Building and Environment 70 (2013) 232–244

Contents lists available at ScienceDirect

Building and Environment

journal homepage: www.elsevier.com/locate/buildenv

A Gaussian process emulator approach for rapid contaminant characterization with an integrated multizone-CFD model☆

Piyush M. Tagade, Byeong-Min Jeong, Han-Lim Choi*

Division of Aerospace Engineering, KAIST, Daejeon 305-701, Republic of Korea

ARTICLE INFO

Article history:
Received 24 May 2013
Received in revised form
24 July 2013
Accepted 13 August 2013

Keywords:
Bayesian framework
Gaussian process emulator
Multizone models
Integrated Multizone-CFD
CONTAM
Rapid source localization and characterization

ABSTRACT

This paper explores a Gaussian process emulator based approach for rapid Bayesian inference of contaminant source location and characteristics in an indoor environment. In the pre-event detection stage, the proposed approach represents transient contaminant fate and transport as a random function with multivariate Gaussian process prior. Hyper-parameters of the Gaussian process prior are inferred using a set of contaminant fate and transport simulation runs obtained at predefined source locations and characteristics. This paper uses an integrated multizone-CFD model to simulate contaminant fate and transport. Mean of the Gaussian process, conditional on the inferred hyper-parameters, is used as a computationally efficient statistical emulator of the multizone-CFD simulator. In the post event-detection stage, the Bayesian framework is used to infer the source location and characteristics using the contaminant concentration data obtained through a sensor network. The Gaussian process emulator of the contaminant fate and transport is used for Markov Chain Monte Carlo sampling to efficiently explore the posterior distribution of source location and characteristics. Efficacy of the proposed method is demonstrated for a hypothetical contaminant release through multiple sources in a single storey seven room building. The method is found to infer location and characteristics of the multiple sources accurately. The posterior distribution obtained using the proposed method is found to agree closely with the posterior distribution obtained by directly coupling the multizone-CFD simulator with the Markov Chain Monte Carlo sampling.

1. Introduction

Safety systems in the modern building environments use sensors that monitor atmospheric parameters and alert in the eventuality of an accident. With the present day increased threat of use of chemical and biological warfare by terrorist organizations, such a scenario has become a real danger. Importance of adverse effects of the atmospheric and indoor contaminants on the overall building lifecycle and health of the occupants is already identified by the researchers [1]. Current sensor systems for buildings are designed to monitor critical parameters like temperature, humidity, carbon dioxide etc [2]. The sensor system can also detect accidental/deliberate release of hazardous contaminant, and also suggest an appropriate evacuation plan to ensure safety of occupants [3]. Since prolonged exposure of the occupants to the hazardous contaminants may result in serious health conditions including death [4], rapid source localization by the sensor system is essential. Considering that majority of individuals are expected to spend upto 90% of time in an indoor environment, it is imperative to design a sensor system that can detect, characterize and rapidly locate the accidental or deliberate contaminant release. The system is expected to aid in detection of airborne contaminant, real-time interpretation of the information to characterize and localize the contaminant source, computationally efficient prediction of contaminant dispersion with associated uncertainty quantification, and subsequent evacuation decisions based on the predictions.

1.1. Background

The sensor system often uses contaminant fate and transport models to predict the contaminant dispersion that can aid in source localization and characterization. Multizone, zonal and computational fluid dynamics (CFD) models are used for simulation of indoor airflow and contaminant dispersion patterns [2,5,6]. Owing to

☆ This work was supported in part by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2010-0025484), and in part by the KI Project through KAIST Institute for Design of Complex Systems.

* Corresponding author. Tel.: +82 423503727.
E-mail addresses: piyush.tagade@kaist.ac.kr (P.M. Tagade), bmjeong@lics.kaist.ac.kr (B.-M. Jeong), hanlimc@kaist.ac.kr (H.-L. Choi).

0360-1323/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.buildenv.2013.08.023

ease of implementation and computational efficiency, multizone models are most widely used for predicting the contaminant dispersion and source localization/characterization [4,7–10]. A multizone model represents any building as a network of well-mixed zones connected by flow paths like doors, windows, leaks etc. The airflow and contaminant transport between the zones is calculated using adjustment of zone pressures that balances mass flow through the paths [7,11,12]. The outdoor environment is modeled as an additional unbounded zone. Although used widely, limitations of the multizone models, especially related to the well-mixed assumption, are extensively reported in the literature [7,13,14]. Zonal models represent intermediate fidelity between multizone and CFD models, wherein large well-mixed zones are further divided into smaller subzones [15]. Zonal models use conservation of mass, conservation of energy and pressure gradients to model airflow and contaminant dispersion [16]. Computational fluid dynamics (CFD) models numerically solves governing equations of fluid flow and contaminant dispersions [4]. The CFD models provide detailed airflow and contaminant distribution inside a room [17]. Although most accurate amongst three, computational cost requirement prohibits use of CFD models for rapid source localization and characterization [15]. There are recent research efforts to integrate CFD with multizone models [18–21] (termed hereafter as mutlizone-CFD model). The multizone-CFD modeling approach models one of the zones using CFD, while the resultant solution is coupled with other well-mixed zones using appropriate boundary conditions. This paper uses the integrated multizone-CFD model for rapid source localization and characterization.

1.2. Motivation

Traditional deterministic approaches for sensor data fusion and interpretation, like optimization [22], Kalman filtering [23] and backward methods [24], are found inappropriate by the researchers in the context of rapid contaminant source localization and characterization [25]. Owing to the ability to provide the event probability distribution, and associated ease in the uncertainty analysis post event detection, current state of the art for source localization and characterization mainly focusses on probabilistic methods [26]. Liu and Zhai [27] have explored adjoint probability method for rapid contaminant source localization. The method derives adjoint equations for backward probability calculations using the multizone contaminant fate and transport model. Efficacy of the method is demonstrated for contaminant release in a multi-room residential house and a complex institutional building. Lin and Wang [28] have explored Ensemble Kalman Filter for data assimilation to forecast indoor contaminant transport. The ensemble Kalman filter assimilates the contaminant transport predictions obtained using the multizone model with the sensor observations from different rooms. However, the work does not consider the source localization and characterization, instead, focusses on the contaminant transport prediction conditional on the uncertain source location and characteristics.

Main aim of the present research work is to develop a Markov Chain Monte Carlo (MCMC)-based Bayesian framework that can aid the sensor system to rapidly localize and characterize the contaminant source in case of the event detection. Main advantage of the Bayesian inference method is that it can admit prior information and estimates complete probability distributions of the uncertain parameters, as against point estimates provided by optimization based methods. Sohn et al. [25] have proposed a computationally efficient Bayes Monte Carlo method for real-time data interpretation and rapid source localization. The method is divided in two stages. In first stage, a large database of simulation runs for all the possible scenario is collected that sufficiently represent uncertainty. In the second stage, Bayesian updating of the probability for each collected data is obtained after the event detection. See Sreedharan et al. [7–10] for details of recent applications of the Bayes Monte Carlo method.

Though computationally efficient, the Bayes Monte Carlo method essentially is an approximate formulation of the Bayesian inference which cannot exploit full capabilities of the Bayesian framework, including ability to handle arbitrary priors and uncertainty in the simulation model. Rather, if a large number of simulation runs are possible in real time, the MCMC-based Bayesian inference is preferred over the Bayes Monte Carlo method [7]. However, currently there is no reported exposition of the MCMC-based Bayesian inference for rapid source localization and characterization in the open literature.

Implementation of the MCMC based Bayesian framework for sensor systems is challenging due to: 1) necessity of rapid real-time inference to ensure successful evacuation with minimum losses; 2) transient nature of the underlying phenomenon; and 3) requirement of large number of MCMC samples (often in the range of 10^3–10^6) for acceptable accuracy. The problem is further exacerbated by the often large scale nature of the phenomenon being monitored. Note that items 2) and 3) necessitate large number of dynamic simulator runs, which contradicts with item 1), rendering the MCMC based Bayesian framework intractable for the sensor systems. This paper proposes computationally efficient Gaussian process emulator (GPE) [29] based approach for rapid real-time inference in view of dynamic simulators.

1.3. Proposed method

Considering the improved fidelity of the multizone-CFD model over the multizone model, coupled with the accuracy of the MCMC-based Bayesian inference over the Bayes Monte Carlo method, the MCMC-based Bayesian inference using the multizone-CFD model is expected to provide more accurate source localization and characterization as compared to the multizone model based Bayes Monte Carlo method. However, despite of the significant computational advantage over the CFD implementation, the multizone-CFD model remains computationally prohibitive for MCMC-based rapid source localization and characterization. This paper proposes a Gaussian process emulator (GPE) based Bayesian framework that can use multizone-CFD model in the context of rapid source localization and characterization. The proposed approach follows Bayesian inference method of Kennedy and O'Hagan [30], where computer simulator is calibrated using limited number of experimental observations and simulation runs (see also Higdon et al. [31], Goldstein and Rougier [32]). The proposed approach treats computer output as a random function [33], with the associated probability distribution modeled through a Gaussian process prior. The Gaussian process prior for representation of uncertain simulator outputs is extensively explored in the literature [34,35], with associated hyper-parameters predicted using the maximum likelihood estimates [36] or Bayesian inference [37]. Conditional on the hyper-parameters and a set of simulator outputs obtained at different input settings, mean of the Gaussian process acts as a computationally efficient statistical emulator of the simulator. See O'Hagan [38] for detailed tutorial on building the GPE for a simulator, while Kennedy et al. [29] may be referred for discussion on some of the case studies. However, these approaches concern statistical emulation of single-output static simulators. Conti and O'Hagan [39] have extended the GPE method for statistical emulation of dynamic simulators.

This paper adapts the GPE for dynamic simulators proposed by Conti and O'Hagan [39] to the multizone-CFD model. The resultant emulator is used in the Bayesian framework, wherein

computational efficiency of the emulator over the simulator is used for rapid source localization and characterization. The proposed method first uses dynamic simulator output data to derive the GPE, which is then used in the Bayesian framework to infer source location and characteristics using the experimental observations.

The method proposed in this paper advances the current state of the art as follows: a) the method provide MCMC-based Bayesian inference using multizone-CFD model, whereas earlier methods reported in the literature are limited to Bayes Monte Carlo approaches using the multizone models; b) Gaussian process emulator based approach is proposed for efficient Bayesian inference; c) the method provide ability to consider model structural uncertainty, which is not treated in the earlier expositions.

Rest of the paper is organized as follows: detailed problem formulation is presented in section 2. Section 3 provides details of the emulator for dynamic system simulators. In section 4, the proposed Bayesian framework for rapid source localization and characterization is discussed in detail. In section 5, efficacy of the proposed method is demonstrated for a synthetic test case of a hazardous contaminant release in a single storey seven room building. The paper is summarized and concluded in section 6.

2. Problem formulation

This paper concerns a sudden accidental/deliberate release of contaminant in a building that may cause serious health hazards, including death, to the occupants if exposed over a prolonged period of time. Although released locally, the contaminant diffuses rapidly through flow paths like doors, windows and leakages, affecting occupants throughout the building. The building is often equipped with sensors that can detect and measure the amount of contaminant present in a room. The sensor data is collected over a period of time, which is then used to decide the evacuation strategy and the containment plan, including appropriate air-handling unit actions and source extinguishing strategies. However, success of the control and evacuation strategy depends on the knowledge of source location and characteristics, which is inferred using the Bayesian framework. Typically, the source is characterized by specifying the time of activation, S_t, and the amount released, S_a. Present paper demonstrates the proposed method for possibly multiple number of sources, S_N, while each source is localized by specifying the zone in which the sources are active, Z, and xy-coordinate of each source in the zone, (x_i,y_i). Note that the Bayesian framework relies on ability to accurately predict the contaminant fate and transport for a given source location and characteristics.

2.1. Integrated multizone-CFD model

Multizone model represents a building using a network of well-mixed zones, each zone often representing a room or compartment connecting to rest of the building through flow paths. The model accounts for influences of the internal air flows, which are generated by pressure differences between the zones. The multizone model uses internal air flows, coupled with the atmospheric and outdoor wind conditions, to predict contaminant dispersion inside a building. Wang et al. [19–21] have coupled a multizone model CONTAM [40] with a zero-turbulence CFD model. The program define one of the zone as a CFD-zone, where full CFD analysis is used, while the resultant air and contaminant properties are linked with other zones to embed the CFD-zone with CONTAM. Further, an external coupling is provided to link information on outdoor air pressure and contaminant concentration to indoor building. This subsection briefly describes the integrated multizone-CFD model.

The multizone model estimates the airflow and the contaminant dispersion between the zones i and j, through the flow path ij, using the pressure drop across the path ΔP_{ij}. The model uses a power-law function to calculate the airflow rate, F_{ij}, through the flow path ij as [19]

$$F_{ij} = c_{ij}\left(\frac{P_i - P_j}{|P_i - P_j|}\right)|\Delta P_{ij}|^{n_{ij}}, \tag{1}$$

where c_{ij} is flow coefficient, n_{ij} is flow exponent while P_i and P_j are total pressures in zone i and j respectively. For each zone j, the multizone model evaluates steady state air mass balance using

$$\sum_i c_{ij}\left(\frac{P_i - P_j}{|P_i - P_j|}\right)|\Delta P_{ij}|^{n_{ij}} + F_j = 0, \tag{2}$$

where F_j is the air mass source in the zone j. Contaminant steady state mass balance for a species α is similarly obtained by

$$\sum_i F_{ij}C_\alpha + S_j = 0, \tag{3}$$

where S_j is the contaminant source in the zone j, while C_α is a contaminant concentration defined such that

$$C_\alpha = \begin{cases} C_{\alpha_i}, & \text{if airflow is from zone } i \text{ to } j \\ C_{\alpha_j}, & \text{if airflow is from zone } j \text{ to } i \end{cases}; \tag{4}$$

C_{α_i} and C_{α_j} are the contaminant concentrations in zone i and j respectively. The resultant transient contaminant transport is solved using either forward or backward Euler method with fixed time step. Lorenzetti et al. [41] have investigated an efficient variable time step solver for contaminant transport prediction using CONTAM.

The CFD model solves a set of partial differential governing equations for conservation of mass, momentum and energy inside the CFD zone. The governing equations for steady state flow are given by

$$\nabla(\rho \boldsymbol{V} u) - \Gamma_u \nabla^2 u = S_u, \tag{5}$$

where u is a variable of conservation equations, ρ is density, $\boldsymbol{V}$ is velocity vector, Γ_u is diffusion coefficient, and S_u is source. At each time step, CFD model solves steady state conservation equation (5).

Let the CFD zone, c, be connected to a zone, i, using a flow-path ic. For each grid point of the discretized flow path ic, CFD model calculates mass flow rate normal to the cell p, f_p, by

$$f_p = c_{L,p}(P_i + d_{ic} - P_p), \tag{6}$$

where $c_{L,p}$ is a linear flow coefficient, P_i is pressure in zone i, d_{ic} is a pressure difference between zones i and c, while P_p is pressure at a grid point p. Thus, the total mass flow through flow path ic predicted by the CFD model is given by

$$F_{ic}^C = \sum_{p=1}^{n_g} f_p, \tag{7}$$

where n_g is total number of grid points for the flow path. The multizone model predicts the total mass flow through flow path ic as

$$F_{ic}^M = c_{L,ic}(P_i + d_{ic} - P_{d,ic}), \tag{8}$$

where $P_{d,ic}$ is the average downwind total pressure for path ic. Thus, the coupling between CFD and multizone models is obtained by ensuring

$$\sum_k \left(\left|F_{ic}^M - F_{ic}^C\right|\right)_k \leq \varepsilon \tag{9}$$

for all connecting flow paths k, where ϵ is a convergence criterion. Using the total mass flow, contaminant concentration in each zone is estimated using Eq. (3).

In the present paper, the coupled multizone-CFD model available with CONTAM [19–21,40,42] is used to simulate the contaminant fate and transport. The room containing active contaminant sources is always defined as a CFD-zone, while other rooms are simulated using multizone model. Transient contaminant concentration in each zone is output of the multizone-CFD model. To motivate the choice of multizone-CFD model over the multizone model, it is imperative to investigate the difference between transient contaminant concentration predictions, as shown in Fig. 1. From the figure, significant difference between predictions can be observed, which may result in erroneous localization and characterization of contaminant sources. The main motivation for the present research work is to develop a Bayesian inference method that can use the more accurate multizone-CFD model for rapid localization of contaminant source in an indoor environment.

2.2. Bayesian framework

This subsection presents reformulation of the rapid source localization and characterization problem in the Bayesian inference terminology. For notational convenience and brevity, the formulation is presented for a single contaminant species, however, the method can be extended without any change for multiple species. Let q_2, represent the multizone-CFD model, where $\boldsymbol{x}\in$ is a set of deterministic inputs, λ_j is a set of uncertain parameters, while $y_j(t) = \{C_j(t)\}$ is a contaminant concentration in the j^{th} zone at time t. For the multizone-CFD model, $\boldsymbol{x}$ typically consists of building description including rooms and flow path specifications, air-handling unit, atmospheric and wind conditions, etc., while, the uncertain parameters are $\boldsymbol{\theta} = [S_N, Z, S_a, S_t, \{(x_i, y_i); i = 1,..,S_N\}]$. For further notational convenience, define $\boldsymbol{y}_j = \{y_j(t); t\in \mathscr{R}^+\}$ as a function representing the transient contaminant concentration, such that

$$\boldsymbol{y}_j = T(\boldsymbol{x}, \boldsymbol{\theta}). \tag{10}$$

Note that Eq. (10) represents a simulator with function as output, thus, explicit dependence on t is removed.

Let $\widehat{\boldsymbol{\theta}}$ be the set of 'true' but unknown source location and characteristics, that need to be inferred for future decisions, including control and evacuation strategies. To account for possible deficiencies of the multizone-CFD model, the simulator output is assumed to deviate from the 'true' system response even on specification of $\widehat{\boldsymbol{\theta}}$. In the present paper, this deviation is modeled as [30]

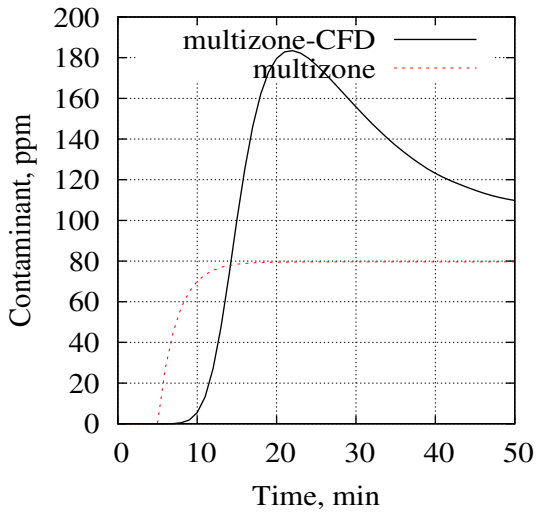


Fig. 1. Comparison of prediction by multizone and multizone-CFD model.

$$\boldsymbol{\zeta}_j = T\left(\boldsymbol{x}, \widehat{\boldsymbol{\theta}}\right) + \boldsymbol{\delta}_j, \tag{11}$$

where $\boldsymbol{\zeta}_j = \{\zeta_j(t); t\in R^+\}$ is the 'true' system response, while $\boldsymbol{\delta}_j = \{\delta_j(t); t\in R^+\}$ is known as a discrepancy function.

Let the building be equipped with sensors in M zones that detect and measure the contaminant concentration. Let the sensor data be collected for N discrete time instances. The relationship between the sensor measurement and the 'true' contaminant concentration for j^{th} zone at i^{th} time instance is given by

$$y_{ej}(t_i) = \zeta_j(t_i) + \varepsilon_j(t_i), \tag{12}$$

where $\varepsilon_j(t_i)$ denotes the sensor measurement uncertainty. For notational convenience, define a set of sensor observations $\boldsymbol{Y_e} = \{y_{ej}(t_i); i = 1,\ldots,N; j = 1,\ldots,M\}$. Similarly define $\boldsymbol{\delta} = \{\delta_j(t_i); i = 1,\ldots,N; j = 1,\ldots,M\}$. Using $\boldsymbol{Y_e}$, $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\delta}$ can be inferred through the Bayes theorem as

$$p\left(\widehat{\boldsymbol{\theta}}, \boldsymbol{\delta}\middle|\boldsymbol{Y_e}\right) \propto p\left(\boldsymbol{Y_e}\middle|\widehat{\boldsymbol{\theta}}, \boldsymbol{\delta}\right) \times p\left(\widehat{\boldsymbol{\theta}}, \boldsymbol{\delta}\right), \tag{13}$$

where $p(\widehat{\boldsymbol{\theta}}, \boldsymbol{\delta})$ is the prior, $p(\boldsymbol{Y_e}|.\widehat{\boldsymbol{\theta}}, \boldsymbol{\delta})$ is the likelihood, and $p(\widehat{\boldsymbol{\theta}}, \boldsymbol{\delta}|.\boldsymbol{Y_e})$ is the posterior probability distribution.

In the present paper, $\varepsilon_j(t_i)$ is assumed to be a zero-mean normally distributed random variable with covariance function

$$\Sigma_{e_j} = \sigma_{e_j}^2 I_N, \tag{14}$$

where σ_{e_j} is the standard deviation of uncertain experimental observations, while I_N is the $N \times N$ identity matrix. The prior uncertainty in $\boldsymbol{\delta_j}$ is specified using a zero-mean Gaussian process with covariance function

$$\Sigma_{\delta_j}(t_1, t_2) = \sigma_{\delta_j}^2 \exp\left(-\lambda_j (t_1 - t_2)^2\right), \tag{15}$$

where $\sigma_{\delta_j}^2$ and λ_j are uncertain hyper-parameters. In the full Bayesian analysis, $\sigma_{\delta_j}^2$ and λ_j are also inferred using the Bayes theorem. Since the method presented in this paper concerns inference of parameters for decisions involving control/evacuation strategies, $\sigma_{\delta_j}^2$ and λ_j are assumed to be fixed.[1] Using the probability distribution of ε and marginalization of $\boldsymbol{\delta}$, the posterior probability distribution is given by

$$p\left(\widehat{\boldsymbol{\theta}}, |\boldsymbol{Y_{e_j}}, \boldsymbol{\sigma_\delta^2}, \boldsymbol{\lambda}\right) \propto |\Sigma_j|^{-\frac{1}{2}} \prod_{j=1}^{M} \exp\left(-\frac{1}{2}\boldsymbol{d_j}^T \Sigma_j^{-1} \boldsymbol{d_j}\right) \times p\left(\widehat{\boldsymbol{\theta}}\right), \tag{16}$$

where $\boldsymbol{d_j} = \{y_{e_j}(t_i) - T(\boldsymbol{x}, \widehat{\boldsymbol{\theta}}; t_i); i = 1, \ldots, N; j = 1, \ldots, M\}$, $\boldsymbol{\sigma_\delta^2} = \{\sigma_{\delta_j}^2; j = 1, \ldots, M\}$, $\boldsymbol{\lambda} = \{\lambda_j; j = 1,\ldots,M\}$ and $\Sigma_j = \Sigma_{\delta_j} + \Sigma_{e_j}$. Solution of Eq. (16) require sampling from the posterior distribution using the MCMC method. MCMC method can be implemented to sample from the probability distribution of a random vector ϕ, $p(\phi)$, using the Metropolis-Hastings algorithm as follows [43,44]:

Algorithm 1. Metropolis-Hastings Algorithm for MCMC Sampling

1. Initialize the chain at $\phi = \phi_0$
2. **for** $i = 1$ **to** total_no_samples **do**

[1] Note that full Bayesian analysis can be used a-priory to infer the hyper-parameters $\sigma_{\delta_j}^2$ and λ_j.

3. Sample a trial point ϕ^* from proposal distribution $f(\phi_*|.\phi_{i-1})$
4. Calculate acceptance probability

$$A(\phi_*,\phi_{i-1}) = \min\left\{1,\frac{p(\phi_*)f(\phi_*|\phi_{i-1})}{p(\phi_{i-1})f(\phi_{i-1}|\phi_*)}\right\} \tag{17}$$

5. Generate a uniform random variable $\mathcal{U}$
6. **if** $\mathcal{U} < A(\phi_*,\phi_{i-1})$ **then**
7. $\phi_i = \phi^*$
8. **else**
9. $\phi_i = \phi_{i-1}$
10. **end if**
11. **end for**

Note that implementation of the MCMC require solution of $T(\boldsymbol{x},\cdot)$ for each sample, rendering the Bayesian framework intractable for computationally expensive simulators. The method proposed in this paper uses Gaussian process emulator (GPE) of the simulator in the MCMC sampling for rapid real time inference. Following section provide details of building a GPE for the dynamic simulator $T(\boldsymbol{x},\cdot)$.

3. Gaussian process emulator for dynamic simulator

For a given $\boldsymbol{x}$, the simulator $T(\boldsymbol{x},\boldsymbol{\theta})$ maps a d-dimensional input $\boldsymbol{\theta}\in\boldsymbol{\Theta}\subset\mathcal{R}^d$ to a transient output $\boldsymbol{\theta}\in\boldsymbol{\Theta}\subset\mathcal{R}^d$, where $\boldsymbol{y}$ is a function of continuous time. An emulator is built using a subset of transient response $\widehat{\boldsymbol{y}}\subset\boldsymbol{y}$, which is treated as a q-variate output of the simulator, thus $\widehat{\boldsymbol{y}}\in\mathcal{R}^q$. The simulator is deterministic in a sense that repeated simulation runs at a given input setting always returns same output. However, the simulator output is considered uncertain as the simulator runs at all the possible values of $\boldsymbol{\theta}$ cannot be obtained for computationally intensive simulators. Thus, $T(\boldsymbol{x})$ is treated as a random function,[2] with a probability distribution quantified using a Gaussian process [30,34,37,38,45]. Following Conti and O'Hagan [39], uncertainty in the random function is specified using a q-dimensional Gaussian process as

$$T(\boldsymbol{x},\cdot)\sim\mathcal{N}_q(\mathbf{m}(\cdot),c(\cdot,\cdot)\Sigma), \tag{18}$$

where $\mathbf{m}(\cdot)$ is mean and $c(\cdot,\cdot)\Sigma$ is a covariance structure of the Gaussian process. Often, the mean is modeled as

$$\mathbf{m}(\cdot) = \mathscr{B}^T\mathbf{h}(\cdot), \tag{19}$$

where $\mathbf{h}(\cdot)=[h_1(\cdot),h_2(\cdot),\ldots,h_m(\cdot)]^T$ is a vector of m regression functions, while $\mathscr{B}\in\mathscr{R}^{m\times q}$ is a matrix of regression coefficients with each column given by $\boldsymbol{\beta}=[\beta_1,\beta_2,\ldots,\beta_m]^T$. Though an arbitrary regression model can be used, literature suggests a linear model suffice for majority of the applications [30], thus $\mathbf{h}=[1,\boldsymbol{\theta}]^T$ and $m=d+1$.

Covariance function of the Gaussian process is given by

$$cov(T(\mathbf{x},\boldsymbol{\theta}_1),T(\mathbf{x},\boldsymbol{\theta}_2)) = c(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)\Sigma, \tag{20}$$

where $c(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)$ is a positive-definite correlation function, while $\Sigma\in\mathcal{R}_+^{q\times q}$ is a $q\times q$ positive definite matrix. In the present work, a square exponential correlation function is used

$$c(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) = \exp\Big(-(\boldsymbol{\theta}_1-\boldsymbol{\theta}_2)^T\Lambda(\boldsymbol{\theta}_1-\boldsymbol{\theta}_2)\Big), \tag{21}$$

where Λ is a diagonal matrix with diagonal elements given by a vector of d correlation length parameters $\boldsymbol{\lambda}$. Parameters $\mathscr{B}$, Σ and $\boldsymbol{\lambda}$ are treated as uncertain hyper-parameters. Weak non-informative prior is used for $\mathscr{B}$ and Σ,

$$p(\mathscr{B},\Sigma|\boldsymbol{\lambda})\propto\left|\Sigma\right|^{-\frac{q+1}{2}}, \tag{22}$$

while prior for $\boldsymbol{\lambda}$ is left unspecified.

A set of n simulation runs at design points $\mathbf{S}=[\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\ldots,\boldsymbol{\theta}_n]\subset\boldsymbol{\Theta}$ is used to build an emulator. Let $\boldsymbol{D}\in\mathscr{R}^{n\times q}$ define a $n\times q$ matrix of simulator outputs. From the Bayesian perspective, an emulator is defined as posterior distribution of the random function $T(\boldsymbol{x})$ given a set of simulation runs $\boldsymbol{D}$. Conditional on hyper-parameters $\mathscr{B}$, Σ and $\boldsymbol{\lambda}$, probability distribution of $\boldsymbol{D}$ is given by Ref. [39]

$$p(\boldsymbol{D}|\mathscr{B},\Sigma,\boldsymbol{\lambda})\sim\mathcal{N}_{n,q}(\mathscr{H}\mathscr{B},\mathcal{A}\Sigma), \tag{23}$$

where $\mathscr{H}^T=[\mathbf{h}(\boldsymbol{\theta}_1),\mathbf{h}(\boldsymbol{\theta}_2),\ldots,\mathbf{h}(\boldsymbol{\theta}_n)]\in\mathscr{R}^{m\times n}$ and $\mathcal{A} = c(\boldsymbol{\theta}_i,\boldsymbol{\theta}_j)\in\mathcal{R}^{n\times n}$ is a correlation matrix for a design set $\mathcal{S}$. Using Eq. (23) as likelihood and prior given by Eq. (22), posterior distribution of hyper-parameters is given by

$$p(\mathscr{B},\Sigma,\boldsymbol{\lambda}|\boldsymbol{D})\propto\mathcal{N}_{n,q}(\mathscr{H}\mathscr{B},\mathcal{A}\Sigma)|\Sigma|^{-\frac{q+1}{2}}p(\boldsymbol{\lambda}). \tag{24}$$

Conditional on the posterior distribution of hyper-parameters and $\boldsymbol{D}$, the emulator is defined as [39]

$$p(T(\boldsymbol{x},\cdot)|\mathscr{B},\Sigma,\boldsymbol{\lambda},\boldsymbol{D})\sim\mathcal{N}_q(\mathbf{m}^*(\cdot),c^*(\cdot,\cdot)\Sigma), \tag{25}$$

where

$$\begin{aligned}\mathbf{m}^*(\boldsymbol{\theta}) &= \mathscr{B}^T\mathbf{h}(\boldsymbol{\theta})+(\boldsymbol{D}-\mathscr{H}\mathscr{B})^T\mathcal{A}^{-1}\mathbf{r}(\boldsymbol{\theta})\\ c^*(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) &= c(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)-\mathbf{r}^T(\boldsymbol{\theta}_1)\mathcal{A}^{-1}\mathbf{r}(\boldsymbol{\theta}_2),\end{aligned} \tag{26}$$

while $\mathbf{r}^T(\cdot)=[c(\cdot,\boldsymbol{\theta}_1),\ldots,c(\cdot,\boldsymbol{\theta}_n)]\in\mathscr{R}^n$. Equation (25) is a statistical emulator of the simulator [29,34,36,38,39], with mean and covariance (Eq. (26)) acting as interpolator and associated expected error, respectively.

Note that Eq. (26) requires sampling from posterior distribution of hyper-parameters, imposing significant computational cost. Thus, if the analytical form is available, marginalization of hyper-parameters can render implementation of the statistical emulator computationally tractable [39]. First, marginalization of $\mathscr{B}$ from Eqs. 22–25 gives

$$p(T(\boldsymbol{x},\cdot)|\Sigma,\boldsymbol{\lambda},\boldsymbol{D})\sim\mathcal{N}_q(\mathbf{m}^{**}(\cdot),c^{**}(\cdot,\cdot)\Sigma) \tag{27}$$

where

$$\begin{aligned}\mathbf{m}^{**}(\boldsymbol{\theta}) &= \widehat{\mathscr{B}}^T\mathbf{h}(\boldsymbol{\theta})+(\boldsymbol{D}-\mathscr{H}\widehat{\mathscr{B}})^T\mathcal{A}^{-1}\mathbf{r}(\boldsymbol{\theta})\\ c^{**}(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) &= c^*(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)+\left[\mathbf{h}(\boldsymbol{\theta}_1)-\mathscr{H}^T\mathcal{A}^{-1}\mathbf{r}(\boldsymbol{\theta}_1)\right]^T\\ &\quad\left(\mathscr{H}^T\mathcal{A}^{-1}\mathscr{H}\right)^{-1}\left[\mathbf{h}(\boldsymbol{\theta}_2)-\mathscr{H}^T\mathcal{A}^{-1}\mathbf{r}(\boldsymbol{\theta}_2)\right].\end{aligned} \tag{28}$$

Here, $\widehat{\mathscr{B}}$ is a generalized least square estimate of $\mathscr{B}$ given by

$$\widehat{\mathscr{B}} = \left(\mathscr{H}^T\mathcal{A}^{-1}\mathscr{H}\right)^{-1}\mathscr{H}^T\mathcal{A}^{-1}\boldsymbol{D}. \tag{29}$$

Further integrating out Σ from Eq. (27) to obtain Ref. [39]

$$p(T(\boldsymbol{x},\cdot)|\boldsymbol{\lambda},\boldsymbol{D})\sim\mathcal{T}_q\left(\mathbf{m}^{**}(\cdot),c^{**}(\cdot,\cdot)\widehat{\Sigma};n-m\right), \tag{30}$$

where $\mathcal{T}_q$ is a Student's T process, while $\widehat{\Sigma}$ is generalized least square estimator of Σ, which is given by

[2] For a function with univariate output, a random function can be considered as a sample from a stochastic process $\mathcal{F}:\mathcal{X}\mathcal{R}^{\mathcal{R}}$. A q-variate random function is a generalization $\mathcal{F}:\mathcal{X}(\mathcal{R}^q)^{\mathcal{R}}$. See Adler [33] for details.

$$\widehat{\Sigma} = \frac{\left(\boldsymbol{D} - \mathscr{H}\widehat{\mathscr{B}}\right)^T \mathcal{A}^{-1}\left(\boldsymbol{D} - \mathscr{H}\widehat{\mathscr{B}}\right)}{n-m}. \tag{31}$$

Final step to build an emulator involves marginalization of λ, however, analytical solution for the resultant integration is not available and requires use of sampling techniques. Instead, the literature suggests fixing the values of correlation length parameters using Maximum Posteriory Estimate (MPE), Eq. (30) conditional on MPE of λ being an emulator of $T(\boldsymbol{x})$.

Posterior distribution of λ is obtained by marginalization of $\mathscr{B}$ and Σ from Eq. (24), which gives

$$p(\lambda|\boldsymbol{D}) \propto |\mathcal{A}|^{-q/2}\left|\mathscr{H}^T\mathcal{A}^{-1}\mathscr{H}\right|^{-q/2}\left|\boldsymbol{D}^T\mathcal{G}\boldsymbol{D}\right|^{-(n-m)/2}, \tag{32}$$

where

$$\mathcal{G} = \mathcal{A}^{-1} - \mathcal{A}^{-1}\mathscr{H}\left(\mathscr{H}^T\mathcal{A}^{-1}\mathscr{H}\right)^{-1}\mathscr{H}^T\mathcal{A}^{-1}. \tag{33}$$

The MPE of λ is obtained by maximizing Eq. (32) with respect to λ. For the emulator, the mean $\mathbf{m}^{**}(\cdot)$ works as an interpolator providing predictions at an unsampled $\boldsymbol{\theta}$, while $c^{**}(\cdot,\cdot)$ provide estimate of uncertainty in the predictions. Thus, the Gaussian process emulator for the dynamic simulator can be implemented using the following algorithm:

Algorithm 2. Gaussian process emulator for the dynamic simulator

1. Select n and a set of design points $\mathbf{S} = [\boldsymbol{\theta}_1,\boldsymbol{\theta}_2,\ldots\boldsymbol{\theta}_n]$ using design of experiments
2. Select q temporal locations
3. **for** $i = 1$ **to** n **do**
4. Simulate $\boldsymbol{y} = T(\boldsymbol{x},\boldsymbol{\theta}_i)$
5. Define $\boldsymbol{D}$ with i^{th} row given by $\boldsymbol{D}_i = \{y(t_j); j = 1,\ldots,q\}$, where $y(t_j)$ represents the simulator output, $T(\boldsymbol{x},\boldsymbol{\theta}_i;t_j)$, at the time instance t_j.
6. **end for**
7. Estimate GLS $\widehat{\mathscr{B}}$ using Eq. (29)
8. Estimate GLS of $\widehat{\Sigma}$ using Eq. (31)
9. Estimate MPE of λ by maximizing Eq. (32) with respect to λ
10. Using the estimates of $\widehat{\mathscr{B}}$, $\widehat{\Sigma}$ and λ, the emulator is defined by

$$\mathbf{m}^{**}(\boldsymbol{\theta}) = \widehat{\mathscr{B}}^T\mathbf{h}(\boldsymbol{\theta}) + \left(\boldsymbol{D} - \mathscr{H}\widehat{\mathscr{B}}\right)^T\mathcal{A}^{-1}\mathbf{r}(\boldsymbol{\theta})$$

4. Proposed method

4.1. Gaussian process emulator for multizone-CFD simulator

In the present paper, efficacy of the proposed method is demonstrated for localization and characterization of multiple sources in a building. Since the current version of coupled multizone-CFD simulator allows only one zone as CFD-zone, the method assumes all the sources be active in a single zone. For a given number of active sources in the zone, multizone-CFD simulator provides averaged transient contaminant concentration in each zone. Thus, each transient response is indexed by number of active sources (a), the zone in which sources are active (b), and the zone in which contaminant concentration is measured (c). Separate GPEs are built for each combination of (a,b,c).

An initial design set, $\boldsymbol{S}_{ini} = \{\boldsymbol{\theta}_i; i = 1,\ldots,n_{ini}\}$, is selected using Latin hypercube sampling [46–48] and transient simulator responses are obtained for each design point. A typical response of the simulator is shown in Fig. 2. Each transient is divided into two parts; first part consists of q_1 closely spaced data points collected just after the source activation, while the second part consists of much more coarsely spaced q_2 points. A set of $n_{ini} \times q_1$ data points, $\boldsymbol{D}_1^{ini}$, is defined using transients obtained at $\boldsymbol{S}_{ini}$. Conditional on $\boldsymbol{D}_1^{ini}$, MPE estimate of λ are obtained by maximizing Eq. (32). Conditional on the MPE estimate of λ, an additional set of design points $\boldsymbol{\theta}^{new}$ is selected as

$$\arg\max_{\theta\in\Theta} c^{**}(\boldsymbol{\theta},\boldsymbol{\theta}). \tag{34}$$

The additional set of design points is generated sequentially till the maxima of $c^{**}(\cdot,\cdot)$ is below certain pre-defined value. It may be noted that during the process of selecting additional design points, λ is kept constant, while generalized least square estimates $\widehat{\mathscr{B}}$ and $\widehat{\Sigma}$ are calculated after addition of each new design point. For this enhanced design set $\boldsymbol{S}$, set of $n \times q_1$ data points, $\boldsymbol{D}_1$, and $n \times q_2$ data points, $\boldsymbol{D}_2$, are defined. Conditional on λ, generalized least square estimates $\widehat{\mathscr{B}_1}$ and $\widehat{\Sigma_1}$ are calculated using $\boldsymbol{D}_1$. Using the same value of λ, estimates of $\widehat{\mathscr{B}_2}$ and $\widehat{\Sigma_2}$ are similarly calculated using $\boldsymbol{D}_2$.[3]

4.1.1. Reconstruction of transient contaminant concentration

Let $\mathcal{T}_1 = \{t_i; i = 1,\ldots,q_1\}$ and $\mathcal{T}_2 = \{t_{q_1+i}; i = 1,\ldots,q_2\}$ be the time instances at which data sets $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ are defined, respectively. For an arbitrary $\boldsymbol{\theta}$, let $\boldsymbol{\mu}_1(\boldsymbol{\theta}) = \{m^{**}_{1,i}(\boldsymbol{\theta}); i = 1,\ldots,q_1\}$ and $\boldsymbol{\mu}_2(\boldsymbol{\theta}) = \{m^{**}_{2,i}(\boldsymbol{\theta}); i = 1,\ldots,q_2\}$ define the predicted contaminant concentration obtained using GPEs at time instances $\mathcal{T}_1$ and $\mathcal{T}_2$ respectively. Further, define a vector, $\boldsymbol{\mu}(\boldsymbol{\theta}) = \{\boldsymbol{\mu}_1(\boldsymbol{\theta}),\boldsymbol{\mu}_2(\boldsymbol{\theta})\}$, and a matrix

$$\mathcal{A} = \begin{bmatrix} \widehat{\Sigma}_1 & 0 \\ 0^T & \widehat{\Sigma}_2 \end{bmatrix}, \tag{35}$$

where $\mathbf{0}$ is a $q_1 \times q_2$ matrix of zeroes. Conditional on $\boldsymbol{\mu}(\boldsymbol{\theta})$ and $\mathcal{A}$, the contaminant concentration at any time t is given by

$$y(t;\boldsymbol{\theta}) \sim \mathcal{N}(\mu_*, \nu_*). \tag{36}$$

Using multivariate normal theory, mean and variance of the normal distribution (36) are given by

$$\begin{aligned} \mu_* &= \mathbf{r}_*(t)\mathcal{A}^{-1}\boldsymbol{\mu} \\ \nu_* &= r_{**} - \mathbf{r}_*(t)\mathcal{A}^{-1}\mathbf{r}_*^T(t) \end{aligned} \tag{37}$$

where, $\mathbf{r}_*(t) = \{cov(t,t_i); i = 1,q_1 + q_2\}$ and $r_{**} = cov(t,t)$. Equation (37) is used as an emulator to predict long term fate and transport of the contaminant. The overall procedure for building the proposed GPE is summarized in Algorithm 3.

Algorithm 3. GPE for Multizone-CFD

1. Select $\boldsymbol{S}_{ini} = \{\boldsymbol{\theta}_i; i = 1,\ldots,n_{ini}\}$ using Latin hypercube sampling
2. Run multizone-CFD for each $\boldsymbol{\theta}_i \in \boldsymbol{S}_{ini}$
3. Using transient response at q_1 time instances, create $\boldsymbol{D}_1^{ini}$
4. Estimate λ by maximizing Eq. (32) conditional on $\boldsymbol{D}_1^{ini}$
5. **while** $c^{**}(\cdot,\cdot) \geq$ tolerance **do**
6. Conditional on λ and $\boldsymbol{D}_1^{ini}$,

$$\arg\max_{\theta\in\Theta} c^{**}(\boldsymbol{\theta},\boldsymbol{\theta}). \tag{38}$$

[3] Note that the sensor system is expected to detect the contaminant soon after the source activation, thus, the densely spaced points are used for source localization. The coarse q_2 points can then be used for predicting the long term fate and transport of the contaminant.

7. $\boldsymbol{S}_{ini} = \boldsymbol{S}_{ini} \cup \boldsymbol{\theta}^{new}$
8. Create $\boldsymbol{D}_1^{ini}$ using $\boldsymbol{S}_{ini}$
9. **end while**
10. $\boldsymbol{S} = \boldsymbol{S}_{ini}$ and $\boldsymbol{D}_1 = \boldsymbol{D}_1^{ini}$
11. Create $\boldsymbol{D}_2$ using transient response at q_2 time instances for all $\boldsymbol{\theta}_i \in \boldsymbol{S}$
12. Conditional on λ and $\boldsymbol{D}_1$, calculate $\widehat{\mathscr{B}_1}$ and $\widehat{\Sigma_1}$
13. Conditional on λ and $\boldsymbol{D}_2$, calculate $\widehat{\mathscr{B}_2}$ and $\widehat{\Sigma_2}$
14. Use λ, $\boldsymbol{D}_1$, $\boldsymbol{D}_2$, $\widehat{\mathscr{B}_1}$, $\widehat{\mathscr{B}_2}$, $\widehat{\Sigma_1}$ and $\widehat{\Sigma_2}$ to predict long term transient contaminant concentration.

4.2. Rapid source localization and characterization

Consider a building with total N_z zones, with N_s maximum possible active sources in each zone. For each possible combination of $a \in N_s$, $b \in N_z$ and $c \in N_z$ the emulator $\varepsilon_{a,b,c}(\boldsymbol{x})$ is built using Algorithm 3. The proposed GPE is used in the Bayesian framework for rapid source localization and characterization in the indoor building environment.

In the present paper, the proposed method is demonstrated for maximum possible 3 sources in a zone. The prior uncertainty in number of sources, S_n, is given by

$$p(S_n) = \frac{1}{N_s}. \tag{39}$$

Location of each source is assumed to be completely unknown with prior given by uniform distribution. Thus,

$$p(x_i, y_i, Z) = p(x_i, y_i|Z) \times p(Z) = \frac{1}{A_z} \times \frac{1}{N_z}, \tag{40}$$

where A_z is area of zone Z. S_a and S_t are assumed to be completely unknown with the range of possible values as only available information. Let $S_a \in I_a$ and $S_t \in I_t$ be the ranges of S_a and S_t. Thus,[4]

$$p(S_t, S_a) = \frac{1}{I_a} \times \frac{1}{I_t}. \tag{41}$$

Let the sensors be placed in $\mathcal{O} \subset \{Z; Z = 1, \ldots, N_o\}$ zones, where N_o represents total number of sensors, while the observations are collected at time instances $T_{\mathcal{O}} = \{t_i\}$. The observations are used in the Bayesian inference given by Eq. (16), with prior defined using Eqs. 39–41, for rapid source localization and characterization. In the MCMC implementation of the Bayesian inference, the multizone-CFD simulator is replaced by an appropriate GPE emulator. Details of the implementation are provided in Algorithm 4. To ensure ergodicity, the chain is restarted after initial burn-out period.

Algorithm 4. MCMC Sampling for GPE based Bayesian Inference

Input: Sensor locations $\mathcal{O}$ and observations $\boldsymbol{Y}_e$ at time instances $T_{\mathcal{O}}$

Define: $\phi = \{\phi^i\} = \{r_s \in [0,1], r_z \in [0,1], (x_i, y_i), S_a, S_t\}$

1. Initialize the chain at $\phi = \phi_0$
2. **for** $k = 1$ **to** total no samples **do**
3. **for** $i = 1$ **to** cardinality(ϕ) **do**
4. Generate a random number $\mathcal{U} \in [-1, 1]$
5. $\phi_*^i = \phi_{k-1}^i + \mathcal{U}$
6. **end for**
7. $a = \text{int}(\phi_*^1 \times N_S + 1)$, $b = \text{int}(\phi_*^2 \times N_Z + 1)$ and $\boldsymbol{\theta} = \{\phi_*^i; i = 3, \ldots, \text{cardinality}(\phi)\}$

[4] Although demonstrated for specific priors, the proposed method is not limited for these choices and can admit arbitrary priors.

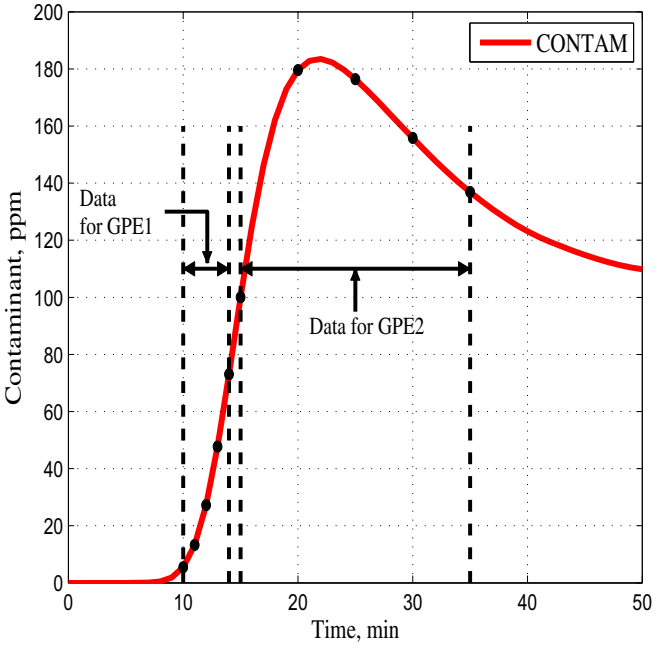


Fig. 2. Result of a CONTAM simulation run.

8. **for all** $c \in \mathcal{O}$ do
9. Predict contaminant concentration at time instances $\mathcal{T}_{\mathcal{O}}$ using emulator $\varepsilon_{a,b,c}(\boldsymbol{x}, \boldsymbol{\theta})$
10. **end for**
11. Calculate posterior probability $p(\phi_*)$ using $\boldsymbol{Y}_e$ and emulator prediction in Eq. (16)
12. Calculate acceptance probability

$$A(\phi_*, \phi_{k-1}) = \min\left\{1, \frac{p(\phi_*)}{p(\phi_{k-1})}\right\} \tag{42}$$

13. Generate a uniform random variable $\mathcal{U}$
14. **if** $\mathcal{U} \leq A(\phi_*, \phi_{k-1})$ **then**
15. $\phi_k = \phi_*$
16. **else**
17. $\phi_k = \phi_{k-1}$
18. **end if**
19. **end for**

5. Results and discussion

Efficacy of the proposed method is demonstrated for localization and characterization of a hypothetical pollutant release in a seven room building. The building plan is shown in Fig. 3.

Case study is carried out for a single storey 3 m high building with one hallway, three bedrooms, a bathroom, a kitchen and a 1 m wide open passage. Rooms are connected internally by doors, while each bedroom is connected to the outside environment by two windows each. Further, the hallway is connected to the outside environment by a main door. At the time of contaminant release, all the doors and windows are assumed to be open. Outside temperature is assumed to be 20 °C with the wind blowing at 3 m/s.

5.1. GPE for multizone-CFD

To build an emulator for the multizone-CFD simulator, an initial set of 121 design points is selected using Latin hypercube sampling

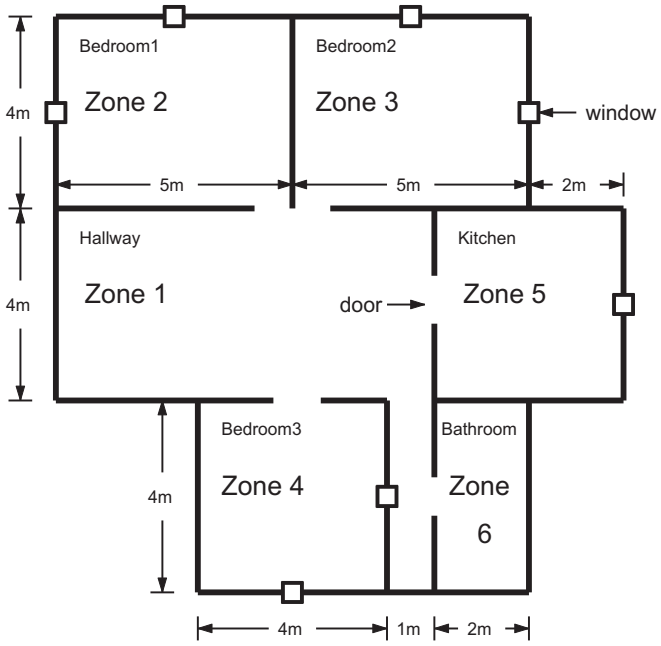


Fig. 3. Plan of the building.

[46–48]. For each design point, contaminant concentration at five temporal locations (i.e. $q_1 = 5$) in the interval of one minutes, starting from one minute after the source activation, is used as a set of initial simulator outputs $\boldsymbol{D}_1^{ini}$. Conditional on $\boldsymbol{D}_1^{ini}$, correlation length parameters $\boldsymbol{\lambda}$ are estimated by maximizing Eq. (32). In the present work, Complex Box method [49] is used for optimization. To avoid local optima, the optimizer is repeatedly run for predetermined number of times and the best point amongst the resultant optima is chosen as an estimate of $\boldsymbol{\lambda}$. The initial set of 121 design points is further augmented by sequentially selecting 29 points as described in the Algorithm 3. The resultant set of 150 design points, $\boldsymbol{S}$, is used to build the GPE. For each design point from $\boldsymbol{S}$, a second set of simulator outputs, $\boldsymbol{D}_2$, is created by using contaminant concentration values at five temporal locations ($q_2 = 5$) in the interval of four minutes, starting from $q_1 + 1$. Conditional on $\boldsymbol{\lambda}$, $\widehat{\mathscr{B}_1}$ and $\widehat{\Sigma_1}$ are estimated using $\boldsymbol{D}_1$, while $\widehat{\mathscr{B}_2}$ and $\widehat{\Sigma_2}$ are estimated using $\boldsymbol{D}_2$. Fate and transport of the contaminant for first five minutes after the source activation is reconstructed by using estimates of the emulator conditional on $\boldsymbol{D}_1$. The long term contaminant fate and transport for six minutes onwards from the source activation is reconstructed using estimates of the emulator conditional on $\boldsymbol{D}_2$ along with Eq. (37). Fig. 4 shows comparison of transient contaminant concentration obtained using the proposed method with multizone-CFD simulator.

5.2. Source localization and characterization with full sensor network

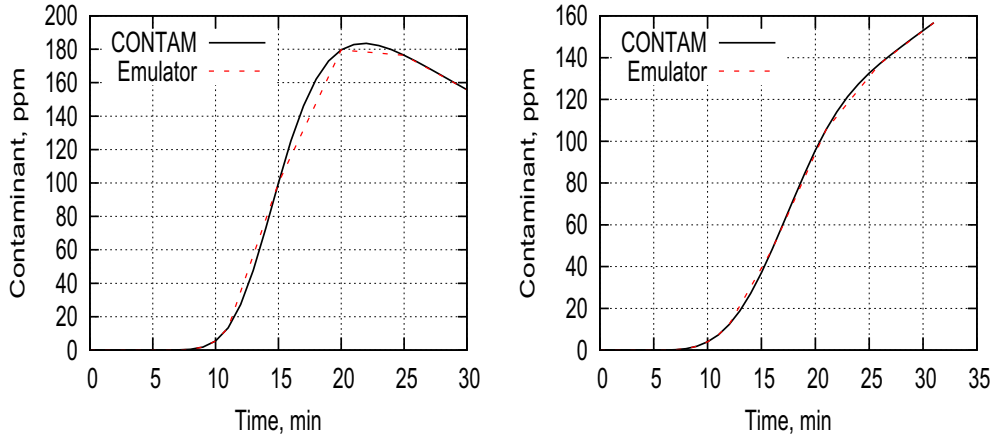
Efficacy of the proposed Bayesian framework for rapid source localization and characterization is investigated for a release of contaminant in Hallway (zone 1). For the present test case, two sources are assumed to be activated at time $T = 18$ min, with each source releasing carbon monoxide (CO) at a rate of 0.09 g/s. Inside the Hallway, source 1 is located at (4.0,1.36), while source 2 is located at (1.44,3.6). Sensors are assumed to be presented in six zones (zones 1–6, except in Passage, zone 7). All the sensors are assumed to be collaborating with each other. Sensor measurement is simulated by running the multizone-CFD with specified source characteristics and location. Transient multizone-CFD prediction in the time-step of 1 min is used as sensor observations, while experimental uncertainty in each sensor observation is assumed to be 1%. Total 5 data points per sensor (i.e., 5 min of data) are used for source localization and characterization. Note that for the present test case, all the zones are connected with the Hallway, thus the contaminant is detected by the sensors in all the zones. Bayesian inference is used after collecting sensor data for five minutes. To investigate the efficacy of the proposed method, the Bayesian inference is also implemented using the direct MCMC sampling, where the integrated multizone-CFD model is used in the Metropolis-Hastings algorithm (in Algorithm 1) to sample from the posterior distribution. Total 20,000 samples are collected after burnout period of 10,000 samples. The resultant posterior distribution is compared with the posterior distribution obtained using the proposed method. Table 1 summarizes posterior probability of source located inside a given zone and the posterior probability of number of active sources. For the present test case, the method infers zone and number of sources accurately with probability one.

The posterior probability contours of source locations obtained using the direct MCMC sampling is shown in Fig. 5(a), while, Fig. 5(b) shows the posterior probability contours obtained using the proposed method. Actual location of the sources is also indicated in the figure. From the figure, it may be concluded that the method accurately infers the source location with high probability. Further, the posterior probability contour obtained using the proposed method matches closely with the direct MCMC sampling.

Fig. 6 shows posterior probability distribution of the time of source activation and the amount of contaminant release by each source. Posterior probability distributions obtained using the proposed method and the direct MCMC sampling are shown in the figure, which are found to match closely with each other. Posterior probability distribution of the release time is a non-symmetric one-sided distribution with high probability near the time of detection and rapidly decreasing away from the detection time, which is similar to the exponential distribution. Note that this behavior is expected as the sensors detect contaminants quickly after the source activation, thus the posterior probability near the detection time is high. Further, as the contaminant accumulates over time, probability of source release at earlier time is low. Posterior probability of contaminant amount release is symmetric with high probability near 0.09, which is a true value of contaminant amount release.

5.3. Effect of varying number of sensors

In this subsection, the proposed method is implemented using the different number of sensors and sensor data points. All the test cases are presented for two active sources ($S_N = 2$) in the Hallway (zone 1), activated at $S_t = 18$ mins and releasing the carbon monoxide at the rate of $S_a = 0.09$ g/s. Sensors are assumed to collect the data in the interval of one minute. The test cases are presented using observations after one minute (1 data point), three minutes (3 data points) and five minutes (5 data points). All the observations are used concurrently for the Bayesian inference. Fig. 7(a) shows the posterior probability of sources located in zone 1 using different number of sensors. When the Bayesian inference is implemented after one minute, the correct zone is inferred with high probability using observation from one sensor, which itself is located in zone 1. As the number of sensors increases, the posterior probability of sources located in zone 1 increases, with the method inferring the correct zone with probability 1 when three or more sensors are used. However, when the proposed method is implemented using three or five minutes of data, the correct zone (zone 1) is inferred



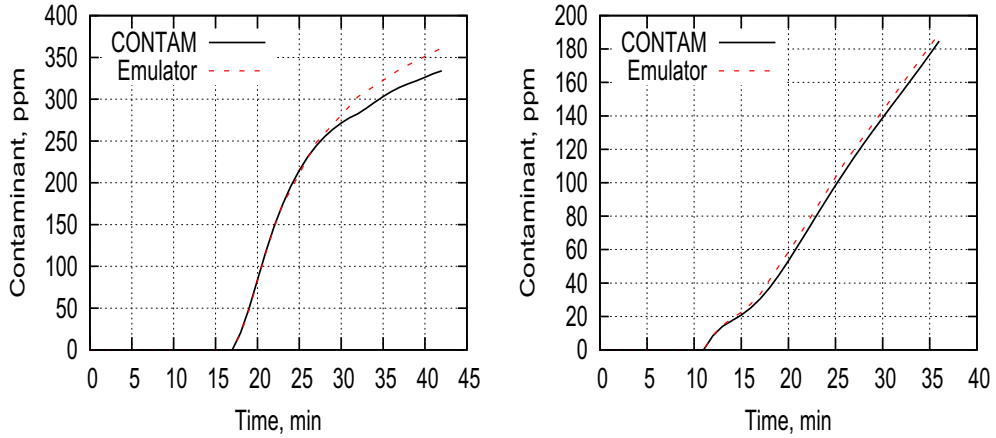


Fig. 4. Comparison of multizone-CFD and emulator predictions. Figure a) shows comparison at some of the $\theta \in \boldsymbol{S}$. Figure b) shows comparison at $\theta \notin \boldsymbol{S}$.

with low probability, while, the posterior probability is $p(Z=1)=1$ when the observations from two sensors are used. As pointed out earlier, zone 1 is connected to other zones, thus, the contaminant released in zone 1 disperses to all the other zones. Hence, when the contaminant is observed in more than one zone, the posterior probability, $p(Z=1)$, increases rapidly.

Fig. 7(b) shows the posterior probability of $S_N=2$. For a given number of sensors, the posterior probability, $p(S_N=2)$,increases with increase in the number of observations used. The inference obtained using 3 min and 5 min of data matches closely with each other. When the Bayesian inference is implemented using 1 observation (after 1 min of data), the number of active sources is inferred correctly with probability one when observations from four or more sensors is used. However, when 3 or 5 min of data is used, number of active sources is inferred with probability one using three or more sensors.

Fig. 8 shows computational time for the proposed method. The computational time for the direct MCMC implementation is also indicated in the figure. All the test cases are implemented on a desktop computer with Intel Core i5 CPU. The computational time of the implementation is obtained using a FORTRAN intrinsic routine *cpu_time*. The implementation of direct MCMC method, when 5 min of data observed by sensors in 6 zones is used, takes more than 120 h of computational time. The computational cost of the proposed method is significantly lower than the direct MCMC, demonstrating the possibility of real-time rapid source localization and characterization. The computational cost of the proposed method increases linearly with the number of sensors used, however, the computational cost does not increase noticeably with the number of observations used. Note that for each sensor, a separate emulator needs to be evaluated, however, each evaluation of the emulator provides the complete reconstruction of the transient contaminant concentration. Thus the computational cost increases with the increase in number of sensors used, whereas, increase in the computational cost is minimal for increased number of observations.

Table 1
Posterior probabilities of room & number identification.

	Posterior probability of sources in a zone, $p(Z)$						
	1	2	3	4	5	6	7
Direct MCMC	1	0	0	0	0	0	0
GPE based MCMC	1	0	0	0	0	0	0

	Posterior probability of no. of sources, $p(S_N)$		
	1	2	3
Direct MCMC	0	1	0
GPE based MCMC	0	1	0

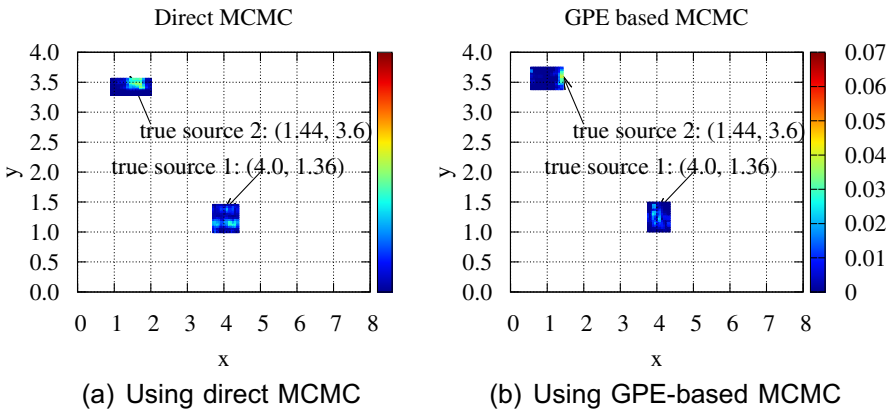


Fig. 5. Conditional posterior probability distribution of source location.

5.4. Inference with dynamic incremental sensor network

Results presented in the previous subsection demonstrate the need for a collaborative sensor network for accurate inference. Computational cost of the proposed method increases with the number of sensors used, limiting the number of sensors for rapid source localization. However, to ensure that the contaminant is detected in any zone, placement of a sensor in each zone is necessary. This subsection investigates the proposed method for a possible dynamic sensor network. The network consists of the sensors placed in six zones (only excluding the passage). In the event of contaminant detection by any of the sensor, one minute of data of the single sensor is used to infer the posterior probability of zone location and characteristics. The sensor subsequently requests next three minutes of data from the two adjacent zones with non-zero probability of the source presence, while, the resultant three minutes of observations from three sensors is used for the Bayesian inference. In the final step, five minutes of observations from all the six sensors are used. Note that posterior probability of each step is used as a prior for the next step. Implementation of the resultant sensor network is explained in Algorithm 5.

Algorithm 5. Dynamic Collaborative Sensor Network

1. Let the contaminant be detected by the sensor in zone j at time t
2. Define $\mathcal{O} = \{j\}$ and $T_{\mathcal{O}} = \{t+1\}$
3. Specify priors given by Eqs. 39–41
4. Use Algorithm 4 to sample from the posterior distribution
5. $k = 1$
6. **for** $i = 1$ **to** no of zones **do**
7. **if** $P(i) \neq 0$ **then**
8. $k = k + 1$
9. $\mathcal{O}_k = i$
10. **end if**
11. **end for**
12. **for** $i = 1$ **to** 3 **do**
13. $T_{\mathcal{O}i} = t + 1 + i$
14. **end for**
15. Use posterior distribution 4 as prior
16. Use Algorithm 4 to sample from the posterior distribution
17. Define $\mathcal{O} = \{j; j = 1, \ldots, 6\}$ and $T_{\mathcal{O}} = \{t + 4 + i; i = 1, \ldots, 5\}$
18. Use posterior distribution 19 as prior
19. Use Algorithm 4 to sample from the posterior distribution

In this subsection, efficacy of the collaborative sensor network, described in Algorithm 5, is investigated for different test cases. Fig. 9 compares the effect of number of sources active in different zones on the source localization. The results are presented for sources activated at 18 min releasing 0.09 g/s of carbon monoxide. The top row of Fig. 9 shows posterior probability of the 'true' zone,

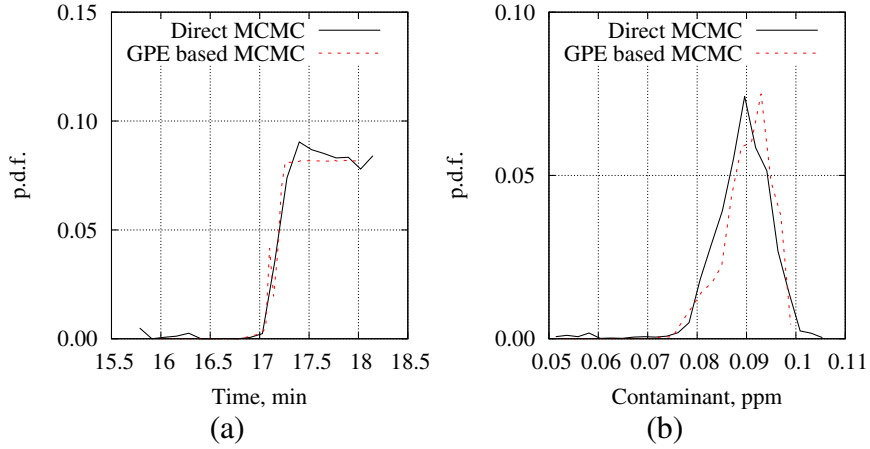


Fig. 6. (a) Posterior probability distribution of time of source activation (for the test case, source is activated at 18 min); (b) posterior probability distribution of amount of contaminant released (for the test case, 0.09 g/s CO contaminant is released).

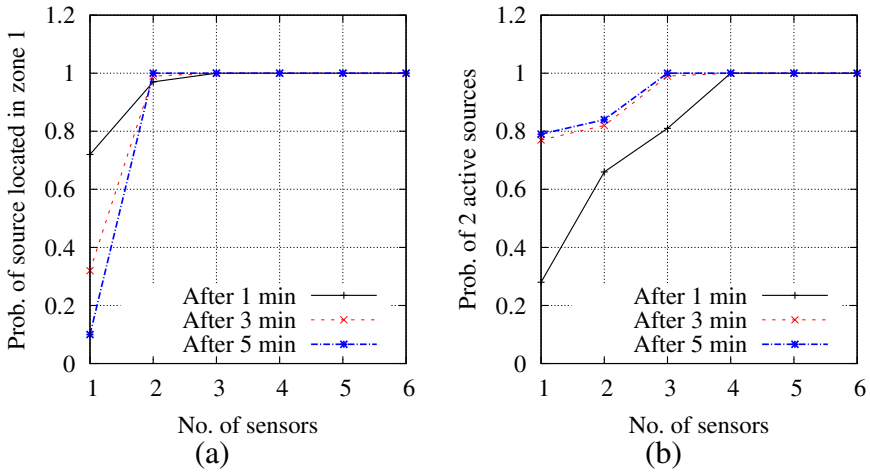


Fig. 7. Inference results with varying number of sensors: (a) posterior probability sources active in zone 1; (b) shows posterior probability of 2 active sources.

$p(Z)$, where each test case represent sources located in different zones. Left row show results for one active source, middle row show results for two active sources, while the right row show the results for three active sources. The bottom row show similar results for posterior probability of the 'true' number of sources, $p(S_N)$. As can be observed from the figure, the 'true' zone of active sources is correctly located by the collaborative sensor network after four minutes (i.e., after using observations from three sensors), for all the test cases except when the single source is active in zone 4, which is inferred correctly with probability one after nine minutes. The 'true' number of active sources is inferred after four minutes with varying posterior probability, however after nine minutes, the 'true' number of active sources is inferred with $p(S_N)$ approaching one for all the test cases. Thus, the collaborative sensor network explained in Algorithm 5 can accurately localize the source using the proposed method with four minutes of sensor observations from three collaborating sensors, while the number of active sources is also inferred accurately after using the nine minutes of data from six sensors.

Fig. 10 investigates the efficacy of the proposed method, using the collaborative sensor network (Algorithm 5), to infer the time and amount of contaminant release. The results are presented for $Z = 1$ and $S_N = 2$. Top row of Fig. 10 shows the posterior probability distribution of time of source activation S_t, with $S_a = 0.09$ g/s, while, the bottom row shows the posterior probability of S_a when $S_t = 18$ min. As the more observations are used from increasing number of sensors, the posterior probability distribution becomes narrow around the 'true' value. For all the test cases presented, time and amount of contaminant release is inferred correctly with high probability after four minutes using three collaborating sensors, while the probability of the 'true' values increases when the network of six collaborating sensors is used after nine minutes. The results presented in this subsection have demonstrated the feasibility of using the proposed method for rapid source localization and characterization for a possible dynamic incremental sensor network. The method can similarly be applied for investigating other sensor networks.

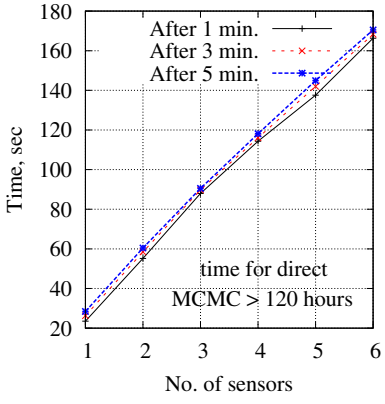


Fig. 8. Comparison of computational time for different sensor networks and data points.

6. Concluding remarks

This paper has presented a Gaussian process emulator (GPE)-based Bayesian framework for rapid contaminant source localization and characterization in the indoor environment. The framework can be used with a computationally expensive integrated multizone-CFD model. The framework approximates the multizone-CFD model using a GPE during the pre-event detection stage, which is used for Bayesian inference of the source location and characteristics after the contaminant detection by the sensors. The framework provides a methodology for rapid localization and characterization of multiple sources. In conjunction with the rapidly advancing digital and sensor technologies, the framework can be used for planning the evacuation and the source extinguishing strategies in an indoor building environment in view of sudden contaminant release. The framework can also be used to test different sensor networks and investigate the performance tradeoffs.

In the present paper, efficacy of the framework have been investigated for a hypothetical contaminant release in a single storey seven room building. The posterior distribution of the uncertain parameters obtained using the proposed method is found to match closely with the direct MCMC implementation, at a significantly lower computational cost. Performance and the robustness of the proposed method have been investigated for a dynamic incremental sensor network. Various test cases presented in the paper have demonstrated the robustness of the proposed method, although in a limited sense for one of a possible sensor network. In future, authors propose to investigate the presented approach as an inference machine for informative sensor planning.

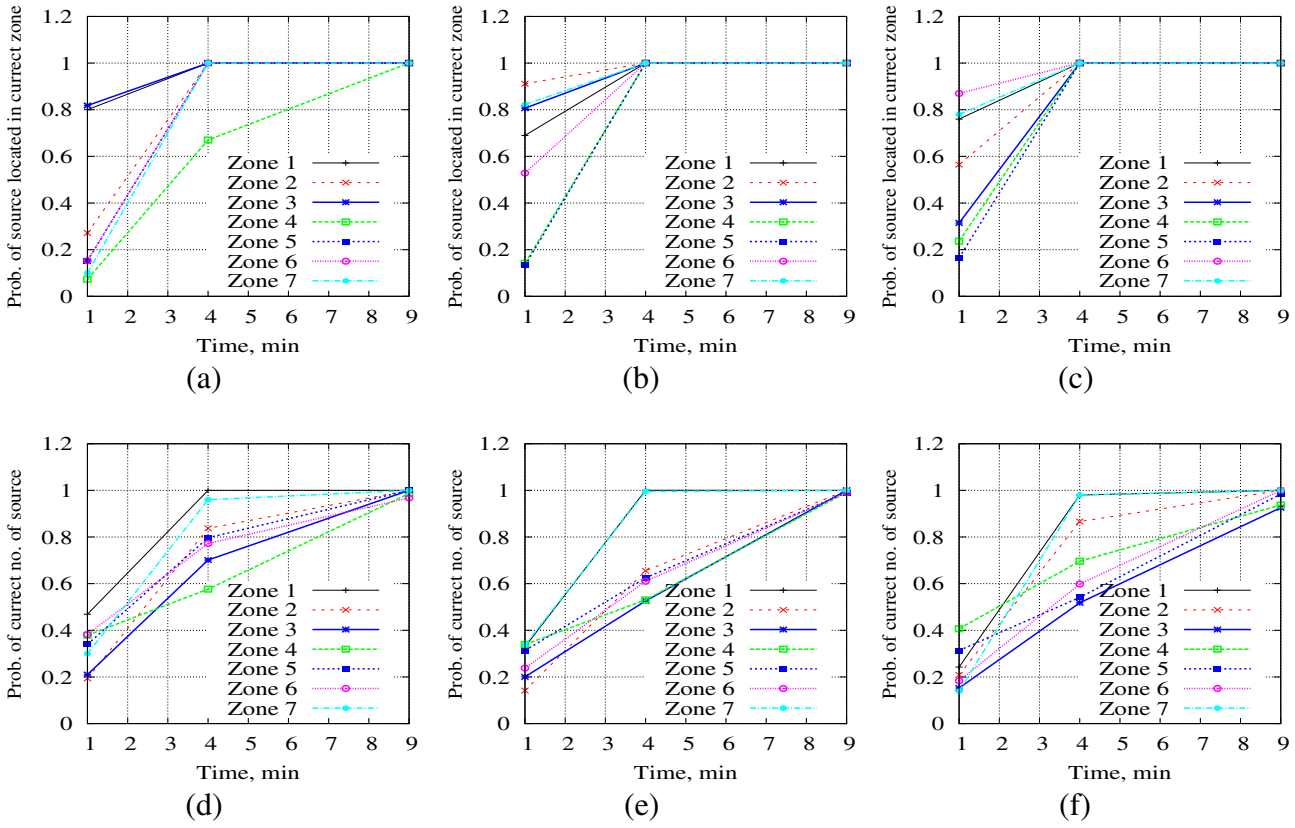


Fig. 9. Source localization using collaborative sensor network. Results are shown for sources located in different zones. Top row shows posterior probability of true zone and bottom row shows posterior probability of true number of sources. Left column shows results for one active source, middle column shows results for two active sources and right column shows results for three active sources.

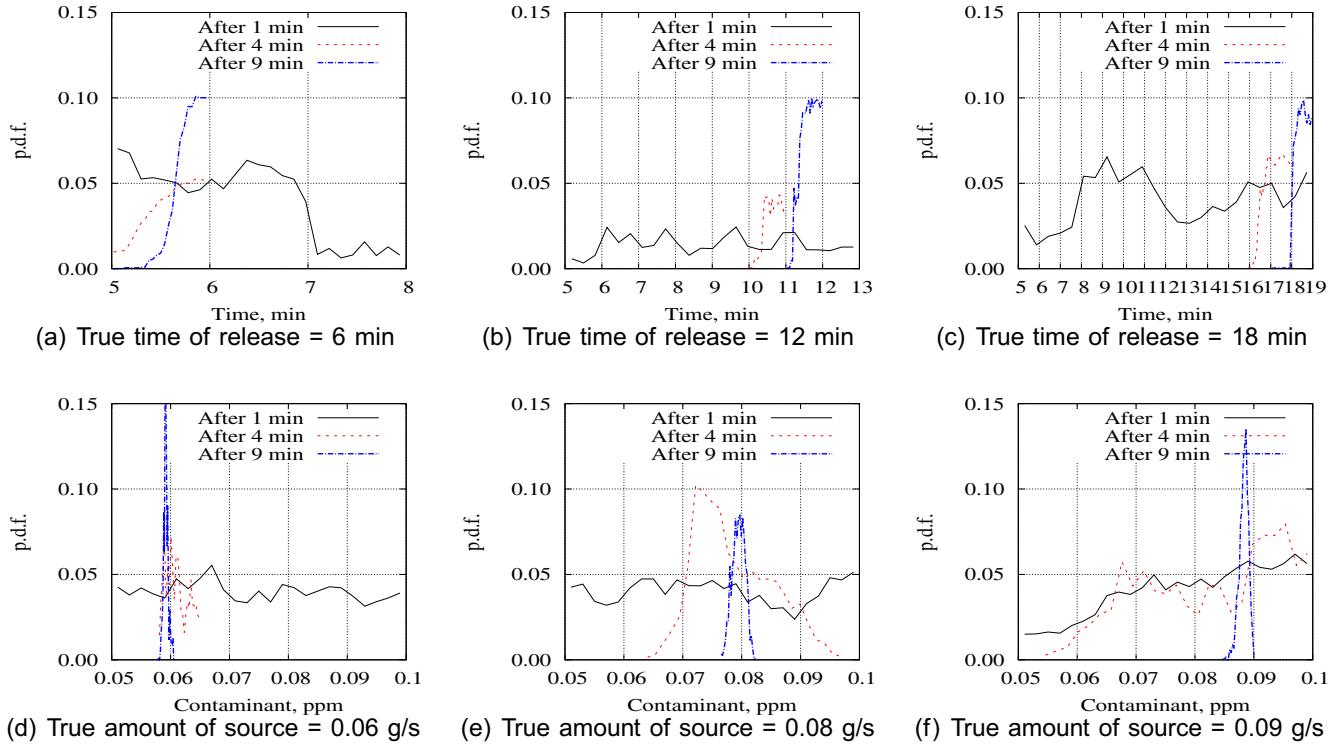


Fig. 10. Source characterization collaborative sensor network. Top row shows posterior probability of time of source activation and bottom row shows posterior probability of amount of source released.

References

[1] Collinge W, Landis A, Jones A, Schaefer L, Bilec M. Indoor environment quality in a dynamic life cycle assessment framework for whole buildings: focus on human health chemical impacts. Build Environ 2013;62:182–90.
[2] Chen Y, Wen J. Comparison of sensor systems designed using multizone, zonal and CFD data for protection of indoor environments. Build Environ 2010;45:1061–71.
[3] Eliades D, Michaelides M, Panayiotou C, Polycarpou M. Security-oriented sensor placement in intelligent buildings. Build Environ 2013;63:114–21.
[4] Zhai Z, Sebric J, Chen Q. Application of CFD to predict and control chemical and biological agent dispersion in buildings. Int J Vent 2003;3:251–64.
[5] Chen Y, Wen J. Sensor system design for building indoor air protection. Build Environ 2008;43:1278–85.
[6] Chen Y, Wen J. The selection of the most appropriate airflow model for designing indoor air sensor systems. Build Environ 2012;50:34–43.
[7] Sreedharan P. Bayesian based design of real-time sensor systems for high-risk indoor contaminants [Ph.D. thesis]. Berkeley: University of California; 2007.
[8] Sreedharan P, Sohn M, Gadgil A, Nazaroff W. Systems approach to evaluating sensor characteristics for real-time monitoring of high-risk indoor contaminant release. Atmos Environ 2006;40:3490–502.
[9] Sreedharan P, Sohn M, Gadgil A, Nazaroff W. Influence of indoor transport and mixing time scales on the performance of sensor systems for characterizing contaminant release. Atmos Environ 2007;41:9530–42.
[10] Sreedharan P, Sohn M, Nazaroff W, Gadgil A. Towards improved characterization of high-risk release using heterogeneous indoor sensor systems. Build Environ 2011;46:438–47.
[11] Axley J. Indoor air quality modeling - phase II report. NBSIR87–3661; 1987.
[12] Feustel H. COMIS - an international multizone air-flow and contaminant transport model. Energy Build 1999;30:3–18.
[13] Baughman A, Gadgil A, Nazaroff W. Mixing of a point-source pollutant by natural-convection flow within a room. Indoor Air 1994;4:114–22.
[14] Richmond-Bryant J, Eisner A, Brixey L, Wiener R. Short-term dispersion of indoor aerosols: can it be assumed the room is well mixed? Build Environ 2006;41:156–63.
[15] Mora L., Gadgil A., Wurtz E., Inard C. Comparing zonal and.
[16] Wurtz E, Nataf J, Winkelmann F. Two and three-dimensional natural and mixed convection simulation using modular zonal models in buildings. Int J Heat Mass Trans 1999;42:923–40.
[17] Nielsen P. Computational fluid dynamics and room air movement. Indoor Air 2004;14:134–43.
[18] Tan G, Glicksman L. Application of integrating multi-zone model with CFD simulation to natural ventilation prediction. Energy Build 2005;37:1049–57.
[19] Wang L. Coupling of multizone and CFD programs for building airflow and contaminant transport simulations [Ph.D. thesis]. West Lafayette: Purdue University; 2007.
[20] Wang L, Chen Q. Theoretical and numerical studies of coupling multizone and CFD models for building air distribution simulations. Indoor Air 2007;17:348–61.
[21] Wang L, Dols W, Chen Q. Using CFD capabilities of CONTAM 3.0 for simulating airflow and contaminant transport in and around building. HVAC&R Res 2010;16:749–63.
[22] Mahar P, Datta B. Optimal monitoring network and ground-water pollution source identification. J Water Resour Plan 1997;123:199–207.
[23] Federspiel C. Estimating the inputs of gas transport processes in buildings. IEEE Trans Control Systems Technol 1997;5:480–9.
[24] Atmadja J, Bagtzoglou A. State of the art report on mathematical methods for groundwater pollution source identification. Environ Foren 2001;2:205–14.
[25] Sohn M, Reynolds P, Singh N, Gadgil A. Rapidly locating and characterizing pollutant releases in buildings. J Air Waste Manag Assoc 2002;52:1422–32.
[26] Liu X, Zhai Z. Inverse modeling methods for indoor airborne pollutant tracking: literature review and fundamentals. Indoor Air 2007;17:419–38.
[27] Liu X, Zhai Z. Prompt tracking of indoor airborne contaminant source location with probability-based inverse multi-zone modeling. Build Environ 2009;44:1135–43.
[28] Lin C, Wang L. Forecasting simulations of indoor environment using data assimilation via an Ensemble Kalman Filter. Build Environ 2013;64:169–76.
[29] Kennedy M, Anderson C, Conti S, O'Hagan A. Case studies in Gaussian process modelling of computer codes. Reli Engi System Safety 2006;91:1301–9.
[30] Kennedy M, O'Hagan A. Bayesian calibration of computer models. J R Stat Soc Series B Stat Methodol 2001;63(3):425–64.
[31] Higdon D, Kennedy M, Cavendish J, Cafeo J, Ryne R. Combining field data and computer simulations for calibration and prediction. SIAM J Scienti Comput 2005;26(2):448–56.
[32] Goldstein M, Rougier J. Probabilistic formulations for transferring inferences from mathematical models to physical systems. SIAM J Scienti Comput 2005;26(2):467–87.
[33] Adler R, Taylor J. Random fields and geometry. New York: Springer; 2007.
[34] Sacks J, Welch W, Mitchell T, Wynn H. Design and analysis of computer experiments. Stat Science 1989;4:409–23.
[35] Oakley J. Eliciting Gaussian process priors for complex computer codes. The Statistician 2002;51:81–97.
[36] Welch W, Buck R, Sacks J, Wynn H, Mitchell T, Morris M. Screening, predicting, and computer experiments. Technometrics 1992;34:15–25.
[37] Currin C, Mitchell T, Morris M, Ylviskar D. Bayesian prediction of deterministic functions, with applications to the design and analysis of computer experiments. J Am Stat Assoc 1991;86:953–63.
[38] O'Hagan A. Bayesian analysis of computer code outputs: a tutorial. Reliability Engineering and System Safety 2006;91:1290–300.
[39] Conti S, O'Hagan A. Bayesian emulation of complex multi-output and dynamic computer models. J Statis Plan Infer 2010;140:640–51.
[40] Wang L., Dols W., Chen Q. An introduction to the cfd capabilities in.
[41] Lorenzetti D, Dols W, Persily A, Sohn M. A stiff, variable time step transport solver for CONTAM. Build Environ 2013;67:260–4.
[42] CONTAM: multizone airflow and contaminant transport analysis softwar-ehttp://www.bfrl.nist.gov/IAQanalysis/CONTAM/index.htm; 2013.
[43] Metropolis N, Rosenbluth A, Rosenbluth M, Teller A, Teller E. Equation of state calculations by fast computing machines. J Chem Phys 1953;21(6):1087–92.
[44] W.K. H. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 1970;57:97–109.
[45] Rasmussen C. Evaluation of Gaussian processes and other methods for non-linear regression [Ph.D. thesis]. University of Toronto; 1996.
[46] McKay M, Beckman R, Conover W. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics 1979;21:239–45.
[47] Munholland P, Borkowski J. Simple Latin square sampling +1: a spatial design using quadrats. Biometrics 1996;52(1):125–36.
[48] Helton J, Davis F. Latin hypercube sampling and propagation of uncertainty in analyses of complex systems. Sandia Report SAND2001–0417; 2001.
[49] Box M. A new method of constrained optimization and a comparison with other methods. Computer J 1965;8(1):42–52.