Bayesian inference for source determination in the atmospheric environment

by

W. Andrew Keats

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Mechanical Engineering

Waterloo, Ontario, Canada, 2009

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

In the event of a hazardous release (chemical, biological, or radiological) in an urban environment, monitoring agencies must have the tools to locate and characterize the source of the emission in order to respond and minimize damage. Given a finite and noisy set of concentration measurements, determining the source location, strength and time of release is an ill-posed inverse problem. We treat this problem using Bayesian inference, a framework under which uncertainties in modelled and measured concentrations can be propagated, in a consistent, rigorous manner, toward a final probabilistic estimate for the source.

The Bayesian methodology operates independently of the chosen dispersion model, meaning it can be applied equally well to problems in urban environments, at regional scales, or at global scales. Both Lagrangian stochastic (particle-tracking) and Eulerian (fixed-grid, finite-volume) dispersion models have been used successfully. Calculations are accomplished efficiently by using adjoint (backward) dispersion models, which reduces the computational effort required from calculating one [forward] plume per possible source configuration to calculating one [backward] plume per detector. Markov chain Monte Carlo (MCMC) is used to efficiently sample from the posterior distribution for the source parameters; both the Metropolis-Hastings and hybrid Hamiltonian algorithms are used.

In this thesis, four applications falling under the rubric of source determination are addressed: dispersion in highly disturbed flow fields characteristic of built-up (urban) environments; dispersion of a nonconservative scalar over flat terrain in a statistically stationary and horizontally homogeneous (turbulent) wind field; optimal placement of an auxiliary detector using a decision-theoretic approach; and source apportionment of particulate matter (PM) using a chemical mass balance (CMB) receptor model. For the first application, the data sets used to validate the proposed methodology include a water-channel simulation of the near-field dispersion of contaminant plumes in a large array of building-like obstacles (Mock Urban Setting Trial) and a full-scale field experiment (Joint Urban 2003) in Oklahoma City. For the second and third applications, the background wind and terrain conditions are based on those encountered during the Project Prairie Grass field experiment; mean concentration and turbulent scalar flux data are synthesized using a Lagrangian stochastic model where necessary. In the fourth and final application, Bayesian source apportionment results are compared to the US Environmental Protection Agency's standard CMB model using a test case involving PM data from Fresno, California. For each of the applications addressed in this thesis, combining Bayesian inference with appropriate computational techniques results in a computationally efficient methodology for performing source determination.

Acknowledgements

My supervisors, Drs Fue-Sang Lien and Eugene Yee, are directly responsible for making this phase of my personal, academic and professional development an incredibly rewarding experience. Dr Lien gave me the tools, freedom and support to carry my work with me (around the world), and I strongly believe that Dr Yee's philosophical and scientific guidance (and editorial prowess!) will have a positive, lasting impact on the rest of my career. Thank you both!

I would also like to thank my old office mates, Drs Johan Larsson and Andrea Scott, whose relative seniority, canny perspectives and good friendship helped me along the way (more than you know). As for my time at National Chung-Hsing University in Taiwan in 2007, I would like to thank Drs Ben-Jei Tsuang and Man-Ting Cheng, as well as all of their graduate students, especially Pei-Hsuan Kuo and Yung-Yao Lan, who looked after me extremely well and made my time there (in Lab 701) so enjoyable. Of the CMC in Dorval, I would like to thank Najat Benbouta and Alexandre Leroux for their support and friendship, as well as Richard Hogue and Michel Jean for facilitating my stay.

I gratefully acknowledge receiving financial support from the University of Waterloo (President's awards and graduate scholarships), the Natural Sciences and Engineering Research Council of Canada (Canada Graduate Scholarship), the Ontario Graduate Scholarship, and from the CRTI[1] under project number CRTI-02-0093RD.

Finally, my entire family has always been highly supportive during this endeavour, and my wife Sonia Singh has been my point of reference, helping me to gain perspective when it was necessary, and providing me with love and encouragement all along.

[1] Chemical, Biological, Radiological-Nuclear, and Explosives (CBRNE) Research and Technology Initiative (CRTI)

Contents

List of Figures

Chapter 1

Introduction

Accurately predicting how pollutants (both natural and anthropogenic) are dispersed in the atmospheric environment has become an important area of research over the past several decades. One immediately recognizable application involves modelling the way stack emissions and effluent releases are transported through the environment, so that 'safe' levels of pollutant concentrations can be maintained. In addition to such regulatory exercises, dispersion modelling can be used as a tool for public security. Releases (either intentional or accidental) of toxic gas can pose great danger to nearby populations, as evidenced by the sheer scale of the impact of the Bhopal disaster in 1984 (Ramesh, 2004).

Dispersion modelling is an example of a 'forward' problem in that the source of the emission is considered known. Of equal importance is the inverse problem: locating and characterizing the source of the pollutant using a finite and possibly noise-corrupted set of concentration measurements. In this thesis, the inverse problem is referred to as *source determination*, a term which reflects the general nature of the problem under consideration. In addition to the noun 'determination', other terms used by authors to describe specific applications include: identification, inversion, estimation, apportionment, localization, and reconstruction. The word 'determination' has been chosen because it applies more generally than the alternatives in terms of which properties of the source we wish to ascertain. For example, 'source localization' implies interest in the location but not necessarily the strength (or other properties) of the source; 'source apportionment' implies awareness of a number of different sources existing at different locations, for whom the relative magnitudes of their emissions must be estimated. However, in a problem of 'source determination', we are interested in any or all parameters describing the source (or sources), for example: location, strength, size, area, turn-on time, and even the identity of the chemical being released.

1.1 Theme and objective

In this thesis, Bayesian inference acts as the hub to which various pieces of computational and analytical machinery are attached with the objective of solving a variety of problems in source determination.

Inverse problems (such as source determination) are frequently described as being 'ill-posed' as though there is something wrong with the way they have been formulated. Adopting Bayesian inference (viz., probability theory as logic) requires a fundamental shift in our philosophical interpretation of such problems. Rather than regarding them as being posed incorrectly, one should interpret these problems as involving states of incomplete information, or more simply, uncertainty. Therefore, the solution to a source determination problem ought to take the form of a function which is capable of describing one's best available information after accounting for errors and uncertainties which originate from measurements, physical models, and prior information regarding the phenomenon under study. It turns out that the rules of probability theory form a 'calculus of inductive logic' for manipulating uncertain propositions, and so the final answer to a source determination problem is best expressed using a probability density function (PDF).

The central theme of this thesis is the use of the Bayesian methodology in formulating and solving several different problems in the area of source determination. For each of the cases considered, the problem is first formulated using probability theory before computations are performed (using dispersion models and statistical sampling techniques). The thesis is broken into two parts: Part I outlines the basics of the probability theory and the computational techniques needed to solve the source determination problem; Part II explores a number of applications, in each case stepping through the methodology while discussing implementational issues and presenting contributions.

The primary objective of this thesis is to demonstrate to the reader the power and flexibility that the Bayesian apparatus lends to the solution of the problems presented in Part II. For each case, either the flows, dispersion models, or applications themselves are quite different; however, probability theory allows each problem to be solved in a logically consistent way, under the same framework. Contributions made in Part II consist of the the development of new methods for solving source determination problems efficiently (such as the statistical tracer decay treatment discussed in chapter 6), as well as the novel adaptation of existing methods (e.g., Bayesian adaptive exploration as applied to the problem of optimal detector placement in chapter 7).

1.2 Contextual setting

While this thesis advances the available set of methodology and techniques for source determination, it is important to recognize that our particular inverse problem is only one out of the multitude of inverse problems encountered in the physical sciences (and in virtually every other discipline). For example, one influential text describing the Bayesian approach to

inverse problems is written from a geophysical viewpoint (Tarantola, 2005), while another focuses on information theory (MacKay, 2003). Astronomy is another field which has benefited significantly from adopting a Bayesian approach, with one key application being extrasolar planet finding (Gregory, 2005).

The applications contained in this thesis consider atmospheric transport over relatively short distances (tens or hundreds of kilometers at most). However, the techniques which have been developed could also be applied to transport at the continental or global scale. Yee et al. (2008) addressed a transient release being dispersed over continental Europe; monitoring the transport and characterizing the source of radionuclides released from a nuclear test is one example of an application involving global transport (Geer, 1996; Hourdin and Issartel, 2000).

1.3 Thesis outline

The work presented in later chapters is multidisciplinary in nature and is intended for an audience with a wide range of interests and perspectives. With this in mind, it is worthwhile to give a brief overview of the thesis content so that the individual reader may focus attention on the parts of greatest interest.

As mentioned above, Part I provides background information on Bayesian inference as required to solve source determination problems. Chapter 2 discusses both philosophical aspects and the fundamentals of Bayesian probability theory. In chapter 3, a simplified but general formulation of the source determination problem is presented. This formulation results in multidimensional PDFs for the source parameters which must be sampled or integrated efficiently. Computational techniques for accomplishing this are presented in chapter 4.

Part II of this thesis delves into a number of different applications, introducing specific dispersion models and sampling techniques as required for each individual case. Chapter 5 demonstrates that source determination can be performed in turbulent flows around obstacles, namely the Mock Urban Setting Test (MUST) array, and buildings in downtown Oklahoma City. An adjoint Eulerian dispersion model is used as a tool to calculate modelled concentrations in a computationally efficient way.

Chapters 6 and 7 involve simpler, horizontally homogeneous, parameterized wind fields, based on those found during the Project Prairie Grass (PPG) dispersion experiment. In chapter 6, a Lagrangian stochastic (LS) particle model is modified to be useful for estimating the rate of decay of a nonconservative tracer. Although the flow is simpler, this problem involves more parameters (the source location, strength and rate of decay are considered unknown) and the original LS model must be augmented in order to perform calculations efficiently. Chapter 7 uses the same LS model and flow fields to examine the problem of optimally placing additional detectors so that the information yielded by them is maximized. A simple example is presented which shows how Bayesian inference can be used in an information-theoretic context to address this problem.

The final case, presented in chapter 8, is a problem of source apportionment where a number of different sources contribute additively to particulate matter (PM) measured at a given

location. The PM is broken down into its individual chemical constituents and its chemical profile is compared to those of several known pollution sources (e.g., road dust, oil refinement). The chemical mass balance (CMB) model is used to relate sources to measurements, with the objective of determining the relative contribution of each source to the measured PM. This problem is entirely statistical in nature and does not involve a transport or dispersion model. However, an advanced Markov chain Monte Carlo (MCMC) method is used (hybrid Hamiltonian) which demonstrates that source determination problems with very large numbers of unknown parameters (hundreds and perhaps thousands) can still be solved efficiently.

Publications directly related to this work

Three of the chapters in Part II are adapted from papers that were published earlier:

Chapter 5: A. Keats, E. Yee, and F.-S. Lien. Bayesian inference for source determination with applications to a complex urban environment. *Atmospheric Environment*, 41:465–479, 2007.

Chapter 6: A. Keats, E. Yee, and F-S. Lien. Efficiently characterizing the origin and decay rate of a nonconservative scalar using probability theory. *Ecological Modelling*, 205: 437–452, 2007.

Chapter 8: A. Keats, M-T. Cheng, E. Yee, and F-S. Lien. Bayesian treatment of a chemical mass balance receptor model with multiplicative error structure. *Atmospheric Environment*, 43:510–519, 2009.

Part I

Source determination: theory and formulation

Chapter 2

Bayesian inference for inverse problems

This chapter begins by outlining the distinguishing features of inverse problems. Toward the end of Section 2.1, the rationale behind using probability theory as a method of solution is presented and placed in context with other, more traditional techniques.

In Section 2.2, we address the way in which probability theory is applied as a method for doing logical inference. At our disposal is Bayes' theorem, a mathematical tool which provides a consistent way for us to update our 'state of knowledge' about a model or system, upon the arrival of new information. Bayes' theorem is discussed in detail in Section 2.3.

2.1 Inverse problems

Inverse problems are pervasive in many areas of mathematics and science. They usually arise from a need to determine (or, 'infer') unknown or loosely constrained model parameters which best characterize a system whose output (be it numerical or physically measured) is known (Tarantola, 2005).

Wherever the available system output is subject to measurement or numerical error, or wherever the system output does not effectively constrain the model parameters, the inverse problem can be said to be 'ill-posed'. Hadamard (1902) characterized ill-posed problems as suffering from one or more of the following conditions:

1. An inverse transformation does not exist;

2. The inverse transformation is not unique;

3. The transformation is unstable (i.e., small changes in the data imply arbitrarily large changes in the model).

Hadamard went on to argue that such problems are actually incorrectly formulated and 'artificial'; however, it is now generally acknowledged that undertaking to solve ill-posed problems is a worthwhile cause. Examples of such inverse problems include:

Extrasolar planet-finding: Given measured data in the form of Doppler shifts in the absorption lines of a star's spectrum, determine the parameters (such as orbital period, eccentricity, planet mass, among others) which describe a Keplerian orbit of the hypothesized planet (or planets). Tinney et al. (2003) used a least-squares approach to solve this problem, while Gregory (2005) later reapplied a Bayesian probabilistic framework.

X-ray tomography: This is the process of inferring the composition (in terms of, e.g., density or attenuation coefficients) of some solid object by transmitting X-rays from different sources at varying angles through the object and measuring the attenuated ray on the other side (see Figure 2.1) (Tarantola, 2005). Assuming that the integrated ray attenuation is subject to some experimental noise, the problem is ill-posed and no unique solution exists.

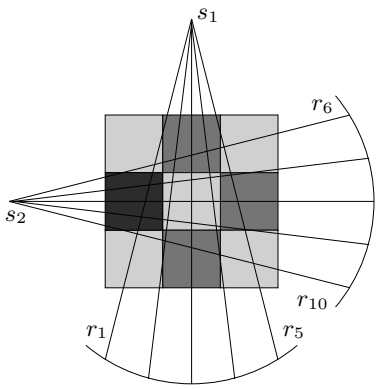


Figure 2.1: Example 2-D tomography problem of inferring the composition of 3×3 blocks with different attenuation coefficients (adapted from Tarantola (2005)). s_i: X-ray sources; r_j: receptor locations

Inverse heat conduction: As presented by Wang and Zabaras (2004), in this problem we wish to infer the heat flux on some part of the boundary of a conducting material, given a limited set of temperature measurements as well as knowledge of the heat flux at the remainder of the boundary (see Figure 2.2).

2.1.1 General formulation and common solution techniques

In general, forward and inverse problems are related through an operator (possibly nonlinear, denoted by $\mathbb{A}$) which relates the system output (e.g., noise-corrupted measured data, $\mathbf{d}$) to the input (e.g., parameters $\mathbf{m}$ defining an idealized system model):

$$\mathbf{d} = \mathbb{A}(\mathbf{m}) \; . \tag{2.1}$$

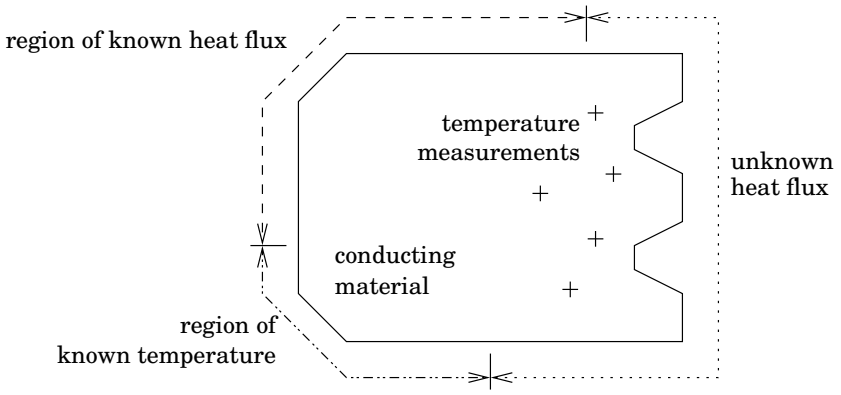


Figure 2.2: Example inverse heat conduction problem (adapted from Wang and Zabaras (2004)). The '+' signs represent temperature measurement locations.

If the data $\mathbf{d}$ and parameters $\mathbf{m}$ belong to the Hilbert spaces $\mathcal{D}$ and $\mathcal{M}$ respectively, the operator $\mathbb{A}$ performs the mapping:

$$\mathbb{A} : \mathcal{M} \rightarrow \mathcal{D} \ . \tag{2.2}$$

The inverse problem is then

$$\mathbf{m} = \mathbb{A}^{-1}(\mathbf{d}) \ , \tag{2.3}$$

where

$$\mathbb{A}^{-1} : \mathcal{D} \rightarrow \mathcal{M} \ , \tag{2.4}$$

whose solution can be approached in several different ways. The following three techniques constitute the more common approaches, in order of ascending general applicability:

Least-squares approach: We seek to minimize the residual by solving the problem

$$\hat{\mathbf{m}} = \arg \min_{\mathbf{m} \in \mathcal{M}} \|\mathbb{A}(\mathbf{m}) - \mathbf{d}\|^2 \ , \tag{2.5}$$

where $\|\cdot\|$ denotes the Euclidean norm, and $\hat{\mathbf{m}}$ denotes the optimal $\mathbf{m}$. For a well-conditioned system of linear equations, the solution is straightforward and can be accomplished analytically or using numerical optimization techniques (e.g., gradient-based, stochastic, etc.). However, when faced with an ill-conditioned operator, or data which does not effectively constrain the model parameters, the solution to this problem may be nonunique or unstable.

Regularization: When the operator $\mathbb{A}$ is ill-conditioned, it may still be possible to obtain a useful approximation to the solution by minimizing an objective function (e.g., the squared residual) which has been better-conditioned through the use of one or more 'regularization parameters'. A common scheme is Tikhonov regularization (Tikhonov, 1977):

$$\hat{\mathbf{m}}_\lambda = \arg \min_{\mathbf{m} \in \mathcal{M}} \|\mathbb{A}(\mathbf{m}) - \mathbf{d}\|^2 + \lambda^2 \|\mathbf{m}\|^2 \ , \tag{2.6}$$

where λ is a regularization parameter. This effectively applies a stable approximation to the inverse operator,

$$\mathbb{A}_{\lambda}^{-1} : \mathcal{D} \rightarrow \mathcal{M} \ , \tag{2.7}$$

resulting in a family of approximate solutions parameterized by λ. One main drawback of regularization techniques is that the choices of both λ and the regularization functional are somewhat subjective, although for a given functional, the 'optimal' value of λ can be determined according to various criteria. The choice of optimality criterion is also, however, subjective.

Bayesian probabilistic: This method has a solid philosophical grounding in terms of its connection to logic (see Section 2.2) and can provide a more comprehensive solution than least-squares or regularization because:

1. All prior knowledge of the parameters can be incorporated in a coherent manner;
2. The full solution takes the form of a probability density function (PDF) which summarizes the complete state of knowledge (including any uncertainty) of the parameter values, given all relevant information.

Under the Bayesian methodology, arbitrary solution parameters (such as λ) should no longer require manual tuning. Indeed, by choosing λ to be a function of the noise variance for cases where the data $\mathbf{d}$ is subject to Gaussian noise, Tikhonov regularization can be shown to be a special case of the more general Bayesian approach. The Bayesian probabilistic approach is adopted in this thesis and is explained in detail in the following sections.

2.2 Probability theory as extended logic

Using probability theory to formulate the source determination problem is sanctioned by the fact that the rules of probability theory form a 'calculus of inductive logic' (or inference) which allows us to manipulate proposals whose plausibility can be represented by a spectrum of real number values (e.g., $P \in [0, 1]$), rather than simply by True or False (e.g., 0 or 1). Cox (1946) began with simple desiderata;

1. Degrees of plausibility can be represented using real numbers;
2. The calculus should be consistent (viz., different methods of calculation should yield the same results);

and showed that the rules of probability calculus are equivalent to the rules for conducting inference. His derivation, known as 'Cox's theorem,' remains somewhat philosophically contentious, and there is disagreement over its range of applicability.[1] However, from a pragmatic

[1] A detailed philosophical and mathematical discussion of controversy surrounding the use of probability theory as extended logic is beyond the scope of this thesis. Examples of popular arguments against it are provided by Colyvan (2004) and Shafer (2004).

point of view, most scientific and mathematical reasoning is founded on classical deductive logic, and probability theory as extended logic has historically been able to provide acceptable results in a consistent fashion. Furthermore, no calculus has yet been proposed based on non-classical systems of logic, and Jaynes (2003) argues that such an approach, if it existed, would either be inconsistent with or isomorphic to classical deductive logic.

Before continuing with a description of the Bayesian interpretation of probability, we define the notation used in this thesis to express combinations of propositions and their probabilities. The rules for manipulating these probabilities are also defined below.

Notation: For propositions (or, hypotheses) A, B, and C, we adopt the notation used by Gregory (2005):

$P(A \mid B, C)$	Probability that A is true given ("|") B and C are true
A, B	Logical product (both A and B are true)
$A + B$	Logical sum (at least one of A and B is true)
$\overline{A}$	Negation (A is false)

Rules: The sum and product rules form the 'grammar' of the probability calculus. All legitimate relationships between probabilities can be derived from these rules:

Product rule:	$P(A, B \mid C) = P(A \mid C)P(B \mid A, C)$ $= P(B \mid C)P(A \mid B, C)$
Sum rule:	$P(A \mid C) + P(\overline{A} \mid C) = 1$ or $P(A + B \mid C) = P(A \mid C) + P(B \mid C) - P(A, B \mid C)$

For propositions whose plausibilities are definitively True (1) or False (0), the sum and product rules remain consistent with the axioms of two-valued symbolic logic.

2.2.1 The nature of probability

In order to use probability theory as a way of conducting inference, we require that the definition of 'probability' be expanded beyond its classical interpretation as a measure of the relative frequency with which the outcome of some event occurs over a very large number of trials. Historically, this approach has been labelled the Frequentist interpretation, while a more general approach (which does not require the event to be the realizable outcome of some theoretical ensemble) has been labelled (over the last 50-60 years) the Bayesian interpretation (after the Reverend Thomas Bayes, due to his essay, posthumously communicated in 1763). Formally, the relative frequency definition of probability is:

> If an experiment is repeated n times under identical conditions and n_x outcomes yield a value of the random variable $X = x$, the limit of n_x/n as n becomes very large, is defined as $P(x)$, the probability that $X = x$.
>
> (Gregory, 2005)

Note that this definition requires either a repeatable experiment or a large population from which to draw samples. It does not permit the term 'probability' to be used to describe more abstract situations where one is faced with an incomplete state of knowledge. In contrast, the Bayesian definition stipulates that probabilities express *degrees of belief* in any logical proposition or hypothesis:

> $P(A \mid B)$ is a real number measure of the plausibility of a proposition or hypothesis A, given (conditional on) the truth of the information represented by proposition B. "A" can be any logical proposition, and is not restricted to propositions about random variables.
>
> (Gregory, 2005)

The Bayesian interpretation is pragmatic in that it does not preclude the use of propositions of a frequentist nature. Furthermore, it allows one to consider propositions beyond the [possibly imaginary] realm of repeatable experiments, such as $A =$ "it will rain tomorrow", or $B =$ "there are 4 planets orbiting the star Betelgeuse". The usage of such propositions leads to criticism of Bayesian probability as being subjective; however, as long as model assumptions and prior information remain consistent, Bayesian probability theory yields consistent results.

In this thesis, we have adopted a Bayesian probabilistic approach because it permits us to interpret parameters as proposals. Indeed, in his original essay, Bayes used probability theory to calculate the distribution for the parameter of the binomial distribution. In the frequentist interpretation, parameters are not considered to have distributions associated with them, because they do not represent repeatable experimental outcomes. However, in the Bayesian interpretation, we consider our *state of knowledge* of the parameter value to be a probability, conditional upon all available data, assumptions, and theoretical models.

2.2.2 Confidence or credibility

Under the frequentist definition of probability, confidence intervals are used to describe our degree of certainty in an estimate. For example, consider the problem of estimating the average height of a male human. Using available data (the heights of some smaller subset of the world population), one could obtain an estimate for the height, $\hat{h}$, along with a surrounding interval based on knowledge or assumptions about the sampling distribution. A 95% confidence interval specifies that if the survey were repeated a large number of times, 95% of the intervals surrounding $\hat{h}$ would contain the true (but unknown) average height.

The Bayesian equivalent is known as the credible interval, and provides a more direct way of expressing one's degree of certainty in the parameter value. Since knowledge about the parameter is expressed as a PDF, a 95% credible interval can be specified by calculating points between which 95% of the probability mass lies. Further comparisons between the frequentist and Bayesian interpretations of confidence vs. credibility intervals can be found in (Brooks, 2003; Gregory, 2005).

2.3 Bayes' theorem

Bayes' theorem can be obtained from the product rule of probability, and is commonly expressed in the literature as:

$$\underbrace{P(M \mid D, I)}_{\textit{Posterior}} = \frac{\overbrace{P(M \mid I)}^{\textit{Prior}} \overbrace{P(D \mid M, I)}^{\textit{Likelihood}}}{\underbrace{P(D \mid I)}_{\textit{Evidence}}} . \tag{2.8}$$

In Eq. (2.8), M could represent, for example, a model and its concomitant parameterization. D might represent data (e.g., experimental or numerical) which we hope to use to improve our knowledge of the model parameters. I describes the background context – any available underlying information about our retrieval of the data and the applicability of the model. Note that for the moment, we are considering the terms M, D and I in their most general sense: solely as logical propositions. For the purpose of Bayes' theorem, M and D need not specifically refer to models and data. However, this thesis is concerned with estimating model parameters, so we assign $M \equiv$ model and $D \equiv$ data by way of example.

Bayes' theorem effectively provides a way to 'update' our current state of knowledge of M, after the arrival of some data (or more generally, information) D. Proceeding term-by-term:

The **prior** distribution $P(M \mid I)$ expresses our state of knowledge about M prior to the arrival of data D. For example, suppose M describes the parameters for a model of some physical phenomenon. If we are originally ignorant of the parameter values, the prior distribution should properly reflect this state of ignorance. A technique for choosing such priors is discussed in Section 2.3.1.

The quantity $P(D \mid M, I)$ is termed the **likelihood** function when considered as a function of M, but is known as the **sampling** distribution when considered as a function of D. $P(D \mid M, I)$ is normalized with respect to D, but not M; therefore, the likelihood function is technically not a PDF[2]. However, for a given inference problem, the likelihood function is used to obtain the sampling distribution. For fixed parameters M, the sampling distribution defines the probability of obtaining D. For example, consider that we have chosen a specific set of values which parameterize a model of the aforementioned physical phenomenon. The sampling distribution then expresses the probability that the data D were obtained, given this parameterization.

The **posterior** distribution $P(M \mid D, I)$ is the full solution to the inference problem and, converse to the likelihood, expresses the probability of M given D. Our final goal is to conduct inference over the parameters which define M, and the posterior expresses

[2] Singpurwalla and Wilson (2009) point out that "... the notion of probability is germane only for events that have yet to occur, or for events that have occurred but whose disposition is not known to you." In the case of the likelihood function, the data are known. Singpurwalla and Wilson go on to interpret the likelihood as a function "that prescribes the weight of evidence provided by the data for the different values that [the parameters] can take."

our complete state of knowledge of these parameters given all of the available data. Depending, however, on the dimensionality of M, it can be difficult to post-process or extract useful summary information from the posterior.

The **evidence** (sometimes known as the marginal likelihood) $P(D \mid I)$ is so-named because it measures the support for the hypothesis of interest (namely, the choice of model M). The term 'evidence' has recently been popularized by Skilling (2006) and MacKay (2003). In the literature, it is alternatively known as the 'marginal likelihood' (which describes its construction), or as the 'prior predictive' (which describes its use). For inference problems where only a single hypothesis (or, model) has been or will ever be considered, the evidence is an unimportant constant of proportionality.

Having given a brief overview of the terms present in Bayes' theorem, it is worthwhile to discuss the nature of two of the more interesting and misunderstood terms; namely, the prior and the evidence.

2.3.1 Prior information and ignorance

Objections to Bayesian probabilistic methods frequently center around a perceived 'subjectivity' of prior information. To counter this objection, one can argue that it is precisely because Bayes' theorem is able to account for prior information that the methodology is so powerful (Brooks, 2003). In any case, it is possible to place the selection of prior probabilities on firmer theoretical ground by requiring that they account properly for 'ignorance' regarding the proposal they describe.

In this thesis, in order to objectively specify prior probabilities, we adopt an information based approach known as the maximum entropy principle (MaxEnt), due to Jaynes (2003). Under the MaxEnt approach, prior distributions are chosen such that they are maximally non-committal (least informative) with respect to assumptions made about the nature of the distribution. This principle is most easily explained by considering the case of a six-sided die. If it is not known beforehand whether the die is weighted or biased, then the only constraint which can be used to specify the prior distribution is

$$\sum_{i=1}^{6} p_i = 1 \ , \tag{2.9}$$

where p_i is the probability that side i is rolled. Given this constraint, we find that by maximizing the entropy,

$$\arg\max_{p} \left(-\sum_{i=1}^{6} p_i \log p_i \right) \ , \tag{2.10}$$

the distribution which expresses maximum ignorance about the outcome is uniform, with $p_i = 1/6$ for all i.

The MaxEnt principle can also be applied to continuous distributions which describe, e.g., noise. Consider some measurement equipment which generates a signal ϕ. If the only known

characteristics of ϕ (i.e., constraints describing the distribution of ϕ) are its mean μ and variance σ^2, then it turns out that the distribution with maximum entropy which describes the noise is Gaussian:

$$P(\phi \mid I) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{\phi - \mu}{\sigma}\right)^2\right] . \tag{2.11}$$

Information about means and variances aside, we often require a continously defined prior which represents a state of total ignorance (within the bounds of the problem domain, of course) about some parameter. Such ignorance is typically expressed using a uniform distribution, although some care must be taken with regards to the invariance properties of the quantity being considered. Tarantola (2006) distinguishes two classes of physical quantities:

Cartesian: Quantities whose values lie in the interval $(-\infty, \infty)$, such as the components of velocity or position;

Jeffreys: Named for Sir Harold Jeffreys, who conducted the first investigations into the properties of positive quantities (Jeffreys, 1931, 1939), these are quantities such as temperature, density, frequency, etc., whose values span the range $(0, \infty)$. Tarantola reckons that in physics there are far more Jeffreys quantities than there are Cartesian.

One interesting feature of Jeffreys quantities is that they are commonly defined by their inverses, depending on the situation or problem at hand. Consider, for example, the following pairs: inverse temperature – temperature ($\beta = 1/T$); conductance – resistance ($C = 1/R$); specific volume – density ($\nu = 1/\rho$); frequency – period ($f = 1/T$). Tarantola goes on to argue that the aforementioned physical quantities are actually coordinates over quality spaces; for example, 'temperature' describes points in the *cold-hot* quality space. An appropriate distance metric for these quality spaces should remain invariant to a change in coordinates. Using density ρ and specific volume ν as an example, the appropriate metric which measures distances between points in the *empty-full* quality space is:

$$D_{\text{empty-full}}(\Omega_1, \Omega_2) = \left|\log \frac{\rho_2}{\rho_1}\right| = \left|\log \frac{\nu_2}{\nu_1}\right| . \tag{2.12}$$

Note that this definition of distance is additive,

$$D(\Omega_1, \Omega_3) = D(\Omega_1, \Omega_2) + D(\Omega_2, \Omega_3) . \tag{2.13}$$

Prior probabilities which express ignorance about Jeffreys quantities ought to satisfy scale invariance. For example, an ignorance prior for a frequency parameter with units of Hz should retain a constant probability mass from octave to octave, or from decade to decade. In other words, rather than considering the prior distribution of the quantity to be constant over a certain interval (as with Cartesian quantities), the logarithm of the quantity should remain constant. Therefore, for a Jeffreys quantity γ, one possible ignorance prior would be

$$P(\log \gamma \mid I) \propto \text{constant} , \tag{2.14}$$

or equivalently,

$$P(\gamma \mid I) \propto \gamma^{-1} \, , \tag{2.15}$$

with $\gamma \in [\gamma_{\min}, \gamma_{\max}]$ in order to render the prior proper (normalizable). For a Cartesian quantity θ, location invariance implies the familiar flat prior distribution:

$$P(\theta \mid I) \propto \text{constant} \, , \; \theta \in [\theta_{\min}, \theta_{\max}] \, . \tag{2.16}$$

Having made the distinction between location invariant (Cartesian) and scale invariant (Jeffreys) priors, it should be mentioned that in general, the practitioner is more concerned with the sensitivity of the posterior distribution to prior information than with the precise form of the prior distribution. Often, if the data sufficiently constrain the parameters under investigation, other prior distributions may be used (e.g., for analytical convenience) if they are sufficiently diffuse. In the source determination applications addressed in this thesis, we find that final estimates obtained for the source strength (a Jeffreys quantity) are relatively insensitive to a choice of Cartesian or Jeffreys prior.

2.3.2 Evidence

It was mentioned in Section 2.3 that the evidence term, $P(D \mid I)$, is an unimportant constant of proportionality when only one model is being considered. In many situations, this is not the case, and Skilling (2006) has argued that even if only one model is under investigation, the evidence should still be calculated as a courtesy to future researchers.

The evidence is obtained by marginalizing (integrating) the likelihood over the entire hypothesis space (where each 'hypothesis' parameterizes the model M):

$$P(D \mid I) = \int_{\text{all } M} P(D \mid M, I) \; P(M \mid I) \; \mathrm{d}M \, . \tag{2.17}$$

Essentially, the evidence is a numerical value which measures the suitability of the model to the data under consideration. If a competing model is able to better predict the data, then the evidence value will be higher. For example, calculating and comparing evidence values might be useful in a problem involving multiple point sources, where each model corresponds to a different number of hypothetical sources. If the measured concentration data are sufficiently descriptive, then the evidence value should be highest for the model which accounts for the correct number of sources.

Evidence values can be very difficult to calculate. If the number of parameters encountered in the model is high, then the integral in Eq. (2.17) will be of high dimensionality. Such integrals are best approached using stochastic techniques such as nested sampling (Sivia and Skilling, 2006; Skilling, 2006).

Chapter 3

Source determination

As discussed in the previous chapter, inverse problems commonly require model parameters to be estimated. In this chapter we use a simple illustrative example to demonstrate how Bayesian inference is applied toward the estimation of model parameters for a source determination problem.

3.1 Literature review

In the literature, problems involving source determination are addressed in many different contexts of varying scope. What follows is a review of the some of the literature surrounding source determination; for the most part, this literature pertains to relatively small-scale applications in atmospheric dispersion. Throughout the review we give particular attention to which of the source parameters were determined, as well as the method used to estimate the parameters. Further, specific literature surveys can be found in the introductory sections of the application chapters in Part II.

Research devoted to estimating the strength (emission rate) of a contaminant source (whose location is known) using a fixed network of concentration measurements was undertaken by Wilson and Shum (1992), who estimated the rate of ammonia volatilization from field plots using a Lagrangian trajectory model, and also by Flesch et al. (1995), who used a backward-time Lagrangian model to estimate the emission rate of a sustained surface area source in horizontally homogeneous turbulence. In both cases, the source strengths were determined using integral approaches based on information from a single sensor.

The problem of characterizing the strengths of a number of recognized sources (source apportionment) was addressed by Skiba (2003) who developed an adjoint pollution transport model to be used for estimating the emission rates of various industrial plants in Mexico. Skiba dealt with issues surrounding the posedness (in the Hadamard sense) of several variations on the problem and proposed non-probabilistic techniques (analytical, regularization, and least-squares) to estimate the emission rates for each variation. Another example of source apportionment involving a transport model is the work by Tsuang et al. (2003), who

implemented a Gaussian plume model with the aim of determining contributions of various sources (e.g., roads, power plants) to concentrations of SO_2, NO_x, particulate matter (PM), and secondary nitrate and sulfate aerosols. Source apportionment may also proceed without a transport model, by chemically analyzing captured PM and comparing it to known source profiles. This was performed in a Bayesian framework by Keats et al. (2009).

The 'source localization' problem was addressed by Nehorai et al. (1995) and by Jeremić and Nehorai (2000), who determined the physical location of a point source with the objective of finding land mines by sensing the vapours shed. They modelled the dispersion of the contaminant exclusively using a diffusion mechanism and used data from a static network of detectors measuring concentration over time. Finally, they estimated the location of the source using a maximum-likelihood method. Matthes et al. (2005) further refined the solution to the problem by taking advection into account, and used an analytical solution to the advection-diffusion equation in conjunction with a least-squares approach to estimate the source location. All of the aforementioned approaches assume idealized, undisturbed homogeneous flow fields, and treat the problem in an optimization framework using Monte Carlo or gradient-based methods.

At this point we digress briefly in order to provide background information describing how dispersion models are utilized for performing source determination. Specific dispersion models are introduced gradually in the applications presented in Part II of this thesis; suffice it to say that for a given source configuration, there are two main ways to calculate source-receptor relationships (the expected detector concentrations for a given source configuration). The first and more simplistic method involves running a forward dispersion model (e.g., numerically solving the advection-diffusion equation over a grid of points) to generate a concentration field, from which the expected detector concentrations can be extracted (e.g., by choosing the appropriate element from the resulting array of concentration values) for later comparison to measured data. The second option is to run a backward (or, adjoint) dispersion model once for each detector (e.g., solve the adjoint advection-diffusion equation) and obtain the detector concentration by computing the inner product of the source distribution and the 'adjoint concentration' (or, residence-time density) field. The advantage of using the second approach is that theoretical detector concentrations can be easily calculated (the inner product calculation is trivial compared to solving the advection-diffusion equation) for many different possible source configurations, once adjoint concentration fields have been generated. In a typical source determination problem there are far fewer detectors than potential source locations, so using backward dispersion models instead of forward ones will be computationally much more efficient.

The work of Yee (2005) and Keats et al. (2007a) advanced the field of source determination by introducing a Bayesian methodology for the simultaneous estimation of source strength, location, and turn-on and turn-off times, using adjoint dispersion models in the complex flow fields characteristic of urban (built-up) environments. Recent work of Keats et al. (2007c) has extended the methodology to account for nonconservative (reacting or decaying) tracers, and Yee (2008) successfully solved a problem involving an unknown number of point sources. Although the flows in the aforementioned cases are statistically stationary, the

Bayesian methodology is also applicable to time-varying fields, and Yee et al. (2008) were able to address a transient release being dispersed at a continental scale. Chow et al. (2005), Kosović et al. (2005) and Rapley et al. (2005) also recognized the value of using a Bayesian inferential framework to perform source determination in urban environments, but they did not implement adjoint dispersion equations for quickly computing source-receptor relationships. As a result, their calculations were computationally intensive. Nevertheless, as noted by Chow et al. (2008), adjoint equations rely on linearity in concentration, so nonlinear effects which occur during the dispersion process may be difficult to account for. Furthermore, effort is required to convert an existing forward dispersion model into its adjoint or backward form. Depending on the availability of computational power, a practitioner may simply prefer to use forward models. For example, Senocak et al. (2008) used forward Gaussian plume models to estimate not only source location and strength, but also wind field parameters describing transport and dispersion coefficients. Gaussian plume models predict concentrations using analytical formulae and can therefore be run very quickly; for a given (hypothesized) source location, it is not necessary to generate an entire concentration field in order to determine individual detector concentrations.

Estimation of both the location and strength of a source was also performed by Pudykiewicz (1998), who considered the problem of determining the source of a radioactive tracer. He recognized the problem as probabilistic in nature, and solved adjoint dispersion equations backwards in time in a global context. However, he mistakenly interpreted the solution to the adjoint equations as representing a probability density function (PDF) of the source location. In Section 3.3 we explain why this interpretation is invalid.

3.2 Problem formulation

By way of illustration, consider a single point source (whose height is known) located at a position given by coordinates (x_s, y_s), and whose emission rate q_s is assumed to be constant with time. These source properties act as a vector of model parameters:

$$\mathbf{m} \equiv (x_s, y_s, q_s) \ . \tag{3.1}$$

Detectors measuring mean concentration data are arranged downwind of the source, and the terrain is uniformly flat. The wind field is statistically stationary with a mean velocity of U m/s. The mean wind direction is aligned with the x-axis, as seen from the overhead view presented in Fig. 3.1.

For now we consider only a single model for the source distribution (i.e., that there is only a single point source, defined by the parameters $\mathbf{m}$); therefore, we are not yet concerned with calculating the evidence term. Furthermore, by using the MCMC method (to be discussed in chapter 4) we are able to draw samples from the posterior PDF without knowing the normal-

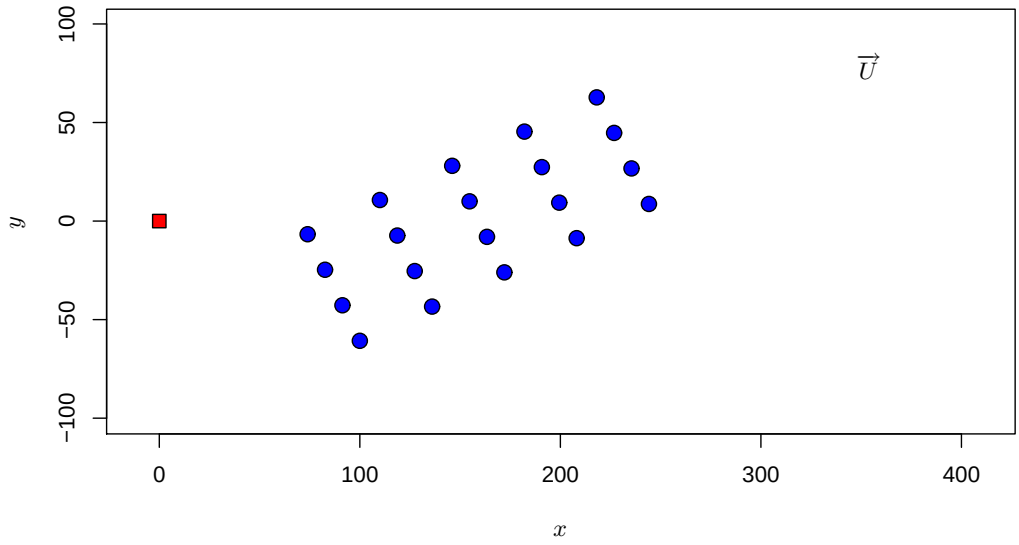


Figure 3.1: Example source-receptor configuration. Circular dots are detectors; the square dot is the [unknown] source.

ization constant, which allows us to use a simplified version of Eq. (2.8):

$$\underbrace{P(\mathbf{m} \mid \mathbf{d}, I)}_{\textit{Posterior}} \propto \underbrace{P(\mathbf{m} \mid I)}_{\textit{Prior}} \underbrace{P(\mathbf{d} \mid \mathbf{m}, I)}_{\textit{Likelihood}} \quad , \tag{3.2}$$

where $\mathbf{d}$ represents the concentration data measured by the detectors.

3.2.1 Likelihood of the parameters

Given that the source is described by the parameters $\mathbf{m}$, we require the probability that an array of detectors observes a certain set of concentrations $\mathbf{d}$. The likelihood function is used to quantify the probability of the discrepancy between the measured concentrations $\mathbf{d}$ and a corresponding set of modelled concentrations, $\mathbf{r}$, termed the theoretical source-receptor relationship. r_i is the value that detector i would theoretically measure if the source were characterized correctly by the parameters $\mathbf{m}$, and is determined using a dispersion model.

The discrepancy between the measured and modelled concentrations at the i^{th} detector, d_i and r_i, arises from [at least] two sources: measurement and model error[1]. First, consider that

[1] A detailed discussion of error sources and propagation in dispersion models can be found in the work by Rao (2005).

the measured mean concentration is subject to additive noise, e_i^{meas}:

$$d_i = d_i^{true} + e_i^{meas} \,, \tag{3.3}$$

where d_i^{true} is the (unknown) true value of the mean concentration at the i^{th} detector. For now we assume the noise e_i to be normally distributed, although other distributions which take into account the positivity and skewness of the noise [such as the lognormal, chosen by Goyal et al. (2005)] can also be used. Nevertheless, Sohn et al. (2002) have performed similar analyses (source localization) using Gaussian distributions, which we adopt both here and later in the thesis. The discrepancy e_i^{model} between the modelled and true concentrations is also assumed to be normally distributed:

$$r_i = d_i^{true} + e_i^{model} \,. \tag{3.4}$$

Both noise components e_i^{meas} and e_i^{model} are assumed to have a mean of zero and variances of $\sigma_{D,i}^2$ and $\sigma_{R,i}^2$, respectively. Furthermore, the measurement and the model errors for any detector are statistically independent, and the measurement and model errors across different detectors are also statistically independent. The measurement error is then codified as

$$P(\mathbf{d} \mid \mathbf{d}^{true}, \mathbf{m}, I) \propto \exp\left[-\frac{1}{2}\sum_i \frac{\left(d_i - d_i^{true}\right)^2}{\sigma_{D,i}^2}\right] \,, \tag{3.5}$$

which represents the probability that the observed data are measured as $\mathbf{d}$ when the true values are actually $\mathbf{d}^{true}$. Note that $\mathbf{m}$ appears in the PDF of Eq. (3.5) purely for accounting purposes; there is no logical dependence of measured concentration data on the model. This term could be removed but is technically necessary for evaluating an integral to follow. The model error is codified as

$$P(\mathbf{d}^{true} \mid \mathbf{m}, I) \propto \exp\left[-\frac{1}{2}\sum_i \frac{\left(d_i^{true} - r_i(\mathbf{m})\right)^2}{\sigma_{R,i}^2}\right] \,, \tag{3.6}$$

which states the probability that the true data are predicted by the model for the source-receptor relationship when the source parameters are $\mathbf{m}$. The likelihood is then obtained by marginalizing the joint PDF of $\mathbf{d}$ and $\mathbf{d}^{true}$ with respect to $\mathbf{d}^{true}$:

$$\begin{aligned} P(\mathbf{d} \mid \mathbf{m}, I) &= \int\limits_{all\ \mathbf{d}^{true}} P(\mathbf{d}, \mathbf{d}^{true} \mid \mathbf{m}, I)\ \mathrm{d}\mathbf{d}^{true} \\ &= \int\limits_{all\ \mathbf{d}^{true}} P(\mathbf{d} \mid \mathbf{d}^{true}, \mathbf{m}, I)\ P(\mathbf{d}^{true} \mid \mathbf{m}, I)\ \mathrm{d}\mathbf{d}^{true} \,. \end{aligned} \tag{3.7}$$

Evaluating the integral of Equation (3.7) yields the likelihood:

$$P(\mathbf{d} \mid \mathbf{m}, I) \propto \exp\left[-\frac{1}{2}\sum_i \frac{\left(d_i - r_i(\mathbf{m})\right)^2}{\sigma_{D,i}^2 + \sigma_{R,i}^2}\right] . \tag{3.8}$$

Assuming that we know $\mathbf{d}$, σ_D and σ_R, calculating $r_i(\mathbf{m})$ for various $\mathbf{m}$ provides $P(\mathbf{d} \mid \mathbf{m}, I)$. Note that Eq. (3.8) is Gaussian in the data, but is non-Gaussian in the source parameters $\mathbf{m}$.

3.2.2 Prior probability of the model parameters

The prior probability encompasses any information known about the source parameters prior to the arrival of the detector information. For example, if the effects of a toxic gas were to be qualitatively observed in some region of space, e.g., $\{x, y\} \subset \Omega$, then the value of the prior probability could be increased in this region according to the reliability of the observations.

In this thesis, it is assumed that nothing is known about the source parameters beforehand, and that the parameters are independent, in that knowledge of one parameter does not imply anything about the others. According to the principle of maximum entropy (which reduces to Laplace's principle of indifference in this case), the PDF whose distribution expresses complete ignorance about the parameter values is flat (Jaynes, 2003); therefore, the prior PDF is assigned a uniform distribution over the domain of definition Ω for the source parameters:

$$P(\mathbf{m} \mid I) = P(x_s \mid I)\ P(y_s \mid I)\ P(q_s \mid I) \propto \text{constant}, \qquad \mathbf{m} \in \Omega\ . \tag{3.9}$$

Of course, any PDF must integrate to unity, which is accomplished in this case by bounding the parameters $\mathbf{m}$. For a bounded computational domain Ω, we have $(x_s, y_s, q_s) \in \Omega$; for example, the source strength q_s is assumed to be greater than zero but less than some practical upper limit. The prior probability is also used to discount the possibility that the source lies within a building. It is set to zero in all of the within-building regions.

Alternatively, if we consider q_s to be a Jeffreys quantity, then we should assign a scale-invariant prior:

$$P(q_s \mid I) \propto q_s^{-1}, \qquad q_s \in [q_{\min}, q_{\max}]\ . \tag{3.10}$$

Typically, the data sufficiently constrain q_s (e.g., to within an order of magnitude) such that the posterior distribution is relatively insensitive to the choice of prior, and in the examples presented later in this thesis, both uniform and Jeffreys priors are used.

3.2.3 Posterior probability of the parameters

Assuming a constant prior distribution, the posterior PDF is essentially proportional to the likelihood:

$$\begin{aligned} P(\mathbf{m} \mid \mathbf{d}, I) &\propto P(\mathbf{m} \mid I) P(\mathbf{d} \mid \mathbf{m}, I) \\ &\propto \mathcal{I}(\mathbf{m} \in \Omega) \exp\left[-\frac{1}{2} \sum_i \frac{(d_i - r_i(\mathbf{m}))^2}{\sigma_{D,i}^2 + \sigma_{R,i}^2} \right]\ , \end{aligned} \tag{3.11}$$

where $\mathcal{I}(\blacksquare)$ denotes the indicator function, which returns '1' when the argument is true, and '0' otherwise.

3.3 Source-receptor relationship

While the datum d_i corresponds to a measured mean (time-averaged) concentration value, the source-receptor relationship $r(\mathbf{m})$ denotes the corresponding mean concentration predicted by a dispersion model for a given detector. For a detector located at (x_d, y_d), we have

$$r(\mathbf{m}) \equiv C(x_d, y_d; \mathbf{m}) \ , \tag{3.12}$$

where C is a concentration field obtained by running a dispersion model using $\mathbf{m}$ to define the source. Simply put, dispersion models provide a way to predict contaminant concentrations given a description of the source (e.g., shape, strength, location) and the physical domain (e.g., wind field, turbulence statistics, boundary conditions). Depending on the application, computed concentrations may be called for as time averaged quantities, higher moments, fluxes, or instantaneous values.

There are two main types of dispersion model: Eulerian advection-diffusion and Lagrangian stochastic. Each possesses its own set of strengths and limitations. Moreover, each can be used in either forward or backward mode. In forward mode, a dispersion model fulfills the capability described above (calculating a concentration field $C(x, y, z)$ given a source specification). However, depending on the computational requirements of the model, it may be infeasible to use a forward dispersion model to solve the source determination problem. In general, the source location is unknown a priori, and running a forward dispersion model for each possible source location could be untenable if the wind field is complex (as in a built-up urban area, for example). Backward models are essentially forward models which have been recast into a receptor-oriented (rather than source-oriented) mode. These models operate in the space which is dual to the concentration field, and are also known as adjoint models (because the adjoint advection-diffusion equation is solved). Qualitatively, a backward model is run backwards in time and space, and generates a 'dual' concentration field (e.g., $C^*(x, y, z)$) which has no physical meaning other than as a residence-time density field, or 'region of influence'. Issartel and Baverel (2003) use the term 'retroplumes' to describe these fields, since they take the same form as concentration plumes, except that they emanate upwind (backwards in time and space) from the detectors, not downwind from the source. Figure 3.2 illustrates the difference between the forward and backward (adjoint) problems as applied to a continuous point source release. Note that the region of influence ($C_2^* \cap C_3^* \setminus C_1^*$) is significantly smaller than the size of the plume, C. C_1^* is excluded from the region of influence because according to the shape of C, detector d_1 measures zero concentration.

As alluded to above, the primary motivation behind using backward models for source determination is a reduction in computational requirements. Numerically solving the adjoint advection-diffusion equation requires the same resources as the forward model, but for source determination we only require one model run per detector, rather than one run per potential source location. Details are presented later in the applications (Part II), but essentially, once the C^* field has been computed at a detector, concentrations expected at that detector can be calculated for any possible source configuration using an inner product, which reduces to a simple multiplication in the case of a point source. Typically, the number of available detectors

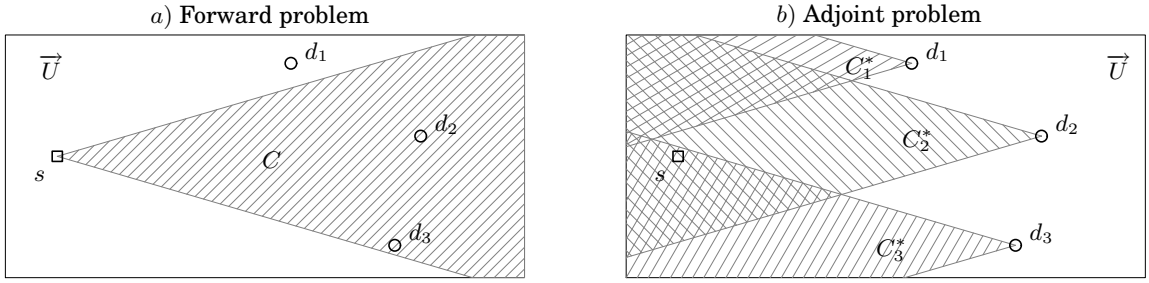


Figure 3.2: Illustration showing the relationships between the source, detectors, C field (plume), and C^* fields (retroplumes) for a 2-D problem layout with one source (denoted by s) and three detectors (d_1, d_2, d_3).

is far less than the number of possible source locations, so adopting a backward model may significantly reduce the computational requirements.[2] The idea of using a backward approach was probably first proposed by Gifford in 1959, who wrote a letter to the editor suggesting that a system of transparent plastic overlays with ground-level concentration isopleths could be placed over a map showing locations of sources and receptors. The concentration isopleths (which would correspond to a source of unit strength) were to be oriented upwind, with the plume origin lying at the receptor location. For source locations lying within the plume envelope, multiplying the isopleth values by the corresponding source strengths and summing the results over all relevant sources would yield an estimate for the concentration at the receptor. Gifford's system is essentially the same as that used in this thesis; plastic transparencies have been replaced by computational dispersion models.

Section 3.1 referred to work by Pudykiewicz (1998) in which the intersection of dual concentration fields (C^*) were used inappropriately as a PDF for the source location. This statement has generated some controversy [as witnessed by the comment-reply (Pudykiewicz, 2007; Keats et al., 2007b)] and is worth briefly clarifying here. It is certainly true that a zone of intersecting C^* fields actually demarcates the region where, according to the dispersion model, a potential source could have contributed to [non-zero] concentration measurements made at all detectors. However, summing these 'influence fields' does not yield any additional information about the source location beyond its potential existence. Moreover, the C^* fields are generated using a model which only approximates reality and are therefore subject to model errors with the result that the true source may not necessarily lie within the intersection of the fields. In effect, the PDF adopted by Pudykiewicz for the source location arbitrarily depends on the locations where detectors are placed, rather than the concentrations they mea-

[2] The computational savings discussed above rely on the wind field and/or emission being constant in time. When either the wind field or the emission changes in time, the workload increases by a factor of the mean number of 'time averaging intervals' during which each detector measures some mean concentration value. This is discussed further in chapter 5.

sure. Generally speaking, individual C^* fields are incompatible and cannot be agglomerated with the goal of obtaining information about the source.

3.4 Obtaining source parameter estimates

For posterior distributions defined over a low-dimensional space (viz., when the number of parameters in $\mathbf{m}$ is small), source parameter estimates can be obtained from marginal distributions. As suggested by the name, these distributions are obtained (numerically, or in some cases analytically) through a procedure called 'marginalization' in which irrelevant variables are integrated out of the PDF. For example, suppose we wish to estimate the x-location of the source. The marginal posterior PDF for the parameter x_s is:

$$P(x_s \mid \mathbf{d}, I) = \int_{\text{all } y_s} \int_{\text{all } q_s} P(x_s, y_s, q_s \mid \mathbf{d}, I) \, \mathrm{d}q_s \, \mathrm{d}y_s \; . \tag{3.13}$$

A qualified estimate for x_s could be obtained by calculating the mean and standard deviation of $P(x_s \mid \mathbf{d}, I)$. The precise choice of estimator is left to the practitioner, as there are cases where a median and credible interval might be desired instead. Correlations existing in the posterior distribution (e.g., x_s is frequently correlated with q_s) can be explored through multidimensional marginal distributions [e.g., $P(x_s, q_s \mid \mathbf{d}, I)$].

For the simple three parameter example given earlier, the posterior distribution can easily be computed over a three dimensional grid of points (with each point in parameter space representing a potential source hypothesis) for subsequent post-processing (e.g., marginalization). However, as source determination problems become more complex and involve more parameters, the computational effort and storage requirements increase exponentially. It is partially for this reason that Bayesian inference has only recently gained in popularity, with the advent of efficient sampling techniques such as Markov chain Monte Carlo (MCMC). In the literature, MCMC is overwhelmingly the numerical method of choice for solving problems formulated in a Bayesian framework. The linkage of Bayesian inference to MCMC is found in virtually all disciplines, and the development and improvement of MCMC algorithms is currently a topic of intensive research. The next chapter presents the basics of MCMC as used for source determination.

Chapter 4

Markov chain Monte Carlo for Bayesian inference

Although the value of the posterior distribution can be obtained directly to within a constant of proportionality, this calculation is nevertheless computationally expensive when conducted millions or billions of times. Conventional Monte Carlo integration is relatively inefficient for computing multidimensional integrals (or in the present work, sampling a multidimensional PDF), because samples are taken at random throughout the entire domain of the parameter space. Gregory (2005) provides the following illustration: suppose that for a one-parameter problem, the fraction of time spent sampling regions of high probability is 10^{-1}. Then in an N-parameter problem, this fraction could fall to 10^{-N}. In contrast, MCMC algorithms generate samples (e.g., $\mathbf{m}^{(k)} \in \mathcal{R}$; $\mathbf{m}^{(k)}$ is the k^{th} sample) in proportion to the value of the PDF, so time is not wasted generating samples from regions in parameter space which contribute very little.

Algorithms implementing MCMC are described by Hastings (1970), Gregory (2005), Gilks et al. (1996), and generally work in the following way:

1. Given some initial values for the parameters (e.g., $\mathbf{m}^{(0)}$), for each parameter we take a series of random steps and either accept or reject them depending on the transition probability [which in turn depends on the proposal distribution, as in the Metropolis-Hastings algorithm, as well as the properties (e.g., value) of the target distribution $P(\mathbf{m} \mid \mathbf{d}, I)$] at each step. Because each new point is chosen to be in a specific neighbourhood of the previous point, each new step $\mathbf{m}^{(k+1)}$ depends only on the previous step $\mathbf{m}^{(k)}$.
2. The series of samples generated by the MCMC method is a Markov chain; the distribution of these samples tends asymptotically to the distribution of $P(\mathbf{m} \mid \mathbf{d}, I)$.
3. Different MCMC algorithms will produce different Markov chains, but they must obey certain criteria in order for the chains to correctly represent the distribution of $P(\mathbf{m} \mid \mathbf{d}, I)$.

For example, consider the Markov chain associated with the i^{th} parameter, $\mathbf{m}_i$. If the k^{th} value of the chain is $m_i^{(k)}$, then subsequent steps could be defined as $m_i^{(k+1)} = m_i^{(k)} + \xi$, where ξ is

a random number drawn from a proposal distribution with mean 0. A normal distribution is commonly chosen, with the variance specified by the user. The proposal distribution does not necessarily need to be symmetric; the Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970) accounts for asymmetrical proposal distributions in the expression for the acceptance probability. This algorithm is often the starting point for applications of MCMC to Bayesian inference, due to its flexibility and ease in programming, and is outlined in Algorithm 4.1 in pseudocode. The following notation is used:

- $q(\cdot \mid \mathbf{m}^{(k)})$ is a *proposal distribution* from which we draw a *candidate* $\tilde{\mathbf{m}}$. For example, it might be a normal distribution with mean $\mathbf{m}^{(k)}$.
- $\alpha(\mathbf{m}^{(k)}, \tilde{\mathbf{m}})$ is the *acceptance probability*. This is the probability that the proposal $\tilde{\mathbf{m}}$ will be accepted as the next point in the Markov chain.

The acceptance probability is defined as

$$\alpha(\mathbf{m}^{(k)}, \tilde{\mathbf{m}}) = \texttt{min}\left[1, \frac{P(\tilde{\mathbf{m}} \mid \mathbf{d}, I)}{P(\mathbf{m}^{(k)} \mid \mathbf{d}, I)} \frac{q(\mathbf{m}^{(k)} \mid \tilde{\mathbf{m}})}{q(\tilde{\mathbf{m}} \mid \mathbf{m}^{(k)})}\right] \tag{4.1}$$

For symmetric proposal distributions ($q(\mathbf{m}^{(k)} \mid \tilde{\mathbf{m}}) = q(\tilde{\mathbf{m}} \mid \mathbf{m}^{(k)})$), the acceptance probability is simplified:

$$\alpha(\mathbf{m}^{(k)}, \tilde{\mathbf{m}}) = \texttt{min}\left[1, \frac{P(\tilde{\mathbf{m}} \mid \mathbf{d}, I)}{P(\mathbf{m}^{(k)} \mid \mathbf{d}, I)}\right] \tag{4.2}$$

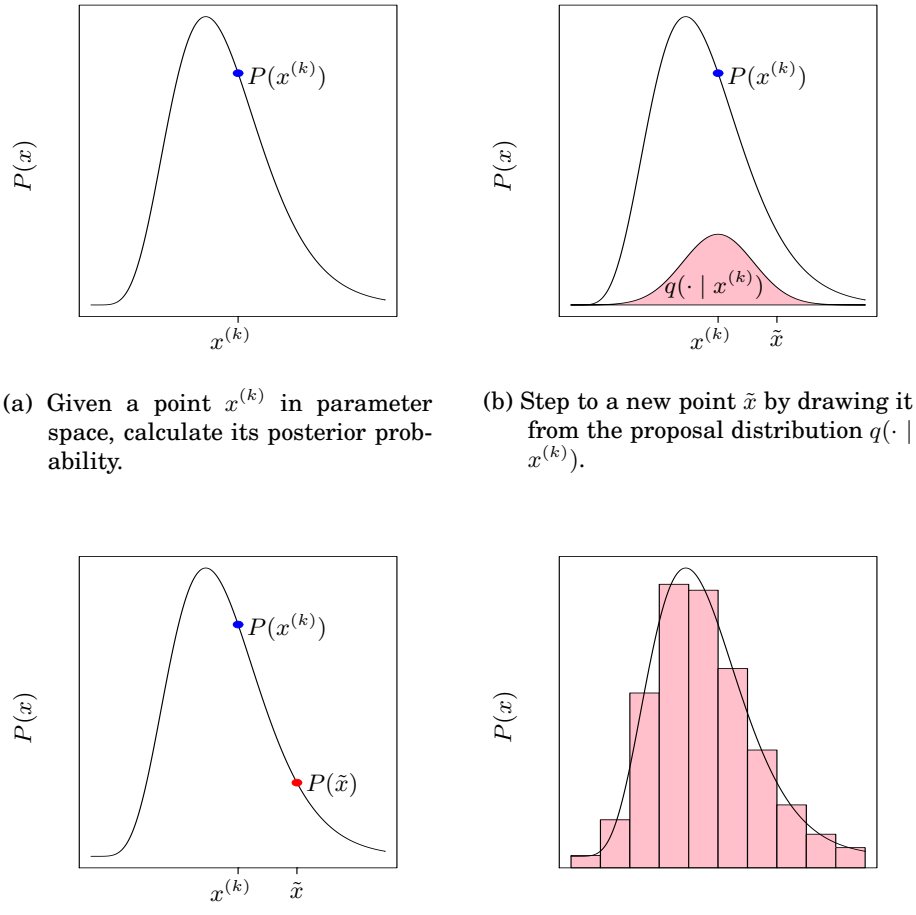
Algorithm 4.1 Metropolis-Hastings MCMC algorithm

select initial parameter values: $\mathbf{m}^{(0)}$
FOR $k = 0, 1, 2, \ldots$ DO
  $\tilde{\mathbf{m}} \leftarrow$ *sample from* $q(\cdot \mid \mathbf{m}^{(k)})$ {propose a new sample}
  $\alpha \leftarrow \min\left[1, \frac{P(\tilde{\mathbf{m}} \mid \mathbf{d}, I)}{P(\mathbf{m}^{(k)} \mid \mathbf{d}, I)} \frac{q(\mathbf{m}^{(k)} \mid \tilde{\mathbf{m}})}{q(\tilde{\mathbf{m}} \mid \mathbf{m}^{(k)})}\right]$ {calculate acceptance probability}
  $u \leftarrow$ *sample from* $\texttt{uniform}(0, 1)$
  IF $u < \alpha$ THEN
    $\mathbf{m}^{(k+1)} \leftarrow \tilde{\mathbf{m}}$ {accept the sample}
  ELSE
    $\mathbf{m}^{(k+1)} \leftarrow \mathbf{m}^{(k)}$ {reject the sample}
  END IF
END FOR

A crude sketch of the MCMC procedure as it might be applied to a one-dimensional posterior distribution (compactly denoted $P(x)$) can be found in Figure 4.1.

4.1 Chain mixing

In practice, the variance (or, width) of the proposal distribution q has a great impact on the ability of the algorithm to generate a representative chain of samples $\mathbf{m}^{(k)}$ in a reasonable amount of time. If the width is too large, the chain may stay at one point for a large number of steps (because the acceptance probability remains low). Alternatively, if the proposal distribution is too small, the chain may explore regions of high probability very slowly. This ability of the Markov chain to explore the parameter space is known as 'mixing'. The effect of the proposal distribution's width on chain mixing is demonstrated in Figure 4.2. The his-



(a) Given a point $x^{(k)}$ in parameter space, calculate its posterior probability.

(b) Step to a new point $\tilde{x}$ by drawing it from the proposal distribution $q(\cdot \mid x^{(k)})$.

(c) Calculate $\tilde{x}$'s posterior probability and compare it to $x^{(k)}$'s. Accept or reject the new point with some probability α. If acceptance, $x^{(k+1)} = \tilde{x}$. Otherwise, $x^{(k+1)} = x^{(k)}$

(d) After repeating steps (a–c), the histogram corresponding to the Markov chain $\{x^{(k)}, \forall k\}$ will converge toward the true marginal posterior distribution.

Figure 4.1: One-dimensional MCMC demonstration for a Metropolis-Hastings type algorithm.

tograms in the middle row indicate that the chain with the narrowest proposal width does not approximate the target distribution well. In this respect, the other two chains perform better, although the chain with the widest proposal distribution suffers from 'shot noise', which manifests itself as irregularity in the histogram bar heights. Trace plots (shown in the top row of Fig. 4.2) can be a good way for the practitioner to visually assess the progress towards convergence of a MCMC algorithm, especially if the number of parameters is reasonably low. However, it is also important to quantitatively assess chain convergence; one possible method for doing so is presented in the next section.

4.2 Chain convergence

For MCMC algorithms, 'convergence' expresses the degree of agreement between the chain's samples and the true (but unknown) target distribution. Many techniques exist for analyzing whether Markov chains have sufficiently converged; some involve comparing the sample means and variances of multiple chains (Gilks et al., 1996), while others examine the shape of the power spectrum (Dunkley et al., 2005). Later in this work we adopt Dunkley et al.'s convergence criterion, partly due to the fact that it only requires information from a single realization of a Markov chain. Briefly, Dunkley et al. analyzed the spectral properties of a Markov chain which had progressed beyond the initial 'burn-in' stage[1], testing for convergence by examining parameter values obtained through the fitting of a model template power spectrum. They imposed the following two requirements:

1. The distribution is not biased by correlated points, and the chain is drawing points throughout the full region of high probability. This is equivalent to saying that for the spectral representation of the chain (obtained using, e.g., a fast Fourier transform), frequencies lying below some pre-determined cut-off value have entered the white noise regime (viz., spectrally flat).

2. The 'convergence ratio' is specified by an estimate for the ratio of the sample mean variance ($\sigma_{\bar{x}}^2$) to the variance of the underlying distribution (σ_0^2). This ratio is required to lie below some cut-off value, such as 0.01.

Power spectra for three different Markov chains (all of which are exploring the same target distribution) are shown in Fig. 4.2. The middle chain, for which the mixing ratio is appropriate, has clearly entered the white noise regime; this is apparent from the delayed roll-off into the region of correlation at higher wavenumbers. Full details regarding the convergence criteria can be found in the work of Dunkley et al. (2005).

In practice, MCMC methods often require manual tuning in order to ensure that proposal widths are appropriate and that the individual chains (there is one chain per model parameter) will converge in a reasonable amount of time. Besides adjustable parameters, the actual

[1] The burn-in period occurs at the beginning of the chain and consists of the initial set of samples which are generated as the chain moves from the starting conditions toward the target distribution. These samples are typically discarded.

choice of MCMC algorithm can greatly affect the convergence rate. As the dimensionality of parameter space increases, the Metropolis-Hastings algorithm experiences increasing difficulty maintaining a reasonable acceptance rate (the ratio of the number of accepted samples to total chain length). For high-dimensional problems, 'guided' methods such as Hamiltonian MCMC become indispensable. Guided methods use additional information such as posterior gradients to steer the random walk and improve acceptance rates, while maintaining the correct mathematical properties of the MCMC method (asymptotic convergence to the target distribution). The Hamiltonian MCMC method is applied to solve the source apportionment problem in chapter 8.

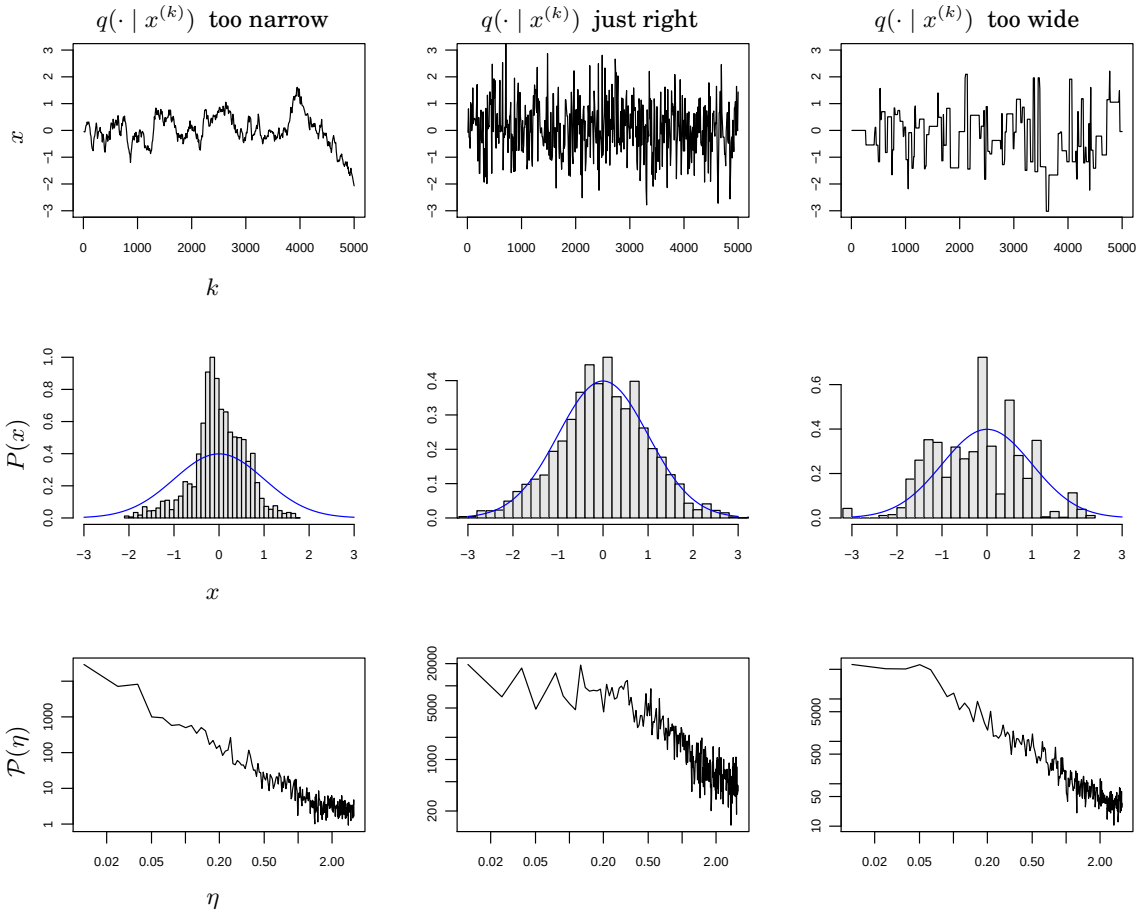


Figure 4.2: The effect of the proposal width on Markov chain mixing. The top row of 'trace plots' (thinned to show every tenth sample) show the Markov chains. The middle row compares histograms of chain samples to the target distribution (a standard Gaussian). Power spectra are plotted along the bottow row. Each chain's samples have a similar standard deviation. Goldilocks-inspired vernacular subtly hints toward the obligatory artful human meddling.

4.3 Post-processing samples

Once a series of MCMC samples has been obtained, summary statistics related to each variable can be found (e.g., sample means and variances). In the event that the target distribution is highly irregular (e.g., asymmetrical with multiple peaks or spikes), histograms are of descriptive value. In the present research, either the mean or median is used as the summary statistic of choice, since the distributions encountered may be somewhat irregular. By comparison, the commonly used maximum a posteriori estimator faces the potential problem of ignoring the bulk of the probability density by describing only an isolated local maximum.

Uncertainties associated with parameter histograms can be effectively communicated through credible intervals (Sec. 2.2.2), which are straightforward to determine given histogram information. Credible intervals can be constructed in a number of different ways; one possibility is to use the highest posterior density (HPD) interval. An HPD interval demarcates the region of the posterior distribution which contains $h\%$ of the total probability mass, such that the values of the PDF within the interval are everywhere larger than outside it. It should be noted that for multimodal distributions with large separations between modes (viz., 'disconnected' modes), the HPD region may be split into multiple intervals, each of which contains a sufficient proportion of the probability mass.[2]

[2] For this reason, the term 'HPD interval' may not be the most appropriate, since the HPD region need not consist of a single contiguous interval.

Part II

Applications and Case Studies

Chapter 5

Source determination in a complex urban environment

The material presented in this chapter is adapted from an earlier paper:

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

This chapter mainly quotes the methods and results from the paper. Some material which was present in the original article pertaining to probability theory, Bayesian inference and existing literature, has been removed, since it is now available in expanded form in Part I of this thesis.

5.1 Introduction

Determining the emission source of a contaminant released into the atmosphere has recently become a topic of intensive study because it carries important implications for both emergency and environmental management. The US Department of Homeland Security envisions using city-wide detector networks capable of sensing chemical, biological or radiological (CBR) emissions as a tool for mitigating potential acts of terrorism involving the release of CBR agents. Using a robust source determination methodology in conjunction with CBR detector measurements would provide emergency services with valuable additional information about the location and nature of a threat. On wider spatial and temporal scales, the problem of environmental contamination warrants the use of source determination methodology in a number of settings. Above ground, industrial sites issuing spurious emissions can be pinpointed (Skiba, 2003; Goyal et al., 2005), and regional sources of air pollution identified (Lin and Chang, 2002). Below ground, aquifer contamination history can be inferred (Aral et al., 2001; Michalak and Kitanidis, 2003). Another application of global importance is the enforcement of the Comprehensive Test Ban Treaty (CTBT). A world-wide network of

radionuclide detectors is in place as a verification tool and can be used to potentially isolate the location of clandestine nuclear testing (Geer, 1996; Hourdin and Issartel, 2000).

The main objective of this work is to show how Bayesian inference can be applied to solve a source determination problem in a consistent and computationally efficient way through the combined use of adjoint dispersion equations and MCMC. Section 5.2 formulates the problem for a time-dependent release (as well as the special case of a steady release) from a point source and explains the rationale behind using the adjoint equation and MCMC approaches. The methodology is validated using two different test cases in Sections 5.3 and 5.4, in which we simultaneously estimate the source location and strength for dispersion experiments performed in built-up environments.

5.2 Problem formulation and solution

Consider a vector of parameters, $\mathbf{m}$, which describe the properties of a transient point source:

$$\mathbf{m} = (x_s, y_s, z_s, q_s, t_{\text{on}}, t_{\text{off}}) \ , \tag{5.1}$$

where $\{x_s, y_s, z_s\}$ represent the spatial location of the source, q_s is its strength (of dimension $[MT^{-1}]$), and $\{t_{\text{on}}, t_{\text{off}}\}$ are turn-on and turn-off times. Bayes' theorem [reproduced from Eq. (2.8) below] provides a way to manipulate the conditional PDFs of the vector of source parameters $\mathbf{m}$, the concentration data $\mathbf{d}$, and background information I:

$$\underbrace{P(\mathbf{m} \mid \mathbf{d}, I)}_{\textit{Posterior}} = \frac{\overbrace{P(\mathbf{m} \mid I)}^{\textit{Prior}} \overbrace{P(\mathbf{d} \mid \mathbf{m}, I)}^{\textit{Likelihood}}}{\underbrace{P(\mathbf{d} \mid I)}_{\textit{Evidence}}} \ . \tag{5.2}$$

Using PDFs in the problem formulation allows us to account for bias and inaccuracy in our experimental and model data. It also provides a way to account for the fact that although many different source configurations may be plausible, some will be more probable than others. However, while the Bayesian approach provides the overall framework for solving this inverse problem, other techniques are required for our calculations to be of practical use (viz., both timely and sufficiently representative of our 'state of knowledge' of the source). Calculating the theoretical source-receptor relationship (the modelled mean concentration expected by a detector for a given source configuration) is rapidly accomplished using the adjoint advection-diffusion equation, which is described in Section 5.2.2. Furthermore, the posterior PDF, whose dimensionality may be high, must be sampled. This is accomplished using MCMC, a stochastic sampling technique which was described in chapter 4. Combining Bayesian inference with the adjoint and MCMC techniques results in an efficient and effective method for determining the source of a dispersion.

5.2.1 Bayesian formulation

The full solution to the source determination exercise is the posterior PDF, which represents the probability that the source parameters $\mathbf{m}$ take certain values, given $i)$ a set of concentration measurements $\mathbf{d}$, and $ii)$ any other background information I that is applicable to the problem. In order to be of practical use, this PDF must be marginalized for each source parameter m_i, and suitable summary statistics (e.g., the mean and standard deviations, $\overline{m_i}$ and σ_{m_i}) must be extracted.

Using MCMC, we are able to draw samples from the posterior PDF without knowing the normalization constant, so a simplified version of Equation (5.2) is used:

$$\underbrace{P(\mathbf{m} \mid \mathbf{d}, I)}_{\textit{Posterior}} \propto \underbrace{P(\mathbf{m} \mid I)}_{\textit{Prior}} \underbrace{P(\mathbf{d} \mid \mathbf{m}, I)}_{\textit{Likelihood}} . \tag{5.3}$$

This avoids the need to calculate the evidence term, a complicated multidimensional integral required for normalization. The posterior is proportional to the product of the likelihood and prior distributions, which are provided in the following section.

5.2.1.1 Assignment of the likelihood function

The likelihood function was formulated in Section 3.2.1 for the case where detector measurements and modelled concentrations are subject to additive Gaussian noise. It takes the following form:

$$P(\mathbf{d} \mid \mathbf{m}, I) \propto \exp\left[-\frac{1}{2}\sum_i \frac{(d_i - r_i(\mathbf{m}))^2}{\sigma_{D,i}^2 + \sigma_{R,i}^2}\right] . \tag{5.4}$$

Given measured concentration data $\mathbf{d}$, and specified uncertainties σ_D and σ_R, by calculating $r_i(\mathbf{m})$ for various $\mathbf{m}$ we are able to obtain $P(\mathbf{d} \mid \mathbf{m}, I)$ to within a constant of proportionality.

5.2.1.2 Assignment of the prior probability

In this chapter, we assume the prior for the source strength q_s to be a constant, resulting in an overall prior:

$$P(\mathbf{m} \mid I) = \text{constant}, \qquad \mathbf{m} \in \Omega , \tag{5.5}$$

which was given in Eq. (3.9).

5.2.1.3 The posterior probability density function

Since the prior is constant, the posterior PDF is essentially proportional to the likelihood:

$$\begin{aligned} P(\mathbf{m} \mid \mathbf{d}, I) &\propto P(\mathbf{m} \mid I) P(\mathbf{d} \mid \mathbf{m}, I) \\ &\propto \mathcal{I}(\mathbf{m} \in \Omega) \exp\left[-\frac{1}{2} \sum_i \frac{\left(d_i - r_i(\mathbf{m})\right)^2}{\sigma_{D,i}^2 + \sigma_{R,i}^2} \right] , \end{aligned} \tag{5.6}$$

where $\mathcal{I}(\blacksquare)$ denotes the indicator function. The dimensionality of this PDF is the same as that of the vector $\mathbf{m}$.

5.2.2 Source-receptor relationship

Efficiently calculating the source-receptor relationship $r(\mathbf{m})$ is crucial to the practical success of the source determination methodology. Under a brute-force approach, the forward advection-diffusion equation (5.8) must be solved for every desired combination of source parameters $\mathbf{m}$ (of which there may be potentially hundreds of millions). This approach yields an entire concentration field when only a limited set of modelled concentration measurements are required. In this paper we adopt an adjoint approach in which the adjoint advection-diffusion equation is solved only once for each detector, and the resulting conjugate concentration field is used to rapidly calculate the expected detector concentration for every desired combination of source parameters. Solving the adjoint advection-diffusion equation requires approximately the same computational time as does the forward advection-diffusion equation.

Consider a transient point source Q $[ML^{-3}T^{-1}]$ which releases material at a steady rate of q_s $[MT^{-1}]$ and whose turn-on and turn-off times are t_{on} and t_{off}:

$$Q = q_s \delta(\mathbf{x} - \mathbf{x}_s)[H(t - t_{\text{on}}) - H(t - t_{\text{off}})] , \tag{5.7}$$

where $\delta(\cdot)$ and $H(\cdot)$ are the Dirac delta and Heaviside unit step functions, and $\mathbf{x}_s \equiv \{x_s, y_s, z_s\}$ is the source location. The transport equation for the mean concentration,

$$\begin{aligned} &\frac{\partial C}{\partial t} + \mathbf{U} \cdot \nabla C - \nabla \cdot (\Gamma \nabla C) + \nabla \cdot \overline{\mathbf{u}' c'} = Q , \\ &\text{subject to} \qquad \nabla_{\mathbf{n}} C = 0 \quad \text{at } \partial\mathcal{R} , \\ &\qquad\qquad\quad\; C(\mathbf{x}, t = t_{\text{on}}) = 0 , \end{aligned} \tag{5.8}$$

models the release of the source Q over a space-time domain $\mathcal{R} \times [0, T]$ through the time evolution of the concentration field, C. Here, C $[ML^{-3}]$ denotes a Reynolds averaged (or, mean) concentration and the elements of $\mathbf{U}$ are Reynolds averaged wind velocities in the x, y, z directions. Γ is a molecular diffusivity, and $\overline{\mathbf{u}' c'}$ are the turbulent scalar fluxes. The boundary-normal direction is given by $\mathbf{n}$, $\nabla_{\mathbf{n}}$ is a directional derivative, and $\partial\mathcal{R}$ is the boundary of the

spatial domain. The scalar fluxes can be modelled using the gradient diffusion hypothesis:

$$\overline{\mathbf{u}'c'} = -\frac{\nu_t}{\mathrm{Sc}}\nabla C \ , \tag{5.9}$$

where ν_t is the kinematic eddy viscosity and Sc is the turbulent Schmidt number, a dimensionless quantity which corresponds to the ratio of momentum diffusivity to mass diffusivity. Under the modelling assumption (5.9), the transport equation for the mean concentration can be rewritten as the following advection-diffusion equation:

$$\frac{\partial C}{\partial t} + \mathbf{U}\cdot\nabla C - \nabla\cdot(K\nabla C) = Q \ , \tag{5.10}$$

where

$$K = \Gamma + \frac{\nu_t}{\mathrm{Sc}} \tag{5.11}$$

is the sum of a molecular diffusivity and an eddy diffusivity used to model the turbulent scalar flux rate. In flows where turbulent diffusion dominates molecular diffusion, $\Gamma \ll \nu_t/\mathrm{Sc}$. In the technical report by Yee et al. (2007) which describes the code used in this thesis to solve the forward and adjoint advection-diffusion equations, Sc is assigned a fixed value of 0.63.

Next, consider the modelled concentration measurement at the i^{th} detector, r_i. This specific value is a linear functional of the concentration field, C, and is determined by the inner product of C and a 'detector response function', h $[L^{-3}T^{-1}]$, at a specific location and measurement time:

$$r_i = \langle C, h\rangle \equiv \int_0^T dt \int_{\mathcal{R}} C\ h\ \mathrm{d}\mathcal{R} \ , \tag{5.12}$$

where $h = h(\mathbf{x}-\mathbf{x}_r, t-t_r)$ for a detector which measures the concentration at location $\mathbf{x}_r$ and time t_r. The function h acts as a space-time filter and would be, e.g., a delta function for an ideal detector with infinite resolving power. According to the duality relationship, r_i can also be obtained using the inner product of the conjugate concentration field C^* $[L^{-3}]$ and the source function Q:

$$r_i = \langle Q, C^*\rangle \equiv \int_0^T dt \int_{\mathcal{R}} Q\ C^*\ \mathrm{d}\mathcal{R} \ , \tag{5.13}$$

where the C^* field evolves according to the adjoint advection-diffusion equation:

$$\begin{aligned} -\frac{\partial C^*}{\partial t} - \mathbf{U}\cdot\nabla C^* - \nabla\cdot(K\nabla C^*) &= h \ , \\ \text{subject to} \quad K\nabla_{\mathbf{n}}C^* + \mathbf{U}\cdot\mathbf{n}C^* &= 0 \quad \text{at } \partial\mathcal{R} \ , \\ C^*(\mathbf{x}, t=t_r) &= 0 \ . \end{aligned} \tag{5.14}$$

The general procedure for obtaining the adjoint of a linear operator is outlined by Estep (2004), and is briefly illustrated below for the present case. The adjoint equation (5.14) is obtained by first multiplying the forward advection-diffusion equation (5.8) by a test function,

C^*, and integrating over the space-time domain:

$$\int_0^T \mathrm{d}t \int_{\mathcal{R}} C^* \frac{\partial C}{\partial t} \,\mathrm{d}\mathcal{R} + \int_0^T \mathrm{d}t \int_{\mathcal{R}} C^* \mathbf{U} \cdot \nabla C \,\mathrm{d}\mathcal{R} \\ - \int_0^T \mathrm{d}t \int_{\mathcal{R}} C^* \nabla \cdot (K \nabla C) \,\mathrm{d}\mathcal{R} = \int_0^T \mathrm{d}t \int_{\mathcal{R}} C^* Q \,\mathrm{d}\mathcal{R} \,. \tag{5.15}$$

Integrating by parts and taking advantage of the divergence theorem where applicable, the derivative terms can be rearranged to yield an expression which is compatible with

$$\langle C, h \rangle = \langle Q, C^* \rangle + \text{boundary terms} \,, \tag{5.16}$$

and which obeys the boundary conditions associated with the forward problem. The term h manifests itself in this expression as the left hand side of Equation (5.14) and can be extracted by inspection. Boundary conditions for the adjoint advection-diffusion equation are chosen such that the boundary terms of Equation (5.16) vanish, resulting in the duality relationship between C and C^*,

$$\begin{aligned} \langle C, \mathbb{L}^* C^* \rangle &= \langle \mathbb{L} C, C^* \rangle \\ \text{or } \langle C, h \rangle &= \langle Q, C^* \rangle \,. \end{aligned} \tag{5.17}$$

The linear operators $\mathbb{L}$ and $\mathbb{L}^*$ are defined by:

$$\begin{aligned} \mathbb{L}(\blacksquare) &\equiv \frac{\partial}{\partial t}(\blacksquare) + \mathbf{U} \cdot \nabla(\blacksquare) - \nabla \cdot \big(K \nabla(\blacksquare)\big) \,, \\ \Rightarrow\ \mathbb{L}(C) &= Q \,, \end{aligned} \tag{5.18}$$

$$\begin{aligned} \mathbb{L}^*(\blacksquare) &\equiv -\frac{\partial}{\partial t}(\blacksquare) - \mathbf{U} \cdot \nabla(\blacksquare) - \nabla \cdot \big(K \nabla(\blacksquare)\big) \,, \\ \Rightarrow\ \mathbb{L}^*(C^*) &= h \,. \end{aligned} \tag{5.19}$$

The inner product of Equation (5.17) can be rapidly calculated in order to find the concentration at a detector for any choice of Q using the value of the C^* field at the location of the point source. For a line, area or volume source, this calculation would be more involved, but still simpler than re-solving Equation (5.8) for a new source term.

In practice, detectors do not obtain concentration readings continuously; rather, they measure for a period of time and provide the average concentration over that period. Therefore, the concentration information at detector i typically takes the form of a time series, with each data point centered around a time of measurement, $t_j^{(i)}$. The concentration read by detector i during time period j is denoted by $d_i^{(j)}$. Figure 5.1 shows how a smooth (theoretical) concentration curve might be sampled by a detector.

It is necessary to solve Equation (5.14) once for every possible (i, j) in order to find $C_i^{*(j)}$. Substituting the assumed source distribution of Equation (5.7) into the duality relationship

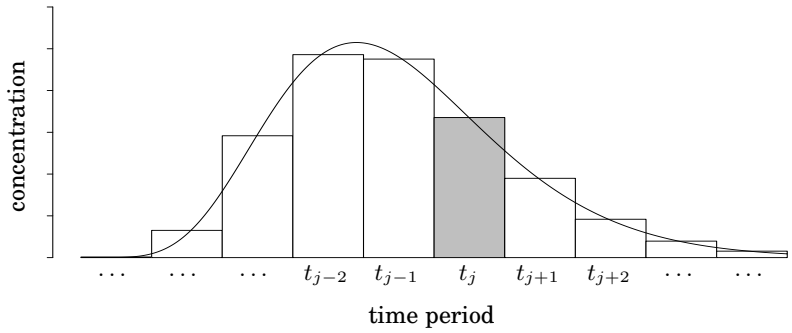


Figure 5.1: A detector measures an average value of the concentration over time period j, resulting in a series of concentration readings at times t_j.

of Equation (5.17), one obtains the source-receptor relationship:

$$r_i^{(j)}(\mathbf{m}) = q_s \int_{t_{\text{on}}}^{\min(t_j^{(i)}, t_{\text{off}})} C_i^{*(j)}(\mathbf{x}_s, t_s)\, \mathrm{d}t_s \tag{5.20}$$

where $C_i^{*(j)}$ is a time-varying field corresponding to the measurement taken by detector i during time period $t_j^{(i)}$, and $r_i^{(j)}(\mathbf{m})$ is the averaged concentration that detector i would expect to have measured during $t_j^{(i)}$ if the source were correctly characterized by $\mathbf{m}$. The total number of unsteady C^* fields that need to be generated is equal to the product of the number of detectors (indexed by i) multiplied by the number of time intervals $t_j^{(i)}$ sampled at detector i. Finally, it is important to note that calculating the posterior distribution involves a summation over all possible (i, j):

$$P(\mathbf{m} \mid \mathbf{d}, I) \propto \mathcal{I}(\mathbf{m} \in \Omega) \exp\left[-\frac{1}{2} \sum_{i,j} \frac{\left(d_i^{(j)} - r_i^{(j)}(\mathbf{m}) \right)^2}{\sigma_{D,i,j}^2 + \sigma_{R,i,j}^2} \right] . \tag{5.21}$$

5.2.2.1 Continuous releases

The problem of a point source which releases material continuously in time into a statistically stationary wind field is a special case of the more general formulation presented above. Here we consider a point source of the form:

$$Q = q_s \delta(\mathbf{x} - \mathbf{x}_s) , \tag{5.22}$$

which is a special case of Equation (5.7) for $t_{\text{on}} \to -\infty$ and $t_{\text{off}} \to \infty$. The steady forward and adjoint advection-diffusion equations are obtained by integrating Equations (5.8) and (5.14), respectively, over all time:

$$\mathbf{U} \cdot \nabla C - \nabla \cdot (K \nabla C) = Q \tag{5.23}$$

$$-\mathbf{U} \cdot \nabla C^* - \nabla \cdot (K \nabla C^*) = h . \tag{5.24}$$

The units of C^* and h become $[TL^{-3}]$ and $[L^{-3}]$, respectively, and the duality relationship, Equation (5.17), remains applicable. Thus, the source-receptor relationship is calculated using:

$$r_i(\mathbf{m}) = q_s C_i^*(x_s, y_s, z_s) \ , \tag{5.25}$$

where the set of parameters has been reduced to $\mathbf{m} = \{x_s, y_s, z_s, q_s\}$.

5.3 Mock Urban Setting Test (MUST) array

The Mock Urban Setting Test (MUST) is a transport and dispersion experiment which took place at Dugway Proving Ground in Utah during September 2001 (Yee and Biltoft, 2004). It was designed to simulate dispersion in a built-up (urban) area using an array of shipping containers (or building-like obstacles). This array consisted of 12 rows of obstacles in the streamwise x-direction and 10 columns in the spanwise z-direction. Propylene gas was used as a tracer and released from various locations within the array both continuously and intermittently, and concentration time series were obtained at detectors spaced throughout the array. A physical model of the MUST field experiment was conducted at a scale of 1:205 in a boundary-layer water channel operated by Coanda Research & Development Corporation (Burnaby, British Columbia, Canada). A detailed description of the experiment is provided by Hilderman and Chong (2004) and Yee et al. (2006). The inverse problem is solved using individual concentration measurements $\mathbf{d}$ which are extracted from the concentration profiles that were measured during the water channel experiment.

The MUST array test case is useful for validating the source determination methodology because it possesses the following attributes:

1. Detailed continuous source concentration data are available from a number of experiments at regularly-spaced locations within the array of obstacles.

2. The experiments simulate an urban environment, which tests the ability of the methodology to locate a source lying in a built-up area.

3. The flow encounters obstacles, which affect the expected shape of the posterior distribution (the probability of a source lying within a building is considered to be zero). Furthermore, obstacles induce recirculation zones, which enhance the mixing of the tracer and obscure the results of the inference procedure (i.e., the uncertainty of the source location is increased in these zones).

5.3.1 Procedure

5.3.1.1 C^* field generation

The C^* fields relevant to the MUST array are generated by solving the steady adjoint advection-diffusion equation (5.24). The mean fluid velocities $\mathbf{U}$ and turbulent diffusivity

ν_t are generated on a structured grid of points using the urbanSTREAM code of Lien et al. (2005a). In order to reduce the computational workload, however, the adjoint code is only solved over a three-row subset of the domain. This domain size is sufficient, given that the bulk flow is in the x-direction and the plume width is smaller than the spanwise (z-direction) width of the domain. Figure 5.2 shows an x-z slice of the mesh used for calculation of the C^* fields, with obstacles added to provide context.

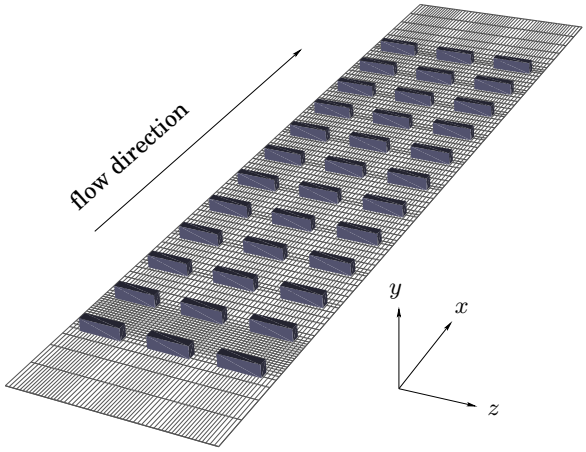


Figure 5.2: A ground-level ($y = 0$) slice of the mesh on which the adjoint advection-diffusion equation was solved.

Note that the mesh is significantly more dense between the first two rows of obstacles – this is a remnant of the fact that this mesh was originally designed for a forward dispersion calculation (the tracer was released from the region between the first two rows of buildings), and was not redesigned to accomodate adjoint (backwards) calculations.

The C^* fields generated by the adjoint advection-diffusion code do not depend on time because the velocity field is steady, and at present only continuous releases have been considered. However, the flow is non-homogeneous in all three spatial directions, so one C^* field must be generated for each detector.

5.3.1.2 Detector selection

Concentration profiles were experimentally measured at several x-locations and heights. Figure 5.3 shows an array of detector locations which lie along the paths of the experimental profiles. Starting from the left ($x = 0$), each spanwise line of detectors is numbered according to its position relative to the leftmost spanwise line of obstacles. The convention used here is to refer to the lines of detectors as belonging to rows 2.5, 3.5, 4.5, 6.5, 9.5 and 12.5. Experimental concentration profiles are not available at other x-locations.

By comparing the profiles generated by the forward dispersion model to the experimental profiles, it is clear that the model error dominates the measurement error. This comparison is shown for rows 2.5 and 4.5 in Figure 5.4. Although the detector concentrations obtained using

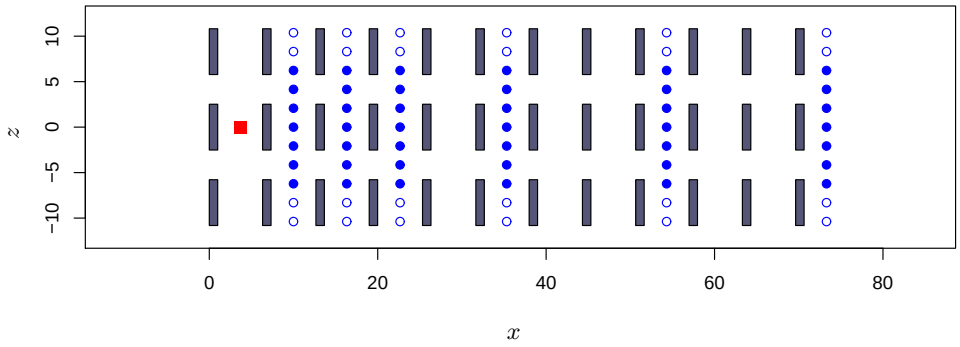
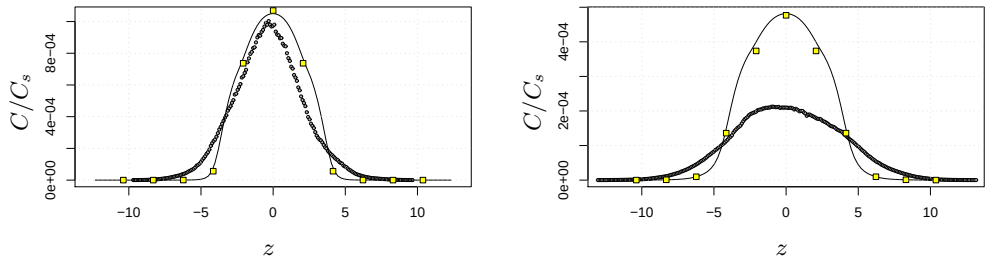


Figure 5.3: MUST array source and detector configuration. The x, y, and z coordinates are normalized by the width of an obstacle in the streamwise (or, x) direction. The source (marked by the square) lies at ground level while the detectors (marked by circles) lie at $y = 0.5$ building heights in rows 2.5, 3.5, 4.5, 6.5, 9.5, and 12.5. In the spanwise direction, the detectors are placed at $z = 0.0, \pm 2.075, \pm 4.150, \pm 6.225, \pm 8.305, \pm 10.380$ building heights. The filled circles depict the 42 detectors used to determine the source location.

the C^* fields agree well with the forward dispersion model, the plume center is considerably overpredicted with increasing x-distance when compared with the experimental data. The lumped theoretical and measurement uncertainties for each detector, $\sigma_{L,i} = (\sigma_{D,i}^2 + \sigma_{R,i}^2)^{1/2}$, are assigned values based on the ability of the dispersion model to predict the experimental concentrations. These values lie between 10% and 270% of the mean concentration measured at a given detector. In general, detector uncertainties increase along the centerline in the streamwise x-direction as the flow is disturbed and the model increasingly overpredicts the concentration. The 42 detector locations depicted in Figure 5.3 were chosen as measurement stations to be used for the source reconstruction. Each detector is assigned an uncertainty according to its position as described above.

5.3.2 Results

The source reconstruction was performed using experimentally measured concentration data. The Metropolis-Hastings algorithm was used to generate 10^7 MCMC samples from various initial conditions. Since the acceptance rate (the proportion of accepted proposals out of the total number of samples drawn) was low (approximately 15%), these chains were thinned to obtain smaller chains consisting of every 100^{th} sample in order to avoid autocorrelation. Although the chain was qualitatively found to quickly converge to the area of interest, we conservatively chose to discard the first half of these samples to avoid 'burn-in'. In all, 5×10^4 samples were used to generate the histograms shown in Figure 5.5. It should be noted that the burn-in period is often chosen by inspecting the results of the Markov chain in order to determine whether it has reached a stationary distribution. The thinning interval can be



(a) Concentration profile at row 2.5 (b) Concentration profile at row 4.5

Figure 5.4: Normalized concentration profiles. Small circles represent experimental measurements, the solid line represents the solution to the forward advection-diffusion equation, and the squares represent detector concentrations reconstructed using C^* fields according to Equation (5.25).

selected by computing the autocorrelation function of the chain samples. Automatic methods for determining optimal burn-in and thinning parameters have been proposed throughout the MCMC literature (e.g., Gilks et al., 1996; Dunkley et al., 2005); however, in the present case the chains were inspected manually. In order to verify the MCMC results, the posterior distribution was evaluated directly, and the marginal distribution of each source parameter was obtained by numerically integrating the full posterior distribution over the remaining source parameters. The marginal distributions are also shown in Figure 5.5. The stair-step appearance of the marginal x_s, y_s, z_s distributions is a remnant of the fact that the value of the distribution is only calculated at the center of each cell in the computational grid, and is not linearly interpolated at points in between. This appearance is not present in the graph for the q_s parameter because it is not considered discrete by the MCMC algorithm.

The marginal distribution corresponding to the streamwise source location, x_s, is obtained by integrating the full posterior PDF according to:

$$\begin{aligned} P(x_s \mid \mathbf{d}, I) &= \int_{\text{all } y_s} \int_{\text{all } z_s} \int_{\text{all } q_s} P(x_s, y_s, z_s, q_s \mid \mathbf{d}, I)\, \mathrm{d}y_s\, \mathrm{d}z_s\, \mathrm{d}q_s \\ &\approx \sum_{j=1}^{N_y} \sum_{k=1}^{N_z} \sum_{l=1}^{N_q} P(x_s, y_{s,j}, z_{s,k}, q_{s,l} \mid \mathbf{d}, I)\, \Delta y_s\, \Delta z_s\, \Delta q_s \end{aligned} \tag{5.26}$$

where the grid cell dimensions $\Delta x_s, \Delta y_s, \Delta z_s$ vary in space, but Δq_s is held constant. The marginal distributions for the other parameters are found similarly. The agreement between the MCMC results and the marginal distributions is good; small discrepancies are due to the parameter domain discretization used for the numerical integration.

The results shown in Figure 5.5 demonstrate that the Bayesian methodology provides an honest assessment of the source-receptor relationship in that the source strength is under-

predicted based on the available experimental concentration data and the performance of the dispersion model. The modelled concentrations near the centerline generally overpredict the measured concentrations and this is reflected in the fact that the methodology expects the source to be of lesser strength. The source position in the spanwise coordinate, z_s, is very well resolved, as the bulk of the probability mass lies in the width of a single cell of the computational grid. This is an ideal result because the C^* fields are not linearly interpolated, so the value of the marginal posterior distribution does not change within a single cell for a fixed source strength q_s. It is also interesting to note that the x_s histogram is smeared over most of the length of the canyon between the first and second rows of obstacles. This demonstrates that the rapid mixing which occurs in the canyon where the source is located erodes the quality of the inference made using data from detectors located in the canyons downstream.

5.4 Joint Urban 2003 atmospheric dispersion study

Mean concentration data were obtained at locations in and around downtown Oklahoma City, Oklahoma, US, during the Joint Urban 2003 atmospheric dispersion study which was conducted from June 28 to July 31, 2003 (Allwine et al., 2004). A sulfur hexafluoride (SF_6) tracer was released continuously for 30 minutes and sampled at several locations around the city. For the present case, we used nine sampler sites to perform the source reconstruction. Figure 5.6a shows the source and sampler locations with respect to the buildings present in the flow field. The dark contours surrounding the source (the area magnified in Figure 5.6b) represent the marginal posterior distribution for (x_s, y_s), obtained using direct evaluation of the full posterior PDF.

Before solving the inverse problem, the wind field was found for the domain shown in Figure 5.6a using the urbanSTREAM code of Lien et al. (2005a). The domain was subdivided into $98 \times 138 \times 68$ grid cells in the x, y, z directions, with the most refined cells located in the built-up area. The buildings in this area were explicitly resolved, while outside this area, the drag-force approach was used to simulate the effect of buildings on the flow by introducing a drag-force term into the spatially-averaged momentum equation (Lien et al., 2005b). For each detector location, the adjoint advection-diffusion equation was solved, resulting in a set of nine C^* fields. The problem is steady-state, so the wind field statistics do not change with time, and the source height (z-location) is assumed known. The units of distance in the x and y directions correspond to the system of units used internally by the flow solver, and the source strength q_s is measured in grams per second.

Detector measurements from the experiment were used for the d_i, and the model and measurement uncertainties were assigned realistic quantities based on the ability of the forward and adjoint dispersion simulations to correctly predict the detector concentrations. Depending on the detector, these uncertainties range from 2% to 70% of the mean concentration measured at the detector.

5.4.1 Results

In Figure 5.6b, the distribution of MCMC samples $(x_s^{\text{MCMC}}, y_s^{\text{MCMC}})$ is plotted as a set of points in contrast with a two-dimensional contour map of the marginal (x_s, y_s) distribution. The points occasionally stray from the bulk of the probability mass outlined by the contours, which is an indication that the Markov chain experiences difficulty traversing regions of very low probability, i.e., through buildings. In this case it is necessary to manually adjust MCMC algorithm parameters (e.g., increase the width of the proposal distribution) in order to encourage the chain to behave in a more exploratory fashion.

Both the MCMC histograms and marginal posterior distributions of the source parameters are shown in Figure 5.7. The parameter estimates in the table of Figure 5.7 show that the MCMC results are able to isolate the source parameters to within one standard deviation. In contrast to the MUST array case, the streamwise source coordinate y_s is well-estimated, with the bulk of the probability mass lying in the two grid cells which surround the source.

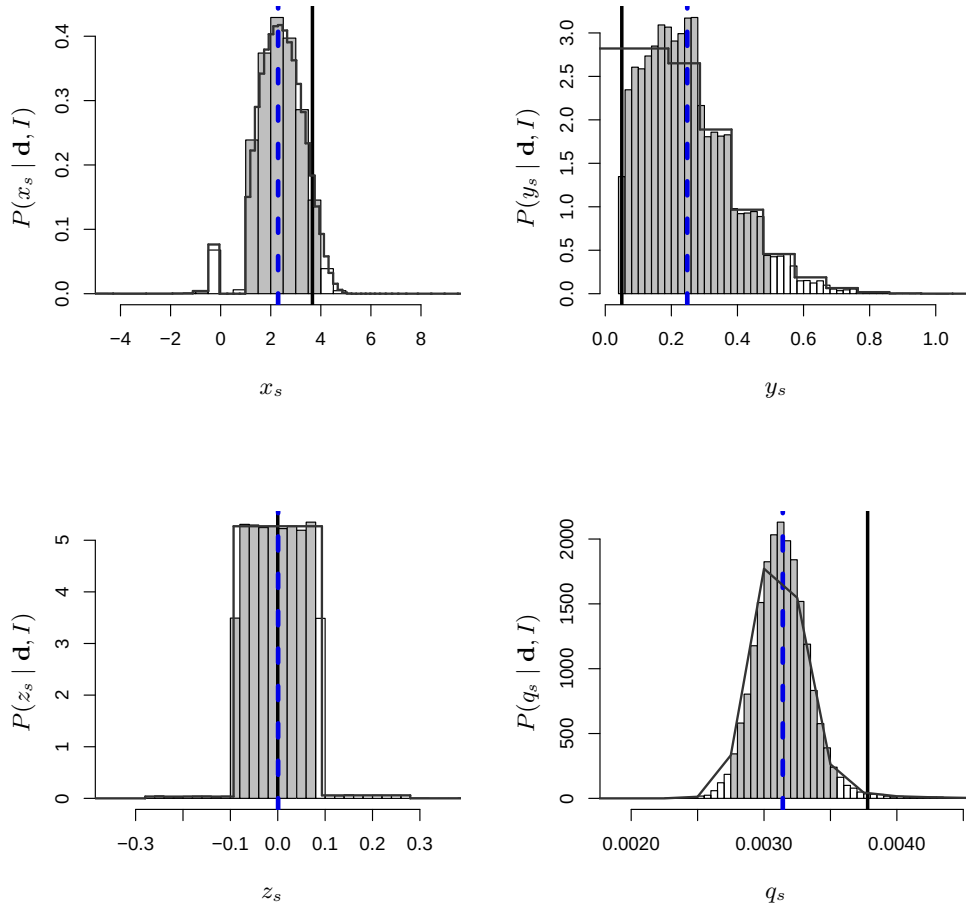
5.5 Conclusions

Bayesian probability theory has been successfully applied to solve the problem of source determination in a flexible and consistent way. By using the adjoint advection-diffusion equation along with MCMC sampling, calculations which yield an accurate picture of the full posterior distribution for the source parameters can be performed in a reasonable amount of time. Using these two techniques, the computational effort required for a fixed spatial domain size scales linearly with both the number of detectors and the number of source parameters. Without the adjoint approach, the computational effort would be proportional to the number of possible source locations, which is typically far greater than the number of detectors. Without using MCMC, the time required to sample from or directly evaluate the posterior PDF grows as a power of the dimensionality of the source parameters. The methodology is capable of simultaneously inferring at least four separate source parameter values: the location and strength are generally well estimated. The quality of the inference truthfully reflects the quality (in terms of the uncertainty) of the modelled and measured concentration data.

The present work assumes detector and model uncertainties to follow Gaussian distributions – alternative error distributions (such as lognormal) may be preferred depending on the available data and models. The primary source of the model uncertainty is our limited understanding of the physics of turbulent flows. Improving the turbulence modelling component of the flow solver and dispersion model would significantly improve the determination of the source location through a reduction in the discrepancy between the model dispersion predictions and the measured detector data.

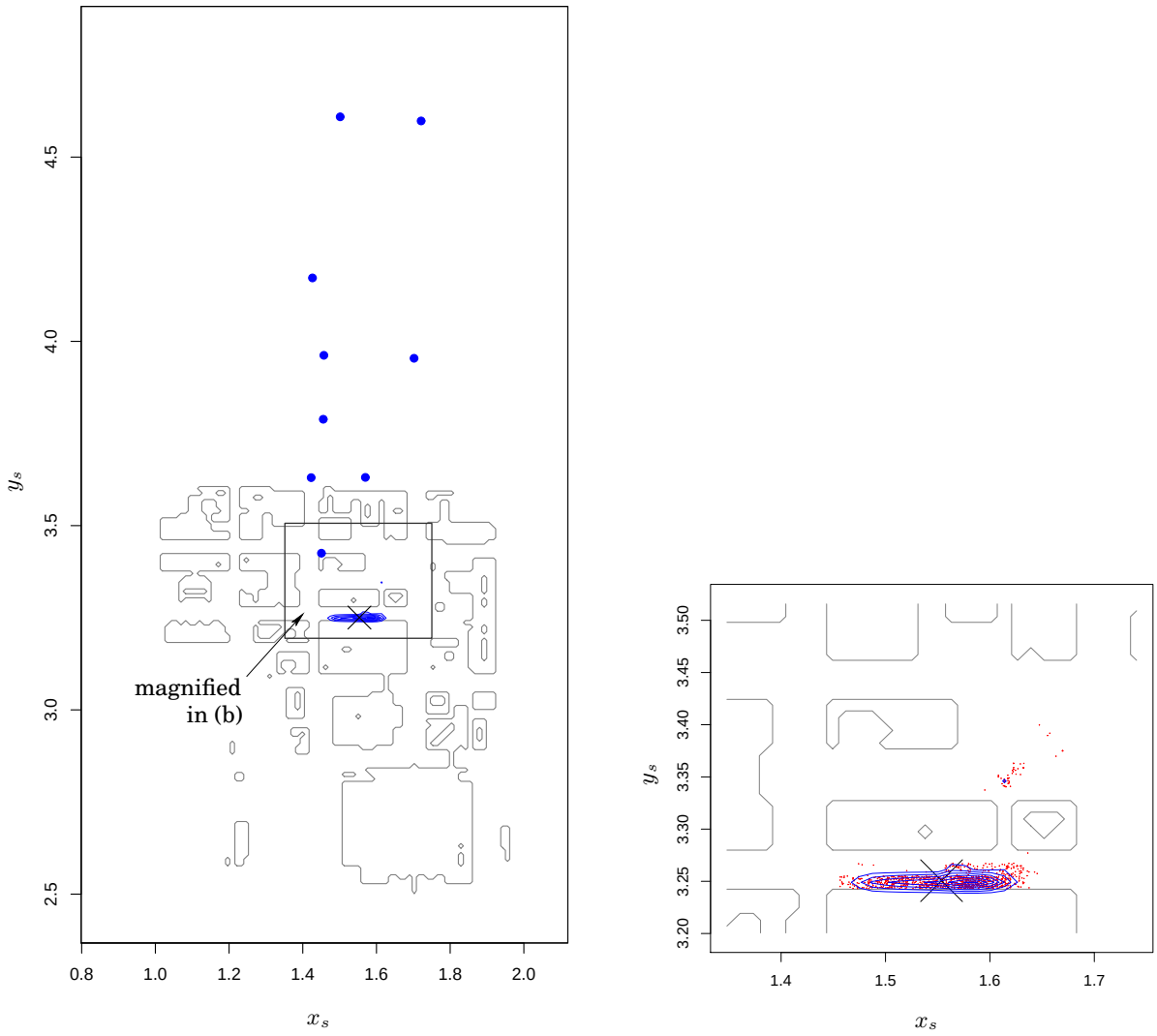
Both of the test cases considered involve dispersion in built-up areas and demonstrate the utility of the method for practical applications in emergency and environmental management. Generating the flow field is by far the greatest computational burden, and adding the ability

to perform source determination to an existing dispersion or CFD tool does not increase this burden significantly.



m_i	x_s	y_s	z_s	q_s
actual m_i	3.665	0.05	0.0	3.780×10^{-3}
mean(m_i^{MCMC})	2.294	0.248	0.001	3.141×10^{-3}
mean(m_i^{direct})	2.315	0.198	0.001	3.124×10^{-3}
sd(m_i^{MCMC})	1.127	0.134	0.060	2.123×10^{-4}
sd(m_i^{direct})	1.227	0.135	0.025	2.190×10^{-4}
95% HPD (m_i^{MCMC})	$[0.98, 4.21]$	$[0.05, 0.50]$	$[-0.09, 0.09]$	$[2.7, 3.5] \times 10^{-3}$

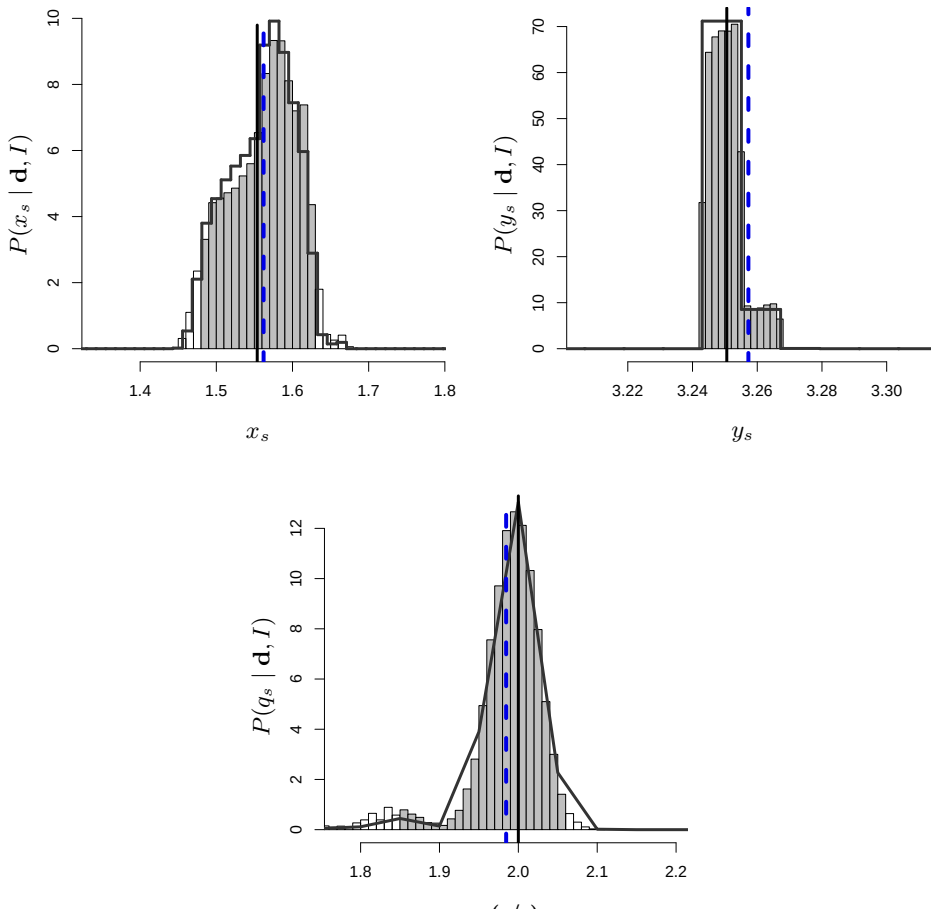
Figure 5.5: Marginal parameter distributions and summary statistics (mean and standard deviation) generated from both MCMC samples and direct marginalization of the posterior PDF. The histograms are generated from MCMC samples and the solid lines are generated using a direct calculation of the posterior distribution. The solid vertical line represents the true parameter value, and the dashed line is the mean of the MCMC samples. Shaded regions represent 95% HPD intervals based on the MCMC samples. The parameter q_s was non-dimensionalized based on the flow rate, reference length, and reference velocity associated with the experiment. The 95% credible [HPD] interval does not contain the true value of q_s; however, the 99% interval does.



(a) Overhead view of the buildings (light contours), source (cross) and detectors (circles) for Oklahoma City. The dark contours represent the marginal posterior distribution $P(x_s, y_s \mid \mathbf{d}, I)$. The prevailing wind direction is aligned with the positive y-axis.

(b) Close-up view of the marginal $P(x_s, y_s \mid \mathbf{d}, I)$ contours, overlaid on the MCMC samples (dots).

Figure 5.6: Oklahoma City building outlines overlaid with samples and contours from the marginal posterior distribution for the source x-y location.



m_i	x_s	y_s	q_s (g/s)
actual m_i	1.554	3.251	2.0
mean(m_i^{MCMC})	1.562	3.257	1.985
mean(m_i^{direct})	1.559	3.254	1.990
sd(m_i^{MCMC})	0.044	0.025	0.049
sd(m_i^{direct})	0.042	0.019	0.041
95% HPD (m_i^{MCMC})	[1.48, 1.63]	[3.24, 3.35]	[1.85, 2.06]

Figure 5.7: Source parameter estimates as determined using MCMC (histograms) and direct calculation of the marginal posterior distributions (dark lines). The true parameter value is shown by the solid vertical line, and the mean of the MCMC samples [mean(m_i^{MCMC})] is shown by the dashed line. Shaded regions represent 95% HPD intervals based on the MCMC samples.

Chapter 6

Determining the origin and decay rate of a nonconservative scalar

The material presented in this chapter is adapted from an earlier paper:

> A. Keats, E. Yee, and F-S. Lien. Efficiently characterizing the origin and decay rate of a nonconservative scalar using probability theory. *Ecological Modelling*, 205:437–452, 2007.

6.1 Introduction

Accurately predicting the dispersion of pollutants in the environment has important implications for both emergency and environmental management, and has been a topic of intensive study over the last several decades. Equally important is the inverse problem: determining the characteristics of the source of the pollutant, whether it be natural or anthropogenic, given a finite and noisy set of concentration measurements. Bayesian inference has recently gained popularity as a framework for solving these problems of parameter estimation and model selection; for example, Borsuk and Stow (2000) and Qian et al. (2003) estimated the magnitude, rate and reaction order of a biochemical oxygen demand model using experimentally measured wastewater data. Estimating rate parameters is not only useful for model selection and verification, but it can also be used to find evidence for specific types of chemistry which occur in the environment. Ariya et al. (1998) isolated the presence of halogen chemistry in the troposphere by examining ozone and nonmethane hydrocarbon (NMHC) depletion in the Arctic boundary layer. They applied linear regression to estimate the removal rates of various NMHCs and deduced a reaction mechanism based on information about Cl, Br and HO radical chemistry. The air mass under consideration remained stagnant which allowed the authors to treat it as a 'smog chamber' reactor. In contrast, the present work explicitly considers the advection of species and would be useful in situations involving both transport and chemistry.

Assuming the existence of a single model used to represent the source (or sources), the problem of source determination becomes one of parameter estimation in which the parameters describing the source could include its strength (emission rate), location, rate of decay in the environment, and if the source is transient, its turn-on and turn-off times. Hanna et al. (1990) used Eulerian-based models to estimate the source strength for the Project Prairie Grass (PPG) experiments (Barad, 1958; Haugen, 1959), while Flesch et al. (1995) used a backward-time Lagrangian stochastic (LS) model to estimate the emission rate of a sustained surface area source in horizontally homogeneous turbulence. The work of Flesch et al. is important because it presents the backward Lagrangian stochastic model used in the present research. A similarly motivated but more involved study was carried out by Lin et al. (2003) who developed a backward LS model for determining surface CO_2 fluxes from aircraft measurements made in the planetary boundary layer. The LS models described by Flesch et al. and Lin et al. both take a 'receptor-oriented' approach to determining detector concentrations, an approach which is also adopted here. However, they do not exploit statistics related to particle travel times for treating potentially nonconservative tracers.

The source apportionment problem[1] was addressed by Skiba (2003) who used an adjoint pollution transport model (the Eulerian equivalent of the backward Lagrangian stochastic model) to identify industries operating in violation of emissions regulations. In this case, a limited set of possible source locations were known a priori, whereas in the present work, no such assumptions need be made about the position of the source. Penenko et al. (2002) and Liu et al. (2005) use a similar method to perform a sensitivity analysis and risk assessment for populated areas which could potentially suffer from the effects of a chemical or radiological accident.

The Bayesian methodology we have adopted for this work is flexible in that any number or type of source parameters may be considered for estimation; however, the main contribution of this research is the attachment of a statistical method (described in Section 6.3.3) to a Lagrangian stochastic dispersion model. This permits the efficient estimation of the first-order decay rate of a dispersed tracer. Therefore, for simplicity we consider only a single point source which continuously emits material into a statistically stationary and horizontally homogeneous atmospheric surface layer. Although this may seem like a special case of limited relevance, the method can easily be generalized to account for transient sources and wind fields. Furthermore, many practical transport-related problems in the field of ecological modelling can be addressed or at the very least approximated by assuming a continuous release of a contaminant into a statistically stationary and horizontally homogeneous (turbulent) flow field. In certain wind conditions, the assumption of stationarity (independence of mean variables and turbulence statistics on time) may be considered valid, which allows tracer transport (emanating from a briefly sustained source) over small distances (100m – 1000m) to be modelled using a continuous release. For example, the assumption of a continuous release and horizontal homogeneity of the turbulent wind field was adopted by Meyers et al. (1998) who inferred dry deposition rates for SO_2, O_3, HNO_3, as well as particulate matter. In fact, whereas Meyers et al. used eddy correlation methods to estimate the deposition rates, the

[1] In the source apportionment problem, we are aware of the locations of a number of pre-existing sources and must estimate the relative magnitudes of their emissions.

method presented in this work could also potentially be used (directly or by augmenting a separate experimental procedure) in the estimation of deposition wherever it can be modelled as a first-order decay process.

It is important to distinguish the present work from studies which have been done on global atmospheric transport inversion, such as Rödenbeck et al. (2003) and the TransCom studies, e.g., Denning et al. (1999). While it is true that they incorporate backward trajectories (or adjoint equations) and Bayesian inference techniques, these investigations are driven by global-scale transport models with the objective of determining surface fluxes (of CO_2 and SF_6) and do not require near-field models to estimate location or reaction rate parameters.

Contaminant source identification in groundwater flows has been addressed by a number of authors, including Aral et al. (2001) who used an optimization approach (based on genetic algorithms) to infer the release history and source location of a contaminant. Michalak and Kitanidis (2002) adopted a Bayesian approach and used Markov chain Monte Carlo to sample the posterior distribution for the source parameters. While both investigations share the goal of inferring the source location of a contaminant, they differ from the present work in that they did not consider scenarios in which the first-order decay coefficient of the contaminant was unknown.

Nonconservative tracers (which decay or grow in mass over time either through mechanical, chemical or photolytic processes) represent an important subset of dispersion cases. A naïve treatment of these cases will result in computationally challenging and data-intensive calculations. Accordingly, in this chapter we present the efficient numerical solution of an inverse problem in which parameters describing the source location and strength, as well as the rate of tracer transformation (growth or decay), are simultaneously estimated. In the next section, we formulate the solution to the source determination problem in terms of the comprehensive probabilistic expression for the source parameters. Efficiently evaluating and interpreting this expression involves techniques which are described in Section 6.3. In Section 6.4, we validate the overall methodology using concentration measurements made during Project Prairie Grass (where the scalar was assumed to be conservative), and also against data obtained from a solution to the forward problem using a Lagrangian stochastic model of short-range dispersion in the atmospheric surface layer for a nonconservative scalar.

6.2 Bayesian problem formulation

In this chapter, we consider the following vector of source parameters:

$$\mathbf{m} = (x_s, y_s, z_s, q_s, k_s) \ , \tag{6.1}$$

where $\{x_s, y_s, z_s\}$ represent the spatial location of the source, q_s is its strength (of dimension $[MT^{-1}]$), and k_s $[T^{-1}]$ is the rate of tracer transformation. Bias and inaccuracy in the experimental and numerical data (e.g., measured concentration data are subject to experimental

uncertainty, while modelled concentration measurements are affected by the accuracy of the numerical model) are taken into account by using PDFs[2].

The posterior distribution expresses the plausibility of all possible hypotheses (a single hypothesis consists of a single set of values for the source parameters) and, when evaluated for a specific set of source parameters, is a scalar quantity whose domain of definition has the same dimensionality as $\mathbf{m}$. Therefore, for an n-dimensional problem where each parameter is rendered into s discrete values, the entire posterior distribution might be represented by s^n numbers. For high-dimensional $\mathbf{m}$, this may be an impossibly large number of data points to calculate, which motivates the use of the MCMC technique for exploring only the significant regions of the posterior PDF. Furthermore, MCMC draws samples (i.e., selects sample parameter values) from the posterior PDF without requiring the evidence term as a normalization constant. The relationship between the posterior, prior and likelihood PDFs can then be simplified:

$$\underbrace{P(\mathbf{m} \mid \mathbf{d}, I)}_{\textit{Posterior}} \propto \underbrace{P(\mathbf{m} \mid I)}_{\textit{Prior}} \underbrace{P(\mathbf{d} \mid \mathbf{m}, I)}_{\textit{Likelihood}} . \tag{6.2}$$

The bulk of the time required to calculate the value of the posterior PDF for a single hypothesis is determined by the calculation of the likelihood function, which relates modelled to measured concentration data. Using a backward Lagrangian particle model in conjunction with an inventory of averaged particle travel times significantly mitigates this effort; these techniques are described in detail in Section 6.3.

6.2.1 Assignment of the likelihood function

Both the physical concentration measurements and the theoretical source-receptor relationship are subject to uncertainties, which are assumed to have the following properties:

1. We adopt the basic assumption that the measurement error for detector i can be characterized as additive Gaussian noise with root-mean square (RMS) experimental error $\sigma_{D,i}$.

2. The model error associated with the source-receptor relationship may be characterized in a similar way, having RMS error $\sigma_{R,i}$.

3. Both the measurement and the model errors are independent; i.e., measurements at one detector do not affect measurements at another detector, and measurement errors do not affect model errors at each detector location.

[2] In this work, we represent errors by simple variances and do not account for bias in the measurement and model uncertainties. However, the present work could easily be extended to account for bias by adding additional offset (bias) variables to the Gaussian PDF. These variables might initially be unknown and could later be removed using marginalization.

Under these assumptions, the likelihood function takes the same form as in the previous chapter:

$$P(\mathbf{d} \mid \mathbf{m}, I) \propto \exp\left[-\frac{1}{2}\sum_i \frac{(d_i - r_i(\mathbf{m}))^2}{\sigma_{D,i}^2 + \sigma_{R,i}^2}\right]. \tag{6.3}$$

6.2.2 Assignment of the prior probabilities

For the present case, we assume a state of ignorance[3] with respect to each of the parameters. Ignorance regarding the location and decay parameters, $\{x_s, y_s, z_s, k_s\}$, is expressed using a uniform distribution:

$$P(x_s \mid I) = P(y_s \mid I) = P(z_s \mid I) = P(k_s \mid I) = \text{constant}, \qquad \mathbf{m} \in \Omega\,, \tag{6.4}$$

and the remaining parameter, q_s, is assigned a prior which remains invariant under transformations of scale[4]:

$$P(q_s \mid I) \propto q_s^{-1}, \qquad q_s \in [q_{\min}, q_{\max}]\,. \tag{6.5}$$

Using a scale-invariant prior ensures that $P(q_s \mid I) = P(aq_s \mid I)$ for any constant a (Jaynes, 2003). The interval $[q_{\min}, q_{\max}]$ ensures that the prior PDF is normalizable; in practice, $q_{\max}$ is chosen to be some finite, reasonable upper bound.

6.2.3 The posterior probability density function

The posterior PDF is proportional to the product of the prior and the likelihood:

$$\begin{aligned} P(\mathbf{m} \mid \mathbf{d}, I) &\propto P(\mathbf{m} \mid I) P(\mathbf{d} \mid \mathbf{m}, I) \\ &\propto \mathcal{I}(\mathbf{m} \in \Omega) \frac{1}{q_s} \exp\left[-\frac{1}{2}\sum_i \frac{(d_i - r_i(\mathbf{m}))^2}{\sigma_{D,i}^2 + \sigma_{R,i}^2}\right], \end{aligned} \tag{6.6}$$

where $\mathcal{I}(\blacksquare)$ denotes the indicator function.

6.3 Modelling and numerical approach

Numerically predicting r_i, the modelled concentration at the i^{th} receptor, requires the use of a model which, when given a specific source configuration (location, strength and decay rate), is capable of providing a concentration value at each of the detector (receptor) locations. Rather than running a forward dispersion model for every possible combination of source parameters, using a backward (or adjoint) dispersion model requires less computational time when

[3] Since we do not consider hypotheses where parameters lie outside of the computational domain, our state of ignorance is not total.

[4] The scale-invariant prior was also discussed in Section 3.2.2.

the number of detectors is significantly less than the number of possible source locations. Lagrangian stochastic particle dispersion models are routinely applied in the field of meteorology to simulate the dispersion of species in environmental flows. Backward Lagrangian models are structurally very similar to their corresponding forward models (Seibert and Frank, 2004) and are used in the present work to generate the required dual (adjoint) concentration fields.

6.3.1 Source-receptor relationship

In this chapter, we consider an ideal continuous point source of the form:

$$Q = q_s \delta(\mathbf{x} - \mathbf{x}_s) \ . \tag{6.7}$$

Q is a source density distribution $[ML^{-3}T^{-1}]$ which releases material continuously at a steady rate of q_s $[MT^{-1}]$ from location $\mathbf{x}_s$. The mean concentration field C resulting from this release can be found using a forward dispersion model, which relates Q to C through the linear operator $\mathbb{L}$ in the following way:

$$\mathbb{L}\, C = Q \ . \tag{6.8}$$

The definition of the operator $\mathbb{L}$ is flexible; in the Eulerian framework, Equation (6.8) becomes the steady advection-diffusion equation,

$$\mathbf{U} \cdot \nabla C - \nabla \cdot (K \nabla C) = Q \ , \tag{6.9}$$

where

$$\mathbb{L}(\blacksquare) \equiv \mathbf{U} \cdot \nabla(\blacksquare) - \nabla \cdot \big(K \nabla(\blacksquare)\big) \ . \tag{6.10}$$

In a Lagrangian framework, $\mathbb{L}$ effectively describes a forward Lagrangian stochastic (fLS) particle model:

$$C(\mathbf{x}) = \int_{\mathcal{R}} G(\mathbf{x} \mid \mathbf{x}_0)\ Q(\mathbf{x}_0)\ d\mathbf{x}_0 \ , \tag{6.11}$$

where $G(\mathbf{x} \mid \mathbf{x}_0)$ is the integral kernel of $\mathbb{L}^{-1}$ and is a function of the specific LS model chosen. In LS models, $G(\mathbf{x} \mid \mathbf{x}_0)$ represents a transition probability density.

The adjoint operator, $\mathbb{L}^*$, relates a dual (or adjoint) concentration field (viz., C^* field) to a 'detector response' function, h $[L^{-3}]$. For the i^{th} detector, the relationship is:

$$\mathbb{L}^*\, C_i^* = h_i \ , \tag{6.12}$$

where $h = h(\mathbf{x} - \mathbf{x}_d)$ models the detector response function of a receptor which measures the concentration at location $\mathbf{x}_d$. The function h acts as a spatial filter and would be, e.g., a delta function for an ideal detector with infinite resolving power. In an Eulerian framework, Equation (6.12) becomes the adjoint advection-diffusion equation. In a Lagrangian framework, $\mathbb{L}^*$ describes a backward Lagrangian stochastic (bLS) model. The C^* field has units of $[TL^{-3}]$ and can be interpreted as a residence-time density field. Here, we assume that the release is continuous and the flow is statistically stationary, so transient terms are absent in both $\mathbb{L}$ and $\mathbb{L}^*$.

The source-receptor relationship is used to obtain r_i, the modelled concentration value at the location of the i^{th} detector. r_i is defined through the inner product:

$$r_i = \langle C, h_i \rangle \equiv \int_{\mathcal{R}} C \; h_i \; d\mathcal{R} \; , \tag{6.13}$$

where $\mathcal{R}$ defines the spatial domain (or a space-time domain if the problem were transient). The duality relationship, defined by Equation (6.14), connects the forward and adjoint operators and provides an alternative way to calculate the source-receptor relationship:

$$\langle C, \mathbb{L}^* \, C^* \rangle = \langle \mathbb{L} \, C, C^* \rangle \tag{6.14}$$

$$\Rightarrow \quad r_i = \langle C, h_i \rangle = \langle Q, C_i^* \rangle \equiv \int_{\mathcal{R}} Q \; C_i^* \; d\mathcal{R} \; . \tag{6.15}$$

For a point source, the inner product reduces to a simple multiplication:

$$r_i(\mathbf{m}) = q_s C_i^*(x_s, y_s, z_s, k_s) \; . \tag{6.16}$$

In general, one C^* field must be generated per receptor, with the response function h treated as a 'source'. Once all of the C^* fields have been generated, the source-receptor relationship can be rapidly calculated using Equation (6.16) for any combination of the parameter values $\{\mathbf{x}_s, q_s\}$. Varying the tracer decay rate, k_s, introduces difficulties which are addressed in Section 6.3.3.

6.3.2 Forward and backward Lagrangian stochastic dispersion model

Lagrangian stochastic (LS) particle models provide an alternative to Eulerian methods for simulating the dispersion of a tracer in a wind flow. Whereas Eulerian methods directly calculate concentration fields using the advection-diffusion equation discretized over a grid of fixed locations, LS methods track individual 'particles' or 'parcels of fluid' through a flow field and generate a set of particle trajectories which can then be manipulated to yield concentration fields (and other information).

Both the fLS and bLS models used in the present research calculate particle trajectories by solving for velocity increments du_i which evolve according to the Langevin equation (Rodean, 1996):

$$du_i = a_i(\mathbf{x}, \mathbf{u}, t)dt + b_i(\mathbf{x}, \mathbf{u}, t)dW_t \; , \tag{6.17}$$

where dW_t denotes an increment of the standard Wiener process; $\mathbf{x}$ is a vector indicating the position of the particle; $\mathbf{u}$ is the Lagrangian velocity vector of the particle, and $(u_1, u_2, u_3) = (u, v, w)$ are the streamwise, spanwise and vertical components of the velocity vector. Particle positions are calculated using the following equation:

$$dx_i = u_i dt \; . \tag{6.18}$$

The a_i terms govern the deterministic component of the particle trajectory and represent a combination of:

1. Damping coefficients which relax the velocity increments du_i back toward the mean flow;
2. A 'drift correction' term which satisfies the 'well-mixed criterion':

> If a species of passive "marked particles" is initially mixed uniformly in position and velocity space in a turbulent flow, it will stay that way (Thomson, 1987). In other words, if the concentration of a species is initially uniform in a flow, it will remain uniform if there are no sources or sinks for the species (Rodean, 1996).

In the present model, the a_i are functions of velocity u_i, Reynolds stresses τ_{ij}, and their respective z-derivatives, and the dissipation rate ϵ. With respect to the test cases addressed later in this chapter, parameterizations for the mean wind profile, Reynolds stresses and dissipation rate can be found in Section 6.4.1. The b_i terms govern the stochastic component of the trajectory and represent acceleration increments generated by random pressure fluctuations with very short correlation times. The present work assumes horizontally homogeneous and statistically stationary flow in the atmospheric surface layer, for which functional forms of the a_i and b_i coefficients for Gaussian turbulence consistent with the 'well-mixed criterion' and Kolmogorov's theory of local isotropy are:

$$\begin{aligned} a_1 = &\left[-\frac{C_k\epsilon}{2}[\lambda_{11}(u_1-U_1)+\lambda_{13}u_3]+\frac{\partial U_1}{\partial x_3}u_3+\frac{1}{2}\frac{\partial \tau_{13}}{\partial x_3}\right] \\ &+\left[\frac{\partial \tau_{11}}{\partial x_3}[\lambda_{11}(u_1-U_1)+\lambda_{13}u_3]+\frac{\partial \tau_{13}}{\partial x_3}[\lambda_{13}(u_1-U_1)+\lambda_{33}u_3]\right]\frac{u_3}{2}\ , \end{aligned} \tag{6.19a}$$

$$a_2 = \left[-\frac{C_k\epsilon}{2}(\lambda_{22}u_2)+\frac{\partial \tau_{22}}{\partial x_3}(\lambda_{22}u_2)\frac{u_3}{2}\right]\ , \tag{6.19b}$$

$$\begin{aligned} a_3 = &\left[-\frac{C_k\epsilon}{2}[\lambda_{13}(u_1-U_1)+\lambda_{33}u_3]+\frac{1}{2}\frac{\partial \tau_{33}}{\partial x_3}\right] \\ &+\left[\frac{\partial \tau_{13}}{\partial x_3}[\lambda_{11}(u_1-U_1)+\lambda_{13}u_3]+\frac{\partial \tau_{33}}{\partial x_3}[\lambda_{13}(u_1-U_1)+\lambda_{33}u_3]\right]\frac{u_3}{2}\ , \end{aligned} \tag{6.19c}$$

$$b_i = (C_k\epsilon)^{1/2}\ , \tag{6.19d}$$

where the λ_{ij} are the components of the inverse Reynolds stress tensor τ_{ij}^{-1}:

$$\lambda_{11} = (\tau_{11}-\tau_{13}^2/\tau_{33})^{-1}\ , \tag{6.20a}$$

$$\lambda_{22} = \tau_{22}^{-1}\ , \tag{6.20b}$$

$$\lambda_{33} = (\tau_{33}-\tau_{13}^2/\tau_{11})^{-1}\ , \tag{6.20c}$$

$$\lambda_{13} = (\tau_{13}-\tau_{11}\tau_{33}/\tau_{13})^{-1}\ , \tag{6.20d}$$

and the Reynolds stresses are $\tau_{ij} = \overline{(u_i-U_i)(u_j-U_j)} = \overline{u_i'u_j'}$, where U_j are components of the Reynolds averaged wind velocity. For the horizontally homogeneous wind field used later

in this chapter, the flow direction is aligned with the x-axis, so $U_2 = U_3 = 0$. Furthermore, in the atmospheric surface layer, $\tau_{12} = \tau_{23} = 0$. The explicit Euler scheme is used for time integration, with the time step chosen to be significantly smaller than the integral Lagrangian time scale:

$$\Delta t = 0.01\tau_L(z) \ , \tag{6.21}$$

$$\tau_L(z) = \frac{2\overline{w'^2}}{C_k\epsilon(z)} \ , \tag{6.22}$$

$$\text{with} \quad \epsilon(z) \propto \frac{1}{z} \ , \tag{6.23}$$

where $C_k = 4.8$ is a constant. General expressions (including the changes necessary for implementation in a backward LS model) for the a_i and b_i terms (e.g., for cases where flows and turbulence are not homogeneous) are given by Flesch et al. (1995) and Yee (2008). Rodean (1996) lists a number of alternate values that various authors have proposed for the constant C_k, and Yee and Wilson (2007) discuss issues related to the stability of both the dynamical and time integration scheme.

With respect to the fLS model, dispersion from a continuous source is modelled by releasing a large number N_p of particles from the source location, with each particle's initial 'pseudo-mass' representing a fraction of the source strength q_s. As particles spend time in the flow field, they may undergo transformations which alter their pseudo-mass. These mechanisms are modelled according to a first-order decay process which is described in Section 6.3.3. For the moment, we consider that the j^{th} particle's pseudo-mass (viz., source strength fraction), q_j, is a function of the amount of time, t_j, that it has spent in the field:

$$q_j = q_{j,0}\, f(k_s, t_j) \ , \tag{6.24}$$

where $q_{j,0} = q_s/N_p$ is the particle's pseudo-mass at $t_j = 0$, and k_s is a coefficient characterizing the decay process. Similarly, with reference to the bLS model, N_p particles are released from each of the detector locations. Since the detector response function, h, integrates to unity, each particle's initial dual pseudo-mass is $q^*_{j,0} = 1/N_p$, and the decay process is modelled in the same way:

$$q^*_j = q^*_{j,0}\, f(k_s, \tau_j) \ . \tag{6.25}$$

The temporal frame of reference has been reversed (particles originate from detectors and are travelling backward through the flow field), so $\tau_j = -t_j + T$, where T is an arbitrary transformation constant.

In order to obtain a discrete representation of the particle trajectory defined by Equations (6.17) and (6.18), particle locations are recorded at discrete time intervals. Between the locations $\mathbf{x}_j^{(n)}$ and $\mathbf{x}_j^{(n+1)}$, and within a volume enveloping $\{\mathbf{x}_j^{(n)}, \mathbf{x}_j^{(n+1)}\}$, the j^{th} particle spends a 'residence time', $\Delta\tau_j = \tau_j^{(n+1)} - \tau_j^{(n)}$. To first-order accuracy, the individual contribution of the

particle to the residence-time density in a grid cell whose centroid is x can be obtained using:

$$C^*(\mathbf{x}, k_s) = \sum_{j:\mathbf{x}_j \in \mathcal{D}} \psi(\mathbf{x}_j; \mathbf{x})\, q_j^*\, \Delta\tau_j \ , \tag{6.26}$$

where $\psi(\mathbf{x}_j, \mathbf{x})$ is a mass-conserving kernel function with finite support over spatial domain $\mathcal{D} \subset \mathcal{R}$. It assigns a relative contribution to the C^* value in the cell based on the distance between the j^{th} particle's position $\mathbf{x}_j$ and the center of the grid cell. The bandwidth (a measure of the region of support) of this kernel is typically on the order of the edge length of a grid cell.

6.3.3 Tracer decay treatment

The species being dispersed is assumed to undergo transformation (decay or growth) by, e.g., reaction, radiological decay, or scavenging, which can be modelled by the first-order mechanism:

$$\frac{dC}{dt} = -k_s C \ , \tag{6.27}$$

where k_s is a rate constant with units of $[T^{-1}]$. Here, we assume k_s to be positive for transformations in which the concentration of the tracer decays over time as it is transported. The solution to Equation (6.27) is:

$$C(t) = C_0 \exp(-k_s t) \ , \tag{6.28}$$

where C_0 is the concentration at time $t = 0$. The adjoint decay mechanism is then modelled using:

$$\frac{dC^*}{d\tau} = -k_s C^* \ , \tag{6.29}$$

whose solution is:

$$C^*(\tau) = C_0^* \exp(-k_s \tau) \ , \tag{6.30}$$

assuming that particles were released from the detector at time $\tau = 0$. Note that $\tau > 0$ refers to earlier times relative to $\tau = 0$.

With respect to the bLS model, 'tagging' the j^{th} particle with its 'accumulated travel time' τ_j enables us to rapidly quantify its expected transformation for any k_s:

$$q_j^*(\tau_j) = q_{j,0}^* \exp(-k_s \tau_j) \ , \tag{6.31}$$

where $q_{j,0}^* \equiv q_j^*(\tau_j = 0)$. Treating dual pseudo-masses (as opposed to dual concentrations) in this manner is appropriate given the linearity of Equation (6.29).

Complications arise when we attempt to calculate the C^* value in a grid cell through which a large number N of particles have passed. Figure 6.1 demonstrates how different [dual] particles travelling upstream from a given detector could take different amounts of time to reach the same grid cell. Strictly speaking, the correct C^* value for the cell is obtained by re-calculating Equation (6.31) using the desired k_s value for all N particles, and then substituting q_j^* into Equation (6.26). However, this requires us to maintain and manipulate lists

of all the particles generated from each detector by the bLS model, which would be computationally intractable and excessively data-intensive for large simulations consisting of millions or billions of particles moving in flow fields which might evolve over time. Fortunately, by

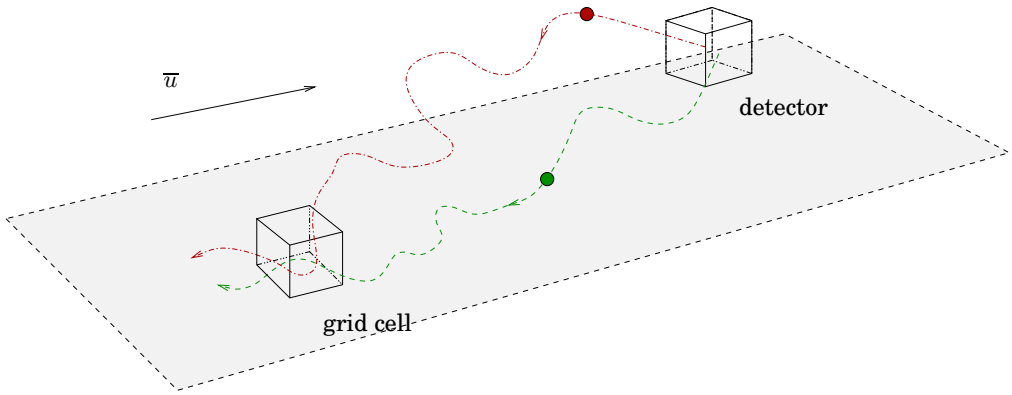


Figure 6.1: Dual particles are 'released' from detectors and travel upstream through the wind field, generating retroplumes. A number of particles will pass through the same grid cell (a potential source location) and will have taken different lengths of time to travel there, by virtue of the stochastic nature of their trajectories.

phrasing the problem in a statistical sense, i.e., by considering the distribution of the individual particle travel times τ_j, we can estimate $C^*(\mathbf{x}, k_s)$ with an accuracy determined by the value of k_s together with the properties of the distribution of the τ_j. In Section 6.4.4 we assess the computational savings made through the use of the statistical method, relative to solving the problem exactly using all available trajectory information.

In order to simplify the analysis of our estimate for $C^*(\mathbf{x}, k_s)$, we will assume a very basic kernel (with a top-hat function) which arithmetically averages both the travel times and dual pseudo-masses of particles passing though the domain $\mathcal{D}$ defined by a single grid cell centered on x. Assuming a kernel of this form results in simple expressions for the sample mean travel time, $\hat{\tau}$, and the conservative C^* field value, C_0^*:

$$\psi(\mathbf{x}_j, \mathbf{x}) = \frac{1}{\Delta x \Delta y \Delta z} \qquad \mathbf{x}_j \in \mathcal{D} \ , \tag{6.32}$$

$$\hat{\tau}(\mathbf{x}) = \frac{1}{N} \sum_{j:\mathbf{x}_j \in \mathcal{D}} \tau_j \ , \tag{6.33}$$

$$C_0^*(\mathbf{x}) = \frac{1}{\Delta x \Delta y \Delta z} \sum_{j:\mathbf{x}_j \in \mathcal{D}} q_{j,0}^* \, \Delta\tau_j \ . \tag{6.34}$$

If we assume that the only information available describing the distribution of particle travel times, τ_j, is their mean and variance, then the principle of maximum entropy (Jaynes, 2003) asserts that the maximally non-committal (least informative) PDF used to describe the particle travel times should be the Gaussian distribution. Given that the τ_j are obtained directly from the Lagrangian stochastic particle model, obtaining their mean and standard deviation is straightforward.

The true C^* value in a given grid cell is a consequence of the decaying pseudo-mass, q^*, and is obtained directly using the arithmetic mean of the exponentiated decay coefficients:

$$C^*(\mathbf{x}, k_s) = C_0^* \frac{1}{N} \sum_{j:\mathbf{x}_j \in \mathcal{D}} \exp\left(-k_s \tau_j\right) \ . \tag{6.35}$$

Maintaining lists of individual τ_j is impractical, so we must find a way to estimate C^* using the sample mean and standard deviation of the τ_j. Assuming that the individual particle travel times τ are distributed normally,

$$\tau \sim N(\hat{\tau}, \sigma_\tau) \ , \tag{6.36}$$

where $N(\hat{\tau}, \sigma_\tau)$ is a normal (Gaussian) distribution with mean $\hat{\tau}$ and standard deviation σ_τ, then the exponentiated decay coefficients are distributed log-normally:

$$\phi \equiv \exp\left(-k_s \tau\right) \sim LN\left(-k_s \hat{\tau}, k_s \sigma_\tau\right) \ , \tag{6.37}$$

with probability density function given by

$$P(\phi) = \frac{1}{k_s \sigma_\tau \phi \sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\log(\phi) + k_s \hat{\tau}}{k_s \sigma_\tau}\right)^2\right) \ . \tag{6.38}$$

Here, $LN(-k_s\hat{\tau}, k_s\sigma_\tau)$ is a log-normal distribution such that the logarithm of the random variate results in a Gaussian distribution whose mean and standard deviation are $-k_s\hat{\tau}$ and $k_s\sigma_\tau$. The mean of this log-normal distribution provides the following estimate for C^*:

$$\hat{C}^*(\mathbf{x}, k_s; \hat{\tau}, \sigma_\tau) = C_0^* \exp\left(-k_s \hat{\tau} + \frac{1}{2} k_s^2 \sigma_\tau^2\right) \ . \tag{6.39}$$

Before proceeding to use this estimate, it is worthwhile to examine its accuracy in view of the fact that the limited availability of computational power constrains the number of particle trajectories that can be simulated (within a reasonable amount of time) during a given bLS model run. Consider a detector for which the bLS model is used to generate a corresponding C^* field. With increasing upstream distance from this detector, particle trajectories are spread thinly over more grid cells, resulting in lower individual cell particle counts (and in turn lower dual concentration [C^*] values). For grid cells experiencing low tagged particle counts, the law of large numbers does not necessarily guarantee the accuracy of the mean particle travel time, $\hat{\tau}$, which might vary significantly across several different realizations of the same (in the parametric sense) bLS model run. Since the estimate, $\hat{C}^*$, is a function of $\hat{\tau}$, the impact of variability in $\hat{\tau}$ on the variability of $\hat{C}^*$ must be analyzed.

Based on the earlier assumption that particle travel times are normally distributed, their sum, and thus the estimated mean travel time, $\hat{\tau}$, is also normally distributed[5],

$$\hat{\tau} \sim N\left(\bar{\tau}, \sigma_\tau/\sqrt{N}\right) \ , \tag{6.40}$$

where $\bar{\tau}$ is the 'true' mean travel time, and the standard deviation is $\sigma_\tau/\sqrt{N}$, where N is the number of particles passing through a given grid cell. N could change for different bLS model realizations, but here we assume for simplicity and with no loss in generality that it remains the same. As with τ, the fact that $\hat{\tau}$ is normally distributed leads to a log-normal distribution for the estimate, $\hat{C}^*$:

$$\begin{aligned} \hat{C}^*(\mathbf{x}, k; \hat{\tau}, \sigma_\tau) &= C_0^* \exp\left(-k\hat{\tau} + \frac{1}{2}k^2\sigma_\tau^2\right) \\ &\sim LN\left(-k\bar{\tau} + \log\left[C_0^* \exp\left(\frac{1}{2}k^2\sigma_\tau^2\right)\right], \frac{k\sigma_\tau}{\sqrt{N}}\right) \ , \end{aligned} \tag{6.41}$$

whose standard deviation is:

$$\mathbf{sd}(\hat{C}^*) = C_0^* \exp\left(\frac{k^2\sigma_\tau^2}{N}\left(\frac{N+1}{2}\right) - k\bar{\tau}\right)\left(\exp\left(\frac{k^2\sigma_\tau^2}{N}\right) - 1\right)^{\frac{1}{2}} \ . \tag{6.42}$$

It is clear from Equation (6.42) that as k and σ_τ decrease, and as N increases, the standard deviation of the sample mean decreases. However, for certain values of k, $\bar{\tau}$, σ_τ and N, $\hat{C}^*$ may vary significantly. This variance could be treated as part of the overall model (theoretical) uncertainty, σ_R, information which can be encoded into the likelihood function. It should also be noted that the above analysis assumes that the variability in $\hat{\tau}$ as encoded in σ_τ is known exactly; in practice, σ_τ must be estimated from the sample of N particles that 'move' through the given grid cell.

6.4 Short-range dispersion in the atmospheric surface layer

Here we apply the source determination methodology to a test case whose wind field and geometry match those used for Project Prairie Grass (PPG), a benchmark tracer dispersion experiment that was conducted over flat terrain with no obstacles. This experiment is described in detail in the original reports (Barad, 1958; Haugen, 1959), and more recently by other authors such as Venkatram and Du (1997), and Hanna et al. (2004).

In PPG, sulfur dioxide (SO_2) was released from a small tube placed 46 cm above the ground. Seventy 20-minute releases were conducted during July and August 1956, in a wheat field near O'Neil, Nebraska. The wild hay was trimmed to a uniform height of 5 to 6 cm. Samplers were positioned on concentric semi-circular arcs centred on the release, at downwind distances of 50, 100, 200, 400, and 800 m. The samplers were positioned 1.5 m above

[5] For more general distributions of τ, Equation (6.40) is only true in the limit as $N \to \infty$, but for the present case, the relationship is also valid for small N.

the ground, and provided 10-minute (averaged) concentration values. Towers for measuring vertical profiles of mean concentration were also available along the arc with a radius of 100 m.

After stating the parameterization used to define the wind field (Section 6.4.1), we present a reference solution based on part of the actual PPG experiment (Section 6.4.2). This is followed by a validation of the statistical tracer decay treatment and a discussion of its performance. We validate the overall source determination methodology in two stages. First, in Section 6.4.5.1 the source reconstruction approach is tested using real concentration data measured during the PPG experiment (in which the scalar was considered to be conservative). During the source reconstruction approach, it is assumed that the rate of decay is unknown. In the second stage (Section 6.4.5.2), the reconstruction approach is applied to two sets of synthetic concentration data generated using a forward Lagrangian stochastic model operating under the same atmospheric conditions as PPG, with decay of particle mass being modelled by the first-order mechanism described by Equation (6.27). These synthetic measurements are then chosen to play the role of $\mathbf{d}$ in the inverse problem, and the bLS model is applied to generate the required C^* fields.

It should be noted that the Lagrangian stochastic particle model described in this chapter has already been validated against PPG using parameterized wind statistics, appropriate for a horizontally homogeneous neutrally-stratified atmospheric surface layer (or, adiabatic wall shear layer), for the case of a passive, conservative tracer (Wilson et al., 1981).

6.4.1 Wind field

The wind field is fully-developed and horizontally homogeneous, so all velocity and turbulence statistics are functions of z (height above the ground surface) only. The mean wind velocity is aligned with the x-axis. The fLS and bLS models require the wind field to be supplied in terms of its mean velocity and turbulence statistics. For the present case, the wind field can be described analytically by semi-empirical relationships developed for a horizontally homogeneous neutrally-stratified surface layer. Parameterizations of wind statistics also exist for describing non-neutral (e.g., stably stratified and convective) boundary layers, but they are not considered in this work. The components used to describe the turbulent wind field are outlined below. These expressions are commonly used for LS models applied to the surface layer, and are similar to those found in Flesch et al. (1995) and Rodean (1996). They are parameterized in terms of u_*, the friction velocity, and z_0, the roughness length.

Mean wind velocity profile

The average wind speed in the x (streamwise) direction is assumed to follow a log-law profile in the surface layer:

$$U_1(z) \equiv \frac{u_*}{\kappa} \ln \frac{z}{z_0} \,, \tag{6.43}$$

where $\kappa \approx 0.4$ is von Kármán's constant. The mean y and z velocity components ($\overline{v}$ and $\overline{w}$) are both zero.

Velocity variances

The velocity variances, $\overline{u'^2}$, $\overline{v'^2}$, $\overline{w'^2}$, and the covariance, $\overline{u'w'}$, are constant within the surface layer:

$$\begin{aligned} \overline{u'^2} = \overline{v'^2} &= 4.5u_*^2 \ , \\ \overline{w'^2} &= 1.69u_*^2 \ , \\ \overline{u'w'} &= -u_*^2 \ . \end{aligned} \tag{6.44}$$

It is assumed that the velocity covariances $\overline{u'v'}$ and $\overline{v'w'}$ vanish in the surface layer.

Dissipation rate

The turbulence kinetic energy dissipation rate is determined as follows:

$$\epsilon(z) = \frac{u_*^3}{\kappa z} \ . \tag{6.45}$$

6.4.2 Reference solution: forward dispersion

The fLS model was used to simulate dispersion under the conditions encountered in Run 24 of the PPG experiment. Using a roughness length of $z_0 = 0.006$ m, a friction velocity of $u_* = 0.38$ m s^{-1}, and a source strength of $q_s = 41.2$ g s^{-1}, particle trajectories were used to generate a three-dimensional concentration field over a $420 \times 200 \times 10$ m^3 domain. A horizontal slice of this domain ($z = 1.5$ m), along with the locations of the source and detectors, is shown in Figure 6.2. The concentration field was generated for a grid of cells of dimension $\Delta x = \Delta y = \Delta z = 1$ m.

Along each arc, detectors were spaced at 2° intervals, and the streamwise flow direction (aligned with the x-axis) was determined using the maximum concentration measurement. Experimentally measured concentration data for each arc (at radii $r = \{50, 100, 200, 400\}$ m from the source) are plotted in Figure 6.3 together with concentration profiles generated with the fLS model using decay coefficient values of $k_s = \{0, 0.03\}$ s^{-1}.

6.4.3 Validation: tracer decay treatment

Before proceeding to solve the inverse problem (Section 6.4.5), we first examine the suitability of the tracer decay treatment which was presented in Section 6.3.3. Without loss of generality, consider the detector located at $(x, y, z) = (400, 0, 1.5)$. The C^* field (or 'retroplume') emanating upwind from this detector was calculated by using the bLS model to determine the trajectories

of 1×10^5 tagged particles. For the six grid cell locations marked in Figure 6.4, tagged particle travel times were recorded and used to generate the normal probability plots shown in Figure 6.5. The Kolmogorov-Smirnov test was applied to each set of travel times. P-values for each set except τ_a met or exceeded 0.1. Despite τ_a failing the test, we consider the assumption of normality to be vindicated by the high quality of the statistical tracer decay approximation in regions near to the source. This property is quantified in Figure 6.7.

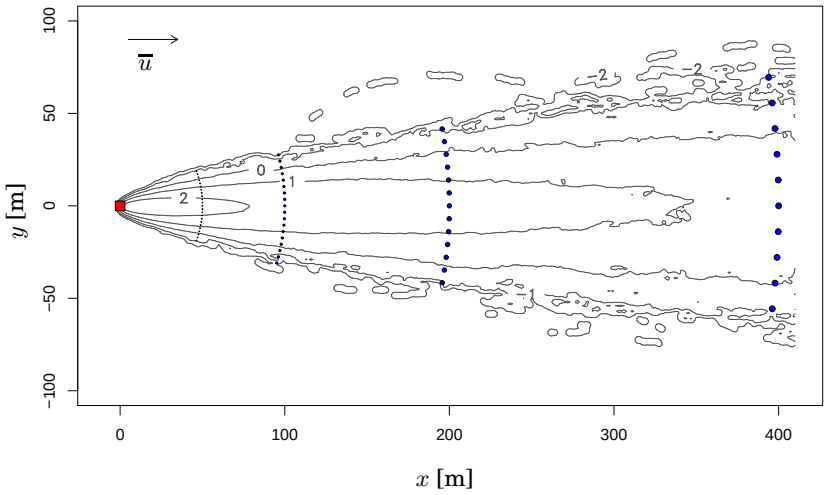


Figure 6.2: Arrangement of the point source (square) and detector arcs (circular dots) for Project Prairie Grass. The detectors shown measured non-zero concentrations during the PPG experiment. Contours of $\log_{10}(C\ [\mu\text{g/m}^3])$ obtained using the fLS model ($k_s = 0, z = 1.5$ m) are also plotted.

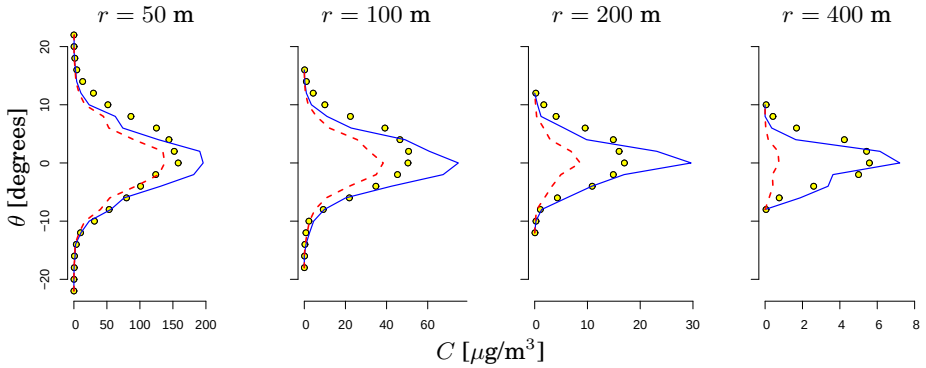


Figure 6.3: Experimentally measured circumferential concentrations (circles), simulated concentration profiles (solid line, obtained using fLS model), and decayed concentration (dashed line, $k_s = 0.03$ s^{-1}).

Having assessed the validity of assuming normally-distributed particle travel times, we turn our attention to estimating the relative error incurred by using particle travel time

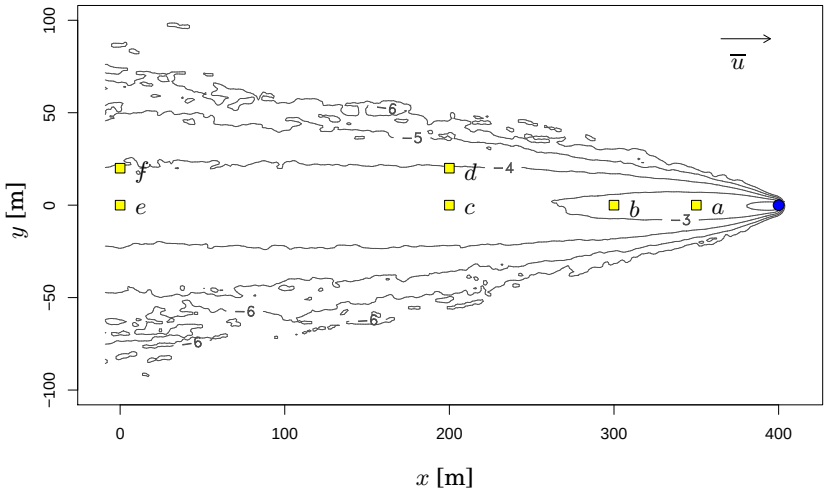


Figure 6.4: Detector (circle) and grid cell locations (squares labelled a–f) in which particle travel times were binned. Contours of $\log_{10}(C^*)$ obtained using the bLS model ($k_s = 0, z = 1.5$ m) are plotted in the background.

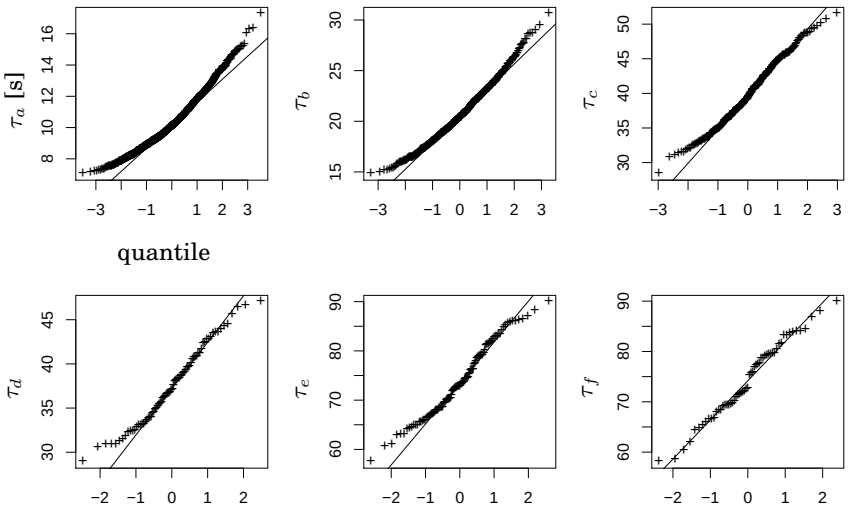


Figure 6.5: Normal probability plots of particle travel times recorded at the grid cells shown in Figure 6.4.

statistics (mean and variance) to approximate the 'true' C^* field[6]. The estimate, $\hat{C}^*$, is defined by Equation (6.39). We define the absolute value of the relative error incurred by the approximation as:

$$\mathcal{E}_{C^*} = \frac{\mid \hat{C}^* - C^* \mid}{C^*}, \quad C^* > 0 \, . \tag{6.46}$$

For the case of the C^* field shown in Figure 6.4, the error term $\mathcal{E}_{C^*}$ was calculated in all cells (where $C^* > 0$) throughout the three-dimensional domain $\mathcal{R}$ for the cases of $k_s = \{0.03, 0.3\}$ s^{-1}. Histograms showing the distribution of $\mathcal{E}_{C^*}$ over the domain as a whole are presented in Figure 6.6. For the present test case, the error is clearly significant for the larger

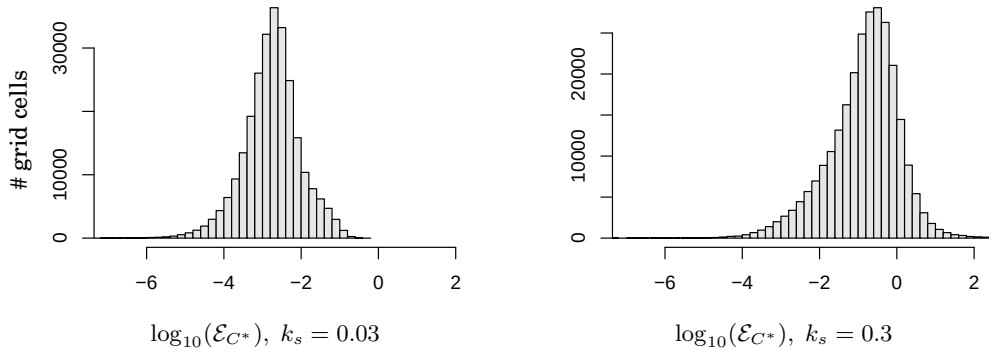


Figure 6.6: Histograms of the error incurred by the approximation, Eq. (6.39). The error is evaluated once per grid cell over the three-dimensional spatial domain.

value of k_s, and Figure 6.7 demonstrates that $\mathcal{E}_{C^*}$ generally grows with upstream distance and, by extension, increasing particle travel time.

While Figures 6.6 and 6.7 characterize $\mathcal{E}_{C^*}$ specifically with respect to the estimate of the C^* field of Figure 6.4, it remains desirable to characterize $\mathcal{E}_{C^*}$ for more general cases. When considering bLS simulations that are not problem-specific, the standard deviation of $\hat{C}^*$ as defined by Equation (6.42) is indicative of the accuracy of the approximation. This quantity depends on k_s, $\bar{\tau}$, σ_τ and N, but can be characterized effectively by approximating the ratio $\sigma_\tau/\bar{\tau}$ using a constant value. For the bLS test case outlined above, Figure 6.8 presents the distribution (over all grid cells) of this ratio, and shows a pronounced mean value of approximately 0.10.[7] In Figure 6.9, we plot contours of sd$(\hat{C}^*)$ as a function of k_s and N, conservatively assuming that the ratio $\sigma_\tau/\bar{\tau} = 0.15$. For low N and large k_s, sd$(\hat{C}^*)$ increases drastically, indicating that bLS simulations involving high decay rates should be made more accurate by increasing the number of particles released from the detector (in the hope of increasing N, the number of particles passing through the grid cell).

[6] By 'true', we refer to a C^* field calculated using Eqn. (6.35), not necessarily a field generated using a large enough number of particles to ensure small statistical error.

[7] Empirical observation suggests that this figure remains more or less constant for different scenarios.

We recapitulate that the approximation $\hat{C}^*$ is based on the assumption that particle travel times are normally distributed, which leads to the consequence that decayed pseudo-masses q^* are distributed log-normally. Alternatively, if we assume that pseudo-masses q^* are in fact normally distributed, numerical experiments show that accuracy of the approximation suffers ($\mathcal{E}_{C^*}$ grows, and the histograms of Figure 6.6 are shifted to the right by a significant amount).

6.4.4 Performance of the statistical tracer decay treatment

The above analysis has shown that the statistical tracer decay approximation is valid under the condition that the ratio of the average particle travel time to the decay coefficient is relatively low. Applying this statistical treatment to an existing LS model requires only a little

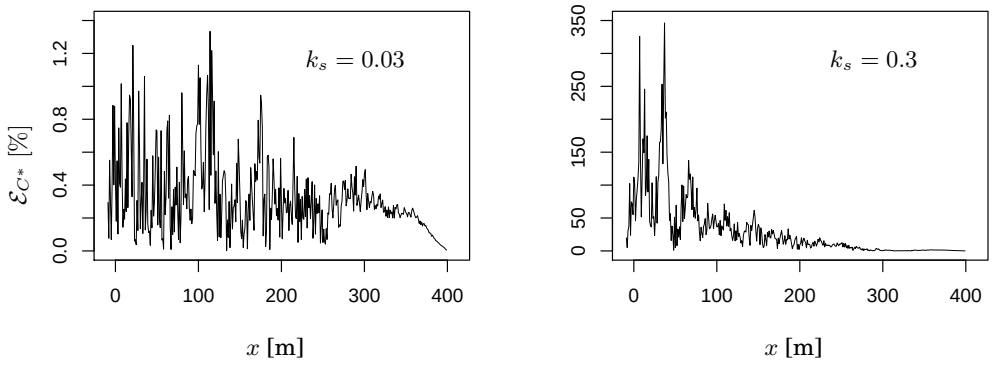


Figure 6.7: The error $\mathcal{E}_{C^*}$ as a percentage, evaluated along the centerline of the C^* field ($y = 0, z = 1.5$), upstream of the detector.

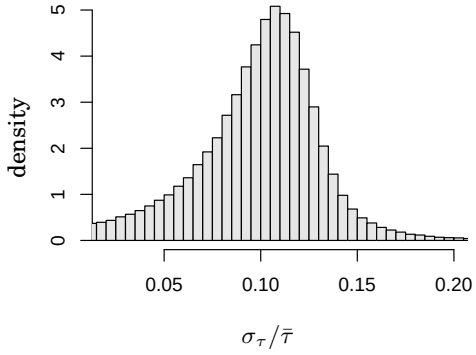


Figure 6.8: Histogram of the ratio of travel time standard deviation (σ_τ) to mean particle travel time ($\bar{\tau}$) for all grid cells in the PPG domain.

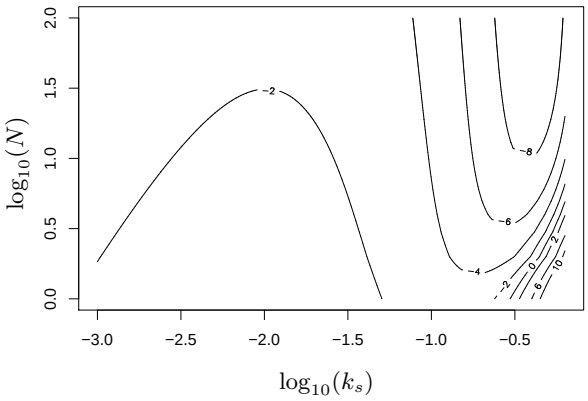


Figure 6.9: Contours of $\log_{10}\left(\mathrm{sd}(\hat{C}^*)\right)$ for varying k_s and N. Here, $\sigma_\tau/\bar{\tau} = 0.15$, and $\bar{\tau} = 100$ s.

code modification and results in a much faster and more memory-efficient calculation of the expected dual concentration in a grid cell, compared to using an exact approach in which all particle trajectory information would have to be retained.

Due to the considerable variation in computer storage techniques and data structures, there is no point in performing a side-by-side comparison of two LS codes, one which incorporates the statistical treatment and another which does not. Instead, we consider how the required CPU time and memory (storage) requirements relate to typical LS model parameters governing the length and spatiotemporal resolution of a dispersion simulation.

Consider a bLS dispersion model run in which N_p particles are released from a single detector. Each of these N_p particles follows a trajectory which is recorded either on disk or in memory by saving the individual particle's position ($\mathbf{x}_j$), pseudo-mass (q_j^*), and cumulative travel time (τ_j). Trajectory data is built up by saving this positional information at every time step, and we can assume that the average particle spends N_t time steps in the domain. Thus, an LS simulation will typically write $(N_p\, N_t)$ entries in a trajectory file.

The desired end product of a bLS simulation is usually a C^* field. Consider such a field, generated from a trajectory file using the kernel described in Equations (6.32–6.34), and discretized over a Cartesian grid which contains $(N_x N_y N_z)$ grid cells. In order to calculate the hypothetical concentration expected by the detector using Equation (6.16), we extract the value of the C^* field at a potential source location (x_s, y_s, z_s). When the rate of tracer decay (k_s) is known a priori, a C^* field need be generated only once from the trajectory data using the first-order decay equation (6.31). However, in the present scenario, k_s is unknown, which means that a new C^* field (or at the very least, the C^* value at all potential source locations) must be recalculated for many possible values of k_s.

Because LS models are inherently stochastic, N_p is required to be very large, large enough to generate reasonably smooth C^* (or concentration, in the fLS case) data on what may be a high resolution grid of the problem domain. For a given grid cell, especially if it is near the

plume centerline, a very high number N of particles will pass through, contributing to the C^* value in the cell. By way of illustration, the examples addressed in this work use $N_p = 1 \times 10^5$ and for many grid cells, N is on the order of 1000 particles. Larger-scale problems solved over higher-resolution grids may result in the generation of a thousand times as much trajectory data. It should be noted that the C^* fields which are generated from trajectory data are, for the present example, on the order of $1/1000$ the size of the trajectory data. In general, it is desirable to avoid manipulating raw trajectory data once it has been used to generate gridded data fields.

With this background in mind, Table 6.1 summarizes the savings in computational time and memory requirements obtained by using the statistical decay treatment. From a practical standpoint, three fields must be generated from the trajectory data in order to use the statistical approximation: $C_0^*(\mathbf{x})$ (a C^* field for which $k_s = 0$); $\hat{\tau}(\mathbf{x})$ (the average particle travel time for each grid cell); and $\sigma_\tau^2(\mathbf{x})$ (the variance of the particle travel times). However, once these fields have been generated, it is no longer necessary to store or manipulate the raw trajectory data.

Computational task	**Exact approach**	**Statistical treatment**
Trajectory data storage	$\propto N_p N_t$	0
Field storage	$\propto N_x N_y N_z$	$\propto 3 \times N_x N_y N_z$ (for each of $C_0^*, \hat{\tau}, \sigma_\tau^2$)
Data retrieval required for $C^*(\mathbf{x}_s, k_s)$ calculation	List of N particles which passed through grid cell centred on $\mathbf{x}_s$	3 array entries: $C_0^*(\mathbf{x}_s)$, $\hat{\tau}(\mathbf{x}_s)$, $\sigma_\tau^2(\mathbf{x}_s)$
Calculation of $C^*(\mathbf{x}_s, k_s)$	N evaluations of $\exp(-k_s \tau_j)$	1 evaluation of $C_0^* \exp(-k_s \hat{\tau} + \frac{1}{2} k_s^2 \sigma_\tau^2)$

Table 6.1: Comparison of computational effort required by the exact vs. statistical tracer decay treatments.

6.4.5 Inverse problem: source determination

We begin by assessing the performance of the overall source determination methodology using concentration data measured during the PPG experiment, Run 24. In this case, the tracer was considered conservative ($k_s = 0$); however, we have included k_s as an unknown parameter to be estimated. Following this assessment, in Section 6.4.5.2 we apply the methodology to two problems involving synthetic, decayed ($k_s > 0$) concentration data generated using the fLS model. For all three test cases we assume the following parameter bounds for the

computational domain Ω:

$$\begin{aligned} x_s &\in [-10, 410] \text{ m} , & q_s &\in [1, 200] \text{ g s}^{-1} , \\ y_s &\in [-100, 100] \text{ m} , & k_s &\in [-1, 1] \text{ s}^{-1} , \\ z_s &\in [0, 10] \text{ m} . \end{aligned} \tag{6.47}$$

6.4.5.1 PPG experimental data (conservative tracer)

The experimentally measured concentration data used here is extracted from a subset of the arc-based measuring stations shown in Figure 6.2. Four stations from each arc were used and are shown in Figure 6.10. In the $\{50, 100\}$ m arcs, detectors are located off-centerline by $\{-6, -2, 2, 6\}^\circ$. In the $\{200, 400\}$ m arcs, detectors are located off-centerline by $\{-4, -1, 1, 4\}^\circ$.

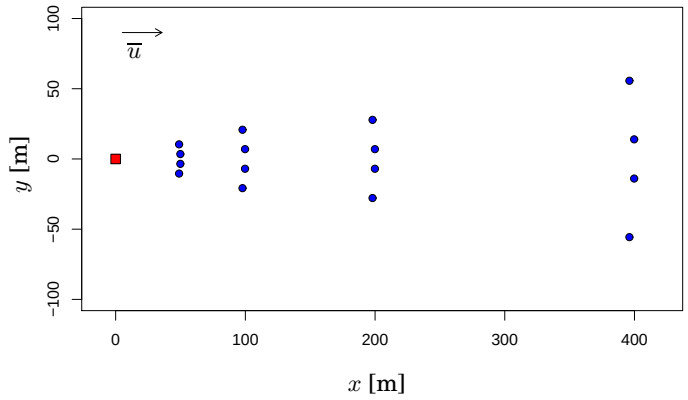


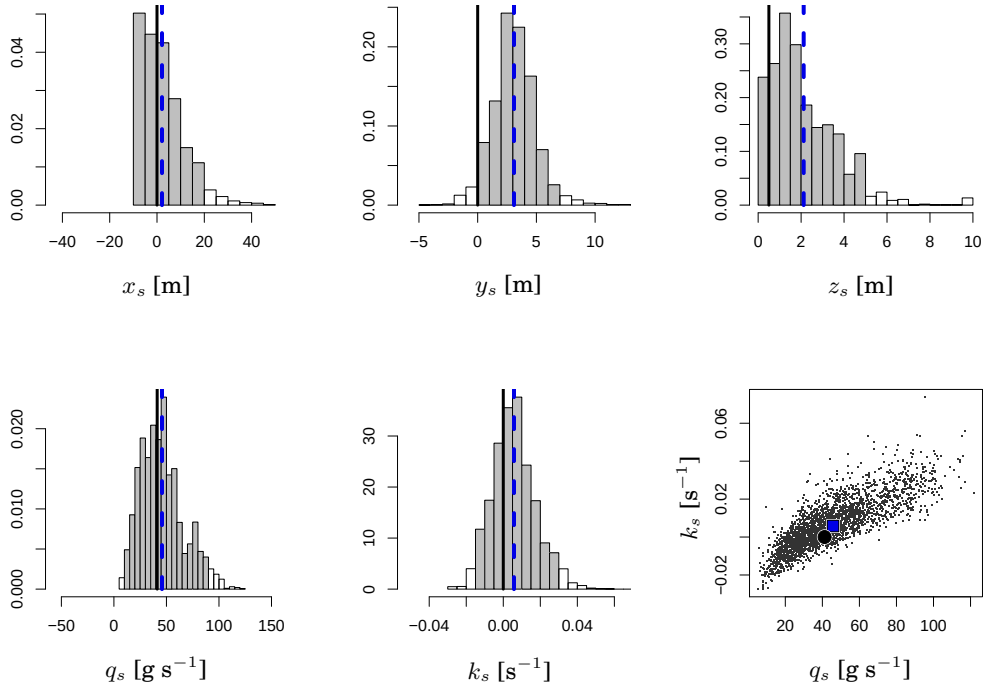
Figure 6.10: Layout, Case 1 (measured data). Unknown source (square) and detector (circular dots) arrangement for determining the source parameters.

The combined uncertainty in the model and measurement noise for each detector was conservatively set such that the model results shown in Figure 5.4 lie within 3 standard deviations of the measurements. This led to the assignment of $\{100, 50, 50, 50\}\%$ of the mean concentration measured at the detectors in each of the four $\{50, 100, 200, 400\}$ m arcs, respectively. Dual concentration (C^*), and particle travel time mean and variance ($\bar{\tau}$ and σ_τ^2) fields were generated for one of the rightmost detectors using the bLS model (10^5 particles were released), and were subsequently translated in space to all of the other detectors in the array. This translation is admissible due to the horizontally homogeneous nature of the flow encountered in PPG. Vertical translation of the C^* field is inadmissible owing to the vertical inhomogeneity of the wind statistics, and for general flows which lack homogeneity in any one direction, C^* fields and particle travel time statistics cannot be translated.

The posterior PDF was sampled using the Metropolis-Hastings algorithm, and MCMC samples for each parameter were binned. 10^5 points were generated using normal proposal distributions whose width was decided based on observations of each chain's progress. His-

tograms and corresponding summary statistics for the MCMC samples are shown in Figure 6.11. The scatter plots in the lower-right corner of the set of histograms show the MCMC samples drawn in $\{q_s, k_s\}$ parameter space. A clear positive correlation is evident in the spread and density of these samples, indicating that difficulty could potentially be encountered when attempting to isolate both source strength and decay rate simultaneously. Detector measurements must be relatively unambiguous in their representation of the effects of decay rate and source strength in order for the posterior PDF to yield meaningful information about these two parameters.

In this case, all parameters are estimated such that the true values are enclosed within at most two standard deviations of the means of the MCMC samples. This includes the decay rate coefficient, which was known a priori to be zero.



m_i	x_s [m]	y_s [m]	z_s [m]	q_s [g s^{-1}]	k_s [s^{-1}]
actual m_i	0.0	0.0	0.46	41.20	0.0
mean(m_i^{MCMC})	2.11	3.09	2.12	45.78	0.006
sd(m_i^{MCMC})	9.12	1.84	1.58	20.53	0.012
95% HPD (m_i^{MCMC})	$[-9.47, 19.17]$	$[-0.44, 6.62]$	$[0.01, 4.85]$	$[12.12, 88.01]$	$[-0.015, 0.031]$

Figure 6.11: Case 1: Marginal parameter distributions and summary statistics generated from MCMC samples. The true parameter value is represented by the solid vertical line in the histograms, and the circular dot in the scatter plot. The mean of the MCMC samples is represented by the dashed vertical line in the histograms, and the square dot in the scatter plot. Shaded regions represent 95% HPD intervals based on the MCMC samples.

6.4.5.2 Synthetic data (nonconservative tracer)

In Figures 6.12 and 6.13, detector arrangements are shown for two source determination test cases where the unknown decay rate k_s differs by an order of magnitude. For the first case, in which the decay rate is very low, detectors which provide meaningful data (given their susceptibility to noise) are generally located closer to the source. By 'meaningful data' we refer to data points which are in general representative of the decay and dispersion of the plume. Hence, the streamwise spread of the detectors in the second case (Figure 6.13) is approximately half that of those for the first case (Figure 6.12). In both cases, the detector array is asymmetric about the plume centerline, with all detectors placed at a height of $z = 1.5$ m (as in the PPG field experiment). Synthetic concentration data were obtained for the two decay rates using the concentration field generated by the fLS model after releasing 5×10^4 particles. This data was then subjected to additive Gaussian noise whose standard deviation was 50% of the measured concentration.

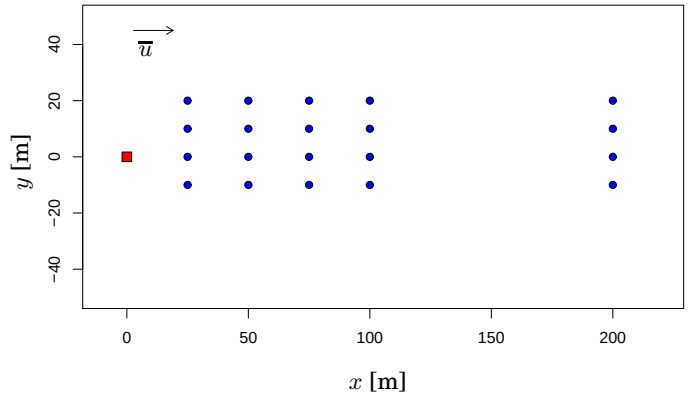


Figure 6.12: Layout, Case 2. Unknown source (square) and detector (circular dots) arrangement for determining the source parameters (for small k_s).

For both cases, uncertainties at all detectors were assumed to be 50%, and the same C^* field was generated and horizontally translated in the x-y plane to each of the synthetic detector locations. The posterior PDF was sampled using the same MCMC approach as with case 1. Histograms and summary statistics for the second case ($k_s = 0.03$) are shown in Figure 6.14, and the results of the third case ($k_s = 0.30$) are shown in Figure 6.15. The scatter plots of the MCMC samples drawn from the $\{q_s, k_s\}$ parameter space once again demonstrate a positive correlation. In both cases, parameters are generally well estimated; the true parameter values are enclosed within two standard deviations of the means of the MCMC samples.

6.5 Conclusions

Combining the following techniques;

1. a statistical approach to reconstructing the [dual] concentration field for a given decay coefficient (in LS particle models);

2. the adjoint approach; and

3. Markov chain Monte Carlo

results in a computationally efficient method for solving the source determination problem for a nonconservative tracer in a Bayesian probabilistic framework.

The detector positions used in the three test cases might be construed to be arranged based on prior knowledge of the source location, as opposed to being randomly spread about the domain (as might be expected in a real-life scenario). However, our choice of detector arrangement is designed to elicit knowledge about the capabilities and limitations of the source determination methodology. In a random array, detectors upwind of the source would be expected to measure zero concentration, information which would effectively constrain the potential source location (but not the decay rate or the strength). It could be considered that we have utilized a 'subset' of detectors which sample only part of the plume (downwind of the source), which actually results in a more challenging inference.

A prior understanding of the expected scale of tracer decay is clearly important when considering an inverse problem in which the decay coefficient and source strength could vary by orders of magnitude. For the case of a tracer undergoing rapid decay, detectors will yield useful information (in terms of their signal-to-noise ratio) only when placed relatively close to the source. Lagrangian stochastic simulations must be run using relatively large numbers of particles in order to improve model concentration estimates for detectors which lie far from the source. Conversely, for tracers which decay slowly in time, the spread of detectors must be wide enough to capture the relative behaviours of dispersion (indicative of the source strength) and decay. The MCMC samples presented in the previous section demonstrate that while

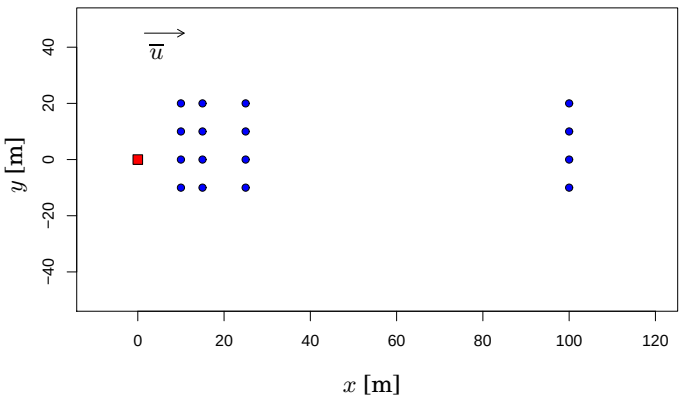
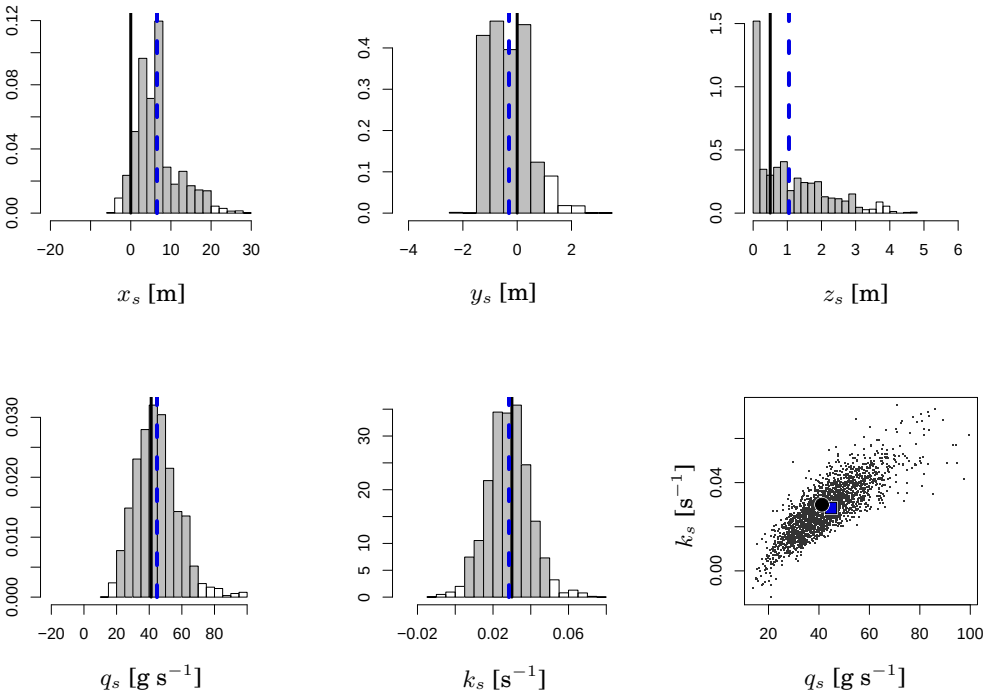


Figure 6.13: Layout, Case 3. Unknown source (square) and detector (circular dots) arrangement for determining the source parameters (for large k_s).

these parameters are closely correlated (i.e., the concentration measured by a single detector could be reduced either by reducing the source strength or by increasing the decay rate), they can nevertheless be simultaneously estimated (using multiple detectors), since they are not linearly dependent. The k_s vs. q_s MCMC sample scatter plots shown in Figures 6.11–6.15 match the trend seen in the analogous plot of Rate coefficient vs. Ultimate biochemical oxygen demand presented by Qian et al. (2003).

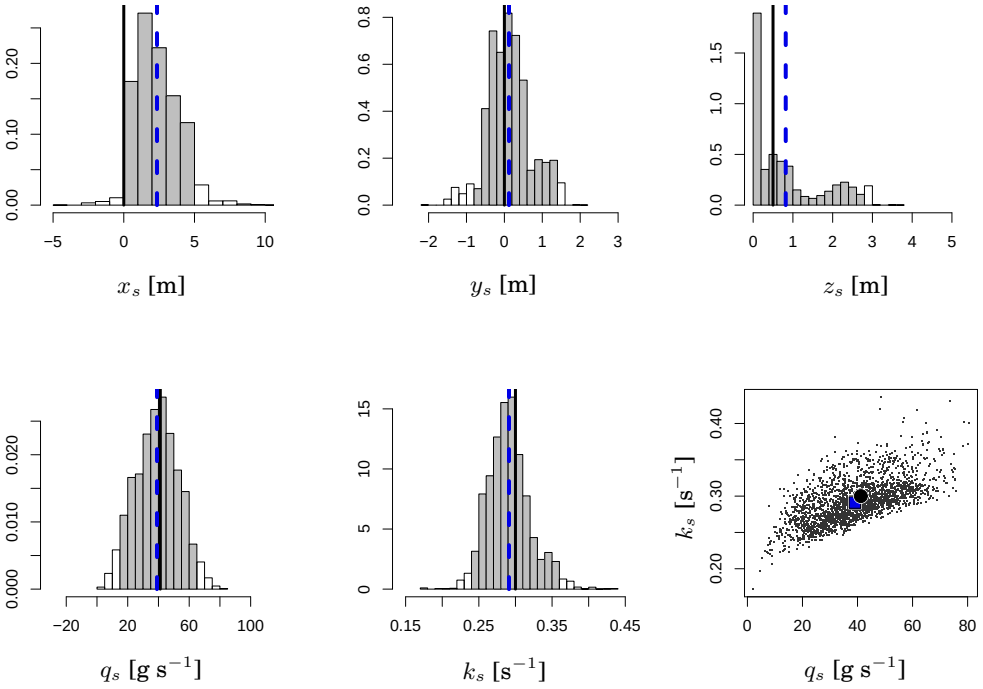
The test cases demonstrate that the method can be applied to environmental flows in which several of the source parameters are unknown. Consider a scenario where tracer decay or scavenging is unexpected, but experimentally unconfirmed. Using an inference procedure based on the 'overparameterized' model (i.e., the model which includes k_s) will re-



m_i	x_s [m]	y_s [m]	z_s [m]	q_s [g s^{-1}]	k_s [s^{-1}]
actual m_i	0.0	0.0	0.46	41.20	0.030
mean(m_i^{MCMC})	6.53	-0.30	1.05	44.76	0.028
sd(m_i^{MCMC})	5.37	0.79	1.05	13.22	0.011
95% HPD (m_i^{MCMC})	$[-1.85, 19.00]$	$[-1.50, 1.15]$	$[0.01, 3.17]$	$[20.51, 68.19]$	$[0.006, 0.050]$

Figure 6.14: Case 2: Marginal parameter distributions and summary statistics generated from MCMC samples. The true parameter value is represented by the solid vertical line in the histograms, and the circular dot in the scatter plot. The mean of the MCMC samples is represented by the dashed vertical line in the histograms, and the square dot in the scatter plot. Shaded regions represent 95% HPD intervals based on the MCMC samples.

sult in more truthful estimates (in terms of their accuracy) of the other source parameters (Reichert and Omlin, 1997). In other words, uncertainty about the persistence of a tracer in the environment will be reflected in additional uncertainty about its origin and strength. The statistical approach to particle travel times described in this work significantly mitigates the computational effort required to include k_s as a source parameter in the inference procedure.



m_i	x_s [m]	y_s [m]	z_s [m]	q_s [g s^{-1}]	k_s [s^{-1}]
actual m_i	0.0	0.0	0.46	41.20	0.30
mean(m_i^{MCMC})	2.35	0.12	0.82	39.05	0.29
sd(m_i^{MCMC})	1.53	0.57	0.92	14.03	0.029
95% HPD (m_i^{MCMC})	[0.48, 5.40]	[−0.79, 1.50]	[0.01, 2.73]	[13.44, 65.78]	[0.239, 0.355]

Figure 6.15: Case 3: Marginal parameter distributions and summary statistics generated from MCMC samples. The true parameter value is represented by the solid vertical line in the histograms, and the circular dot in the scatter plot. The mean of the MCMC samples is represented by the dashed vertical line in the histograms, and the square dot in the scatter plot. Shaded regions represent 95% HPD intervals based on the MCMC samples.

Chapter 7

Optimal auxiliary detector placement for source determination

7.1 Introduction

Recent efforts made toward solving inverse problems in atmospheric dispersion have mainly considered situations which involve a fixed network of detectors measuring mean concentration data. When finding the location of the release is time-critical (due to toxic or other hazardous materials being emitted continuously at some undiscovered incident site), a fixed detector network (if one exists) may not yield data capable of effectively isolating the source location to within an easily searchable area. In these cases, it may become necessary to augment the available concentration data using a mobile detection capability. In an alternate scenario, regulatory considerations may necessitate placement of additional detectors for better isolating a single source of pollution where many potential sources exist.

Bayesian approaches to source determination commonly treat the problem as one of parameter estimation (Keats et al., 2007a; Yee, 2008), where the source is defined by a set of parameters $\mathbf{m}$ (location, strength, time of release, etc.) which must be inferred probabilistically. Uncertainties associated with concentration data $\mathbf{d}$ and model predictions $\mathbf{r}$ are essentially propagated through the Bayesian apparatus, resulting in a posterior probability density function (PDF) for the source parameters. Rather than extracting estimates $\hat{\mathbf{m}}$ from this posterior PDF, our goal is to strategically place one or more additional detectors so that the entropy (as a measure of the uncertainty) of the posterior distribution is minimized (or equivalently, its Shannon information content is maximized). This should have the desirable effect of improving our degree of certainty in potential estimates $\hat{\mathbf{m}}$.

In the literature, entropy-based criteria are often applied to the design and analysis of fixed environmental monitoring networks. One type of application involves determining subsets of existing networks which are optimal in the sense of the information they yield (Wu and Zidek, 1992; Silva and Quiroz, 2003; Fuentes et al., 2007). The problem of extending an existing monitoring network (by adding additional detectors) was examined by Zidek et al.

(2000), who considered not only the potential information gain (entropy decrease), but also the cost of operating the new network. The present research differs from these efforts by considering a more specific case: rather than extending our network of detectors to provide an expected information gain for general release scenarios (where weather patterns might change, and pollutant sources are inventoried), we presume a single release which happens under a distinct set of meteorological conditions. Our goal is not to provide a general monitoring capability, but rather to act to improve our state of knowledge regarding this specific, unknown release.

Our work lies along a similar vein to that of Patan et al. (2008), who formulated the problem within the framework of optimal control in order to determine the best trajectory that a set of mobile sensors should take. Patan et al. (2008) consider a time-varying advection-diffusion-reaction equation and perform the trajectory optimization while accounting for constraints on sensor path lengths. By contrast, we examine the problem of a quasi-steady-state release, without reaction or decay of the tracer. We focus on calculating and analyzing the expected information in the solution based on a Bayesian approach to parameter estimation (rather than the optimization approach adopted by Patan et al.).

We adopt a methodology put forward by Loredo (2004) known as Bayesian adaptive exploration (BAE). This methodology is founded on information-theoretic principles, and essentially provides a way to decide how future experiments should be performed so that information about the phenomenon of interest in maximized. Loredo's method falls under the rubric of optimal Bayesian experimental design, which has a rich background in the literature (Lindley, 1956; Bernardo, 1979; Chaloner and Verdinelli, 1995; Sebastiani and Wynn, 2000).

7.2 Bayesian adaptive exploration for source determination

Bayesian adaptive exploration (BAE), as described by Loredo (2004), is an iterative process which involves the following three stages:

Observation: Data are collected in an experiment that is designed to elicit information about the phenomenon under investigation. Here, the data consist of concentration measurements obtained at a limited set of locations.

Inference: Bayesian inference is applied to calculate probabilities for hypotheses about the phenomenon, given the data collected in the previous stage. In the present work, each hypothesis takes the form of a set of parameters defining a possible source location.

Design: A subsequent experiment must be proposed (i.e., designed, or decided upon[1]) which we expect will maximally improve our state of information regarding the phenomenon. Here, we wish to propose a supplementary detector location where we hope that the measured concentration, once it has been processed during the next inference stage, will yield maximum information about the source location.

[1] Bayesian 'decision theory' is synonymous with experimental design.

7.2.1 Bayesian inference for source parameters

Given mean (time-averaged) concentration data $\mathbf{d}$ obtained from a set of detectors, we wish to estimate the parameters $\mathbf{m}$ which characterize the source of the dispersion. From Bayes' theorem, information regarding the source parameters is encapsulated in the posterior distribution:

$$\underbrace{P(\mathbf{m} \mid \mathbf{d}, I)}_{\textit{Posterior}} \propto \underbrace{P(\mathbf{m} \mid I)}_{\textit{Prior}} \underbrace{P(\mathbf{d} \mid \mathbf{m}, I)}_{\textit{Likelihood}} . \tag{7.1}$$

In this chapter we consider the following set of source parameters:

$$\mathbf{m} \equiv (x_s, y_s, q_s) , \tag{7.2}$$

where (x_s, y_s) represent the location of the source, and q_s is the release rate. The source height is assumed known[2], and neither the release rate nor the wind field change in time.

7.2.1.1 Prior and likelihood

The prior distribution expresses our state of knowledge about the parameters $\mathbf{m}$ before the arrival of data $\mathbf{d}$. During the first iteration of the BAE process, this distribution would typically be chosen to express a state of ignorance.[3] Distinguishing between location parameters (x_s, y_s) and scale parameter q_s (Jaynes, 2003), and assuming a priori that parameters are logically independent, the prior distribution takes the form:

$$P(x_s, y_s, q_s \mid I) \propto q_s^{-1} , \quad (x_s, y_s, q_s) \in \Omega , \tag{7.3}$$

where Ω bounds the computational domain (parameter space), ensuring that the prior integrates to unity.

For a fixed set of source parameters $\mathbf{m}$, $P(\mathbf{d} \mid \mathbf{m}, I)$ defines the probability that data $\mathbf{d}$ are measured. When data $\mathbf{d}$ are fixed, $P(\mathbf{d} \mid \mathbf{m}, I)$ defines the likelihood of $\mathbf{m}$ (MacKay, 2003). For the problem of source determination, different authors have adopted different likelihood functions based on varying assumptions made about the noise prior (viz., the level of disagreement between modelled and measured concentrations). In this chapter we adopt a Gaussian form for the likelihood,

$$P(\mathbf{d} \mid \mathbf{m}, I) = \prod_{i=1}^{N} \frac{1}{\sigma_i \sqrt{2\pi}} \exp\left[-\frac{1}{2} \left(\frac{d_i - r_i(\mathbf{m})}{\sigma_i} \right)^2 \right] . \tag{7.4}$$

Other distributions such as log-normal can also be used (Goyal et al., 2005; Senocak et al., 2008; Keats et al., 2009), but for now we adopt the Gaussian for simplicity and compatibility with previous work (Keats et al., 2007a,c). In Eq. (7.4), d_i are the measured concentration

[2] This is a reasonable assumption since in many scenarios, one is concerned with sources located at or very near to the ground.

[3] In subsequent iterations, however, the prior would be replaced by the posterior distribution from the previous iteration.

data (indexed by detector i), while r_i are their modelled counterparts. The r_i are obtained using a source-receptor relationship which could take the form of either a forward or adjoint dispersion model. The source-receptor relationship is discussed further in Sec. 7.2.1.3. At this point, however, it should be noted that the Bayesian formulation remains independent of the choice of dispersion model used to obtain the r_i. The σ_i are a measure of the error associated with each (d_i, r_i) pair. In the present work, this quantity is considered to be specified a priori, either as a constant or as a function of the mean concentration.

7.2.1.2 Posterior distribution

The posterior PDF is proportional to the product of the prior and likelihood:

$$P(\mathbf{m} \mid \mathbf{d}, I) \propto q_s^{-1} \exp\left[-\frac{1}{2}\sum_{i=1}^{N}\left(\frac{d_i - r_i(\mathbf{m})}{\sigma_i}\right)^2\right] . \tag{7.5}$$

7.2.1.3 Source-receptor relationship

Bayesian inference for source determination has already been addressed in the literature for both building-resolving and regional-scale flows (Keats et al., 2007a; Yee et al., 2008). In the present work we wish to focus attention on the design stage of BAE, so we adopt the Lagrangian stochastic (LS) dispersion model used earlier in chapter 6. The LS model was selected over the simpler (but well-known) steady-state Gaussian plume model (Csanady, 1973; Arya, 1999) because we are not only interested in the potential information gain yielded by mean concentrations, but also by turbulent scalar fluxes. As with mean concentration, the scalar flux 'seen' by a detector can be calculated efficiently (for many different source hypotheses) using a backward LS model. Turbulent scalar fluxes behave differently to mean concentration, and we seek to quantify their value as measurables within the BAE framework. The LS model used in this work has already been described in chapter 6, so we shall simply outline the additional facility required for calculating the scalar fluxes (in both backward and forward modes).

In forward mode, a LS model generates particle trajectories which can be post-processed to generate a grid of concentration data. Elaborate kernel smoothing techniques aside, the Reynolds averaged concentration C in a grid cell can be estimated simply by averaging particle residence times:

$$C(\mathbf{x}_d) \approx \frac{1}{\mathcal{V}\, N_p}\sum_{p=1}^{N_p} q^{(p)}\delta t^{(p)} , \tag{7.6}$$

where $\mathcal{V} = \Delta x \Delta y \Delta z$ is the volume of the (sensor or detector) grid cell centered on $\mathbf{x}_d$, $q^{(p)}$ is the mass flow rate (source strength) associated with the particle released from the source, and N_p is the number of particles released from the source. The 'residence time' that particle p spends within the grid cell is given by $\delta t^{(p)}$. The Reynolds averaged scalar fluxes are calculated similarly using the forward model, by weighting particles by their associated velocity

fluctuations:

$$\begin{aligned}\overline{u_j' c'}(\mathbf{x}_d) \approx \phi_j(\mathbf{x}_d;\mathbf{x}_s) &\equiv \frac{1}{\mathcal{V}\, N_p}\sum_{p=1}^{N_p} u_j'^{(p)} q^{(p)} \delta t^{(p)} \ , \quad j=1,2,3, \\ &= \frac{q_s}{\mathcal{V}\, N_p}\sum_{p=1}^{N_p} u_j'^{(p)} \delta t^{(p)} \quad \text{(constant mass flow rate } q^{(p)} = q_s\text{)} \ ,\end{aligned} \tag{7.7}$$

where $(u_1', u_2', u_3')^{(p)} = (u', v', w')^{(p)}$ are the streamwise, spanwise and vertical components of the particle velocity fluctuation in the grid cell centered on $\mathbf{x}_d$; $c'(\mathbf{x}_d)$ denotes the concentration fluctuation at $\mathbf{x}_d$ (this is not directly computed by the LS model). The true (but unknown) scalar flux, $\overline{u_j' c'}$, is approximated by the computed quantity $\phi_j(\mathbf{x}_d;\mathbf{x}_s)$. The quality of this approximation depends on N_p.

In order to calculate scalar fluxes in backward mode, we first define a 'conjugate' scalar flux, ϕ_j^*, where particle pseudo-mass flow rates and residence times are weighted by the velocity fluctuations seen at a detector:

$$\begin{aligned}\phi_j^*(\mathbf{x}_s;\mathbf{x}_d) &\equiv \frac{1}{\mathcal{V}\, N_p}\sum_{p=1}^{N_p} u_j'^{(p)} q^{*(p)} \delta \tau^{(p)} \ , \quad j=1,2,3, \\ &= \frac{1}{\mathcal{V}\, N_p}\sum_{p=1}^{N_p} u_j'^{(p)} \delta \tau^{(p)} \quad (q^* \text{ unity}) \ , \\ &\approx \frac{1}{q_s}\overline{u_j' c'}(\mathbf{x}_d) \ ,\end{aligned} \tag{7.8}$$

where $\mathcal{V}$ is now the volume of the grid cell centered on $\mathbf{x}_s$, and N_p refers to the number of particles released from the detector (located at $\mathbf{x}_d$). At this point it is worthwhile to compare the terms in Eqs. (7.7) and (7.8). In Eq. (7.7), u_j' and δt refer to the marked particle's velocity fluctuation and residence time in the detector volume centered on $\mathbf{x}_d$. The scalar fluxes ϕ_j are essentially 'field variables' in the traditional sense. By contrast, for the backward LS calculation of eq. (7.8), the conjugate flux in a potential source grid cell (located at $\mathbf{x}_s$) is determined using velocity fluctuations at the detector (which is fixed at location $\mathbf{x}_d$ with respect to the grid, and acts as the 'source' from which pseudo-particles are released), but particle pseudo-mass release rates q^* and residence times $\delta\tau$ are taken from the grid cell at $\mathbf{x}_s$. Barring removal processes such as those discussed in chapter 6, q^* is typically unity, and so the conjugate flux ϕ_j^* is equivalent to the normalized scalar flux at the detector location $\mathbf{x}_d$. Multiplying ϕ_j^* by the source strength q_s then yields the scalar flux ϕ_j at the detector location.

We are now equipped with a backward LS model which permits the rapid calculation of both mean concentrations C and scalar fluxes ϕ_j at a given detector location for any arbitrary source location and strength. In chapters 5 and 6, backward (adjoint) models were used to generate one C^* field per detector location. Now, in addition to the C^* field, we require an additional three fields to be generated for each detector: ϕ_j^*, $j = 1, 2, 3$. Scalar fluxes are easily integrated into the overall Bayesian framework for source determination; we simply use an

extended version of the likelihood (8.20) which incorporates scalar flux data:

$$P(\mathbf{d} \mid \mathbf{m}, I) = \prod_{i=1}^{N} \frac{1}{\sigma_i^C \sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{d_i^C - r_i^C}{\sigma_i^C}\right)^2\right] \times \prod_{j=1}^{3} \frac{1}{\sigma_{i,j}^{\phi} \sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{d_{i,j}^{\phi} - r_{i,j}^{\phi}}{\sigma_{i,j}^{\phi}}\right)^2\right] . \tag{7.9}$$

In Eq. (7.9), d_i^C is the concentration and $d_{i,j}^{\phi}$ is the j^{th} scalar flux measured by the i^{th} detector. Their modelled counterparts are $r_i^C \equiv C(\mathbf{x}_{d_i}; \mathbf{m})$ and $r_{i,j}^{\phi} \equiv \phi_j(\mathbf{x}_{d_i}; \mathbf{m})$. The uncertainties surrounding model–data agreement are specified by σ_i^C and $\sigma_{i,j}^{\phi}$.

7.2.2 Experimental design

Consider a scenario where there are N_d receptors (all measuring mean concentration levels d_i, $i = 1, 2, \ldots, N_d$) spread about the domain of interest. The design stage of the BAE process requires that we attempt to maximize the expected utility of a future experiment in which an additional receptor is placed somewhere in the domain to measure an additional concentration datum (or some combination of concentration and flux data), $d_\star$. For experiments whose main purpose is to gain knowledge about a phenomenon, and when the particular choice of experiment does not affect the overall cost of performing it (or the cost is of no concern), an appropriate utility function that measures the information content of the 'final' posterior distribution for $\mathbf{m}$, after the receipt of the future datum $d_\star$ (Lindley, 1956) is:

$$U(d_\star, e) = \int_{\text{all } \mathbf{m}} \mathrm{d}\mathbf{m}\, P(\mathbf{m} \mid d_\star, \mathbf{d}, I_e)\ \log\ P(\mathbf{m} \mid d_\star, \mathbf{d}, I_e) , \tag{7.10}$$

where e specifies the choice of experiment (i.e., the location at which $d_\star$ is to be obtained), and I_e signifies that our background information I accounts for e.

Obviously, the future measurement has not yet been made, so we are faced with maximizing the expected utility, which is the average of the utility function (7.10) weighted by the predictive distribution for $d_\star$ conditioned on known data $\mathbf{d}$ and choice of experiment e:

$$EU(e) = \int_{\text{all } d_\star} \mathrm{d}d_\star\, P(d_\star \mid \mathbf{d}, I_e)\, U(d_\star, e) . \tag{7.11}$$

Following the notation of Loredo (2004), we denote the information content of a distribution by $\mathcal{I}[\blacksquare]$, and the expected utility (7.11) becomes the expected information:

$$E\mathcal{I}(e) = \int_{\text{all } d_\star} \mathrm{d}d_\star\, P(d_\star \mid \mathbf{d}, I_e)\, \mathcal{I}[\mathbf{m} \mid d_\star, \mathbf{d}, I_e] . \tag{7.12}$$

The expected information is easier to evaluate once decomposed into its constituent parts (Sebastiani and Wynn, 2000):

$$\begin{aligned} E\mathcal{I}(e) &= \mathcal{I}[\mathbf{m} \mid \mathbf{d}, I_e] + \int \mathrm{d}\mathbf{m}\; P(\mathbf{m} \mid \mathbf{d}, I_e)\; \mathcal{I}[d_\star \mid \mathbf{m}, \mathbf{d}, I_e] - \mathcal{I}[d_\star \mid \mathbf{d}, I_e] \\ &= \int \mathrm{d}\mathbf{m}\; P(\mathbf{m} \mid \mathbf{d}, I_e) \log P(\mathbf{m} \mid \mathbf{d}, I_e) \\ &\quad + \int \mathrm{d}\mathbf{m}\; P(\mathbf{m} \mid \mathbf{d}, I_e) \int \mathrm{d}d_\star\; P(d_\star \mid \mathbf{m}, \mathbf{d}, I_e) \log P(d_\star \mid \mathbf{m}, \mathbf{d}, I_e) \\ &\quad - \int \mathrm{d}d_\star\; P(d_\star \mid \mathbf{d}, I_e) \log P(d_\star \mid \mathbf{d}, I_e)\; . \end{aligned} \tag{7.13}$$

The first term in Eq. (7.13) is the information in the posterior from the previous iteration of BAE (or in the prior, if this is our initial iteration), and is independent of the choice of experiment e. We disregard this term as an unimportant constant. The second term is the average information contained in the sampling distribution for the new datum $d_\star$. When the shape and spread of the sampling distribution are fixed, this term is also constant. However, if the sampling distribution depends upon the data being measured (e.g., the uncertainty might be a function of the concentration or flux seen by a detector), the second term will depend on e and must therefore be calculated. The final term represents the entropy of the predictive distribution, and also depends upon e.

7.3 Computational approach

For small problems possessing relatively few parameters and where sampling distributions for future data are canonical (e.g., Gaussian, log-normal, Weibull), $E\mathcal{I}(e)$ can be evaluated either analytically or through quadrature. In general, however, the nested sample space and parameter space integrals in Eq. (7.13) may need to be calculated using efficient sampling approaches. In Sections 7.3.1–7.3.3 we outline procedures for sampling from the posterior and for computing the second and third terms in the expected information.

7.3.1 Sampling from the posterior distribution

Estimating the second and third terms of $E\mathcal{I}(e)$ will require that we possess a set of N samples, $\{\mathbf{m}^{(k)}\}$, drawn from the posterior distribution, $P(\mathbf{m} \mid \mathbf{d}, I)$. In this work, we obtain these samples using Metropolis-Hastings Markov chain Monte Carlo (MCMC). MCMC approaches are commonly used for source determination (Keats et al., 2007a; Yee, 2008), and will not be discussed in depth here. Details can be found in chapter 4 and in the books by Gilks et al. (1996) and MacKay (2003). At this point it should be noted that the posterior distribution does not account for the arrival of future datum $d_\star$, and so notationally, $P(\mathbf{m} \mid \mathbf{d}, I_e) = P(\mathbf{m} \mid \mathbf{d}, I)$.

7.3.2 Estimating $\mathcal{I}[d_\star \mid \mathbf{d}, I_e]$

Having obtained a set of samples $\{\mathbf{m}^{(k)}\}$ using MCMC[4], we adopt a technique known as posterior sampling to estimate the information in the predictive PDF for $d_\star$. This technique, outlined by Loredo (2004), is summarized below and adapted for the task of optimal detector placement.

The PDF $P(d_\star \mid \mathbf{d}, I_e)$ can be expressed as the product of the sampling distribution for new data and the posterior distribution from the previous BAE iteration:

$$\begin{aligned} P(d_\star \mid \mathbf{d}, I_e) &= \int_{\text{all } \mathbf{m}} \mathrm{d}\mathbf{m}\; P(d_\star \mid \mathbf{m}, I_e)\; P(\mathbf{m} \mid \mathbf{d}, I_e) \\ &\approx \frac{1}{N} \sum_{k=1}^{N} P(d_\star \mid \mathbf{m}^{(k)}, I_e) = \tilde{P}(d_\star)\ , \end{aligned} \tag{7.14}$$

recognizing that $P(d_\star \mid \mathbf{m}, I_e) = P(d_\star \mid \mathbf{m}, \mathbf{d}, I_e)$. Posterior sampling provides a way to obtain M samples $d_\star^{(j)}$, from which the information $\mathcal{I}[d_\star \mid \mathbf{d}, I_e]$ can be estimated. Details are given in Algorithm 7.1.

Algorithm 7.1 Posterior Sampling

1: $\{\mathbf{m}^{(k)}\} \leftarrow$ *draw N samples from* $P(\mathbf{m} \mid \mathbf{d}, I_e)$
2: *choose integer* $M \leq N$
3: FOR $j = 1, 2, ..., M$ DO
4: $\quad \mathbf{m}^{(j)} \leftarrow$ *draw uniformly from* $\{\mathbf{m}^{(k)}\}$
5: $\quad r_\star(\mathbf{m}^{(j)}; e) \leftarrow$ *calculate source-receptor relationship*
6: $\quad d_\star^{(j)} \leftarrow$ *draw from sampling dist,* $P(d_\star \mid \mathbf{m}^{(j)}, I_e)$
7: $\quad \tilde{P}(d_\star^{(j)}) \leftarrow \frac{1}{N} \sum_{k=1}^{N} P(d_\star^{(j)} \mid \mathbf{m}^{(k)}, I_e)$
8: END FOR
9: $\mathcal{I}[d_\star \mid \mathbf{d}, I_e] \leftarrow \frac{1}{M} \sum_{j=1}^{M} \log\ \tilde{P}(d_\star^{(j)})$

With regard to steps 5 and 6 of the posterior sampling algorithm, note that the sampling distribution depends on the modelled concentration for a specific source-receptor configuration, $r_\star(\mathbf{m}^{(j)}; e)$. The sampling distribution for new data is effectively the noise prior (which was used to establish the likelihood):

$$P(d_\star \mid \mathbf{m}, I_e) = \frac{1}{\sigma_\star \sqrt{2\pi}} \exp\left[-\frac{1}{2} \left(\frac{d_\star - r_\star}{\sigma_\star} \right)^2 \right]\ . \tag{7.15}$$

[4] The samples $\{\mathbf{m}^{(k)}\}$ might only constitute a subset of the entire collection of MCMC samples.

7.3.3 Estimating $\int \mathrm{d}\mathbf{m}\; P(\mathbf{m} \mid \mathbf{d}, I_e)\; \mathcal{I}[d_\star \mid \mathbf{m}, \mathbf{d}, I_e]$

Obtaining an estimate for the posterior-averaged information in the sampling distribution is relatively straightforward if an analytical expression for $\mathcal{I}[d_\star \mid \mathbf{m}, \mathbf{d}, I_e]$ is available. Recognizing that $\mathcal{I}[d_\star \mid \mathbf{m}, \mathbf{d}, I_e] = \mathcal{I}[d_\star \mid \mathbf{m}, I_e]$ (conditioning on $\mathbf{m}$ obviates dependence on $\mathbf{d}$), the information in the sampling distribution (7.15) can be obtained analytically:

$$\begin{aligned} \mathcal{I}[d_\star \mid \mathbf{m}, I_e] &= \int \mathrm{d}d_\star\; P(d_\star \mid \mathbf{m}, I_e) \log P(d_\star \mid \mathbf{m}, I_e) \\ &= -\log \sigma_\star \sqrt{2\pi \exp(1)}\;. \end{aligned} \tag{7.16}$$

The posterior-averaged information can then be approximated with the help of existing MCMC samples $\{\mathbf{m}^{(k)}\}$:

$$\int \mathrm{d}\mathbf{m}\; P(\mathbf{m} \mid \mathbf{d}, I_e)\; \mathcal{I}[d_\star \mid \mathbf{m}, \mathbf{d}, I_e] \approx \frac{1}{N} \sum_{k=1}^{N} \mathcal{I}[d_\star \mid \mathbf{m}^{(k)}, I_e]\;. \tag{7.17}$$

As mentioned in Section 7.2.2, when $\sigma_\star$ is constant, the term (7.16) is also constant, and does not need to be calculated as part of the optimization procedure (where we try to maximize the expected information). However, in this work we shall consider a situation where $\sigma_\star$ is dependent upon the choice of experiment; for example, one might expect the error to depend on the magnitude of the modelled or measured data.

7.4 Short-range dispersion in the atmospheric surface layer

Here we present test cases for which the wind field and the heights of the source and detectors are the same as those used in the Project Prairie Grass (PPG) dispersion experiment (which was described in Sections 6.4 and 6.4.1 of the previous chapter). However, data obtained from samplers used in the PPG experiment are not used.[5] Instead, a reference solution (mean concentration and scalar flux fields) is obtained using a forward LS model in order to provide 'synthetic' concentrations and fluxes for later use in both the inverse problem (sec. 7.4.2) and the subsequent Bayesian adaptive exploration (sec. 7.4.3).

7.4.1 Reference solution: forward dispersion

Figures 7.1 and 7.2 present contours of the concentration field obtained using the forward LS model, generated from the trajectories of 10^6 marked particles. The dimensions of each grid cell in the domain are $(\Delta x, \Delta y, \Delta z) = (1/2, 1/2, 1/16)$ m. As in Section 6.4, the source is located at $(x_s, y_s, z_s) = (0, 0, 0.46)$ m. The present results differ from the previous chapter in that the source strength is now set to be unity.

[5] Scalar fluxes were not measured during the PPG experiment.

In addition to mean concentration data, we are also interested in obtaining a reference solution for the scalar fluxes $\overline{u_j' c'}$. In the near-field, scalar fluxes exhibit a more interesting behaviour than the concentration, as seen in figures 7.3 and 7.4. The vertical flux profiles compare favourably to those obtained by Fackrell and Robins (1982).

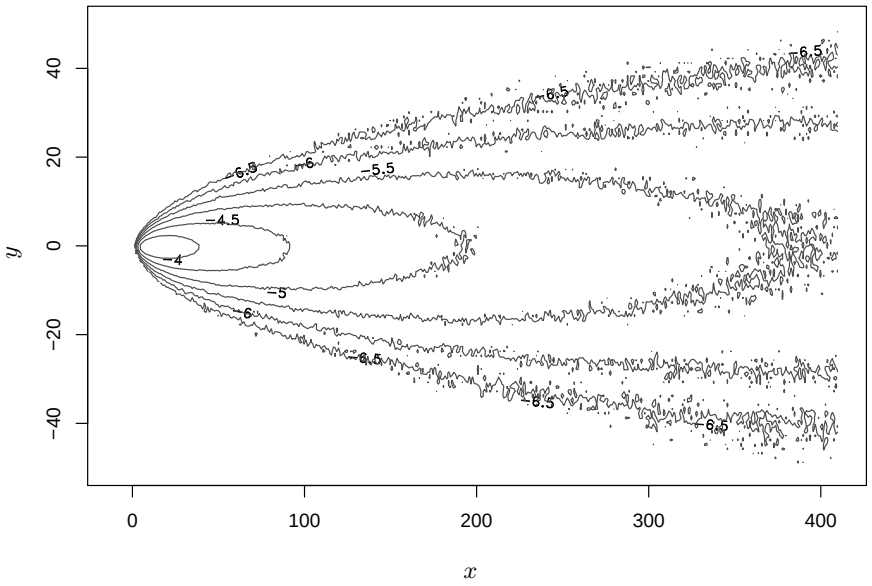


Figure 7.1: Contours of $\log_{10} C$ obtained using a forward LS model with a source strength of unity. Noise increases with downstream distance due to the relative paucity of marked particles passing through grid cells. Contours are drawn at a height of $z_s = 0.46$ m in the $x - y$ (horizontal) plane.

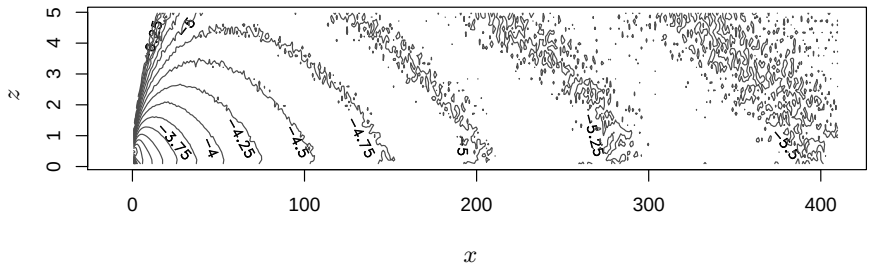


Figure 7.2: Contours of $\log_{10}$ mean concentration, drawn at the plume centerline ($y = 0$) in the $x - z$ (vertical) plane.

Unfortunately, the scalar fluxes decay far more rapidly than the mean concentration, as seen in figure 7.5. Their rate of decay with downstream distance is roughly twice that of the mean concentration, and we observe an even more precipitous decrease in the signal to noise ratio obtained from the LS simulation results. The implication here is that if scalar flux data is to aid in solving the inverse problem, it should be obtained at a location sufficiently close to the true (but unknown) source.

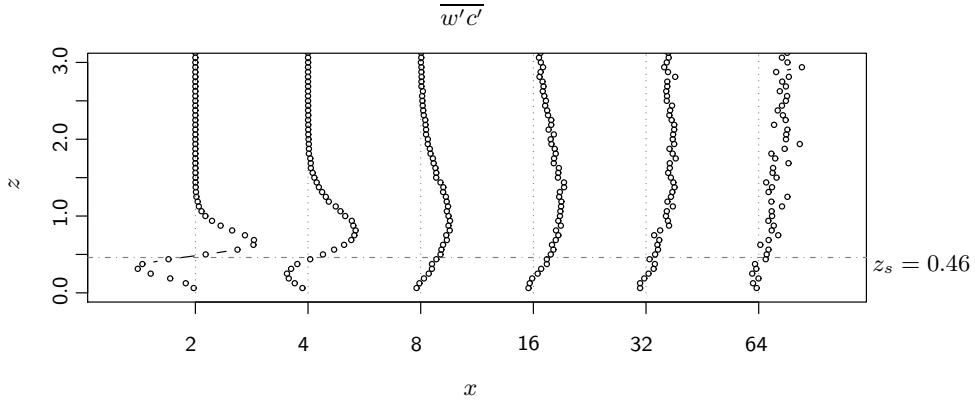


Figure 7.3: Profiles showing the vertical structure of the scalar flux $\overline{w'c'}$, normalized by $u_* C^{\max}$, where $C^{\max}$ is the maximum centerline concentration at each x location.

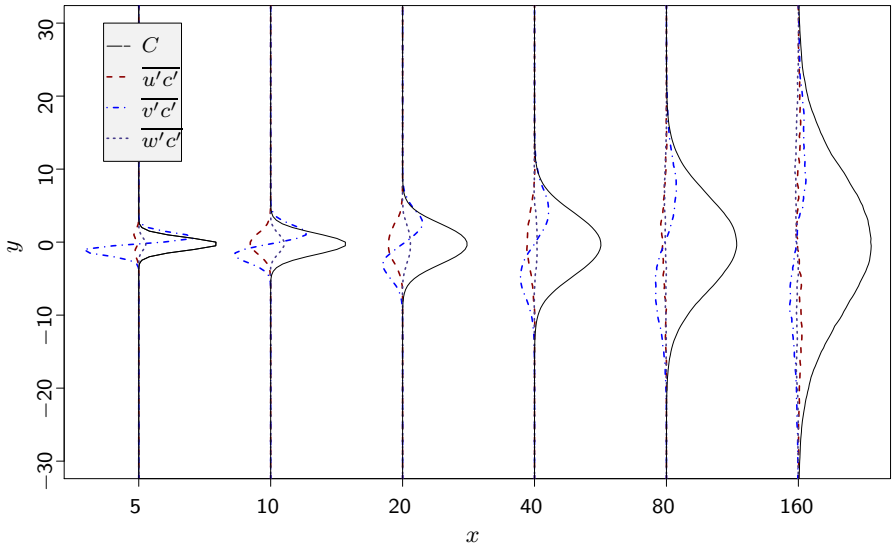


Figure 7.4: Crosswind profiles of mean concentration and scalar fluxes, drawn at source height (0.46 m). Concentration data are normalized by $C^{\max}$, while fluxes are normalized by $u_* C^{\max}$. Profiles have been smoothed.

7.4.2 Inverse problem: source determination

In order to demonstrate the utility of the BAE methodology (sec. 7.4.3), we must first solve an inverse problem where detectors and their corresponding measured concentrations d_i^C are already specified. At this point, scalar fluxes are not used; however, the information that they could potentially yield at a supplementary detector site will be assessed in the next section.

From equations (7.14) and (7.16), it can be seen that the information functional $E\mathcal{I}$ is relatively sensitive to the choice of sampling distribution (or equivalently, noise prior). It is therefore worthwhile to consider, at a minimum, the following two cases:

1. $\sigma_i^C = 10^{-6}\ \forall\ i$: detector uncertainties are constant;
2. $\sigma_i^C = \sqrt{(0.3\, d_i^C)^2 + (10^{-6})^2}\ \forall\ i$: uncertainties are approximately proportional to measured values.

In each case the noise is assumed to be Gaussian.

Figure 7.6 illustrates the layout of detectors used to solve this inverse problem. A single C^* field, generated at the far end of the domain, was copied and transposed to each of the detector locations. Mean concentration data were obtained from the forward LS simulation presented in Section 7.4.1 and subjected to additive Gaussian noise with the variances prescribed above.

The posterior PDF (7.5) was sampled using MCMC, and histograms and corresponding summary statistics for the cases of constant and variable σ_i^C are shown in figures 7.7 and 7.8 respectively. For both cases, the chosen detector array apparently does a poor job of distinguishing the source's x-location and strength, as evidenced by the strong correlation and wide spread seen in the contour plots of the marginal posterior distribution for x_s and q_s. The con-

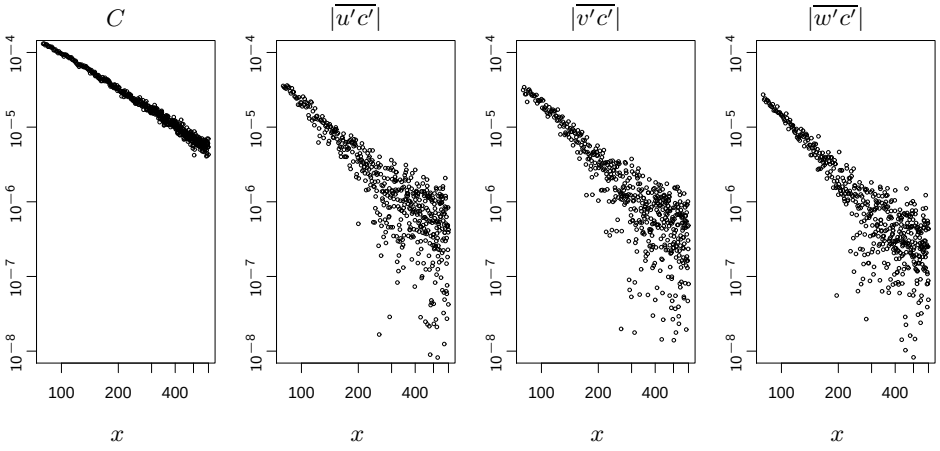


Figure 7.5: Along-plume profiles of mean concentration and absolute values of scalar fluxes. Taken at detector height ($z = 1.5$ m). The quantities $C, |\overline{u'c'}|$, and $|\overline{w'c'}|$ are plotted along the plume centerline ($y = 0$), while $|\overline{v'c'}|$ is plotted along $y = 1.5$ m, slightly off-centerline.

tour plots were obtained from the MCMC samples using a binned kernel density estimation tool.

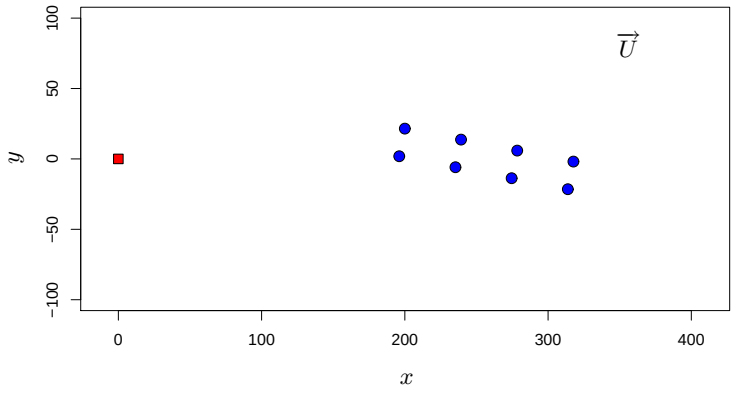


Figure 7.6: Source-receptor configuration. Circular dots are detectors; the square dot is the [unknown] source. Detector array is oriented at $11.25°$ from the horizontal.

7.4.3 Bayesian adaptive exploration: design stage

At this point we wish to evaluate $\mathcal{I}[d_\star \mid \mathbf{d}, I_e]$ and $\int \mathrm{d}\mathbf{m}\; P(\mathbf{m} \mid \mathbf{d}, I_e)\; \mathcal{I}[d_\star \mid \mathbf{m}, \mathbf{d}, I_e]$, the components of the expected information which depend on the placement of an additional detector. Adding these two components and obtaining their maximum will reveal the location where the additional detector must be placed in order to optimally reduce uncertainty in the posterior distribution for the source parameters.

Due to the fact that the wind field is horizontally homogeneous, the hypothetical source-receptor relationship can be rapidly calculated for arbitrary detector placements. This allows us to examine the $E\mathcal{I}$ surface over a relatively large grid of possible detector locations, instead of relying on an optimization approach to locate a single maximum.[6] As mentioned above, the overall shape of the $E\mathcal{I}$ surface is quite sensitive to the noise specification, and plotting it for a range of possible (x_d, y_d) coordinates offers insight into this sensitivity.

We specify the sampling distribution for new data, $P(d_\star \mid \mathbf{m}, I_e)$, separately for concentrations and fluxes. For concentrations, the standard deviation of the sampling distribution is defined to be compatible with the aforementioned definitions:

$$\sigma_\star^C = 10^{-6} \qquad \text{(constant detector uncertainty)}\,, \tag{7.18}$$

$$\sigma_\star^C = \sqrt{(0.3\, r_\star^C)^2 + (10^{-6})^2} \qquad \text{(variable detector uncertainty)}\,. \tag{7.19}$$

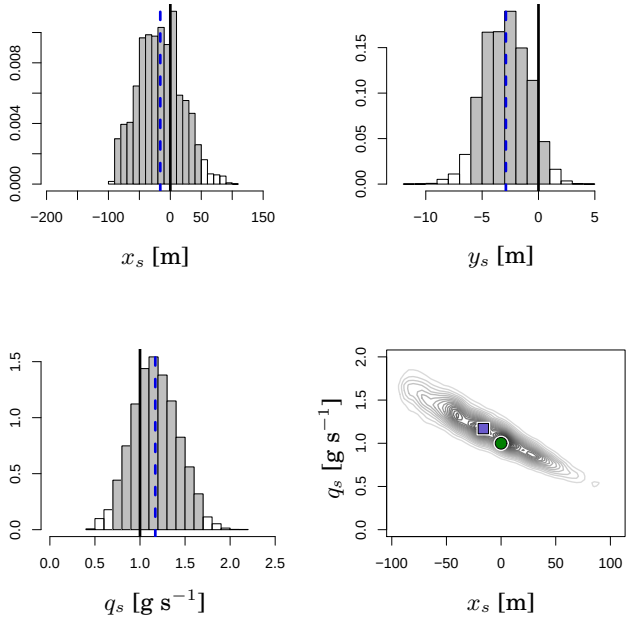
[6] This maximum may also be corrupted by noise, since the procedure for computing $E\mathcal{I}$ (outlined in Sections 7.3.2 and 7.3.3) is stochastic.

These standard deviations are then scaled by a factor of 10% in order to obtain expressions for the scalar fluxes:

$$\sigma^{\phi}_{\star,j} = 10^{-7} \qquad \text{(constant detector uncertainty)}\,, \tag{7.20}$$

$$\sigma^{\phi}_{\star,j} = \sqrt{(0.1 \times 0.3\, r^{\phi}_{\star,j})^2 + (10^{-7})^2} \qquad \text{(variable detector uncertainty)}\,. \tag{7.21}$$

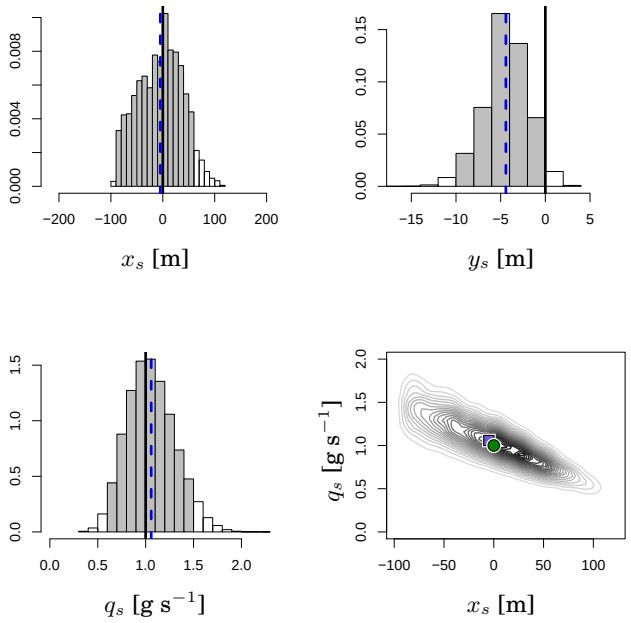
Using the above definitions, we calculate the components of EI using the techniques described in Sections 7.3.2 and 7.3.3. EI surfaces corresponding to the constant and variable uncertainty cases are presented in figures 7.9 and 7.10, respectively. Contributions from concentration and flux data are separated and shown in the upper two panels of each figure; the lower panel presents the sum of the two components.



m_i	x_s [m]	y_s [m]	q_s [g s^{-1}]
actual m_i	0.0	0.0	1.00
mean(m_i^{MCMC})	-14.4	-2.9	1.16
sd(m_i^{MCMC})	35.5	2.0	0.25
95% HPD (m_i^{MCMC})	$[-87.2, 45.7]$	$[-6.4, 1.0]$	$[0.69, 1.67]$

Figure 7.7: MCMC results for the case where detector uncertainties are constant: $\sigma_i^C = 10^{-6}\ \forall\ i$. The true parameter value is represented by the solid vertical line in the histograms, and the circular dot in the contour plot. The mean of the MCMC samples is represented by the dashed vertical line in the histograms, and the square dot in the contour plot. Shaded regions represent 95% HPD intervals based on the MCMC samples.

Additional noise is present in the surfaces shown in figure 7.10, where $\sigma_\star$ is not held constant. This is due to the fact that the posterior-weighted information in the sampling distribution ($\int \mathrm{d}\mathbf{m}\ P(\mathbf{m} \mid \mathbf{d}, I_e)\ \mathcal{I}[d_\star \mid \mathbf{m}, \mathbf{d}, I_e]$) is no longer constant[7] and must be estimated using the stochastic technique described in Section 7.3.3. Evidently, the maximum information yielded by an additional detector measuring scalar fluxes is approximately three times that obtained from a detector measuring concentration alone (the j^{th} scalar flux yields about as much information as a single concentration measurement). It is also interesting to observe how the shapes of the $E\mathcal{I}$ surfaces depend on the specification of the sampling distribution for new data. While the maxima are similarly located, the expected information surface for the case of variable $\sigma_\star$ exhibits 'wings' which follow regions of relatively large concentration



m_i	x_s [m]	y_s [m]	q_s [g s^{-1}]
actual m_i	0.0	0.0	1.00
mean(m_i^{MCMC})	-2.4	-4.3	1.05
sd(m_i^{MCMC})	42.4	2.4	0.25
95% HPD (m_i^{MCMC})	$[-87.5, 64.4]$	$[-9.3, 0.4]$	$[0.59, 1.55]$

Figure 7.8: MCMC results for the case where detector uncertainties are approximately proportional to measured values: $\sigma_i^C = \sqrt{(0.3\, d_i^C)^2 + (10^{-6})^2}$. The true parameter value is represented by the solid vertical line in the histograms, and the circular dot in the contour plot. The mean of the MCMC samples is represented by the dashed vertical line in the histograms, and the square dot in the contour plot. Shaded regions represent 95% HPD intervals based on the MCMC samples.

[7] For the case of constant $\sigma_\star$ seen in figure 7.9, $\mathcal{I}[d_\star \mid \mathbf{m}, \mathbf{d}, I_e]$ is also a constant.

gradient (along the plume edges). In these regions, the signal-to-noise ratio as determined by $d_\star/\sigma_\star$ does not degrade rapidly, and so the behaviour seen in figure 7.10 is entirely expected.

Having calculated the $E\mathcal{I}$ surfaces, for each of the two cases we place a single additional detector measuring both concentration and scalar fluxes at the location of the maximum $E\mathcal{I}$ value. The forward LS model results are used to generate synthetic concentration and scalar flux measurements (which are then perturbed by Gaussian noise) at this new location. With this additional datum, MCMC is used to sample from the new posterior distribution based on the extended likelihood given by Eq. (7.9). Histograms and corresponding summary statistics are presented in figures 7.11 and 7.12. These new results show that uncertainties in x_s and q_s are drastically reduced, and correlations previously seen in the contour plot corresponding to the marginal posterior distribution for (x_s, q_s) are now negligible.

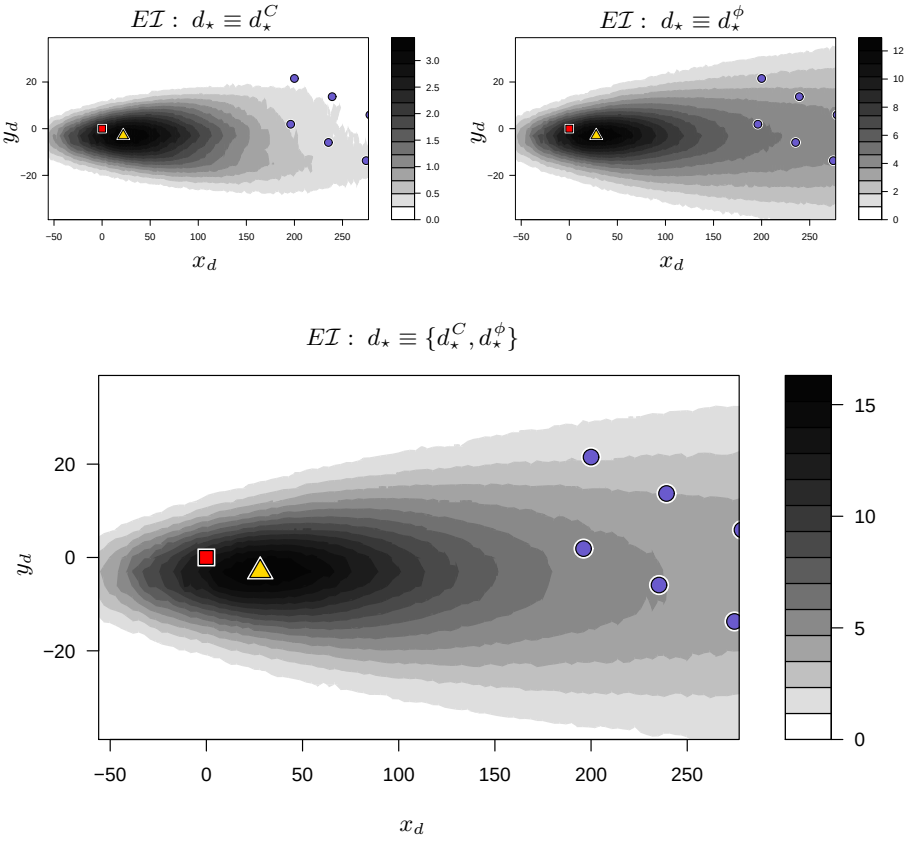


Figure 7.9: $E\mathcal{I}$ surfaces, $\sigma_\star$ held constant. The square marks the true source location, and the circles mark detector locations. The triangle marks the location where $E\mathcal{I}$ reaches a maximum.

7.5 Conclusions

The material in this chapter has demonstrated the viability and some of the issues behind using an information-based approach (BAE) to address the optimal detector placement problem. Furthermore, the fusion of additional data types (such as scalar fluxes) adds potential value (informationally) for solving the source determination problem. Clearly, much future work remains to be done in the area of optimal detector placement. Three key recommendations can be made:

1. It is necessary to perform further research into properly defining detector uncertainties (both modelled and measured), and by extension, the sampling distribution for new data. The sampling distributions used in this chapter were chosen arbitrarily, and while the location of maximum $E\mathcal{I}$ was not found to depend greatly on the choice, the overall shape was. This carries implications where potential optimization algorithms are concerned (for cases where the problem is too complicated to calculate $E\mathcal{I}$ over a grid of points).

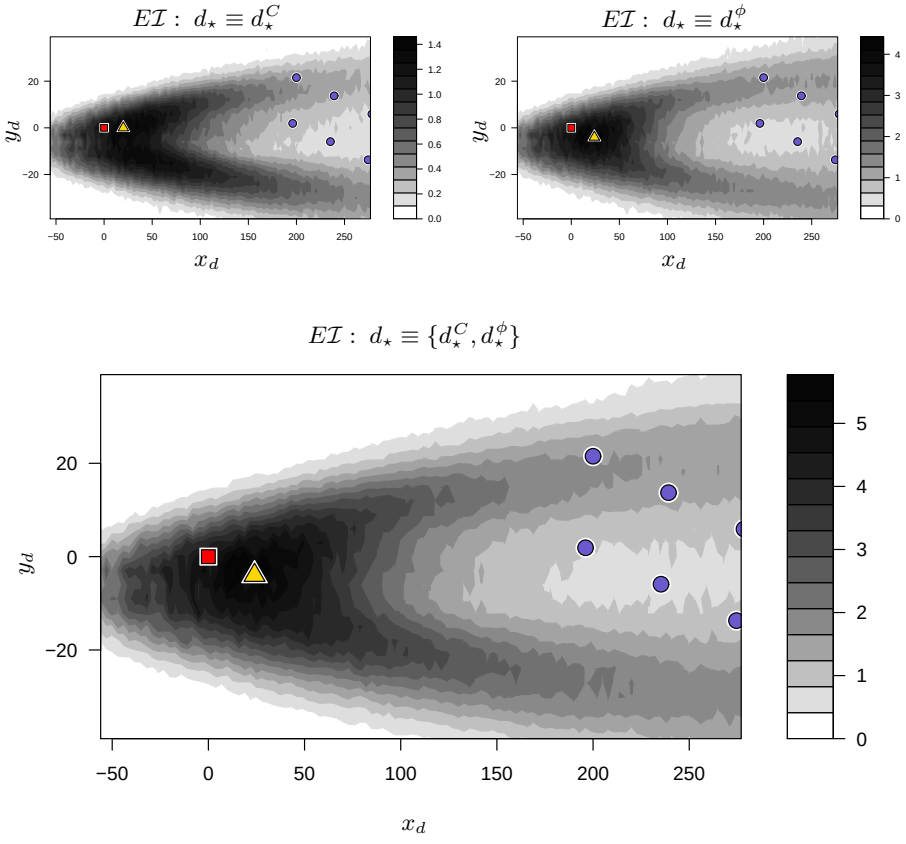
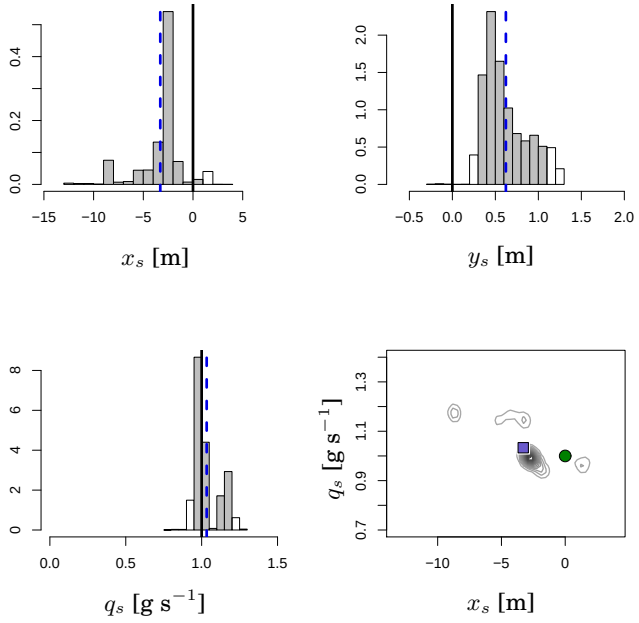


Figure 7.10: $E\mathcal{I}$ surfaces, $\sigma_\star$ variable depending on modelled concentration. The triangle marks the location where $E\mathcal{I}$ reaches a maximum.

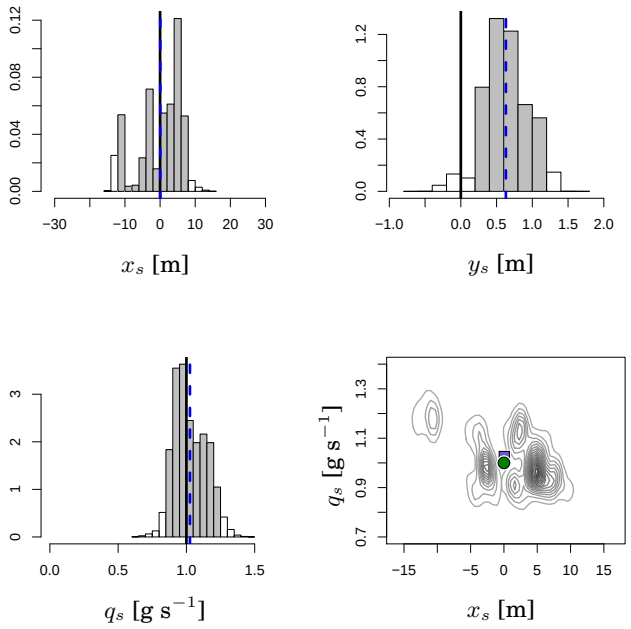
2. Optimization algorithms are needed which are capable of maximizing (or minimizing) a noisy objective function such as that which arises when $E\mathcal{I}$ is calculated using stochastic techniques. Such algorithms might be genetic, or take the form of simulated annealing methods.

3. Further research is required into mitigating the computational effort that would be needed when flows are spatially inhomogeneous and time dependent. Currently, for every proposed detector location that a given optimization technique proposes, a backward dispersion model must be run in order to calculate the concentrations that the detector would expect to see. Judiciously adopting heuristics in aid of the chosen optimization procedure (such as searching regions of high expected concentration gradient) may help to minimize the number of backward model runs needed.



m_i	x_s [m]	y_s [m]	q_s [g s^{-1}]
actual m_i	0.0	0.0	1.00
mean(m_i^{MCMC})	-3.1	0.6	1.03
sd(m_i^{MCMC})	2.3	0.3	0.08
95% HPD (m_i^{MCMC})	$[-9.0, 1.1]$	$[0.25, 1.14]$	$[0.93, 1.20]$

Figure 7.11: Post-BAE MCMC results for the case where detector uncertainties are constant. An extra detector measuring both concentration and scalar fluxes has been added at the location of the maximum $E\mathcal{I}$ value as shown in figure 7.9.



m_i	x_s [m]	y_s [m]	q_s [g s^{-1}]
actual m_i	0.0	0.0	1.00
mean(m_i^{MCMC})	1.3	0.6	1.01
sd(m_i^{MCMC})	5.6	0.3	0.11
95% HPD (m_i^{MCMC})	$[-12.9, 8.0]$	$[0.1, 1.2]$	$[0.83, 1.26]$

Figure 7.12: Post-BAE MCMC results for the case where detector uncertainties are approximately proportional to measured values. An extra detector measuring both concentration and scalar fluxes has been added at the location of the maximum $E\mathcal{I}$ value as shown in figure 7.10.

Chapter 8

Source apportionment using the Chemical Mass Balance model

The material presented in this chapter is adapted from an earlier paper:

> A. Keats, M-T. Cheng, E. Yee, and F-S. Lien. Bayesian treatment of a chemical mass balance receptor model with multiplicative error structure. *Atmospheric Environment*, 43:510–519, 2009.

8.1 Introduction

Source apportionment studies performed using receptor models employ information about the chemical makeup of sampled particulate matter (PM) in order to infer the relative contributions made by emission sources. Given a set of relevant source profiles (which indicate the relative proportions of constituent species present in the PM released by each individual emission source) as well as knowledge of the constituent elemental and molecular species which comprise the sampled PM, the receptor model provides a source-receptor relationship which is used to estimate the most important individual contributors to the sample.

The chemical mass balance (CMB) receptor model (Watson et al., 1990), which is commonly used to perform source apportionment (Watson and Chow, 2001; Kim and Henry, 1999; Chio et al., 2004; Marmur et al., 2007), essentially reduces the problem to one of constrained multiple linear regression. In the standard CMB model, mass concentrations of individual species are assumed to be linear combinations of emission source contributions:

$$y_i = \sum_{j=1}^{N} X_{ij}\ \beta_j + \epsilon_i\ , \tag{8.1}$$

where N is the number of source profiles; y_i is the measured concentration of the i^{th} species, and X_{ij} is the fractional amount of the i^{th} species originating from the j^{th} source. The target

quantities, β_j, are the calculated contributions of the j^{th} source to the receptor. Given that there are M species of interest (both elemental and molecular), the X_{ij} constitute a matrix, $\mathbb{X}^{M\times N}$, which is usually not square. The vector $\boldsymbol{\beta}$ can be obtained using, e.g., a least-squares method (Watson et al., 1990). The CMB model accounts for measurement uncertainty (and to a certain extent, model uncertainty) through the term ϵ_i, which is commonly assumed to be additive, uncorrelated and Gaussian in nature.

In the literature, the CMB model has been reinterpreted and implemented in several different but related ways, with the primary differences lying in the treatment of its associated measurement and model errors. Aside from the ordinary least-squares approach, Christensen and Gunst (2004) outline four alternative approaches to source apportionment based on the CMB model, each of which yields a specific estimator for $\boldsymbol{\beta}$ along with its variance. One popular method, known as the effective variance solution, explicitly accounts for uncertainty in the source composition matrix $\mathbb{X}$ (we hereafter denote this uncertainty by $\sigma_{X_{ij}}$) in addition to measurement uncertainties (Watson et al., 1990). For the methods described by Christensen and Gunst which do account for model error explicitly, this error is assumed to be the mass of the i^{th} species that is not accounted for by all sources in the model. By contrast, the present work presumes a more holistic (although less precise) interpretation of the model error – we assume it to be representative of any prospective failure of the CMB model to calculate the true species concentration y_i, given the source contributions β_j. As with any other predictive model, this failure might arise as a consequence of physical reality's deviation from any or all of the basic assumptions behind the CMB model (lists of these assumptions can be found in the works of Christensen and Gunst (2004) or Seinfeld and Pandis (2006)).

Bayesian approaches to source apportionment are becoming increasingly common. Bayesian inference has already proven to be a useful tool for multivariate statistical analysis, and its application to receptor models is a part of this pattern. Chan et al. (1996) applied Bayesian inference to a receptor modelling problem in Taipei and used Markov chain Monte Carlo (MCMC) to sample from a joint posterior distribution for the source contributions $\boldsymbol{\beta}$ and the combined model and measurement uncertainties. In a similar vein to the present work, they log-transformed the data and used a modified CMB equation of the form

$$\log y_i = \log\left(\sum_{j=1}^{N} X_{ij}\ \beta_j\right) + \epsilon_i\ , \tag{8.2}$$

with ϵ_i being normally distributed and uncorrelated. We expand on their analysis, however, to consider model, measurement and source profile errors separately, and adopt informative prior distributions for these errors.

A more general problem based on the chemical mass balance approach is source identification (and subsequent apportionment), frequently conducted using techniques of multivariate receptor modelling. Bayesian inference provides a valuable framework for addressing this problem, as seen in the work of Park et al. (2002). They consider both measurement and model parameter uncertainties and formulate a posterior distribution based on truncated normal and conjugate priors for these quantities. Such an approach presumes additive behaviour for

the errors, whereas the present work adopts a multiplicative assumption (although Park et al. did conduct investigations using log-normally distributed data). Park et al. (2001) extended the Bayesian approach to account for temporal correlations in the data, improving estimates for source compositions.

Billheimer (2001) and Kashiwagi (2004) also address source apportionment in a Bayesian framework. They both treat the data compositionally (viz., as vectors which convey relative proportions). Kashiwagi examines the case where one source is unknown, and compares a few different distributions for the model error (including log-normal and truncated-normal), evaluating their suitability using estimates of the posterior mean and variance. The present work is more similar to that of Billheimer in that positivity is maintained and prior information about the source profiles and measurement errors is incorporated in a cogent way. However, we do not explicitly treat the data as being compositional, and whereas Billheimer uses the logistic normal distribution for model parameters, we adopt a log-normal distribution. It is interesting to note that Billheimer obtains source contribution estimates based on posterior medians, a reflection of the fact that the median is a better indicator of central tendency in such positively skewed distributions (Slob, 1994). Recent work of Lingwall et al. (2008) adopts a Bayesian approach which maintains positivity while treating data compositionally through the use of a generalized Dirichlet distribution. Lingwall et al. address both exploratory aspects (source profiles considered unknown) and confirmatory (CMB apportionment) aspects of the problem.

The question of how to treat errors in the chemical mass balance model (additively vs. multiplicatively) has not received a great deal of attention in the literature, perhaps because the log-normal distribution is not as firmly embedded in the scientific consciousness as the normal distribution. Some would argue that the log-normal occurs just as frequently as the normal (Limpert et al., 2001), as a consequence of multiplicative error behaviour. Such errors require alternate techniques of analysis, such as using the coefficient of variation (CV) and median as measures of spread and central tendency, respectively. In the work by Watson and Chow (2001), source profile entries for certain elements can be seen to display uncertainties which are almost two orders of magnitude larger than the average abundances. Although these uncertainties (and averages) are commonly expressed in additive terms, we advocate reconsidering them in a multiplicative sense.

In the next section, we phrase the source apportionment problem in a Bayesian probabilistic sense by giving consideration to the types of uncertainty present. Quantifying these uncertainties leads to a comprehensive probabilistic expression for the source and profile parameters (this expression is known as the posterior distribution), which must be sampled in order to obtain statistics (e.g., estimates of expected source contributions). Section 8.3 describes the specific MCMC technique used for sampling from the posterior distribution. Since there are potentially hundreds of parameters present, conventional Markov chain Monte Carlo sampling techniques remain challenging to execute (due to the high dimensionality of the parameter space). Indeed, we employ a Hamiltonian MCMC algorithm, utilizing the gradients of the log-posterior in order to accomplish the sampling in a reasonable amount of time. The overall

methodology is evaluated in Section 8.4 using $PM_{2.5}$ (particulate matter with an aerodynamic diameter less than 2.5 μm) data from Fresno, California.

8.2 Bayesian formulation

Phrasing the source apportionment problem in a Bayesian framework allows one to obtain an expression for the probability of the source contributions which is consistent with any assumptions made about the nature of measurement and model errors. The Bayesian methodology culminates in an expression for the posterior distribution,

$$P(\boldsymbol{\beta}, \mathbb{X} \mid \mathbf{y}, I) \; , \tag{8.3}$$

where $\boldsymbol{\beta}$ is the vector of N unknown source contributions β_j, $\mathbb{X}^{M \times N}$ is the matrix of source profile information X_{ij}, $\mathbf{y}$ is the vector of M elemental and chemical species measurements y_i, and I represents background information pertaining to our problem. In traditional CMB-based source apportionment, the source profiles $\mathbb{X}$ are often considered to be fixed, and uncertainty surrounding them enters the calculation in the form of the standard deviations $\sigma_{X_{ij}}$ (which are reported along with the profiles). However, for the present approach we consider $\mathbb{X}$ to be unknown (although our prior knowledge about the entries X_{ij} and their standard deviations acts to 'constrain' the range of plausible values for $\mathbb{X}$) and requiring inference. According to Bayes' theorem, the posterior probability is proportional to the likelihood multiplied by the prior:

$$P(\boldsymbol{\beta}, \mathbb{X} \mid \mathbf{y}, I) \propto \underbrace{P(\mathbf{y} \mid \boldsymbol{\beta}, \mathbb{X}, I)}_{\text{likelihood}} \; \underbrace{P(\boldsymbol{\beta}, \mathbb{X} \mid I)}_{\text{prior}} \; . \tag{8.4}$$

Here we assume a priori that $\boldsymbol{\beta}$ and $\mathbb{X}$ are statistically independent[1]. In this case, the prior factors as follows:

$$P(\boldsymbol{\beta}, \mathbb{X} \mid I) = P(\boldsymbol{\beta} \mid I) P(\mathbb{X} \mid I) \; . \tag{8.5}$$

When the species measurements y_i are statistically independent (in this work we do not consider the dependent case), the likelihood factors as follows:

$$P(\mathbf{y} \mid \boldsymbol{\beta}, \mathbb{X}, I) = \prod_{i=1}^{M} P(y_i \mid \boldsymbol{\beta}, \mathbf{X}_i, I) \; , \tag{8.6}$$

[1] Note that a prior assumption of independence does not rule out posterior correlation.

where $\mathbf{X}_i$ is a vector of length N consisting of the i^{th} row of the matrix $\mathbb{X}$. Under similar assumptions of independence, the prior distributions for $\mathbb{X}$ and $\boldsymbol{\beta}$ factor as follows:

$$P(\mathbb{X} \mid I) = \prod_{i=1}^{M} \prod_{j=1}^{N} P(X_{ij} \mid I) \; , \tag{8.7}$$

$$P(\boldsymbol{\beta} \mid I) = \prod_{j=1}^{N} P(\beta_j \mid I) \; . \tag{8.8}$$

Note that the Bayesian formulation presented above is independent of the specific choice of model used for source apportionment. Before proceeding with the model definition, however, we reiterate that in this work, both measurement and model errors are considered separately, so at this point it is important to distinguish y_i, the measured concentration of the i^{th} species, from $\check{y}_i$ $(= \sum_{j=1}^{N} X_{ij} \; \beta_j)$ the modelled concentration of the i^{th} species.

8.2.1 Sources and types of uncertainty

The probability $P(\mathbf{y} \mid \boldsymbol{\beta}, \mathbb{X}, I)$ is effectively determined by quantifying the model and measurement errors, while the prior probability $P(\mathbb{X} \mid I)$ is specified based on the degree of uncertainty surrounding the source profiles. Before proceeding with the expressions for the likelihood, prior and posterior PDFs (Sections 8.2.3, 8.2.4 and 8.2.5, respectively), we first explain the roles of errors in the CMB approach by relating the modelled and measured quantities to their true (but unknown) counterparts.

8.2.1.1 Measurement and model errors

Under a multiplicative noise assumption, the logarithm of the noise is additive and Gaussian, rendering the untransformed noise log-normally distributed. The measured species concentrations y_i are therefore related to the true quantities in the following way:

$$\begin{aligned} \log(y_i) = \log(y_i^{\text{true}}) + \log(\epsilon_i) \; , \qquad & \log(\epsilon_i) \sim N(0, \sigma_i) \; ; \\ y_i = \epsilon_i \; y_i^{\text{true}} \; , \qquad & \epsilon_i \sim LN(0, \sigma_i) \; , \end{aligned} \tag{8.9}$$

where $LN(0, \sigma_i)$ is a log-normal distribution based on a Gaussian distribution parameterized by a zero mean and a standard deviation of σ_i. We also assume that model uncertainty, like the measurement uncertainty, can be characterized as log-normal noise:

$$\begin{aligned} \log(\check{y}_i) = \log(y_i^{\text{true}}) + \log(\check{\epsilon}_i) \; , \qquad & \log(\check{\epsilon}_i) \sim N(0, \check{\sigma}_i) \; ; \\ \check{y}_i = \check{\epsilon}_i \; y_i^{\text{true}} \; , \qquad & \check{\epsilon}_i \sim LN(0, \check{\sigma}_i) \; . \end{aligned} \tag{8.10}$$

The probability density function (PDF) for the errors ϵ_i is given by

$$P(\epsilon_i) = \frac{1}{\epsilon_i \sigma_i \sqrt{2\pi}} \exp\left[-\frac{\log^2(\epsilon_i)}{2\sigma_i^2} \right] \; . \tag{8.11}$$

One consequence of assuming a multiplicative error structure is that existing data on measurements, source profiles and their variances must be converted to a form compatible with the log-normal distribution. For example, in the PDF of Eq. (8.11), the quantity σ_i expresses the standard deviation of the log-transformed ϵ_i. In order for Eq. (8.11) to properly describe the untransformed ϵ_i, tabulated measurement errors must be converted to express coefficients of variation (CV's) instead of standard deviations. The σ_i should then be chosen based on the CV's, instead of being extracted directly from the data. This conversion will be addressed in Section 8.2.2.

8.2.1.2 Source profile uncertainty

Source profile entries X_{ij} are assumed to be distributed log-normally about median values $X_{\circ,ij}$:

$$P(X_{ij} \mid I) = \frac{1}{X_{ij}\sigma_{X_{ij}}\sqrt{2\pi}} \exp\left[\frac{-\log^2\left(X_{ij}/X_{\circ,ij}\right)}{2\sigma^2_{X_{ij}}}\right] . \tag{8.12}$$

In Watson and Chow (2001), source profile estimates are stated in terms of average fractional abundances and their variances. We wish to choose the $X_{\circ,ij}$ to be equivalent to the median fractional abundances, so a conversion from mean to median fractional abundances is necessary. This conversion depends on the stated source profile variances, and is given by Eq. (8.15) in Section 8.2.2.1.

For certain source profiles, the list of species present is exhaustive (profile entries sum to 1). For these species, we specify the constraint equation:

$$\sum_{i=1}^{M} X_{ij} = 1 \ , \text{ for fully specified source profile } j. \tag{8.13}$$

8.2.1.3 Multiplicative vs. additive error

To justify using log-normal distributions to characterize both our measurement and model noise, we invoke the principle of maximum entropy (MaxEnt), as put forward by Jaynes (2003). MaxEnt provides a way of selecting distributions which are maximally noncommittal with respect to any information about them which remains unknown. If all that is known about our noise are its mean and variance, then according to the MaxEnt principle, the least informative distribution that can be chosen is the Gaussian.

In the CMB problem, we manipulate concentration and source profile data which may only take positive values. Uncertainties related to such information are commonly dealt with in a 'scaling fashion' in which uncertainties are specified as relative percentages of the measurement, rather than absolute plus-or-minus values (Jaynes, 2003; Sivia and Skilling, 2006). Indeed, the source profiles themselves are vectors of positive, fractional amounts. It therefore makes sense to deal with concentration data and source profiles in a logarithmic setting

(Tarantola, 2006). Adopting a Gaussian distribution for the logarithmic noise implies that the untransformed noise should be considered to belong to a log-normal distribution.

In some cases, individual measurements and source profile entries are specified as being zero. While the information contained in such zero values is undoubtedly important (such data could have a large impact on the source apportionment by permitting the outright rejection of source profiles identified by certain elements which act as markers), directly adopting a log-normal error distribution (noise prior) in these cases leads to an overemphasis on these zero values within the Bayesian framework. This overemphasis can be rectified by recognizing that measurements and source profiles are subject to lower limits of detection, which are almost certainly non-zero. We therefore replace zero entries by an estimate for the lower limit of detection and supply an appropriate value for the uncertainty.

8.2.2 Assigning distribution parameters

Although we wish to characterize uncertainties as log-normal, the measured concentration data and source profile information are actually supplied as sample mean and standard deviation information. Furthermore, the model uncertainty is not explicitly known and must be either assumed or inferred.

8.2.2.1 Measurement error σ_i and source profile parameters $\sigma_{X_{ij}}, X_{\circ,ij}$

Out of the available measures of central tendency (e.g., the mean, median, and mode), the mean is typically the more popular. However, when the data are known to be log-normally distributed, the median is a more appropriate measure of central tendency than the mean, for the following reasons (Slob, 1994):

1. The median of the log-normal distribution is the 'multiplicative analogue' of the mean: if the mean of $\log(x)$ is μ, then the median of x is $x_\circ = \exp(\mu)$. In contrast, the mean of x is $\exp(\mu + \sigma^2/2)$, which depends on the variance of the untransformed data [$\log(x)$].

2. The median is less sensitive than the mean to extreme values in the data. Log-normally distributed data are positive, and may be spread over several orders of magnitude. Interpreting such data from an absolute rather than a logarithmic standpoint can result in a mean value which is significantly higher than the location of the distribution's central region.

Slob goes on to argue that "the coefficient of variation is a natural measure of uncertainty in the log-normal distribution, since $\exp(\sigma^2)$ [...] can be recognized as the multiplicative analogue of σ^2, which has an additive nature." The coefficient of variation is the ratio of the standard deviation to the mean, which for the log-normal distribution is:

$$\mathrm{CV} = \frac{\text{standard deviation}}{\text{mean}} = \sqrt{\exp(\sigma^2) - 1}\,. \tag{8.14}$$

Adopting the CV as our preferred measure of uncertainty is especially useful in the context of the CMB model, where uncertainties are best represented as fractions of measured values (with the exception of zero-measurements, in which case the uncertainty would be representative of the lower limit of detection). We therefore assign measurement uncertainties σ_i using Eq. (8.14), with the CV taking the percentage value of the measurement standard deviation. For example, if a measurement takes a value of 100 units, and the measurement standard deviation is specified as 10 units, then σ_i is chosen such that CV$= 10\%$.

The source profile distribution parameters $X_{\circ,ij}$ are specified based on quoted estimates for the average abundances, which we denote $\mu_{X_{ij}}$; in other words, $\mu_{X_{ij}}$ represents the mean of the $(i,j)^{\text{th}}$ log-normal distribution for X_{ij}. Manipulating relationships for the mean and CV of a log-normal distribution, an expression is obtained which allows one to obtain the median value given the mean and CV:

$$X_{\circ,ij} = \mu_{X_{ij}} \left(1 + \text{CV}_{X_{ij}}^2\right)^{-1/2} \ . \tag{8.15}$$

The parameters $\sigma_{X_{ij}}$ are determined using Eq. (8.14), which implies that:

$$\sigma_{X_{ij}}^2 = \log(1 + \text{CV}_{X_{ij}}^2) \tag{8.16}$$

8.2.2.2 Model error $\check{\sigma}_i$

The CMB model rests on a number of assumptions which may be violated in unpredictable ways. It is difficult to derive a single useful uncertainty estimate for the model because it is implemented across many different scenarios. Traditional least-squares approaches to estimating source contributions through the CMB model carry a built-in assumption of a Gaussian, independent error structure. This assumption often leads to negative mass estimates which must be truncated or otherwise remedied in an ad hoc manner. The present approach avoids this problem by characterizing errors as multiplicative (log-normal), and guarantees that the estimated masses are positive. However, the scale of these errors remains unknown, and we require a methodology to deal with this 'uncertainty of the uncertainty'. The following is a partial list of possible approaches (in order of decreasing subjectivity):

1. Arbitrarily specify quantities for the model errors $\check{\sigma}_i$. Equivalently, specify the CV's for the model errors.

2. Select model errors to be proportional to a percentage of the concentration measurement errors (with both errors expressed as CV's). Use a single constant of proportionality to relate all of the model errors (we have just added a parameter to the inference). Specify a prior probability for this constant and incorporate it into the Bayesian inference scheme.

3. Consider each model error to be an unknown parameter which must be inferred. The parameter values can either be sampled using MCMC, or else analytically marginalized (if possible). Whether these parameters are sampled or marginalized will not affect the marginal posterior distributions for any of the other parameters.

In this work we select option #2 to account for model error. The posterior distribution must therefore incorporate a prior which assumes a state of ignorance with respect to the 'scale' parameter ρ. In a manner similar to Sivia and Skilling (2006), we adopt a form for the prior which does not require the specification of a finite upper bound on ρ:

$$P(\rho \mid I) = \frac{\rho_{\min}}{\rho^2} , \quad \rho \in [\rho_{\min}, \infty) , \tag{8.17}$$

where $\rho_{\min}$ is a conservative lower bound for ρ (selected by the user). This limit ensures that the PDF of Eq. (8.17) is normalizable. The model CV's are rewritten in terms of the measurement CV's: $\check{\text{CV}}_i \to \rho \text{CV}_i$, $i = 1, 2, \ldots, M$.

8.2.3 Assignment of the likelihood $P(\mathbf{y} \mid \boldsymbol{\beta}, \mathbb{X}, I)$

The likelihood is obtained by marginalizing the joint PDF of the measured (y_i) and true (y_i^{true}) data given the modelled ($\check{y}_i = \sum X_{ij}\beta_j$) data:

$$\begin{aligned} P(y_i \mid \boldsymbol{\beta}, \mathbf{X}_i, I) &= \int\limits_{\text{all } y_i^{\text{true}}} \mathrm{d}y_i^{\text{true}} \; P(y_i, y_i^{\text{true}} \mid \boldsymbol{\beta}, \mathbf{X}_i, I) \\ &= \int\limits_{\text{all } y_i^{\text{true}}} \mathrm{d}y_i^{\text{true}} \; P(y_i \mid y_i^{\text{true}}, \boldsymbol{\beta}, \mathbf{X}_i, I) \; P(y_i^{\text{true}} \mid \boldsymbol{\beta}, \mathbf{X}_i, I). \end{aligned} \tag{8.18}$$

Expanding and combining both expressions in the integrand (each is a log-normal PDF[2]), the likelihood can be written as:

$$\begin{aligned} P(y_i \mid \boldsymbol{\beta}, \mathbf{X}_i, I) = \int\limits_{\text{all } y_i^{\text{true}}} \mathrm{d}y_i^{\text{true}} \; & \frac{1}{2\pi\sigma_i\check{\sigma}_i y_i \check{y}_i} \\ & \times \exp\left[-\frac{\log^2(y_i/y_i^{\text{true}})}{2\sigma_i^2} - \frac{\log^2(\check{y}_i/y_i^{\text{true}})}{2\check{\sigma}_i^2} \right] , \end{aligned} \tag{8.19}$$

where $\check{\sigma}_i^2$ is the variance pertaining to the log of the i^{th} modelled datum ($\log \check{y}_i$) and σ_i^2 is the variance pertaining to the log of the i^{th} measured datum ($\log y_i$). Integration over $y_i^{\text{true}} \in (0, \infty)$

[2] A detailed derivation would show that each expression is obtained through a convolution integral, $\int \mathrm{d}\epsilon_i \; P(\epsilon_i) \; \delta(y_i - \epsilon_i y_i^{\text{true}})$, with $P(\epsilon_i)$ specified by Eq. (8.11).

yields the likelihood:

$$
\begin{aligned}
P(y_i \mid \boldsymbol{\beta}, \mathbf{X}_i, I) &= a_i \sqrt{\frac{\pi}{r_i}} \exp\left[\frac{(q_i+1)^2}{4r_i} - p_i\right] \ , \\
\text{where} \qquad a_i &= \frac{1}{2\pi\sigma_i \check{\sigma}_i y_i \check{y}_i} \ , \\
p_i &= \frac{\log^2 y_i}{2\sigma_i^2} + \frac{\log^2 \check{y}_i}{2\check{\sigma}_i^2} \ , \\
q_i &= \frac{\log y_i}{\sigma_i^2} + \frac{\log \check{y}_i}{\check{\sigma}_i^2} \ , \\
r_i &= \frac{1}{2\sigma_i^2} + \frac{1}{2\check{\sigma}_i^2} \ .
\end{aligned}
\tag{8.20}
$$

8.2.4 Assignment of the prior probabilities

We assume a state of ignorance with respect to each parameter β_j and therefore adopt a Jeffreys' prior (Sivia and Skilling, 2006), constrained by our estimates for the lower and upper bounds for β_j:

$$
P(\beta_j \mid I) = \frac{1}{\beta_j \log\left(\beta_j^{\max}/\beta_j^{\min}\right)} \ , \quad \beta_j \in [\beta_j^{\min}, \beta_j^{\max}] \ . \tag{8.21}
$$

Note that under this prior, $P(\log \beta_j \mid I) \sim$ constant for $\beta_j \in [\beta_j^{\min}, \beta_j^{\max}]$.

The prior distribution $P(X_{ij} \mid I)$ was given in Eq. (8.12), and as mentioned above, the prior distribution $P(\rho \mid I)$, Eq. (8.17), must also form part of the posterior distribution.

8.2.5 The full posterior distribution

Calculations are most easily performed using the logarithm of the posterior distribution, which takes the following form:

$$
\begin{aligned}
\log P(\boldsymbol{\beta}, \mathbb{X}, \rho \mid \mathbf{y}, I) = &\sum_{i=1}^{M} \left[\log\left(a_i \sqrt{\frac{\pi}{r_i}}\right) + \frac{(q_i+1)^2}{4r_i} - p_i\right] \\
&- \sum_{i=1}^{M}\sum_{j=1}^{N} \left[\log\left(X_{ij}\sigma_{X_{ij}}\sqrt{2\pi}\right) + \frac{\log^2\left(X_{ij}/X_{\circ,ij}\right)}{2\sigma_{X_{ij}}^2}\right] \\
&- \sum_{j=1}^{N} \log\left[\beta_j \log\left(\beta_j^{\max}/\beta_j^{\min}\right)\right] \\
&- 2\log\rho + \log\rho_{\min} + C \ ,
\end{aligned}
\tag{8.22}
$$

where C is a constant derived from taking the logarithm of the (unknown) normalization constant for the posterior distribution.

8.2.5.1 Gradients of the negative log-posterior

As explained below in Section 8.3, we require knowledge of the partial derivatives of the negative log-posterior PDF in order to effectively sample from it. Fortunately, these expressions are available analytically, and are reproduced below for completeness.

$$\frac{\partial}{\partial \beta_j}\left(-\log P\right) = \sum_{i=1}^{M} \frac{X_{ij}}{\check{y}_i}\left(\frac{\check{\sigma}_i^2 + \log \check{y}_i - \log y_i}{\check{\sigma}_i^2 + \sigma_i^2}\right) + \frac{1}{\beta_j} \tag{8.23a}$$

$$\frac{\partial}{\partial X_{ij}}\left(-\log P\right) = \frac{\beta_j}{\check{y}_i}\left(\frac{\check{\sigma}_i^2 + \log \check{y}_i - \log y_i}{\check{\sigma}_i^2 + \sigma_i^2}\right) + \frac{\log\left(X_{ij}/X_{\circ,ij}\right)}{\sigma_{X_{ij}}^2 X_{ij}} + \frac{1}{X_{ij}} \tag{8.23b}$$

$$\begin{aligned}\frac{\partial}{\partial \rho}\left(-\log P\right) = \frac{2}{\rho} + \sum_{i=1}^{M} & \frac{\rho \mathrm{CV}_i^2}{\left(\rho^2 \mathrm{CV}_i^2 + 1\right)\left[\sigma_i^2 + \log(\rho^2 \mathrm{CV}_i^2 + 1)\right]^2} \\ & \times \left[-\sigma_i^4 + \sigma_i^2\left(2\log(\check{y}_i/y_i) + 1\right) - \log^2\left(y_i/\check{y}_i\right) + \log(\rho^2 \mathrm{CV}_i^2 + 1)\right]\end{aligned} \tag{8.23c}$$

8.3 Exploring the posterior distribution with Markov chain Monte Carlo

For a source apportionment involving N sources and M species, the posterior distribution involves up to $N \times M + N + 1$ parameters. For typical problems, this number could lie between 10^2 and 10^3. Clearly, tesselating and evaluating the posterior PDF over such a high-dimensional parameter space is computationally impractical, so here we use Markov chain Monte Carlo (MCMC) as a technique for sampling from the posterior distribution.

MCMC methods are well-documented in the literature (Neal, 1993; Gilks et al., 1996; MacKay, 2003; Gregory, 2005) and provide a way to explore high-dimensional parameter spaces more efficiently than conventional Monte Carlo integration. MCMC algorithms work by generating a Markov chain of samples whose distribution tends asymptotically to a target distribution (in our case, the posterior PDF). This property ensures that time is not wasted generating samples from areas of the parameter space which contribute negligibly to the overall probability mass. We denote the sequence by $\mathbf{m}^{(k)} \in \mathbb{R}^{N_d}$, where N_d is the dimensionality of $\mathbf{m}^{(k)}$; $\mathbf{m}^{(k)}$ is the k^{th} sample; and m_i is the i^{th} component of the model parameter vector, $\mathbf{m}$.

The Metropolis-Hastings MCMC algorithm (which was described in chapter 4) is difficult to implement for the present source apportionment problem because, as Hajian (2007) points out, the efficiency[3] of this algorithm is inversely proportional to the number of parameters involved. Hajian advocates the use of Hamiltonian MCMC (originally due to Duane et al. (1987), the method is also known as 'hybrid Monte Carlo'), for which the efficiency remains constant with dimensionality. This method is slightly more difficult to implement than Metropolis-

[3] The statistical efficiency of a chain is basically a measure of the MCMC method's effectiveness at generating samples which accurately describe the target PDF. An inefficient chain will suffer from large sample mean variance. Details can be found in Gilks et al. (1996) and Hajian (2007).

Hastings, as it requires the use of auxiliary momentum variables and knowledge of the partial derivatives of the logarithm of the posterior PDF. Furthermore, all of our parameters of interest, $\mathbf{m} = (\boldsymbol{\beta}, \mathbb{X}, \rho)$, are necessarily positive and must be transformed so that exploration of the posterior distribution (which must also be transformed) can be carried out in log-parameter space. However, due to the hundreds of parameters present in our source apportionment problem, and due to the fact that partial derivatives of the log-posterior are available analytically, Hamiltonian MCMC remains a valuable technique for exploring our parameter space.

8.3.1 Hamiltonian MCMC: implementation

In the Hamiltonian MCMC method, we augment the model parameters $\mathbf{m}$ with a vector of auxiliary momentum variables $\mathbf{p}$ and construct the Hamiltonian,

$$H(\mathbf{m}, \mathbf{p}) = U(\mathbf{m}) + K(\mathbf{p}), \tag{8.24}$$

where U, the 'potential energy', is the negative logarithm of the target (posterior) PDF, and K, the 'kinetic energy', is a quadratic function of the momenta: $K(\mathbf{p}) = \frac{1}{2}\mathbf{p}^T\mathbf{p}$. The Hamiltonian (8.24) is explored in the way described by Algorithm 8.1.

The leapfrog moves serve to transport the proposal $\tilde{\mathbf{m}}$ through the Hamiltonian along trajectories of constant energy. In this way, mixing (the tendency of the chain to explore different regions of parameter space) happens with more intensity than in a chain generated by the Metropolis-Hastings algorithm. In theory, the acceptance rate for the method should be 100%; however, because the Hamiltonian dynamics are not simulated exactly (the leapfrog algorithm discretely approximates the true trajectory; step sizes ς are finite), the acceptance rate does not necessarily reach 100%. Good acceptance rates[4] are obtained when each step size is chosen to be proportional to the standard deviation of the chain: $\varsigma_i \propto \text{sd}(m_i)$. The number of leapfrog steps taken, N_{leap}, is also determined based on the desired rates of acceptance and chain exploration, and is subjected to small random perturbations at each step k in order to guarantee ergodicity. These issues, along with an analysis of Hamiltonian MCMC, are discussed in detail by Neal (1993).

8.3.2 Assessing chain convergence

Because the Markov chain of samples generated by an MCMC algorithm tends asymptotically to the target distribution, it is necessary to assess how well a chain of finite length approximates the target distribution. Hajian (2007) compares power spectra obtained using both Hamiltonian and Metropolis-Hastings MCMC, demonstrating the ability of the Hamiltonian method to achieve convergence more rapidly than Metropolis-Hastings for high-dimensional parameter spaces. For the rest of this chapter, we adopt the method of Dunkley et al. (2005), which was briefly described in Sec. 4.2 for assessing chain 'convergence'.

[4] In practice, a compromise is reached between acceptance rate and the speed of each chain's exploration. For this method applied to the source apportionment problem, we generally aim for acceptance rates higher than 50%.

Algorithm 8.1 Hamiltonian MCMC

select initial parameter values: $\mathbf{m}^{(0)}$
select leapfrog step sizes: $\boldsymbol{\varsigma}$
FOR $k = 1, 2, \ldots$ DO
  $\mathbf{p}^{(0)} \leftarrow$ *sample from* $N(0, 1)$
  $\boldsymbol{\tau} \leftarrow \mathbf{0}$
  $\tilde{\mathbf{m}}^{(\boldsymbol{\tau})} \leftarrow \mathbf{m}^{(k-1)}$
  FOR $n = 1, 2, \ldots, N_{\text{leap}}$ DO {execute leapfrogging}
    $\mathbf{p}^{(\boldsymbol{\tau}+\boldsymbol{\varsigma}/2)} \leftarrow \mathbf{p}^{(\boldsymbol{\tau})} - \frac{1}{2}\boldsymbol{\varsigma}^T \nabla_{\mathbf{m}} U(\tilde{\mathbf{m}}^{(\boldsymbol{\tau})})$
    $\tilde{\mathbf{m}}^{(\boldsymbol{\tau}+\boldsymbol{\varsigma})} \leftarrow \tilde{\mathbf{m}}^{(\boldsymbol{\tau})} + \boldsymbol{\varsigma}^T \mathbf{p}^{(\boldsymbol{\tau}+\boldsymbol{\varsigma}/2)}$
    $\mathbf{p}^{(\boldsymbol{\tau}+\boldsymbol{\varsigma})} \leftarrow \mathbf{p}^{(\boldsymbol{\tau}+\boldsymbol{\varsigma}/2)} - \frac{1}{2}\boldsymbol{\varsigma}^T \nabla_{\mathbf{m}} U(\tilde{\mathbf{m}}^{(\boldsymbol{\tau}+\boldsymbol{\varsigma})})$
    $\boldsymbol{\tau} \leftarrow \boldsymbol{\tau} + \boldsymbol{\varsigma}$
  END FOR {end leapfrogging}
  calculate acceptance probability:
  $\alpha \leftarrow \min\left(1, \exp\left[H(\mathbf{m}^{(k-1)}, \mathbf{p}^{(0)}) - H(\tilde{\mathbf{m}}^{(\boldsymbol{\tau})}, \mathbf{p}^{(\boldsymbol{\tau})})\right]\right)$
  $u \leftarrow$ *sample from* $\texttt{uniform}(0, 1)$
  IF $u < \alpha$ THEN
    $\mathbf{m}^{(k)} \leftarrow \tilde{\mathbf{m}}^{(\tau)}$ {accept the sample}
    ELSE
    $\mathbf{m}^{(k)} \leftarrow \mathbf{m}^{(k-1)}$ {reject the sample}
  END IF
END FOR

8.4 Test case: San Joaquin Valley Fine (SJVF) data

We evaluate the proposed Bayesian source apportionment methodology using a subset of the data collected at the Fresno site in California's San Joaquin Valley (SJV) as part of the 1988-1989 Valley Air Quality Study (VAQS). Details pertaining to this study (including the area's geography, dominant sources and measurement procedures) are presented in the work of Chow et al. (1992), where a source apportionment is performed using the US Environmental Protection Agency's (EPA) CMB software version 7.0 (Watson et al., 1990). In a similar vein to the study of Christensen and Gunst (2004), we perform a source apportionment using the $PM_{2.5}$ data and compare our results to those supplied in the EPA's CMB version 8.2 software users manual (Coulter, 2004). It should be noted that while Christensen and Gunst performed the source apportionment using data from each time period, and Seinfeld and Pandis (2006) performed the apportionment using annual average concentrations, our results are based on data from only a single observation (Feb 27, 1989, as listed in the manual). The full data set is available at the US EPA's internet site[5].

8.4.1 Source contribution estimates

Estimates for the contributions of each individual source are obtained directly from the chains of samples generated by the MCMC algorithm. Figure 8.1 illustrates the sample histograms corresponding to each source contribution parameter β_j. For each histogram, the median value along with 50% and 95% credible intervals (each credible interval contains a given percentage of the probability mass) are shown.

A comparison between the apportionment results illustrated in Figure 8.1 and those obtained using the EPA's CMB software (Coulter, 2004) is graphed and tabulated in Figure 8.2. The results are similar, although the uncertainty bounds are generally decreased. Furthermore, the estimate for crude oil burning lies much closer to zero.

8.4.2 Source profile estimates

One of the strengths of the proposed method lies in the way it accounts for uncertainty in the source profiles. The Bayesian methodology allows us to obtain new estimates for the source profile concentrations ($\hat{X}_{\circ,ij}$) and their uncertainties, based on prior information which consists of the profiles reported in the literature. Many source profile species concentrations are listed in the literature as being zero; however, as mentioned in Section 8.2.1.3, we replace these entries by an estimate for the lower limit of detection. For the present test case, this lower limit is assumed to be equal to the lowest non-zero quoted measurement uncertainty. Replacing these zero entries permits one to compare prior distributions for source profile species to their posterior distributions. Figure 8.3 compares prior and posterior distributions of species for the motor vehicles profile. Shifts in median estimates for certain species are

[5] `http://www.epa.gov/scram001/receptor_cmb.htm`

evident (e.g., silicon) while for other species, the posterior distribution appears much narrower than the prior.

Figure 8.4 illustrates the ratio $(\hat{X}_\circ / X_\circ)_{ij}$ for two different profiles; namely, wood burning and motor vehicles. Relatively large differences exist between the estimated and reported profiles for motor vehicles, whereas in contrast, the posterior estimate for the wood burning profile does not deviate greatly from its prior specification.

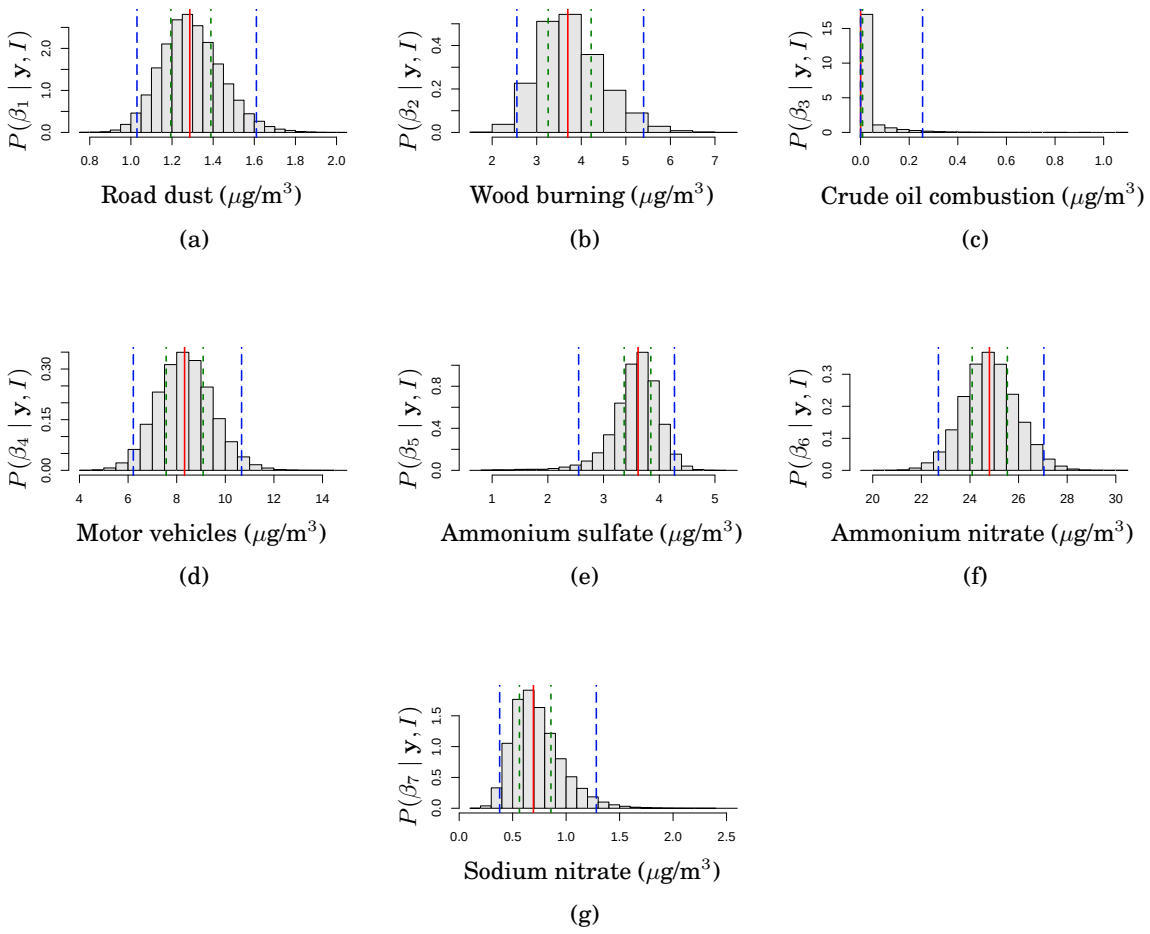
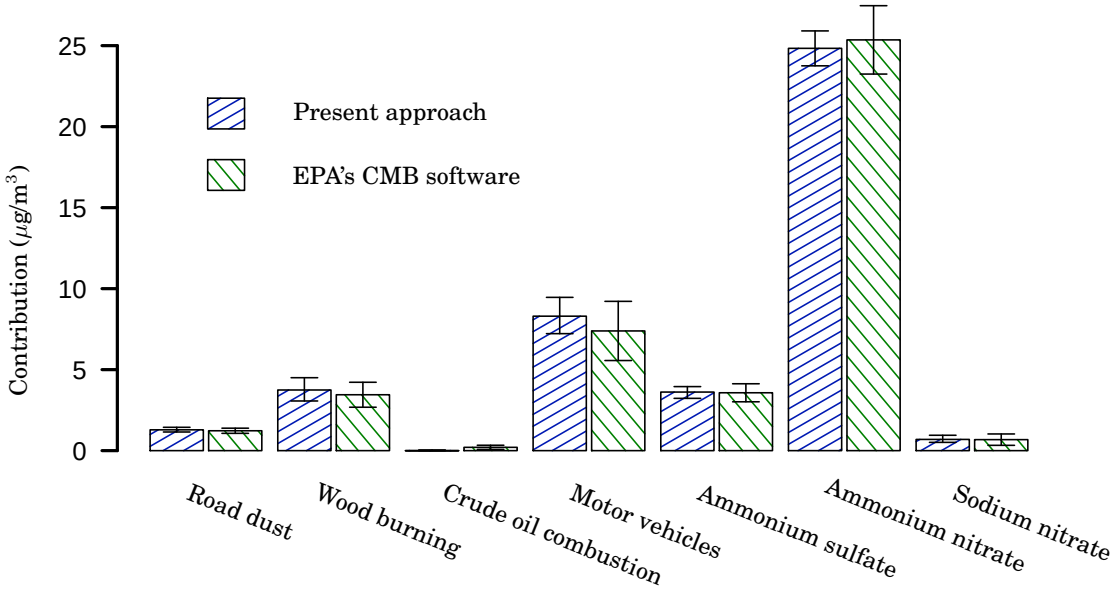


Figure 8.1: Marginal parameter distributions generated from MCMC samples. The solid lines are at median values, and the long and short dashed lines delineate 95% and 50% credible intervals (CI's), respectively.

8.4.3 Markov chain convergence

To complete this section on the validation of our Bayesian methodology, we present some details regarding the application of the MCMC algorithm (described in Section 8.3.1) to the SJVF test case.

The Hamiltonian MCMC method used to obtain the samples (binned and summarized in Figure 8.1) was run for 525×10^3 iterations, with each leapfrogging loop performing approximately 15 steps (this number was randomized to lie uniformly between 11 and 19). The



Source	Present approach		EPA's CMB software	
	Median value	68% CI width	Estimate	2×uncertainty
Road dust	1.29	0.29	1.23	0.32
Wood burning	3.74	1.43	3.45	1.54
Crude oil combustion	1.3×10^{-5}	4.24×10^{-2}	0.20	0.26
Motor vehicles	8.30	2.24	7.39	3.65
Ammonium sulfate	3.62	0.73	3.57	1.11
Ammonium nitrate	24.8	2.16	25.4	4.22
Sodium nitrate	0.693	0.438	0.680	0.702

Figure 8.2: Comparison between apportionments obtained using the present approach vs. those obtained using the EPA's CMB software. Error bars delineate 68% credible intervals (for the present approach) and estimated standard deviations (quoted by the EPA's CMB software).

algorithm resulted in an overall acceptance ratio of 69%. Initial values and step sizes ς were arrived at by running the algorithm iteratively using shorter chain lengths (on the order of 10^3–10^4 samples) in order to drive the chain behaviour to achieve a desirable acceptance ratio. The total number of iterations was chosen to be slightly larger than a power of two ($2^{19} = 524\,288$) due to the fact that the chain convergence criteria are assessed using a spectral technique which requires a fast Fourier transform (FFT). Commonly available FFT routines operate most efficiently on data whose length is a power of two, and convergence was not found to happen after 2^{17} iterations.

Convergence was assessed using the criteria of Dunkley et al. (2005), and convergence ratios were found to lie well below 0.01 for all parameters β and $\mathbb{X}$. A diagram of the thinned

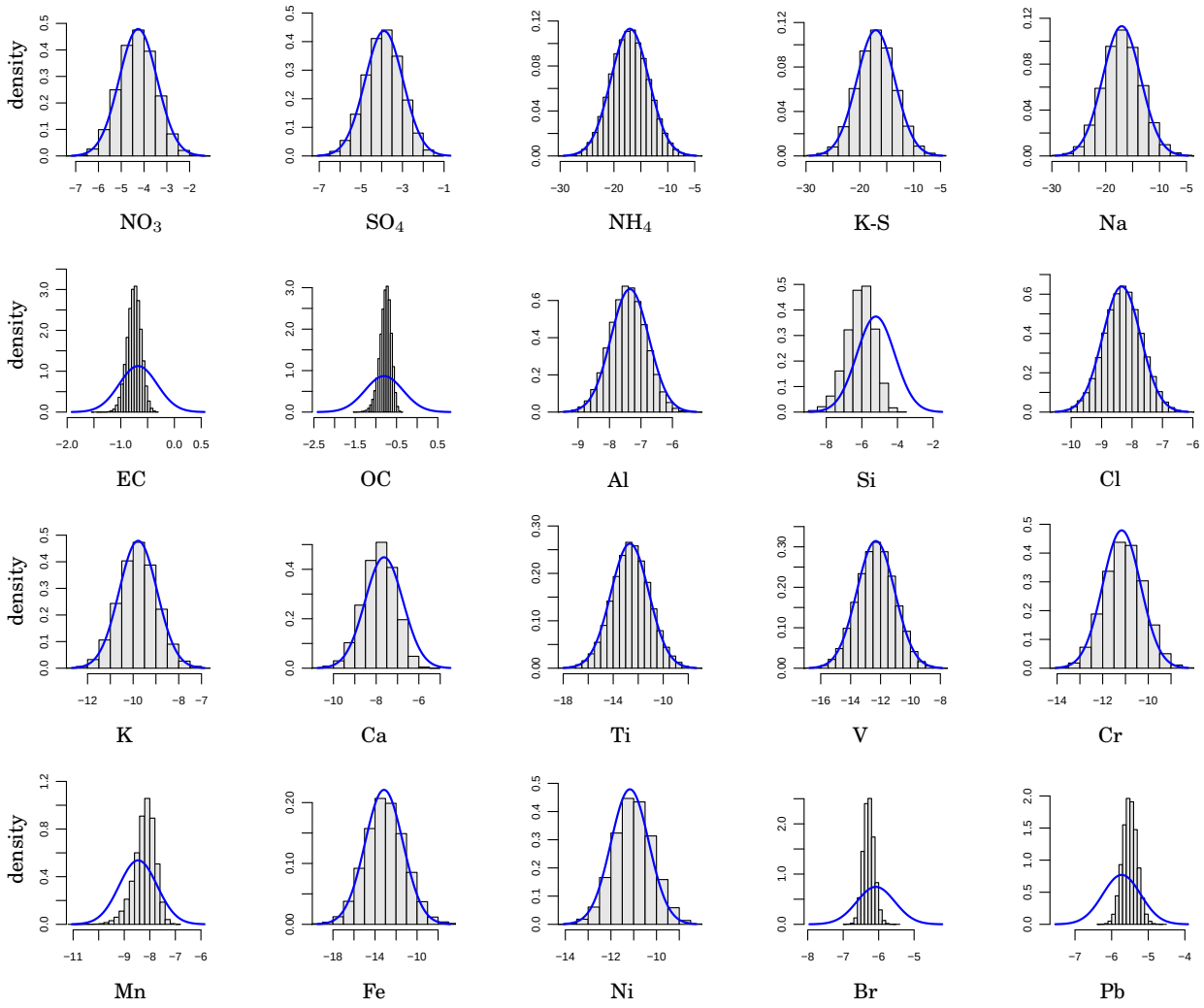
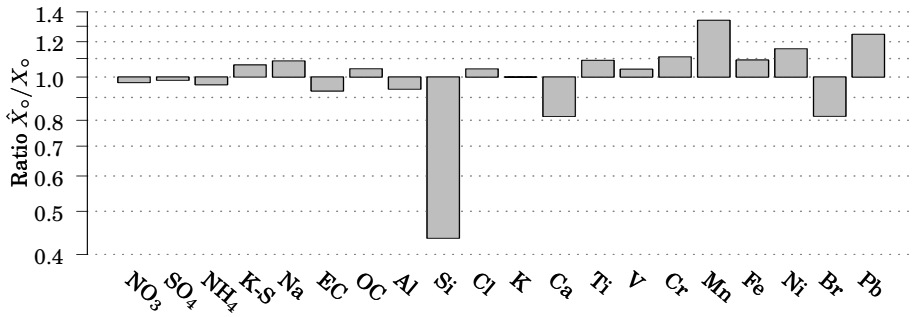


Figure 8.3: Motor vehicles source profile: comparison of prior (solid line) to posterior density (histogram) for each of the species in the profile. Horizontal axes are logarithmic.

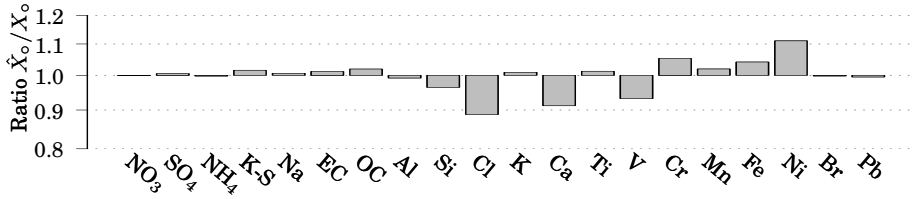
power spectrum[6] for the chain corresponding to β_1 (road dust), along with a parametric fit to the spectrum, can be seen in Figure 8.5. The parametric fit models the power spectrum according to a template proposed by Dunkley et al.:

$$\mathcal{P}(\eta) = \mathcal{P}_0 \frac{(\eta^*/\eta)^\gamma}{(\eta^*/\eta)^\gamma + 1} \ , \tag{8.25}$$

which effectively divides the spectrum into 'white noise' (the flat region to the left) and 'correlated noise' (the downward-sloping region at the right). Dunkley et al. showed that estimating $\mathcal{P}_0$ (or, $\mathcal{P}(\eta)$ as $\eta \to 0$) is equivalent to estimating the sample mean variance of a long chain. The parametric fit allows one to estimate this quantity easily, as well as the quantity η^*, which indicates the extent of the white noise regime. Chains which have entered the white noise regime no longer experience correlations at the largest scales, and are likely to be exploring the full region of high posterior probability.



(a) Profile: motor vehicles



(b) Profile: wood burning

Figure 8.4: Ratios of inferred to reported source profiles for motor vehicles and wood burning sources. Vertical scale is logarithmic.

[6] The full spectrum was obtained using 2^{19} samples.

8.5 Conclusions

Applying the multiplicative error assumption to the CMB model within a Bayesian framework has proven to yield estimates for the source contributions (and their associated uncertainties) which are consistent with results obtained using more traditional methods. Despite the apparent similarity between these sets of estimates, the following attributes of the proposed methodology are noteworthy:

1. The source profiles form part of the inference. We obtain posterior estimates for the source profiles which effectively improve model concentration estimates $\check{\mathbf{y}}$ with respect to the measured PM data, $\mathbf{y}$. It should be mentioned that source profile estimates cannot be obtained if the posterior distribution is analytically marginalized over the parameters $\mathbb{X}$.

2. The ability to account for an unknown model uncertainty. It is often difficult to objectively assess the impact that intrinsic features and assumptions of the model will have on the agreement between model results and the true (but unknown) data. Accounting for model uncertainty through an extra parameter ρ provides insight into how appropriate the CMB model is for a given scenario. For example, if the posterior estimate for the 'model error factor' (ρ) was large, it might indicate that the tabulated source profiles ($\mathbb{X}_\circ$) are inappropriate for the geographical area or data set under study (examination of the posterior estimates for the source profiles themselves would help to support or refute this hypothesis). Furthermore, many other sources of model uncertainty are possible,

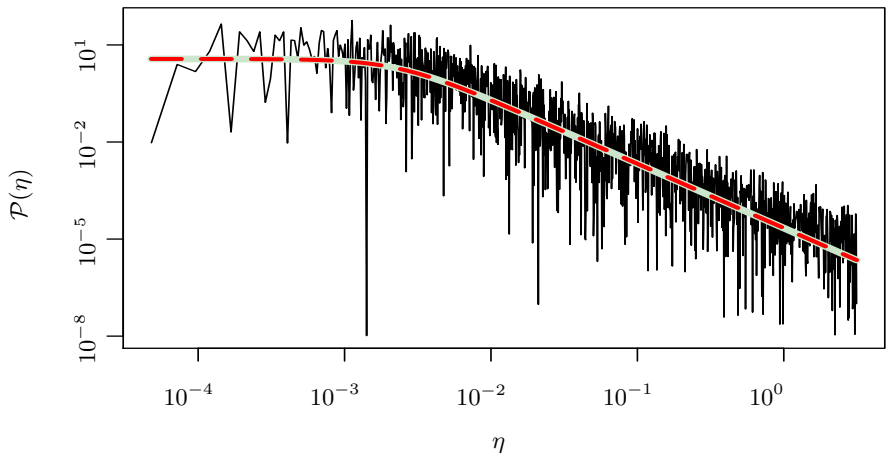


Figure 8.5: Power spectrum ($\mathcal{P}(\eta)$, cosmetically thinned) for the chain corresponding to the parameter β_1, generated by the Hamiltonian MCMC alogrithm. The horizontally flat, leftmost portion of the parametric fit (dashed line) indicates that the chain has entered the white noise regime.

such as the assumption that species do not react as they are transported from source to receptor. The Bayesian methodology permits one to account for model uncertainty in any number of ways, allowing competing hypotheses (regarding the applicability of the model) to be tested.

3. Hamiltonian MCMC permits the resulting high-dimensional parameter space to be sampled efficiently. A fortunate consequence of using the CMB receptor model is that gradients of the posterior distribution are available analytically, allowing one to use 'directed' MCMC methods such as the Hamiltonian method described in Section 8.3. For this method, efficiency remains constant with dimensionality, which is especially important since the SJVF test case involved 148 parameters.

To conclude, the proposed approach is of greater generality and flexibility than traditional methods, yet manages to yield estimates with competing (or better) degrees of uncertainty. While it is beyond the scope of this thesis to address issues surrounding source identification and selection, Bayesian approaches to model selection and hypothesis testing could be applied in a way compatible with the present framework.

Chapter 9

Conclusions and recommendations

The material presented in this thesis constitutes a kernel which could be grown or extended in several possible directions throughout the taxonomy of source determination problems. Each of the chapters in Part II has addressed an individual application along with its solution, obtained by coupling a Bayesian inferential framework together with suitable analytical and computational techniques. The contributions offered in this thesis lie in the blending of techniques to achieve the said solutions, as well as the fact that a number of problems pertaining to source determination in the atmospheric environment have now been collected and treated under a single global framework. Accordingly, this thesis concludes with a summary of the contributions presented in Part II, followed by suggestions for further investigation into problems related to source determination.

9.1 Summary of contributions

1. The main contribution made in this thesis is the demonstration of how source determination problems, when posed in a Bayesian framework, can be solved efficiently through the use of adjoint or backward dispersion models, together with MCMC for drawing samples from the posterior distribution. Chapter 5 demonstrated the feasibility of this approach for complex turbulent flows in built-up environments.

   Our commentary on the correct usage of the adjoint approach for source determination (Sec. 3.3) had the beneficial effect of generating discussion, and a comment-response was later published which allowed us to further illuminate the key issues behind the connection of the adjoint 'influence fields' to the PDF for the source location.

2. The source determination methodology was expanded in chapter 6 to involve nonconservative tracers–agents expected to undergo a first-order removal process[1] at an unknown or uncertain rate k_s. To this end, we introduced a novel extension to the Lagrangian

[1] Many physical processes are modelled using a first-order scheme; e.g., deposition, scavenging (dry or wet), and radioactive decay.

stochastic dispersion model, allowing influence field (C^*) information to be rapidly recalculated (estimated) for arbitrary rate constants.

3. A decision-theoretic framework (Bayesian adaptive exploration, or BAE) was applied to solve the optimal detector placement problem in chapter 7. Starting with a toy problem involving synthetic data, the potential usefulness of BAE was demonstrated, but at the same time a number of issues were uncovered which will need to be addressed in order for the method to be practical in more general scenarios. These issues are described below (Sec. 9.2).

4. Finally, an advanced MCMC method (hybrid Hamiltonian) was exploited to solve a source apportionment problem involving hundreds of unknown parameters. All relevant uncertainties in the model and data were taken into account at the outset, yet the posterior uncertainties associated with the source apportionment estimates were either comparable to or less than those obtained using traditional methods.

9.2 Recommendations for future work

Uncertainty quantification and model error specification

For all of the examples presented in this thesis, the functional forms for the model and measurement uncertainties are fixed (as either Gaussian or log-normal) and of either predetermined or unknown width. In the absence of available data and assumptions, these functional forms are appropriate (according to the MaxEnt principle). However, potential improvement in each parameter estimate (in terms of reducing its uncertainty) can only be obtained through improvement in the initial measurement and model error specifications, and for this reason it is recommended that further investigation be performed into uncertainty quantification (and characterization) for dispersion models.

The problem formulation presented in Sec. 3.2 began with the a priori assumption that the only information available about the noise (model error) was its mean and mean-square value. Application of the MaxEnt principle to this state of information (i.e., given mean and mean-square noise values) yielded a Gaussian distribution as the joint PDF for the noise (this being the maximally uninformative distribution given our 'state of knowledge' of the noise). Moreover, the form of this Gaussian distribution for the noise does not contain correlations (since this information was assumed to be unknown a priori). The reason for this is that given our assumed state of knowledge about the noise structure, a further (a priori known) constraint on the correlations must lower the entropy. In other words, a constraint on the noise correlation structure must lower the entropy (resulting in a more informative distribution), which in turn makes more precise the estimates of the (unknown) source parameters, if the correlation coefficients themselves are known.[2]

[2] However, if we do not know the correlation structure of the noise a priori, and introduce a specific form (model) for the correlation structure with some unspecified parameters which also need to be inferred from the concentration data, this case may not lead to more precise estimates for the source parameters (since now the same data must

From this perspective, the particular form of the Gaussian distribution for the noise assigned using the MaxEnt principle given the a priori information on only the mean and mean-square values of the noise (correlations unknown) should not be viewed as not containing correlations – rather, it is more precise to state that this form of the Gaussian distribution for the noise effectively allows for every possible (noise) correlation structure that could be present and, as a consequence, is less informative than a correlated distribution. In other words, the functional form for the Gaussian distribution makes no specific assumptions about the unknown noise correlation structure (for which it was assumed that no information was available a priori).

Given, however, the observed nature of model errors (as seen from, e.g., the concentration profiles obtained from the MUST array experiment, Fig. 5.4), the assumption that correlations are unknown a priori is somewhat naïve and ought to be improved upon. We would begin by recognizing that in reality, for the atmospheric dispersion problems considered in this thesis, the structure and causes of dependency between model errors are expected to be highly complex. It is probably the case that errors in mean concentration depend strongly on the accuracy of the flow field, which depends in turn on the underlying flow model's ability to resolve turbulence quantities and mean wind velocities. This ability depends further on the specification of initial conditions such as wind direction and turbulent velocity profiles, as well as boundary conditions such as surface roughness and temperature.

Having shortlisted potential sources of model uncertainty, one could consider, as a first-order approximation, that there exists a temporal or spatial correlation structure to the model errors, quantified by a characteristic time or length scale. In such a case, it is interesting to examine the problem from a MaxEnt viewpoint as seen in Sec. 8.5 of Sivia and Skilling (2006). Sivia and Skilling demonstrate that when all that is known about the prospective error covariance is a characteristic nearest-neighbour correlation strength, maximizing the entropy of the likelihood function results in a multivariate Gaussian whose covariance matrix exhibits a power-law type of decay. In other words, starting with a specification for only the diagonal and first set of off-diagonal terms, the MaxEnt principle can be used to obtain a complete specification for the rest of the off-diagonal entries in the covariance matrix. Alternative covariance matrices could be specified as a result of different assumptions being made regarding the off-diagonal terms. Having specified a form for the covariance matrix, the next tasks would be to estimate (infer) specific values for its entries, and to assess the suitability of its specified form (e.g., power-law vs. linear decay).

Evaluating competing hypotheses pertaining to model error structure is a problem related to uncertainty quantification, and may call for techniques of Bayesian model comparison to be applied. For example, consider the case where a sufficiently large data set of measured concentration and flow field data of known quality is available to be used for model validation. A Bayesian approach to the uncertainty quantification problem might begin by estimating the parameters of a model error specification (e.g., the terms in a covariance matrix) using available data. However, the terms in the covariance matrix may be of a given form or struc-

also be used to constrain the unknown noise parameters). For example, in chapter 6, lack of knowledge of decay rate resulted in an increase in the uncertainty in the source strength parameter.

ture and we wish to assess whether or not a competing specification is more appropriate. By calculating and comparing evidence terms, $P(D \mid I_1)$ vs. $P(D \mid I_2)$, the relative merits of each different specification can be assessed. As mentioned in Sec. 2.3.2, however, calculating evidence terms is a nontrivial exercise, and it is expected that this particular aspect of uncertainty quantification would require substantial research effort.

Complex source configurations

While the material in chapter 5 formulated the source determination case for a time-dependent source, and while problems involving multiple point sources have already been examined by Yee (2008), more general source models remain to be addressed under the Bayesian framework. The adjoint approach described in Sec. 5.2.2 improves the efficiency of source-receptor computations not only for point sources, but also for line, area, and volume sources. Issues arise, however, when we wish to characterize and parameterize such multidimensional sources. For example, a set of basis functions may be required in order to characterize the source for a given application. In such a case, the source would be parameterized by basis function coefficients, of which there may be a varying number, which might in turn necessitate the use of a dimension-jumping MCMC technique [as was used in the multiple source problem of Yee (2008)]. Depending on the dimensionality of the parameter space, a hybrid Hamiltonian MCMC technique (or other guided method) may also be required in order to ensure timely convergence. Clearly, the issue has very quickly become complicated and therefore further investigation into the efficient solution of problems involving arbitrary source configurations is recommended.

Improvement in computational efficiency for experimental design

As mentioned in the previous section, investigating the optimal detector placement problem (chapter 7) yielded a number of issues which remain open. Of the three issues described in Sec. 7.5, two stand out and bear restating. Firstly, the way that detector uncertainties (both modelled and measured) are specified has a relatively large effect on the shape of the surface defined by the expected information ($E\mathcal{I}$) as a function of auxiliary detector location. This underscores the importance of our first recommendation, which was that properly defining the uncertainties involved in source determination is an important part of the problem. Secondly, the optimal detector placement problem is computationally challenging in that many potential detector locations must be proposed in order for an optimization algorithm to lead the proposal to the location where $E\mathcal{I}$ is maximized. Future research in this area could perhaps concentrate on developing heuristic criteria to aid the optimization procedure.

Appendix A

Nomenclature

Propositions

A, B, M, D, I

Spaces

Ω	computational domain (parameter space)
$\mathcal{R}$	physical domain
$\mathcal{D}$	subset of physical domain (grid cell or support of kernel function)

Operators and matrices

$\mathbb{L}, \mathbb{L}^*$	forward and adjoint dispersion operators
$\mathbb{X}$	source profile matrix
$\mathbf{X}_i$	row of $\mathbb{X}$ corresponding to i^{th} species
X_{ij}	entry of $\mathbb{X}$; quantity of i^{th} species contributed by j^{th} source

Vectors and components

$\mathbf{d}, d_i$	measured data (mean concentration or scalar fluxes)
$\mathbf{r}, r_i$	modelled data (mean concentration or scalar fluxes)
$\mathbf{m}, m_i$	source parameters (indexed by i but of different dimension from data)
$\boldsymbol{\beta}, \beta_j$	source contributions
$\mathbf{y}, y_i$	measured species concentrations
$\check{\mathbf{y}}, \check{y}_i$	measured species concentrations
σ_i, CV_i	standard deviation, coefficient of variation; i^{th} modelled/measured concentration

$\mathbf{p}, p_i$	auxiliary momentum variables used for Hamiltonian MCMC
$\boldsymbol{\varsigma}, \varsigma_i$	step sizes used for leapfrog method

Note: $\mathbf{m}$, $\mathbf{p}$ and $\boldsymbol{\varsigma}$ have been indexed by i but are of different dimension to $\mathbf{d}$ and $\mathbf{r}$.

$\mathbf{x}_s$	source location
$\mathbf{x}_d$	detector location

Field variables and functions

$C(\mathbf{x})$	concentration field
$C^*(\mathbf{x})$	conjugate concentration field
$\hat{\tau}(\mathbf{x})$	sample average LS particle travel time field
$\sigma_\tau^2(\mathbf{x})$	particle travel time variance field
$\overline{\mathbf{u}'c'}(\mathbf{x})$	turbulent scalar fluxes
$\mathbf{U}(\mathbf{x})$	Reynolds averaged wind velocity
$Q(\mathbf{x}; \mathbf{x}_s, q_s)$	source density distribution
$h(\mathbf{x}; \mathbf{x}_d)$	detector response function
$\mathcal{I}(\blacksquare)$	indicator function; returns '1' if argument is true, otherwise '0'
$P(A \mid B, I)$	probability of A given B and I (probability density function)
$\mathcal{I}[A \mid B, I]$	information content of PDF: $\int \mathbf{d}A \; P(A \mid B, I) \; \log \; P(A \mid B, I)$
$E\mathcal{I}(e)$	expected information as a function of chosen experiment e
$H(\mathbf{m}, \mathbf{p})$	Hamiltonian function
$\mathcal{P}(\eta)$	power spectrum as a function of wavenumber η

Individual parameters and other scalar quantities

x_s, y_s, z_s	source x, y, z locations
q_s	source strength
k_s	rate of decay of nonconservative tracer
ρ	scale parameter (unknown scale of model errors)
α	acceptance probability (MCMC algorithms)

References

K. J. Allwine, K. L. Clawson, M. J. Leach, D. Burrows, R. Wayson, J. Flaherty, and E. Allwine. Urban dispersion processes investigated during the Joint Urban 2003 study in Oklahoma City. Vancouver, British Columbia, Canada, Aug 2004. 5th Conference on Urban Environment.

M. M. Aral, J. Guan, and M. L. Maslia. Identification of contaminant source location and release history in aquifers. *Journal of Hydrologic Engineering*, 6:225–234, 2001.

P. A. Ariya, B. T. Jobson, R. Sander, H. Niki, G. W. Harris, J. F. Hopper, and K. G. Anlauf. Measurements of C_2–C_7 hydrocarbons during the Polar Sunrise Experiment 1994: Further evidence for halogen chemistry in the troposphere. *Journal of Geophysical Research*, 103: 13169–13180, 1998.

S. Pal Arya. *Air Pollution Meteorology and Dispersion*. Oxford University Press, New York, 1999.

M. L. Barad, editor. *Project Prairie Grass, A Field Program in Diffusion*. Geophysical Research Papers No. 59, Vols. I and II. Report AFCRC-TR-58-235, Air Force Cambridge Research Center, 439pp., 1958.

T. Bayes and R. Price. An Essay towards Solving a Problem in the Doctrine of Chances. By the Late Rev. Mr. Bayes, F. R. S. Communicated by Mr. Price, in a Letter to John Canton, A. M. F. R. S. *Philosophical Transactions of the Royal Society of London*, 53:370–418, 1763.

J.M. Bernardo. Expected information as expected utility. *Annals of Statistics*, 7:686–690, 1979.

D. Billheimer. Compositional receptor modeling. *Environmetrics*, 12:451–467, 2001.

M. E. Borsuk and C. A. Stow. Bayesian parameter estimation in a mixed-order model of BOD decay. *Water Research*, 34:1830–1836, 2000.

S.P. Brooks. Bayesian computation: a statistical revolution. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 361:2681–2697, 2003.

K. Chaloner and I. Verdinelli. Bayesian experimental design: a review. *Statistical Science*, 10:273–304, 1995.

C-C. Chan, C-K. Nien, and J-S. Hwang. Receptor modeling of VOCs, CO, NO_x, and THC in Taipei. *Atmospheric Environment*, 30:25–33, 1996.

C-P. Chio, M-T. Cheng, and C-F. Wang. Source apportionment to PM_{10} in different air quality conditions for Taichung urban and coastal areas, Taiwan. *Atmospheric Environment*, 38: 6893–6905, 2004.

F. K. Chow, B. Kosović, and S. Chan. Source inversion for contaminant plume dispersion in urban environments using building resolving simulations. Fairfax, Virginia, USA, Jul 2005. 9th Annual George Mason University Conference on Atmospheric Transport and Dispersion Modeling.

F. K. Chow, B. Kosović, and S. Chan. Source Inversion for Contaminant Plume Dispersion in Urban Environments Using Building-Resolving Simulations. *Journal of Applied Meteorology and Climatology*, 47:1553–1572, 2008.

J. C. Chow, J. G. Watson, D. H. Lowenthal, P. A. Solomon, K. L. Magliano, S. D. Ziman, and L. W. Richards. PM_{10} source apportionment in California's San Joaquin Valley. *Atmospheric Environment*, 26A:3335–3354, 1992.

W. F. Christensen and R. F. Gunst. Measurement error models in chemical mass balance analysis of air quality data. *Atmospheric Environment*, 38:733–744, 2004.

M. Colyvan. The philosophical significance of Cox's theorem. *International Journal of Approximate Reasoning*, 37:71–85, 2004.

C. T. Coulter. EPA-CMB8.2. Users manual EPA-452/R-04-011, Office of Air Quality Planning & Standards, Research Triangle Park, NC, Dec 2004.

R. T. Cox. Probability, frequency and reasonable expectation. *American Journal of Physics*, 14:1–13, 1946.

G.T. Csanady. *Turbulent Diffusion in the Environment*. Kluwer Academic Pub, 1973.

A. S. Denning, M. Holzer, K. R. Gurney, M. Heimann, R. M. Law, P. J. Rayner, I. Y. Fung, S-M. Fan, S. Taguchi, P. Friedlingstein, Y. Balkanski, J. Taylor, M. Maiss, and I. Levin. Three-dimensional transport and concentration of SF_6. A model intercomparison study (TransCom 2). *Tellus B*, 51:266–297, 1999.

S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth. Hybrid Monte Carlo. *Physics Letters B*, 195:216–222, 1987.

J. Dunkley, M. Bucher, P. G. Ferreira, K. Moodley, and C. Skordis. Fast and reliable MCMC for cosmological parameter estimation. *Monthly Notices of the Royal Astronomical Society*, 356:925–936, 2005.

D. Estep. A short course on duality, adjoint operators, Green's functions, and a posteriori error analysis. Course notes, Department of Mathematics, Colorado State University, Aug 2004. URL `http://www.math.colostate.edu/~estep/research/preprints/adjointcourse_final.pdf`.

J.E. Fackrell and A.G. Robins. Concentration fluctuations and fluxes in plumes from point sources in a turbulent boundary layer. *Journal of Fluid Mechanics*, 117:1–26, 1982.

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

M. Fuentes, A. Chaudhuri, and D.M. Holland. Bayesian entropy for spatial sampling design of environmental data. *Environmental and Ecological Statistics*, 14:323–340, 2007.

L.-E. De Geer. Sniffing out clandestine tests. *Nature*, 382:491–492, 1996.

F.A. Gifford. Computation of pollution from several sources. *International Journal of Air Pollution*, 2:109–110, 1959.

W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, editors. *Markov Chain Monte Carlo in Practice*. Chapman & Hall/CRC, London, 1996.

A. Goyal, M. J. Small, K. von Stackelberg, D. Burmistrov, and N. Jones. Estimation of fugitive lead emission rates from secondary lead facilities using hierarchical Bayesian models. *Environmental Science and Technology*, 39:4929–4937, 2005.

P. C. Gregory. *Bayesian Logical Data Analysis for the Physical Sciences: A Comparative Approach with Mathematica® support*. Cambridge University Press, Cambridge, 2005.

J. Hadamard. Sur les problèmes aux dérivées partielles et leur signification physique. *Princeton University Bulletin*, pages 49–52, 1902.

A. Hajian. Efficient cosmological parameter estimation with Hamiltonian Monte Carlo technique. *Physical Review D*, 75:083525 (11 pages), 2007.

S. R. Hanna, J. S. Chang, and D. G. Strimaitis. Uncertainties in source emission rate estimates using dispersion models. *Atmospheric Environment*, 24A:2971–2890, 1990.

S. R. Hanna, O. R. Hansen, and S. Dharmavaram. FLACS CFD air quality model performance evaluation with Kit Fox, MUST, Prairie Grass, and EMU observations. *Atmospheric Environment*, 38:4675–4687, 2004.

W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109, 1970.

D. A. Haugen, editor. *Project Prairie Grass, A Field Program in Diffusion*. Geophysical Research Papers No. 59, Vol. III. Report AFCRC-TR-58-235, Air Force Cambridge Research Center, 673pp., 1959.

T. Hilderman and R. Chong. A laboratory study of momentum and passive scalar transport and diffusion within and above a model urban canopy – MUST Array report. Report CRDC00327c, Prepared by Coanda Research and Development Corporation for Defence Research and Development Canada – Suffield, Aug 2004.

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

J.-P. Issartel and J. Baverel. Inverse transport for the verification of the Comprehensive Nuclear Test Ban Treaty. *Atmospheric Chemistry and Physics*, 3:475–486, 2003.

E. T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, Cambridge, 2003.

H. Jeffreys. *Scientific Inference*. Cambridge University Press, Cambridge, 1931.

H. Jeffreys. *Theory of Probability*. Clarendon Press, Oxford, 1939.

A. Jeremić and A. Nehorai. Landmine detection and localization using chemical sensor array processing. *IEEE Transactions on Signal Processing*, 48:1295–1305, 2000.

N. Kashiwagi. Chemical mass balance when an unknown source exists. *Environmetrics*, 15: 777–796, 2004.

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

A. Keats, E. Yee, and F-S. Lien. Reply to the 'Comments on: "Bayesian inference for source determination with applications to a complex urban environment" '. *Atmospheric Environment*, 41:5547–5551, 2007b.

A. Keats, E. Yee, and F-S. Lien. Efficiently characterizing the origin and decay rate of a nonconservative scalar using probability theory. *Ecological Modelling*, 205:437–452, 2007c.

A. Keats, M-T. Cheng, E. Yee, and F-S. Lien. Bayesian treatment of a chemical mass balance receptor model with multiplicative error structure. *Atmospheric Environment*, 43:510–519, 2009.

B. M. Kim and R. C. Henry. Diagnostics for determining influential species in the chemical mass balance receptor model. *Journal of the Air and Waste Management Association*, 49: 1449–1455, 1999.

B. Kosović, G. Sugiyama, F. K. Chow, K. Dyer, W. Hanley, G. Johannesson, S. Larsen, G. Loosmore, J. K. Lundquist, A. Mirin, J. Nitao, R. Serban, and C. Tong. Stochastic source inversion methodology and optimal sensor network design. Fairfax, Virginia, USA, Jul 2005. 9th Annual George Mason University Conference on Atmospheric Transport and Dispersion Modeling.

F-S. Lien, E. Yee, H. Ji, W. A. Keats, and K-J. Hsieh. Development of a high-fidelity numerical model for hazard prediction in the urban environment. St. John's, Newfoundland, Canada, Aug 2005a. CFD Society of Canada, 13th Annual Conference of Computational Fluid Dynamics.

F.-S. Lien, E. Yee, and J. D. Wilson. Numerical modelling of the turbulent flow developing within and over a 3-D building array, Part II: A mathematical foundation for a distributed drag force approach. *Boundary-Layer Meteorology*, 114:245–285, 2005b.

E. Limpert, W. A. Stahl, and M. Abbt. Log-normal distributions across the sciences: Keys and clues. *BioScience*, 51:341–352, 2001.

C.-H. Lin and L.-F. W. Chang. Relative source contribution analysis using an air trajectory statistical approach. *Journal of Geophysical Research*, 107:ACH6.1–10, 2002.

J. C. Lin, C. Gerbig, S. C. Wofsy, A. E. Andrews, B. C. Daube, K. J. Davis, and C. A. Grainger. A near-field tool for simulating the upstream influence of atmospheric observations: The Stochastic Time-Inverted Lagrangian Transport (STILT) model. *Journal of Geophysical Research*, 108:ACH6.1–13, 2003.

D. V. Lindley. On a measure of the information provided by an experiment. *Annals of Mathematical Statistics*, 27:986–1005, 1956.

J.W. Lingwall, W.F. Christensen, and C.S. Reese. Dirichlet based Bayesian multivariate receptor modeling. *Environmetrics*, 19:618–629, 2008.

F. Liu, Y. Zhang, and F. Hu. Adjoint method for assessment and reduction of chemical risk in open spaces. *Environmental Modeling and Assessment*, 10:331–339, 2005.

T. J. Loredo. Bayesian Adaptive Exploration. In G. J. Erickson and Y. Zhai, editors, *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, volume 707 of *American Institute of Physics Conference Series*, pages 330–346, April 2004. doi: 10.1063/1.1751377.

D. J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, Cambridge, 2003.

A. Marmur, J. A. Mulholland, and A. G. Russell. Optimized variable source-profile approach for source apportionment. *Atmospheric Environment*, 41:493–505, 2007.

J. Matthes, L. Gröll, and H. B. Keller. Source localization by spatially distributed electronic noses for advection and diffusion. *IEEE Transactions on Signal Processing*, 53:1711–1719, 2005.

N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equations of state calculations by fast computing machine. *Journal of Chemical Physics*, 21:1087–1091, 1953.

T. P. Meyers, P. Finkelstein, J. Clarke, T. G. Ellestad, and P. F. Sims. A multilayer model for inferring dry deposition using standard meteorological measurements. *Journal of Geophysical Research*, 103:22645–22661, 1998.

A. M. Michalak and P. K. Kitanidis. Application of Bayesian inference methods to inverse modeling for contaminant source identification at Gloucester Landfill, Canada. *Computational Methods in Water Resources*, 2:1259–1266, 2002.

A. M. Michalak and P. K. Kitanidis. A method for enforcing parameter nonnegativity in Bayesian inverse problems with an application to contaminant source identification. *Water Resources Research*, 39:SBH7.1–14, 2003.

R. M. Neal. Probabilistic inference using Markov chain Monte Carlo methods. Technical Report CRG-TR-93-1, Department of Computer Science, University of Toronto, Sep 1993.

A. Nehorai, B. Porat, and E. Paldi. Detection and localization of vapor-emitting sources. *IEEE Transactions on Signal Processing*, 43:243–253, 1995.

E. S. Park, M-S. Oh, and P. Guttorp. Multivariate receptor models and model uncertainty. *Chemometrics and Intelligent Laboratory Systems*, 60:49–67, 2002.

E.S. Park, P. Guttorp, and R.C. Henry. Multivariate Receptor Modeling for Temporally Correlated Data by Using MCMC. *Journal of the American Statistical Association*, 96:1171–1183, 2001.

M. Patan, C. Tricaud, and Y. Chen. Resource-constrained sensor routing for parameter estimation of distributed systems. In *Proceedings of the 17th World Congress, The International Federation of Automatic Control, Seoul, Korea, July 6–11, 2008.*, pages 7772–7777, 2008.

V. Penenko, A. Baklanov, and E. Tsvetova. Methods of sensitivity theory and inverse modeling for estimation of source parameters. *Future Generation Computer Systems*, 18:661–671, 2002.

J. A. Pudykiewicz. Application of adjoint tracer transport equations for evaluating source parameters. *Atmospheric Environment*, 32:3039–3050, 1998.

J. A. Pudykiewicz. Comments on: "Bayesian inference for source determination with applications to a complex urban environment" by A. Keats, E. Yee, and Fue-Sang Lien. *Atmospheric Environment*, 41:5541–5551, 2007.

S. S. Qian, C. A. Stow, and M. E. Borsuk. On Monte Carlo methods for Bayesian inference. *Ecological Modelling*, 159:269–277, 2003.

R. Ramesh. Bhopal still suffering, 20 years on. *The Guardian*, Nov 2004. URL `http://www.guardian.co.uk/world/2004/nov/29/india.randeepramesh`.

K. S. Rao. Uncertainty Analysis in Atmospheric Dispersion Modeling. *Pure and Applied Geophysics*, 162:1893–1917, 2005.

R. Rapley, P. Robins, and P. A. Thomas. Source term estimation and probabilistic sensor modelling. Fairfax, Virginia, USA, Jul 2005. 9th Annual George Mason University Conference on Atmospheric Transport and Dispersion Modeling.

P. Reichert and M. Omlin. On the usefulness of overparameterized ecological models. *Ecological Modelling*, 95:289–299, 1997.

H. C. Rodean. Stochastic Lagrangian models of turbulent diffusion. *Meteorological Monographs*, 26, August 1996.

C. Rödenbeck, S. Houweling, M. Gloor, and M. Heimann. CO_2 flux history 1982–2001 inferred from atmospheric data using a global inversion of atmospheric transport. *Atmospheric Chemistry and Physics*, 3:1919–1964, 2003.

P. Sebastiani and H. P. Wynn. Maximum entropy sampling and optimal Bayesian experimental design. *Journal of the Royal Statistical Society: Series B*, 62:145–157, 2000.

P. Seibert and A. Frank. Source-receptor matrix calculation with a Lagrangian particle dispersion model in backward mode. *Atmospheric Chemistry and Physics*, 4:51–63, 2004.

J. H. Seinfeld and S. N. Pandis. *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change*. John Wiley and Sons, Hoboken, 2006.

I. Senocak, N. W. Hengartner, M. B. Short, and W. B. Daniel. Stochastic event reconstruction of atmospheric contaminant dispersion using Bayesian inference. *Atmospheric Environment*, 42:7718–7727, 2008.

G. Shafer. Comments on "Constructing a logic of plausible inference: a guide to Cox's Theorem", by Kevin S. Van Horn. *International Journal of Approximate Reasoning*, 35:97–105, 2004.

C. Silva and A. Quiroz. Optimization of the atmospheric pollution monitoring network at Santiago de Chile. *Atmospheric Environment*, 37:2337–2345, 2003.

N. Singpurwalla and A. Wilson. Probability, chance and the probability of chance. *IIE Transactions*, 41:12–22, 2009.

D. S. Sivia and J. Skilling. *Data Analysis: A Bayesian Tutorial*. Oxford University Press, 2006.

Y. N. Skiba. On a method of detecting the industrial plants which violate prescribed emission rates. *Ecological Modelling*, 159:125–132, 2003.

J. Skilling. Nested sampling for general Bayesian computation. *Bayesian Analysis*, 1:833–860, 2006.

W. Slob. Uncertainty analysis in multiplicative models. *Risk Analysis*, 14:571–576, 1994.

M. D. Sohn, P. Reynolds, N. Singh, and A. Gadgil. Rapidly locating and characterizing pollutant releases in buildings. *Journal of the Air and Waste Management Association*, 52: 1422–1432, 2002.

A. Tarantola. *Inverse Problem Theory and Methods for Model Parameter Estimation*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 2005.

A. Tarantola. *Elements for Physics: Qualities, Quantities, and Intrinsic Theories*. Springer, Berlin, 2006.

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

A. N. Tikhonov. *Solutions of ill-posed problems*. Winston, Washington, 1977.

C. G. Tinney, R. P. Butler, G. W. Marcy, H. R. A. Jones, A. J. Penny, C. McCarthy, B. D. Carter, and J. Bond. Four new planets orbiting metal-enriched stars. *Astrophysical Journal*, 587: 423–428, 2003.

B-J. Tsuang. Quantification on the source/receptor relationship of primary pollutants and secondary aerosols by a Gaussian plume trajectory model: Part I–theory. *Atmospheric Environment*, 37:3981–3991, 2003.

B-J. Tsuang, C-L. Chen, C-H. Lin, M-T. Cheng, Y-I. Tsai, C-P. Chio, R-C. Pan, and P-H. Kuo. Quantification on the source/receptor relationship of primary pollutants and secondary aerosols by a Gaussian plume trajectory model: Part II. Case study. *Atmospheric Environment*, 37:3993–4006, 2003.

A. Venkatram and S. Du. An analysis of the asymptotic behavior of cross-wind-integrated ground-level concentrations using Lagrangian stochastic simulation. *Atmospheric Environment*, 31:1467–1476, 1997.

J. Wang and N. Zabaras. A Bayesian approach to the inverse heat conduction problem. *International Journal of Heat and Mass Transfer*, 47:3927–3941, 2004.

J. G. Watson and J. C. Chow. Source characterization of major emission sources in the Imperial and Mexicali Valleys along the US/Mexico border. *The Science of the Total Environment*, 276:33–47, 2001.

J. G. Watson, N. F. Robinson, J. C. Chow, R. C. Henry, B. M. Kim, T. G. Pace, E. L. Meyer, and Q. Nguyen. The USEPA/DRI Chemical Mass Balance receptor model, CMB 7.0. *Environmental Software*, 5:38–49, 1990.

J. D. Wilson and W. K. N. Shum. A re-examination of the integrated horizontal flux method for estimating volatilisation from circular plots. *Agricultural and Forest Meteorology*, 57: 281–295, 1992.

J.D. Wilson, G.W. Thurtell, and G.E. Kidd. Numerical simulation of particle trajectories in inhomogeneous turbulence, III: Comparison of predictions with experimental data for the atmospheric surface layer. *Boundary-Layer Meteorology*, 21:443–463, 1981.

S. Wu and J.V. Zidek. An entropy-based analysis of data from selected NADP/NTN network sites for 1983-1986. *Atmospheric environment. Part A, General topics*, 26:2089–2103, 1992.

E. Yee. Probabilistic inference: an application to the inverse problem of source function estimation. Melbourne, Australia, 31 January – 4 February 2005. The Technical Cooperation Program (TTCP) Chemical and Biological Defence (CBD) Group Technical Panel 9 (TP-9) Annual Meeting, Defence Science and Technology Organization.

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

E. Yee and C. A. Biltoft. Concentration fluctuation measurements in a plume dispersing through a regular array of obstacles. *Boundary-Layer Meteorology*, 111:363–415, 2004.

E. Yee and J.D. Wilson. Instability in Lagrangian stochastic trajectory models, and a method for its cure. *Boundary-Layer Meteorology*, 122:243–261, 2007.

E. Yee, R. M. Gailis, A. Hill, T. Hilderman, and D. Kiel. Comparison of wind tunnel and water channel simulations of plume dispersion through a large array of obstacles with a scaled field experiment. *Boundary-Layer Meteorology*, 121:389–432, 2006.

E. Yee, F.S. Lien, and H. Ji. Technical Description of Urban Microscale Modeling System: Component 1 of CRTI Project 02-0093RD. DRDC Suffield TR 2007-067, Defence R&D Canada – Suffield, Ralston, Alberta, 2007.

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

J.V. Zidek, W. Sun, and N.D. Le. Designing and integrating composite networks for monitoring multivariate Gaussian pollution fields. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 49:63–79, 2000.