

AMERICAN UNIVERSITY OF BEIRUT

Stochastic Transport and Identification of
Moving Passive Tracer Sources with Application
to Marine Traffic in the Mediterranean Sea

by

Alexios Boutros Rustom

A thesis

submitted in partial fulfillment of the requirements
for the degree of Master of Engineering
to the Department of Mechanical Engineering
of the Maroun Semaan Faculty of Engineering and Architecture
at the American University of Beirut

Beirut, Lebanon
August 2020

AMERICAN UNIVERSITY OF BEIRUT

Stochastic Transport and Identification of
Moving Passive Tracer Sources with Application
to Marine Traffic in the Mediterranean Sea

by
Alexios Boutros Rustom

Approved by:

Prof. Issam Lakkis, Professor
Mechanical Engineering

Advisor



Prof. Omar Knio, Professor
Applied Mathematics and Computational Science, KAUST

Member of Committee



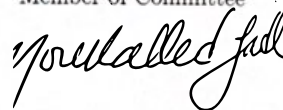
Prof. Ibrahim Hoteit, Professor
Earth Science and Engineering, KAUST

Member of Committee



Prof. Fadl Moukalled, Professor
Mechanical Engineering

Member of Committee



Date of thesis defense: August 10, 2020

AMERICAN UNIVERSITY OF BEIRUT

THESIS, DISSERTATION, PROJECT
RELEASE FORM

Student Name: Rustom Alexios Boutros
Last First Middle

Master's Thesis Master's Project Doctoral Dissertation

I authorize the American University of Beirut to: (a) reproduce hard or electronic copies of my thesis, dissertation, or project; (b) include such copies in the archives and digital repositories of the University; and (c) make freely available such copies to third parties for research or educational purposes.

I authorize the American University of Beirut, to: (a) reproduce hard or electronic copies of it; (b) include such copies in the archives and digital repositories of the University; and (c) make freely available such copies to third parties for research or educational purposes after: **One** **year from the date of submission of my thesis, dissertation or project.**
Two ___ years from the date of submission of my thesis, dissertation or project.
Three ___ years from the date of submission of my thesis, dissertation or project.

Alexios Rustom

Signature

August 17, 2020

Date

Acknowledgements

”Not to us, O LORD, not to us but to your name be the glory, because of your love and faithfulness.” (Psalm 115:1)

First and foremost, I would like to thank God for His endless graces and His presence throughout my life. I have always felt His guiding hand leading me to make the right choices and take the right actions.

I would like to express my sincere appreciation to my advisor Prof. Issam Lakkis for allowing me to conduct research under his supervision. I am truly grateful for his confidence and the freedom he gave me during my studies at AUB. Without his illuminating advices and his coherent and insightful instruction, this thesis work would not have reached its current form.

I would also like to thank Prof. Omar Knio for giving me the opportunity to intern within his group at the King Abdullah University of Science and Technology (KAUST). I am grateful for his advice and experience which provided an invaluable learning environment that was tremendously helpful.

I would also like to thank my thesis committee Prof. Ibrahim Hoteit and Prof. Fadl Moukalled for their encouragement and follow-up during my research work.

Finally, I would like to extend, from the bottom of my heart, my gratitude to my parents Boutros and Antoinette for their countless sacrifices, love and support. I am also grateful for having my sister Michella by my side throughout my whole life. Words cannot express how grateful I am for my aunt Sayde kindness and welcoming heart during my master’s degree.

An Abstract of the Thesis of

Alexios Boutros Rustom for Master of Engineering
Major: Mechanical Engineering

Title: Stochastic Transport and Identification of Moving Passive Tracer Sources with Application to Marine Traffic in the Mediterranean Sea

Source Reconstruction Problems are widely used for the aim of determining the sources of contaminants and pollutants in the case of deliberate or accidental release scenarios. This involves the inverse problem and inference of the sources parameters given a set of observed and measured data for direct emergency actions and services while estimating at the same time the nature of the threat in order to avoid and manage escalating consequences.

The aim of this thesis report is to introduce the methodology and results of the inference and source reconstruction problems with an application to contaminants and pollutants transport in the Mediterranean Sea in the presence of a stochastic velocity field. This will allow immediate and accurate actions in emergency cases such as any unexpected release scenario.

The forward Lagrangian model, adopted in transporting the pollutants, is first introduced followed by a detailed discussion of the full implementation of the stochastic transport of moving passive tracers in the Mediterranean Sea in the presence of a stochastic velocity field. This also involves building the probability maps for various release scenarios along a given ship path.

In addition, a new sampling based approach, similar to MCMC algorithms and based on the evaluation of the cost function between the modelled likelihood of the contributing events and the observation patch, is introduced. This algorithm allows the inference of single and multiple sources in the Mediterranean Sea with a quantitative measure of the relative contribution of these sources to the observation patch while quantifying the uncertainty in the solution. The results obtained from this proposed inference algorithm are illustrated and are in accordance with the true solution obtained using a deterministic optimization approach.

Contents

Acknowledgements	v
Abstract	vi
1 Introduction	1
2 Literature Review	5
2.1 Problem Statement	5
2.1.1 Source Term Estimation	5
2.1.2 Forward and Inverse Problems	6
2.2 Inverse Problem Approaches	6
2.2.1 Deterministic Optimization Approaches	7
2.2.1.1 Genetic Algorithm (GA)	8
2.2.1.2 Pattern Search Method (PSM)	8
2.2.1.3 Summary on Deterministic Optimization Approaches	9
2.2.2 Bayesian Probabilistic Approaches	9
2.2.2.1 Baye’s Theorem	10
2.2.2.2 Markov Chain Monte Carlo Sampling	11
2.2.2.3 Summary on Bayesian Probabilistic Approaches	12
3 Methodology	14
3.1 Problem Statement	14
3.1.1 Stochastic Velocity Field	14
3.1.2 Observations	15
3.1.3 Moving Source	15
3.2 Building the likelihood	17
3.2.1 Simulations	19
3.2.1.1 Studied Cases for speed improvement of the sim- ulations	22
3.3 Observation Patches	28
3.3.1 Deterministic simulation using the mean (MEAN) of the stochastic velocity field	28
3.3.2 Synthetic Observation patches	28

3.3.3	Scaling Coefficient Calculation	29
3.4	Cost Function	29
3.5	Sampling Algorithm	30
3.6	Optimization Algorithm	36
4	Results and Discussion	37
4.1	Forward Problem	37
4.1.1	Impact of Adaption of λ and σ_r on the inference problem .	37
4.2	Inference Problem	40
4.2.1	Inference of a Single Source	40
4.2.2	Inference of Multiple Sources	41
5	Conclusion and Future Work	58
A	Abbreviations	59

List of Figures

1.1	Examples of oil spills accidents.	3
1.2	Oil spills locations due to oil tankers according to ITOPF.	4
2.1	Steps required for any inverse problem or source reconstruction approach.	13
3.1	Advection of a particle at time t_a	15
3.2	Flow chart of the forward advection algorithm.	16
3.3	Density Maps in the Mediterranean Sea.	17
3.4	Path of the ship from Suez Canal until the Strait of Gibraltar. . .	18
3.5	Discretized Ship Path.	19
3.6	Difference between the observation and injection times for the 1037 events and 22 trajectories under study.	20
3.7	Ship path (white line) and the full stochastic probability map for Trj. DayShift0 ($N_{max} = 100 \times 10^6$, $\Delta t_{adv} = 1$ hr).	24
3.8	Ship path (white line) and the full stochastic probability map for Trj. DayShift7 ($N_{max} = 10^6$, $\Delta t_{adv} = 3$ hr).	24
3.9	Comparison between the three cases (Trj. DayShift7) for Event300. . .	25
3.10	Pure Stochastic Probability maps of Trj. DayShift21 for Event400. . .	26
3.11	Deterministic Probability maps of Trj. DayShift21 for Event400. . .	27
3.12	Flow Chart of the minimization of the calculation of the distance between a satellite image (a) and a probability map (c).	31
4.1	Probability Maps in Trj. DayShift17 along the ship path (white line).	38
4.2	Effect of the adaptation parameters on the cost function variation (minimum cost function shown in red dashed line).	39
4.3	Investigation of the effect of the size of the box on the cost function in Case 1.	43
4.4	Summary of Results in Case 1.	44
4.5	Synthetic Observation Patch of Case 2 obtained from the probability maps of Events 549 and 599.	45
4.6	Results in Case 2.	46
4.7	Correlation Maps in Case 2.	47

4.8	Marginal Posterior Probabilities in Case 2.	48
4.9	Synthetic Observation Patch of Case 3 obtained from the probability maps of Events 300 and 400.	49
4.10	Results in Case 3.	50
4.11	Correlation Maps in Case 3.	51
4.12	Marginal Posterior Probabilities in Case 3.	52
4.13	4 Separate Patches in Trj. DayShift17 generated using the MEAN of the velocity field.	53
4.14	Inferred Results in Case 4 with adaptivity of σ_r	54
4.15	Correlation Maps in Case 4.	55
4.16	Marginal Posterior Probabilities of sources 1 and 2 in Case 4.	56
4.17	Marginal Posterior Probabilities of sources 3 and 4 in Case 4.	57

List of Tables

3.1	Some considered stochastic trajectories	21
3.2	Several Cases for the Probability Map for Event400 of Trj. DayShift21.	22
4.1	Single Source Inference problem in Trj. DayShift17.	40
4.2	Sampling Algorithm parameters of the Single Source Inference problem in Trj. DayShift17.	40
4.3	Multiple Source Inference problems in Trj. DayShift17.	42
4.4	Sampling Algorithm parameters of the Multiple Source Inference problems in Trj. DayShift17.	42

Chapter 1

Introduction

In the case of a chemical, biological, and radiological release events, either deliberately or accidentally, extensive research studies have been conducted in the purpose of determining quickly and accurately the probable sources of contaminants as well as their corresponding characteristics (release time, strength, etc.) for direct emergency actions and services with an estimate of the nature of the threat in order to manage the consequences. Some examples of catastrophic releases are listed below:

- In 1995, 12 people were killed and more than 50 were injured in a Tokyo subway where a nerve agent Sarin was intentionally released [1].
- It has been reported by the China Statistical Yearbook that 9339 events of water contamination took place between 1997 and 2008 and caused many damages on the social and economic levels [2].
- In November 2005, around 100 tons of benzene were spilled in the Songhua-jiang River in China [2].
- The Jiyeh oil spill after explosion of the storage tanks at the thermal power station in 2006. Figure 1.1a illustrates the contaminated beaches of Beirut.
- In 2012, 3 million deaths were caused by air pollution [3].
- In 2019, a pipeline firm *Plains All American Pipeline* gets \$3.3 million fine for causing in 2015 the worst California coastal spill in 25 years in Refugio State Beach in Santa Barbara County. This has blackened popular beaches, killed wildlife, and hurt both tourism and fishing. This was due to the fact that the firm failed to quickly detect the ruptured pipeline and responded slowly to this release event. Figure 1.1b illustrates the effects of the crude oil spill on the Refugio State Beach.

Furthermore, according to the International tanker Owners Pollution Federation (ITOPF), most of the locations of oil spills due to oil tankers are present in the Middle East region and most specifically in the Mediterranean Sea as shown in figure 1.2. Therefore, this report will introduce the development and implementation of source reconstruction approaches with an application to Marine traffic in the Mediterranean Sea for future immediate emergency actions in case of pollutants and contaminants release accidents.



(a) Contaminated beaches of Beirut due to the Jiyeh Oil Spill.



(b) Ruptured pipeline spilled 140,000 gallons of crude oil into the Refugio State Beach.

Figure 1.1: Examples of oil spills accidents.

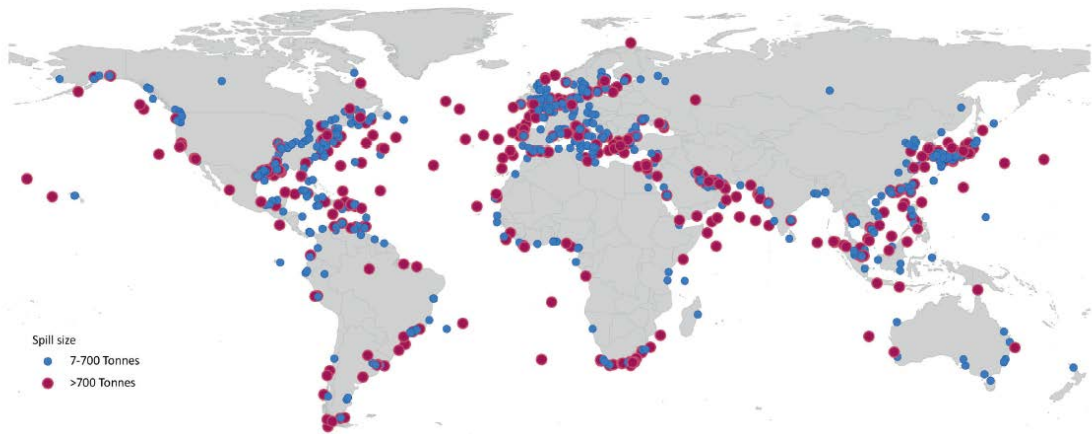


Figure 1.2: Oil spills locations due to oil tankers according to ITOPF.

Chapter 2

Literature Review

The objective of this literature survey is to give a general overview about some important concepts related to the thesis work. An overview of the techniques and concepts used in solving the inverse problem and source term reconstruction is addressed. Several algorithms and methods used in the identification of the source parameters (location, strength, release time, etc.) are introduced. This general literature review will help in directing the work of this thesis and its aim and purpose of identifying probable sources of contaminants with application to marine traffic in the Mediterranean Sea using a stochastic velocity field.

2.1 Problem Statement

2.1.1 Source Term Estimation

The most general term of the problem of interest is called Source Determination that involves the inference of the parameters of some given sources (location, strength, ON-OFF time, size, area, and even the identity of any pollutant being released). This determination is also characterized and known as a source inversion problem or a Source Term Estimation (STE) which corresponds to the inverse problem that consists of characterizing the source and its parameters based on a set of observations [4] and sensor measurements.

Those kind of problems are present in many areas of science and mathematics and are a very challenging research field. They have been applied in many research areas including the localization of pollutants, chemicals, and contaminants in the Atmosphere [3] [5] [6][7][8] [9] Sea [10], ground water aquifer [11], rivers [2], quantification of the emission of certain chemicals from the ground to the atmosphere, determination of the origin and decay rate of non-conservative scalar [12], as well as the determination of the parameters of extra solar planets. These problems were also applied to the estimation of the parameters (mostly the composition) of solid objects using X-rays tomography, determination of explosives

in an airport luggage or in mail packages, and identification of unknown number of Land Mines [13].

2.1.2 Forward and Inverse Problems

Most of the work in the literature focuses on the determination of the dispersion of a certain passive quantity given some known parameter sources. This is called the forward problem and it involves the modelling and dispersion of tracers in the environment (atmosphere, river, ocean, etc.). On the other hand, the estimation and inference of the parameters of an unknown source given a certain output characterized by certain observations and measurements is called the inverse problem. Both forward and inverse problems are related through an operator \mathbb{A} which is non-linear in most cases. This operator relates the system output \mathbf{D} (data measurements and observations) to a system input (parameters \mathbf{M} of the source):

$$\mathbf{D} = \mathbb{A}(\mathbf{M}) \quad (2.1)$$

In other terms, this operator \mathbb{A} maps from the Hilbert space of the parameters \mathbf{M} to the Hilbert space of the data \mathbf{D} :

$$\mathbb{A} : \mathcal{M} \rightarrow \mathcal{D} \quad (2.2)$$

The inverse problem is solved by constructing the inverse operation \mathbb{A}^{-1} , and the determination of the parameters \mathbf{M} from the set of measured data \mathbf{D} using the inverse of the operator \mathbb{A} :

$$\mathbf{M} = \mathbb{A}^{-1}(\mathbf{D}) \quad (2.3)$$

Those inverse problems are ill-posed:

- \mathbb{A} may be singular and an inverse transformation may not exist.
- The solution may not be unique and multiple input solutions for a given system output may exist. Any small perturbation in the measured or observed data may change dramatically the solution.
- The solution to this problem is highly affected by measurement and model errors because of the limited amount of collected data.

2.2 Inverse Problem Approaches

Two different approaches are used in solving the source reconstruction problem: the deterministic optimization approach that seeks to obtain a single optimized solution for the inverse source reconstruction problem while taking input without

uncertainty, and the stochastic Bayesian approach that seeks multiple solutions to the ill-posed problem while evaluating the degree of plausibility of each configuration of the solution.

To note that the goal of the STE methods consists of the estimation and inference of the unknown source parameters by a simple fusion of observations or measurement data with any prior information of the model parameter space.

Before discussing the various approaches in the STE problems, it is important to note that the modelled data is determined using some dispersion and transport models. The most used models in the literature consist of the Lagrangian and Eulerian dispersion models[14].

2.2.1 Deterministic Optimization Approaches

These methods involve the selection of an optimum configuration “best guess” from a possibly an infinite set of configurations of the source distributions in the presence of incomplete and noisy data while optimizing the objective or cost function in order to obtain a deterministic solution of the parameters of the source without any quantification of the uncertainty of this solution.

A broad range of deterministic optimization approaches was derived from the sum of the square difference between the measured and modelled data, and will be introduced sequentially.

The optimum solution \hat{M} of the parameters is determined by minimizing the residual:

$$\hat{M} = arg \min_{\mathbf{M} \in \mathcal{M}} \|\mathbf{A}(\mathbf{M}) - \mathbf{D}\|^2 \quad (2.4)$$

Where $\|\bullet\|$ represents the Euclidean norm. For linear problems, the solution is straightforward. However, in many real-world applications (non-linear problems), the solution to this problem is not unique and unstable given the ill-posed inverse problem under study.

A more stabilized formulation of the cost function is based on the regularized least-square approach where the objective function \mathbb{J} in the residual method is optimized with the use of some regularized parameters β that apply a stable approximation to the inverse operator:

$$\mathbb{J} = \|\mathbf{A}(\mathbf{M}) - \mathbf{D}\|^2 + \beta\Phi(\mathbf{M}) \quad (2.5)$$

The first term indicates the discrepancy between the measured and modelled concentration and the second term $\Phi(\mathbf{M})$ is the regularization functional that is adopted for treating instabilities in the inverse problem, and β is the regularizing parameter.

A common choice for the regularization functional is the Tikhonov regularization and the optimal \hat{M}_β is:

$$\hat{M}_\beta = arg \min_{\mathbf{M} \in \mathcal{M}} \|\mathbf{A}(\mathbf{M}) - \mathbf{D}\|^2 + \beta^2\|\mathbf{M}\|^2 \quad (2.6)$$

In general, gradient search algorithms seek to minimize the discrepancy between the measured and observed data, and the direction of the descent is determined by the gradient (first or second derivative) of the objective function. Such optimization approach is a local search algorithm that is highly dependent on the initial guess of the parameters.

More sophisticated algorithms in the literature are the Genetic Algorithm (GA), Pattern Search Method (PSM), and the hybrid algorithm.

2.2.1.1 Genetic Algorithm (GA)

The GA is a global search approach which is a widely used AI global optimization approach in the source reconstruction problems (especially non-linear and non-convex ones). The intelligent optimization algorithm behind the GA is based on the process of natural evolution. It consists of the following steps [11]:

- a. Generation of a random population of the source parameters called chromosomes (initialization). These source parameters are then encoded.
- b. The cost function is evaluated in order to measure the fitness F of the solutions (selection).
- c. A mating of high quality solutions is performed in order to generate new estimates of the parameters and a second generation population of the solutions that are higher in quality than the previous generation (Mating).
- d. Similar to the process of evolution, a selection of chromosomes are mutated in order to generate another set of solutions (Mutation).
- e. The termination is checked. If convergence is not satisfied, steps (b) to (e) are repeated.

The used model is solved for a number of times that is equal to the population size in order to obtain for example the modelled concentrations at specific locations, and therefore suffers from being time-consuming for a high dimensional parameter space.

To note that in this algorithm, it is of great importance to tune the population size, the mutation rate, the mating strategy, and the range of parameters in order to yield accurate and efficient results for the model parameters.

2.2.1.2 Pattern Search Method (PSM)

The PSM is a local search algorithm, and consists of the following steps:

- A. The theoretical parameters, that will be determined, are defined with their initial guesses.

- B. The algorithm will vary these parameters by increasing or decreasing them using a constant factor.
- C. The objective function is then calculated directly for the new set of parameters without any use of derivatives.
- D. If there is no variation in the cost function from the previous iteration, the pattern moves (step size is halved) and the previous steps are repeated until the convergence criteria is satisfied.

This method was tested in the literature [14] and was highly dependent on the initial guess of the model parameters. For this purpose, a hybrid algorithm was developed in [14], where the PSM was coupled with a Genetic Algorithm (GA) (a global optimization method) that serves as a tool for the generation of a reasonable initial guess for the parameters in the Pattern Search Method PSM. This hybrid algorithm has yielded more accurate and more efficient results compared to the case where the PSM is only used because of benefiting from both the local and global search methods [15].

2.2.1.3 Summary on Deterministic Optimization Approaches

An overview of the most used optimization approaches was addressed in solving the inverse and the Source Term Estimation (STE) problems. In the optimization approach, a requirement of no or little prior information gives advantage over other methods such as the Bayesian Approach. However, the presence of contextual information will yield more efficient and accurate results.

2.2.2 Bayesian Probabilistic Approaches

Probabilistic methods (probability modelling methods) [12] in which the inverse and source reconstruction problems are solved from a Bayesian perspective, and the final solution to this problem takes the form of a probability density function that encapsulates all the relevant information about the parameters of the source with an estimation of the uncertainty in the solution. In addition, these methods provide a logical framework for source determination, and deal well with uncertainty in the input data and the model.

These methods, based on Bayes' theorem, consist of determining the solution to the inverse problem using some incorporated prior knowledge of the parameters and some limited and noisy observations and data measurements.

Furthermore, this method overcomes the convergence issue to local minima that exists in most of the optimization methods. If this method is used with an efficient sampling scheme, it is less sensitive to the starting point and initial guess of the chain of the algorithm.

2.2.2.1 Baye's Theorem

Baye's theorem is formulated in the following equation:

$$P(\underbrace{\mathbf{M}|\mathbf{D}, I}_{\text{Posterior}}) = \frac{\overbrace{P(\mathbf{M}|I)}^{\text{prior}} \overbrace{P(\mathbf{D}|\mathbf{M}, I)}^{\text{Likelihood}}}{\underbrace{P(\mathbf{D}|I)}_{\text{Evidence}}} \quad (2.7)$$

Where \mathbf{M} represents the model parameters, \mathbf{D} represents the data (measured or observed) used to improve our estimates of the model, and I represents the background information related to the data and the model.

The parameters of equation (2.7) are:

- The prior probability $P(\mathbf{M}|I)$ represents our state of knowledge about the model parameters \mathbf{M} given the background information I before the arrival of the collected data \mathbf{D} . Usually, no prior information is known and is required to reflect the ignorance about the proposal they describe (based on the Maximum Entropy Principle). The prior probability is assumed to be a uniform distribution for Cartesian variables (quantities that lie from $(-\infty, +\infty)$ like position and velocity), and should satisfy scale invariance for Jeffreys variables (positive quantities defined by their inverse).
- The likelihood function (as function of \mathbf{M}) quantifies the discrepancy between the synthetic and the measured data. In general, for fixed parameters \mathbf{M} , the likelihood represents the probability of having a certain data \mathbf{D} given a selected model.
- The evidence $P(\mathbf{D}|I)$ is usually independent of \mathbf{M} and plays only the role of a normalization constant when only a single hypothesis is being considered. When for example, the number of sources is unknown (existence of many hypothesis), the evidence is calculated for each model and the higher the value, the better is the model in predicting the data \mathbf{D} . This parameter is obtained by marginalizing the likelihood over the entire hypothesis space M and is used to ensure proper normalization of the posterior distribution:

$$P(\mathbf{D}|I) = \int_{\text{all } M} P(\mathbf{D}|\mathbf{M}, I)P(\mathbf{M}, I)dM \quad (2.8)$$

- The posterior PDF $P(\mathbf{M}|\mathbf{D}, I)$ represents the full solution to the inverse problem and source reconstruction problem. This PDF encapsulates all the information related to our model parameters and expresses our state of knowledge of \mathbf{M} . Low values of the posterior PDF indicate that the numerical value of \mathbf{M} is improbable while high probability values indicate

higher plausibility. In [16], the prior, likelihood and posterior probabilities are called the three amigos.

To note that the Posterior PDF whose dimension can be very high (very large dimension of \mathbf{M}) is not analytically tractable and should be sampled instead of being marginalized for every source parameter. A variety of sampling methods were developed such as the Markov Chain Monte Carlo (MCMC).

2.2.2.2 Markov Chain Monte Carlo Sampling

For a very low dimensional parameter space, the marginal distributions of all the parameters can be obtained by integrating the posterior distribution.

However, in the case of a very high dimensional parameter space, the analytical solution is not tractable and some stochastic sampling techniques are needed such as the Markov Chain Monte Carlo (MCMC) algorithm.

MCMC sampling technique is used to draw samples iteratively from some distributions until convergence to the posterior distribution of the model parameters which is the target distribution.

MCMC algorithms work in the following manner [2] [9] [14] [17]:

- A. Given some initial value for the parameter $\mathbf{M}^{(0)}$, some random steps are done in order to generate some random sample of the parameter \mathbf{M} . These samples will be either accepted or rejected based on some acceptance criteria.
- B. The series of samples generated by the MCMC algorithm are a Markov chain, and the distribution of these samples will tend asymptotically to the target distribution.

One well-known algorithm is the Metropolis-Hasting (MH) algorithm that accounts for asymmetric proposal distributions that is used in the acceptance criteria.

It is of great importance to note that the choice the proposed samples depend on the selection of a proposal distribution which should be chosen adequately in order to generate representative samples of the distribution. If its width is very large, the chain may remain at the same point for a large number of steps. On the other hand, if the width is small, the convergence may be slow and the exploration of regions of high probability will be very slow. Furthermore, it is very important to consider monitoring the evolution of the Markov chain for a given parameter and for a given proposal distribution. This will automatically give an idea about its convergence.

Now, after obtaining all the samples for the marginal distribution of a given parameter, summary statistics can be obtained [18] [19] [20]:

- The maximum a posteriori estimate of M is the mode of the corresponding marginal posterior distribution:

$$\hat{M}_{MAX} = \arg \max_M P(M|\mathbf{D}, I) \quad (2.9)$$

- Posterior mean of each source parameter:

$$\bar{M} = \int M_i P(M|\mathbf{D}, I) dM \quad (2.10)$$

- Posterior standard deviation that measures the uncertainty in the estimate of M_i :

$$\sigma^2(M_i) = \int (M_i - \bar{M}_i)^2 P(M|\mathbf{D}, I) dM \quad (2.11)$$

- A $p\%$ credible interval or highest posterior density (HPD) that contains the source parameter M_i with $p\%$ probability (the values of the PDF inside the interval are everywhere larger than outside it).

To note that the Bayesian Inference was applied to the determination of the origin and decay rate of non-conservative scalar [12] or to a complex urban environment [21] or a real world application [22], and in the investigation of the number of sources as a parameter estimation problem [23] or as a model selection analysis [24].

2.2.2.3 Summary on Bayesian Probabilistic Approaches

Despite all the advantages provided in terms of probabilistic representation of the solution with a quantification of the uncertainty, the probabilistic Bayesian approach suffers from some inaccuracies when applied to a real world application because of the difficulties in modeling the errors (model and measurement errors) [22]. In addition, a major limitation of the Bayesian Probabilistic approach is the expensive computational cost in the case of a highly dimensional parameter space regardless of the improvements because of the time-consuming sampling process, and the need of the prior information in order to evaluate the posterior distribution. This has led to the inefficiency of the Bayesian Probabilistic approach compared to the optimization methods in cases of emergency.

Finally, figure 2.1 illustrates the steps required in either the optimization or Bayesian inference based approaches [14].

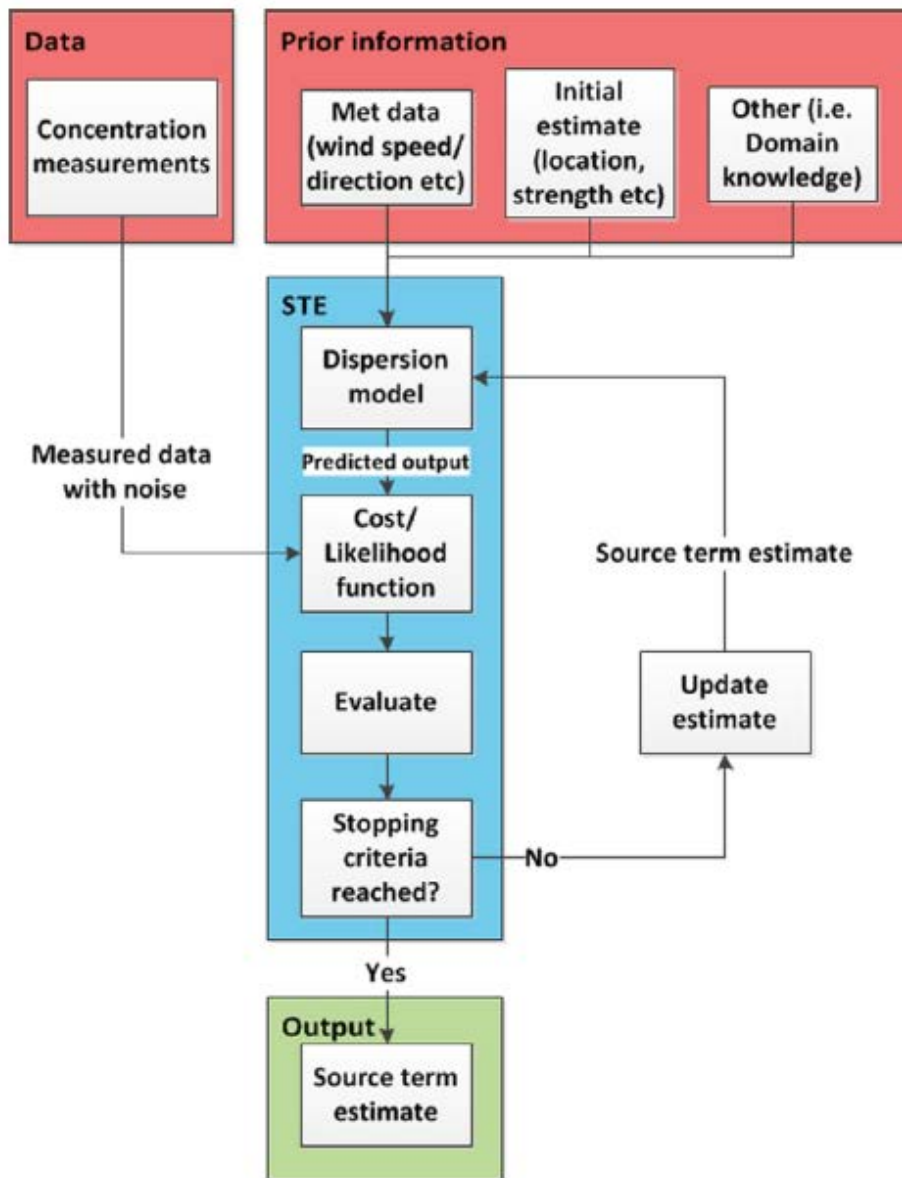


Figure 2.1: Steps required for any inverse problem or source reconstruction approach.

Chapter 3

Methodology

The purpose of this chapter is to layout the framework of the problem statement and the methodology involved in the identification of moving passive tracers in the Mediterranean Sea using a stochastic velocity field.

3.1 Problem Statement

The section consists of introducing the aim of this research work and a brief discussion of the main elements of the inference of sources of pollutants in the Mediterranean Sea.

3.1.1 Stochastic Velocity Field

Given the source location \vec{x}_s and observation time t_0 , the problem is to identify probable locations of a passive particle carried by a stochastic flow field, represented in terms of an ensemble of velocity fields, available at every assimilation step t_a ,

$$\vec{u}_i(\vec{x}_g, t_a), t_a = kT_a, k \in \mathbb{N}, i = 1, \dots, N_e \quad (3.1)$$

where T_a is the assimilation time and is equal to 1 day, \vec{x}_g denotes the grid coordinates, and $N_e = 50$ is the ensemble size, and i is the realization of the velocity.

An ensemble of daily flow fields for the month of January 2006 was generated, resulting in a dataset of 30 sampled time steps that are used in the advection of passive tracers in the Mediterranean Sea. This will allow building the velocity map enabling by that the advection of passive tracers in the Mediterranean Sea.

Basically, an initial particle located at $\vec{x}_i^{t_a}$ at $t = t_a$ will arrive at N_e^2 equally probable destinations, $\vec{x}_{ij}^{t_a + \Delta t}$ at $t_a + \Delta t$. Note that N_e^2 is the product of N_e equally probable velocities at t_a and N_e equally probable velocities at $t_a + T_a$, as illustrated in figure 3.1. Note that the calculation of the velocity at intermediate

time steps is accomplished by using linear interpolation between the velocities at $t = t_a$ and $t = t_a + T_a$.

Furthermore, a binning procedure was implemented in order to control the exponential growth in the number of elements. This procedure enabled the creation of probability maps that provide a quantitative measure of the probability of having a particle at a given location and time. This binning methodology merges all advected particles that fall in the same bin, and are associated with a particular realization, into a single particle that belongs to the same realization while conserving the total probability, the mean position, and the variance.

Figure 3.2 illustrates the steps of Lagrangian model developed in [25].

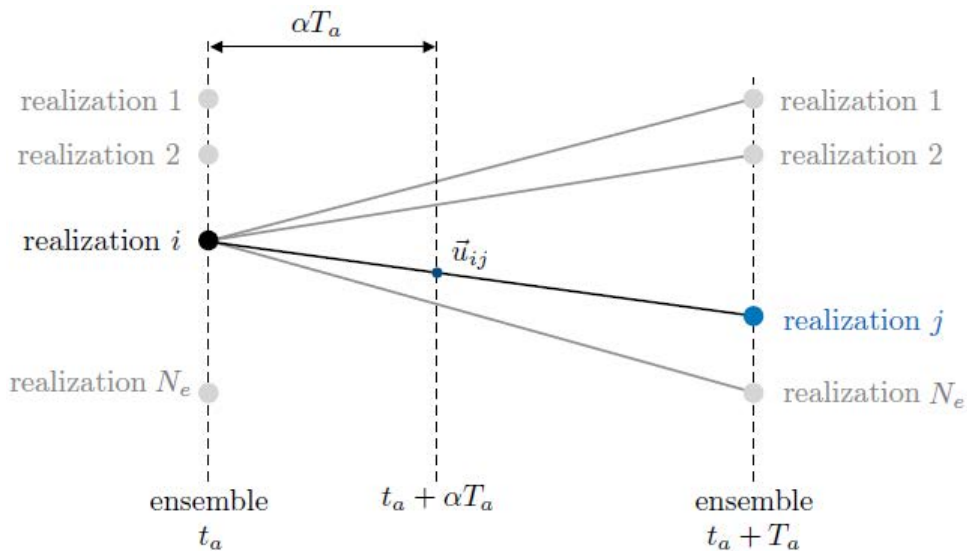


Figure 3.1: Advection of a particle at time t_a .

3.1.2 Observations

The observations are described as $\vec{x}_o^{(m)}(t_o)$, $m = 1, \dots, N_o$, where $\vec{x}_o^{(m)}(t_o)$ is the location of observation m at time t_o , and N_o is the number of observations.

These observation locations would be typical satellite images that will be discussed in details in further sections.

3.1.3 Moving Source

The type of scenario that is addressed in this work is related to the identification of moving passive tracers along the same path with a release of contaminants and sources of pollutants at different times. This investigation is referred to by the *Same path - different time* case.

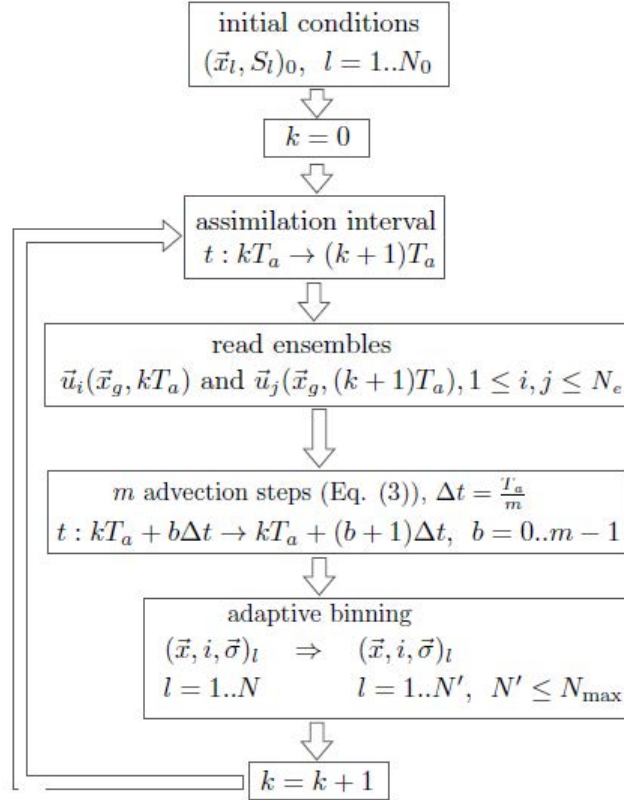


Figure 3.2: Flow chart of the forward advection algorithm.

The path of the ship is selected from the *MarineTraffic: Global Ship Tracking Intelligence* website (<https://www.marinetraffic.com>) where the real-time information about ships trajectories around the world can be accessed. In this work, the zone of interest is the Mediterranean Sea where only oil tankers are considered.

The selection of a given trajectory is based on the density maps of ships provided in the website. By varying the opacity of these maps as shown in figure 3.3, the coordinates of the path (decomposed into 5 stages) are chosen in approximated locations of high densities starting from the Suez Canal through Sardegna (Italy), Spain, and the Strait of Gibraltar. Figure 3.4 illustrates the different stages of the ship path.

Now, Given observations $\vec{x}_o^{(m)}(t_o)$, $m = 1, \dots, N_o$, a stochastic velocity map $\vec{u}_i(\vec{x}_g, t_a)$, $t_a = kT_a$, $k \in \mathbb{N}$, $i = 1, \dots, N_e$, and a set of moving sources $\vec{x}_s^{(j,k)}(t)$, $k = 0, \dots, (N_\tau - 1)$, $j = 0, \dots, (N_s(k) - 1)$, the question is the following: What are the most likely sources and their relative contributions $\hat{q}^{(j,k)}$ that have caused the observed spill in the Mediterranean Sea?

Note that N_τ is the number of trajectories and $N_s(k)$ is the number of sources

along a given trajectory k . Note that all the release events in all the carried experiments will be instantaneous. Now, the pdf of observation $\vec{x}_o^{(m)}(t_o)$ due to a ship $\vec{x}_s^{(j,k)}(t)$ is defined as $f(\vec{x}_o^{(m)}(t_o)|\vec{x}_s^{(j,k)}(t))$. This likelihood function is constructed from the forward simulations following a source oriented approach.

To note that the problem can be also formulated using a one index notation instead of a double index notation where:

$$r = kN_s(k) + j \quad (3.2)$$

Throughout the whole report, the double index notation (j, k) is replaced by the one index notation (r) .

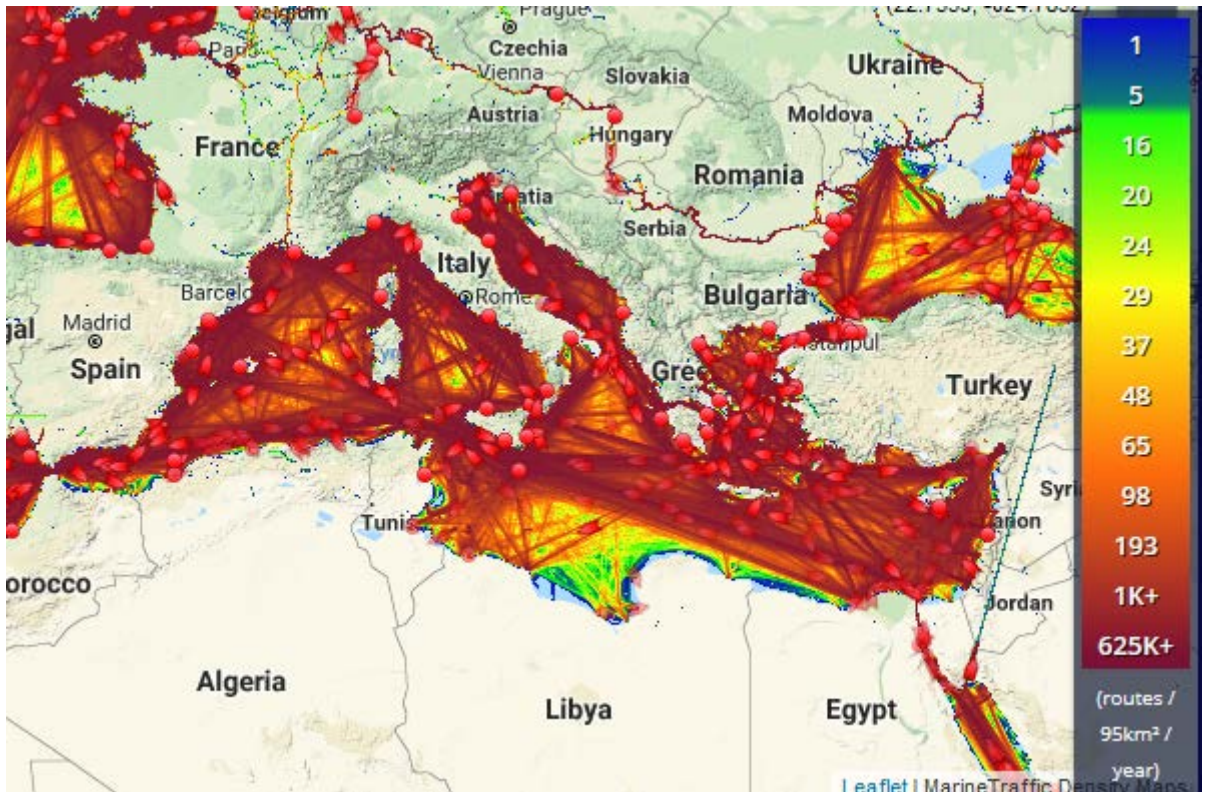


Figure 3.3: Density Maps in the Mediterranean Sea.

3.2 Building the likelihood

The likelihood function is determined from the forward simulations for a single path. Typical speeds of oil tankers are in the range of 12-15 knots, which corresponds to 22.22 - 27.78 km/hr.

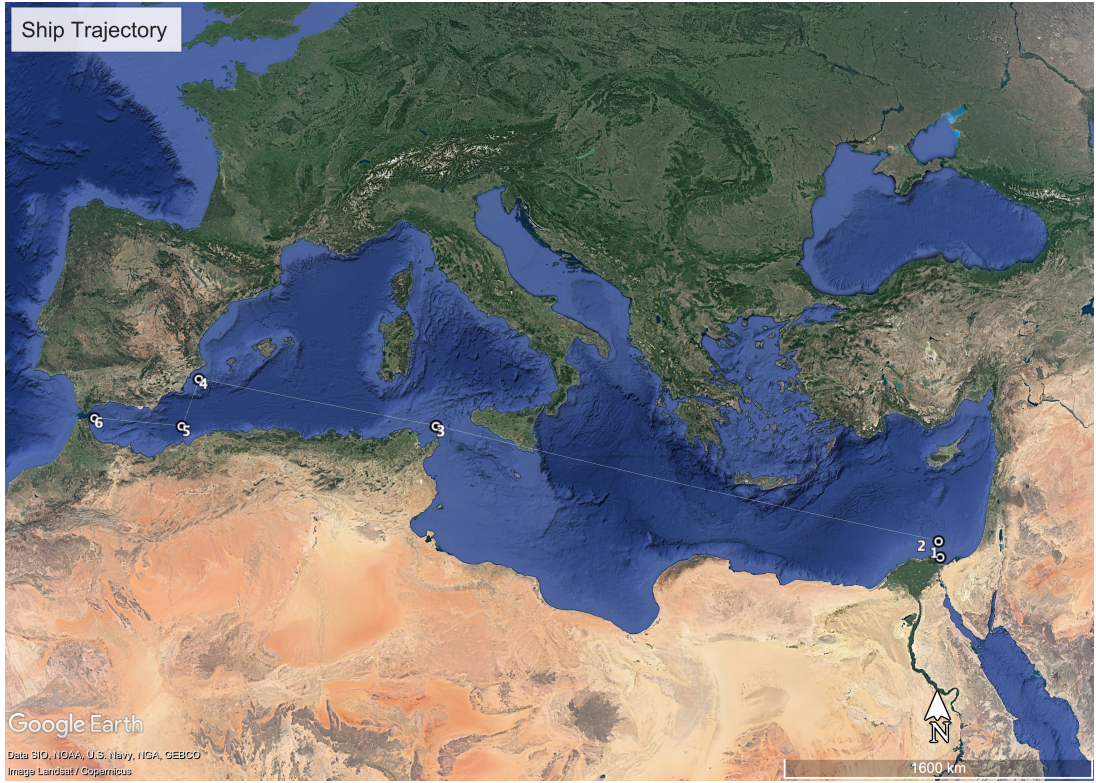
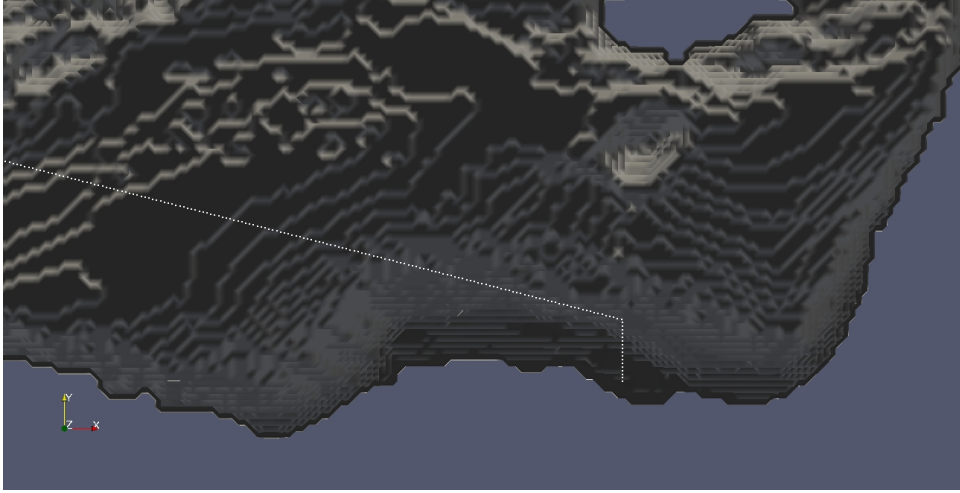
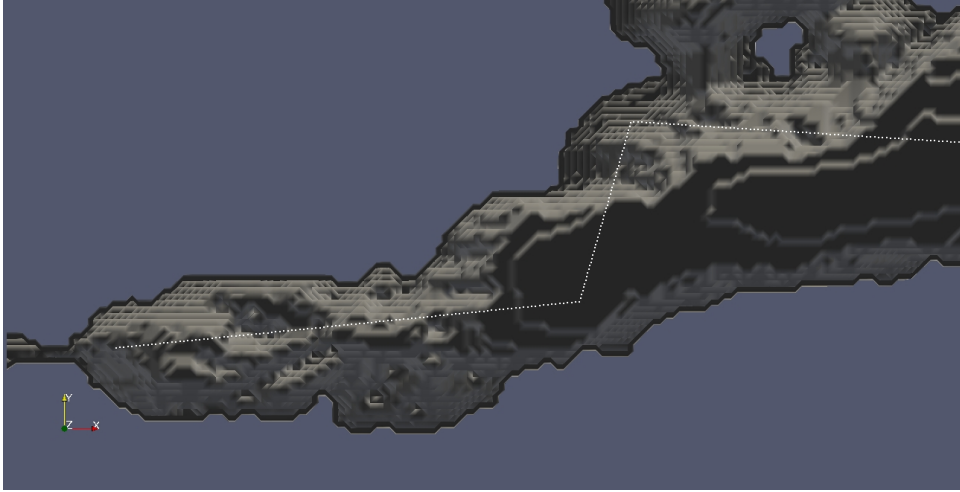


Figure 3.4: Path of the ship from Suez Canal until the Strait of Gibraltar.

- A. Divide the path of the trajectory, \vec{x}_s , into $N_s = 1037$ segments of approximate length $\Delta s \simeq 4\text{km}$.
- B. For the speed range 22.22 - 27.78 km/hr, the tanker traverses the 4 km in 8.639 - 10.81 minutes. The time is adjusted to be 10 minutes, so that it is an integer fraction of the assimilation interval. The tanker speed is set equal to 4 km/10 minutes = 24 km /hr.
- C. Therefore, a tanker is used moving at a speed of $V_s = 24$ km/hr. The tanker traverses the trajectory segment $\Delta s = 4$ km over a time period $\Delta T_R = 10$ minutes.
- D. $N_s N_\tau$ release events are simulated where $N_\tau = 22$, one event at a time, characterized by release locations $\vec{x}_s^{(j,k)}$ and release times $t_R^{(j,k)}$. Where $k = 0, 1, 2, \dots, (N_\tau - 1)$ and $j = 0, 1, 2, \dots, (N_s(k) - 1)$.
- E. The release locations correspond to the centers of the segments Δs comprising the trajectory.
- F. Each trajectory is denoted by DayShift k where $k = 0, 1, 2, \dots, (N_\tau - 1)$.



(a) Discretization of the first stage of the ship path.



(b) Discretization of the last stage of the ship path.

Figure 3.5: Discretized Ship Path.

G. The release times are $t_R^{(j,k)} = t_o - k \times 1day - j\Delta T_R$. Note that t_o in all the carried simulations is set to the 29th day.

Figures 3.5b and 3.5a illustrate the part of the discretized path in the last and first stages of the ship, respectively.

3.2.1 Simulations

In order to reduce the computational cost of the simulations, several strategies were adopted:

- A. Limiting the number of particles using a threshold for the maximum number of particles N_{max} .

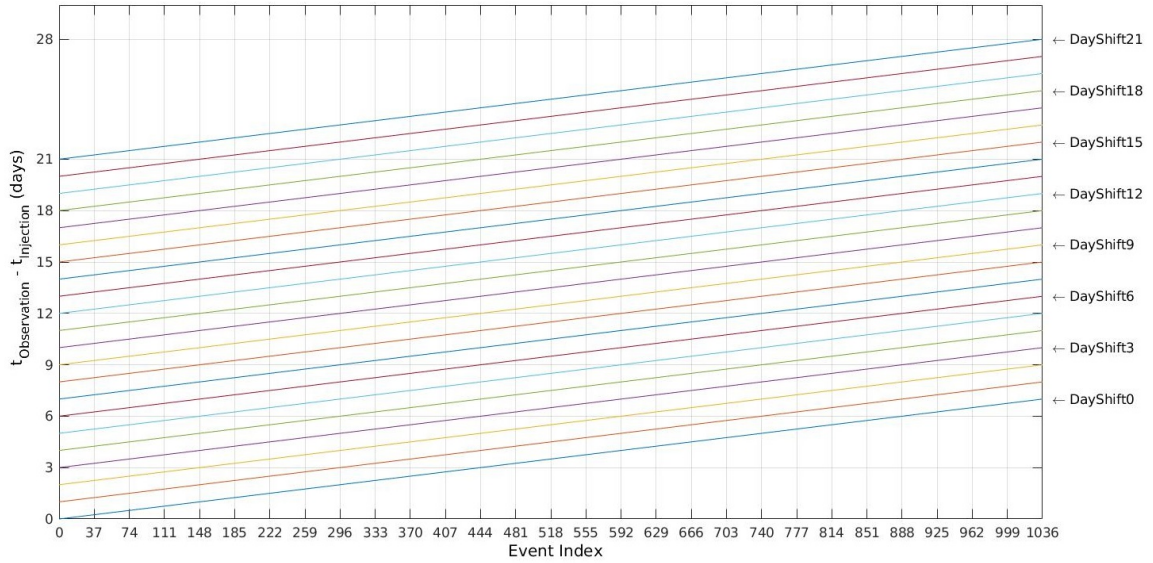


Figure 3.6: Difference between the observation and injection times for the 1037 events and 22 trajectories under study.

- B. Increasing the advection time Δt_{adv} .
- C. Restricting the simulation to the surface flow.

The path, illustrated in figure 3.4, is studied:

- The ship goes from Suez Canal to the Strait of Gibraltar with a total duration of motion around 7 days.
- 1037 simulations are carried out with each simulation corresponding to a release at a given location and time along the path of the ship.
- The releases are 10 minutes apart corresponding to a travelling distance of around 4 km.

Some examples of the trajectories, considered in this study, are listed in table 3.1, and figure 3.6 illustrates the difference between the observation and injection times for the 1037 events and 22 trajectories under study.

- The first mentioned trajectory (Trj. DayShift0) is studied when the ship reaches the Strait of Gibraltar at the same time of the observation time (29 days). This means that the simulation times go from 0 (for a release of the particle occurring at the observation time) to 7 days (for a release of the particle occurring near the Suez Canal). The total real time duration of these simulations was equal to 18.22569 days. Figure 3.7 illustrates the

Table 3.1: Some considered stochastic trajectories

	Type	$t_o - t_R^{(j,0)}$	N_{max}	Δt_{adv} (hour)	Cost (days)
Trj. DayShift0	Stoch.	0	100×10^6	1	18.22569
Trj. DayShift7 - Case 1	Stoch.	7	100×10^6	1	50.80208
Trj. DayShift7 - Case 2	Stoch.	7	10×10^6	1	12.21597
Trj. DayShift7 - Case 3	Stoch.	7	10×10^6	3	9.3368
Trj. DayShift13	Stoch.	13	10×10^6	3	15.02708
Trj. DayShift21	Stoch.	21	10×10^6	3	18.99583

full stochastic probability map obtained from the superposition of the 1037 probability maps.

- The second mentioned trajectory (Trj. DayShift7) (illustrated in figure 3.8) is studied when the ship reaches the Strait of Gibraltar 7 days prior to the observation time. In this case, the simulation times go from 7 days (for an injection near the Strait of Gibraltar) to 14 days (for an injection near the Suez Canal). In this set, three cases were studied:
 - A. The first case (Trj. DayShift7 - Case 1) is studied for a maximum number of particles N_{max} equal to 100×10^6 and an advection time Δt_{adv} equal to 1 hour. These simulations were terminated, and the total real time duration until the termination process was 50.80208 days.
 - B. The second case (Trj. DayShift7 - Case 2) is studied for a maximum number of particles N_{max} equal to 10×10^6 and an advection time Δt_{adv} equal to 1 hour. The simulations were done with a total real time duration equal to 12.21597 days.
 - C. The third case (Trj. DayShift7 - Case 3) is studied for a maximum number of particles N_{max} equal to 10×10^6 and an advection time Δt_{adv} equal to 3 hours. The simulations were done with a total real time duration equal to 9.3368 days.

Figure 3.9 compares some events with different parameters (advection time Δt_{adv} , maximum number of particles N_{max}) along Trj. DayShift7. It can be clearly seen that an advection time $\Delta t_{adv} = 3hrs$ and a maximum number of particles $N_{max} = 10 \times 10^6$ yielded good results similar to the case where $\Delta t_{adv} = 1hr$ and $N_{max} = 100 \times 10^6$. This option of parameters was used throughout the whole simulations of the different trajectories and resulted in a good balance between the accuracy of the probability maps and the computational cost of the simulations.

- The third mentioned trajectory (Trj. DayShift13) is studied when the ship reaches the Strait of Gibraltar 13 days prior to the observation time. In this case, the simulation times go from 13 days (for an injection near the Strait of Gibraltar) to 20 days (for an injection near the Suez Canal). These simulations are done with a total real time duration equal to 15.02708 days.
- The last mentioned trajectory (Trj. DayShift21) is studied when the ship reaches the Strait of Gibraltar 21 days prior to the observation time. In this case, the simulation times go from 21 days (for an injection near the Strait of Gibraltar) to 28 days (for an injection near the Suez Canal). These simulations are done with a total real time duration equal to 18.99583 days.

Note that all probability maps for the different trajectories were obtained using a probability cutoff equal to 10^{-7} .

3.2.1.1 Studied Cases for speed improvement of the simulations

In addition, another set of simulations were carried out in one specific event in (Trj. DayShift21) as listed in table 3.2 in order to investigate further improvements in the speed of the simulations.

Table 3.2: Several Cases for the Probability Map for Event400 of Trj. DayShift21.

	Type	N_{max}	Cost (minutes)
Stoch. Type 1	Stoch.	200000	9
Stoch. Type 2	Stoch.	10^6	10
Stoch. Type 3	Stoch.	10×10^6	24
Determ. Type 1	Determ.	10201	<1
Determ. Type 2	Unopt. Stoch. (x2) + Determ.	125000	1.5
Determ. Type 3	Opt. Stoch. (x2) + Determ.	125000	1.5

The cases for the probability map were studied for Event400 of Trj. DayShift21 in table 3.2:

- The first probability map (Stoch. Type 1) was generated following a **stochastic** simulation using a maximum number of elements N_{max} equal to 200000. The corresponding simulation time was equal to 9 minutes.
- The second probability map (Stoch. Type 2) was generated following a **stochastic** simulation using a maximum number of elements N_{max} equal to 10^6 . The corresponding simulation time was equal to 10 minutes.

To note that there was no decrease in the amount of time between Stoch. Type 2 and Stoch. Type 1 because of reaching a certain overhead time for reading the velocity fields, and any further decrease in the maximum number of particles N_{max} will not affect significantly the simulation time.

- The third probability map (Stoch. Type 3) was generated following a **stochastic** simulation using a maximum number of elements N_{max} equal to 10×10^6 . The corresponding simulation time was equal to 24 minutes.
- The fourth probability map (Determ. Type 1) was generated following a pure **deterministic** simulation (without diffusion) using a total number of elements equal to 10201. The corresponding simulation time was less than 1 minute. This simulation only predicted the high probability region with no further information about regions of low probabilities.
- The fifth probability map (Determ. Type 2) was generated following a **stochastic** simulation over **two assimilation intervals** followed by a **deterministic** simulation with a total number of elements equal to 125000. The corresponding simulation time was equal to 1.5 minutes and the probability map was generated with σ_x , σ_y , and σ_z equal to $\sqrt{t_{pt}}$ where t_{pt} is the particle travel time (artificial diffusion added).
- The sixth probability map (Determ. Type 3) was generated following a **stochastic** simulation over **two assimilation intervals** followed by a **deterministic** simulation with a total number of elements equal to 125000. The corresponding simulation time was equal to 1.5 minutes and the probability map was generated in an optimized way with σ_x , σ_y , and σ_z equal to $\sqrt{D \times t_{pt}}$ where D is the diffusion coefficient, and t_{pt} is the particle travel time. The optimized D coefficient was determined using the Golden Search algorithm by finding the minimum of the l^2 -norm between the optimized deterministic and stochastic probability maps.

For this specific event in Trj. DayShift21, the optimized diffusion coefficient D was found equal to 1416 and the corresponding optimized probability map was generated.

Figures 3.10 and 3.11 illustrate the pure stochastic and deterministic probability maps, respectively. To note that the hybrid case (Determ. Type 3) did not yield desirable results in terms of minimizing the computational cost of the whole simulations because of the need to calculate the artificial diffusion coefficient for each stochastic probability map.

This is why, the only simulations that were carried out are Stoch. Type 3 which provided a good balance and trade-off between accuracy and computational cost.

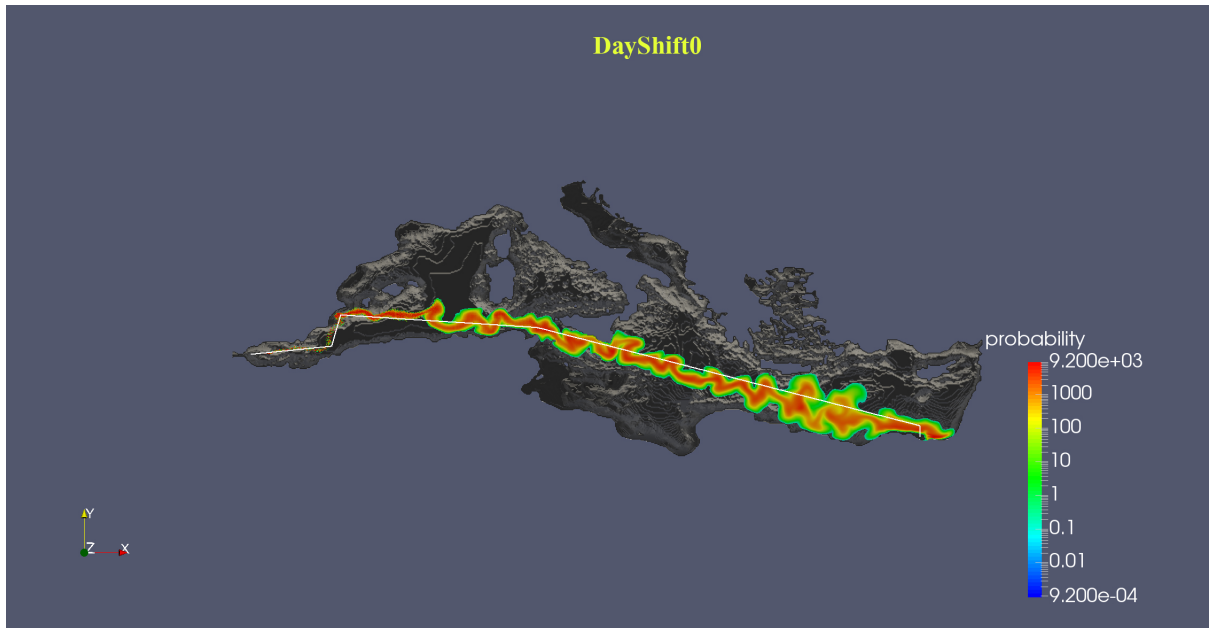


Figure 3.7: Ship path (white line) and the full stochastic probability map for Trj. DayShift0 ($N_{max} = 100 \times 10^6$, $\Delta t_{adv} = 1$ hr).

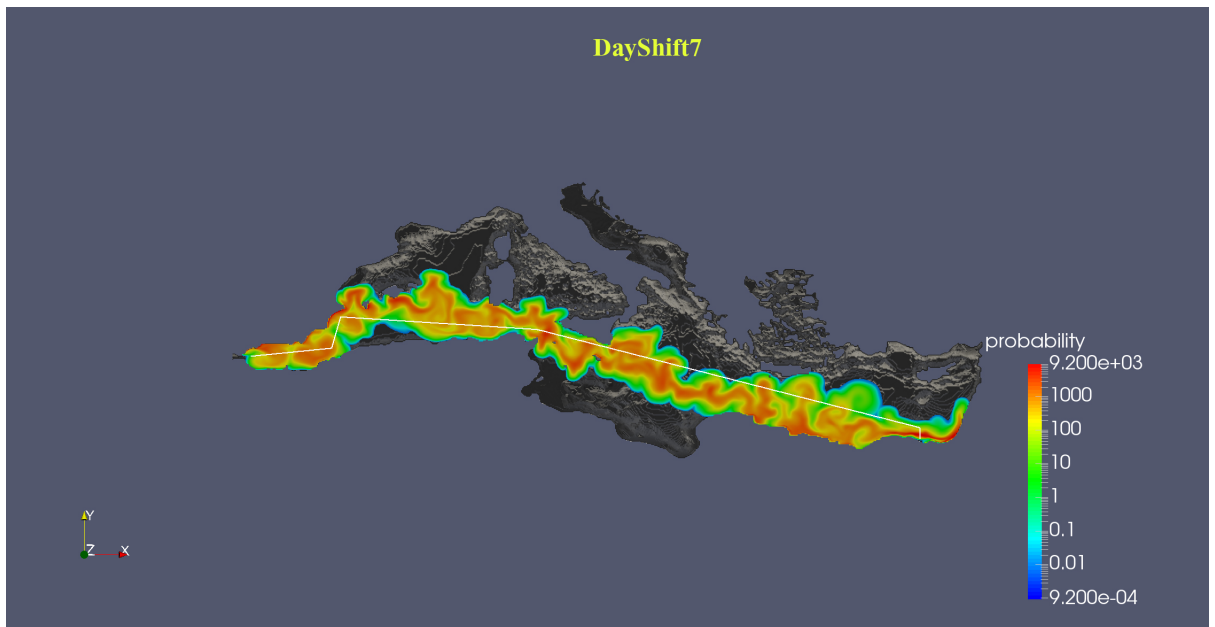


Figure 3.8: Ship path (white line) and the full stochastic probability map for Trj. DayShift7 ($N_{max} = 10^6$, $\Delta t_{adv} = 3$ hr).

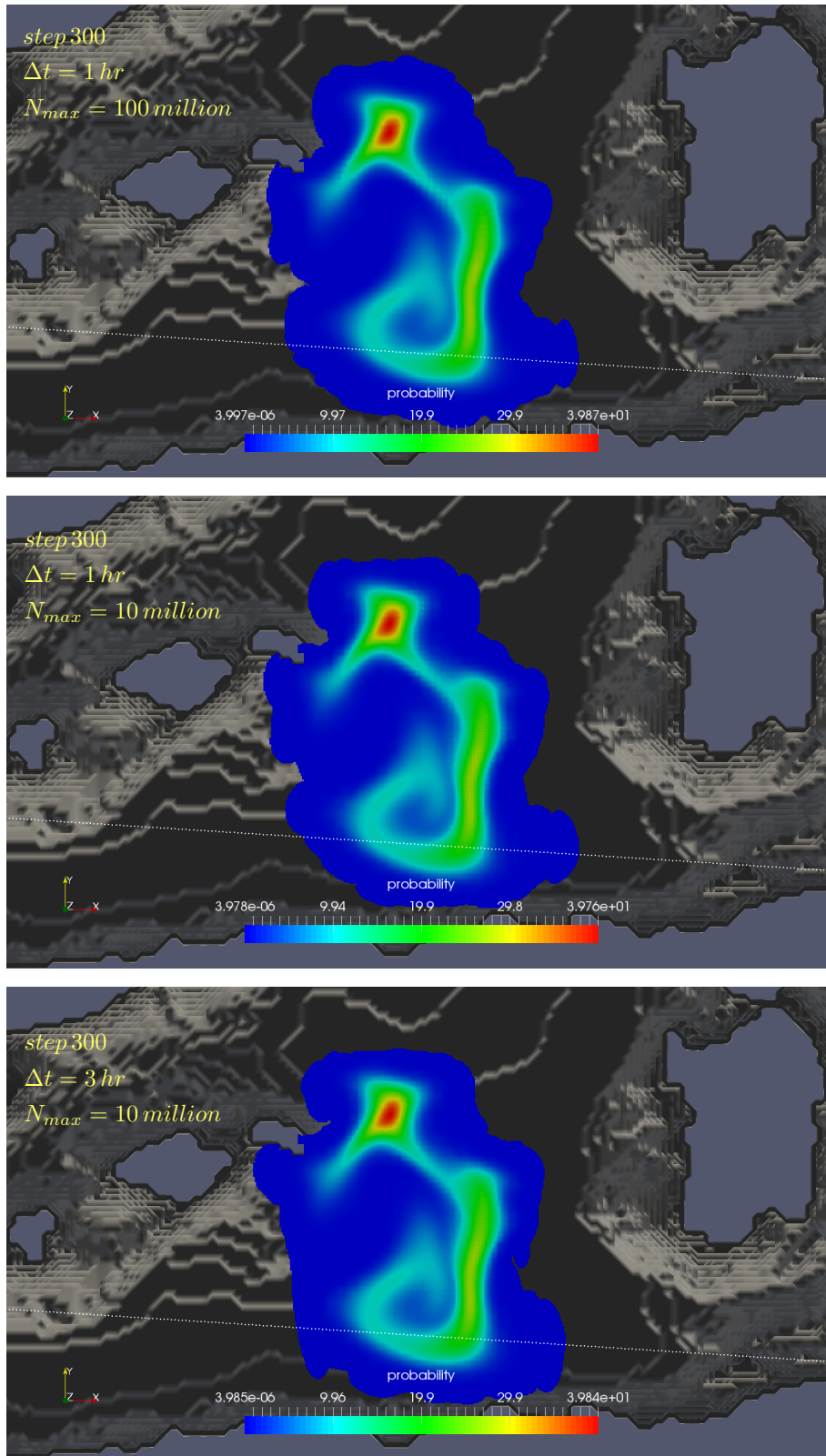
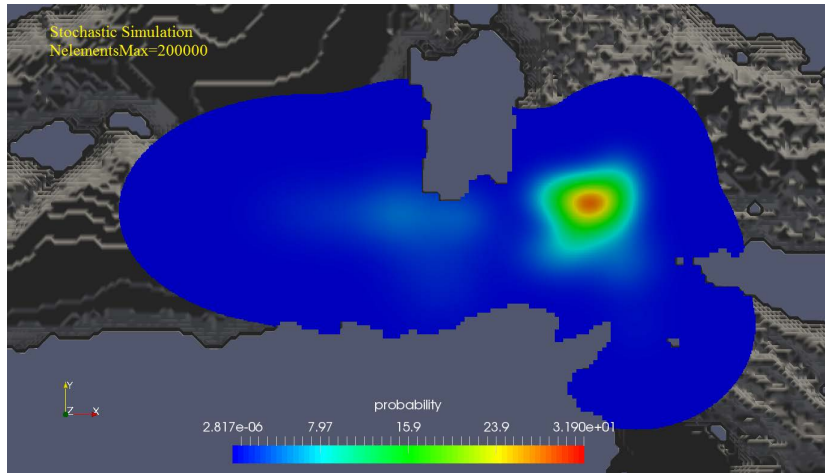
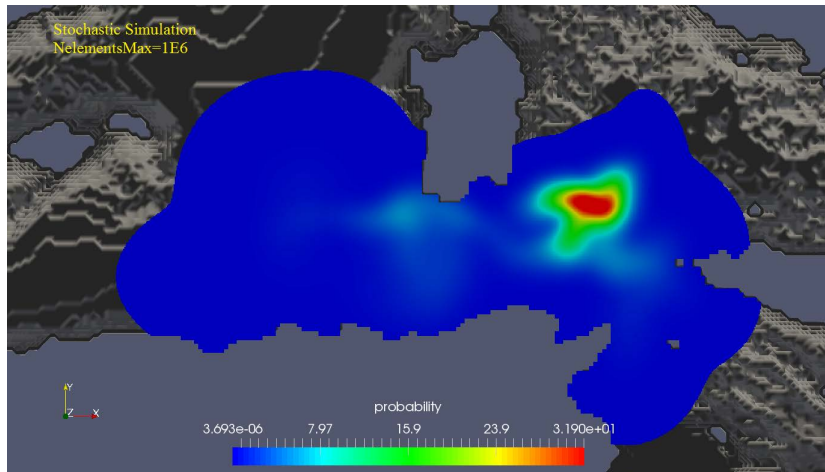


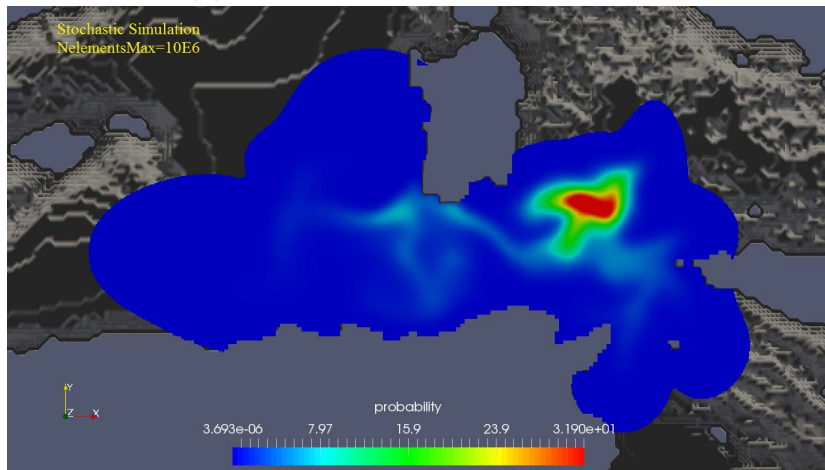
Figure 3.9: Comparison between the three cases (Trj. DayShift7) for Event300.



(a) Stoch. Type 1 with $N_{max} = 200000$.

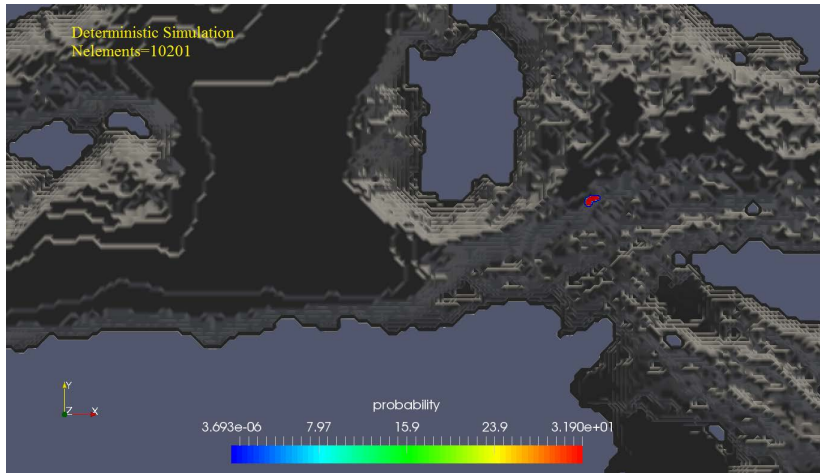


(b) Stoch. Type 2 with $N_{max} = 10^6$.

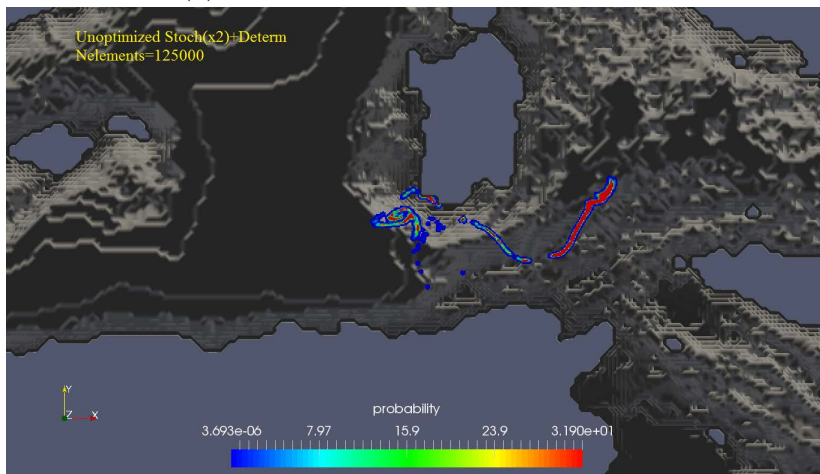


(c) Stoch. Type 3 with $N_{max} = 10 \times 10^6$.

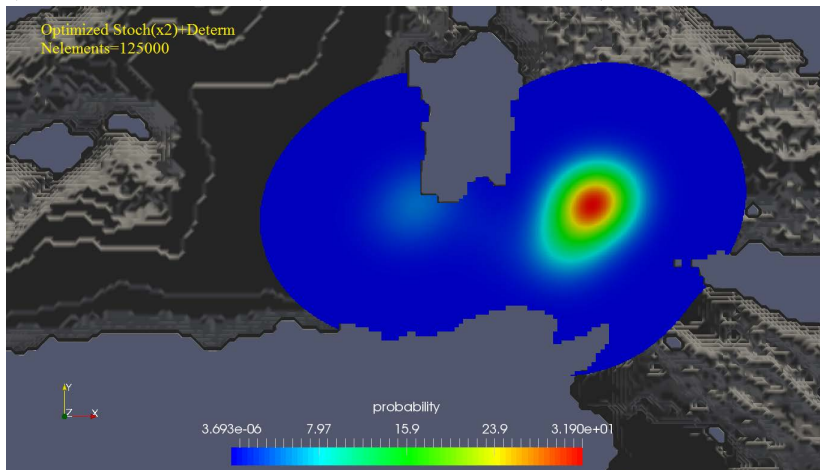
Figure 3.10: Pure Stochastic Probability maps of Trj. DayShift21 for Event400.



(a) Determ. Type 1 with $N = 10201$.



(b) Determ. Type 2 (Unoptimized deterministic) with $N = 125000$.



(c) Determ. Type 3 (Optimized deterministic) with $N = 125000$.

Figure 3.11: Deterministic Probability maps of Trj. DayShift21 for Event400.

3.3 Observation Patches

An observation patch is a typical satellite image which represents the spatial distribution of a physical quantity, and will indicate whether a pollutant is present or not in the Mediterranean Sea. Therefore, there exists a great challenge in the mapping between an observation patch (or a satellite image) and the spacial distribution of the probability obtained using the forward Lagrangian model.

Now, an observation patch is in the form of a binary representation, where $Y = 1$ indicates the presence of a pollutant, whereas $Y = 0$ is related to the absence of a pollutant in a satellite image. This observation patch can be regarded as a probability map with a probability equal to 1 when there exists a pollutant, and a probability equal to 0 when there is no pollutant.

Instead of placing randomly $Y = 1$ observations (presence of pollutants) within our probabilistic model, the observation patch can be synthesized or determined from the probability map generated from the deterministic advection of pollutants using three realizations of the velocity field at every grid point in the Mediterranean Sea. Note that an artificial diffusion is added in the post processing of the probability map. The locations of the presence of pollutants (or $Y = 1$) are selected within the generated observation patch, while for the absence of pollutants (or $Y = 0$), they are within a box of specific size surrounding the generated patch.

The generation of such kind of observation patches is achieved in two different ways that will be discussed sequentially.

3.3.1 Deterministic simulation using the mean (MEAN) of the stochastic velocity field

In order to construct the observation patch, deterministic simulations of a set of events are carried out with an incorporation of an artificial diffusion in order to generate the observation patch. Note that the deterministic simulations are done using the mean of the velocity field at every grid point of the Mediterranean Sea. Mimicing again the presence of a satellite image, the $Y = 1$ and $Y = 0$ observations are selected as discussed previously.

3.3.2 Synthetic Observation patches

In this case, the observation is synthesized using the stochastic probability maps obtained using the forward Lagrangian model.

Therefore, an observation patch can be a synthesized patch or a deterministic probability map which will mimic the availability of a satellite image, where the $Y = 1$ observations are selected over all the grid cells within the generated patch and the $Y = 0$ observations are selected in its outer region and within a pre-defined box. The box and the number of $Y = 0$ observations are defined by the

extension of the box from the synthetic or generated patch, and is denoted by "BoxExtension". Note that BoxExtension0 does not extend from the observation patch, and is associated with the lowest number of $Y = 0$ observations, and used in all the carried inference problems unless otherwise mentioned.

3.3.3 Scaling Coefficient Calculation

A mapping between the satellite image and the stochastic probability map is achieved using a scaling coefficient that will convert the likelihood function into a value between 0 and 1.

In the case of N_s sources contributing with different weights to the observation patch in a given trajectory, the scaling coefficient is calculated as such:

$$\begin{aligned}
& \left[\int_{\mathcal{D}} \hat{q}^{(r_1)} f(\vec{x}_g | \vec{x}_s^{(r_1)}) dA + \int_{\mathcal{D}} \hat{q}^{(r_2)} f(\vec{x}_g | \vec{x}_s^{(r_2)}) dA + \dots + \int_{\mathcal{D}} \hat{q}^{(r_{N_s})} f(\vec{x}_g | \vec{x}_s^{(r_{N_s})}) dA \right] \\
&= \sum_{i=1}^{N_g} [f(\vec{x}_g | \vec{x}_s^{(r_1)}) \Delta x_i \Delta y_i + f(\vec{x}_g | \vec{x}_s^{(r_2)}) \Delta x_i \Delta y_i + \dots + f(\vec{x}_g | \vec{x}_s^{(r_{N_s})}) \Delta x_i \Delta y_i] \\
&= \sum_{p=1}^{N_p} C \Delta x_p \Delta y_p
\end{aligned} \tag{3.3}$$

Where $f(\vec{x}_g | \vec{x}_s^{(r)})$ is the likelihood probability (not normalized) at grid points \vec{x}_g due to ship of index r at location $\vec{x}_s^{(r)}$. N_g is the number of grid points, Δx and Δy are the size of a grid cell in the x and y direction, respectively.

In the case of one single event, the scaling coefficient calculation reduces to:

$$\begin{aligned}
\int_{\mathcal{D}} f(\vec{x}_g | \vec{x}_s^{(r)}) dA &= \sum_{i=1}^{N_g} f(\vec{x}_g | \vec{x}_s^{(r)}) \Delta x_i \Delta y_i \\
&= \sum_{p=1}^{N_p} C \Delta x_p \Delta y_p
\end{aligned} \tag{3.4}$$

3.4 Cost Function

When multiple events with different weights are contributing to an observation patch, the cost function is formulated using the Logistic Regression approach.

Logistic Regression is a classification algorithm that is widely used in machine learning applications. Unlike the linear regression cost function that aims at minimizing in general the square difference between modelled and measured data, the logistic regression one avoids ending up with non-convex function where

the detection of the global minimum, and subsequently the true identity of the inferred sources, becomes quite impossible when dealing with a binary classification problem.

The scaling coefficient discussed previously serves as a tool to map the predictions (likelihood function f which is not normalized) into probabilities in a similar fashion to the sigmoid function that is used in machine learning problems.

Since the problem of interest involves classified observations ($Y = 1$ and $Y = 0$ observations), the logistic regression cost function is used in this case, and is found to yield better results, when dealing with multiple sources, compared to the linear regression one. This cost function is given by:

$$\begin{aligned} \mathbb{J}^{(r)} = & \sum_{p=1}^{N_p} \left[-Y(\vec{x}_o^{(p)}(t_o)) \log \left(\sum_{r=r_1}^{r_{N_s}} \hat{q}^{(r)} \frac{f(\vec{x}_o^{(p)}(t_o) | \vec{x}_s^{(r)}(t))}{C} \right) \right] \\ & + \sum_{n=1}^{N_n} \left[-[1 - Y(\vec{x}_o^{(n)}(t_o))] \log \left(1 - \sum_{r=r_1}^{r_{N_s}} \hat{q}^{(r)} \frac{f(\vec{x}_o^{(n)}(t_o) | \vec{x}_s^{(r)}(t))}{C} \right) \right] \end{aligned} \quad (3.5)$$

Where N_o is the number of observations, Y is the observation (0 or 1), N_p is the number of $Y = 1$ observations, N_n is the number of $Y = 0$ observations, $\hat{q}^{(r)}$ is the relative strength of the contributing source j along trajectory k and $f(\vec{x}_o^{(m)}(t_o) | \vec{x}_s^{(r)}(t))$ is the likelihood function.

Figure 3.12 illustrates the process of the calculation of the distance between the observation and stochastic maps in order to yield the minimum cost function associated with the probable sources contributing to an observation patch. A satellite image (a) will be converted into a binary representation ($Y = 0$ and $Y = 1$) (b) and will be compared to a likelihood map (c) obtained using the forward Lagrangian model. Note that this likelihood map is scaled with a scaling factor and converted into probability map in the same space as the observation patch. The distance between these two images is calculated (d) through the cost function in equation (3.5) that operates in the same space of the observation patch, and the aim is to find the minimum value that will be associated with the most probable sources contributing to an observed patch.

3.5 Sampling Algorithm

In order to infer for the probable sources contributing to the observation patches, as well as their relative contributions, a sampling algorithm is adopted.

This algorithm consists of sampling the model parameters in a similar way to the Metropolis Hastings algorithm, with the use of the logistic cost function in the acceptance criterion instead of the posterior probability. The general idea of this

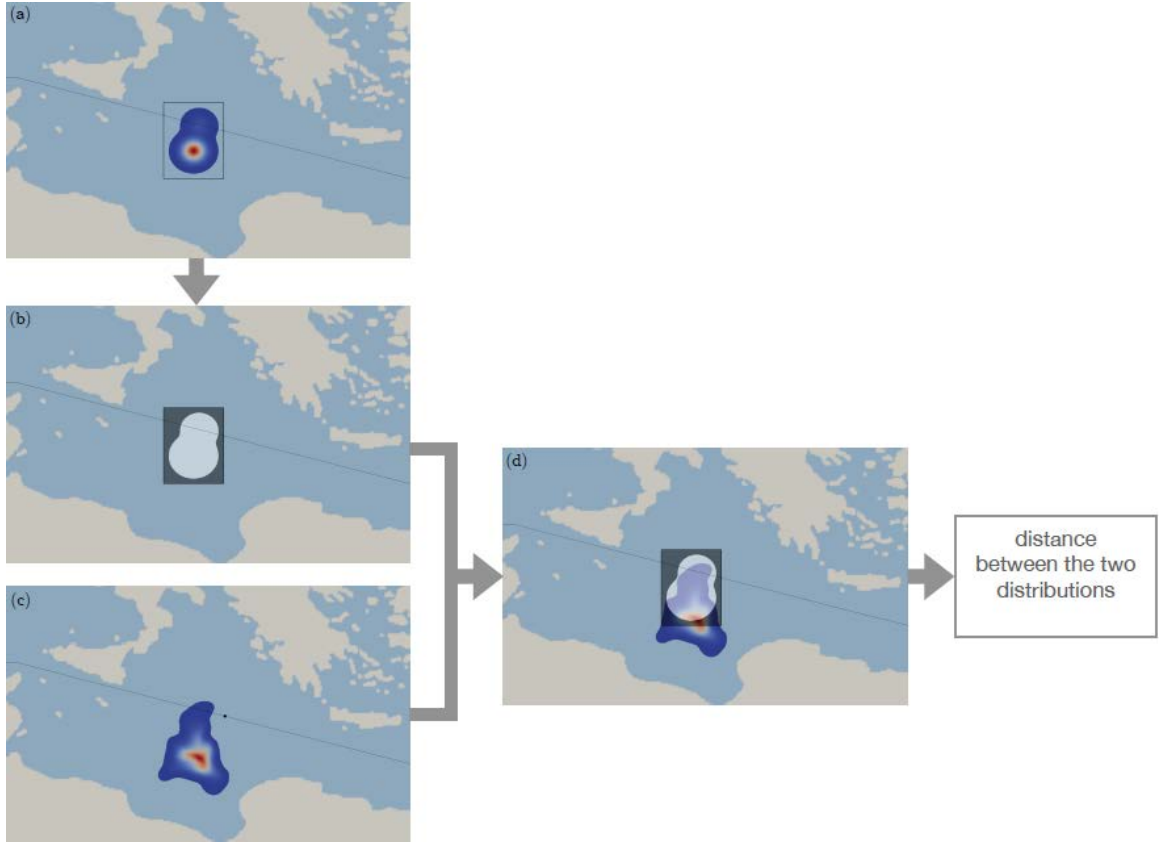


Figure 3.12: Flow Chart of the minimization of the calculation of the distance between a satellite image (a) and a probability map (c).

sampling approach is to adapt the value of the hyperparameter λ and the value of the standard deviation of the sources σ_r in order to yield converged chains that are well-mixed. Note that the standard deviation of the relative weights $\sigma_{\hat{q}}$ is not adapted during sampling since the corresponding chains are already well-mixed for any values of λ and σ_r .

The algorithm consists of the following essential steps that can be carried out over multiple chains N_{chains} :

A. Initialization:

- The initial vector of model parameters $\mathbf{M}^{(0)}$ is randomly initialized. The indices of the sources $r_1^{(0)}$ through $r_{N_s}^{(0)}$ are randomly sampled from a uniform distribution between 0 and 1036 (since 1037 release scenarios occur along the path of the ship). For the relative weights of the sources $\hat{q}_1^{(0)}$ through $\hat{q}_{N_s}^{(0)}$, they are sampled from a uniform distribution between 0 and 1. Note that the sum of the relative weights should be equal to 1. In addition, the user specifies the number of samples M_s required to have a good representation of the model parameters.

- The initial guesses of the standard deviations ($\sigma_r^{(0)}$ and $\sigma_{\hat{q}}^{(0)}$) of the proposal density functions are also set by the user.
- The initial guess of the hyperparameter $\lambda^{(0)}$, as well as its adaptation factors (u_λ and l_λ) are also specified by the user.

B. Adaptation of the hyperparameter λ :

- A sampling subroutine (*sample*) is called. This function takes as an input the initial guess of the model parameters $\mathbf{M}^{(0)}$, as well as the adaptation parameters $\sigma_r^{(0)}$, $\sigma_{\hat{q}}^{(0)}$, and $\lambda^{(0)}$. This function also takes as an input the number of samples M_s required to obtain a good representation of the probability density function of every source parameter. The output of the sampling function is the acceptance rate AR_0 and the chain of the accepted samples $\mathbf{M}^{(M_s)}$.
- The algorithm will converge when the acceptance rate is between 30% and 50%. If the AR is less than 30%, the hyperparameter $\lambda^{(0)}$ is increased by a factor l_λ to enhance the acceptance of samples. On the other hand, if the AR is greater than 50%, the hyperparameter is decreased by another factor u_λ . This way of updating λ will ensure that the AR is in the optimum range.

C. Sampling subroutine:

- The chain of samples $\mathbf{M}^{(M_s)}$ is obtained in this case by using a component-wise sampling algorithm instead of sampling the whole vector of parameters. This will enhance the mixing and convergence of the chains especially for a high dimensional parameter space.
- For every source parameter, the proposed sample is an increment of the current value. If the index of the source r is sampled, the proposed index is sampled from a normal distribution having as a mean the current value of the index and as a standard deviation σ_r . If the relative weight of the source \hat{q} is sampled, the proposed index is sampled from a normal distribution having as a mean the current value of the index and as a standard deviation $\sigma_{\hat{q}}$. Note that vector of relative weights should be always normalized. Note that in the case the proposed model parameters are outside their ranges, we can resample again until the model parameters are in their ranges or we can adapt a reflection strategy that reflects the model parameter to its mirrored value within the range of the model parameters.
- When a model parameter is proposed, the cost function of the new vector of model parameters is compared to the one associated with the current model parameters. The aim is to find the minimum cost

function that will be automatically related to the optimal source parameters. If the acceptance criteria is satisfied, the vector of model parameters in the chain is updated. Otherwise, the model parameters are not modified. This process continues until we sample M_s samples.

- The adaptation of the standard deviation of the indices of the sources σ_r is adapted every M_u samples are drawn. This update of σ_r will enhance the mixing of the chains of the samples. This adaptation is achieved by calculating the autocovariance at lag 0 ($s_0^{(d)}$) for every source parameter over these M_u samples which represents the autocorrelation between the values themselves. If the chain of the samples is locked over the M_u samples (or the algorithm is not accepting new model parameters), $s_0^{(d)}$ is equal to zero. In this case, σ_r is decreased by a factor a_σ specified by the user. If this is not the case, σ_r is not updated. The adaptation step is achieved every M_u steps until M_s samples are obtained. Note that this modified σ_r is used in the normal distribution for proposing the indices of the sources.
- The acceptance rate AR for this sampling process is calculated, and it is required to be between 30% and 50% to reach convergence.

D. Summary Statistics:

- The distribution of the samples obtained from multiple chains should tend asymptotically to the true distribution of the model parameters and the chain convergence and mixing will be visualized.
- Correlation maps are also generated when the number of sources is greater than 1. These maps will illustrate the correlation between the different model parameters, and the regions of high occurrences and probability.
- Posterior probability density functions (or marginal) are also generated representing the uncertainty in the inferred model parameters and how close are the true source parameters (obtained using the global optimization algorithm) to the maximum occurred ones obtained using the sampling algorithm.

The previously discussed steps for the inference of the source indices \mathbf{r} and their relative contributions to the observation patches $\hat{\mathbf{q}}^{(r)}$ are illustrated in the following algorithm:

Run N_{chains} ($c = 1, \dots, N_{chains}$)
 Initialize $\mathbf{M}^{(0)}(c) = [r_1^{(0)}, r_2^{(0)}, \dots, r_{N_s}^{(0)}, \hat{q}_1^{(0)}, \hat{q}_2^{(0)}, \dots, \hat{q}_{N_s}^{(0)}](c)$ and M_s
 Set $(\sigma_r^{(0)}(c), \sigma_{\hat{q}}^{(0)}(c))$
 Set $\lambda^{(0)}(c)$, $u_\lambda (< 1)$, and $l_\lambda (> 1)$

repeat until stop
 if $AR_0 < 0.3$ or $AR_0 > 0.5$ then
 call $\text{sample}(\mathbf{M}^{(0)}(c), M_s, \sigma_r^{(0)}(c), \sigma_{\hat{q}}^{(0)}(c), \lambda^{(0)}(c), AR_0, \mathbf{M}^{(M_s)}(c))$
 if ($AR_0 < 0.3$): $\lambda^{(0)}(c) = l_\lambda \lambda^{(0)}(c)$
 if ($AR_0 > 0.5$): $\lambda^{(0)}(c) = u_\lambda \lambda^{(0)}(c)$
 end if
 end repeat until stop

This is the component by component *sample* subroutine:

```

Subroutine sample( $\mathbf{M}^{(in)}$ ,  $M_s$ ,  $\sigma_r$ ,  $\sigma_{\hat{q}}$ ,  $\lambda$ ,  $AR$ ,  $\mathbf{M}^{(out)}$ )
 $\mathbf{M}^{(out)} = \mathbf{M}^{(in)}$ 
 $k = 0$ 
Set  $a_\sigma$  ( $a_\sigma < 1$ ) and  $M_u$ 
do  $i=1$  to  $M_s$ 
  do  $d = 1$  to  $2N_s$ 
     $\mathbf{M}^* = \mathbf{M}^{(out)}$ 
    If  $1 \leq d \leq N_s$ : Sample  $\mathbf{M}^*(d) \sim \mathcal{N}(\mathbf{M}^{(out)}(d), \sigma_r)$ 
    If  $N_s < d \leq 2N_s$ : Sample  $\mathbf{M}^*(d) \sim \mathcal{N}(\mathbf{M}^{(out)}(d), \sigma_{\hat{q}})$ 
    Sample again or Reflect when out of range
    Normalize  $\mathbf{M}^*(\hat{\mathbf{q}})$ 
    Calculate  $\mathbb{J}(\mathbf{M}^*)$  and  $\mathbb{J}(\mathbf{M}^{(out)})$ 
    Sample  $\alpha \sim \mathcal{U}(0, 1)$ 
    If  $\frac{e^{-\frac{\mathbb{J}(\mathbf{M}^*)}{\lambda}}}{e^{-\frac{\mathbb{J}(\mathbf{M}^{(out)})}{\lambda}}} > \alpha$ 
       $\mathbf{M}^{(out)} = \mathbf{M}^*$ .
       $k = k + 1$ 
    end if
  end do  $d$ 
  If  $\sim \text{mod}(i, M_u)$ :
    Calculate  $s_0^{(d)}$  (autocovariance at lag 0) for  $d = 1, \dots, N_s$ 
    If any  $s_0^{(d)}$  is zero:  $\sigma_r = a_\sigma \sigma_r$ 
  end if
end do  $i$ 
 $AR = \frac{k}{2N_s M_s}$ 

```

3.6 Optimization Algorithm

Formulating this inference exercise using an optimization approach reduces our problem of interest into minimizing equation (3.5) to determine the optimum model parameters with their associated weights. Note that this deterministic algorithm is used for the validation of the results obtained with the sampling algorithm.

This optimization problem is non-convex, and a global optimization algorithm is crucial in this case. This is why, the “Global Search” algorithm implemented in MATLAB is used. It is a scatter-search based global optimization solver.

The aim of this algorithm is to locate the solution with the lowest cost function value. This algorithm starts by generating a set of trial points using a Scatter Search Method, and these points are then filtered based on the values of the cost functions and the constraint filters. It starts by finding solutions from each of the filtered points in order to obtain a global solution vector of the several variables under study.

In order to decrease the computational cost of the inference problem, another similar global optimization algorithm called “Multi-Start” is used. This algorithm enables parallel processing of the algorithm and allows reaching faster the solution of the inference problem. Another optional step that will reduce the computational cost of this algorithm relies in the initial determination of the range of probable sources of contaminants through the identification of the events contributing positively to the observation locations, and the optimization algorithm will try to find the most probable combination of pollution sources out of this reduced range.

It is of great importance to note that the deterministic optimization approach and the sampling approach, inspired from Baye’s Theorem, are directly linked. Assuming that our prior likelihood on the model parameters is uniform, the Maximum A Posteriori (MAP) and the Maximum Likelihood Estimation are the same. This means that the Bayesian Probabilistic approaches are directly linked to an optimization procedure. Therefore, the results obtained using the global optimization algorithm and the sampling algorithm inspired from the Bayesian approach will be compared in the results section.

Chapter 4

Results and Discussion

The purpose of this chapter is to illustrate the results obtained using the sampling algorithm in the source reconstruction problem in the Mediterranean Sea in the presence of a stochastic velocity field.

4.1 Forward Problem

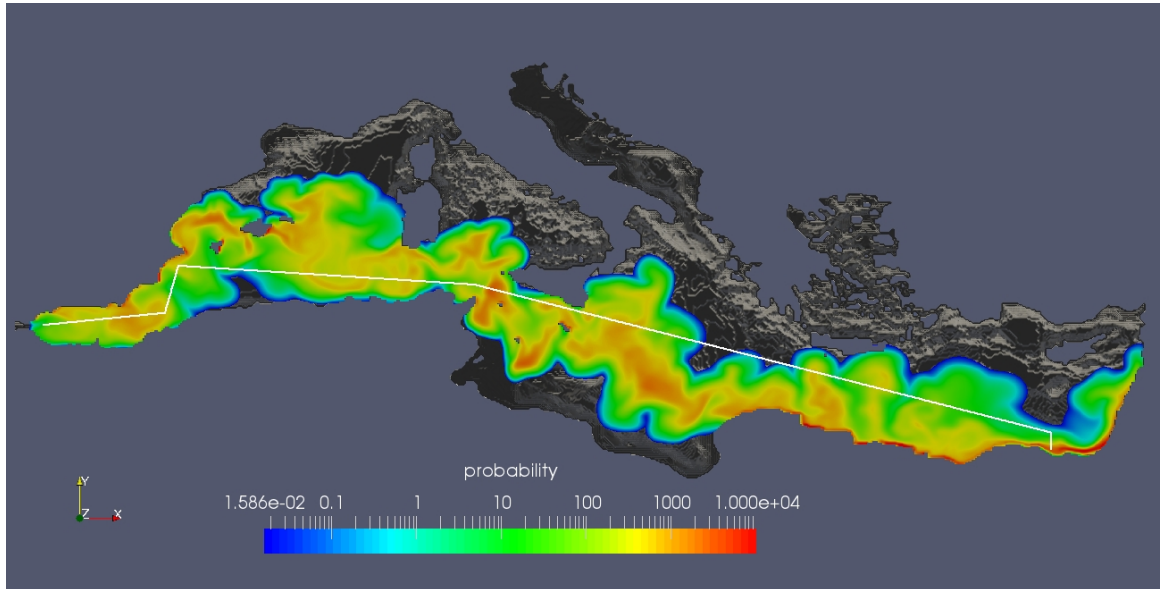
Trajectory DayShift17 is used in the inference problem. This trajectory corresponds to a ship moving for 7 days in the Mediterranean sea and reaches the Strait of Gibraltar 17 days prior to the observation time (Day 29).

The 1037 release events in this trajectory were simulated using a maximum number of particles equal to 10×10^6 and an advection time equal to 3 hours. Note that this combination of parameters yielded a good balance and tradeoff between the computational cost of the simulations and the accuracy of the representation of the probability maps.

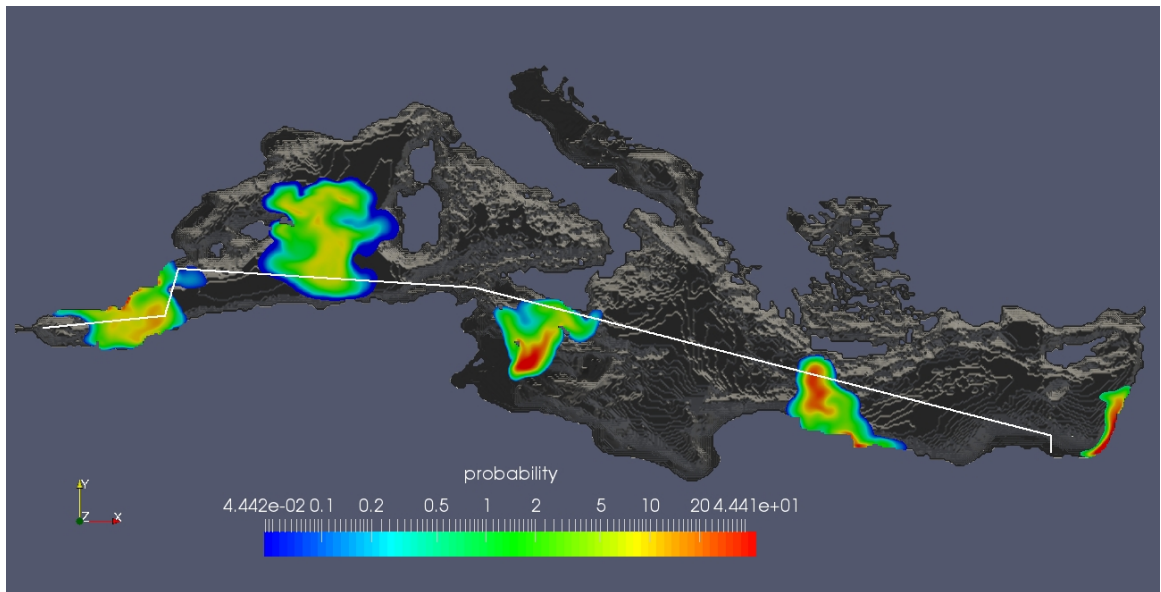
Figure 4.1 illustrates the generated probability map obtained from the superposition of the individual probability maps of the 1037 events, as well as some individual probability maps in trajectory DayShift17 with discretized ship path in white.

4.1.1 Impact of Adaption of λ and σ_r on the inference problem

The implementation of an adaptation of λ and σ_r in the inference algorithm is to ensure that the chains converge to the true sources and relative weights while investigating the regions of high probability and being well-mixed. Figure 4.2 illustrates the behavior of the cost function for multiple input parameters for the algorithm, and these parameters involve the values of λ and whether the adaptation of σ_r is set in the sampling process. Note that these investigations in the inference algorithm were carried out for a double source inference problem.

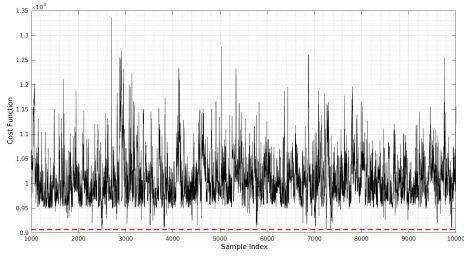


(a) Composite Map of Trj. DayShift17 obtained from the superposition of the 1037 events.

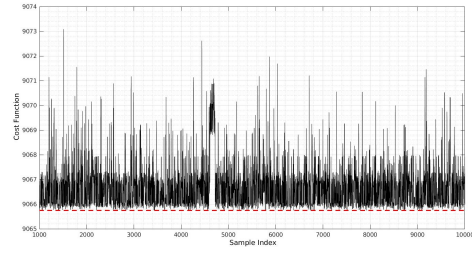


(b) Some selected individual probability maps in Trj. DayShift17 along the ship path.

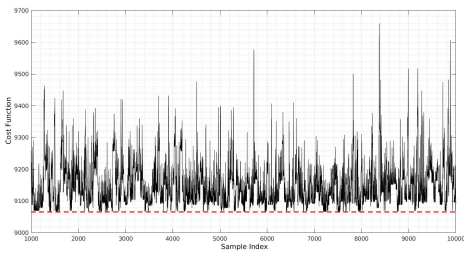
Figure 4.1: Probability Maps in Trj. DayShift17 along the ship path (white line).



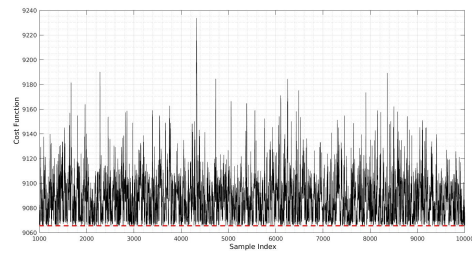
(a) Cost function variation with $\lambda = 500$ with no adaptation of σ_r .



(b) Cost function variation with $\lambda = 1$ with no adaptation of σ_r .



(c) Cost function variation with converged optimum λ with adaptation of σ_r .



(d) Cost function variation with converged optimum λ with adaptation of σ_r .

Figure 4.2: Effect of the adaptation parameters on the cost function variation (minimum cost function shown in red dashed line).

It can be seen from figure 4.2a that the cost function of the accepted samples in the chain of the algorithm would undergo abrupt changes and deviate towards high values when the hyperparameter λ is large. With the absence of effect of λ (equal to unity) (figure 4.2b), the cost function of the samples fluctuates near the minimum value with no significant changes, and the highest encountered difference in cost functions in the samples is around 8. This implies the need for an optimum value of the hyperparameter λ that will allow relatively significant changes in the cost function of the accepted samples, as illustrated in figure 4.2c, without causing the algorithm to deviate towards regions of low probability (when λ and the cost function are high).

Figure 4.2d illustrates that the implementation of the adaptation of σ_r decreases the jumps in the cost function compared to the case where the adaptation of the standard deviation of the sources is not set. However, the fluctuations near the minimum cost function are higher than the case when $\lambda = 1$, and this will enhance the chain mixing as will be illustrated in the following sections.

4.2 Inference Problem

In this section, the inference of source parameters, given a set of observations in the Mediterranean Sea, is applied for single and multiple sources. The identity of the sources and their corresponding weights are determined, and compared to the true values obtained using the global optimization algorithm.

Note that all observation patches consist of $Y = 1$ and $Y = 0$ observations (presence or absence of pollutants, respectively). The $Y = 1$ and $Y = 0$ are represented by white and magenta points, respectively.

4.2.1 Inference of a Single Source

The sampling algorithm was applied to the single source reconstruction problem, where the observation patch was generated by solving an advection-diffusion problem using the MEAN realization of the velocity field at every grid point in the Mediterranean Sea. The events, used to generate this typical satellite image, are 349 to 353, and the global optimization algorithm was used to predict the true identity of the index of the source.

Table 4.1 illustrates the studied case, and the associated figure for the observation patch and the inferred results, and table 4.2 illustrates the parameters of the sampling algorithm that are specified by the user.

Table 4.1: Single Source Inference problem in Trj. DayShift17.

	Events	Observation Patch	Figures
Case 1	349 to 353	Determ.+ Diff. + MEAN	4.3 and 4.4

Table 4.2: Sampling Algorithm parameters of the Single Source Inference problem in Trj. DayShift17.

	n_{Chains}	u_λ	l_λ	M_s	M_u	a_σ	burn-In
Case 1	5	0.1	2	10000	10	0.7	1000

Figures 4.3a illustrates the generated observation patch of the MEAN realization.

It can be seen from the brute force calculations of the cost function for the different indices of the sources while varying the box Extension (or the number of $Y = 0$ observations), that the increase in the $Y = 0$ observations does not affect

significantly the index of the source associated with the lowest value of the cost function. This is illustrated in figure 4.3b.

In addition, it can be seen from figure 4.4a that the chain obtained from the sampling algorithm converged to the true identity of the event obtained using the global optimization algorithm (represented in a red dashed line). With the implementation of the adaptation of the hyperparameter λ and the standard deviation of the index of the contributing source σ_r , the chain is also well-mixed. Note that a MEAN realization, which is a good representation of the stochastic velocity field, yielded event 349 that is associated with the lowest cost function. Note that this event is also a member of the subset of events used to generate the observation patch.

Finally, figure 4.4b illustrates the posterior probability function and the uncertainty in the solution when using the MEAN realization. It is clear that the maximum occurred event in the chain is the true identity of the source obtained using the global optimization algorithm.

4.2.2 Inference of Multiple Sources

In this case, the sampling algorithm was applied to the inference of multiple sources contributing to a given observation patch in the Mediterranean Sea. The indices of the sources as well as their relative weights are inferred. Note that the relative weights represent the relative contributions of the sources to the observation patch under study. Note that in all cases, the algorithm was able to predict the sources and their relative contributions to the observation patches.

Table 4.3 illustrates the studied cases with the identity of the events used to generate the observation patches, as well as the numbers of the figures of the satellite images and the inferred sources and their weights. In addition, table 4.4 illustrates the parameters of the sampling algorithm of the studied cases in the multiple sources inference problems that are specified by the user.

Cases 2 and 3 are associated with a double source inference problem, where the observation patch is common to two events. In this case, the typical satellite image is synthesized using the stochastic probability maps. In any case, the $Y = 1$ observations are common to the two intersecting maps and the $Y = 0$ observations are within a box of Extension0. The generated observation patches of these 2 cases are illustrated in figures 4.5 and 4.9.

In these cases, which are associated with a number of sources N_s equal to 2, the chains of the sampling algorithm converged to the true sources and weights obtained using the global optimization algorithm, and the chains are also well-mixed. This well-mixing of the chains is illustrated in figure 4.6 of Case 2 or 4.10 of Case 3, where the adaptation of σ_r clearly enhances the mixing.

In addition, figures 4.7 and 4.11 illustrate the correlation maps between the different combinations of inferred sources or weights in Cases 2 and 3. Clearly, the true combinations of the sources and weights are in regions of high occurrences

of the inferred results of the chains obtained using the sampling algorithm. This is also validated in the marginal posterior probabilities of each model parameter for these two source inference cases as illustrated in figures 4.8 and 4.12.

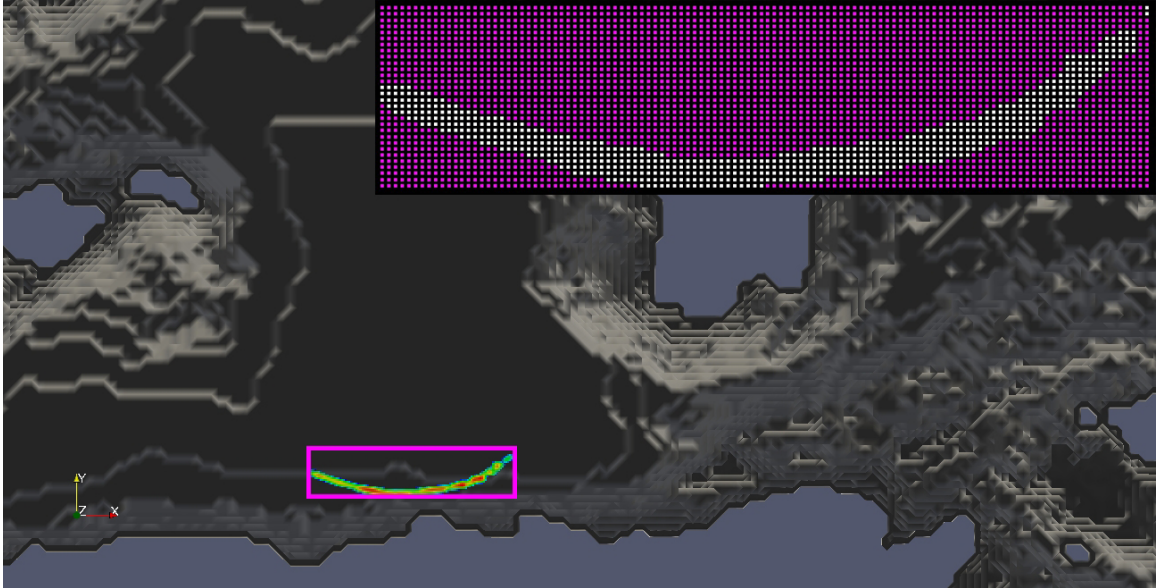
Now, in Case 4, which is associated with the inference of 4 sources contributing to 4 separate patches in the Mediterranean Sea, the sampling algorithm detected the sources and their relative weights while also quantifying the uncertainty in the inferred solution. The generated observation patches are shown in figure 4.13 which were generated by solving an advection diffusion problem with the MEAN realization of the velocity field. Figure 4.14 illustrates the converged and well-mixed chains of each model parameter in this 4 sources inference problem. Furthermore, figure 4.15 illustrates the correlation maps of the different possible combinations of source indices or relative weights obtained using the proposed sampling algorithm where clearly the algorithm predicts robustly the identity of the model parameters contributing to the observation patches. This is also validated in figures 4.16 and 4.17 that illustrate the marginal posterior probability of each source index and its relative weight.

Table 4.3: Multiple Source Inference problems in Trj. DayShift17.

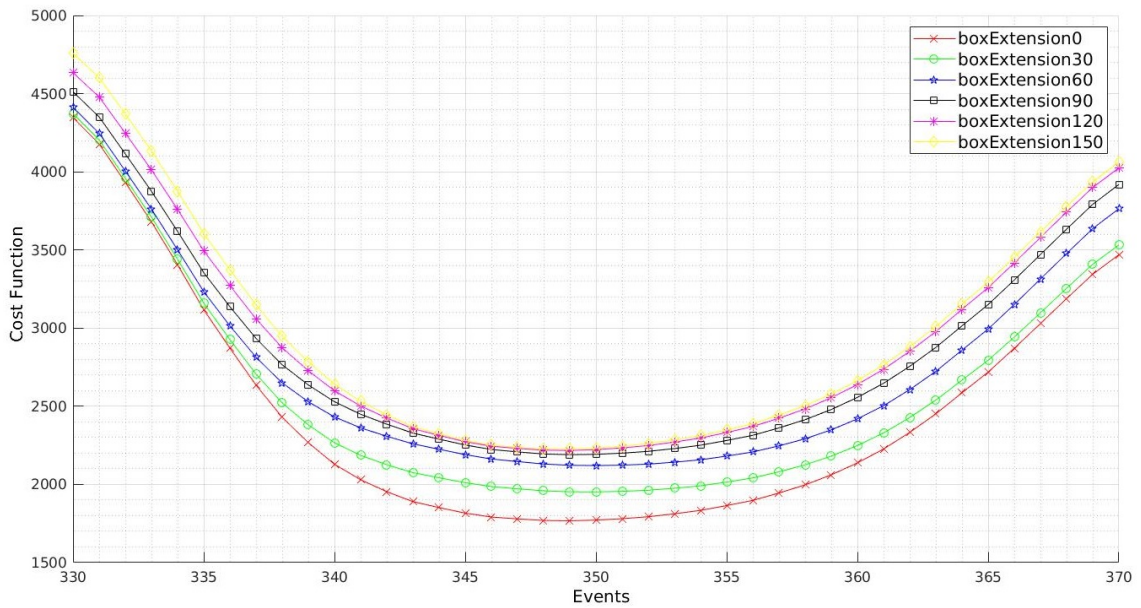
	Events	Observation Patch	Figures
Case 2	549 - 599	Synthetic (stoch. maps)	4.5, 4.6, 4.7, and 4.8
Case 3	300 - 400	Synthetic (stoch. maps)	4.9, 4.10, 4.11, and 4.12
Case 4	200 to 203 349 to 353 668 to 671 799 to 802	Determ.+ Diff. + MEAN	4.13, 4.14, 4.15, 4.16, and 4.17

Table 4.4: Sampling Algorithm parameters of the Multiple Source Inference problems in Trj. DayShift17.

	n_{Chains}	u_λ	l_λ	M_s	M_u	a_σ	burn-In
Case 2	5	0.5	2	10000	50	0.7	1000
Case 3	5	0.5	2	10000	100	0.7	1000
Case 4	5	0.5	2	10000	50	0.3	1000

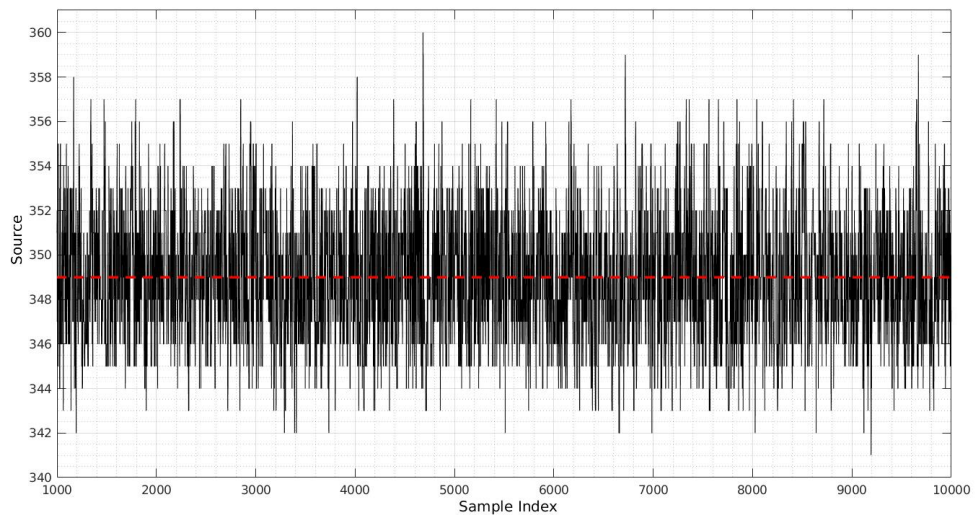


(a) Observation patch due to Events 349 to 353 using the MEAN realization of the velocity field.

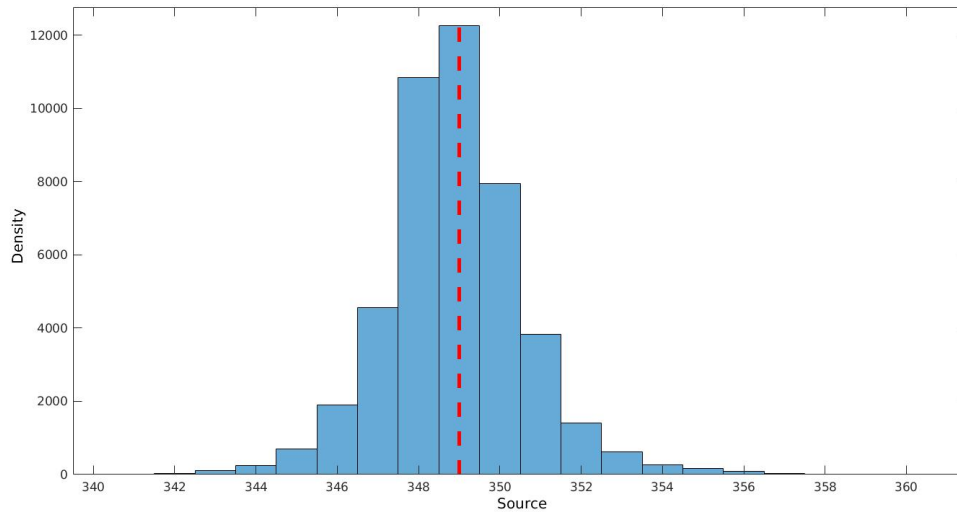


(b) Cost function variation for a subset of events for different box extensions.

Figure 4.3: Investigation of the effect of the size of the box on the cost function in Case 1.



(a) Inferred Events in a given chain in Case 1.



(b) Posterior Probability of the Source Index.

Figure 4.4: Summary of Results in Case 1.

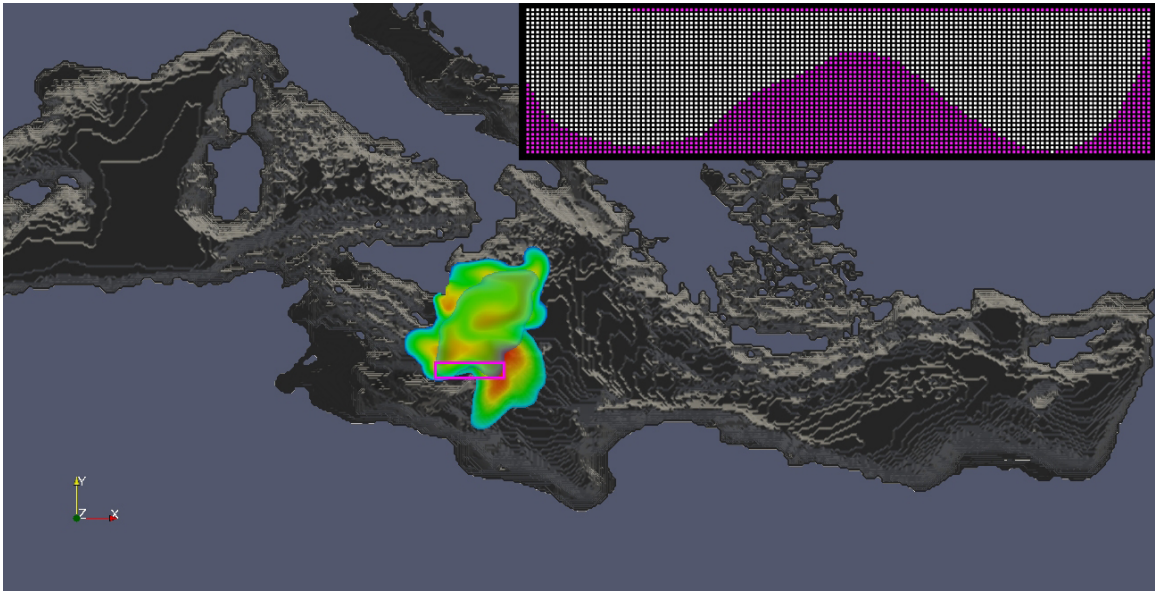
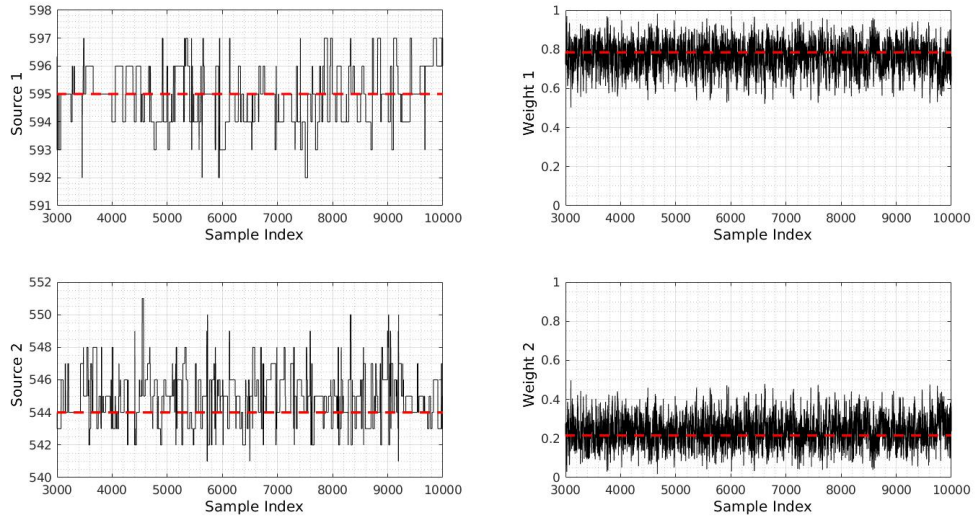
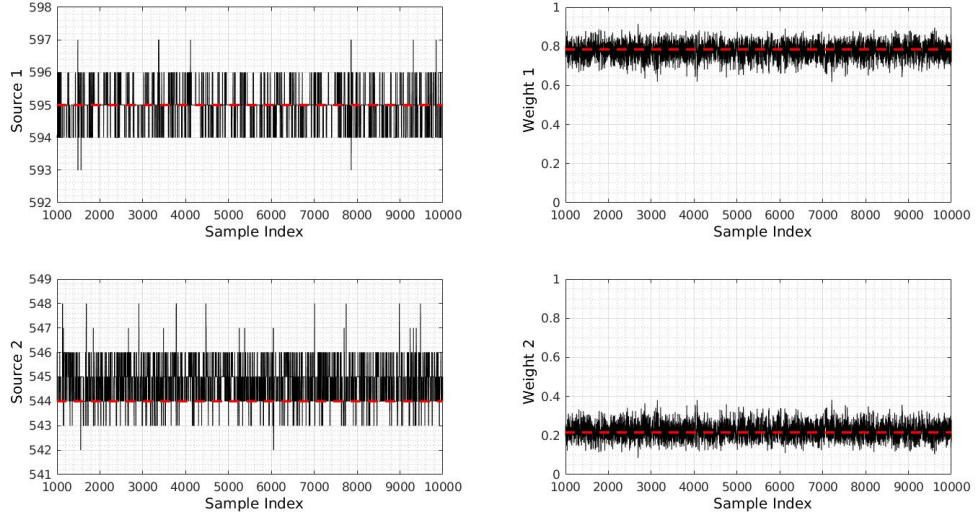


Figure 4.5: Synthetic Observation Patch of Case 2 obtained from the probability maps of Events 549 and 599.



(a) Inferred events and weights in a chain in Case 2 with no adaptivity of σ_r .



(b) Inferred events and weights in a chain in Case 2 with adaptivity of σ_r .

Figure 4.6: Results in Case 2.

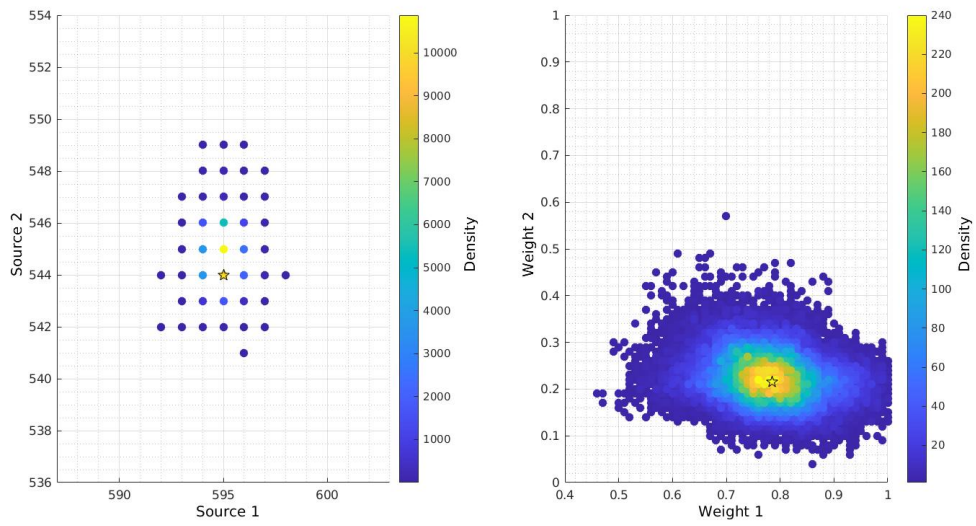
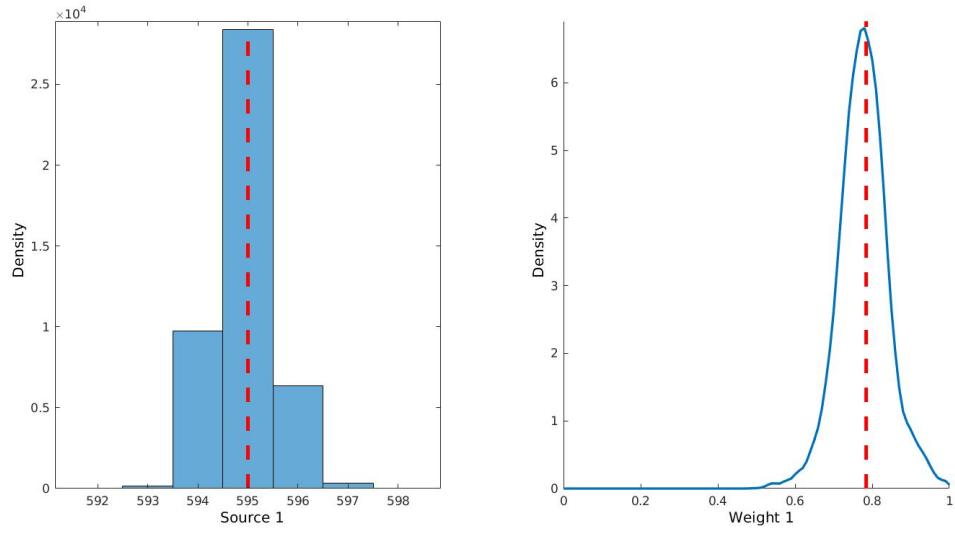
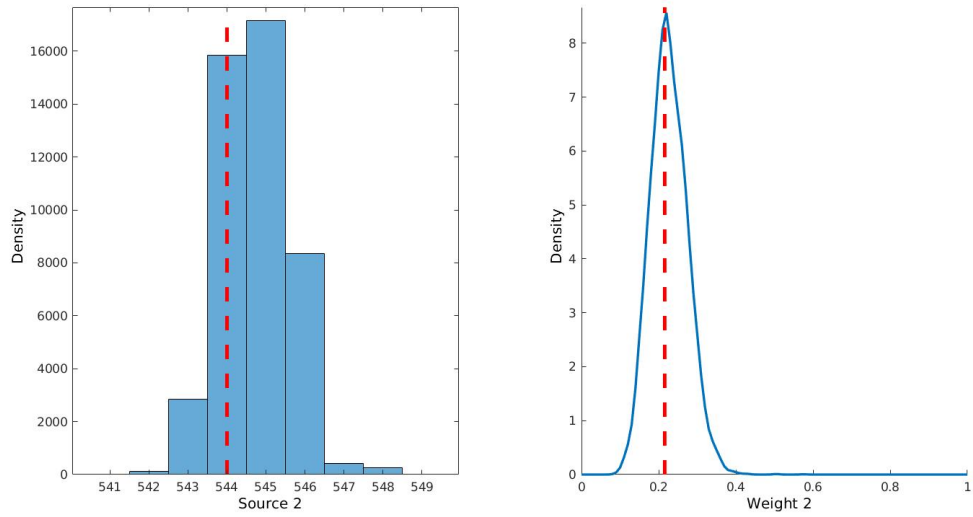


Figure 4.7: Correlation Maps in Case 2.



(a) Marginal Posterior Probabilities of Source 1.



(b) Marginal Posterior Probabilities of Source 2.

Figure 4.8: Marginal Posterior Probabilities in Case 2.

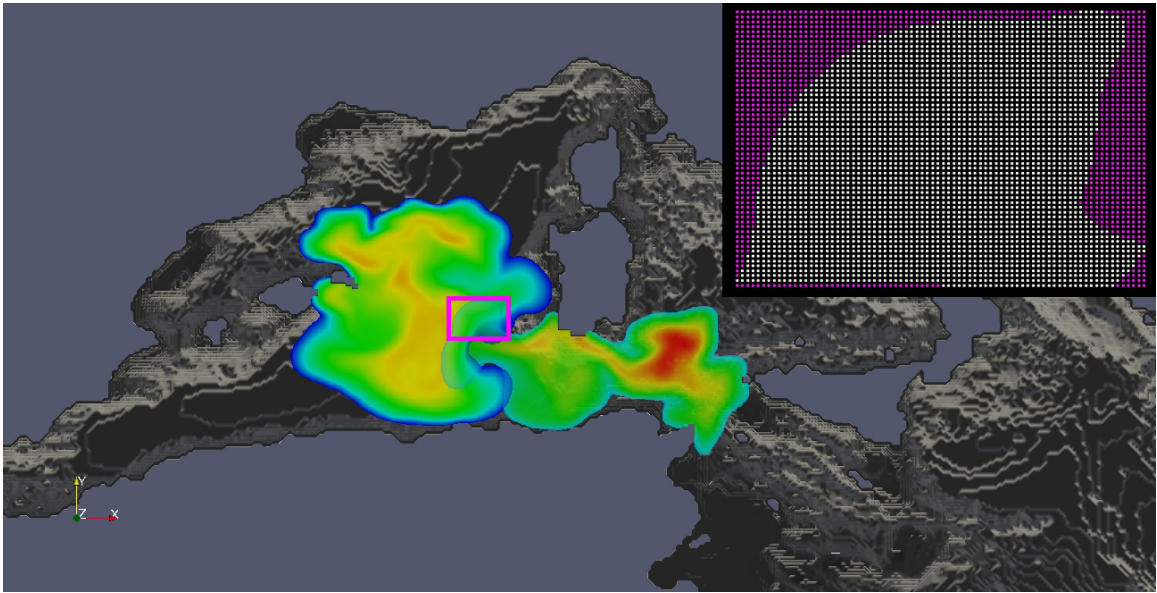
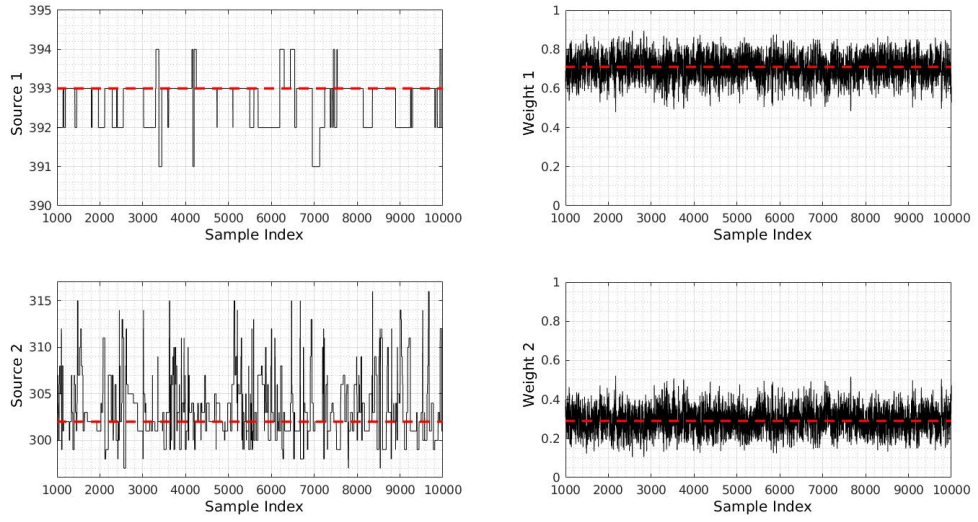
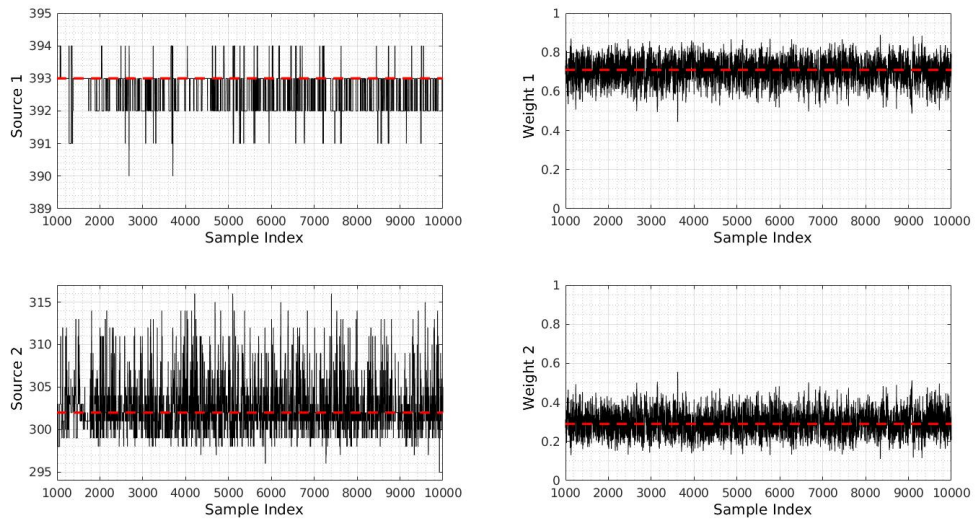


Figure 4.9: Synthetic Observation Patch of Case 3 obtained from the probability maps of Events 300 and 400.



(a) Inferred events and weights in a chain in Case 3 with no adaptivity of σ_r .



(b) Inferred events and weights in a chain in Case 3 with adaptivity of σ_r .

Figure 4.10: Results in Case 3.

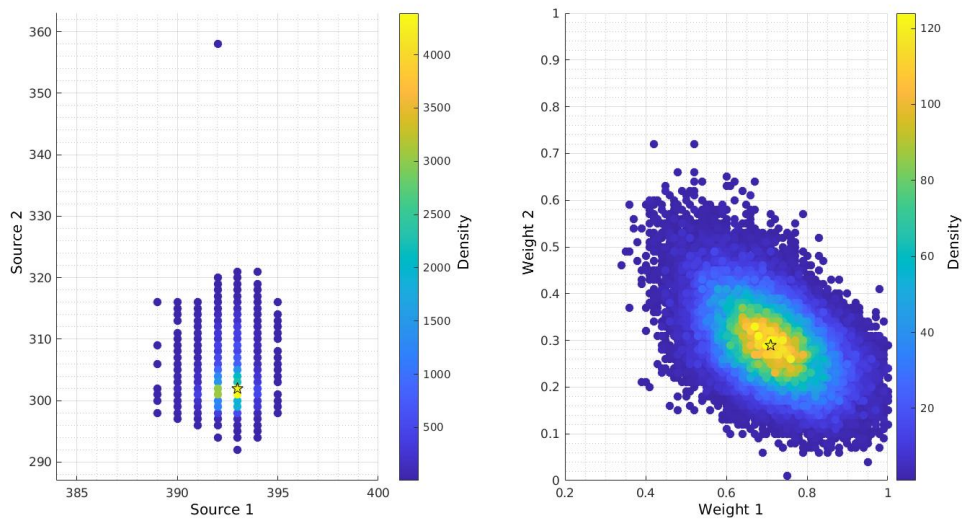
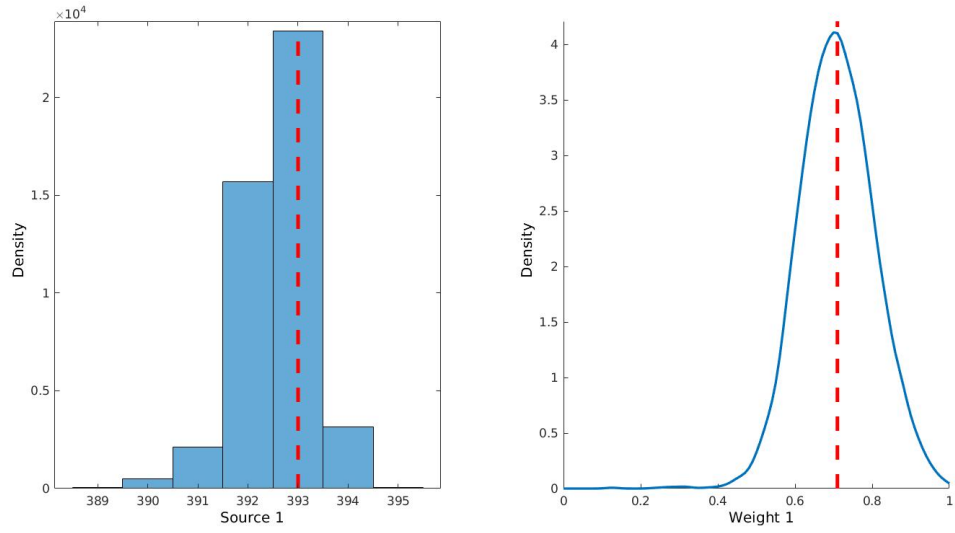
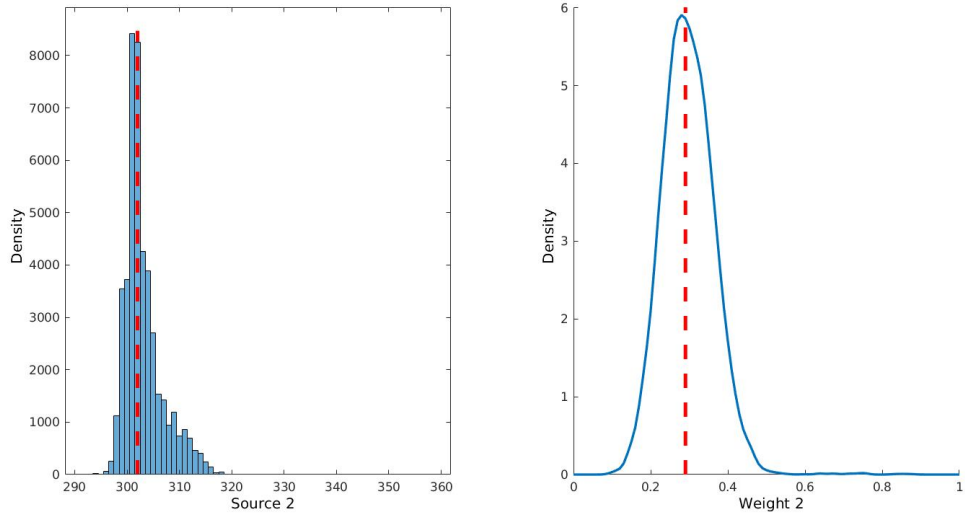


Figure 4.11: Correlation Maps in Case 3.

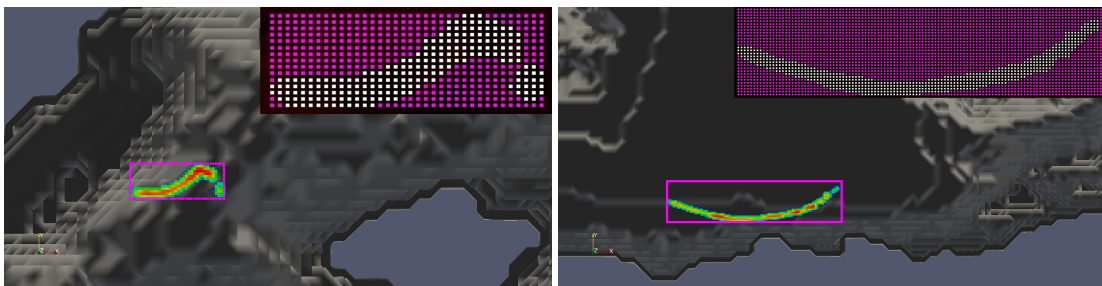


(a) Marginal Posterior Probabilities of Source 1.

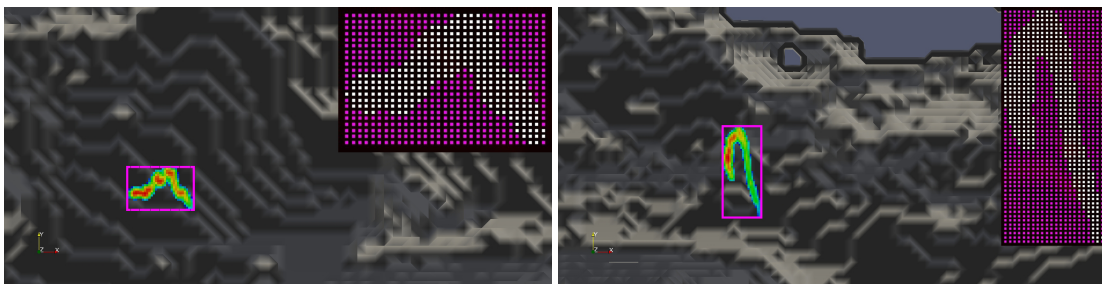


(b) Marginal Posterior Probabilities of Source 2.

Figure 4.12: Marginal Posterior Probabilities in Case 3.

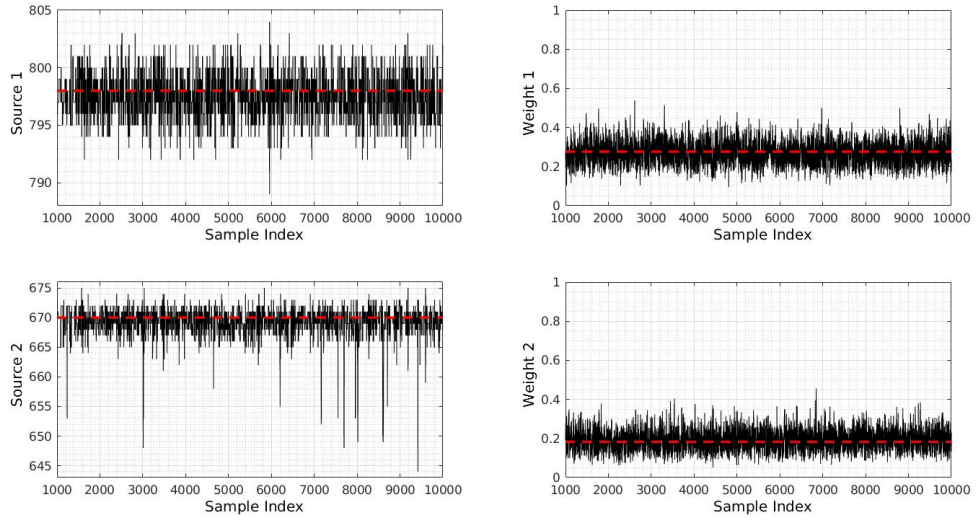


(a) Patch generated using Events 200-203. (b) Patch generated using Events 349-353.

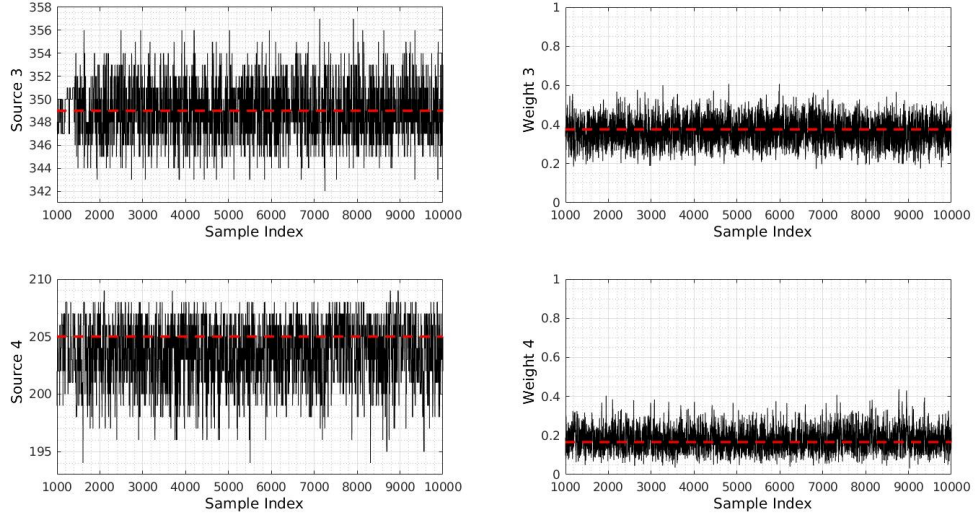


(c) Patch generated using Events 668-671. (d) Patch generated using Events 799-802.

Figure 4.13: 4 Separate Patches in Trj. DayShift17 generated using the MEAN of the velocity field.

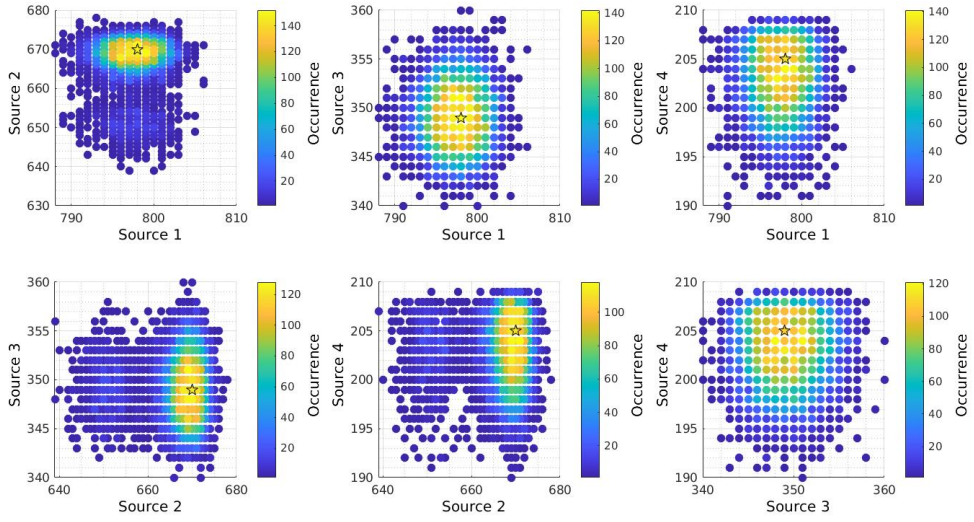


(a) Inferred events and weights for sources 1 and 2 in a chain.

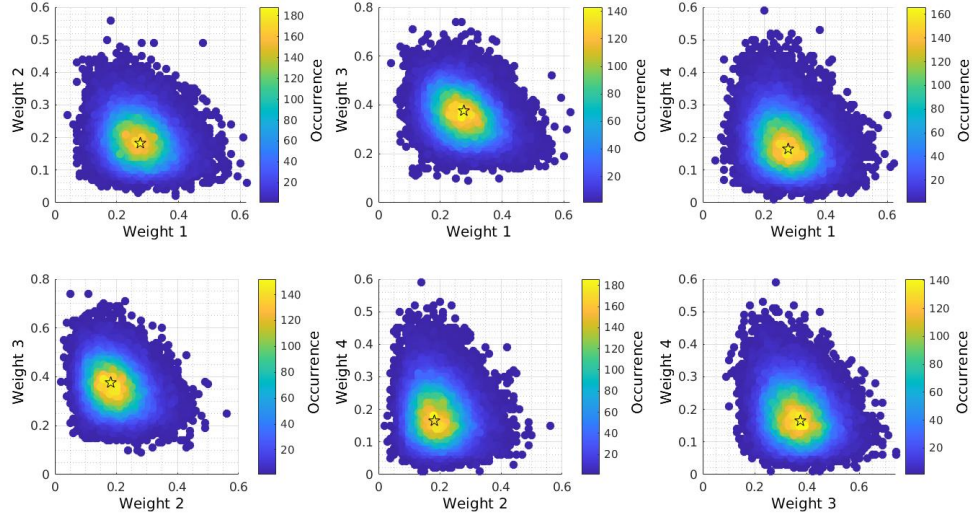


(b) Inferred events and weights for sources 3 and 4 in a chain.

Figure 4.14: Inferred Results in Case 4 with adaptivity of σ_r .

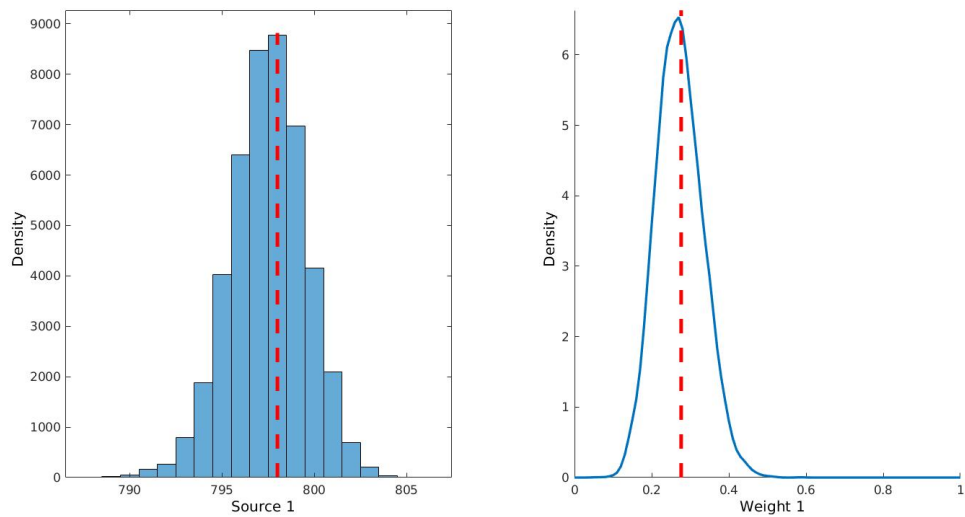


(a) Correlation Maps of the events.

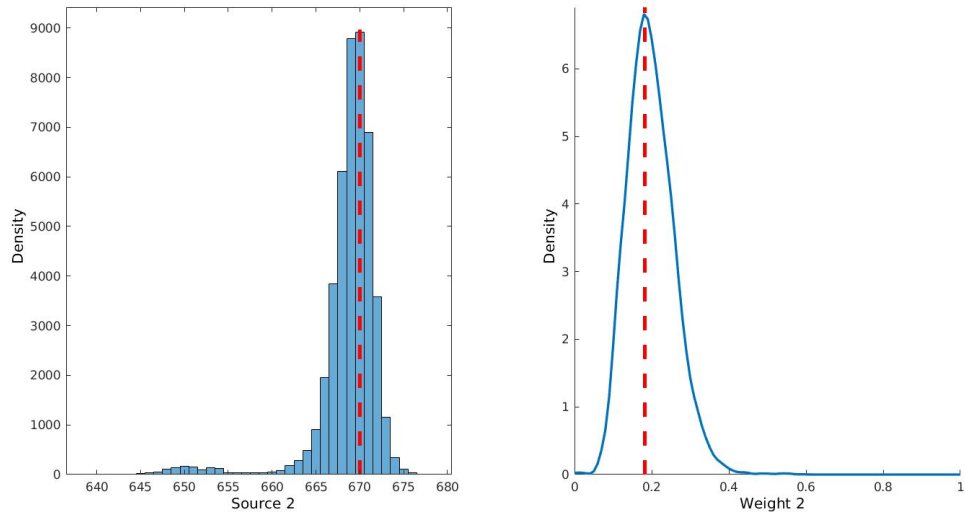


(b) Correlation Maps of the weights.

Figure 4.15: Correlation Maps in Case 4.

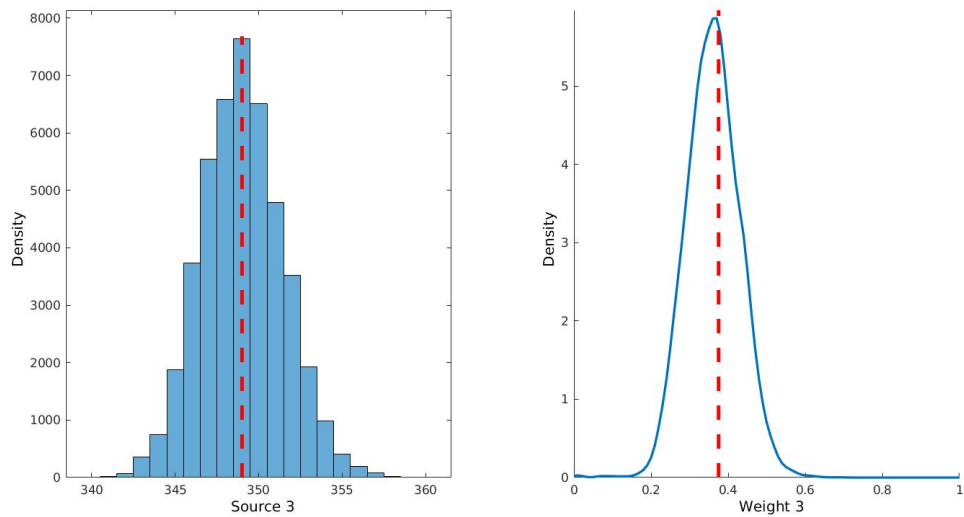


(a) Marginal Posterior Probabilities of Source 1.

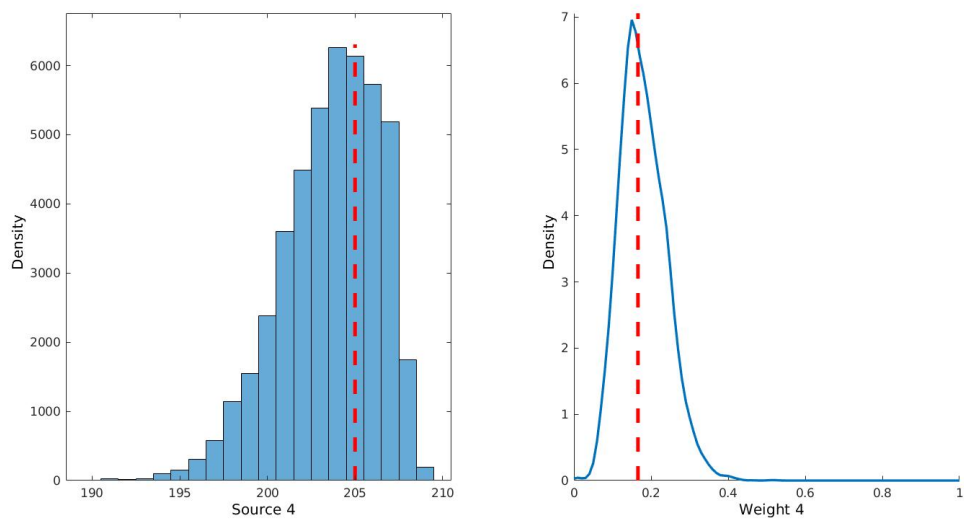


(b) Marginal Posterior Probabilities of Source 2.

Figure 4.16: Marginal Posterior Probabilities of sources 1 and 2 in Case 4.



(a) Marginal Posterior Probabilities of Source 3.



(b) Marginal Posterior Probabilities of Source 4.

Figure 4.17: Marginal Posterior Probabilities of sources 3 and 4 in Case 4.

Chapter 5

Conclusion and Future Work

To sum up, this thesis report included a detailed discussion of the methodology implemented in the stochastic transport of moving passive tracers in the Mediterranean Sea in the presence of a stochastic velocity field. This consisted of the selection of a suitable ship path along which pollutants are released instantaneously at different times and advected using a stochastic velocity field, as well as the generation of the corresponding probability maps.

The work was also extended to the development and implementation of a sampling algorithm that allows the inference of single and multiple sources on a given trajectory while quantifying the uncertainty in the solution for both the identity of source and its relative contribution to a given observation patch. In addition, a deterministic optimization algorithm was also implemented in order to validate the results obtained from the chains of the sampling algorithm.

Note that the observation patches were either synthesized or generated using a deterministic advection-diffusion model with the mean of the velocity field. Each observation patch is a typical satellite image, represented in a binary form, and indicates whether a pollutant is present or not.

Finally, future work will consist of the extension of the inference algorithm to multiple sources in multiple trajectories, as well as, the incorporation of the modelled and measured concentration data in both optimization and sampling approaches.

Appendix A

Abbreviations

\mathbf{D}	Data measurements and observations
\mathbf{M}	Model Parameters
\mathbb{A}	Mapping operator
\mathcal{M}	Hilbert Space of the parameters
\mathcal{D}	Hilbert Space of the data
\mathbb{A}^{-1}	Inverse Mapping Operator
\hat{M}	Optimum Solution of the model parameters
\mathbb{J}	Objective function
β	Regularized parameter
$\Phi(\mathbf{M})$	Regularization functional
F	Fitness function
I	Background information
$P(\mathbf{M} I)$	Prior probability
$P(\mathbf{D} \mathbf{M}, I)$	Likelihood probability
$P(\mathbf{D} I)$	Evidence
$P(\mathbf{M} \mathbf{D}, I)$	Posterior probability
N_{chains}	Number of Chains of the MCMC algorithm
\hat{M}_{MAX}	Maximum a posteriori estimate of M
\bar{M}	Posterior mean of M
$\sigma^2(M_i)$	Posterior standard deviation of M
$p\%$	Credible Interval of M
\mathbf{U}	Velocity field
\vec{x}_s	Source Location
t_a	Assimilation step
T_a	Assimilation time
\vec{x}_g	Grid coordinates
N_e	Ensemble size
\vec{u}_i	Realization i of the velocity
\vec{x}_o	Observation location
t_o	Observation time

m	Observation index
N_o	Number of observations
j	Index of the source
k	Index of the trajectory
t_R	Release time
$\vec{x}_R^{(s)}$	Release Location s
N_s	Number of sources along a given trajectory
N_τ	Number of trajectories
$f(\vec{x}_o^{(m)}(t_o) \vec{x}_s^{(j,k)}(t))$	Pdf of observation $\vec{x}_o^{(m)}(t_o)$ due to a ship $\vec{x}_s^{(j,k)}(t)$
Y	Actual observation (0 or 1)
r	Compact variable combining the indices of the sources and trajectories
Δs	Approximate length of a segment along the ship path
ΔT_R	Ship motion duration over Δs
V_s	Tanker speed
DayShift k	k th trajectory
N_{max}	Maximum number of particles
Δt_{adv}	Advection time
C	Scaling coefficient
Δx_p and Δy_p	x and y observation cell sizes
N_p	Number of $Y = 1$ observations
N_n	Number of $Y = 0$ observations
N_g	Number of grid points
σ_x	x -standard deviation of the generated Gaussian probability
σ_y	y -standard deviation of the generated Gaussian probability
σ_z	z -standard deviation of the generated Gaussian probability
t_{pt}	Particle travel time
D	Artificial diffusion coefficient
n_x	Number of x -grid points
n_y	Number of y -grid points
n_z	Number of z -grid points
P_S	Stochastic probability
P_D	Deterministic Probability
$\hat{q}^{(r)}$	Relative weight of source r
σ_r	Standard deviation of the source index r
$\sigma_{\hat{q}}$	Standard deviation of the source relative weight \hat{q}
λ	Hyperparameter
f_λ	Adaptive factor of the Hyper-parameter λ
AR	Acceptance rate of the MCMC chain
M_s	Number of samples in the MCMC chain
a_σ	Adaptive factor of σ_r
M_u	Number of samples for adapting σ_r
s_0	autocovariance at lag 0

Bibliography

- [1] L. Zeng, J. Gao, B. Du, R. Zhang, and X. Zhang, “Probability-based inverse characterization of the instantaneous pollutant source within a ventilation system,” *Building and Environment*, vol. 143, pp. 378–389, 2018.
- [2] L. Jing, R. Chen, X. Bai, F. Meng, Z. Yao, Y. Teng, and H. Chen, “Utilization of a bayesian probabilistic inferential framework for contamination source identification in river environment,” in *MATEC Web of Conferences*, vol. 246, p. 02035, EDP Sciences, 2018.
- [3] X. Zhou, V. Amaral, and J. D. Albertson, “Source characterization of airborne emissions using a sensor network: Examining the impact of sensor quality, quantity, and wind climatology,” in *2017 IEEE International Conference on Big Data (Big Data)*, pp. 4621–4629, IEEE, 2017.
- [4] A. Stohl, A. Prata, S. Eckhardt, L. Clarisse, A. Durant, S. Henne, N. I. Kristiansen, A. Minikin, U. Schumann, P. Seibert, *et al.*, “Determination of time-and height-resolved volcanic ash emissions and their use for quantitative ash dispersion modeling: the 2010 eyjafjallajökull eruption,” *Atmospheric Chemistry and Physics*, vol. 11, no. 9, pp. 4333–4351, 2011.
- [5] R. Humphries, C. Jenkins, R. Leuning, S. Zegelin, D. Griffith, C. Caldow, H. Berko, and A. Feitz, “Atmospheric tomography: a bayesian inversion technique for determining the rate and location of fugitive emissions,” *Environmental science & technology*, vol. 46, no. 3, pp. 1739–1746, 2012.
- [6] I. V. Kovalets, G. C. Eftimiou, S. Andronopoulos, A. G. Venetsanos, C. D. Argyropoulos, and K. E. Kakosimos, “Inverse identification of unknown finite-duration air pollutant release from a point source in urban environment,” *Atmospheric Environment*, vol. 181, pp. 82–96, 2018.
- [7] C. Jenkins, T. Kuske, and S. Zegelin, “Simple and effective atmospheric monitoring for co2 leakage,” *International Journal of Greenhouse Gas Control*, vol. 46, pp. 158–174, 2016.

- [8] G. Turbelin, S. Singh, P. Ngae, and P. Kumar, “An optimization-based approach for source term estimations of atmospheric releases,” *Earth and Space Science*, vol. 5, no. 12, pp. 950–963, 2018.
- [9] J. Wang, R. Zhang, Y. Yan, X. Dong, and J. M. Li, “Locating hazardous gas leaks in the atmosphere via modified genetic, mcmc and particle swarm optimization algorithms,” *Atmospheric environment*, vol. 157, pp. 27–37, 2017.
- [10] O. Paladino, A. Moranda, and M. Seyedsalehi, “A method for identifying pollution sources of heavy metals and pah for a risk-based management of a mediterranean harbour,” *Scientifica*, vol. 2017, 2017.
- [11] T. Borah and R. K. Bhattacharjya, “Development of an improved pollution source identification model using numerical and ann based simulation-optimization model,” *Water resources management*, vol. 30, no. 14, pp. 5163–5176, 2016.
- [12] A. Keats, E. Yee, and F.-S. Lien, “Efficiently characterizing the origin and decay rate of a nonconservative scalar using probability theory,” *ecological modelling*, vol. 205, no. 3-4, pp. 437–452, 2007.
- [13] H. Chhadé, F. Abdallah, I. Mougharbel, A. Gning, S. Julier, and L. Mihaylova, “Localisation of an unknown number of land mines using a network of vapour detectors,” *Sensors*, vol. 14, no. 11, pp. 21000–21022, 2014.
- [14] M. Hutchinson, H. Oh, and W.-H. Chen, “A review of source term estimation methods for atmospheric dispersion events using static or mobile sensors,” *Information Fusion*, vol. 36, pp. 130–148, 2017.
- [15] X. Zheng and Z. Chen, “Inverse calculation approaches for source determination in hazardous chemical releases,” *Journal of Loss Prevention in the Process Industries*, vol. 24, no. 4, pp. 293–301, 2011.
- [16] E. Yee and T. K. Flesch, “Inference of emission rates from multiple sources using bayesian probability theory,” *Journal of Environmental Monitoring*, vol. 12, no. 3, pp. 622–634, 2010.
- [17] F. Xue, H. Kikumoto, X. Li, and R. Ooka, “Bayesian source term estimation of atmospheric releases in urban areas using les approach,” *Journal of hazardous materials*, vol. 349, pp. 68–78, 2018.
- [18] E. Yee, “An operational implementation of a cbrn sensor-driven modeling paradigm for stochastic event reconstruction,” tech. rep., DEFENCE RESEARCH AND DEVELOPMENT SUFFIELD (ALBERTA), 2010.

- [19] E. Yee, “Source reconstruction: a statistical mechanics perspective,” *International Journal of Environment and Pollution*, vol. 48, no. 1-4, pp. 203–213, 2012.
- [20] E. Yee, “Automated computational inference engine for bayesian source reconstruction: Application to some detections/non-detections made in the ctbt international monitoring system,” *Applied Mathematical Sciences*, vol. 11, no. 32, pp. 1581–1618, 2017.
- [21] 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*, vol. 96, no. 10-11, pp. 1805–1816, 2008.
- [22] E. Yee, I. Hoffman, and K. Ungar, “Bayesian inference for source reconstruction: A real-world application,” *International scholarly research notices*, vol. 2014, 2014.
- [23] E. Yee, “Theory for reconstruction of an unknown number of contaminant sources using probabilistic inference,” *Boundary-layer meteorology*, vol. 127, no. 3, pp. 359–394, 2008.
- [24] E. Yee, “Inverse dispersion for an unknown number of sources: model selection and uncertainty analysis,” *ISRN Applied Mathematics*, vol. 2012, 2012.
- [25] S. El Mohtar, I. Hoteit, O. Knio, L. Issa, and I. Lakkis, “Lagrangian tracking in stochastic fields with application to an ensemble of velocity fields in the red sea,” *Ocean Modelling*, vol. 131, pp. 1–14, 2018.