Approximate Bayesian Computation Methods in the Localization of Atmospheric Contamination Sources in an Urban Area

Piotr Kopka[1,2], Anna Wawrzyńczak[2] and Mieczyslaw Borysiewicz[2]

[1] Institute of Computer Science, Polish Academy of Sciences
ul. Jana Kazimierza 5, 01-248 Warsaw, Poland
[2] National Centre for Nuclear Research,
Otwock, Poland

Abstract. Sudden releases of harmful material into a densely-populated area pose a significant risk to human health. The apparent problem of determining the source of an emission in urban and industrialized areas from the limited information provided by a set of released substance concentration measurements is an ill-posed inverse problem. When the only information available is a set of measurements of the concentration of released substances in urban and industrial areas it is difficult to determine the source of emission. However, the problem can be solved when there is additional information available together with the appropriate tools. A convenient choice is the Bayesian probability framework, which provides a connection between model, observational and additional information about the source. The Bayesian approach was applied in this study to find the posterior probability density function of the contamination source parameters (location and strength) given a set of concentration measurements. The posterior distribution of the source parameters was sampled using an Approximate Bayesian Computation (ABC) algorithm. The stochastic source determination method was validated against the real data set acquired in a highly disturbed flow field in an urban environment. The datasets used to validate the proposed methodology include the dispersion of contaminant plumes in a full-scale field experiment performed within the project "Dispersion of Air Pollutants and their Penetration into the Local Environment in London (DAPPLE)". It demonstrates the use of the proposed approach for the event reconstruction problem in a highly urbanized environment.

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 in cities and the subsequent movement of polluted air. The lack of pertinent experimental information means there is a gap in the understanding of

short range dispersion behavior in highly urbanized areas. Given a gas source and wind field, we can apply an appropriate atmospheric dispersion model to calculate the expected gas concentration for any location. Conversely, given concentration measurements and knowledge of the arrangement of buildings, wind field and other atmospheric air parameters, identifying the actual location of the release source and its parameters is difficult. This problem has no unique solution and can be analyzed using probabilistic frameworks. In the framework of Bayesian approach, all quantities included in the mathematical model are modeled as random variables with joint probability distributions. This randomness can be interpreted as lack of knowledge of parameter values, and is reflected in the uncertainty of the true values as expressed in terms of probability distributions. Bayesian methods reformulate the problem thusly: by comparing data, and efficient sampling of a group of simulations, to find a solution.

The problem of source term estimation has been studied in literature, based on both the deterministic and probabilistic approach. [1] implemented an algorithm based on integrating the adjoint of a linear dispersion model backward in time to solve a reconstruction problem. [2, 3] introduced dynamic Bayesian modeling, and the Markov Chain Monte Carlo (MCMC) sampling approaches to reconstruct a contaminant source for synthetic data. Source reconstruction in an urban environment using building-resolving simulations was studied in [4] and [5]. [4] used an adjoint representation of the source-receptor relationship. They used a Bayesian inference methodology in conjunction with MCMC sampling procedures. This approach was validated using data from water channel simulations and a field experiment (Joint Urban 2003) in Oklahoma City. In [5] the authors applied the methodology presented in [2] to the reconstruction of the flow around an isolated building and the flow during IOP3 (third intensive observation period) and IOP9 of the Joint Urban 2003 Oklahoma City experiment. In these experiments they found the source location $\sim$ 70m from the true location for IOP3 (within the domain $\sim 400m \times 400m$) while for the IOP9 model errors and other uncertainties limit the ability to pinpoint the source location.

Methods of approximate Bayesian computation (ABC) are especially useful for problems in which the likelihood function is analytically intractable or too expensive to compute. The original version of the approximate Bayesian computation with Sequential Monte Carlo (ABC SMC) algorithm was proposed in [6]. Applications of this algorithm have been presented in a variety of areas including population biology [7], genetics [8] and psychology [9]. Also, there has been an increased interest in extensions and improvements of this algorithm, as demonstrated in ([10], [11], [12], [13]). The more advanced form of the algorithm, which relies upon the new idea "Sequential Monte Carlo with Adaptive Weights", is shown in Algorithm 1 section 4 and was originally presented in [14].

Previously [15], we have tested the methodology by combining Bayesian inference with MCMC methods and applied these to the problem of dynamic, data-driven contaminant source localization, based on data from the the synthetic experiment. In [16] various modifications of the MCMC algorithm to estimate the probability distributions of searched parameters were examined. We

have shown the advantages of several algorithms. These algorithms use, in a variety of ways, the probability distributions of the source location parameters obtained based on available measurements. Once the new concentration data are received, the marginal probability distribution of the selected parameters is updated. We have also presented the application of the Sequential Monte Carlo (SMC) methods combined with the Bayesian inference to the problem of locating the atmospheric contamination source based on synthetic experiment data [17].

We propose algorithms to locate the source of contamination based on the data from the central London DAPPLE experiment that was performed in May and June 2007 (see section 2) [18]. We used the fast running QUIC-URB [19] model for computing mean flow fields around buildings and QUIC-PLUME [20] as the forward model to predict the concentrations at the sensor locations (section 3). As a sampling approach in the event reconstruction procedure we used the modern algorithm from the class of likelihood-free Bayesian methods [14] with some extension, described in section 4.

2 Dispersion Experiments in London - DAPPLE

The DAPPLE experiment took place in central London (see fig. 1). The two major roads in the vicinity are Marylebone Road, which runs from west to east, and Gloucester Place, which intersects perpendicularly with Marylebone Road near the Westminster City Council building (the red star in fig. 1) [18]. The mean building height in the study area is $21.6m$ (range 10 to $64m$). The experimental site was chosen so as to have a diameter of approximately $500m$ in order to cover the whole dispersion field. There are over 50 experiment sets of dispersion from point sources in the whole DAPPLE data, but to address the issue of source reconstruction we selected a time-resolved contamination experiment. A selected release was carried out on the fourth day, 28th June 2007, in which a sequence of ten samples was taken over a 30 minute sampling period at each of the 18 receptor positions. The sampling process included the collection of ten $150s$ samples at each of the 18 sites, each sample separated from the next by $30s$. The source locations (green X point) and monitoring sites (numbered yellow points) are shown on the map included in fig. 1. The total mass emitted from point-source release was $323mg$ of *perfluoromethyl-cyclohexane* ($PMCH, C7F14$), in accordance with experimental requirements. The other source locations Y and Z were chosen and fixed for the run of experiments conducted during each tracer day. This choice was based on analysis of the weather forecast on the preceding day and a reconstruction of these sources is not present in this publication. Two sets of long-term reference measurements were taken to generate the wind data sets: the rooftop Westminster City Council (WCC) ($18m$) and tower top ($190m$) winds. In order to not increase the height of the domain in the calculations only data from WCC has been taken into account. All aggregate information of the analyzed experiments and wind condition are shown in table 1.

Fig. 1: The map shows the DAPPLE area of central London and is centered at the focal intersection, that of Marylebone Road and Gloucester Place (at 51.5218N 0.1597W). The sampling receptors are numbered 1-18 (yellow dots). Three fixed-point tracer sources (green dots X,Y and Z); red star - Westminster City Council (WCC). The white rectangle shows the computational domain.

In fig. 1 the rectangle area was separated as a computing domain (white line). The positions of all the objects (sensors, source, buildings, wind direction, etc.) have been rotated by 17^o angle, in order to fix the main streets parallel to the edges of the domain. The $latitude - longitude$ geographic coordinate system was changed to the metric system with a reference point $(0, 0)$. This reference point denotes the lower left corner of the white rectangle, both for the convenience of creating a domain and the presentation of results. The domain after the transformation is presented in fig. 2a. All the information presented above (experiment setup - table 1 and the geometry of the domain fig. 2) have been introduced into the Quick-URB environment, which is described in the next section.

DAPPLE experiment summary	
Date and time	28 Jun 07 13:00
Number of tracers released experiment	3
Number of samples and sample duration experiment	10x3 mins
Number of sampling sites	18
Range of source-receptor separations(m)	22-437
Point-source release total mass (mg)	323
WCC Roof data summary	
Wind speed $\frac{m}{s}$	2.6
Wind direction	+19
Longitudinal turbulence u'/U_H	0.80
Lateral turbulence v'/U_H	0.59
Vertical turbulence w'/U_H	0.27

Table 1: DAPPLE and WCC summary [18]

3 Forward dispersion model - QUIC

The Quick Urban Industrial Complex (QUIC) Dispersion Modeling System is intended for applications where dispersion of air pollutants released near buildings must be computed very quickly [20]. The QUIC system, comprises a wind model - QUIC-URB, a dispersion model QUIC-PLUME, and a graphical user interface. The modelling strategy adopted in QUIC-URB was originally developed by Rockle [21] and uses a 3D mass-consistent wind model to combine properly resolved time-averaged wind fields around buildings [22]. The mass-consistent technique is based on a 3D complex terrain diagnostic wind model. The basic methodology involves first generating an initial wind field that includes various empirical parameterizations to account for the physics of flow around buildings. Next, this velocity field is forced to be divergence free, subject to the weak constraint that the variance of the difference between the initial velocity field and mass consistent final velocity field is minimized. The ability of the QUIC-URB model to produce proper wind fields around buildings is dependent on the empirical wind parameterizations. These parameterizations introduce rotation into the flow field and without these parameterizations the method is essentially a potential flow solver. QUIC-PLUME uses a stochastic Lagrangian random walk approach to estimate concentrations in a gridded domain. The model is designed to use averaged wind fields produced by the QUIC-URB system. Parcels, representing substances, are transported with a vector sum of mean winds from QUIC-URB plus turbulent fluctuating winds computed using the random walk equations. Turbulence parameters required in the random walk equations are estimated from vertical and horizontal gradients in the mean wind. A detailed description of the theory is described in [23]. Fig. 2b shows a 3D domain model of the part of London created in QUIC-GUI environment based on the extracted most important buildings from fig. 2a. On the other hand, figs. 2c and 2d present the output of subsystem QUIC-URB which is a wind flows map between

the buildings obtained from WCC measurements. QUIC-PLUME is a 'forward' model, that is run repeatedly for various parameter sets representing position and sources based on the Bayesian inference tool presented in section 4.

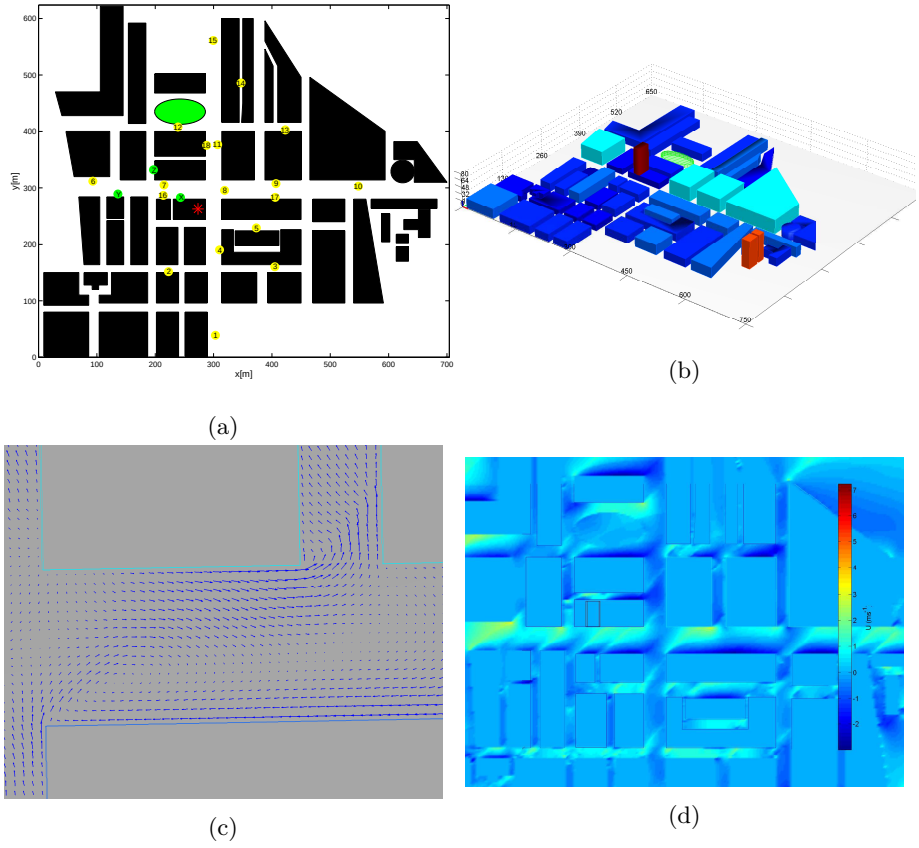

Fig. 2: a) The rotated DAPPLE area, with the selected buildings (black rectangles) and greenery (green ellipses), created using the map from fig. 1 ; sampling receptors are numbered 1-18 (yellow dots), three fixed-point tracer sources (green dots X,Y and Z); red star - Westminster City Council (WCC) b) 3D model of city buildings designed in QUIC-GUI base on the maps. c) map section presenting the wind vectors in the given points d) map presenting the strength of the wind between the buildings in experimental area

4 ABC Methodology

Let θ be a parameter vector, with the prior distribution $\pi(\theta)$. The goal of the Bayesian inference is to approximate the posterior distribution, $\pi(\theta|x) \propto$

$\pi(x|\theta)\pi(\theta)$, where $\pi(x|\theta)$ is the conditional distribution of θ given the data x. The main idea of Approximate Bayesian Computation (ABC) methods is to accept θ as an approximate posterior draw if its associate data x is close enough to the observed data x_{obs}. Accepted parameters are a sample from $\pi(\theta|\rho(x, x_{obs}) < \epsilon)$ where the $\rho(x, x_{obs})$ is the chosen measure of discrepancy, and ϵ is a threshold defining the "closeness margin". If ϵ is sufficiently small then the distribution $\pi(\theta|\rho(x, x_{obs}) < \epsilon)$ will be a good approximation for the posterior distribution $\pi(\theta|x)$. It is often difficult to define an adequate distance function $\rho(x, x_{obs})$ between the simulated and observed data, so in many cases it is replaced with a distance defined by summary statistics, $\rho(S(x), S(x_{obs}))$. However, as we are considering values of concentrations in specific places at a set of time points, we are able to compare those data directly without the use of summary statistics.

In ABC methods, Sequential Monte Carlo (SMC) is used in order to automatically, sequentially "clean" the posterior distribution used to generate proposals for further steps. In $ABCSMC$ methods, the set of samples with weights, called particles, sampled from the population with the prior distribution $\pi(\theta)$, are propagated through a sequence of intermediate posterior distributions $\pi(\theta|\rho(x, x_{obs}) < \epsilon_t)$, $t = 1, ..., T$, until it represents a sample from the target distribution $\pi(\theta|\rho(x, x_{obs}) < \epsilon_T)$. These methods aim to generate draws from $p(\theta|\rho(x, x_{obs}) < \epsilon_t)$, at each of a series of sequential steps t, where ϵ_t defines a series of thresholds. One of the most important issues in $ABCSMC$ is the defining of the particle weights formula correctly. In [14] the authors propose strategies called $ABCSMC$ with Adaptive Weights ($ABCSMCAW$). This method includes a new step where the weights are modified according to the respective values of x. Algorithm 1 shows the description of $ABCSMCAW$ presented in [14].

After initialization of the threshold schedule, first N samples are simulated based on the predefined a priori distribution $\pi(\theta)$ and the corresponding acceptance condition $\rho(x, x_{obs}) < \epsilon_1$. In time step $t = 2$ simple uniform weights are changed based on additional kernel $K_{x,t}(x_{obs}|x_i^{t-1})$ proposed in [14]. Samples, denoted by a tilde are drawn from the previous generation with probabilities v_j^{t-1}. Using perturbation kernel $K_{\theta,t}(\theta_i^t|\tilde{\theta}_i)$ new "fresh" samples θ_i^t are obtained, with the veracity of the condition $\rho(x, x_{obs}) < \epsilon_t$. The weights are calculated according to the formula in step (11); in step (12) the weights are normalized and the time step is increased - $t = t + 1$. The procedure is repeated until $t \leq T$. In the section 4.1 the details are discussed, along with the motivation for choosing specific components of the Algorithm 1 for the problem of stochastic event reconstruction. More information and also theoretical aspects can be found in [14].

4.1 Data and distance measure

In the problem of stochastic event reconstruction all observed data can be split into two types of information: 1) concentration data from the sensor network, and 2) background information. The background information consists of all of the data included in the dispersion model e.g. strength and direction of the wind,

Algorithm 0.1 ABC SMC AW

1. Initialize threshold schedule $\epsilon_1 > \epsilon_2 > ... > \epsilon_T$
2. Set $t = 1$

for $i = 1$ to N **do**
- 3. Simulate $\theta_i^t \sim \pi(\theta)$ and $x \sim \pi(x|\theta_i^t)$
- 4. Until $\rho(x, x_{obs}) < \epsilon_t$
- 5. Set $w_i^t = \frac{1}{N}$

end for

for $t = 2$ to T **do**
- 6. Compute new weights $v_i^{t-1} \propto w_i^{t-1} K_{x,t}(x_{obs}|x_i^{t-1})$ for $i = 1, ..., N$
- 7. Normalize weights v_i^{t-1} for $i = 1, ..., N$
- **for** $i = 1$ to N **do**
  - 8. Pick $\tilde{\theta}_i$ from the set $\{\theta_j^{t-1}\}_{1 \le j \le N}$ with probabilities $\{v_j^{t-1}\}_{1 \le j \le N}$
  - 9. Draw $\theta_i^t \sim K_{\theta,t}(\theta_i^t|\tilde{\theta}_i)$ and $x \sim \pi(x|\theta_i^t)$
  - 10. Until $\rho(x, x_{obs}) < \epsilon_t$
  - 11. Compute new weights as $w_i^t \propto \frac{\pi(\theta_i^t)}{\sum_j v_j^{t-1} K_{\theta,t}(\theta_i^t|\theta_j^{(t-1)})}$
  - 12. Normalize weights w_i^t for $i = 1, ..., N$
- **end for**

end for

temperature and so on. To compute the $\rho(x, x_{obs})$ value we use only data from the sensor network which measures gas concentration $\hat{C}_i^{Sj}$ where i corresponds to the time step and Sj is the sensor identifier. In this test case we have 18 sensors $(S1, S2, ..., S18)$, whose positions are given in fig. 1 and fig. 2a as yellow dots. We assume that the substance concentrations registered by the sensors arrive subsequently at time intervals, hereafter referred to as 'time steps'. It is important to know that for time step t only data $\hat{C}_1^{Sj}\hat{C}_2^{Sj}\dots\hat{C}_t^{Sj}$ are available and finally we have ten time steps $(t = 10)$. The reconstruction algorithm starts to search a source location (x, y) and release rate (q) just after the first 6 minutes $(t = 2)$. To get the predicted concentration a QUIC-PLUME forward model is running and it refers to the procedure $x \sim \pi(x|\theta_i^t)$ in Algorithm 1. To run a dispersion model and obtain data x we use source parameter vector θ_i^t and the information obtained from the QUIC-URB subsystem. The simulated data also have a form of concentration value C_i^{Sj} where Sj corresponds to the known locations of j sensor.

The choice of distance measure or summary statistics is a crucial step in ABC. Since distance measures are not sufficient in many cases, this choice involves a trade-off between loss of information and reduction of dimensionality. In those cases we chose to normalize approximation error between all the data obtained to the current time step t which is also called Fractional Bias (FB) [24] . The FB is used to indicate a bias towards underprediction or overprediction of concentration data by the model. Due to the data type for all sensors in time step t the $\rho(x^t, x_{obs}^t)$ measure is as follows:

$$\rho(x^t, x^t_{obs}) = \frac{1}{18}\sum_{j=1}^{18}(\frac{1}{t}\sum_{i=1}^{t}\frac{|C_i^{Sj} - \hat{C}_i^{Sj}|}{C_i^{Sj} + \hat{C}_i^{Sj}}), \tag{1}$$

under additional definition, that $\frac{|C_i^{Sj} - \hat{C}_i^{Sj}|}{C_i^{Sj} + \hat{C}_i^{Sj}} = 0$ when $C_i^{Sj} = 0$ and $\hat{C}_i^{Sj} = 0$. Given that the concentration $C_i^{Sj} \geq 0$, the value of $\rho(x^t, x^t_{obs})$ is always between 0 and 1 . Let us notice that $\rho(x^t, x^t_{obs}) = 0$ is the situation when our prediction is perfect. In the opposite case, when $\rho(x^t, x^t_{obs}) = 1$ the prediction is wrong. In finding source parameters one of the most important areas is the detection time window, when there is a measurement in the current sensor. The measure (1) supports this approach, because when we have non-zero concentration in some time steps but our model shows that there should be 0 concentration value, the penalty value for this step will be 1. The situation is the same, if the observed value is equal to 0 and the model shows a positive value of the concentration. On the other hand, if $C_i^{Sj} > 0$ and $\hat{C}_i^{Sj} > 0$ then the absolute difference also has an impact on the value of $\rho(x^t, x^t_{obs})$ measure. Finally, the contributions of all time steps are averaged for one sensor. Because $\rho(x^t, x^t_{obs}) \in < 0, 1 >$ one sensor cannot corrupt the overall $\rho(x^t, x^t_{obs})$ value. Also, each sensor has an equal contribution to the $\rho(x^t, x^t_{obs})$ measure, regardless of the level of concentration, which is of course smaller in sensors located further from the source.

4.2 Threshold schedule and weights

The most commonly used adaptive scheme for threshold choice is based on the quantile of the empirical distribution of the distances between the simulated data and observations from the previous population, (see [8], [13]). The method determines ϵ_t at the beginning of the t time-iteration by sorting the measure $\rho(x_i^{t-1}, x_{obs}^{t-1})_{1<i\leq N}$ and setting ϵ_t such that α_t percent of the simulated data $\rho(x_i^{t-1}, x_{obs}^{t-1})_{1<i\leq N}$ are below it, for some predetermined α_t. In [12] the authors show a new strategy based on an acceptance rate curve but also discuss a cumulative number of simulation versus different threshold schedules. In this, and many other cases, quantile-based methods seem to be an easy and appropriate solution of estimating ϵ_t. Based on our own preprocessing experience we set quantile $\alpha_2 = 0.7$ in the second time step, that subsequently decreases to $\alpha_{10} = 0.3$ for $t = 10$ [12]. The additional kernel $K_{x,t}(x_{obs}|x_i^{t-1})$, which is used in calculating the weights, depends on observed and simulated data. Since weights are normalized in step (7), in Algorithm 1 we can simply use the $\rho(x^t, x^t_{obs})$ measure as the proposed kernel. Due to the restriction $0 \leq \rho(x^t, x^t_{obs}) \leq 1$ we can define $K_{x,t}(x_{obs}|x_i^{t-1}) \equiv 1 - \rho(x_i^{t-1}, x_{obs})$, because the greater weight should correspond to a better solution.

4.3 Transition kernel

We chose transition kernel $K_{\theta,t}(\cdot|\cdot)$ to be a Gaussian kernel. Unfortunately in this type of inverse problems the parameters are often highly correlated and

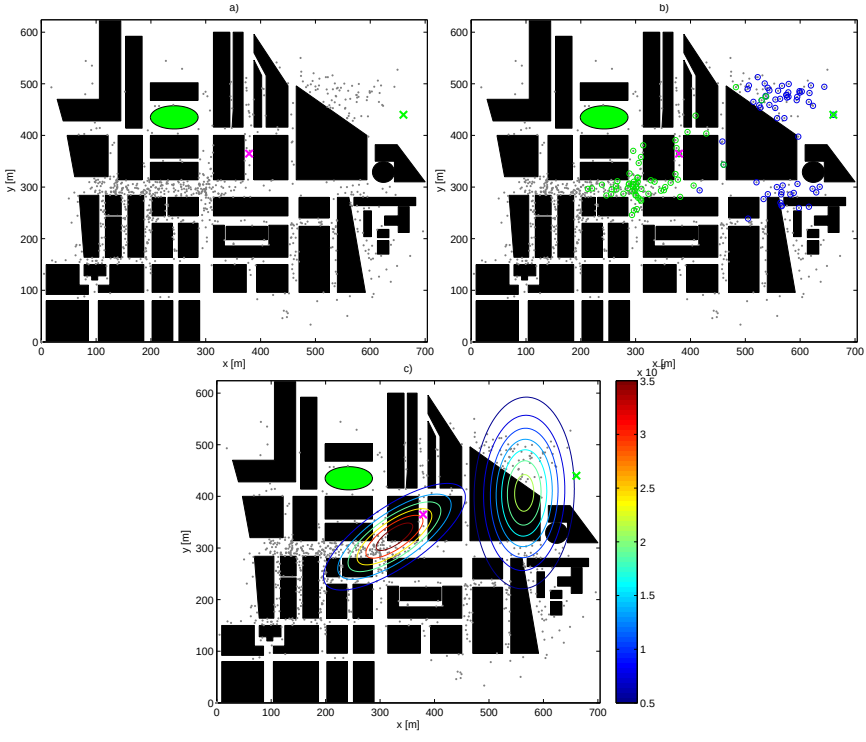

Fig. 3: *a*) all samples - grey point, selected samples - green and magenta crosses; *b*) M-nearest samples to selected green point are marked by blue circles, for magenta green circles; *c*) multivariate normal perturbation kernels evaluated from a set of $M - neighbors$ samples for two selected points.

multimodality is very common. Especially when the (x, y) domain contains a lot of prohibited regions, like buildings. Samples may tend to split in a disjointed group by filling out different street canyons. In such cases it is interesting to consider the use of a local mean and covariance matrix. Instead of computing the covariance matrix based on all the samples from $(t-1)$, a better idea is to use only limited information about the local correlation. In [11] one of the proposed methods is to use the multivariate normal kernel based on the M neighbours. Application of that procedure is presented below:

After the procedure of drawing a new sample from local multivariate normal perturbation kernel, if a new sample is accepted, new weights are also computed using this empirical kernel. The authors in [11] pay attention to the disadvantages of choosing this perturbation kernel. First, the parameter M typically has to be fixed before any of the information about the posterior are known (too small a value of M may lead to a lack of exploration of parameter space, while too large

Algorithm 0.2 ABC SMC AW Step 8

8. Pick $\tilde{\theta}_i$ from the set $\{\theta_j^{t-1}\}_{1\leq j\leq N}$ with the probabilities $\{v_j^{t-1}\}_{1\leq j\leq N}$.
8.1 Select M-nearest samples to $\tilde{\theta}_i$ from the set $\{\theta_j^{t-1}\}$ by using Nearest Neighbors Algorithm.
8.2 Compute the empirical covariance $\sum_{\tilde{\theta}_i,M}^t$ and mean $\bar{\theta}$ from the M nearest neighbours samples of $\tilde{\theta}_i$.
8.4 Set local perturbation kernel $K_{\theta,t}(\theta_i^t|\tilde{\theta}_i) as N(\bar{\theta}, \sum_{\tilde{\theta}_i,M}^t)$.

would offer little or no advantage compared to the standard multivariate normal kernel). In our case the number of samples allocated to one time step is $N = 1000$ samples for each time step. Based on pre-processed experiments we determined the number of neighbors $M = 70$. This kind of procedure may seem to be computationally expensive. However, in experiments the $M - NearestNeighbors$ multivariate normal perturbation kernel minimizes the number of samples needed to be generated, which in the case of stochastic event reconstruction problems is highly preferred. Furthermore, the computation time of running the forward model is much longer than the start-up procedure for finding the nearest neighbors and computing covariance estimation. It is worth mentioning that the choice of the correct determination of the $NearestNeighborsAlgorithms$ is important and depends on the problem.

In the experiment presented in this publication we use classical *M-Nearest Neighbors* algorithm with Mahalanobis distance due to the differences between the various dimensions of the parameters. Results of an experiment using this procedure are presented in fig. 3. This experiment refers to the source location (x, y) but the samples are three-dimensional vectors. We can see that the set of possible solutions is spread among the buildings. Sub-optimal solutions are related to two cases, where the first involves possible sources located in the center of the domain, as contrasted with the north-east location. In fig. 3 *a*) the selected sample is illustrated by a green and magenta cross surrounded by all the samples - i.e. grey points. In fig. 3 *b*) the M nearest samples are marked by blue circles relative to green points and green circles relative to magenta samples. Finally, in c) the subplot shows empirical multivariate normal kernel evaluated from the set of $M - neighbors$ samples for two sets of samples. The shapes of kernel correspond to the correlation between x and y parameters and also support only a single candidate solution. It is worth noting that the locations inside buildings are permitted although the launch dispersion model for these sites is impossible. Consequently, if the drawn sample in step 3) $\theta_i^t \sim \pi(\theta)$ and step 9) $\theta_i^t \sim K_{\theta,t}(\theta_i^t|\tilde{\theta}_i)$ in Algorithm 1 does not satisfy the assumptions then there is a re-drawing of the θ_i^t sample. The next section presents the results for the stochastic parameters reconstruction for the setup described above and the experimental data presented in section 2.

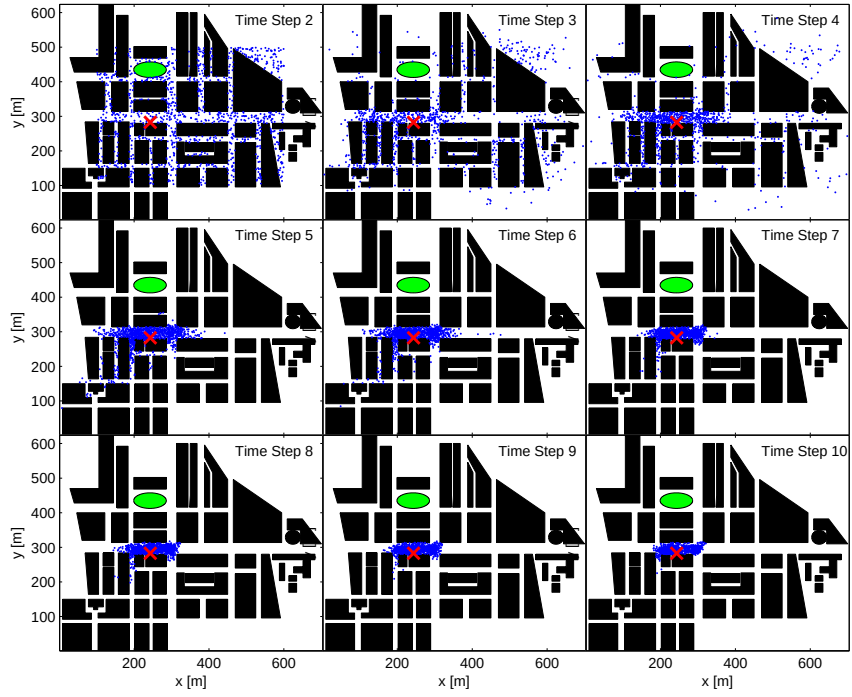

Fig. 4: A scatter plot of all samples generated in the subsequent time steps $t = 2, 3, ..., 10$ in (x, y) space of source location. The red cross marks the true source position

5 Results of DAPPLE reconstruction experiment

Fig. 4 shows the locations of the buildings in the DAPPLE London area, together with all the samples generated in subsequent time steps $t = 2, 3, .., 10$ which are decomposed directly to $6, 9, .., 30$ experimental minutes. As we can see, samples after the $4th$ time step converge from all possible (x, y) space to the vicinity of the actual source location. Using these samples, we construct the marginal probability distributions for the source location and release rate, as shown in fig. 5 for all time intervals. As time goes on, the mass of probability distribution is concentrating in the vicinity of the proper values of x and y. This looks quite different for emissions amounts, where posterior distribution for the parameter q looks like a bimodal distribution. This is better shown in figure fig. 6 where all samples are included.

After limiting the (x, y) domain to the area surrounding the real source, we can see that the distribution is divided into two areas, which suggests two different solutions of the problem. One location is closer to the main intersection

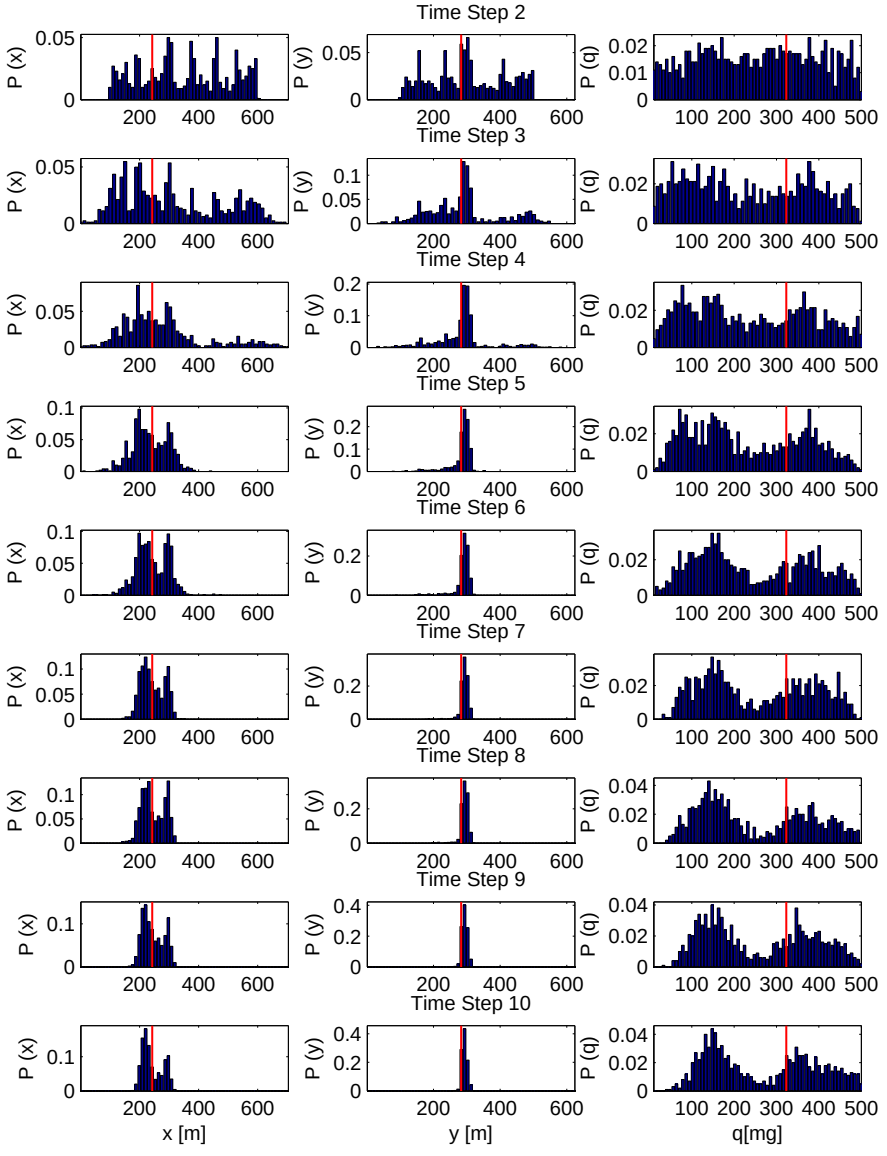

Fig. 5: Evolution of the marginal posteriori probability distribution for x, y and q parameters for time steps $t = 2, 3, .., 10$. The red vertical line represents target value of parameters.

and the second is around the true source. These results can be also noticed in the marginal distributions for x where we note two picks of probabilities. The results from all time steps are summarized in a so-called trellis plot presented in fig. 6, where the parameters reconstruction was started after 6 minutes. A color pattern reflected in fig. 6 was used to show empirical 2D probability distribution of all parameters combinations. The colored contour lines are enveloping higher probability at the joint posterior distributions. The diagonal plots are marginal empirical posterior distributions of the forward model parameters. The real parameter values from the field experiment are highlighted with vertical red lines in diagonal plots and black cross markers on the other subplot, which are successfully captured by the high posterior probability region. The correct position obtained after the transformation of the relative domain is $x = 243m$, $y = 282m$ and $q = 323mg$, where the most probable parameters values are $P(x = 223.0 \pm 7.6m) = 0.0632$, $P(y = 291.4 \pm 6.7m) = 0.1990$ and $P(q = 144.9 \pm 5.3mg) = 0.0218$. To accurately analyze the results for release rate parameter in fig. 7 a) we split the samples into two groups supported by two separate probability masses. After this assumption, two different groups of samples are presented in fig. 7 b). One can see that the green samples corresponding to $q < 250mg$ are distributed closer to the center, while the blue points are closer to the true source (red cross) and the corresponding estimates of $q = 323mg$ group closer to the real value (see fig. 7 a) red vertical line). Fig. 7 c) shows two histograms of weights $1 - \rho(x^{10}, x^{10}_{obs})$ for the green and blue points. As we can see, more points from the blue subset have higher weights (better model fit). As it means that the points have higher probability to be drawn in the next step, we can conclude that with the extension of the reconstruction procedure the "green" solutions should be slowly converging to the other (blue solution) which is close to the true value of the source parameters.

6 Conclusion

A stochastic event reconstruction method for atmospheric contaminant dispersion in an urban environment has been presented. The method described in section 4 is based on Bayesian inference with the Approximate Bayesian Computation (ABC) tool with an extension. Fast-running QUIC-PLUME dispersion models have been adopted as the forward model in the Bayesian framework. The dispersion model has been uniquely enhanced by taking into account empirical wind turbulence between buildings obtained from the QUIC-URB tool. Additional attention was given to the formulation of the distance function to take into account concentration measurements provided in successive time steps that can be available from a sensor network. The event reconstruction method has been successfully validated against the real DAPPLE experiment. In particular, the modeling of a priori distribution based on the threshold schedule substantially improved the results. Also the transition kernel set treated as a local empirical distribution, conformable to the non-standard domain, had an impact on convergence. In the event reconstruction of the DAPPLE tracer experiment, up to three

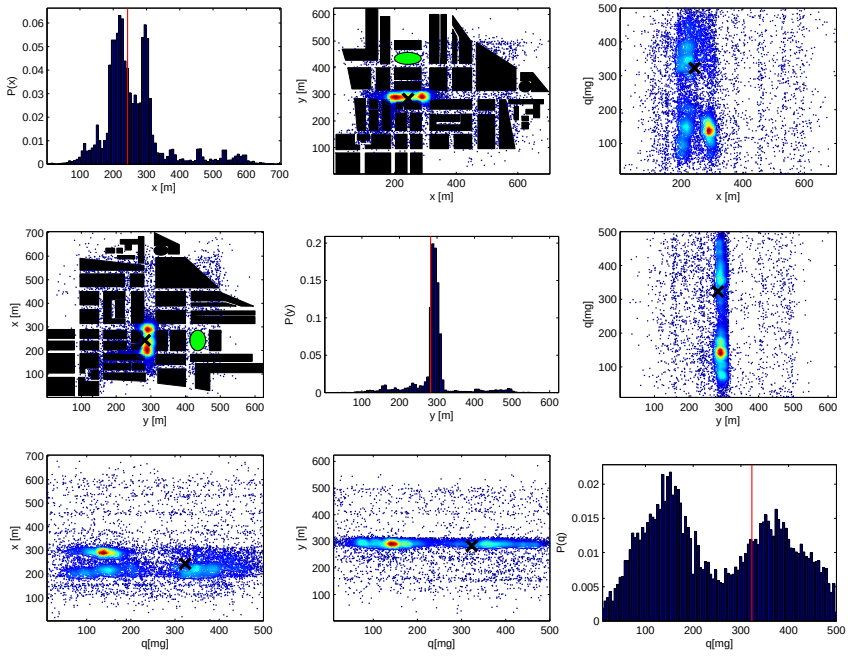

Fig. 6: Bivariate and marginal posterior distributions for all parameters $\theta \equiv (x, y, q)$. The plot is colored according to probability density, where the most probable regions are colored the deepest red (i.e., a heatmap). The vertical red lines in diagonal plots (black cross in bivariate) show the real value of each parameter. The distributions are built based on all the samples generated in the reconstruction procedure.

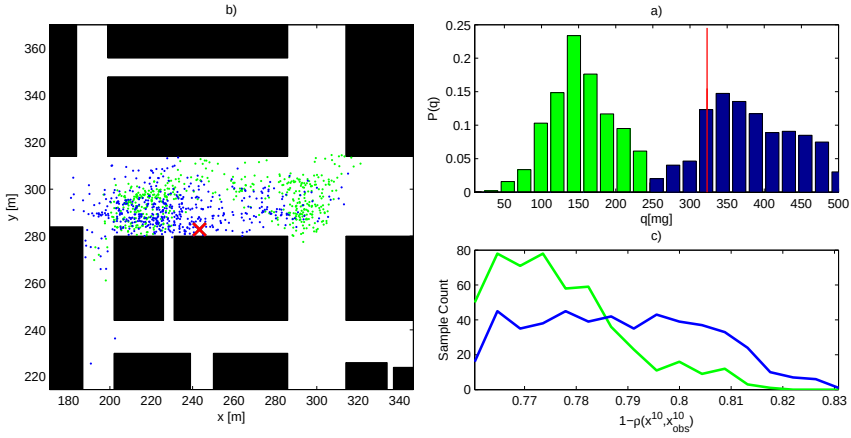

Fig. 7: *a*) Marginal posteriori distribution of q split into two sample sets *b*) Scatter plot of all samples in the (x, y) space - the sample colors correspond to the sample sets in a) *c*) The histogram of weights, which was obtained from the two groups of samples - green and blue.

parameters were estimated. From a practical point of view, release location and emission rates are of the greatest significance to the emergency responders. The present study has shown that the event reconstruction problem can be solved for the urban area without using the time-consuming Computational Fluid Mechanic model. Posterior probability distributions of model parameters were also used to build priori distribution when new concentration data became available. Although the ABC framework is general, a comprehensive operational event reconstruction tool needs to address various release scenarios. The present study focused on steady point source releases in a highly urbanized area. However, possible release scenarios may include moving sources. Furthermore, the scale of the event may range from local sites to areas of greater size. Future work will concentrate on adding new possible hazardous scenarios to the present stochastic event reconstruction tool, not necessarily the release of gases into the atmosphere.

7 Acknowledgments

The study is cofounded by the European Union from resources of the European Social Fund. Project PO KL "Information technologies: Research and their interdisciplinary applications", Agreement UDA-POKL.04.01.01-00-051/10-00.

References

1. Pudykiewicz, J.A.: Application of adjoint tracer transport equations for evaluating source parameters. Atmospheric environment **32**(17) (1998) 3039–3050

2. Johannesson, G., Hanley, B., Nitao, J.: Dynamic Bayesian models via Monte Carlo an introduction with examples. Lawrence Livermore National Laboratory, UCRL-TR-207173 (2004)
3. Johannesson, G., Dyer, K., Hanley, W., Kosovic, B., Larsen, S., Loosmore, G., Lundquist, J., Mirin, A.: Sequential Monte-Carlo based framework for dynamic data-driven event reconstruction for atmospheric release. In: Proc. of the Joint Statistical Meeting, Minneapolis, MN, American Statistical Association and Cosponsors. (2005) 73–80
4. Keats, A., Yee, E., Lien, F.S.: Bayesian inference for source determination with applications to a complex urban environment. Atmospheric environment **41**(3) (2007) 465–479
5. Chow, F.K., Kosovic, B., Chan, S.: Source inversion for contaminant plume dispersion in urban environments using building-resolving simulations. Journal of applied meteorology and climatology **47**(6) (2008) 1553–1572
6. Sisson, S.A., Fan, Y., Tanaka, M.M.: Sequential monte carlo without likelihoods. Proceedings of the National Academy of Sciences **104**(6) (2007) 1760–1765
7. Toni, T., Welch, D., Strelkowa, N., Ipsen, A., Stumpf, M.P.: Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. Journal of the Royal Society Interface **6**(31) (2009) 187–202
8. Beaumont, M.A., Zhang, W., Balding, D.J.: Approximate Bayesian computation in population genetics. Genetics **162**(4) (2002) 2025–2035
9. Turner, B.M., Van Zandt, T.: A tutorial on approximate Bayesian computation. Journal of Mathematical Psychology **56**(2) (2012) 69–85
10. Lenormand, M., Jabot, F., Deffuant, G.: Adaptive approximate Bayesian computation for complex models. Computational Statistics **28**(6) (2013) 2777–2796
11. Filippi, S., Barnes, C.P., Cornebise, J., Stumpf, M.P.: On optimality of kernels for approximate Bayesian computation using sequential Monte Carlo. Statistical applications in genetics and molecular biology **12**(1) (2013) 87–107
12. Silk, D., Filippi, S., Stumpf, M.P.: Optimizing threshold-schedules for approximate Bayesian computation sequential Monte Carlo samplers: applications to molecular systems. arXiv preprint arXiv:1210.3296 (2012)
13. Del Moral, P., Doucet, A., Jasra, A.: An adaptive sequential Monte Carlo method for approximate Bayesian computation. Statistics and Computing **22**(5) (2012) 1009–1020
14. Bonassi, F.V., West, M., et al.: Sequential Monte Carlo with Adaptive Weights for Approximate Bayesian Computation. Bayesian Analysis **10**(1) (2015) 171–187
15. Borysiewicz, M., Wawrzynczak, A., Kopka, P.: Stochastic algorithm for estimation of the model's unknown parameters via Bayesian inference. In: Computer Science and Information Systems (FedCSIS), 2012 Federated Conference on, IEEE (2012) 501–508
16. Borysiewicz, M., Wawrzynczak, A., Kopka, P.: Bayesian-based methods for the estimation of the unknown model's parameters in the case of the localization of the atmospheric contamination source. Foundations of Computing and Decision Sciences **37**(4) (2012) 253–270
17. Wawrzynczak, A., Kopka, P., Borysiewicz, M.: Sequential Monte Carlo in Bayesian Assessment of Contaminant Source Localization Based on the Sensors Concentration Measurements. In: Parallel Processing and Applied Mathematics. Springer (2014) 407–417
18. Wood, C.R., Barlow, J.F., Belcher, S.E., Dobre, A., Arnold, S.J., Balogun, A.A., Lingard, J.J., Smalley, R.J., Tate, J.E., Tomlin, A.S., et al.: Dispersion experi-

ments in central London: the 2007 DAPPLE project. Bulletin of the American Meteorological Society **90**(7) (2009) 955–969
19. Pardyjak, E.R., Brown, M.: QUIC-URB v1. 1 Theory and User's Guide. Los Alamos National Laboratory, Los Alamos, NM (2003)
20. Williams, M.D., Brown, M.J., Singh, B., Boswell, D.: QUIC-PLUME theory guide. Los Alamos National Laboratory (2004)
21. Röckle, R.: Bestimmung der Strömungsverhältnisse im Bereich komplexer Bebauungsstrukturen. na (1990)
22. Sherman, C.A.: A mass-consistent model for wind fields over complex terrain. Journal of applied meteorology **17**(3) (1978) 312–319
23. Williams, M.D., Brown, M., Boswell, D., Singh, B., Pardyjak, E.: Testing of the QUIC-PLUME model with wind-tunnel measurements for a high-rise building. In: 5th AMS Urban Env Conf, Vancouver, BC, Canada LA-UR-04-4296. (2004)
24. Cox, W.M.: Protocol for determining the best performing model. Technical report, Environmental Protection Agency, Research Triangle Park, NC (United States). Technical Support Div. (1992)