Technometrics

ISSN: 0040-1706 (Print) 1537-2723 (Online) Journal homepage: https://www.tandfonline.com/loi/utch20

Bayesian State Space Modeling of Physical Processes in Industrial Hygiene

Nada Abdalla, Sudipto Banerjee, Gurumurthy Ramachandran & Susan Arnold

To cite this article: Nada Abdalla, Sudipto Banerjee, Gurumurthy Ramachandran & Susan Arnold (2020) Bayesian State Space Modeling of Physical Processes in Industrial Hygiene, Technometrics, 62:2, 147-160, DOI: 10.1080/00401706.2019.1630009

To link to this article: https://doi.org/10.1080/00401706.2019.1630009

View supplementary material

Accepted author version posted online: 26 Jun 2019.
Published online: 22 Jul 2019.

Submit your article to this journal

Article views: 65

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=utch20

TECHNOMETRICS
2020, VOL. 62, NO. 2, 147–160
https://doi.org/10.1080/01621459.2019.1630009

Taylor & Francis
Taylor & Francis Group

Check for updates

Bayesian State Space Modeling of Physical Processes in Industrial Hygiene

Nada Abdalla[a], Sudipto Banerjee[a], Gurumurthy Ramachandran[b], and Susan Arnold[c]

[a]Department of Biostatistics, University of California-Los Angeles, Los Angeles, CA; [b]Department of Environmental Health and Engineering, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD; [c]Division of Environmental Health Sciences, School of Public Health, University of Minnesota, Minneapolis, MN

ABSTRACT

Exposure assessment models are deterministic models derived from physical–chemical laws. In real workplace settings, chemical concentration measurements can be noisy and indirectly measured. In addition, inference on important parameters such as generation and ventilation rates are usually of interest since they are difficult to obtain. In this article, we outline a flexible Bayesian framework for parameter inference and exposure prediction. In particular, we devise Bayesian state space models by discretizing the differential equation models and incorporating information from observed measurements and expert prior knowledge. At each time point, a new measurement is available that contains some noise, so using the physical model and the available measurements, we try to obtain a more accurate state estimate, which can be called filtering. We consider Monte Carlo sampling methods for parameter estimation and inference under nonlinear and non-Gaussian assumptions. The performance of the different methods is studied on computer-simulated and controlled laboratory-generated data. We consider some commonly used exposure models representing different physical hypotheses. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received July 2018
Accepted May 2019

KEYWORDS
Bayesian modeling; Exposure assessment; Industrial hygiene; Kalman filters; Physical models; State-space modeling

1. Introduction

In industrial hygiene, estimation of a worker's exposure to chemical concentrations in the workplace is an important concern. In many situations, chemical concentrations are not observed directly and partial noisy measurements are available. Exposure models aim at capturing the underlying physical processes generating chemical concentrations in the workplace. Statistical and mathematical models may provide more accurate exposure estimates than monitoring (Nicas and Jayjock 2002). Industrial hygienists seek to estimate these latent processes from the available measurements as well as quantification of uncertainty in parameter estimation. For example, generation and ventilation rates in a worker's chamber are crucial parameters that are difficult to obtain since most workplaces do not collect information routinely. Traditional approaches involve deterministic physical models that ignore the existence of uncertainty by assigning values to those parameters (Keil, Berge, and AIHA 2009). These approaches do not provide accurate representation in a real workplace.

Bayesian methods (see, e.g., Banerjee et al. 2014) combining professional judgment from experts and direct measurements have been shown to be effective in industrial hygiene decision-making. However, Bayesian inference on processes generating chemical concentrations have received scant attention. Zhang et al. (2009) employed a Bayesian nonlinear regression using the solution of the differential equations representing the underlying physical process using Gaussian errors. The model has some limitations since it ignores extraneous factors and variations and requires a closed-form solution of the differential equations. Monteiro, Banerjee, and Ramachandran (2011) introduced an R package (B2Z), which implements the Bayesian two-zone model proposed by Zhang et al. (2009). Monteiro, Banerjee, and Ramachandran (2014) demonstrated that straightforward Bayesian regression can be ineffective in predicting exposure concentrations in industrial workplaces since the information is limited to partial measurements. They introduced a process-based Bayesian melding approach, where measurements are related to the physical model through a stochastic process capturing the bias in the physical model and a measurement error. The resulting inference suffers from inflated variability because of the additional complexities in the model and cumbersome computations due to Gaussian process random effects.

Physical models for industrial hygiene are represented by differential equations that model the rate of change in concentrations. We propose Bayesian state space models (SSMs) by discretizing differential equations and incorporating information from observed measurements and prior knowledge of experts. This enriches the existing methods as we are no longer restricted to fitting a confined selection of physical models amenable to analytic solutions. Any conceivable physical model, in theory, can be accommodated. Neither will they be restricted to Gaussian data, an assumption that most industrial hygiene practitioners will agree is rarely tenable, especially given the small to moderate number of measurements they have to deal with.

CONTACT Sudipto Banerjee sudipto@ucla.edu Department of Biostatistics, University of California-Los Angeles, Los Angeles, CA.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/TECH.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/TECH.
© 2019 American Statistical Association and the American Society for Quality

At each time point, a new measurement is available that contains some noise, so using the physical model and the available measurements, we try to obtain a more accurate state estimate, which can be called filtering. The importance of filters lies in their ability to produce estimates of the latent process using information generated by the observations which may provide a poor representation of the latent process if used alone. The aim is to evince and statistically quantify the latent process using chamber observations, while incorporating information from the physical model that theoretically describes it and expert knowledge of the posited physical system. We consider Monte Carlo based filtering methods for parameter estimation and inference in SSMs. We also relax the assumption of Gaussian error terms and consider other alternatives.

In particular, we consider different filtering methods under different assumptions. The widely deployed Kalman filter (KF) (Eubank 2005) offers an optimal solution under linearity and normality assumptions. State-by-state update samplers (Fearnhead 2011) can provide state estimates under nonlinear and/or non-Gaussian models. The different models are compared and assessed using computer-simulated as well as lab-generated datasets. In the lab-generated data, most of the model parameters are known up to a considerable level of accuracy. Experiments were conducted in a controlled chamber that mimics real workplace settings, where concentrations were generated at different ventilation and generation rates and under different exposure model settings.

Our contribution in this article expands upon the existing exposure models to allow for better prediction of the quantities of interest. The article is organized as follows. Section 2 provides a brief review of three families of commonly referenced exposure physical models. Section 3 describes the Bayesian approaches used. Section 4 applies our models to the simulated data and lab-generated data. Section 5 concludes the article with an eye toward future work.

2. Physical Models and Their Statistical Counterparts

Bayesian SSMs for exposure assessment incorporate direct measurements of the environmental exposure, deterministic physical models, and prior information from experts. There are several physical models varying in their level of complexity (Ramachandran 2005). Three commonly used families that we consider here are: (i) the well-mixed compartment (one-zone) model; (ii) the two-zone model; and (iii) the turbulent eddy diffusion model. We use discrete approximations to these deterministic models and introduce stochasticity to devise flexible Bayesian versions. This obviates the need for exact analytic solutions to the differential equations, which can be sensitive to the choice of initial conditions. Prior specifications for the model parameters produce Bayesian SSMs. Dynamic steady-state models are composed of (i) a measurement equation that relates the observations (or some function thereof) to the true concentrations; and (ii) a transition equation describing the concentration change from time t to time $t + \delta_t$. We will derive the dynamic models from the respective differential equations for three popular physical models in industrial hygiene.

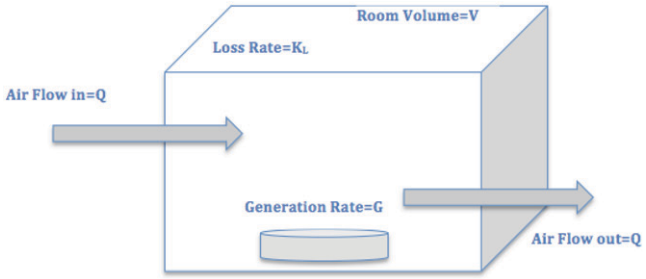

Figure 1. One-zone model schematic showing key model parameters; generation rate G, ventilation rate Q, and loss rate K_L.

2.1. Well-Mixed Compartment (One-Zone) Model

The well-mixed compartment model assumes that a source is generating a pollutant at a rate G (mg/min) in a room of volume V (m^3) with ventilation rate Q (m^3/min). The room is assumed to be perfectly mixed, which means that there is a uniform concentration of the contaminant throughout the room (Figure 1). The loss term K_L (mg/min) measures the loss rate of the contaminant due to other factors such as chemical reactions or the contaminant being absorbed by the room surfaces. The differential equation describing this model is

$$V\frac{d}{dt}C(t) + (Q + K_L V)\,C(t) = G. \tag{1}$$

The exact solution to (1) is

$$C(t) = \exp\{-t(Q + K_L V)/V\}C(t_0) + ((Q + K_L V)/V)^{-1} \left[1 - \exp\{-t(Q + K_L V)/V\}\right] G/V.$$

Theoretically, the steady state concentration is the limit of $C(t)$ as $t \to \infty$ which is $G/(Q+K_L V)$ (mg/m^3). Details of the steady state solution are provided in the supplementary materials. Further specifications yield the Bayesian SSM corresponding to (1) as below,

$$\begin{aligned}
&\text{Measurement: } Z_t = f(C_t) + \nu_t\,, \quad \nu_t \overset{\text{iid}}{\sim} P_{\nu,\theta_\nu};\\
&\text{Transition: } C_{t+\delta_t} = \left(1 - \delta_t \frac{Q + K_L V}{V}\right) C_t + \delta_t \frac{G}{V} + \omega_t,\\
&\qquad\qquad \omega_t \overset{\text{iid}}{\sim} P_{\omega,\theta_\omega}.\\
&Q \sim \text{Unif}(a_Q, b_Q);\quad G \sim \text{Unif}(a_G, b_G);\\
&\qquad K_L \sim \text{Unif}(a_{K_L}, b_{K_L});\quad \sigma^2 \sim \text{IG}(a_\sigma, b_\sigma),
\end{aligned} \tag{2}$$

where Z_t represents measurements (perhaps transformed), $f(\cdot)$ is a function that maps C_t to the scale of Z_t, P_{ν,θ_ν}, and P_{ω,θ_ω} are probability distributions to be specified, while the prior distributions for the physical parameters are customarily specified as uniform within certain fixed physical bounds. The transition equation depends on the parameters only through $(Q/V + K_L)$ and G/V. One can, therefore, reparameterize the transition equation in terms of these functions. However, industrial hygienists are interested in estimating all the parameters $\{Q, K_L, G\}$ (the volume V is usually known) using prior information on these parameters. While constructing a prior on G is equivalent to a prior on G/V, experts find it easier to quantify and construct prior beliefs on the individual parameters $\{Q, K_L\}$ than on the function $(Q/V + K_L)$. Hence, the formulation in (2) is retained in the Bayesian setting developed here.

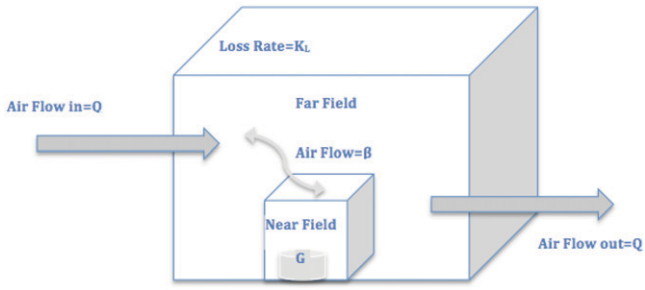

Figure 2. Two-zone model schematic showing key model parameters; generation rate G, ventilation rate Q, airflow β, and loss rate K_L.

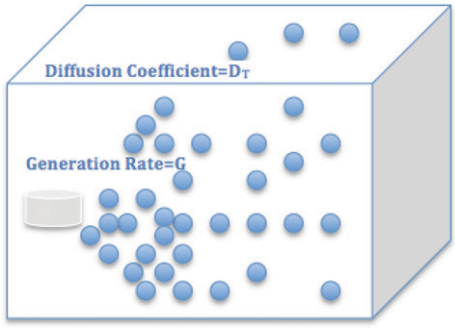

Figure 3. Eddy diffusion model schematic showing key model parameter; diffusion coefficient D_T.

2.2. Two-Zone Model

The two-zone model assumes the presence of a source for the contaminant in the workplace. Two zones or regions are defined: (i) the region closer to the source is called the "near field," while the rest of the room is called the "far field," which completely encloses the near field. Both fields are assumed to be a well-mixed box, that is, two distinct places that are in the same field have equal levels of concentration of the contaminant. Similar to the one-zone model, this model assumes that a contaminant is generated at a rate G (mg/min), in a room with supply and exhaust flow rates (ventilation rate) Q (m^3/min), and loss rate by other mechanisms K_L (mg/m^3). This model includes one more parameter that indicates the airflow between the near and the far field β (m^3/min). The volume in the near field is denoted by V_N (m^3) and the volume in the far field is denoted by V_F (m^3). Figure 2 illustrates the dynamics of the system. The following system of differential equations represents the two-zone model,

$$\overbrace{\frac{d}{dt}\begin{bmatrix} C_N(t) \\ C_F(t) \end{bmatrix}}^{\frac{d}{dt}C(t)} = \overbrace{\begin{bmatrix} -\beta/V_N & \beta/V_N \\ \beta/V_F & -(\beta+Q)/V_F - K_L \end{bmatrix}}^{A} \overbrace{\begin{bmatrix} C_N(t) \\ C_F(t) \end{bmatrix}}^{C(t)} + \overbrace{\begin{bmatrix} G/V_N \\ 0 \end{bmatrix}}^{g}. \tag{3}$$

The solution to (3) is $C(t) = \exp(tA)C(t_0) + A^{-1}\left[\exp(tA) - I\right]g$, where $\exp(tA)$ is the matrix exponential. Theoretically, for large values of t, the steady state concentrations are $\frac{G(\beta + Q + K_L V_F)}{\beta(Q + V_F K_L)}$ (mg/m^3) and $\frac{G}{Q + K_L V_F}$ (mg/m^3) for the near and far fields, respectively. The matrix exponential can be numerically unstable to compute in general. For example, for non-diagonalizable matrices a Jordan decomposition (see, e.g., Banerjee and Roy 2014) may be required, which is very sensitive to small perturbations in A. Hence, we will avoid this approach. Instead, we derive the discrete counterpart of (3) as

$$\begin{aligned}
&\text{Measurement: } Z_t = f(C_t) + v_t\,,\ v_t \overset{\text{iid}}{\sim} P_{v,\theta_v};\\
&\text{Transition: } C_{t+\delta_t} = (\delta_t A(\theta_c; x) + I)\, C_t + \delta_t g(\theta_c; x) + \omega_t;\\
&\qquad \omega_t \overset{\text{iid}}{\sim} P_{\omega,\theta_\omega};\\
&Q \sim \text{Unif}(a_Q, b_Q); \quad G \sim \text{Unif}(a_G, b_G);\\
&\qquad K_L \sim \text{Unif}(a_{K_L}, b_{K_L}); \quad \beta \sim \text{Unif}(a_\beta, b_\beta),
\end{aligned}$$

where Z_t is the 2×1 vector with near-field and far-field measurements (or some function thereof) at time t, C_t is the unobserved concentration state at time t, $A(\theta_c; x) = \begin{bmatrix} -\beta/V_N & \beta/V_N \\ \beta/V_F & -(\beta+Q)/V_F - K_L \end{bmatrix}$ and $g(\theta_c; x) = \begin{bmatrix} G/V_N \\ 0 \end{bmatrix}$. Similar to the one-zone model, we will specify distributions for v_t and for ω_t, where θ_v and θ_ω are parameters in P_{v,θ_v} and P_{ω,θ_ω}, respectively.

2.3. Turbulent Eddy Diffusion Model

In real workplace settings, the rooms may neither be perfectly mixed nor consist of well-mixed zones. Furthermore, the concentration state could depend upon space and time. A popular model for such settings is the turbulent eddy diffusion model, which accounts for a continuous concentration gradient from the source outward. It accounts for the worker's location relative to the source. The concentration $C(s,t)$ is a function of the location $s = (x, y)$ in a two-dimensional Euclidean coordinate frame and time t, where the source of the contaminant is assumed to be at $(0, 0)$. The parameter that is unique to this model is the turbulent eddy diffusion coefficient D_T (m^2/min). It describes how quickly the emission spreads with time (Figure 3) and is assumed to be constant over space and time. There has been very little research on the values of D_T due to the difficulty of measuring it. Some studies suggest a relationship between D_T and air change per hour (ACH) (Shao et al. 2017). We will provide inference for this parameter.

The exact contaminant concentration at location s relative to the source of emission is

$$C(s,t) = \frac{G}{2\pi D_T \|s\|}\left\{1 - \text{erf}\left(\frac{\|s\|}{\sqrt{4D_T t}}\right)\right\}, \tag{4}$$

where $\text{erf}(z) = \frac{2}{\pi}\int_0^z \exp(-u^2)du$. The steady state concentration at location s is theoretically the limit of the concentration as $t \to \infty$, which is $G/(2\pi D_T(s))$ (mg/m^3). The following differential equation represents the change in concentration over time

$$\frac{d}{dt}C(s,t) = \frac{G}{4(D_T \pi t)^{3/2}} \exp\left(-\frac{\|s\|^2}{4D_T t}\right).$$

A general dynamic modeling framework accounting for space and time is as follows

$$\begin{aligned}
&\text{Measurement: } Z(t,s) = f(C(t,s)) + \nu_t(s) + \eta_t,\\
&\qquad \nu_t(s) \sim P_{\nu_t(s),\theta_\nu},\ \eta_t \sim P_{\eta_t,\theta_\eta};\\
&\text{Transition: } C(s,t+\delta_t) = C(s,t) + \delta_t \frac{G}{4(D_T\pi t)^{3/2}}\\
&\qquad \exp\left(-\frac{\|s\|^2}{4D_T t}\right) + \omega(s,t+\delta_t),\ \omega(s,t) \sim P_{\omega_{t,s},\theta_\omega};\\
&D_T \sim \text{Unif}(a_{D_T}, b_{D_T}); \quad G \sim \text{Unif}(a_G, b_G),
\end{aligned} \tag{5}$$

where $P_{\nu_t(s),\theta_\nu}$ and $P_{\omega_{t,s},\theta_\omega}$ are probability laws for spatial-temporal stochastic processes and P_{η_t,θ_η} is the law for a temporally structured or perhaps a white-noise process. Note that $\nu_t(s)$ is a spatial-temporal process discrete in time and continuous in space. This is reasonable because the measurements are taken over discrete time points and the estimation for the latent concentration states are usually desired at those times. On the other hand, $\omega(s,t)$ would ideally be a process continuous in both space and time because it models spatial-temporal associations between concentration states at arbitrary space-time coordinates.

3. Model Implementation and Assessment

For each physical model in Section 2 we will consider two different Bayesian SSMs. We will refer to the first as a Gaussian SSM. Gaussian (linear) SSMs result from specifying $f(C_t) = B_t C_t$, where B_t is a known $p \times p$ design matrix (usually the identity matrix), $P_{\nu,\theta_\nu} \equiv N(0,\Sigma_\nu)$ and $P_{\omega,\theta_\omega} \equiv N(0,\Sigma_\omega)$ are p-variate Gaussian densities. These deliver accessible distribution theory for updating parameters using Kalman-filters or Gibbs samplers. Let $\mathcal{T} = \{t_1,\ldots,t_n\}$ be timepoints where concentration measurements Z_t have been measured. A Bayesian hierarchical SSM is

$$\begin{aligned}
&p(\theta_c) \times \text{IW}(\Sigma_\omega \,|\, r_\omega, S_\omega) \times \text{IW}(\Sigma_\nu \,|\, r_\nu, S_\nu) \times N(C_{t_0} \,|\, M_0 m_0, M_0)\\
&\quad \times \prod_{i=1}^{n} N(C_{t_i} \,|\, A_{t_i}(\theta_c) C_{t_{i-1}} + \delta_i g_{t_i}, \Sigma_\omega)\\
&\quad \times \prod_{i=1}^{n} N(Z_{t_i} \,|\, B_{t_i} C_{t_i}, \Sigma_\nu),
\end{aligned} \tag{6}$$

where $p(\theta_c)$ is the prior distribution on θ_c, $\delta_i = t_i - t_{i-1}$, and the other distributions follow definitions as in Gelman et al. (2013).

We will implement filtering and smoothing on the latent concentration states. Filtering estimates the posterior expectation of the concentration value C_{t_i} given the data up to and including t_i, that is, $\text{E}[C_{t_i} \,|\, Z_{1:i}]$, where $Z_{1:i} = \{Z_{t_j} : j = 1,2,\ldots,i\}]$, for every $t_i \in \mathcal{T}$. For any fixed values of the parameters $\Omega = \{\theta_c, \Sigma_\omega, \Sigma_\nu\}$ filtering can be achieved by adapting the KFs forecast and update steps to our physical models. Given the distribution $p(C_{t_{i-1}} \,|\, \Omega, Z_{1:(i-1)}) = N(C_{t_{i-1}} \,|\, M_{t_{i-1}} m_{t_{i-1}}, M_{t_{i-1}})$, the forecast step computes the predictive distribution $p(C_{t_i} \,|\, \Omega, Z_{1:(i-1)}) = N(C_{t_i} \,|\, \tilde{\mu}_{t_i}, \tilde{\Sigma}_{t_i})$, where $\tilde{\mu}_{t_i} = A_{t_i}(\theta_c) M_{t_{i-1}} m_{t_{i-1}} + \delta_i g_{t_i}$ and $\tilde{\Sigma}_{t_i} = A_{t_i}(\theta_c) M_{t_{i-1}} A_{t_i}(\theta_c)^\top + \Sigma_\omega$. The update step follows the forecast step to compute $p(C_{t_i} \,|\, \Omega, Z_{1:i}) = N(C_{t_i} \,|\, M_{t_i} m_{t_i}, M_{t_i})$, where $M_{t_i}^{-1} = \tilde{\Sigma}_{t_i}^{-1} + B_{t_i}^\top \Sigma_\nu^{-1} B_{t_i}$ and $m_{t_i} = \tilde{\Sigma}_{t_i}^{-1} \tilde{\mu}_{t_i} + B_{t_i}^\top \Sigma_\nu^{-1} Z_{1:i}$. Thus, $\text{E}[C_{t_i} \,|\, \Omega, Z_{1:i}] = M_{t_i} m_{t_i}$ is conveniently computed in sequential fashion starting with fixed values of $m_{t_0} = m_0$ and $M_{t_0} = M_0$ in the prior for C_{t_0} in (6). When the parameters in Ω are unknown and need to be estimated, we simulate samples (using MCMC) from $p(\Omega \,|\, Z_{1:i})$ and then compute $\text{E}_{\Omega \,|\, Z_{1:i}}\left[\text{E}\left\{C_{t_i} \,|\, \Omega, Z_{1:i}\right\}\right] = \text{E}\left[M_{t_i} m_{t_i} \,|\, Z_{1:i}\right]$ as a Monte Carlo average of $M_{t_i} m_{t_i}$ over the samples from $p(\Omega \,|\, Z_{1:i})$. Posterior samples from $p(C_{t_i} \,|\, Z_{1:i})$, if desired for each $t_i \in \mathcal{T}$, can also be easily drawn in sequential fashion using the above distributions drawing one C_{t_i} from $p(C_{t_i} \,|\, \Omega, Z_{1:i})$ for each sampled Ω.

For smoothing, we estimate the posterior expectation of each C_{t_i} given the entire set of observations from timepoints in $\mathcal{T}$, that is, $\text{E}[C_{t_i} \,|\, Z_{1:n}]$. Here, we devise an MCMC sampling algorithm that updates $\{C_{t_i}\}_{i=1}^n$, Σ_ω and Σ_ν using Gibbs updates from their respective full conditional distributions and we update θ_c from its full conditional distribution using Metropolis random walk steps. The full conditional distributions are $p(C_{t_i} \,|\, \cdot) = N(C_{t_i} \,|\, M_{t_i} m_{t_i}, M_{t_i})$ where $M_{t_i}^{-1} = \Sigma_\omega^{-1} + A_{t_{i+1}}(\theta_c)^\top \Sigma_\omega^{-1} A_{t_{i+1}}(\theta_c) + B_{t_i}^\top \Sigma_\nu^{-1} B_{t_i}$ and $m_{t_i} = \Sigma_\omega^{-1}(A_{t_i}(\theta_c) C_{t_{i-1}} + \delta_i g_{t_i}) + A_{t_{i+1}}^\top \Sigma_\omega^{-1}(C_{t_{i+1}} - \delta_{i+1} g_{t_{i+1}}) + B_{t_i}^\top \Sigma_\nu^{-1} Z_{t_i}$, $p(\Sigma_\nu \,|\, \cdot) = \text{IW}(\Sigma_\nu \,|\, r_{\nu|\cdot}, S_{\nu|\cdot})$ and $p(\Sigma_\omega \,|\, \cdot) = IW(\Sigma_\omega \,|\, r_{\omega|\cdot}, S_{\omega|\cdot})$, where $r_{\nu|\cdot} = r_\nu + n$, $S_{\nu|\cdot} = S_\nu + \sum_{i=1}^n (Z_{t_i} - B_{t_i} C_{t_i})(Z_{t_i} - B_{t_i} C_{t_i})^\top$, $r_{\omega|\cdot} = r_\omega + n$ and $S_{\omega|\cdot} = S_\omega + \sum_{i=1}^n (C_{t_i} - A_{t_i}(\theta_c) C_{t_{i-1}} - \delta_i g_{t_i})(C_{t_i} - A_{t_i}(\theta_c) C_{t_{i-1}} - \delta_i g_{t_i})^\top$. The full conditional distribution for θ_c is not a standard distribution and is proportional to $p(\theta_c) \times \prod_{i=1}^n N(C_{t_i} \,|\, A_{t_i}(\theta_c) C_{t_{i-1}} + \delta_i g_{t_i}, \Sigma_\omega)$. The smoothed estimates $\text{E}[C_{t_i} \,|\, Z_{1:n}]$ are obtained by simply averaging the posterior samples (post burn-in) of C_{t_i}.

The two-zone model has $p = 2$, while the one-compartment and eddy-diffusion models have $p = 1$. Gaussian Bayesian SSMs for $p = 1$ specify $P_{\nu,\theta_\nu} \equiv N(0,\sigma^2)$ and $P_{\omega,\theta_\omega} \equiv N(0,\tau^2)$. The measurement equation is linear in the state C_t. The $\text{IW}(\cdot,\cdot)$ priors in (6) are replaced by $\text{IG}(\sigma^2 \,|\, a_\sigma, b_\sigma)$ and $\text{IG}(\tau^2 \,|\, a_\tau, b_\tau)$.

Although Gaussian SSMs are popular in dynamic modeling of physical systems, especially due to convenient updating schemes, the Gaussian assumption for the concentration measurements may be untenable. Our second Bayesian SSM assumes that $Z_t = \log Y_t$ are log-concentration measurements and $f(C_t) = \log C_t$ in the measurement equation. We still specify P_{ν,θ_ν} as Gaussian, which means that Z_t's are log-normal and is probably a more plausible assumption than in Gaussian SSMs. In the transition equation, again the Gaussian assumption on ω_t seems implausible: if the measurements of the state are log-normal, then why should C_t be Gaussian? Since C_t is positive, a Gamma or log-normal specification for P_{ω,θ_ω} seems much more plausible. For $p = 2$, we will specify logarithmic bivariate normal distributions, while for $p = 1$ we will explore Gamma and log-normal densities. We will refer to all of these models as non-Gaussian Bayesian SSMs.

The turbulent eddy-diffusion model requires some further specifications. While the framework in (5) is rich, unfortunately it will not usually be applicable to practical industrial hygiene settings because typically very few measurements are available over distinct locations in a workplace chamber and estimating the processes will be unfeasible. Hence, we will need simpler specifications. For example, we can consider a setting with locations $\{s_1, s_2, \ldots, s_m\}$ and n time-points. We fit the model in (5) with $Z_t(s_i) = \log Y_t(s_i)$ are log-concentration measurements

and $f(C_t(s_i)) = \log C_t(s_i)$. We further specify P_{η_t,θ_η} as a white-noise process, that is, $\eta_t \overset{\text{iid}}{\sim} N(0,\tau^2)$ for every t and s, and $P_{\nu_t(s),\theta_\nu}$ is a temporally indexed spatial Gaussian process with an exponential covariance function, independent across time. This means that the $m \times 1$ vector $\nu_t \overset{\text{ind}}{\sim} N(0,\sigma_t^2 R_t(\phi_t))$, where $R_t(\phi_t)$ is an $m \times m$ matrix with (i,j)th element $\exp(-\phi_t d_{ij})$ and $d_{ij} = \|s_i - s_j\|$.

In theory $P_{\nu_t(s),\theta_\nu}$ can be a continuous-time spatial-temporal process specified through a space-time covariance function (see, e.g., Banerjee, Carlin, and Gelfand 2014). Alternatively, we could treat time as discrete and evolving, for each location s, as an autoregressive process so that $\nu_t(s) = \gamma \nu_{t-1}(s) + \eta_t(s)$ with $\eta_t(s)$ being a spatial process independent across time (see, e.g., Wikle and Cressie 1999; Gelfand, Banerjee, and Gamerman 2005). One could continue to embellish the model in (5) using spatial-temporal structures that represent richer hypotheses and more flexible modeling. However, in industrial hygiene applications such specifications will rarely lead to estimable models given the scarcity of data points. Most realistic settings will provide measurements from only a handful of locations (e.g., $m \sim 5$) and some moderate numbers of time points (e.g., $n \sim 100$). Hence, we will not explore these specifications any further. Moreover, even when we assume independence across time it will be difficult to estimate models with time-varying spatial process parameters. Hence, we let $\nu_t \overset{\text{iid}}{\sim} N(0,\sigma^2 R(\phi))$ so that each $m \times 1$ vector ν_t has the same m-variate Gaussian distribution.

To compare among competing models, we adopt a posterior predictive loss approach (Gelfand and Ghosh 1998), which quantifies how tenable a given model is with respect to the observed data. This approach is composed of two components, one measuring goodness of fit and another penalizing for model complexity. We generate replicated datasets that would be predicted by the model with the values of the parameters estimated from the observations. The uncertainty in the model's parameter estimates is propagated to the replicated datasets. To be specific, we generate samples from the posterior predictive distribution for each data point by sampling from $Z_{\text{rep},i} \sim p(Z_{\text{rep},i} \mid Z_{1:n}) = \int p(Z_{\text{rep},i} \mid \Omega, \{C_{t_i}\}) p(\Omega, \{C_{t_i}\} \mid Z_{1:n}) d\Omega$ for $i = 1, 2, \ldots, n$, where $\{C_{t_i}\}$ is the collection of latent concentrations over the entire time frame, and $Z_{\text{rep},i}$ is a random variable having the same probability distribution as the observed data point Z_{t_i} given Ω and $\{C_{t_i}\}$. Thus, the replicated observations follow the posterior predictive distribution (see, e.g., Gelman et al. 2013, for more on sampling from the posterior predictive distributions) at the observed timepoints.

We will compute the posterior predictive mean, $\mu_{\text{rep},i} = \text{E}[Z_{\text{rep},i} \mid Z_{1:n}]$, and dispersion, $\Sigma_{\text{rep},i} = \text{var}[Z_{\text{rep},i} \mid Z_{1:n}]$, for each $Z_{\text{rep},i}$; these are easily calculated from the posterior samples for each $Z_{\text{rep},i}$. We will prefer models that will perform well under a decision-theoretical *balanced loss function* that penalizes departure of replicated means from the corresponding observed values (lack of fit), and also penalizes model complexity by accounting for variability in model parameters. Using a squared error loss function, the measure for goodness of fit is evaluated as $\text{G} = \sum_{i=1}^{n} \|Z_{t_i} - \mu_{\text{rep},i}\|^2$, where $\|\cdot\|$ denotes the Euclidean norm (two-dimensional for the two-zone; absolute difference for one-zone and eddy diffusion), while the penalty for variability in replicates is given by $\text{P} = \sum_{i=1}^{n} \text{Tr}(\Sigma_{\text{rep},i})$, where $\text{Tr}(A)$ denotes the trace of the matrix A. Models with poorer fits will tend to have higher G and those with less reliable replicates will usually have increased P. We will use the score $\text{D} = \text{G} + \text{P}$ as a model selection criteria, with lower values of D indicating better models. This measure is based upon a sound decision-theoretical principle (see, e.g., Gelfand and Ghosh 1998) and, unlike information criterion such as AIC or BIC, does not rely upon asymptotic justifications.

4. Data Analysis

In this section, we evaluate the performance of the models discussed in Section 3, for the three physical exposure models illustrated in Section 2, using computer-simulated datasets as well as experimental lab-generated data. We generate computer-simulated data by adding noise to the exact solutions of the physical model. The computer-simulated data were generated using the R computing environment. The lab-generated data experiments were conducted in test chambers. Arnold, Shao, and Ramachandran (2017) examined parts of this data using the deterministic one-zone and two-zone models and showed that performance is highly sensitive to the modeling assumptions and knowing the generation (G) and ventilation (Q) rates. Shao et al. (2017) studied the eddy diffusion data using a deterministic model and concluded that it is suitable for indoor spaces with persistent directional flow toward a wall boundary, as well as in rooms where the airflow is solely driven by mechanical ventilation (no natural ventilation involved). These results imply the need for a more flexible model that accounts for uncertainty and also be used for parameter inference.

4.1. Prior Settings

In Bayesian exposure models, reasonable informative priors are usually used, based on expert knowledge and physical considerations (Monteiro, Banerjee, and Ramachandran 2014). We assigned informative priors on the generation rate G, ventilation rate Q, loss rate K_L, airflow rate β, and diffusion coefficient D_T using uniform distributions for the plausible values of the parameters.

For the simulation data, uniform priors were assigned within at least 20% of the true values following the prior settings in Monteiro, Banerjee, and Ramachandran (2011). In the one-zone and two-zone models, we assume that $G \sim \text{Unif}(281, 482)$, $Q \sim \text{Unif}(11, 17)$, $K_L \sim \text{Unif}(0, 1)$, and $\beta \sim \text{Unif}(0, 10)$ in the two-zone model and $D_T \sim \text{Unif}(0, 3)$ in the eddy diffusion model. For the exponential covariance function, the spatial range is given by approximately $3/\phi$ which is the distance where the correlation drops below 0.05. The prior on $\phi \sim \text{Unif}(0.5, 3)$ implies that the effective spatial range, that is, the distance beyond which spatial correlation is negligible, is between 1 and 6 m.

Wider ranges for the prior distributions were considered in the lab-generated data analysis because the exact true values for

some of the parameters were unknown. The ranges of the true values in the well-mixed compartment and two-zone models for G, Q, K_L, and β are (40–120) (mg/min), (0.04–0.77) (m^3/min), <0.01, and (0.24–1.24)(m^3/min), respectively. We assume that $G \sim \text{Unif}(30, 150)$, $Q \sim \text{Unif}(0, 1)$, and $K_L \sim \text{Unif}(0, 1)$ in the one-zone and two-zone models and $\beta \sim \text{Unif}(0, 5)$ in the two-zone model. For the eddy diffusion model, the true value for G is 1318 (mg/sec) and from the literature (Shao et al. 2017) the range for D_T is (0.001–0.2) m^2/sec. Hence, we assigned priors of $G \sim \text{Unif}(1104, 1650)$ and $D_t \sim \text{Unif}(0, 1)$. Weakly informative priors using $IW(3, I)$ were assigned to the variance covariance matrices Σ_ω and Σ_ν.

4.2. Simulation Results

We generate computer-simulated data by fixing parameters in a physical model and adding Gaussian noise to the exact solution of the respective deterministic equation. We fix the deterministic model parameters based upon physical considerations similar to the experiments described in Zhang et al. (2009). Following Zhang et al. (2009), we assumed that the workplace chamber was 1.73 m long, 1.27 m wide, and 1.73 m high, yielding a volume $V = 3.8$ m^3. The one-zone model also assumed that toluene was released at a rate of $G = 351.5$ mg/min, the average flow-rate was $Q = 13.8$ m^3/min and the loss rate was $K_L = 0.1$. For the two-zone model, the near field represents the region very near and around the source and its volume contains the breathing zone of the worker and is equal to half of the volume of a sphere with radius 0.2 m (i.e., $V_N = 10^{-2} \times \pi$), and the volume of the far field is equal to the difference between the volume of the room (3.8 m^3) and the volume of the near field. In two-zone experiments, usually the airflow flow rate is not measured directly, but estimated using the steady-state solution. Based upon Zhang et al. (2009), we assumed that $\beta \to \frac{G}{C_N - C_F} \approx 5$ m^3/min. For the eddy diffusion data, an assigned value of $D_T = 1$ m^2/min agrees with the values in literature reported in Shao et al. (2017). Estimation of physical model parameters and the latent concentration process, and subsequent model assessment are conducted as described in Section 3.

4.2.1. One-Zone Model

We simulated 50 independent datasets, each with $T = 100$ exposure concentrations at equally spaced time points, using the exact solution to the ODE in (1). The measurements $\{y_t, t = 1, \ldots, T\}$ were generated by adding random Gaussian noise with zero mean and variance 0.1 to values obtained from the exact solution at time $t = 1, \ldots, T$. The initial concentration $C(0)$ was assigned a value of 1 mg/m^3. Theoretically, the steady state concentration in the simulated setting is ≈25 mg/m^3. We applied the Gaussian and non-Gaussian SSM models to the synthetic data and compared our results to the simple Bayesian nonlinear regression model (BNLR) proposed by Zhang et al. (2009). The Gaussian SSM in (6) assumes linearity and Gaussian errors, where $A_t(\theta_c) = \left(1 - \delta_t \frac{Q + K_L V}{V}\right)$ and $g = \delta_t \frac{G}{V}$.

Table 1 presents the posterior medians and the 95% credible intervals for the model parameters along with the MSE and the D = G + P score for different models corresponding to one of the simulated datasets. The table also presents the coverage probabilities of the Bayesian credible intervals—calculated as the percentage of cases in which the 95% credible intervals include the true parameter value—and the average lengths of the credible intervals over the 50 replicated datasets. Panels "a" and "b" in Figure 4 present filtered results from the Gaussian and non-Gaussian SSMs, a panel titled "BNLR" presents the model-fitted results from the Bayesian nonlinear regression in Zhang et al. (2009), and a fourth panel titled "Smoothing" presents smoothed results from the non-Gaussian SSM (the best-fitting model). Each panel plots the simulated concentration measurements and the true value from the one-zone ODE, as indicated by "Measurements" and "True." Panels "a," "b," and BNLR present estimates of the posterior means of the (latent) concentrations conditional on all measurements up to the time-points indicated in the x-axis (filtering, indicated as "Estimated"), while the corresponding curve in "Smoothing" presents the estimates of the posterior means of the latent concentrations conditional on all measurements spanning the entire dataset (including before and after the indicated time-points in the x-axis).

A summary of the performances of the different models are as follows:

- Non-Gaussian SSM: The credible intervals include the true values for all the parameters except K_L. The latent state estimates are very close to the true simulated values as shown in Figure 4.
- Gaussian SSM: The credible intervals for the generation rate G and the ventilation rate Q include the true values. The interval for the loss rate K_L does not cover the true parameter value. The model estimates for the latent states are closer to the observed values than the true values, that is, it produced noisy estimates for the state process.

Table 1. Posterior medians, 95% credible intervals, coverage probabilities, and average interval lengths based on the 50 simulated datasets, posterior predictive loss (D = G + P), and MSEs for the non-Gaussian and Gaussian SSMs and the BNLR for measurements simulated using the one-zone physical model.

Parameter	Non-Gaussian SSM	Gaussian SSM	BNLR
G(351.5)	326.8(283.3, 351.7)	363.5(314.2,413.8)	301.9(282.0,347.3)
Coverage and average length	(84%, 64.0)	(78%, 131.0)	(22%, 66.0)
Q(13.8)	12.9(11.1, 14.8)	12.8(11.4, 14.3)	13.0(13.2,16.9)
Coverage and average length	(96%, 4.0)	(94%, 4.7)	(98%, 3.5)
K_L(0.1)	0.34(0.19,0.78)	0.30(0.28, 0.4)	0.35(0.06,0.78)
Coverage and average length	(98%, 0.7)	(4%, 0.02)	(98%, 0.7)
D = G + P	312.2 = 5.9 + 306.3	435.8 = 232.8 + 203.0	727.9 = 371.0 + 356.8
MSE	0.07	2.3	0.3

NOTE: The posterior medians, credible intervals, D score, and MSEs are representatives from a single simulated dataset.

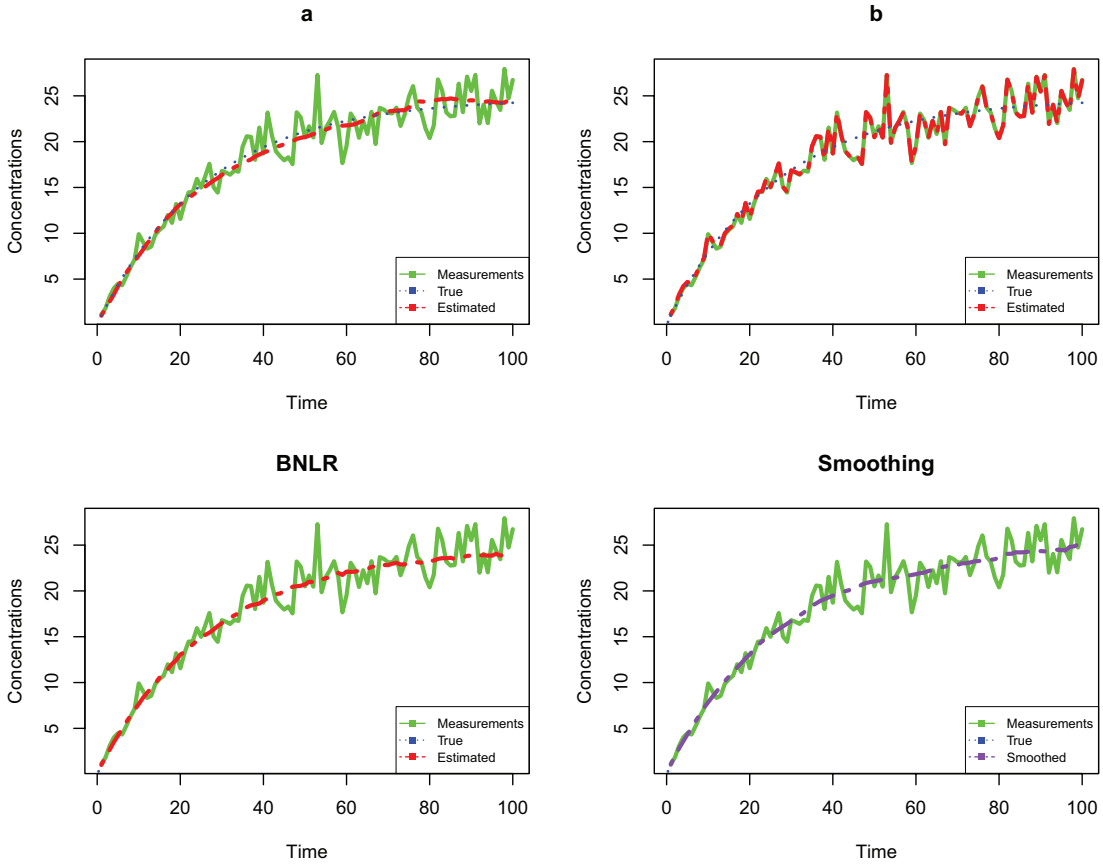

Figure 4. Plots of the simulated measurements, the true concentrations from the physical model, and estimated concentration profiles using filtering. Panels "a" and "b" correspond to the non-Gaussian and Gaussian SSMs, respectively. "Smoothing" presents smoothed concentration estimates from the non-Gaussian SSM. Each panel plots the measurements and the true values from the simulated one-zone ODE for comparisons.

- BNLR: The credible intervals include the true values for all the parameters except for G. The model estimates for the latent states are close to the true values.

The D scores show that the non-Gaussian SSM is preferable to the Gaussian and BNLR models. This is corroborated by the coverage probabilities of Bayesian credible intervals: the percentage of cases among the 50 simulations in which the credible intervals include the true values of the parameters indicate that the non-Gaussian SSMs empirical coverages tend to be much closer to the nominal 95% compared to the Gaussian SSM and the BNLR model. This implies that the non-Gaussian SSM offers better posterior region calibration (Syring and Martin 2018).

4.2.2. Two-Zone Model

We simulated 50 independent datasets, each with $T = 200$ exposure concentrations at the near and far fields at equally spaced time points using the exact solution to the ODE in (3). Random noise was generated from a bivariate Gaussian distribution with zero means and unit variances for each of the two fields and zero correlation between them. The noise was added to the log of the true values to produce measurements $\{y_t, t = 1, \ldots, T\}$. The initial concentrations $C_N(0)$ and $C_F(0)$ were assigned values 0 and 0.5 mg/m^3, respectively. Theoretically, the steady state concentrations are ≈95 mg/m^3 and ≈25 mg/m^3 in the near and far fields, respectively. The Gaussian SSM in (6) assumes linearity and Gaussian errors, so that $A_t(\theta_c) = \delta_t A + I$ and $g = \delta_g$.

Table 2 presents the posterior summaries for the two-zone model analogous to those described in Table 1. Figure 5 has panels corresponding to both near and far fields from the two-zone model. The labels and descriptions of the panels are analogous to those described in Figure 4.

A summary of the performances of the three models are as follows:

- Non-Gaussian SSM: The credible intervals include the true values for all the parameters. The estimates of the latent states are close to the true values in both the near and far fields as shown in Figure 5.
- Gaussian SSM: The credible intervals for all the parameters except the ventilation rate Q and the flow rate β do not include the true values. The estimates of the latent states tend to be closer to the true values in the near field than in the far field. The estimates tend to also exhibit wider credible intervals.
- BNLR: The credible intervals include the true values for all the parameters. The estimates for the latent states are closer to the true values at in the near field than in the far field

The D = G + P scores indicate that the non-Gaussian SSM is preferred to the BNLR and the Gaussian SMM. The MSEs and Figure 5 confirm these results. The coverage probabilities

Table 2. Posterior medians, 95% credible intervals, coverage probabilities, and average interval lengths based on the 50 simulated datasets, posterior predictive loss (D = G + P), and MSEs for the non-Gaussian and Gaussian SSMs and the BNLR for measurements simulated using the two-zone physical model.

Parameter	Non-Gaussian SSM	Gaussian SSM	BNLR
$G(351.5)$	347.3(315.6,379.3)	450.5(395.2, 480.2)	335.1(302.5,382.6)
Coverage and average length	(98%, 41.0)	(90%, 102.0)	(82%,87.5)
$Q(13.8)$	14.7(12.1,16.8)	13.5(11.1, 16.7)	14.4(11.2, 15.8)
Coverage and average length	(96%, 3.6)	(96%, 5.0)	(62%, 3.5)
$K_L(0.1)$	0.38(0.02,0.78)	0.22(0.16,0.35)	–
Coverage and average length	(100%, 0.7)	(86%, 0.7)	–
$\beta(5)$	5.0(4.3,5.8)	4.8(3.3, 5.8)	5.1(4.0, 6.8)
Coverage and average length	(100%,0.6)	(88%,1.4)	(98%, 3.6)
D = G + P	1,049,840 = 1,010,905 + 38,934.0	1,118,550 = 1,033,428 + 85,121.7	2,504,429 = 1,359,016 + 1,145,413
MSE	15.3	116.1	54.9

NOTE: The posterior medians, credible intervals, D score, and MSEs are representatives from a single simulated dataset.

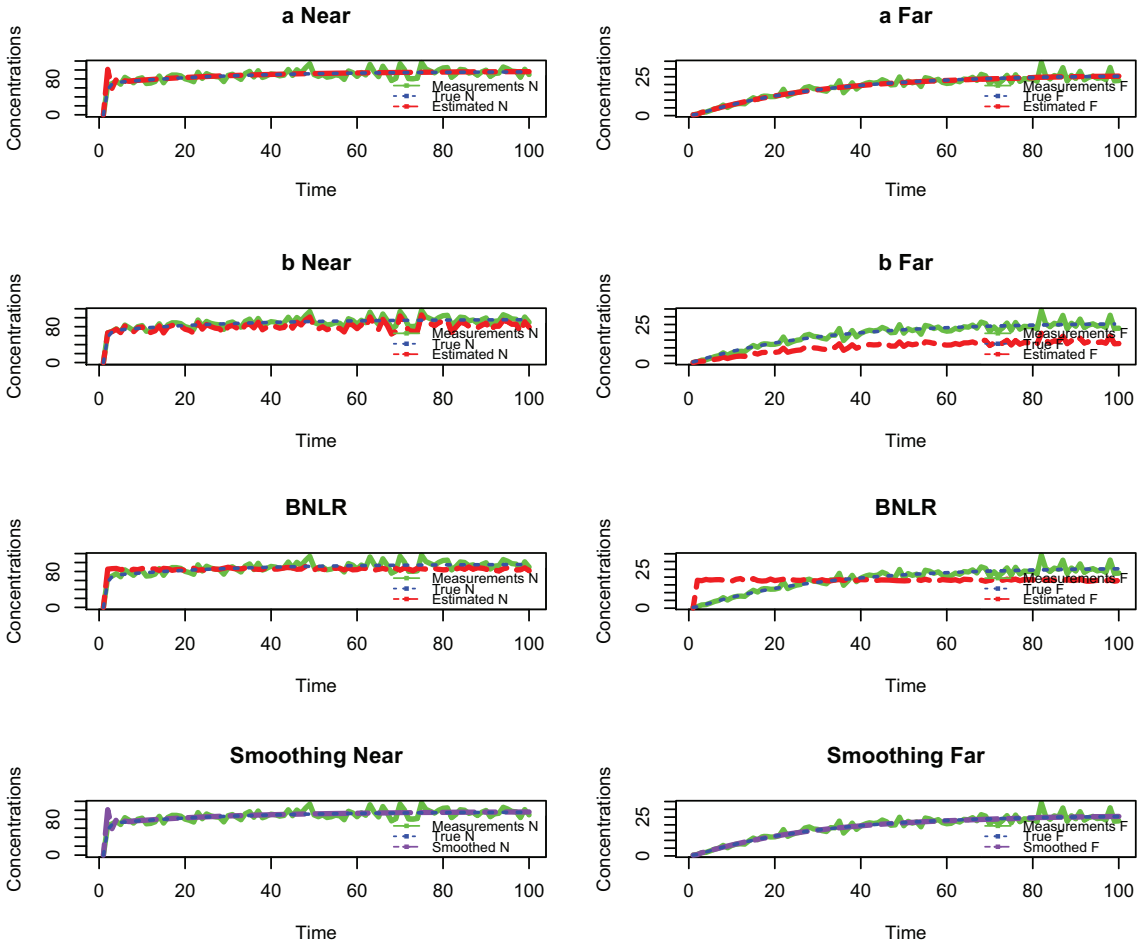

Figure 5. Plots of the simulated measurements, the true concentrations from the physical model, and estimated concentration profiles using filtering. Panels "a" and "b" correspond to non-Gaussian and Gaussian SSMs, respectively, "Near" and "Far" indicates the respective fields. "Smoothing" presents smoothed concentration estimates for the near and far fields from the non-Gaussian SSM. Each panel plots the measurements and the true values from the simulated two-zone ODE for comparisons.

calculated from the 50 simulated datasets indicate that the non-Gaussian SSMs coverages tend to be the closest to the nominal 95%. In addition, the credible intervals tend to be shorter than those from the Gaussian and BNLR models.

4.2.3. *Turbulent Eddy Diffusion Model*

We simulated 50 independent datasets, each comprising 5 different locations and 100 equally spaced time points over which $5 \times 100 = 500$ concentrations were generated using the exact model given in (4). Random Gaussian noise with zero mean and unit variance was added to the log of the generated concentrations to yield measurements $\{y_t : t = 1, 2, \ldots, T\}$. Table 3 describes posterior summaries of the model parameters, model comparison metrics, and coverage probabilities and average credible interval lengths analogous to those in Tables 1 and 2. Figure 6 has panel labels analogous to Figures 4 and 5 but are presented for three different locations indicated as L1 (top row), L2 (middle row), and L3 (bottom row). Figure 7 is an interpolated image of the posterior mean surface of the latent spatial process $v_t(s)$. The plot indicates

Table 3. Posterior medians, 95% credible intervals, coverage probabilities, and average interval lengths based on the 50 simulated datasets, posterior predictive loss (D = G + P), and MSEs for the non-Gaussian and Gaussian SSMs and the BNLR for measurements simulated using the turbulent eddy-diffusion model.

Parameter	Non-Gaussian SSM	Gaussian SSM	BNLR
$G(351.5)$	355.9(284.0,477.5)	449.6(301.0,480.5)	376.5(281.0,480.0)
Coverage and average length	(99%,184.0)	(26%,32.0)	(28%,50.0)
$D_T(1)$	1.2(0.9,1.5)	1.4(1.3,1.6)	1.14(1.03, 1.8)
Coverage and average length	(88%, 1.2)	(14%,0.1)	(28%,0.1)
D = G + P	7062.4 = 1564.5 + 5497.9	22,025.7 = 1112.5 + 20,913.1	27,719.1 = 14,529.7 + 13,189.5
MSE	3.11	5.55	4.22

NOTE: The posterior medians, credible intervals, D score, and MSEs are representatives from a single simulated dataset.

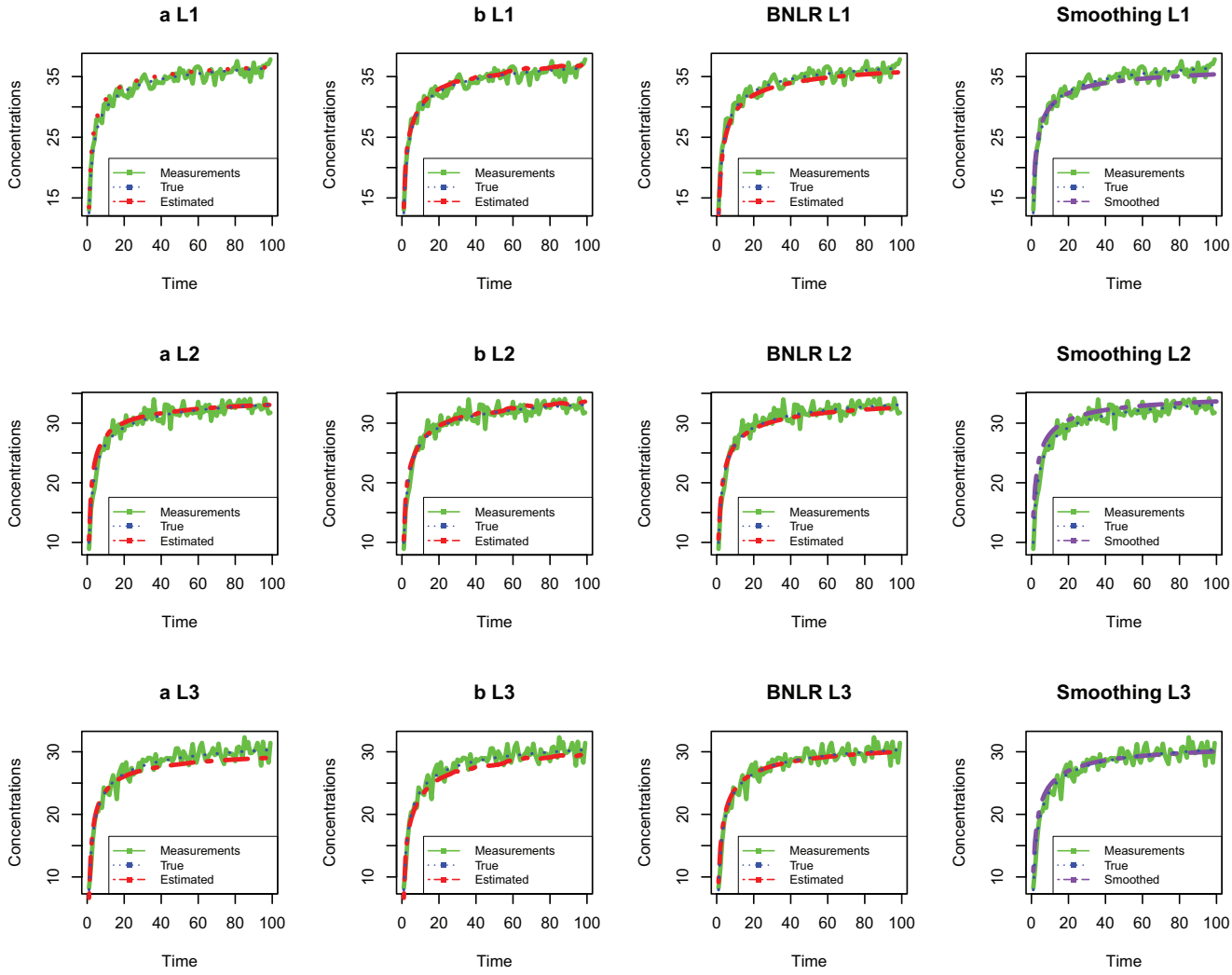

Figure 6. Plots at three locations (L1, L2, and L3) of the simulated measurements, the true concentrations from the physical model, and estimated concentration profiles using filtering. Panel labels "a" and "b" correspond to non-Gaussian and Gaussian SSMs, respectively, while "Smoothing" presents smoothed concentration estimates. Each panel plots the measurements and the true values from the simulated turbulent eddy-diffusion ODE for comparisons.

higher concentration values near the source of emission at the bottom-left corner and lower values away from the source, which is what the turbulent eddy diffusion model posits.

The performance of the three models are summarized as follows:

- Non-Gaussian SSM: The credible intervals include the true values for all the parameters. The estimates of the latent states approximate the true values very well at the five locations.
- Gaussian SSM: The credible intervals include the true value for the generation rate G, but not for the eddy diffusion coefficient D_T. The model estimates for the latent states are closer to the observed values than the true values.
- BNLR: The credible intervals do not include the true value for the eddy diffusion coefficient D_T. The model estimates for the latent states are close to the true values.

The non-Gaussian SSM again seems to produce more accurate parameters estimates. The D scores also prefer the non-Gaussian SSM to either the BNLR or the Gaussian SSM. The

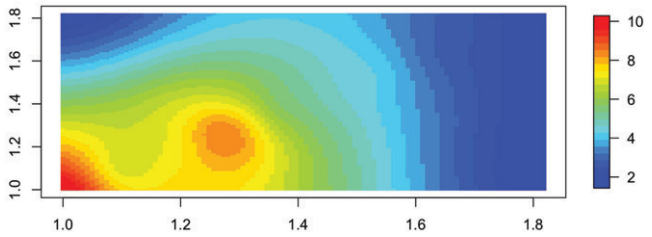

Figure 7. Interpolated surface of the mean of the random spatial effects posterior distribution.

MSE and Figure 6 further affirm these results. Coverage of the credible intervals also seem to be closest to the theoretical 95% for the non-Gaussian SSM, but the average length of the intervals tend to be wider. Thus, the non-Gaussian SSM seems to be more conservative with its parameter estimates for the turbulent eddy-diffusion model.

4.3. Experimental Chamber Data Results

In this section, we study the performance of the non-Gaussian and Gaussian SSMs on controlled lab-generated data in which solvent concentrations have been measured under different scenarios. We are interested in the inference through the posterior distributions of the parameters Q and G in the one-zone model, in addition to β in the two-zone model, and G and D_T in the eddy diffusion model.

4.3.1. One-Zone Model

A series of studies were conducted in an exposure chamber under different controlled conditions. Arnold, Shao, and Ramachandran (2017) constructed a chamber of size ($2.0\,\text{m} \times 2.8\,\text{m} \times 2.1\,\text{m} = 11.8\,\text{m}^3$), where two industrial solvents (acetone and toluene) were released using different generation G (mg/min) and ventilation Q (m^3/min) rates. In particular, three levels of ventilation rates corresponding to ranges of 0.04–0.07 m^3/min, 0.23–0.27 m^3/min and 0.47–0.77 m^3/min were used. The loss rate K_L was determined from empirical studies to be <0.01. Solvent concentrations were measured every 1.5 min. Details of the experiments can be found in Arnold, Shao, and Ramachandran (2017).

Table 4 presents the medians and 95% Bayesian credible intervals from the MCMC posterior samples in addition to the D = G + P scores. The non-Gaussian SSMs credible intervals cover the true values for both G and Q, while the Gaussian SSMs intervals include the true values for G at low and high ventilation levels. BNLRs intervals also include the true values for G at high ventilation levels and Q at all levels. Posterior predictive loss (D = G + P) indicates better fit for the non-Gaussian SSM model followed by the Gaussian model and finally the BNLR. Figure 8 confirms these results.

4.3.2. Two-Zone Model

The near field box of size ($0.51\,\text{m} \times 0.51\,\text{m} \times 0.41\,\text{m} = 0.105\,\text{m}^3$) was constructed within the far field box (Arnold, Shao, and Ramachandran 2017). The volume of the far field is $11.79\,\text{m}^3$, which is the chamber volume minus the near field volume. The airflow parameter β cannot be directly measured, but it was estimated from the local air speed to range from 0.24 to $1.24\,\text{m}^3$/min. Similar to the one-zone model, three different experimental datasets at three different ventilation levels were used. Table 5 shows the posterior medians and 95% credible intervals, and the D = G + P score. The non-Gaussian SSMs 95% credible intervals include the true (experimentally set) values of Q at medium and high ventilation rates, while they include the true values of G only at the medium ventilation rate. The Gaussian SSMs 95% credible intervals cover the true value of Q only at medium ventilation, but not the generation rates G in any of the three ventilation settings. The BNLRs intervals cover the true value of Q at a high ventilation level, but not the G . The true value for β was not directly measured and hence is unknown, however, it was estimated to be between 0.24 and 1.24. In general, the non-Gaussian SSMs estimate for β is closer to experimental range. The D = G + P scores clearly indicate that the non-Gaussian SSM outperforms the BNLR and the Gaussian SSM. These are also further affirmed in Figure 9.

4.3.3. Turbulent Eddy Diffusion Model

Shao et al. (2017) constructed a chamber of size ($2.8\,\text{m} \times 2.15\,\text{m} \times 2.0\,\text{m} = 11.9\,\text{m}^3$), where toluene was released. Measurements were taken at two locations at distances 0.41 m and 1.07 m away from the source every 2 min. Due to the limited spatial information from the two locations, an unstructured covariance for $v_t(s)$ was used instead of the geostatistical exponential covariance specified in the simulation experiments. A weakly informative prior was assigned to the covariance matrix using $IW(3, I)$ (Gelman et al. 2013).

Table 6 shows the posterior medians and 95% credible intervals, and the D = G + P scores. The value of D_T is difficult to measure; hence, the true value is unknown. However, Shao et al.

Table 4. Posterior predictive loss (D = G + P), medians, and 95% credible intervals from the posterior samples of the one-zone model parameters using toluene and acetone solvents.

Parameter	Ventilation level	True value	Non-Gaussian SSM	Gaussian SSM	BNLR
G	Low	43.2	38.1(30.2,62.9)	35.3(30.2, 46.7)	30.1(30.0,30.4)
	Medium	43.2	45.06(30.5,101.9)	72.9(45.6,94.9)	30.9(30.0,34.2)
	High	39.55	81.7(32.9,142.4)	38.1(30.5,51.4)	36.1(30.2,67.6)
Q	Low	0.04–0.07	0.27(0.02, 0.41)	0.20(0.15,0.27)	0.07(0.003,0.19)
	Medium	0.23–0.27	0.50(0.02,0.97)	0.15(0.10,0.21)	0.57(0.02,0.94)
	High	0.47–0.77	0.59(0.03,0.98)	0.30(0.23,0.45)	0.5(0.03,0.97)
D = G + P	Low		129.4 = 88.8 + 40.6	208.0 = 4.3 + 203.7	52,257.53 = 36,044.83 + 16,212.71
	Medium		9.8 = 0.52 + 9.2	77.7 = 0.20 + 77.1	16,256.04 = 3040.128 + 13,215.91
	High		7.5 = 1.0 + 6.5	38.2 = 0.1 + 38.1	4345.8 = 237.4 + 4108.4

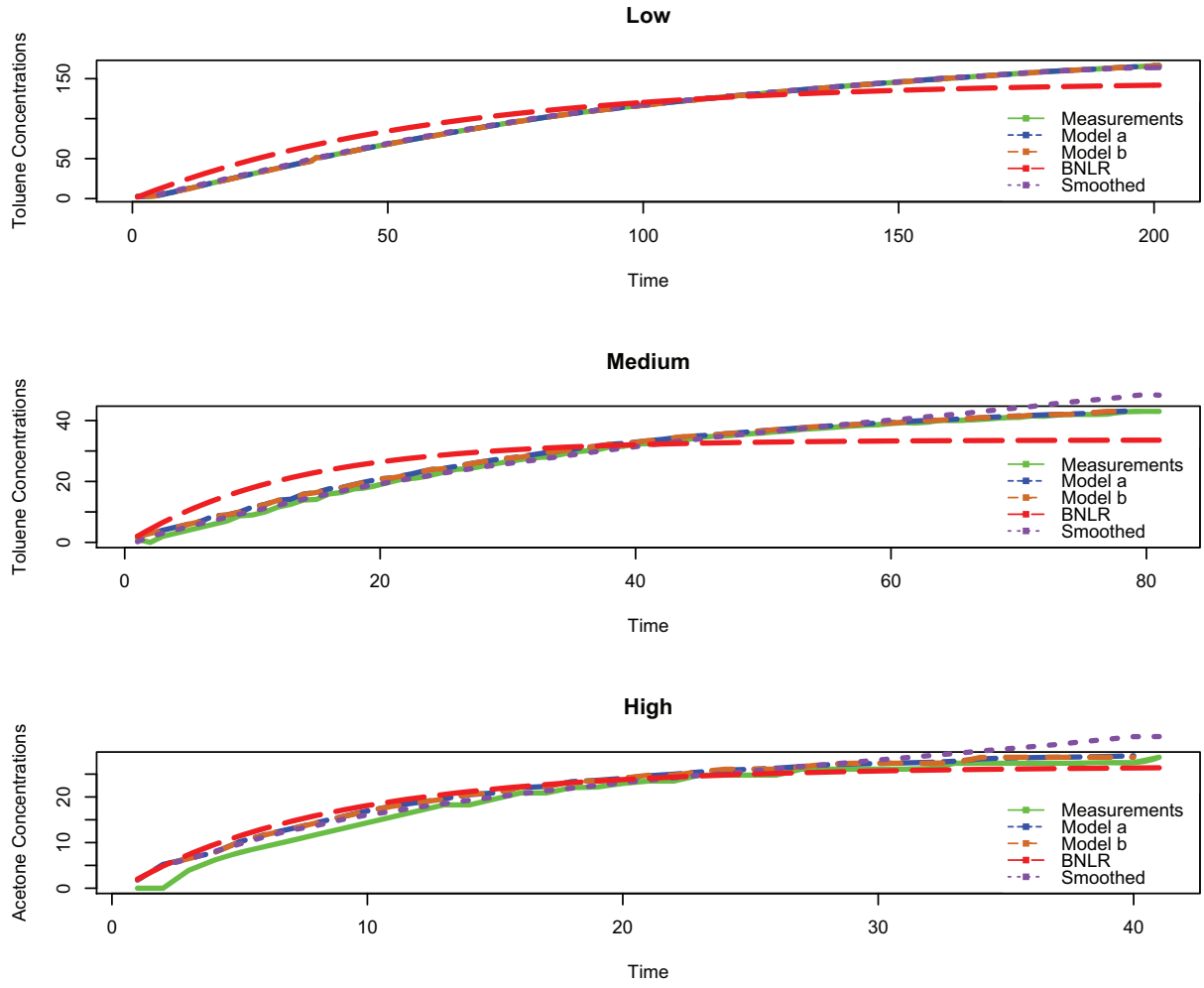

Figure 8. Plots of the measured concentrations and the posterior means of the latent states conditional on the measurements for (a) Non-Gaussian SSM; (b) Gaussian SSM; and BNLR.

Table 5. Posterior predictive loss (D = G + P), posterior medians, and 95% credible intervals for the two-zone model parameters using toluene and acetone solvents.

Parameter	Ventilation level	True value	Non-Gaussian SSM	Gaussian SSM	BNLR
G	Low	43.2	30.4(30.0, 32.2)	115.8(88.9, 143.9)	28.1(28.0,28.4)
	Medium	86.4	73.7(60.2,90.5)	141.6(130.6,149.7)	28.5(28.0,30.8)
	High	120.7	49.8(33.9,68.3)	132.9(121.6,148.0)	43.7(37.8,50.3)
Q	Low	0.04–0.07	0.68(0.09, 0.98)	0.28(0.23,0.36)	0.62(0.60,0.65)
	Medium	0.23–0.27	0.38(0.11,0.50)	0.25(0.20,0.31)	0.38(0.29,0.50)
	High	0.47–0.77	0.46(0.45,0.98)	0.14(0.11,0.16)	0.5(0.30,0.64)
β	Low	0.24–1.24	3.0(2.3,3.7)	5.1(4.1,6.0)	4.9(4.7,5.0)
	Medium	0.24–1.24	2.9(2.5, 3.4)	2.3(2.0,2.8)	4.5(3.4,5.0)
	High	0.24–1.24	2.2(1.5, 2.8)	2.5(2.0,3.0)	4.1(2.7,4.9)
D = G + P	Low		5653 = 189 + 5464	554,650 = 554,234 + 416	248,358 = 73,006 + 175,352
	Medium		22,262 = 10,596 + 11,666	850,014 = 424,452 + 425,562	93,267 = 16,824 + 76,443
	High		20,941 = 4345 + 16,596	479,098 = 240,278 + 238,820	119,212 = 64,968 + 54,244

(2017) demonstrated that most of the reported values of D_T in the literature range from 0.001 to 0.01 m^2/sec. The 95% credible intervals for D_T in the non-Gaussian SSM lie within that range. The 95% credible interval for G includes the true value. The 95% credible intervals from the Gaussian SSM do not include any of the true parameter values. The BNLRs estimates of G does not include the true value and the range for D_T is very narrow. Figure 10 shows that the latent state estimates for both SSMs are closer to the measurements in the first location than in the second location. The BNLR model is clearly biased and that is illustrated in the D score and in Figure 10. D = G + P scores show that the non-Gaussian SSM provides a better fit.

5. Discussion

We have proposed a Bayesian framework for analyzing experimental exposure data specific to industrial hygiene. This approach combines information from physical models

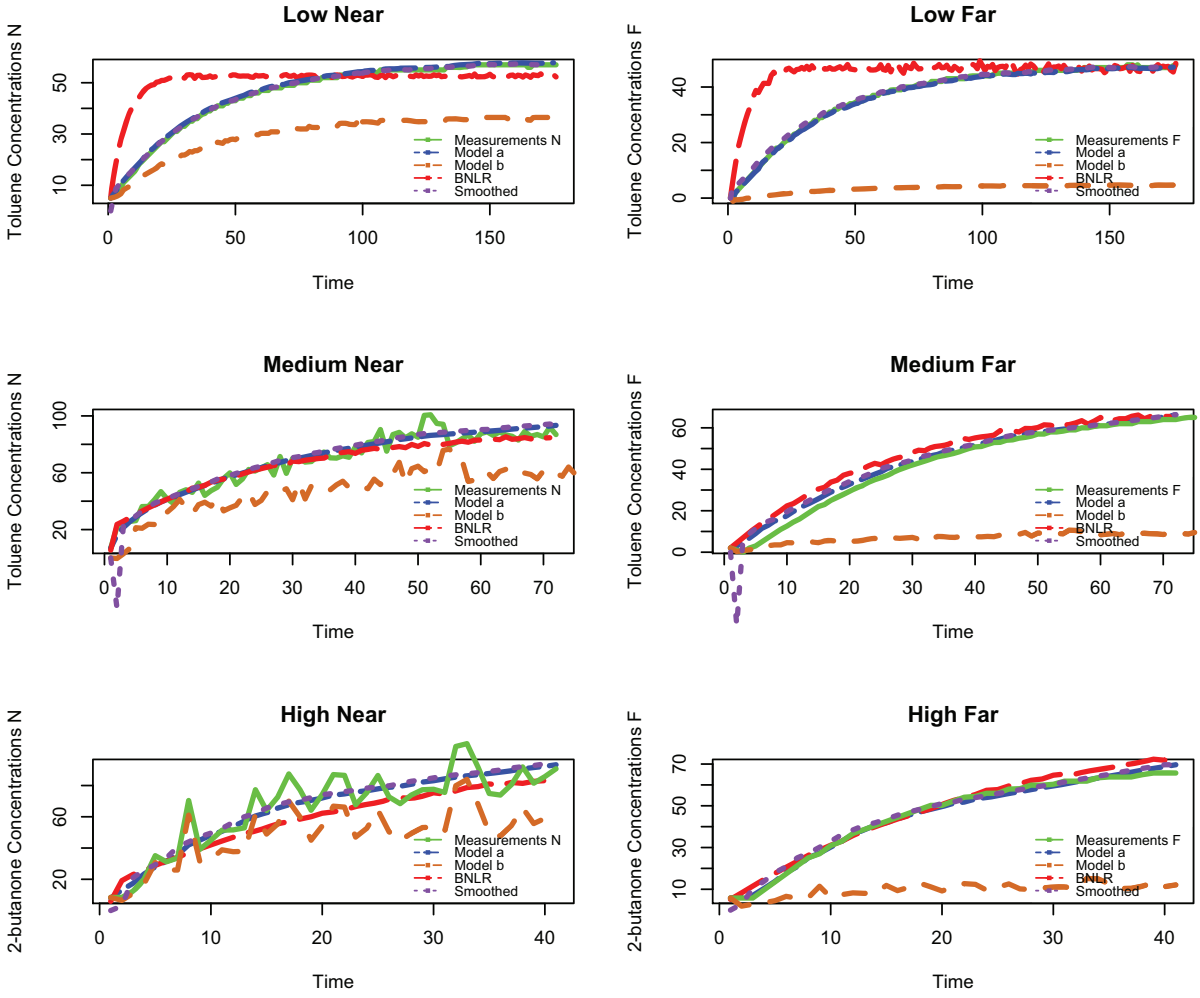

Figure 9. Plot of the measured concentrations and the posterior mean of the latent states conditional on the measurements in the near and far fields for (a) Non-Gaussian SSM; (b) Gaussian SSM; and BNLR.

Table 6. Posterior predictive loss (D = G + P), medians, and 95% CI of the posterior samples of the turbulent eddy diffusion model parameters using toluene.

Parameter	True value	Non-Gaussian SSM	Gaussian SSM	BNLR
G	1318.33	1207.3(1107.2,1371.7)	1118.7(1104.5,1294.3)	1108.4(1104.1,1127.7)
D_T	0.001–0.01	0.007(0.006,0.008)	0.67(0.64,0.78)	0.008(0.008,0.008)
D = G + P		100,877.8 = 59,369.9 + 41,507.9	3,664,659 = 3,660,710 + 3949.3	6,458,521 = 6,289,785 + 168,735.6

of industrial hygiene, observed data and prior knowledge of the physical system. We derive a likelihood by discretizing the physical models and subsequently relax the Gaussian noise assumptions, so that industrial hygienists will not be restricted to Gaussian SSMs.

In practical industrial hygiene settings, Gaussian SSMs are still often used as approximations to analyze possibly non-Gaussian data. To do so, some possibly inappropriate accommodations may need to be made. For example, Hoi, Yuen, and Mok (2008) allowed negative values in estimating PM_{10} concentrations, while Leleux et al. (2002) used KFs to predict gas concentrations by using a tuning parameter to fix σ_ω^2 and σ_v^2 in a one dimensional autoregressive exposure model, rather than pursuing full statistical inference. Our simulation experiments and results demonstrate that Gaussian SSMs may yield extremely poor fits when data are non-Gaussian. This was especially evident for the two-zone analysis. Our results, we hope, will inform the industrial hygiene community about some of the pitfalls of Gaussian SSMs.

Non-Gaussian SSMs tended to perform better than linear Gaussian SSMs, a result that appeared to be consistent across different exposure models and different experimental conditions. Moreover, our analysis revealed that the discretized non-Gaussian models outperform the BNLR method proposed by Zhang et al. (2009). This is unsurprising given that our approach is richer by accommodating stochastic distributions at two levels—one each for the measurement and transition equations—whereas BNLR accommodates only an error distribution from a nonlinear regression and hence performs poorly if the physical model is misspecified. Finally, our proposed approach also enjoys better interpretation than the hierarchical Gaussian process models of Monteiro, Banerjee, and Ramachandran (2014) as they provide greater precisions in estimates because the random effects in the hierarchical models

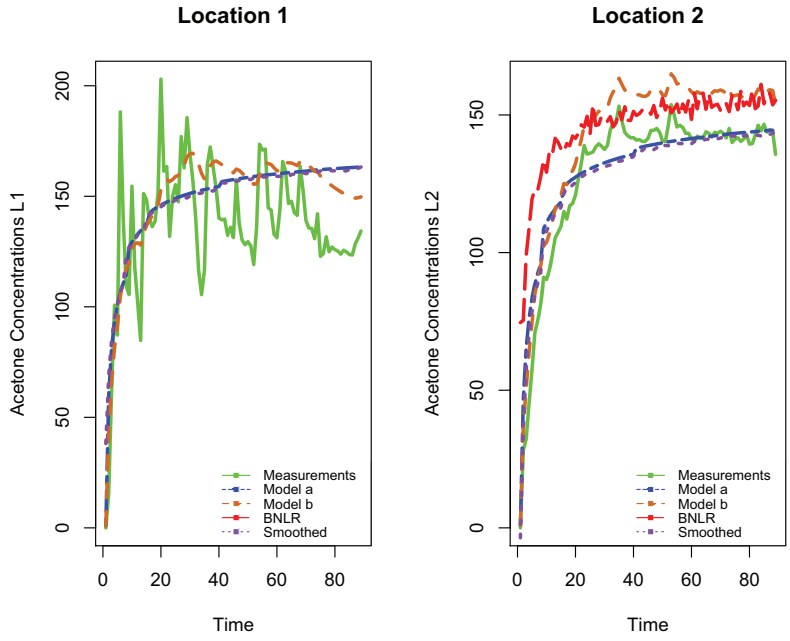

Figure 10. Plots of the measured concentrations and posterior means of the latent states conditional on the measurements at the two locations for (a) Non-Gaussian SSM; (b) Gaussian SSM; and BNLR.

of Monteiro, Banerjee, and Ramachandran (2014) tend to inflate variances.

For the experimental data, the one-zone model results were better compared to the two-zone model and the eddy diffusion model. This is not entirely surprising since each physical experiment was designed for one physical model, and simpler models imply simpler data and assumptions, and possibly fewer parameters. In addition, in the one-zone model, there is only one state at each time point to be estimated, unlike the two-zone and the eddy diffusion models, where there are at least two point estimates at each time point. However, we believe that in a real workplace settings, assuming a uniform concentration of the contaminant across the room may not be realistic and a more flexible model like the eddy diffusion model would yield better results.

The eddy diffusion data has some limitations related to the small size of the chamber, which rendered a small difference between the concentrations in the two locations which also makes it hard to measure the spatial variation for Model (5) implementation. Despite that, in most cases, a nonlinear non-Gaussian Bayesian SSM was able to characterize the data well and the model seems robust to most of the experimental scenarios.

We conclude with some indicators for future research. First, as alluded to earlier, we will need to do a much more comprehensive spatiotemporal analysis for eddy diffusion experiments. While our simulation experiments showed the promise of spatiotemporal SSMs in analyzing eddy diffusion experiments, our chamber data analysis had limited scope because of the very small number of spatial measurements. Second, our current framework relies upon first-order linear approximations of the differential equations. This approximation will likely be improved using second or higher order approximations leading to new classes of dynamic hierarchical models that need to be further investigated. Another important consideration is misaligned data, such as was considered in Monteiro, Banerjee, and Ramachandran (2014) for two zone experiments, where not all measurements for the near and far fields came from the same set of timepoints. An advantage of the Bayesian paradigm is that we can handle missing data, hence misaligned data, very easily and indeed our Bayesian SSMs should be able to handle them as easily as the models in Monteiro, Banerjee, and Ramachandran (2014). Future work will include such analysis and also extensions to spatiotemporal misalignment for eddy-diffusion experiments, where not all timepoints generated measurements for the same set of spatial locations. Future work may also include incorporating the multiresolution method proposed by Kou et al. (2012) in the discretized model. Accuracy of the discretization can be improved by introducing missing data, where more accuracy implies dense discretization and hence computational burden. The multiresolution method can improve the discrete-time approximations of diffusion processes differential equations by employing different discretization schemes at different resolutions that communicate with each other, hence provide fast and accurate approximations. The multiresolution method can be used in the proposed SSM when data is observed over long intervals in time.

Supplementary Materials

R-code for Bayesian SSMs used: R-code to perform the filtering, smoothing and parameters estimation and model assessment methods described in the article. (Bayesian_STSP Rmd and PDF files)

Discretization of the differential equations: We approximate the deterministic physical model through discretization. The Taylor expansion of $C(t)$ at $t = t^*$ is $C(t) = \sum_{n=0}^{\infty} \frac{C^{(n)}(t^*)}{n!}(t - t^*)^n$, where $C^{(n)}(t^*) = \frac{d^n}{dt^n} C(t)\Big|_{t=t^*}$. Let $t = t^* + \delta_t$ hence

$$C(t^* + \delta_t) = \sum_{n=0}^{\infty} \frac{C^{(n)}(t^*)}{n!}(\delta_t)^n = C(t^*) + \frac{C'(t^*)}{1!}\delta_t + o(\delta_t), \quad (7)$$

for small δ_t. From the above equation we can express $C'(t^*)$ as

$$C'(t^*) = \frac{C(t^* + \delta_t) - C(t^*)}{\delta_t} + o(\delta_t). \quad (8)$$

In the applications to the three physical models, we replace the first-order derivative $\frac{d}{dt}C(t)$ at $t = t^*$ with Equation (8) using the appropriate value of δ_t. In the one-zone and two-zone models a value $\delta_t = 0.01$ was found to provide an accurate approximation, while for the eddy diffusion model $\delta_t = 1$ was used.

Steady states derivations: The steady state is achieved as $t \to \infty$ in the exact solution of the ODE.

$$\lim_{t\to\infty} \exp\{tF_t\}C(t_0) + F_t^{-1}[\exp\{tF_t\} - I]g. \quad (9)$$

For the one-zone model $F_t = -(Q + K_L V)/V$ and $g = G/V$ so (9) $= F_t^{-1}[-I]g = G/(Q + K_L V)$.

For the two-zone model, $F_t = A = \begin{bmatrix} -\beta/V_N & \beta/V_N \\ \beta/V_F & -(\beta + Q)/V_F - K_L \end{bmatrix}$ and $g = \begin{bmatrix} G/V_N \\ 0 \end{bmatrix}$. The term $\exp(tF_t)$, where exp() is the matrix exponential, can be written as $\exp(tL\Lambda L^{-1}) = \sum e^{t\lambda_i} G_i$ where $G_i = u_i v_i^T$, u_i is the ith column of L and v_i^T is the ith row of L^{-1}. It easily follows that $e^{tF_t} = \sum_{i=1}^m e^{t\lambda_i} G_i$. The eigenvalues are available in closed form Zhang et al. (2009) as

$$\lambda_1 = \frac{1}{2}\left[-\left(\frac{\beta V_F + (\beta+Q)V_N}{V_N V_F}\right) + \sqrt{\left(\frac{\beta V_F + (\beta+Q)V_N}{V_N V_F}\right)^2 - 4\left(\frac{\beta Q}{V_N V_F}\right)}\right],$$

$$\lambda_2 = \frac{1}{2}\left[-\left(\frac{\beta V_F + (\beta+Q)V_N}{V_N V_F}\right) - \sqrt{\left(\frac{\beta V_F + (\beta+Q)V_N}{V_N V_F}\right)^2 - 4\left(\frac{\beta Q}{V_N V_F}\right)}\right]. \quad (10)$$

As long as β and Q are positive, the sum of the two eigenvalues are negative. Hence, $e^{tF_t} = \sum_{i=1}^m e^{t\lambda_i} G_i \to 0$ as $t \to \infty$ and the first term becomes 0 and the second term becomes $A^{-1}[-I]g$. Also, $\det(A) = (Q\beta + \beta K_L V_F)/V_N V_F$, and

$$A^{-1} = \begin{bmatrix} -((\beta + Q + K_L V_F)/V_F)(V_N V_F/(\beta Q + \beta K_L V_F)) & -(\beta/V_N)(V_N V_F/(\beta Q + \beta K_L V_F)) \\ -(\beta/V_F)(V_N V_F/(\beta Q + \beta K_L V_F)) & -((\beta)/V_N)(V_N V_F/(\beta Q + \beta K_L V_F)) \end{bmatrix}.$$

So the steady state is a 2×1 vector equal to $A^{-1}[-I]g = \begin{bmatrix} \frac{G\beta + QG + K_L V_F G}{\beta Q + \beta V_F K_L} \\ \frac{\beta G}{\beta Q + \beta K_L V_F} \end{bmatrix}$. So as $t \to \infty$ $C_N(t) \approx \frac{G\beta + QG + K_L V_F G}{\beta Q + \beta V_F K_L}$ and $C_F(t) \approx \frac{\beta G}{\beta Q + \beta K_L V_F}$.

The steady state for the eddy diffusion model is theoretically the value of $C(s,t)$ in Equation (4) when $t \to \infty$. Clearly $\lim_{t\to\infty} \frac{G}{2\pi D_T(||s||)}\left(1 - erf\frac{||s||}{\sqrt{4D_T t}}\right) = \frac{G}{2\pi D_T(||s||)}$.

Acknowledgments

The authors thank the editor, associate editor, and two anonymous referees for several constructive suggestions.

Funding

The work of the second author was supported (in part) by federal grants NSF/DMS 1513654, NSF/IIS 1562303, NIH/NIEHS 1R01ES027027, and NIH/NIEHS R01ES030210.

References

Arnold, S., Shao, Y., and Ramachandran, G. (2017), "Evaluating Well-Mixed Room and Near-Field-Far-Field Model Performance Under Highly Controlled Conditions," *Journal of Occupational and Environmental Hygiene*, 14, 427–437. [151,156]

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2014), *Hierarchical Modeling and Analysis for Spatial Data*, Boca Raton, FL: Chapman and Hall/CRC. [151]

Banerjee, S., Ramachandran, G., Vadali, M., and Sahmel, J. (2014), "Bayesian Hierarchical Framework for Occupational Hygiene Decision Making," *The Annals of Occupational Hygiene*, 58, 1079–1093. [147]

Banerjee, S., and Roy, A. (2014), *Linear Algebra and Matrix Analysis for Statistics*, Boca Raton, FL: Chapman and Hall/CRC. [149]

Eubank, R. L. (2005), *A Kalman Filter Primer*, Boca Raton, FL: Chapman and Hall/CRC. [148]

Fearnhead, P. (2011), *MCMC for State-Space Models*, Boca Raton, FL: Chapman and Hall, pp. 513–529. [148]

Gelfand, A. E., Banerjee, S., and Gamerman, D. (2005), "Spatial Process Modelling for Univariate and Multivariate Dynamic Spatial Data," *Environmetrics*, 16, 465–479. [151]

Gelfand, A. E., and Ghosh, S. K. (1998), "Model Choice: A Minimum Posterior Predictive Loss Approach," *Biometrika*, 85, 1–11. [151]

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013), *Bayesian Data Analysis*, Boca Raton, FL: Chapman and Hall/CRC. [150,151,156]

Hoi, K., Yuen, K., and Mok, K. (2008), "Kalman Filter Based Prediction System for Wintertime PM10 Concentrations in Macau," *Global NEST Journal*, 10, 140–150. [158]

Keil, C. B., Berge, W. F. T., and AIHA (2009), *Mathematical Models for Estimating Occupational Exposure to Chemicals*, Fairfax, VA: AIHA Press. [147]

Kou, S. C., Olding, B. P., Lysy, M., and Liu, J. S. (2012), "A Multiresolution Method for Parameter Estimation of Diffusion Processes," *Journal of the American Statistical Association*, 107, 1558–1574. [159]

Leleux, D., Claps, R., Chen, W., Tittel, F. K., and Harman, T. (2002), "Applications of Kalman Filtering to Real-Time Trace Gas Concentration Measurements," *Applied Physics B*, 74, 85–93. [158]

Monteiro, J. V. D., Banerjee, S., and Ramachandran, G. (2011), "B2Z: An R Package for Bayesian Two-Zone Models," *Journal of Statistical Software*, 43, 1–23. [147,151]

——— (2014), "Bayesian Modeling for Physical Processes in Industrial Hygiene Using Misaligned Workplace Data," *Technometrics*, 56, 238–247. [147,151,158,159]

Nicas, M., and Jayjock, M. (2002), "Uncertainty in Exposure Estimates Made by Modeling Versus Monitoring," *AIHA Journal*, 63, 275–283. [147]

Ramachandran, G. (2005), *Occupational Exposure Assessment for Air Contaminants*, Boca Raton, FL: CRC Press. [148]

Shao, Y., Ramachandran, S., Arnold, S., and Ramachandran, G. (2017), "Turbulent Eddy Diffusion Models in Exposure Assessment—Determination of the Eddy Diffusion Coefficient," *Journal of Occupational and Environmental Hygiene*, 14, 195–206. [149,151,152,156,157]

Syring, N., and Martin, R. (2018), "Calibrating General Posterior Credible Regions," *Biometrika*, 106, 479–486. [153]

Wikle, C. K., and Cressie, N. (1999), "A Dimension-Reduced Approach to Space-Time Kalman Filtering," *Biometrika*, 86, 815–829. [151]

Zhang, Y., Banerjee, S., Lungu, C., and Ramachandran, G. (2009), "Bayesian Modeling of Exposure and Airflow Using Two-Zone Models," *Annals of Occupational Hygiene*, 53, 409–424. [147,152,158,160]