Ann. Occup. Hyg., Vol. 53, No. 4, pp. 409–424, 2009

doi:10.1093/annhyg/mep017

Bayesian Modeling of Exposure and Airflow Using Two-Zone Models

YUFEN ZHANG[1], SUDIPTO BANERJEE[1], RUI YANG[2], CLAUDIU LUNGU[3] and GURUMURTHY RAMACHANDRAN[4]*

[1]Division of Biostatistics, University of Minnesota, Minneapolis, MN 55455, USA; [2]Workplace Safety Insurance Board, Toronto, M5V 3J1 Ontario, Canada; [3]Department of Environmental Health Sciences, University of Alabama at Birmingham, Birmingham, AL 35294, USA; [4]Division of Environmental Health Sciences, University of Minnesota, Minneapolis, MN 55455, USA

Received 3 October 2008; in final form 26 February 2009; published online 29 April 2009

Mathematical modeling is being increasingly used as a means for assessing occupational exposures. However, predicting exposure in real settings is constrained by lack of quantitative knowledge of exposure determinants. Validation of models in occupational settings is, therefore, a challenge. Not only do the model parameters need to be known, the models also need to predict the output with some degree of accuracy. In this paper, a Bayesian statistical framework is used for estimating model parameters and exposure concentrations for a two-zone model. The model predicts concentrations in a zone near the source and far away from the source as functions of the toluene generation rate, air ventilation rate through the chamber, and the airflow between near and far fields. The framework combines prior or expert information on the physical model along with the observed data. The framework is applied to simulated data as well as data obtained from the experiments conducted in a chamber. Toluene vapors are generated from a source under different conditions of airflow direction, the presence of a mannequin, and simulated body heat of the mannequin. The Bayesian framework accounts for uncertainty in measurement as well as in the unknown rate of airflow between the near and far fields. The results show that estimates of the interzonal airflow are always close to the estimated equilibrium solutions, which implies that the method works efficiently. The predictions of near-field concentration for both the simulated and real data show nice concordance with the true values, indicating that the two-zone model assumptions agree with the reality to a large extent and the model is suitable for predicting the contaminant concentration. Comparison of the estimated model and its margin of error with the experimental data thus enables validation of the physical model assumptions. The approach illustrates how exposure models and information on model parameters together with the knowledge of uncertainty and variability in these quantities can be used to not only provide better estimates of model outputs but also model parameters.

Keywords: Bayesian statistics; exposure assessment; indoor air modeling; industrial hygiene; Markov chain Monte Carlo; two-zones modeling; worker's exposure

INTRODUCTION

A primary issue in industrial hygiene is the estimation of a worker's exposure to chemical, physical, and biological agents. Usually exposure assessment proceeds from three basic methodologies: (i) subjective estimation using professional judgment, (ii) direct measurement of the environment exposure, and (iii) prediction of exposure through mathematical modeling. Traditionally, subjective judgments made with little transparency have driven most exposure assessments with direct measurements playing a smaller role. Exposure modeling, however, has been a neglected area, with very little emphasis within industry and negligible support for research from governmental funding agencies. However, this situation is expected to change dramatically with the advent of the REACH regulations in the European Union that require assessing exposures in a variety of exposure scenarios where monitoring may not be feasible (Ramachandran, 2008).

*Author to whom correspondence should be addressed.
Tel: 612-626-5428; fax: 612-626-4837;
e-mail: ramac002@umn.edu

Statistical and mathematical modeling have some advantages over direct air monitoring data in certain situations: systematically evaluating retrospective exposure when past monitoring data are poor or nonexistent; predicting current and future exposure in the absence of the working process or operation, and in estimating exposure with only a small number of air samples with possibly high variability. Indeed, Nicas and Jayjock (2002) have argued that with only a few monitoring data points, modeling may provide more precise estimates of exposure than monitoring with only a few data points. With advances in computational methods and inexpensive software implementation, formal modeling is set to become an indispensable tool in the industrial hygienists' armory.

Formal modeling includes a deterministic component describing the physical laws that underlie the relationship between contaminant generation rate, pollutant transportation characteristics, and contaminant concentrations. The precise nature of the models varies in their scope and complexity according to the different experimental settings and certain assumptions that might be made. For example, assumptions on pollutant transportation patterns range from complete instantaneous mixing, to two well-mixed zones within a room, to diffusion resulting in continuous concentration gradients in time and space.

However, predicting exposure in real settings is constrained by lack of quantitative knowledge of exposure determinants and mathematical exposure models that are appropriate for the scenario. Even though the profession has for long paid lip service to the importance of a thorough knowledge of the determinants of exposure on the part of the hygienist, most companies do not collect such information routinely. Even basic data such as ventilation rates, pollutant generation rates, and worker time activity patterns are hard to come by in most situations.

The models employed today in industrial hygiene are typically based upon some simple assumptions about airflow and contaminant transport pattern. Hemeon (1963) and Nicas (1996) have described a two-zone model for concentrations in a 'near field' in the proximity of a source and a 'far field' that is some distance away from the source. While these models are being increasingly employed in the industrial hygiene community (e.g. Nicas, 2003; Nicas *et al.*, 2006), they have not been validated empirically in occupational settings. Little experimental research has been conducted to evaluate the parameters used in the models and assess model performance. For example, the parameter denoting the airflow rate between the near and the far field is poorly understood in the two-zone model. Existing methods for estimating this parameter (Melikov and Zhou, 1996; Nicas and Miller, 1999) ignore possible influencing factors such as presence of human body, body movement, and body temperature that could affect the average air speed inside the near-field zone. Cherrie (1999) investigated the effect of a wide range of general ventilation conditions and room sizes in a simulation study. Given the uncertainties in model parameters in most settings, validating these models becomes a challenging exercise. Not only do the model parameters need to be known, the models also need to predict the output (i.e. the air concentrations in the two zones) with some degree of accuracy.

The present article offers a Bayesian statistical framework for estimating model parameters and exposure concentrations from experimental data. This estimation will provide the margins of statistical confidence, thereby accounting for uncertainty. Finally, comparing the estimated model and its margin of error with the experimental data will enable validation of physical model assumptions. The approach illustrates how exposure models and information on model parameters together with the knowledge of uncertainty and variability in these quantities can be used to not only provide better estimates of model outputs but also model parameters.

Deterministic Equations of Two-Zones Modeling provides the physical background of the two-zone exposure problem leading to the formulation of the differential equations and their solutions. Statistical Modeling provides a description of the statistical modeling framework highlighting the different approaches we investigate. Bayesian Modeling briefly outlines the numerical algorithms for implementing and formally comparing our models. Simulation Study of Two-Zone Data present data analysis from a simulated data set as well as an actual experimental setting, while Experimental Two-Zone Study concludes the paper with some discussion and thoughts.

DETERMINISTIC EQUATIONS OF TWO-ZONES MODELING

The one-compartment model that assumes a uniform concentration at all points in the compartment irrespective of distance from the source can underestimate worker exposures, especially for workers near the contaminant source (AIHA, 2000). To account for this deficiency, a slightly more complicated two-compartment model can be used (Nicas, 1996; Cherrie, 1999; Nicas and Miller, 1999) as in Fig. 1. Conceptually, it is a small step from a one-compartment to a two-compartment model. The region very near and around the source is modeled as one well-mixed box—the so-called near field, and the rest of the room is another well-mixed box that completely encloses the near-field box. This box is called the far field, and there is some amount of air exchange between the two boxes. Figure 1 shows the two zones schematically.

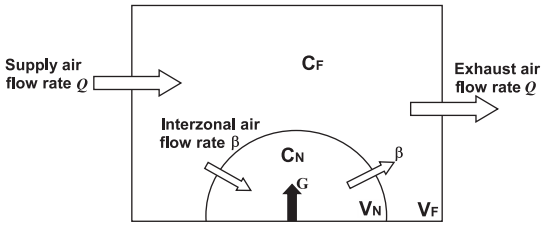


Fig. 1. The near-field and far-field zones.

The volumes of the far- and near-field zones are V_F and V_N, respectively. The supply and exhaust flow rates are the same and equal to Q (e.g. in units of $m^3\ min^{-1}$). The airflow rate between the two zones is β (in units of $m^3\ min^{-1}$). While the determination of Q is straightforward, the determination of β is not. It is dependent, to some extent, on the dimensions of the near-field zone. One approach (Nicas, 1996) is to determine it as the product of the random air speed (v) at the boundary of the near field (reported by Baldwin and Maynard, 1998) and one half of the free surface area (SA) of the near field ($\beta = \frac{1}{2}SA \times v$). The factor of $\frac{1}{2}$ arises because the flow rate of β occurs by air flowing into the near field through one half of the free surface area and air flowing out through the other half of the free surface area.

Certain quantities accounted for in the modeling are measured externally and are treated as 'known'. In this article, we will assume that the volumes of the far-field and near-field zones, V_F and V_N, respectively, are such. Here, we concentrate on a statistical estimation procedure (Bayesian) that combines information from experimental data and prior information (quantified in terms of probability distributions) for the unknown parameters and provide statistical predictions of the near- and far-field concentrations. We will primarily focus upon estimating the airflow rate between the far and near field, denoted by β. Two other parameters governing the dynamics of the system are the contaminant's total mass emission rate, denoted by G, and the rate of airflow into and out of the workplace, denoted by Q. In practical settings, G and Q may also not be known precisely. To illustrate our statistical approach, however, we will first fix G and Q and regard β as the only parameter requiring statistical estimation. We do so to keep the notations simpler and because the essential concepts of the approach do not change when additional parameters are assumed unknown. Subsequently, we will extend the analysis with unknown G and Q, hence with priors on G, Q, and β.

For the sake of simplicity, we assume that the initial concentration in both zones is equal to zero, the supply air is free of contaminant, and the only removal mechanism for the contaminant is by ventilation. Denoting concentrations in the near and far fields by $C_F(\beta;t)$ and $C_N(\beta;t)$, the dynamics of the total contaminant mass in a two-zone field is described as

$$\frac{d}{dt}\begin{bmatrix} C_N(\beta;t) \\ C_F(\beta;t) \end{bmatrix} = \begin{bmatrix} -\beta/V_N & \beta/V_N \\ \beta/V_F & -(\beta+Q)/V_F \end{bmatrix} \times \begin{bmatrix} C_N(\beta;t) \\ C_F(\beta;t) \end{bmatrix} + \begin{bmatrix} G/V_N \\ 0 \end{bmatrix}. \quad (1)$$

This is concisely expressed as $\dot{\mathbf{C}}(\beta;t) = \mathbf{A}(\beta)\mathbf{C}(\beta;t) + \mathbf{g}$, where $\mathbf{C}(\beta;t) = (C_N(\beta;t), C_F(\beta;t))^T$, where T denotes the transpose of the vector, $\dot{\mathbf{C}}$ denotes the derivative $(d/dt)\mathbf{C}(\beta;t)$, $\mathbf{A}(\beta;t)$ is the transformation matrix (as above), and $\mathbf{g} = (G/V_N, 0)$ is a translation vector independent of t. In fact, when $\mathbf{A}(\beta)$ is an invertible matrix, the solution to equation (1) can be written as

$$\mathbf{C}(\beta;t) = \exp(\mathbf{A}(\beta)t)\mathbf{C}(0) + \mathbf{A}^{-1}(\beta)\left[\exp(\mathbf{A}(\beta)t) - I\right]\mathbf{g}. \quad (2)$$

Matrix representations such as in equation (2) are especially useful in solving linear systems of differential equations (see, e.g. Laub, 2005) in computer packages such as MATLAB© and R (http://www.r-project.org/). In the following discussion, variables in bold represent matrices or vectors while the other variables are scalars.

STATISTICAL MODELING

Statistical modeling focuses upon optimal estimation of the parameter β from the experimental data. Since perfect experimental conditions are impossible, observed data may contain significant deviations from the physical two-zone model. Therefore, one seeks a nonlinear statistical equation that will account for noise in the measurements and uncertainties in model estimation and prediction. This statistically estimated model can subsequently help in validating the physical model for the experimental data.

At each time point t, let us consider the response $\mathbf{Y}(t) = (Y_N(t), Y_F(t))^T$ as a 2×1 vector corresponding to the natural logarithm of the concentration measurements from the near and far fields. The observed value of $\mathbf{Y}(t)$ will be the result of two components: the systematic component $\mathbf{C}(\beta;t)$ (also converted to the logarithm scale) that represents the physical two-zone model and a stochastic measurement error process, also known as a 'white-noise', that represents measurement error. The physical two-zone model is denoted by $\mathbf{C}(\beta;t)$ [solution to equation (2)], and the stochastic measurement error process is denoted by the 2×1 vector $\boldsymbol{\epsilon}(t) = (\epsilon_N(t), \epsilon_F(t))^T$, where $\epsilon_N(t)$ and $\epsilon_F(t)$ are the measurement error processes corresponding to the near and far field, respectively. Letting log

$\mathbf{C}(\beta;t) = (\log C_N(\beta;t), \log C_F(\beta;t))^T$, the measurement model is

$$\mathbf{Y}(t) = \log\mathbf{C}(\beta;\, t) \,+\, \boldsymbol{\epsilon}(t). \tag{3}$$

The error $\boldsymbol{\epsilon}(t)$ requires a probability distribution, which we take as normal or Gaussian. In other words, we are modeling two components in equation (3): the mean function $\log \mathbf{C}(\beta;t)$ captures the trend or large scale variation in $\mathbf{Y}(t)$, expressed as a function of time and some experimental parameters (β, G, Q). The second component, $\boldsymbol{\epsilon}(t)$, captures measurement error as well as discrepancies in the response due to imperfect physical conditions and, perhaps, misspecification of the physical model. The parameters in the distribution of $\boldsymbol{\epsilon}(t)$ help us estimate and predict the uncertainty.

A customary specification will have $\epsilon_N(t)$: $^{\text{iid}}$ $N(0, \tau_N)$ and $\epsilon_F(t)$: $^{\text{iid}}N(0, \tau_F)$, implying that the measurement errors for $\mathbf{Y}(t)$ for both the near and far fields are independently and identically distributed as normal distributions with zero means and variances denoted by τ_N and τ_F, respectively. Equation (3) implies that the concentration measurements in their true scale [i.e. $\exp(Y_N(t))$ and $\exp(Y_F(t))$] are assumed to follow a log-normal distribution. We call $\sqrt{\tau_N}$ and $\sqrt{\tau_F}$ the residual standard deviations. The geometric standard deviation (GSD) is given by $\exp\left(\sqrt{\tau_N}\right)$ and $\exp\left(\sqrt{\tau_F}\right)$. This one-one correspondence between the residual standard deviation and the GSD is appealing from a computational and inferential perspective. It is computationally convenient to have prior distributions on the residual variances. Once inference on the residual variances is obtained, we can easily obtain inference on the GSDs.

Another feature that can influence the observed values of the response is the presence of correlation between the near- and far-field measurement error processes. For example, assuming that $\epsilon_N(t)$ and $\epsilon_F(t)$ are independent of each other (and across time), we can write the error distribution as $\boldsymbol{\epsilon}(t) \overset{\text{iid}}{\sim} N\left(\mathbf{0}, \begin{bmatrix} \tau_N & 0 \\ 0 & \tau_F \end{bmatrix}\right)$. This implies that the two components of $\boldsymbol{\epsilon}(t)$ follow a bivariate normal distribution with zero mean and a 2×2 variance–covariance matrix that is simply a diagonal matrix with the two variances. The off-diagonal element that represents the covariance between the fields is 0. More generally, however, we may want to relax this assumption of independence between the fields, in which case we write $\boldsymbol{\epsilon}(t) \overset{\text{iid}}{\sim} N(\mathbf{0}, \Sigma)$, where the matrix Σ is symmetric and positive definite (i.e. all its eigenvalues are real and positive) and is no longer restricted to be diagonal. Positive definiteness of $\Sigma = \begin{bmatrix} \tau_N & \tau_{NF} \\ \tau_{NF} & \tau_F \end{bmatrix}$ is equivalent to $-1 \le \tau_{NF}/\sqrt{\tau_N\tau_F} \le 1$, where the off-diagonal element τ_{NF} represents a parameter that captures the covariance between the two fields and $\tau_{NF}/\sqrt{\tau_N\tau_F}$ represents the statistical 'correlation coefficient' between the near field and far field.

Nonlinear models, as in equation (2), often cause estimation problems since the parameters in the physical model (e.g. the parameter β) are often not easily estimable from the data. Some expert or 'prior' knowledge regarding the plausible ranges for these parameters is often required. One approach is to fit the model for a number of different fixed values of β to gather an idea about what values yield the closest fit. This approach is fairly *ad hoc* and subjective and might deliver spurious estimates that subsequently generate inaccurate predictions for concentrations.

A more principled statistical approach would quantify the plausible range of β through a prior probability distribution, allowing us to quantify the strength of this belief. This information is then combined with the likelihood using Bayes Theorem [see equation (5) in Bayesian Modeling]. Statistical models that combine such prior knowledge with experimental data are known as Bayesian hierarchical models and have become extremely popular over the last two decades (e.g. Gelman *et al.*, 2004; Carlin and Louis, 2008).

BAYESIAN MODELING

In Bayesian statistics, one constructs hierarchical (or multilevel) schemes by assigning probability distributions to parameters *a priori* and inference is based upon the distribution of the parameters conditional upon the data *a posteriori*. By modeling both the observed data and any unknown regressor or covariate effects as random variables, the hierarchical Bayesian approach to statistical analysis offers a cohesive framework for combining complex data models and external knowledge or expert opinion.

More specifically, let $\mathbf{Y} = (\mathbf{Y}(t_1)^T, \ldots, \mathbf{Y}(t_n)^T)^T$ denote the $2n \times 1$ vector of observed concentrations from the field experiment at n time points. Let us denote the collection of unknown parameters in our model by $\boldsymbol{\theta} = (\beta, \Sigma)$. Thus, there are four model parameters in our setting: β and the three distinct elements in Σ (τ_N, τ_F, and τ_{NF}). Then, equation (3) specifies the density (or likelihood) $p(\mathbf{Y} \mid \boldsymbol{\theta})$ as Gaussian for the observed data given $\boldsymbol{\theta}$, written as

$$p(\mathbf{Y} \mid \boldsymbol{\theta}) = \frac{1}{(2\pi\det(\Sigma))^{n/2}} \prod_{i=1}^{n} \exp\Big\{ -\frac{1}{2}(\mathbf{Y}(t_i) - \log\mathbf{C}(\beta;\, t_i))^T \Sigma^{-1} (\mathbf{Y}(t_i) \,-\, \log\mathbf{C}(\beta;\, t_i)) \Big\}, \tag{4}$$

which is also called the 'data likelihood'. The prior distribution for θ is specified by some probability

distribution $p(\boldsymbol{\theta})$ and the 'posterior' distribution is given by

$$p(\boldsymbol{\theta} \mid \mathbf{Y}) = \frac{p(\mathbf{Y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})}{\int p(\mathbf{Y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}}. \quad (5)$$

which incorporates both the data and the possible external (prior or expert) knowledge. In general, posterior distributions are analytically intractable due to the potentially complicated integral in the denominator and are evaluated by drawing samples from the posterior distribution. A suite of methods known as Markov chain Monte Carlo (MCMC) algorithms such as the Gibbs sampler and Metropolis–Hastings algorithms (see, e.g. Gilks *et al.*, 1996; Gelman *et al.*, 2004; Marin and Robert, 2007; Carlin and Louis, 2008) have gained enormous popularity in Bayesian statistics. The power of these methods lies in completely avoiding the numerical integration in equation (5), which, in general, can be high dimensional, by absorbing this integral into a 'proportionality constant' and generating samples from the distribution $p(\boldsymbol{\theta} \mid \mathbf{Y}) \propto p(\boldsymbol{\theta}) p(\mathbf{Y} \mid \boldsymbol{\theta})$. In Appendix B, we briefly describe the Gibbs sampler and Metropolis–Hastings algorithm in the current context; details are available in the aforementioned references. Since there is no *a priori* justification for why β and the elements of Σ should be correlated, we assume they are independent *a priori*, i.e. $p(\boldsymbol{\theta}) = p(\beta) p(\Sigma)$.

MCMC yields samples from a Markov chain that require an initial 'burn-in' time to converge to its stationary distribution, namely the posterior distribution. Customarily, convergence is diagnosed by starting a few different chains (say two or three) from different starting values and noting where they start mixing (e.g. Gelman *et al.*, 2004; Chapter 11). On the basis of these plots, we discard some initial samples as burn-in and retain the rest to draw inference. Furthermore, samples obtained from MCMC are not independent, but autocorrelated. This effect is usually small for models with relatively fewer parameters and drawing a large enough posterior sample ensures consistency in estimation. MCMC algorithms have now been automated in free software projects such as WinBUGS (http://www.mrc-bsu.cam.ac.uk/) and several packages within the R statistical environment (http://www.r-project.org/). These programs not only execute sampling from the posterior distribution but also provide tools for gauging convergence and suggesting effective Monte Carlo sample sizes for posterior inference.

An important advantage of carrying out inference using posterior samples is that they immediately yield inference for functions of the parameters we have sampled. For example, once we have obtained the posterior samples of the elements of Σ from (B), we can immediately obtain the posterior samples of the GSDs by exponentiating each sampled value of the square root of the diagonal elements of Σ. More precisely, suppose we have obtained L posterior sample units $\left\{\Sigma_{(i)}\right\}_{i=1}^{L}$. For each sample $\Sigma_{(i)}$, we extract its diagonal elements $\tau_{\mathrm{N}(i)}$ and $\tau_{\mathrm{F}(i)}$ and compute $\exp\left(\sqrt{\tau_{\mathrm{N}(i)}}\right)$ and $\exp\left(\sqrt{\tau_{\mathrm{F}(i)}}\right)$. The samples $\left\{\exp(\sqrt{\tau_{\mathrm{N}(i)}})\right\}_{i=1}^{L}$ and $\left\{\exp\left(\sqrt{\tau_{\mathrm{F}(i)}}\right)\right\}_{i=1}^{L}$ are now the posterior samples of the GSDs.

Another feature of Bayesian inference is that prediction and model assessments can be carried out from the 'posterior predictive' distribution. To be precise, predicting the response $\mathbf{Y}(t_0)$ at a new location amounts to evaluating $p(\mathbf{Y}(t_0) \mid \mathbf{Y}) = \int p(\mathbf{Y}(t_0) \mid \boldsymbol{\theta}, \mathbf{Y}) p(\boldsymbol{\theta} \mid \mathbf{Y}) \mathrm{d}\boldsymbol{\theta}$, where $p\left(\mathbf{Y}(t_0) \mid \boldsymbol{\theta}_{(i)}, \mathbf{Y}\right)$ further simplifies to $p(\mathbf{Y}(t_0) \mid \boldsymbol{\theta})$ which is the model likelihood at t_0. Again, sampling from $p(\mathbf{Y}(t_0) \mid \mathbf{Y})$ is preferred: having obtained posterior samples, $\left\{\boldsymbol{\theta}_{(i)}\right\}_{i=1}^{\mathrm{M}}$, for each $\boldsymbol{\theta}_{(i)}$, we now draw $\mathbf{Y}_{(i)}(t_0)$ from the distribution $p\left(\mathbf{Y}(t_0) \mid \boldsymbol{\theta}_{(i)}, \mathbf{Y}\right)$. The resulting $\left\{\mathrm{Y}_{(i)}(t_0)\right\}_{i=1}^{\mathrm{M}}$ are the desired samples from the posterior predictive distribution. The mean and standard deviation offer predictive estimates and uncertainty. All uncertainty pertaining to estimation and prediction are accounted for in these samples. Bayesian model assessment can now proceed from this predictive distribution by using the estimated model to predict at observed time points and comparing these predictions with the observed data.

While the highly nonlinear nature of equation (1) precludes effective prediction using simple linear regression models, a least-squares approach that maximizes the likelihood in equation (4), either using non-linear regression or some locally linear approximation, offers a feasible alternative to our Bayesian method for fetching point estimates and predictions. However, quantifying uncertainty of these estimates will rely upon asymptotic normality for large sample sizes. Furthermore, classical nonlinear optimization algorithms can perform poorly with several parameters in equation (1) (we discuss such a setting in Estimates for G and Q, where we estimate not just β, but also G and Q and dependent error structures (see Results with dependent near- and far-field measurements). The sampling-based Bayesian approach absolves us of the above problems: the posterior simulations offer 'exact' inference based upon whatever sample size we have in a numerically stable manner.

SIMULATION STUDY OF TWO-ZONE DATA

Generating data and the methods

We will first illustrate our method using a simulated data set that will help assess the effectiveness of the proposed algorithms. We generate data with all model parameters fixed and known and then to estimate them from the generated data in a Bayesian setting with priors on these parameters. We then check

to see whether the differences between the estimated and true values are within the range of statistical error. This entire procedure is repeated with not one but several, say 100, data sets generated with different parameter values. To be more precise, here we simulate data from the statistical model in equation (3) having the following factors—(i) two different settings for the covariance structure Σ: the first $\begin{bmatrix} 1 & 0 \\ 0 & 0.64 \end{bmatrix}$ assumes independent near- and far-field measurements, while the second $\begin{bmatrix} 1 & 0.5 \\ 0.5 & 0.64 \end{bmatrix}$ imposes a fairly strong association (the correlation coefficient is $0.5/\sqrt{0.64} = 0.625$ to be precise); (ii) β is set to be 0.1, 0.2, . . ., 10 m^3 min^{-1}, yielding a total of 100 simulated data sets for each fixed covariance structure in (i); (iii) a sequence of time points is taken as 0, 0.01, . . ., 4 min, so that each data set has 401 pairs of (C_N, C_F) with a fixed Σ and β. The values of β chosen are in the low to mid-range used by Cherrie (1999) in his simulation study.

The remaining physical parameters are treated as fixed constants. They are assigned values that simulate conditions similar to our real data set (see the next section) with near-field volume $V_N = \pi \times 10^{-3}$ m^3, far-field volume $V_F = 3.8$ m^3, room supply air rate $Q = 13.8$ m^3 min^{-1}, and the constant mass emission rate $G = 351.5$ mg min^{-1}. The data sets are generated using an ordinary differential equation solver from the package odesolve in R. Given the parameters in equation (1), odesolve produces values of $C_N(t)$ and $C_F(t)$ at different time points for fixed β.

For statistical estimation, we worked with the log concentration as our dependent variable and assume Gaussian error distributions in the log scale. In fact, C_N shows serious violation for the normality assumption which the log transformation meliorates. The closed-form expressions we derived for $\mathbf{C}(\beta;t)$ in equation (2) [also equation (A1) in Appendix A] facilitate computations, but even when closed forms were unavailable, we could numerically solve the system in equation (1) using a linear differential equation package (e.g. odesolve in R). This numerical solution must be carried out in every iteration of the MCMC algorithm using the current value of β.

Prior settings

The parameter β represents airflow rate between the near and the far field and is, hence, positive. The prior for β can be log-normal with mean 0 and a large variance. A large prior variance represents vagueness and uncertainty about our confidence in our prior beliefs and allows the data to drive the inference. In our simulation, we set the prior variance to be 2.0, i.e. geometric mean 1, and variance exp(2.0). Based upon practical experience and physical principles, we know β cannot be huge for most of the real situations. Therefore, we can also assume that, β is distributed on a positive interval and another sufficiently vague prior would be the uniform prior $U(0, 50)$. Both these priors gave stable and similar estimates for the testing data set and we chose to present the results from the log-normal prior in the final analysis.

For the error term in equation (3), we first considered the independence error model, where $\Sigma = \text{diag}(\tau_N, \tau_F)$. Therefore, the priors for τ_N and τ_F are set to be inverse Gamma with mean 0 and variance 1000. Next, instead of assuming independence between fields, we also considered the inverse-Wishart prior (see Carlin and Louis, 2008, p. 426, for the Wishart/inverse-Wishart density) for a general Σ matrix. In the simulation, we let Σ have an inverse-Wishart (S, v) prior with the scale matrix whose diagonal elements are each 10 and degrees of freedom $v = 4$.

Simulation results

For each simulated data set, we ran three chains for 10 000 iterations each. Convergence was diagnosed within 2000 iterations using the coda package in R. We discarded the first 6000 (2000×3) samples as burn-in and retained the remaining 24 000 (8000×3) for posterior inference. Using equation (2) directly or numerically solving equation (1) with the odesolve package (which even applies to analytically intractable differential systems) produced practically indistinguishable results. Hence, we only present results from the latter.

Results with independent near- and far-field measurements. Table 1 is the posterior summary for one simulated data set that was generated with the true β equaling 5 and the true $\Sigma = \text{diag}(1, 0.64)$. The table reveals that the values that generated the data are all included in the Bayesian 95% equal-tailed posterior credible intervals (BCI). Estimates of the error term τ_N and τ_F are also quite good for the independent inverse-Gamma priors with the posterior 95% credible interval for τ_N being a little wider than that for τ_F. Figure 2 shows the posterior distribution of coefficient of variation (CV) at different time points for the near and far fields. These are computed using the posterior predictive intervals for $Y_N(t)$ and $Y_F(t)$. With time, the posterior concentrations at the near field and far field become steady, so their CVs quickly become stable as expected.

Figure 3 presents the 95% equal-tailed Bayesian credible intervals (BCI) for β computed from each of the 100 simulated data sets, where true $\beta = 0.1$, 0.2, . . ., 10 m^3 min^{-1} as described in Generating data and the methods and true $\Sigma = \text{diag}(1, 0.64)$. The middle points represent the posterior medians, and the line segments represent the (2.5%, 97.5%) percentiles of the posterior distribution. The overall coverage for the 100 data sets is ~92%, only slightly lower than the exact theoretical coverage of 95%.

Table 1. Parameter estimates for our simulated data example from the model assuming independence between the far and near fields (i.e. $\tau_{NF} = 0$) with G and Q held fixed

	Estimates	2.5%	97.5%	Real value
β	5.10	4.22	6.18	5.00
τ_N	0.92	0.70	1.22	1.00
GSD(τ_N)	2.61	2.31	3.02	2.72
τ_F	0.75	0.57	1.00	0.64
GSD(τ_F)	2.38	2.13	2.72	2.23
$\log(\overline{C_N})$	4.52	2.62	6.43	5.16
$\log(\overline{C_F})$	3.21	1.44	4.91	2.61
β	Interzonal airflow rate (m^3 min^{-1})			
τ_N, τ_F	Variance of measurement error at near and far field (log scaled)			
GSD(τ_N)	Geometric standard deviation for measurement error at near field			
$\overline{C_N}$, $\overline{C_F}$	Steady-state concentration at near and far field (mg m^{-3})			

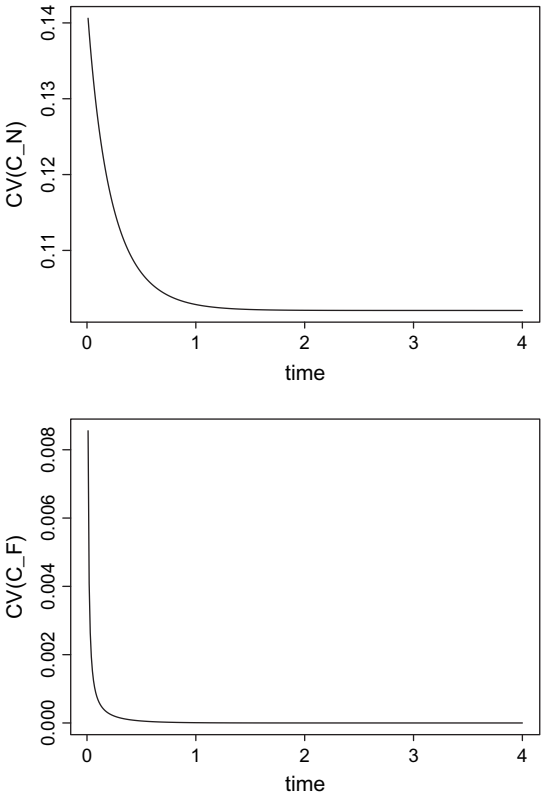


Fig. 2. Posterior means of CVs for a simulated data with true $\beta = 5$, $\tau_N = 1.0$, $\tau_F = 0.64$ and $\tau_{NF} = 0$: upper panel is the near field; lower panel is the far field. X-axis: time (minutes); y-axis: CV(C_N) and CV(C_F).

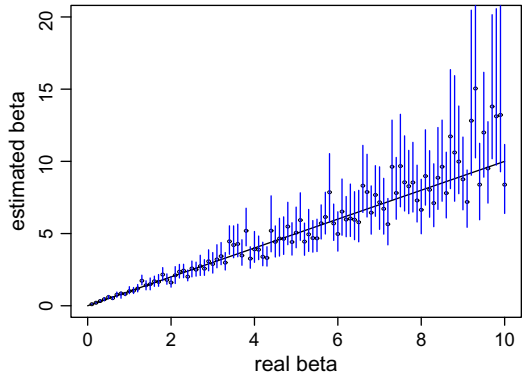
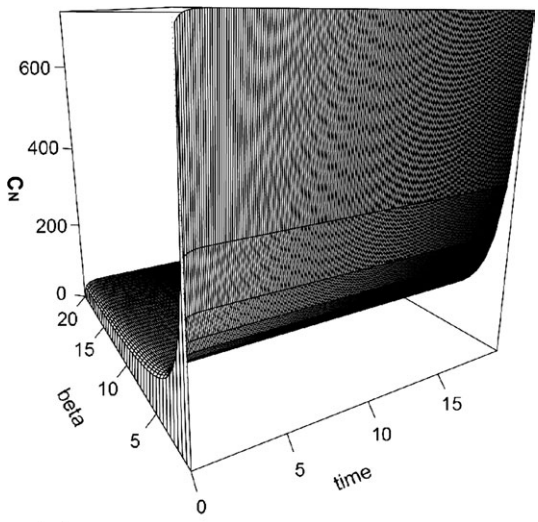


Fig. 3. The 95% equal-tailed posterior credible intervals for 100 simulated data sets, where the true values are $\tau_N = 1$, $\tau_F = 0.64$, $\tau_{NF} = 0$, and $\beta = 0.1, 0.2, \ldots, 10$. X-axis: true β (m^3 min^{-1}); y-axis: estimated β (m^3 min^{-1}).

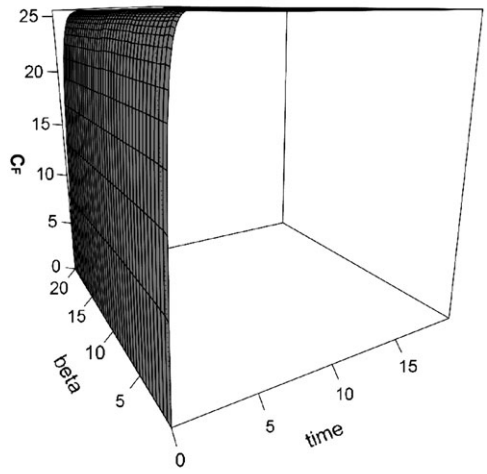
Figure 3 also exhibits an interesting trend: the larger the β, the wider the confidence interval. Additional insight is obtained from the differential system's properties. The 3D plots in Fig. 4 reveal that both C_N and C_F reach steady concentration. For any data set, the smaller the β, the larger its 'steady C_N' will be (in Fig. 4a). On the other hand, for C_F, the data set with different βs always show similar 'steady C_F' (in Fig. 4b) i.e. C_F contains little information about β. Therefore, C_N contains more information than C_F to restrict β in a small interval when the true β is really small, thereby causing narrow Bayesian confidence intervals. On the other hand, even for C_N, the steady concentration becomes more similar for large βs. Consequently, a data set with a large β will have wider confidence intervals.

In Table 1, $\log(\overline{C_N})$ and $\log(\overline{C_F})$ present posterior summaries for the steady-state (when time t is large enough) log concentrations [i.e. $\log(C_N(t))$ and $\log(C_F(t))$] for the simulated data. Figure 5 shows the plots for the predictive estimates of $\log(C_N(t))$ and $\log(C_F(t))$ with time, where the 95% equal-tailed predictive interval is approximately symmetric about the median line. Because the error variance τ_N at the near field is greater than τ_F at the far field, the width of predictive interval for C_N is also larger than C_F. Figure 5 clearly reveals that the predicted concentration (median) stabilizes with time.

Results with dependent near- and far-field measurements. We next generated 100 data sets, where the true covariance matrix Σ was not diagonal, but had off-diagonal elements equaling 0.5. We now

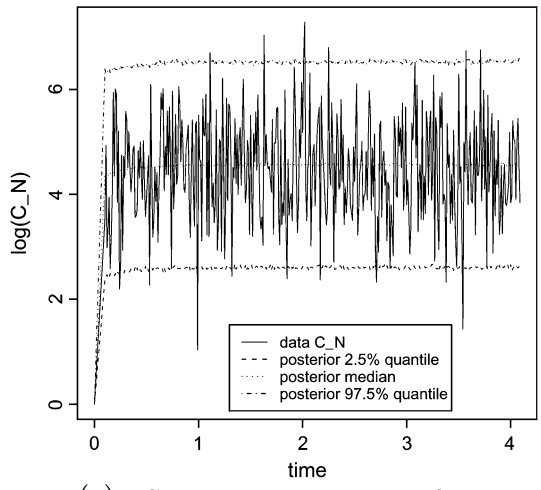


(a) C_N: concentrations at near-field

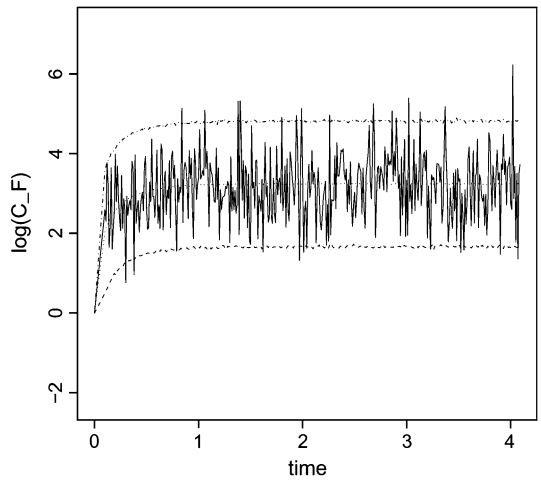


(b) C_F: concentrations at far-field

Fig. 4. 3D plots of concentrations for the two-zone differential equation system. X-axis: time (minutes); y-axis: β (m^3 min^{-1}); z-axis: C_N, C_F (mg m^{-3}).



(a) C_N: concentrations at near field



(b) C_F: concentrations at far field

Fig. 5. Predicted (C_N, C_F) for a simulated data set with true β = 5, τ_N = 1.0, τ_F = 0.64, and τ_{NF} = 0: upper panel is the near field; lower panel is the far field. X-axis: time (minutes); y-axis: log(C_N), log(C_F) (log (mg m^{-3})).

estimated this data set with two models: the first was the independent error model as in the preceding illustration where we put independent inverse-Gamma priors on τ_N and τ_F; the second was the model with the dependent error assumption, where the covariance matrix Σ was assigned an inverse-Wishart prior. Table 2 shows a comparison of posterior summaries for one simulated data set between these two models. The true values that generated this simulated data set are β = 5 m^3 min^{-1}, τ_N = 1, τ_F = 0.64, τ_{NF} = 0.5, i.e. a correlation of ρ = 0.625.

Table 2 reveals that the 95% posterior credible interval (BCI) includes β only for the model with inverse-Gamma prior. The model with the inverse-Wishart prior has a narrower BCI for β than the inverse-Gamma prior and does not include the true β. The inverse-Wishart prior involves an additional parameter estimate that might lead to more conservative intervals for β. The estimation of the error terms is excellent for both priors, even for the covariance term, which has the narrow BCI (0.41, 0.96). We also calculated the BCI coverage of β for the 100 data sets with true τ_{NF} = 0.5 under two different prior models. The overall coverage rate for the inverse-Wishart model is ∼91%, which is slightly higher than the independent inverse-Gamma prior model (89%). This suggests that the independent inverse-Gamma prior model can be robust even for data sets with a fairly strong covariance structure.

Estimates for G and Q. So far, we assumed that only β was unknown in equation (1); we now assume that G and Q are also to be estimated. We did a simulation experiment for the same data as Table 1, but now β, G, and Q are all assumed to be unknown. We

Table 2. Comparisons between parameter estimates from two models: the block on the left fits the model assuming independence between the fields (i.e. $\tau_{NF} = 0$), while the block on the right corresponds to a dependent model that estimates τ_{NF}

	Independent model ($\tau_{NF} = 0$)			Dependent model (τ_{NF} estimated)			
	Estimates	2.5%	97.5%	Estimates	2.5%	97.5%	True value
β	5.24	4.65	6.29	5.31	5.22	5.42	5.00
τ_N	1.00	0.77	1.34	1.12	0.83	1.60	1.00
GSD(τ_N)	2.72	2.40	3.18	2.88	2.49	3.54	2.72
τ_F	0.61	0.47	0.82	0.69	0.50	0.97	0.64
GSD(τ_F)	2.18	1.98	2.47	2.29	2.03	2.68	2.23
τ_{NF}	0	0	0	0.63	0.41	0.96	0.5
$\log(\overline{C_N})$	4.49	2.50	6.42	4.53	1.82	7.25	4.72
$\log(\overline{C_F})$	3.22	1.68	4.78	3.20	−0.26	6.60	4.20
β	Interzonal airflow rate (m^3 min^{-1})						
τ_N, τ_F	Variance for measurement error at near and far field (log scaled)						
τ_{NF}	Covariance between near and far field for measurement error (log scaled)						
GSD(τ_N)	Geometric standard deviation for measurement error at near field						
$\overline{C_N}$, $\overline{C_F}$	Steady-state concentration at near and far field (mg m^{-3})						

The simulated data set was generated using true parameter values indicated in the rightmost column.

then set some 'informative' priors based on practical knowledge. For example, we assigned G a uniform prior within ±20% of the true value $G = 351.5$ mg min^{-1}, that is $U(281, 422)$. For Q, an informative prior, $U(11, 17)$, is given based on true $Q = 13.8$ m^3 min^{-1}. After setting the priors for new parameters, the MCMC procedure is fairly routine. The results are included in Table 3.

From Table 3, all the true parameter values that generated the data are included in the 95% equal-tailed posterior intervals for both dependent and independent models. Treating G and Q as unknown parameters, the estimates for β, τs, $\log(\overline{C_N})$ and $\log(\overline{C_F})$ remain quite close to those in Table 2 for the independent model. For the dependent model, we observed a slight increase in β and a slight decrease in τ_N and τ_F compared to their counterparts in Table 2. The estimates for $G = 351.23$ mg min^{-1} is very close to the true value, 351.5 mg min^{-1} for the independent model, and it (384.84 mg min^{-1}) is a little larger than the true value for the dependent model. Both $Q = 12.58$ m^3 min^{-1} for independent model and $Q = 14.87$ m^3 min^{-1} for dependent model are fairly close to the true value, 13.8 m^3 min^{-1}. It reveals that even with additional unknowns, our method yields relatively stable results. This property is appealing mainly because there are always multiple unknowns that are hard to measure for real problems.

EXPERIMENTAL TWO-ZONE STUDY

Description and design of the experiment

We now turn to an experiment conducted in a test chamber depicted in Fig. 6. The test chamber is connected to the inlet of a centrifugal exhaust fan powered by a 0.75 *HP* (at 2500 *rpm*) motor. The upstream pre-filters increase the level of turbulence in the entering airflow. The test section is 1.73 m long, 1.27 m wide, and 1.73 m high. Thus, the volume of the far field is $V_F = 3.8$ m^3. A mixing fan is mounted midway across the ceiling of the chamber and 20 cm downstream from the air inlet plane of the chamber. The position of the mixing fan was selected after running flow visualization tests to maximize the air-mixing effect. The measured average airflow rate through chamber was $Q = 13.8$ m^3 min.

Toluene in a 200-ml glass impinger was vaporized using an airflow of 15 l min^{-1}, and the outlet of the impinger was connected to a conical metal vessel covered with a mesh placed in the near field (described later). The toluene generation rate was calculated to be $G = 0.404$ ml min^{-1} or 351.5 mg min^{-1} and is constant over time.

It is useful to remember that the model (used to predict worker exposures) makes several simplifying assumptions besides the ones listed in Deterministic Equations of Two-Zones Modeling. These include neglecting the presence of the worker's body, body movement, and body heat. We conducted a series of experiments to explore the effects of each of these factors on concentrations in the two zones. A 25-cm high mannequin measuring 6 cm across the shoulder (14.4% scale mannequin) was used to model the worker. Body movement was simulated by placing the mannequin on a stand powered by a motor that rotated at 10 rpm. For stationary conditions, the mannequin was oriented facing the contaminant source. To simulate heat generated by the human body, the mannequin surface was heated using heating tape that was wrapped around the mannequin. The voltage

Table 3. Parameter estimates for models with G and Q also assumed unknown and estimated: the block on the left fits the model assuming independence between the fields (i.e. $\tau_{NF} = 0$), while the block on the right corresponds to a dependent model that estimates τ_{NF}

	Independent model ($\tau_{NF} = 0$)			Dependent model (τ_{NF} estimated)			
	Estimates	2.5%	97.5%	Estimates	2.5%	97.5%	True value
β	5.21	3.86	6.39	5.61	4.69	6.12	5.00
τ_N	1.01	0.77	1.36	0.92	0.66	1.40	1.00
GSD(τ_N)	2.73	2.40	3.21	2.61	2.25	3.26	2.72
τ_F	0.61	0.47	0.82	0.57	0.40	0.86	0.64
GSD(τ_F)	2.18	1.98	2.47	2.13	1.88	2.53	2.23
τ_{NF}	0	0	0	0.45	0.25	0.82	0.50
G	351.23	298.17	408.55	384.84	320.89	419.10	351.48
Q	12.58	11.07	16.42	14.87	11.18	16.79	13.78
$\log(\overline{C_N})$	4.52	2.51	6.50	4.55	1.96	7.22	4.72
$\log(\overline{C_F})$	3.28	1.70	4.85	3.27	−0.19	6.62	4.20
β	Interzonal airflow rate ($m^3\ min^{-1}$)						
G	Contaminant's mass emission rate ($mg\ min^{-1}$)						
Q	Room supply and exhaust airflow rate ($m^3\ min^{-1}$)						
τ_N, τ_F	Variance for measurement error at near and far field (log scaled)						
τ_{NF}	Covariance between near and far field for measurement error (log scaled)						
GSD(τ_N)	Geometric standard deviation for measurement error at near field						
$\overline{C_N}$, $\overline{C_N}$	Steady-state concentration at near and far field ($mg\ m^{-3}$)						

The simulated data set was generated using true parameter values indicated in the rightmost column.

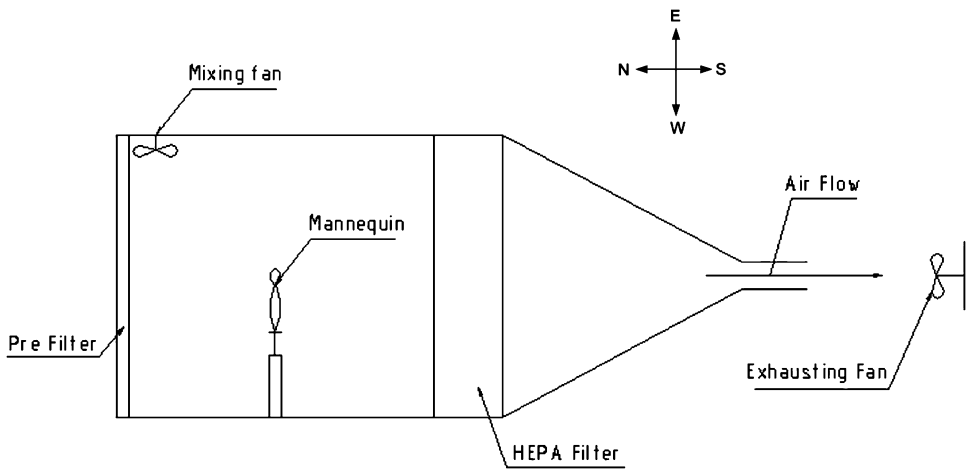


Fig. 6. Schematic of test chamber.

input to the heating tape was regulated using a variable autotransformer that maintained temperature at a constant 33°C under normal conditions (skin temperature is 4°C less than the core temperature of 37°C).

Toluene concentrations were measured with an infrared analyzer (Foxboro MIRAN 1B2, Foxboro, MA, USA). The measured data were logged using a data logger (Model TC 110, Omega Engineering Inc., Stanford, CT, USA) and transferred into a spreadsheet for further analysis. Prior to data collection, the data loggers were calibrated using known toluene concentrations. To study the air contaminant concentration in the test chamber, concentrations were measured at >60 locations around the contaminant source to obtain spatial profiles of toluene concentrations. These measurements were made in four directions (east, west, north, and south) on three horizontal parallel planes at five different distances (10, 15, 20, 30, and 40 cm) from the generation source. The source was located on the middle plane. Concentrations were measured every 5 s and for at least 15 min in each location.

In the previous studies, the near field has been arbitrarily selected as a volume that contains the breathing zone of the worker. We used a different approach. The near and far fields are supposed to have distinctly different concentrations and the model, in fact, predicts a sharp discontinuity at the boundary between the two zones that would not be seen in reality. While the measured concentrations show a more gradual decrease in concentration moving away from the source, the spatial rate of change of concentration is not uniform. The maximum rate of change of concentration occurred at a distance of 10 cm from the source. Accordingly, we used this to define our near field as a 10 cm high cylinder with a radius of 10 cm with its base on the plane of the generation source.

Estimation results for the experimental data

The concentrations in the near and far fields have the same initial states, i.e. $C_N(0) = C_F(0) = 0$ mg m^{-3}. We assume that concentrations at the near and far fields were measured simultaneously at each time point. Although the real data set does not fairly meet this condition, we can estimate β and τ_N using only concentrations at the near field. Later, under more assumptions, we also used $[C_N(t), C_F(t)]$ pairs that were taken under the same experimental condition.

We fit the model in equation (3) to the experimental data set using three parallel MCMC chains of 10 000 iterations each. Convergence diagnostics revealed that the three chains converged fairly quickly and the first 2000 × 3 samples were discarded as burn-in. Table 4 shows the posterior summary of β. The posterior mean of β is $\sim$2.5 m^3 min^{-1} for log-scaled data, while the variance for the log-scaled data at the near field is $\sim$0.7.

After fixing the initial conditions, a unique solution of the equation system in equation (1) can be found [see equation (A1) in Appendix A]. From the equations, we see that $C_N \to \frac{G}{Q} + \frac{G}{\beta}$ and $C_F \to \frac{G}{Q}$ as $t \to \infty$. Therefore, the steady-state solution for β is $\frac{G}{C_N - C_F}$. Theoretically, as time goes by, the measured and modeled concentrations agree reasonably well on steady state as long as the system conditions remain the same. We can calculate β based on the steady-state solution, i.e. $\beta \to \frac{G}{C_N - C_F} = \frac{G}{C_N - \frac{G}{Q}}$. Therefore, the calculated β is $\sim$1.74 m^3 min^{-1} based on the steady concentration C_N close to 60 ppm = 227.44 mg m^{-3} for toluene.

As in Simulation results, we plotted the 95% equal-tailed posterior predicted interval for the near field concentration (log scaled) for the build-up data in Fig. 7. From the figure, the real data show smaller error than the simulated data. There is a rapid increase in the calculated concentration, but a more gradual one in the measured concentration. This might result from the assumption of mixing efficiency of the two-zone model and measurement error. Therefore, our estimated β is not far, but still different from the steady-state solution. When calculating the steady-state solution, we used the steady state of measured concentration. It is hard to define precisely the exact steady-state concentration from experimental data. In our experiment, we used the average concentration of the last 5 min to represent the steady concentration at that point. Therefore, measuring steady concentration involves arbitrary judgments. As a result, the estimated steady-state solutions may also have some bias.

To explore the predictive concentration at the near field, we also plotted the posterior density for C_N at some fixed time points. Figure 8 shows density for posterior C_N at time t = 5, 25, 250, and 1000 s. All the curves have similar shapes as expected. The posterior means for C_N increase rapidly at earlier stages (t = 5 and 25 s), becoming more gradual for later times (t = 250 and t = 1000 s). This agrees well with the 3D data plot (Fig. 4).

As mentioned earlier, we did not measure C_N and C_F simultaneously, but we can assume that separately measured $C_N(t)$ and $C_F(t)$ were taken at the exact same time points. With the same initial conditions, we regarded them as pairs and fit the model. In the following, we consider the complete two-zone data with both near- and far-field measurements. A series of experiments were conducted to explore the effects of the presence of the worker's body, movement of the body, and body heat on C_N, C_F, and β.

Table 4. Parameter estimates for the two-zone data considering C_N measurements only

	Estimates	2.5%	97.5%
β	2.56	2.31	4.22
τ_N	0.75	0.65	0.93
GSD(τ_N)	2.38	2.24	2.62
$\log(\overline{C_N})$	5.09	3.38	6.81
β	Interzonal airflow rate (m^3 min^{-1})		
τ_N	Variance of measurement error at near field (log scaled)		
GSD(τ_N)	Geometric standard deviation for measurement error at near field		
$\overline{C_N}$	Steady-state concentration at near field (mg m^{-3})		

The directional airflow through the chamber also has an effect on the shape of the near field, something that the model ignores. Flow visualization experiments (not reported here) showed that the near field is not symmetrical along the direction of the airflow through the chamber (north–south), while it is symmetrical along the east–west direction. Though measurements taken at northbound and southbound directions were excluded when defining the near-field zone, we wanted to determine if different directions really matter for concentrations and their estimates of β. Table 5a compares the estimates among data sets taken under the same working conditions (without mannequin, body movement, and body heat) but in different directions, where $\hat{\beta}$ refers to the estimates for β from its posterior samples; $CV(\overline{C_N})$ and $CV(\overline{C_F})$ are the posterior estimates of CV for steady-state C_N and C_F; 'β_{Eqm}' refers to the steady-state solution for β as mentioned earlier; $\overline{C_N}$ denotes the posterior estimates for the steady-state C_N; experimental $\overline{C_N}$ is the steady-state C_N that we used to calculate the steady-state solution for β. The steady-state C_N varies for different exposure conditions; therefore, their steady-state solutions $\beta \rightarrow \frac{G}{C_N - \frac{G}{Q}}$ are different too. As expected, all estimated βs are close to the estimated steady-state solution from the differential equation system. The east and west estimates of βs are more similar than those from the south.

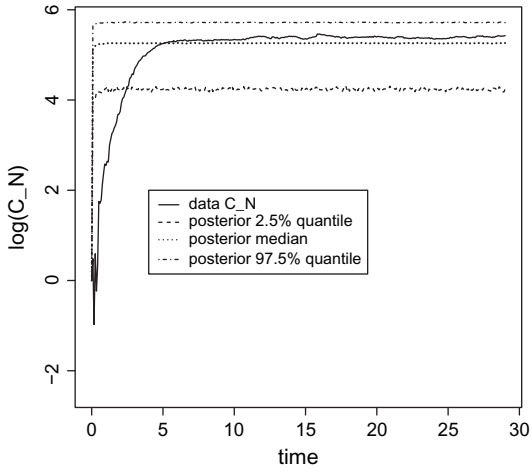


Fig. 7. Posterior predictive median and 95% intervals for $C_N(t)$ computed from the two-zone data considering $C_N(t)$ measurements only. X-axis: time (minutes); y-axis: log(C_N), (log (mg m^{-3})).

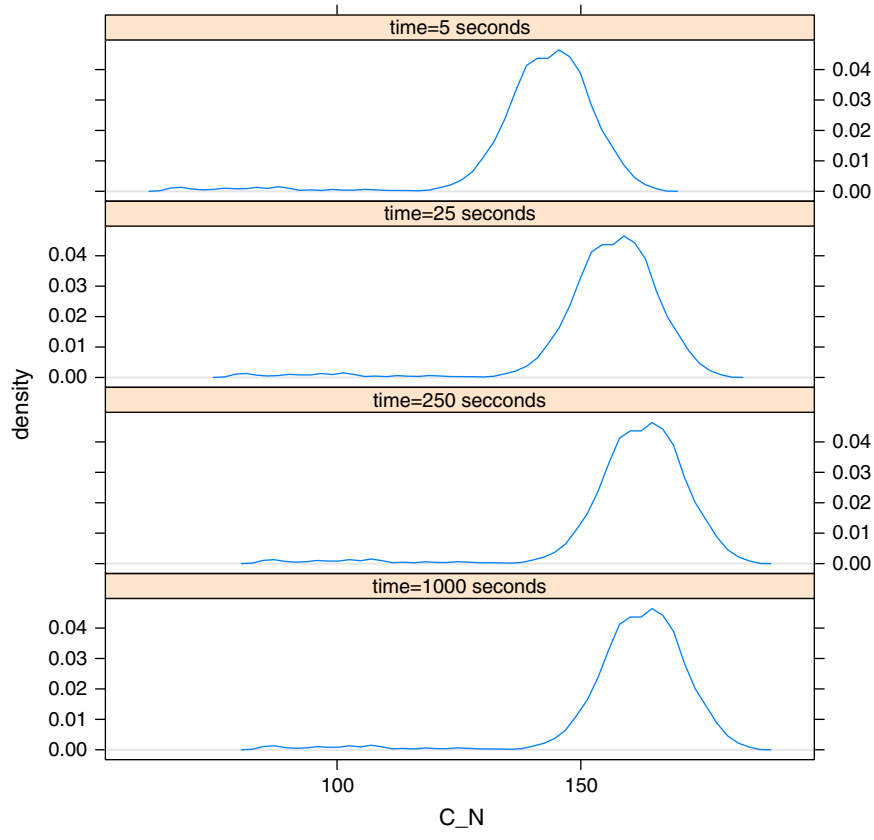


Fig. 8. Posterior predictive density of C_N for the two-zone data considering C_N measurements only. X-axis: C_N (mg m^{-3}); y-axis: density.

Table 5. Comparison of estimated βs and the coefficient of variations for the two-zone experiments: (a) corresponds to different directions under the same working conditions; (b) corresponds to different working conditions but for a fixed direction (east)

(a) Estimated β, CVs and steady-state concentration estimates at different directions under the same working condition

Direction	$\hat{\beta}$ (β_{Eqm})	2.5%	97.5%	$CV(\overline{C_N})$	$\overline{C_N}$(experimental $\overline{C_N}$)	$CV\,\overline{C_F}$	$\overline{C_F}$ (experimental $\overline{C_F}$)
East	4.33 (4.94)	4.26	4.41	0.64	106.68 (96.66)	6.09×10^{-5}	25.51 (66.11)
West	5.16 (5.02)	5.12	5.21	0.32	93.62 (95.52)	1.84×10^{-4}	25.51 (54.69)
South	1.27 (1.41)	1.25	1.30	0.87	302.26 (274.78)	2.38×10^{-4}	25.51 (226.32)

(b) Estimated β, CVs and steady-state concentration estimates at east under different working conditions

Mannequin	Movement	Heat	$\hat{\beta}$ (β_{Eqm})	2.5%	97.5%	$CV(\overline{C_N})$	$\overline{C_N}$ (experimental $\overline{C_N}$)	$CV(\overline{C_F})$	$\overline{C_F}$ (experimental $\overline{C_F}$)
×	×	×	4.33 (4.94)	4.26	4.41	0.64	106.68 (96.66)	6.09×10^{-5}	25.51 (66.11)
√	√	×	3.96 (3.87)	3.90	4.02	0.54	114.26 (116.33)	3.8×10^{-4}	25.51 (50.59)
√	√	×	3.90 (4.13)	3.87	3.92	0.22	115.63 (110.61)	1.56×10^{-4}	25.51 (53.99)
√	×	√	3.97 (4.27)	3.93	4.01	0.42	114.04 (107.82)	2.9×10^{-4}	25.51 (59.59)

$\hat{\beta}$: estimation for the interzonal airflow rate ($m^3\ min^{-1}$); β_{Eqm}: steady-state solution for β ($m^3\ min^{-1}$); $\overline{C_N}$, $\overline{C_F}$: estimation of steady state at near and far field ($mg\ m^{-3}$); $CV(\overline{C_N})$, $CV(\overline{C_F})$: estimation of CV for steady state C_N and C_F; Experimental $\overline{C_N}$, $\overline{C_F}$: steady state C_N, C_F for the experimental data ($mg\ m^{-3}$).

To assess the effects of different exposure conditions, we also experimented with the presence of mannequin, body movement, and body heat. The results are shown in Table 5b. We expect that the presence of the mannequin will reduce air exchange rates between the near and the far field, thereby increasing the toluene concentration. Thus, it decreases the airflow rate β between the near field and far field. The first two rows of Table 5b show that β is ∼4.33 $m^3\ min^{-1}$ without the presence of mannequin and becomes 3.96 $m^3\ min^{-1}$ with the mannequin presented. Comparing with base case (with mannequin present), the body movement and body heat shows a small influence on the concentration level and the airflow rate β as shown in the second, third, and fourth rows of Table 5b. The estimated βs are always ∼3.95 and have similar posterior credible intervals. Thus, their posterior predictive steady-state concentrations also are similar [$\overline{C_N}$ and $\overline{C_F}$ in Table 5b]. The posterior predictive steady-state concentrations for far field are almost the same (≈25.51 mg m^{-3}) for all data, so is their steady-state solution $C_F \to \frac{G}{Q}$. This also implies that C_N contains much more information than C_F. The extremely small $CV(\overline{C_F})$ at far field ($<10^{-3}$) shows the stability of steady-state concentration at far field. Moreover, the small value of $CV(\overline{C_N})$ at near field compared with experimental $\overline{C_N}$ also shows the accuracy of our approach.

Figure 9 shows the 95% posterior predictive intervals for the real data at east without the presence of mannequin, body movement, or body heat. By fitting the complete two-zone data (including C_N and C_F), we found that the predicted C_Fs are farther from the truth than the predicted C_Ns. Obviously, the information from near-field data and far-field data about β do not agree well with each other. It is very likely that there is measurement delay between measured $C_N(t)$ and $C_F(t)$, which we combine together to use as a complete data. Another explanation is that the concentration at the far field is more spatially variable than the near field, e.g. different measurement positions may give very different C_Fs, then our fixed position measurement may not represent the far field very well, thereby producing larger measurement errors for $C_F(t)$ than $C_N(t)$. Consequently, the posterior β has to find a best fit between the two different information sources ($C_N(t)$ and $C_F(t)$). As mentioned earlier in results with independent near- and far-field measurements, C_N has more information about β than C_F and therefore the estimated β agrees more with $C_N(t)$ than $C_F(t)$. As a result, the estimated variance at near field is much smaller than far field, i.e. C_F has much wider posterior predictive intervals.

Validation of the model is essentially a comparison of the model predictions of near and far-field concentrations with actual measurements of the same. This, of course, presupposes that the model input

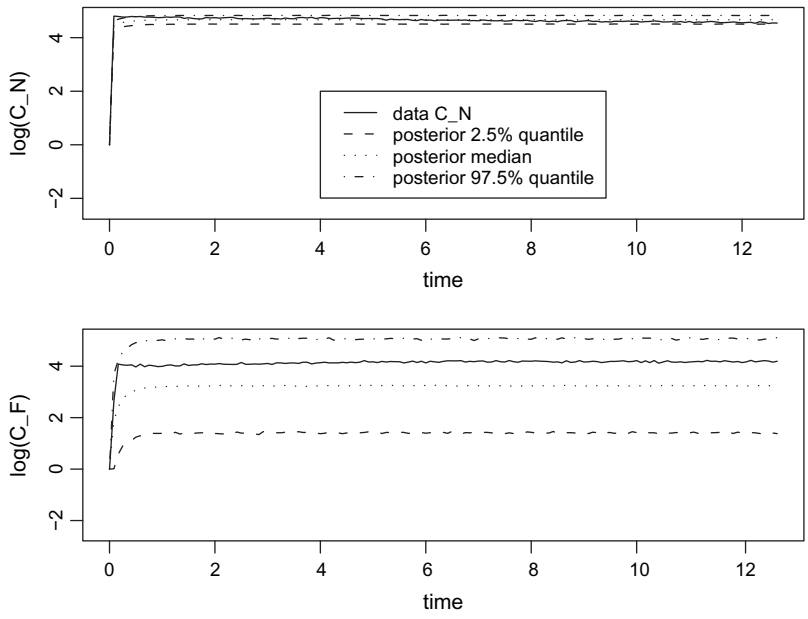


Fig. 9. Prediction of $C_N(t)$ and $C_F(t)$ for the complete two-zone data in a fixed direction (east). X-axis: time (minutes); y-axis: $\log(C_N)$ and $\log(C_F)$ (log (mg m^{-3})).

parameters are known. However if there are uncertainties in the input parameters, validation becomes challenging. The approach here outlined here illustrates how exposure models and information on model parameters together with the knowledge of uncertainty and variability in these quantities are part of the validation process.

It is interesting to compare the values of the parameters chosen in this laboratory study to those in real-life situations reported by Nicas *et al.* (2006) and simulations reported by Cherrie (1999). The ratio of VNF/VFF is $\sim 8 \times 10^{-4}$ in this study while Nicas reports a value of 9×10^{-5}. Thus, there is an order of magnitude difference. However, the air changes per minute in the near field (VNF/beta) varies between $\sim$32 and 3180 in this study. Nicas *et al.* (2006) reports a range of 21–27 air changes per minute for the near field. Thus, the Nicas *et al.* study is in the lower end of the range examined by us. The values of β in this study range between 0.1 and 10 $m^3\ min^{-1}$. This is comparable to the values in Cherrie (3–30 $m^3\ min^{-1}$). Thus, we can conclude that the experimental and simulation conditions examined in this study are roughly of the same order as reported in other studies. At the same time, the validation has been done with a substantial amount of data. The simulations were conducted over a wide range of model input conditions. Realistic effects such as body heat, body movement, and presence of a mannequin were incorporated as part of the experiments. The results show a close match between model predictions and actual measurements.

CONCLUSIONS AND FUTURE WORK

We have proposed a nonlinear regression model within a Bayesian setting for analyzing experimental data from two-zone experiments. This approach combines prior information on the physical model along with the observed data and accounts for uncertainty in measurement and in the rate of airflow between the far and near fields. We recognize the variability arising in experimental data and also predict the concentration from the model and arrive at more pronounced inference regarding the time to achieve system equilibrium.

Our findings reveal that direction and exposure conditions affect the concentration levels differently. Indeed, direction has a pronounced effect on β and the concentration level. On the other hand, the presence of the mannequin has greater effect than body movement and body heat. We found body movement and body heat to only slightly affect β and the concentration level. Our estimates of β are always close to the estimated steady-state solutions, implying that our method works efficiently. The predictions of near-field concentration for both the simulated and experimental data show nice concordance between the observed and predicted values, indicating that the two-zone model assumptions agree with the reality to a large extent and the model is suitable for predicting the contaminant concentration.

Our approach also helps validate the underlying physical process using experimental data. In other words, we need to test if the experimental data

support the nonlinear relationship between concentration and time until it attains steady state. Simply predicting the average concentration using linear relationships, as is done by normal linear regression models, will not allow such validations as they do not estimate the physical model. Our framework can also be applied to other physical models to estimate parameters or validate the model itself.

APPENDIX

Solution for the two-zone differential equations

A unique solution of equation (1) is

$$C_N = \frac{G}{Q} + \frac{G}{\beta} + G\left(\frac{\beta Q + \lambda_2 V_N(\beta + Q)}{\beta Q V_N(\lambda_1 - \lambda_2)}\right)e^{(\lambda_1 t)} - G\left(\frac{\beta Q + \lambda_1 V_N(\beta + Q)}{\beta Q V_N(\lambda_1 - \lambda_2)}\right)e^{(\lambda_2 t)}, C_F = \frac{G}{Q} + G\left(\frac{\lambda_1 V_N + \beta}{\beta}\right)\left(\frac{\beta Q + \lambda_2 V_N(\beta + Q)}{\beta Q V_N(\lambda_1 - \lambda_2)}\right)e^{(\lambda_1 t)} - G\left(\frac{\lambda_1 V_N + \beta}{\beta}\right)\left(\frac{\beta Q + \lambda_1 V_N(\beta + Q)}{\beta Q V_N(\lambda_1 - \lambda_2)}\right)e^{(\lambda_2 t)}, \quad \text{(A1)}$$

where λ_1 and λ_2 are eigenvalues of the matrix $\mathbf{A}(\beta)$ in equation (2). These are available in closed form:

$$\lambda_1 = 0.5\left[-\left(\frac{\beta V_F + (\beta + Q)V_N}{V_N V_F}\right) + \sqrt{\left(\frac{\beta V_F + (\beta + Q)V_N}{V_N V_F}\right)^2 - 4\left(\frac{\beta Q}{V_N V_F}\right)}\right],$$

$$\lambda_2 = 0.5\left[-\left(\frac{\beta V_F + (\beta + Q)V_N}{V_N V_F}\right) - \sqrt{\left(\frac{\beta V_F + (\beta + Q)V_N}{V_N V_F}\right)^2 - 4\left(\frac{\beta Q}{V_N V_F}\right)}\right]. \quad \text{(B1)}$$

Bayesian algorithms: technical details

We provide some details on the MCMC algorithm for estimating β and Σ. The setting with G and Q unknown is analogous. The algorithm proceeds iteratively: we first assign starting values $\beta_{(0)}$ and $\Sigma_{(0)}$ to the parameters; then the i-th iteration updates these parameters by drawing from their full conditionals $\beta_{(i)} \sim p(\beta|\Sigma_{(i-1)}, \mathbf{Y}) \propto p(\beta)p(\mathbf{Y}|\beta, \Sigma_{(i-1)})$ and $\Sigma_{(i)} \sim p(\Sigma|\beta_{(i)}, \mathbf{Y}) \propto p(\Sigma)p(\mathbf{Y}|\beta_{(i)}, \Sigma)$. The mathematical expressions for these 'full conditional distributions' are required only up to a proportionality constant.

For β, we note that the prior or expert knowledge regarding this parameter would usually be summarized either as a uniform or a log-normal distribution (note that $\beta > 0$). These do not yield a standard distribution to draw from and we update $\beta_{(i)}$ using a Metropolis–Hastings step (Gelman *et al.*, 2004). We first draw a β^* from a log-normal 'proposal' distribution, denoted $J(v|u)$, with mean u and a fixed variance σ^2. We then calculate an 'acceptance ratio' $\alpha = \frac{p(\beta^*|\Sigma,\mathbf{Y})/J(\beta^*|\beta)}{p(\beta|\Sigma,\mathbf{Y})/J(\beta|\beta^*)}$. If $\alpha > 1$, then we set $\beta_{(i)} = \beta^*$. If $\alpha \in (0, 1)$, then we draw a random number from $U(0, 1)$ and set $\beta_{(i)} = \beta^*$ if the drawn number is less than α; otherwise we 'reject' the proposed value and set $\beta_{(i)} = \beta_{(i-1)}$.

After obtaining $\beta_{(i+1)}$, we turn to the parameter Σ, which has a prior distribution $IW(S, \nu)$ (Carlin and Louis, 2008, p. 426). Here, S is a 2×2 positive definite scale matrix and ν is the degrees of freedom. We let S be diagonal implying no prior assumption of dependence between the near and far-field measurements. Then $\Sigma_{(i+1)}$ is updated using a Gibbs update: $\Sigma_{(i+1)} \sim IW(S_{1(i+1)}, \nu_1)$, where $S_1 = S + \sum_{j=1}^{n}\left(\mathbf{Y}(t_j) - \log(\mathbf{C}(\beta_{(i+1)}); t_j)\right)\left(\mathbf{Y}(t_j) - \log(\mathbf{C}(\beta_{(i+1)}); t_j)\right)^T$ and $\nu_1 = \nu + n$. For the independence model, $\tau_{NF} = 0$. Assuming that τ_N and τ_F have $IG(a_N, b_N)$ and $IG(a_F, b_F)$ priors, respectively, we obtain Gibbs updates for τ_N and τ_F:

$$\tau_{N(i+1)} \sim IG\left(a_N + \frac{n}{2}, b_N + \frac{1}{2}\sum_{j=1}^{n}(Y_N(t_j) - \log(C_N(\beta_{(i+1)}; t_j)))^2\right) \text{ and } \tau_{F(i)} \sim IG\left(a_F + \frac{n}{2}, b_F + \frac{1}{2}\sum_{j=1}^{n}(Y_F(t_j) - \log(C_F(\beta_{(i+1)}; t_j)))^2\right).$$

REFERENCES

American Industrial Hygiene Association (AIHA). (2000) Mathematical models for estimating occupational exposure to chemicals. Fairfax, VA: AIHA Exposure Assessment Strategies Committee, AIHA Press.

Baldwin PE, Maynard AD. (1998) A survey of wind speeds in indoor workplaces. Ann Occup Hyg; 42: 303–13.

Carlin BP, Louis TA. (2008) Bayesian methods for data analysis. Boca Raton, FL: Chapman & Hall/CRC Press.

Cherrie JW. (1999) The effect of room size and general ventilation on relationship between near and far-field concentration. Appl Occup Environ Hyg; 14: 539–46.

Gelman A, Carlin JB, Stern HS *et al.* (2004) Bayesian data analysis. Boca Raton, FL: Chapman & Hall/CRC Press.

Gilks W, Richardson S, Spiegelhalter D. (1996) Markov chain Monte Carlo in practice. London: Chapman and Hall.

Hemeon WC. (1963) Plant and process ventilation. 2nd. New York: Industrial Press; pp. 235–45.

Laub AJ. (2005) Matrix analysis for scientists and engineers. Philadelphia, PA: Society for Industrial and Applied Mathematics.

Marin J, Robert CP. (2007) Bayesian core: a practical approach to computational Bayesian statistics. New York: Springer.

Melikov A, Zhou G. (1996) Air movement at the neck of the human body. Proceedings, 7th international conference on indoor air quality and climate: indoor air '96. Vol. 1. Nagoya, Japan; pp. 209–14.

Nicas M. (1996) Estimating exposure intensity in an imperfectly mixed room. Am Ind Hyg Assoc J; 57: 542–50.

Nicas M. (2003) Estimating methyl bromide exposure due to offgassing from fumigated commodities. App Occup Environ Hyg; 18: 200–10.

Nicas M, Jayjock M. (2002) Uncertainty in exposure estimates made by modeling versus monitoring. AIHA J; 63: 275–83.

Nicas M, Miller SL. (1999) A multi-zone model evaluation of the efficacy of upper-room air ultraviolet germicidal irradiation. Appl Occup Environ Hyg; 14: 317–28.

Nicas M, Plisko MJ, Spencer JW. (2006) Estimating benzene exposure at a solvent parts washer. J Occup Environ Hyg; 3: 284–91.

Ramachandran G. (2008) Toward better exposure assessment strategies—the new NIOSH initiative. Ann Occup Hyg; 52: 297–301.