In search of an effective Monte Carlo method for identification of atmospheric contamination source

Piotr Kopka\textsuperscript{1} and Anna Wawrzynczak\textsuperscript{1,2}
\textsuperscript{1} National Centre for Nuclear Research, Świerk-Otwock, Poland
\textsuperscript{2} Institute of Computer Science, Siedlce University, Poland
E-mail: piotr.kopka@ncbj.gov.pl

Abstract. We present the Bayesian framework able to identify the source of the airborne contaminant. We compare the effectiveness of framework utilizing the three popular Monte Carlo algorithms with the use of data from the full-scale Field experiment - Copenhagen Tracer Experiments. The input data are the on-line arriving concentrations of released substance registered by sensors, while the output are the posterior probabilities of four parameters characterizing the contamination source, i.e.: contamination source position (x,y,z) and release rate (q). As the forward model to predict the concentrations at the sensors locations, we use fast response Second-order Closure Integrated Puff model. The performer study present that among the Markov Chain Monte Carlo (MCMC), Sequential Monte Carlo (SMC) and Sequential Monte Carlo Approximate Bayesian Computation (SMC ABC) the last algorithm performed the best in the proposed framework localizing the airborne contaminant source parameters.

1. Introduction

In emergency response management it is important to know the extent of the area that might become contaminated following the release of dangerous material. Given a gas source and wind field, we can apply an appropriate atmospheric dispersion model to calculate the expected gas concentration for any downwind location. However, precise modeling the atmospheric contaminant transport is not trivial. On the another side, given concentration measurements and knowledge of the atmospheric air parameters identifying the release source is very difficult. This task can be understood as showcase of the dispersion model reproducing the encountered contamination. In the Bayesian approach, all quantities included in the mathematical model are modeled as random variables with joint probability distributions. Bayesian methods formulate this problem into searching for a posterior distribution based on efficient sampling of an ensemble of simulations using prior knowledge and observed data.

A very important factor influencing the effectiveness of the stochastic event reconstruction method is the selection of the sampling algorithm. Choosing the right sampling method is currently one of the most important issues. In most publications, authors choose classic algorithms, which then modified and adopted to the inverse problem (see [1]). Often as concentration data, a synthetic data sets are used (see [3]). [2] present an analysis of different algorithms from the area of global optimization, using Fusion field trial 2007 data. Currently, there have been no research based on dynamic data driven model for currently used probabilistic algorithms in problems. In [4] and [5] we presented an inverse procedure utilizing Sequential Monte Carlo Approximate Bayesian Computation (SMC ABC) to characterize the source of...
contamination in urban scale and reconstruction of the moving source position. In contrast to the likelihood-free approach we also used commonly applied in the data-driven contaminant source localization algorithm like Marcov Chain Monte Carlo (MCMC) and Sequential Monte Carlo (SMC) ([6]).

In the publication, we tested the three most popular algorithms: MCMC, SMC and Approximate Bayesian Computation (ABC) based on your own probabilistic computing framework to characterize the source of contamination in real experiment. To validate propose methodology we use data from Copenhagen Tracer Experiments (see section 2 and [7]). The sequential concentration measurements entails the use of a more advanced atmospheric dispersion model than Gaussian plume model. Thus, in the test cases, we apply the three-dimensional time-dependent Second-order Closure Integrated PUFF Model (see [8]). In addition, we added sequential modifications of the Metropolis-Hastings algorithm, due to the fact that it does not have a natural mechanism for transmitting information in time steps. The reconstructed parameters set includes \((x, y, z)\) source position, mass of release \((q)\) (see section 3) and we showed the possibility of comparing probabilistic algorithms in STE field by using credible regions. A thorough analysis of the reconstruction procedure results and the discussion is presented in section 4.

Figure 1. A topographic map of the CTE 19 October experiment. The red \(\times\) mark the points of the release, the sampling points are marked with green dots that are arranged in three arches \(ARC1\), \(ARC2\) and \(ARC3\).
2. Test Case: Copenhagen Tracer Experiments

The Copenhagen Tracer atmospheric dispersion experiments (CTE) were carried out in the Copenhagen area to explore the dispersion process of a tracer released from an elevated source in the mix of urban and residential area with roughness length of 0.1 m (see figure 1 and [7]). The tracer sulphurhexafluoride ($SF_6$) was released from a tower at a height of 115 meters (red cross at figure 1) and then the data are collected at elevation of 2-3 meters above ground-level, positioned 2-6 km from the point of release (green points at figure 1). As a single reconstructed event, the release of 19 October 1978 was chosen, based on analyzes of all 10 releases and determination of the set of data with significant concentration readings. In the figure 1 the significant sensors, have assigned numbers according to the [7] documentation. The gas concentration measurements were registered in time period between 12:13 and 13:13 and are divided into three 20 minute steps. While, release was continuous and amounted to 0.0032 kg/s of $SF_6$ gas. The figure 2 presents a graph of concentration values that has been registered by individual sensors.

3. Stochastic reconstruction methodology

In Bayesian inference, posterior distribution is related to prior distribution through a likelihood function:

$$\pi(\theta^{1:t}|d^{1:t}_{\text{obs}}) \propto \pi(d^{1:t}_{\text{obs}}|\theta^{1:t}, I)\pi(\theta^{1:t}, I)$$  \hspace{1cm} (1)

where $\theta^{1:t}$ and $d^{1:t}_{\text{obs}}$ denote the source parameters of interest and the measurement for subsequent time steps until $t$. The goal of Bayesian inference is to compute the posterior distribution $\pi(\theta^{1:t}|d^{1:t}_{\text{obs}}, I)$, where $\pi(\theta^{1:t}, I)$ is prior and $\pi(d^{1:t}_{\text{obs}}|\theta^{1:t}, I)$ is the likelihood of $\theta^{1:t}$ given the observed data $d^{1:t}_{\text{obs}}$ and background data $I$. The following parameters vector describe the source of the release $\theta = (x, y, z, q)$. The $(x, y)$ is source position within computing domain, $(z)$ is the height of source location above ground level, $(q)$ is a release rate. The prior probability distribution $\pi(\theta^{1}, I)$ express confidence about model parameters $\theta^{1}$ before first evidence data $d^{1}_{\text{obs}}$ is taken into account. We declare the following initial priori distribution on particular parameters: $\pi(\theta^{1}|I) \equiv \{x \sim \mathcal{U}(-2000, 8000), y \sim \mathcal{U}(-5000, 5000), z \sim \mathcal{U}(10, 200), q \sim \mathcal{U}(0.0002, 0.02)\}$. 

Figure 2. Measurements of the gas concentration in time intervals from 12:13-13:13 during the field tracer experiment conducted on 19 October. The sensor numbers corresponds to the sensors presented in Fig. 1 (numbered green points).
Source position $U(\left[ -2000, 8000 \right], \left[ -5000, 5000 \right])$ denotes uniform distribution over a domain rectangle. A right choice for the prior distributions of the source height is a family of continuous probability distributions fulfilling the condition $z > 0$. Thus, we set up the uniform distribution $z \sim U(10, 200)$ for the $z$ parameter, guaranteeing a realistic probability distribution over possible height.

3.1. Probabilistic dynamic model

Fig. 3 presents the graphical representation of the probabilistic model passing the inference from $\pi(\theta^1)$, $\pi(\theta^{1:2}|d_{obs}^1)$, ..., $\pi(\theta^{1:t}|d_{obs}^1)$, where $t = 3$. By using this bayesian dynamic model we take the advantage from the dynamic nature of the estimated posterior. This approach is often called recursive bayesian modeling, in which a posteriori distribution from previous time step $t-1$ becomes a prior distribution for the present step $t$. Fig. 3 show the relation between the model elements being a random variables $\theta^1$, $\theta^2$, $\theta^3$ and the sequence of posteriors $\pi(\theta^{1:t}|d_{obs}^{1:t-1})$.

To compute the $\pi(d_{obs}|\theta, I)$ value we use data from the sensor network which measure gas concentration $\hat{C}_{i}^{S_j}$, where $i$ corresponds to the time step and $S_j$ is the sensor identifier. In chosen test case concentrations where measured by sensors, so $d_{obs}^{1:t} \equiv [\hat{C}_{1}^{S_1}, \hat{C}_{2}^{S_2}, \hat{C}_{3}^{S_3}]$ for $j = 1, 2, \ldots, J$. The position of sensors are marked in fig. 1 by green dots. We assume that sensors report the substance concentrations subsequently at time intervals, hereafter referred to as ‘time steps’. In the inverse model we determined the error that occurs at each sensor $\hat{C}_{i}^{S_j} = \bar{C}_{i}^{S_j} + e_{obs}$ and is in the form of normal distributions $e_{obs} \sim N(0, \sigma^2)$, where $\bar{C}_{i}^{S_j}$ is mean concentration value (see fig. 2). The $\sigma^2$ value was estimated based on the parameters specified in the documentation [7], $\sigma^2 = 19.38$. Value of likelihood for a sample is computed by running a forward dispersion model with the given source parameters $\theta$ and comparing the model predicted concentrations $d_{obs}^{1:t}$ in the points of sensors location with actual observations $d_{obs}^{1:t}$. In probabilistic model the expected concentrations data $d_{obs}^{1:t}$ are calculated by forward dispersion model. As an input to dispersion model we use source parameter vector $\theta^t$ and the known background meteorological data $I$ presented in [7] table 17.

We have applied the atmospheric dispersion SCIPUFF model (see [8]), as the forward
model to compute the $SF_6$ concentrations at the sensors locations. It uses a collection of Gaussian puffs to represent an arbitrary three-dimensional, time-dependent concentration field and incorporates an efficient scheme for splitting and merging puffs. These features make the SCIPUFF model appropriate to apply it as a forward model in the reconstruction of the CTE experiment. The SCIPUFF output have a form of expected concentration values $SCIPUFF(\theta^t, I) \rightarrow d^{1:t} \equiv [C^{S_1}_t, C^{S_2}_t, \ldots, C^{S_J}_t]$ for $j = 1, 2, ..., J$ where $S_j$ corresponds to the $j - th$ sensor location. The variances of the modeling error for $\epsilon_{mod}$ factor are also estimated by the SCIPUFF program. The generation of samples from single probability distribution is done according to the testing algorithms.

3.2. Stochastic sampling alghorithms for Bayesian inference

The most popular method to calculate the posterior distribution is the Markov Chain Monte Carlo (MCMC) procedure based on the Metropolis Hastings sampling algorithm. In the first time step Markov chains are initialized by taking samples from the prior distribution, which is generally developed from a priori knowledge of the reconstructed event. Base on the simulated data and observed concentration the value of log-likelihood function is calculated using the formula:

$$L_{S_j}^{l-N}(d_{obs}^{1:t}, \theta^{1:t}, I) = \prod_{t=1}^{T} \prod_{j=1}^{J} \frac{1}{\sigma_{S_j}} \sqrt{2\pi} C_i^{S_j} \exp\left[-\frac{[\log(C_i^{S_j}) - \log(C_i^{S_j})]^2}{2\sigma_{S_j}^2}\right]. \tag{2}$$

If the proposed likelihood value of source term parameters sets is lower than that corresponding to the previous chain location, the proposal is accepted. If the comparison is worse, a random variable is used to decide whether or not to accept the new state. The main drawback of the MCMC methods for dynamic bayesian models is it does not have a natural way of carrying the probabilistic information available from the already generated sample $\theta^{1:i}$ to next time step. In SEQ-MCMC setup the prior distribution starting with second time step is basically approximation of posterior from past $p(\theta^{1:2} | d_{obs}^{1})$. The exact description of the algorithm used is in [3].

In contrast to the classical MCMC, Sequential Monte Carlo (SMC) is properly designed to sample from dynamic posterior distributions. Sequential Monte Carlo aims at using Importance Sampling to seamlessly generate samples from a sequence of distributions, $p(\theta^t), p(\theta^{1:2} | d_{obs}^{1}, I), \ldots, p(\theta^{1:i} | d_{obs}^{1}, I)$. To generate a sample set of size $N$ from the distribution $p(\theta)$ without having direct access to an method to do so, but we are able to evaluate $p(\theta)$ up to a proportionality constant. Importance sampling algorithm accomplishes this by using a proposal distribution $q(\theta)$, which is easy to generate samples and is close to the target distribution $p(\theta)$. This makes SMC methods particularly efficient for dynamically evolving models, like the one presented in the model 3. A detailed description of the algorithms can be found in [6] and [3].

The idea of Approximate Bayesian Computation (ABC) methods is to accept $\theta$ as an approximate posterior draw if its associate data $d$ is close enough to the observed data $d_{obs}$. The accepted parameters are a sample from $\pi(\theta | p(d, d_{obs}) < \epsilon)$ where the $p(d, d_{obs})$ is the chosen measure of discrepancy, and $\epsilon$ is a threshold defining as closeness margin. If $\epsilon$ is sufficiently small then the distribution $\pi(\theta | p(d, d_{obs}) < \epsilon)$ will be a good approximation for the posterior distribution $\pi(\theta | d_{obs})$. In ABC methods, SMC is used in order to sequentially clean approximation of posterior distribution which generate proposals for further steps. These methods aim to generate draws from $p(\theta^t | p(d^{1:t}, d_{obs}^{1:t}) < \epsilon_t)$, at each of a series of sequential steps $t$, where $\epsilon_t$ define a series of thresholds not single margin value as it was in ABC. In problem presented in this paper we use normalize approximation error between all the data obtained up
Table 1. Summary of the most important statistics: $\theta^*$ target values, $\theta^{MAP}$ maximum a posteriori estimation, $CI$ credible interval, $CR$ percent of credible regions.

| a)   | methods       | $x[m]$   | $y[m]$   | $z[m]$   | $q[kg/s]$ |
|------|---------------|----------|----------|----------|------------|
| $\theta^*$ | MCMC-SEQ     | 1093.75 ± 53.00 | -406.25 ± 52.41 | 143.59 ± 3.25 | 0.0010 ± 0.000001 |
|       | SMC           | 514.45 ± 53.00   | 57.80 ± 52.41   | 176.30 ± 3.25   | 0.0012 ± 0.000001 |
|       | SMC-ABC       | 616.28 ± 53.00   | 58.14 ± 52.41   | 183.43 ± 3.25   | 0.0023 ± 0.000001 |

| $\theta^{MAP}$ | MCMC-SEQ   | [-1913.67 1786.68] | [-1271.09 2601.55] | [69.68 200.06] | [0.0002 0.0013] |
|                | SMC         | [-156.20 1532.17] | [-561.66 1228.49] | [109.22 198.95] | [0.0007 0.0021] |
|                | SMC-ABC     | [-1894.11 4250.13] | [-1465.67 2651.15] | [47.85 199.54] | [0.0001 0.0157] |

| $CR^{50\%}(\theta)$ | MCMC-SEQ | [-468.01 919.27] | [-346.21 751.66] | [152.36 193.58] | [0.0004 0.0019] |
|                     | SMC       | [-1825.30 848.79] | [-174.42 2267.72] | [26.78 198.27] | [0.0047 0.0088] |
|                     | SMC-ABC   | [-58.14 732.28]  | [-58.14 732.28]  | [84.38 198.76] | [0.0014 0.0037] |

| $CR^{50\%}(\theta)$ | MCMC-SEQ | 40.48% | 56.98% | 34.97% | 27.81% | 34.39% | 64.81% |
|                     | SMC       | 23.68% | 31.85% | 25.2%  | 20.04% | 22.19% | 59.85% |
|                     | SMC-ABC   | 5.28%  | 26.11% | 25.2%  | 20.04% | 22.19% | 59.85% |

| $CR^{90\%}(\theta)$ | MCMC-SEQ | 3.41%  | 15.63% | 13.61% | 8.42%  | 7.79%  | 18.8% |
|                     | SMC       | 2.98%  | 6.14%  | 5.42%  | 2.39%  | 2.12%  | 3.95% |
|                     | SMC-ABC   | 0.82%  | 9.38%  | 6.79%  | 2.42%  | 1.89%  | 9.9%  |

The exact description of the procedure (SMC-ABC) was presented in the works [4], where original idea was presented in [9].

4. Reconstruction results

The summarized results of reconstruction procedure are presented in figure 4-6 as a spatial trellis plot. Recall that a color corresponds to the empirical 2D probability distribution of all parameters combinations. The colored contour lines are enveloping higher probability of the joint posterior distributions. The diagonal plots are marginal empirical posterior distributions of the parameters. The parameters target values $\theta^*$ are highlighted with a vertical red line in diagonal subplots and black cross markers on others subplots. Areas marked with yellow and orange color define the Bayesian credible interval ($CR^{50\%}$, $CR^{90\%}$) and credible regions ($CR^{50\%} CR^{90\%}$). A credible parameters are an intervals/regions within which an unobserved parameter value $\theta$ falls with a particular subjective bayesian probability. In table 1 a summary of the most important statistics obtained by running all and percentages of two-dimensional parameter subspaces covering the regions $CR^{90\%}$ and $CR^{50\%}$.

The proper estimation of the source position is crucial to correctly point out other parameters. The values of $\theta^{MAP}$ (Maximum a Posteriori Estimation) differs from target parameter $\theta^*$, depending on the algorithm from 1093.75m to 514.45m for the parameter $x$ and 406.25m to 57.80m for $y$. For the $y$ parameter, the absolute difference in the distance between $\theta^{MAP}(y)$ is even more pronounced and equals $\approx 350$m. Summing up, taking into account the MAP estimator for source location indicators, the SMC algorithm is best with the results of $p(x = 514.45m) = 0.024$ and $p(y = 57.80m) = 0.028$. The result obtained by the approximation method, despite slightly larger discrepancies in the values of $\theta^{MAP}(x)$ and $\theta^{MAP}(y)$ is characterized by a greater concentration of probability mass around the target location $(x, y) = (0, 0)$. This is especially visible if we compare the diagrams on the diagrams of the figure 5 and the figure 6. Areas with significant probability values, determined by MCMC-SEQ and SMC algorithms marked with red contours on panels above diagonal in drawings 4 - 5 are much less concentrated than those...
shown in the figure 6. This has a significant effect on the percentage values of the confidence areas that for the SMC-ABC algorithm include $CR^{90\%}(x,y) = 5.28\%$ and $CR^{50\%}(x,y) = 0.82\%$ of the total space $(x, y)$.

From the point of view of the SER methodology, $q$ parameter is the most difficult to reproduce characteristics of source. In the figures 4-5, where the posterior distributions for the MCMC-SEQ and SMC algorithms are shown, the $q$ parameter has not been properly determined. The consequence of the wrong estimation of the emission targets of $q$ and other parameters can be seen in the lower panels of the figure 5, where the $CR^{50\%}$ areas are visible. When using the SMC-ABC algorithm, the value $\theta^{MAP}(q)$ is $p(q = 0.0023\, \text{kg/s}) = 0.055$, and the empirical distribution from the figure 6 despite moderate asymmetry and a relatively "heavy" tail, it well defines the target value of the emission level of $q = 0.0032\, \text{kg/s}$. In addition, on the lower panels of the figure 6 areas $CR^{90\%}$ and the much more stringent $CR^{50\%}$ cover the searched values. The estimated by the algorithm most probable release mass is $P(q = 265.7 \pm 10.4\, \text{mg/s}) = 0.042$, while the target value is $q = 323\, \text{mg/s}$. So, the release mass is underestimated about $\sim 50\, \text{mg/s}$. The altitude at which the $z$ source was located was not correctly determined by any of the algorithms. We can only observe a tendency to increase the probability as the height increases.

5. Summary
We have developed an dynamic atmospheric release event reconstruction methodology based on Bayesian inference and we thoroughly tested the most popular and general sampling procedures setups: SEQ-MCMC, SMC, SMC-ABC. The results of the all inversion indicate the probability of a source being found at a particular location with a particular release rate in the form of the join posterior probability distribution. By using the advanced a time-dependent
Figure 5. The bivariate and marginal posterior distributions for all searched parameters obtain by SMC algorithm.

Figure 6. The bivariate and marginal posterior distributions for all searched parameters obtain by SMC-ABC algorithm.
SCIPUFF dispersion model along with the evaluated sampling methods, performed calculations yield an accurate posterior distribution, whereby it is possible to compare the effectiveness of sampling algorithms in dynamic data driven problem. We also developed stochastic algorithms SEQ-MCMC for relatively fair comparison. The Copenhagen Treces Experiment dataset was used for a study into the design event reconstruction comparative methods based on Credible regions. We initiated the development of an SMC-ABC methodology and demonstrated its performance advantages for treating time dependent data models.

References

[1] Rajaona, H. at al 2015 An adaptive Bayesian inference algorithm to estimate the parameters of a hazardous atmospheric release. *Atmospheric Environment* **122** 748-762

[2] Platt N and DeRiggi D 2012 Comparative investigation of Source Term Estimation algorithms using FUSION field trial 2007 data: linear regression analysis. *International Journal of Environment and Pollution* **48** 13-21

[3] Johannesson G, Hanley B and Nitao J 2004 *Dynamic bayesian models via monte carlo-an introduction with examples* (Lawrence Livermore National Laboratory, UCRL-TR-207173)

[4] Kopka P and Wawrzynczak A 2018 Framework for stochastic identification of atmospheric contamination source in an urban area. *Atmospheric Environment* **195** 63-77

[5] Kopk P, Wawrzynczak A and Borysiewicz M 2016 Application of the Approximate Bayesian Computation methods in the stochastic estimation of atmospheric contamination parameters for mobile sources. *Atmospheric Environment* **145** 201-212

[6] Borysiewicz M, Wawrzynczak A and Kopka P 2012 Stochastic algorithm for estimation of the model’s unknown parameters via Bayesian inference. *FedCSIS, Federated Conference on IEEE* 501-508

[7] Gryning S E and Lyck E 1998 *The Copenhagen tracer experiments: Reporting of measurements* (Copenhagen: Riso National Laboratory)

[8] Sykes RI at al 1996 *SCIPUFF-A generalized hazard dispersion model* (Boston: American Meteorological Society)

[9] Bonassi F V and West M 2015 Sequential Monte Carlo with adaptive weights for approximate Bayesian computation. *Bayesian Analysis* **10** 171-187