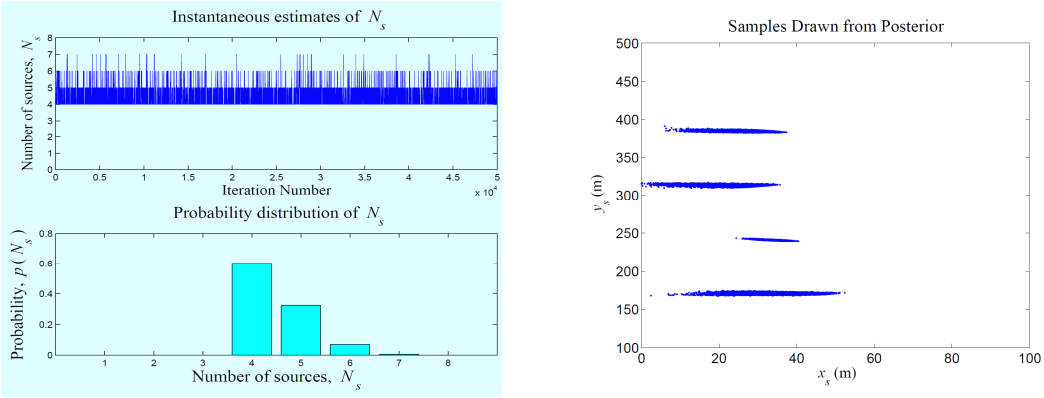

Figure 1. Left panel: Trace plot (top) of the number of discrete sources N_s in the source distribution model samples drawn from $p(\Theta|\mathbf{D},I)$ during the probabilistic exploration phase of the stochastic sampling algorithm, and the posterior distribution (bottom) for the number of sources, $p(N_s) \equiv p(N_s|\mathbf{D},I)$. Right panel: Density plot consisting of samples of source distribution models obtained for $N_s = 4$ (most probable value for number of sources) projected onto the (x_s, y_s) subspace (x_s and y_s are in the alongwind and crosswind directions, respectively).

Figure 1 (left panel, top) shows a trace plot for the number of discrete sources in a source distribution model sample against the sample (or, iteration) number. From this plot it is seen that the samples of source distribution models drawn from $p(\Theta|\mathbf{D},I)$ generally mix well over N_s. Note that annihilation moves for models from $N_s = 4$ to 3 do not occur. However, dimension-changing moves from $N_s = 4$ to 5 (and, vice-versa), as well as higher-order transitions such as from $N_s = 6$ to 7 and its reverse occur also (albeit with smaller probability). Figure 1 (left panel, bottom) displays the marginal posterior distribution for the number of sources. Note that the most probable number of sources ($N_s = 4$) is favoured with a probability of about 0.6. Figure 1 (right panel) displays samples of all source distribution models with $N_s = 4$. We note that there are four clusters of points, with the centroids of these clusters coinciding (approximately or better) with the true location of the four actual sources.

Figure 2 shows the marginal posterior distribution (histogram) of the parameters (x_s, y_s) [source location] and q_s [emission rate] for each of the four discrete sources identified in Figure 1 (right panel). The posterior mean and standard deviation, as well as the lower and upper bounds for the 95% HPD interval, of the parameters for each of these four identified discrete sources are summarized in Table 1. For this example, it is seen that generally the estimates for the source parameters are good and, certainly the true values of the parameters (when these are known) lie within the stated errors.

Table 1. The posterior mean, posterior standard deviation, and lower and upper bounds of the 95% HPD interval of the source location (x_s, y_s) and emission rate q_s of each of the four sources identified in Figure 1 (right panel).

	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)	–

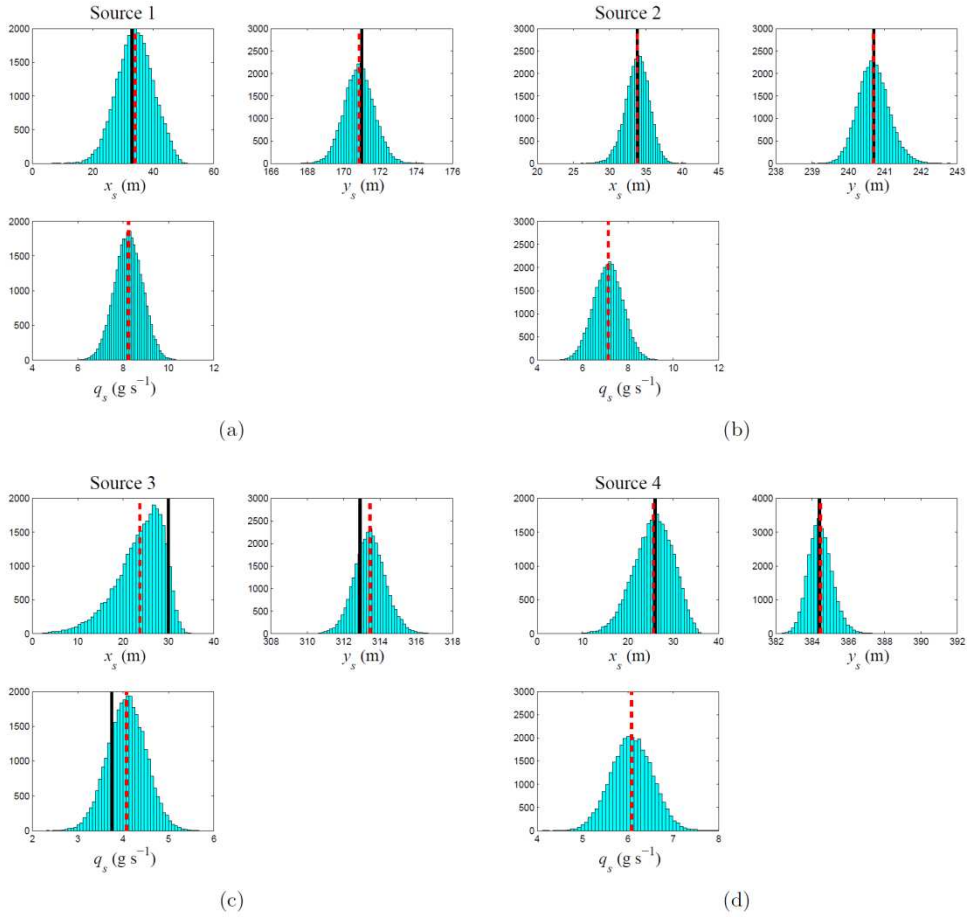

Figure 2. Histograms for the three parameters, alongwind location x_s, crosswind location y_s, and emission rate q_s that characterize the four sources identified in Figure 1 (right panel). In each frame, the solid vertical line indicates the true value of the parameter (if known) and the dashed vertical line corresponds to the best estimate of the parameter obtained as the posterior mean of the associated marginal posterior distribution.

CONCLUSION

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*. The source reconstruction methodology has been successfully validated against a real dispersion field experiment involving a multiple-source release with measurements of the resulting concentration field obtained from an array of detectors. The example illustrates the effectiveness of the proposed methodology and demonstrates 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.

REFERENCES

Gamerman, D. and H. F. Lopes, 2006: Markov chain Monte Carlo: Stochastic simulation for Bayesian inference, 2nd Edition (Texts in Statistical Science), CRC Press, Chapman and Hall, Boca Raton, Florida, 245 pp.

Keating, A., E. Yee and F.-S. Lien, 2007a: Bayesian inference for source determination with applications to a complex urban environment, *Atmos. Environ.*, **41**, 465-479.

Keating, A., E. Yee and F.-S. Lien, 2007b: Efficiently characterizing the origin and decay rate of a non-conservative scalar using probability theory, *Ecol. Modelling*, **205**, 437-452.

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 the *Russian-Canadian Workshop on Modelling of Atmospheric Dispersion of Weapon Agents*, Karpov Institute of Physical Chemistry, Moscow, Russia.

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

Yee E., 2008: Theory for reconstruction of an unknown number of contaminant sources using probabilistic inference. *Boundary-Layer Meteorol.* **127**, 359-394.

Yee E., F.-S. Lien, A. Keats and R. D'Amours, 2008: Bayesian inversion of concentration data: Source reconstruction in the adjoint representation of atmospheric diffusion. *J. Wind Eng. Indust. Aerodyn.* **96**, 1805-1816.