Title

Parametric and Non-parametric Bayesian Modeling of Spatio-temporal Exposure Data in Industrial Hygiene

Author

Abdalla, Nada

Publication Date

2018

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Parametric and Non-parametric Bayesian Modeling of Spatio-temporal Exposure Data in Industrial Hygiene

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Biostatistics

by

Nada Ahmed Abdelfattah Abdalla

2018

ABSTRACT OF THE DISSERTATION

Parametric and Non-parametric Bayesian Modeling of Spatio-temporal Exposure Data in Industrial Hygiene

by

Nada Ahmed Abdelfattah Abdalla
Doctor of Philosophy in Biostatistics
University of California, Los Angeles, 2018
Professor Sudipto Banerjee, Chair

In industrial hygiene, prediction of a worker's exposure to chemical concentrations at the workplace is important for exposure management and prevention. The objective of this dissertation is to consider and address challenges in the statistical analyses of exposure data in industrial hygiene. We outline flexible Bayesian frameworks for parameter inference and exposure prediction. In particular, we will focus on two applications of the Bayesian approach on exposure data.

The first application is spatial interpolation of chemical concentrations at new locations when measurements are available from coastlines, as is the case in coastal clean-up operations in oil spills. We present a novel yet simple methodology for analyzing spatial data that is observed over a coastline. We demonstrate four different models using two different representations of the coast. The four models were demonstrated on simulated data and two of them were also demonstrated on a dataset from the GuLF STUDY. Our contribution here is to offer practicing hygienists and exposure assessors with a simple and easy method to implement Bayesian hierarchical models for analyzing and interpolating coastal chemical concentrations.

The second application is inference and prediction of chemical concentrations at the workplace using state space models. Exposure assessment models are deterministic models that

are usually derived from physical-chemical laws that explain the workplace under theoretically ideal conditions. We propose Bayesian parametric and nonparametric approaches for modeling exposure data in industrial hygiene using a state space model framework which combines information from observations, physical processes and prior knowledge. Posterior inference is obtained via easy implementable Markov chain Monte Carlo (MCMC) algorithms. The performance of the different methods will be studied on computer-simulated and controlled laboratory-generated data. We will consider three commonly used occupational exposure physical models varying in complexity.

The dissertation of Nada Ahmed Abdelfattah Abdalla is approved.

Ronald S. Brookmeyer

Michael Leo Brenna Jerrett

Donatello Telesca

Sudipto Banerjee, Committee Chair

University of California, Los Angeles

2018

To my beloved parents who gave me the love of knowledge

and taught me hard work and patience

and to my beloved husband who has been and always is my source of strength

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I am extremely grateful for the support, mentoring and guidance of my advisor, Dr.Sudipto Banerjee. I am very thankful to all the brilliant ideas, encouragement and all the time and effort he put in to ensure my success. His dedication, brilliance, modesty, quality of work and enthusiasm have always been and will always be an inspiration to me in my career. I am truly honored to have had the opportunity to work with him. I would also like to thank Dr.Donatello Telesca for his close supervision and guidance in my recent research work. I am very thankful for his support, contributions and for the valuable opportunity to work with him and learn so much from all his interesting ideas. Furthermore, I would like to thank Dr.Ron Brookmeyer who has been a constant source of encouragement during my PhD. I am very grateful and honored to have had the opportunity to work with him. His work ethics will always guide me in the future. I could not have asked for better advisors. I would also like to thank Dr.Rob Weiss for planting the love of Bayesian statistics in my heart and Dr.Tom Belin and Dr.Damla Senturk for always having my best interest in their minds. I would like to thank the very valuable member of my committee Dr.Micheal Jerrett for his insightful feedback and guidance.

I would also like to thank my classmates who helped me get through the PhD successfully. First, I would like to thank Aaron Scheffler for being a good and a very supportive friend, Eric Kawaguchi, Alec Chan-Golston and Alex Klomhaus for helping me get through the first years in the PhD. I would also like to express my gratitude to Roxy Narnajo for her constant support and having my best interest in her heart since my first day in the program. Last but not least, I would like to express my gratitude to my family for their continuing support and encouragement. I would like to thank my parents who planted in me the love of knowledge and education and who taught me the importance of hard work and persistence, and my sisters who always encouraged and supported me. Lastly, to my husband, Mohamed, I cannot thank you enough for the years of love, encouragement and support throughout my graduate career, for always giving me from your positivity and for making me happy.

VITA

2003–2007 B.Sc. (Statistics), Cairo University, Cairo, Egypt

2008–2012 M.Sc. (Demography and Biostatistics), Cairo University, Cairo, Egypt

2013–2015 M.Sc. (Biostatistics), UCLA, Los Angeles, California

2013–2018 Teaching Assistant, Biostatistics Department, UCLA. Taught labs and discussions of BIOS100A, 100B, 201A, 201B, 200B, 200C.

2015–2018 Graduate Student Researcher, Biostatistics Department, UCLA.

PUBLICATIONS

Abdalla, N., Banerjee S, Ramachandran G, Arnold S. (2018) Bayesian Forecasting and Dynamic Modeling of Physical Processes in Industrial Hygiene. *JSM Proceedings, Section on Bayesian Statistical Science Section.* Alexandria, VA: American Statistical Association.

Abdalla N, Banerjee S, Ramachandran G, Arnold S. (2018) Bayesian State Space Modeling of Physical Processes in Industrial Hygiene. *Technometrics* (Submitted: Under Review)

Abdalla N, Banerjee S, Ramachandran G, Stenzel M and Stenzel M. (2018) Coastal kriging: a Bayesian approach. *Annals of Work Exposures and Health*; 62(7):818-827.

Brookmeyer R, Abdalla N. (2018) Design and Sample Size Considerations for Alzheimer's Disease Prevention Trials Using Multistate Models, *Clinical Trials* (Accepted for publication).

Brookmeyer R, Abdalla N. (2018) Multistate Models and Lifetime Risk Estimation: Application to Alzheimer's Disease. *Statistics in Medicine* (Accepted for publication).

Brookmeyer R, Abdalla N. (2018) Estimation of lifetime risks of Alzheimer's disease dementia using biomarkers for preclinical disease. *Alzheimer's and Dementia: The Journal of Alzheimer's Association*; 14(8):981-988.

Brookmeyer R, Abdalla N, Kawas C and Corrada M.(2017) Forecasting the prevalence of pre-clinical and clinical Alzheimer's disease in the United States. *Alzheimer's and Dementia: The Journal of Alzheimer's Association*; 14(2):121-129.

Brookmeyer R, Kawas C, Abdalla N, Paganini-Hill A, Kim R and Corrada M (2016). Impact of interventions to reduce Alzheimer's disease pathology on the prevalence of dementia in the oldest-old. *Alzheimer's and Dementia: The Journal of Alzheimer's Association*; 12(3): 225-232.

CHAPTER 1

Introduction

In industrial hygiene, estimation of a worker's exposure to chemical concentrations in the workplace is an important concern. Prediction of exposure through statistical and mathematical modeling is gaining popularity, especially with the advent of the REACH regulations in European Community that requires assessing exposure in a variety of scenarios when monitoring may not be feasible [Ram08]. An accurate representation will produce better concentration estimates and facilitates decision making in exposure management. However, this is challenging because the workplace is complex and no physical model is likely to deliver complete representation. Therefore, accounting for parameter and model uncertainty is crucial and a synergy of physical and statistical models is needed to better estimate the processes in the workplace.

We address common problems associated with the evolution of the underlying processes generating observed concentration levels. In other words, finding the predictive distribution of the unknown physical process X given the data Y. In particular, the dissertation addresses the problem of parameter inference and prediction of chemical exposures that evolve over space and time. In many situations, chemical concentrations are unobserved directly and partial noisy measurements are available. The aim is to infer the latent process using those observations, along with the physical model that theoretically describes it, as well as incorporating professional knowledge. This data assimilation approach employs the spatio-temporal data Y to predict X. Markov chain Monte Carlo is used to learn about both the processes and the model parameters and for uncertainty quantification. Another problem that may arise is when chemical concentrations evolve along a coastline, as is the case in coastal clean-up operations in oil spills. In that situation, the spatial distribution of the

chemical concentrations depends on the representation of the coast using curves.

More specifically, this dissertation will focus on some statistical approaches for the above problems that are motivated by two applications in industrial hygiene. First, we propose a simple approach for statistical interpolation of chemical concentrations at new locations when measurements are available from coastlines which we call "coastal kriging". This method is demonstrated on simulated data and on a dataset from the GuLF STUDY (Gulf Long-term Follow-up Study). The second application is the use of parametric and nonparametric state space models for parameter inference and prediction using noisy chemical concentration measurements at the workplace and the physical model describing the process. Posterior inference is obtained via easy implementable Markov chain Monte Carlo (MCMC) algorithms. The performance of the different models will be studied on computer-simulated data and controlled laboratory-generated data. Three commonly used occupational exposure physical models varying in complexity will be considered.

The remainder of this chapter is organized as follows. Section 1.1 provides a brief background introduction to "coastal kriging". In Section 1.2, we provide a background on state space modeling in industrial hygiene, where exposure modeling using state space models are discussed in Section1.2.1 and common physical models in industrial hygiene are discussed in Section 1.2.2, with more details to be covered in subsequent chapters.

1.1 Background to Coastal Kriging

Statistical interpolation at new locations based upon a set of observed measurements at known locations is often referred to as "Kriging" in the geostatistical literature [Cre93]. Kriging customarily uses spatial analytic tools such as variograms or covariance functions to construct best linear unbiased predictors for data collected from a bounded region of interest with positive area. However, in our application the chemicals are sampled mostly along a coastline and interpolation is sought at new locations along the coast. Thus, all measurements are collected along a curve (approximating the coastline) and prediction is sought at new points on this curve. We call this "coastal kriging."

Models for waterway stream networks using moving averages have been developed by [HP10]. They used stream distance rather than Euclidean distance. These models are flexible and adequate for stream networks and account for the volume and direction of flowing water, but they are more complicated and difficult to fit than is necessary for coastal kriging. Unlike networks, where we have a complex structure of line-segments and joints, in simple coastal kriging we approximate the coastline with a single curve or a sequence of line segments. Therefore, a simple parametrization of the coast will suffice and lead to easily implementable statistical models.

We will illustrate our models using a specific dataset extracted from the GuLF STUDY database. In April 2010 an explosion of the *Deepwater Horizon* oil rig resulted in an oil spill in the Gulf of Mexico which was the largest oil spill in the US history. Thousands of workers were involved in stopping and cleaning up the oil release. The GuLF STUDY was conducted by the National Institute of Environmental Health Sciences (NIEHS) and sponsored by the National Institute of Health (NIH) [KEM17]. One specific task in assessing exposures of workers cleaning the coastline is to statistically interpolate the chemical concentrations at new locations along the coast.

Here we outline a Bayesian hierarchical modeling framework to implement coastal kriging. This offers easier interpretability for the uncertainty estimates and can be easily executed using several software packages within the `R` statistical computing environment.

1.2 Background to State Space Models in Industrial Hygiene

1.2.1 Exposure Modeling using Bayesian State Space Models

Exposure models aim at capturing the underlying physical processes generating chemical concentrations in the workplace. Exposure modeling through statistical and mathematical models may provide more accurate exposure estimates than monitoring [NJ02]. Occupational hygienists seek to infer these latent processes from the available measurements as well as quantification of uncertainty in parameter estimation. For example, generation and ven-

tilation rates are crucial parameters that are difficult to obtain since most workplaces do not collect information routinely.

Traditional approaches involve using deterministic physical models that ignore the existence of uncertainty. Bayesian methods combining professional judgment from experts and direct measurements [GCS13] were successful in different settings [BRV14]. For example, [ZBL09] introduced a nonlinear regression on the solution of the differential equations representing the underlying physical model within a Bayesian setting for the two-zone model 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, which severely limits the number of applicable physical models. [MBR11] introduced an `R` package (B2Z), which implements the Bayesian two-zone model proposed by [ZBL09]. [MBR14] demonstrated that straightforward Bayesian regression can be ineffective in predicting exposure concentrations in industrial workplaces since the information is limited to partial measurements and does not take into account the "bias" between the physical model and reality. They introduced a process-based Bayesian melding approach where measurements are related to the physical model through a stochastic process that captures 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, cumbersome computations and opaque interpretation. Bayesian formulation that utilized Gaussian process (GP) models was also provided by [HGW08] which allows for highly multivariate output.

There are main issues with the current practice in exposure assessment. First, the existing methods tend to assume Gaussian distributions for errors and random effects. This is usually less appropriate for concentration measurements and even if they are transformed to resemble normality, such transformations exacerbate inconsistencies with the underlying physical models. Second, the methods are restricted to a rather confined class of physical models whose solutions are analytically tractable or efficiently computed [MBR14]. One needs the solution to the nonlinear differential equations representing the physical model. This precludes fitting computationally demanding but richer physical models that could have yielded better estimation of physical parameters and concentrations.

We offer a principled Bayesian approach to efficiently and effectively synergize information from the three sources of information (a) professional judgment from experts, and (b) direct measurements of the environment exposure in the workplace and (c) scientific physical models representing the state in the workplace in theoretically ideal conditions. Furthermore, the approach we propose here will completely obviate the need to solve the nonlinear equations governing the physical model. We achieve this by deriving a dynamic statistical model by discretizing the deterministic physical model and incorporating stochastic measurement error. The prior knowledge about the model parameters are encoded using prior distributions (including variance components attributed to measurement error and model approximations). The parameters are estimated by sampling from the posterior distribution [GCS13]. We consider a number of Monte Carlo based filtering methods for parameter estimation and inference in state space models and uncertainty quantification. We also relax the assumption of Gaussian error terms and consider nonparametric alternatives.

The different models are compared and assessed using computer-simulated data and lab-generated data. 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. Concentrations were generated at different ventilation and generation rates and under different exposure physical models that are discussed in the following section.

1.2.2 Physical Models in Industrial Hygiene

Physical models represent the physical processes generating chemical concentrations in the workplace. They are usually derived from physical-chemical laws assuming a workplace under theoretically ideal conditions. We will consider three of the most popular physical models for exposure assessment in industrial hygiene [Ram05]. These are the one-zone (well-mixed compartment) model, two-zone model and the eddy diffusion model.

The one-zone model assumes a single compartment with one source of chemical emission. It assumes that the source is generating the contaminant at a generation rate G(mg/min) in

a room of volume $V(\mathrm{m}^3)$ with a supply and exhaust flow rates (ventilation rate) $Q(\mathrm{m}^3/\mathrm{min})$. The room is assumed to be perfectly mixed which means that there is a uniform concentration of the contaminant throughout the room (Figure 1.1). The loss term $K_L(\mathrm{mg/min})$ measures the loss rate of the contaminant due to other factors such as chemical reactions or the contaminant being absorbed by the room surfaces.

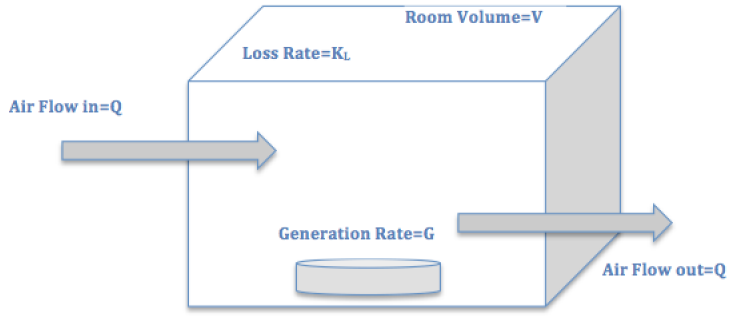

Figure 1.1: One-zone model schematic showing key model parameters; generation rate G, ventilation rate Q and loss rate K_L

The differential equation describing this model is

$$V\frac{d}{dt}C_t + (Q + K_L V)\, C_t = G.$$

The two-zone model [Nic96] assumes different physical behavior near the source of emission *"near field"* from that far from the source *"far field"*. Both fields are assumed to be a well-mixed box, i.e., 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 the contaminant is generated at a rate $G(\mathrm{mg/min})$, in a room with ventilation rate $Q(\mathrm{m}^3/\mathrm{min})$ and loss rate by other mechanisms $K_L(\mathrm{mg/m}^3)$. This model includes one more parameter that indicates the airflow between the near field and the far field $\beta(\mathrm{m}^3/\mathrm{min})$. The volume in the near field is denoted by $V_N(\mathrm{m}^3)$ and the volume in the far field is denoted by $V_F(\mathrm{m}^3)$. Figure 1.2 illustrates the dynamics of the system.

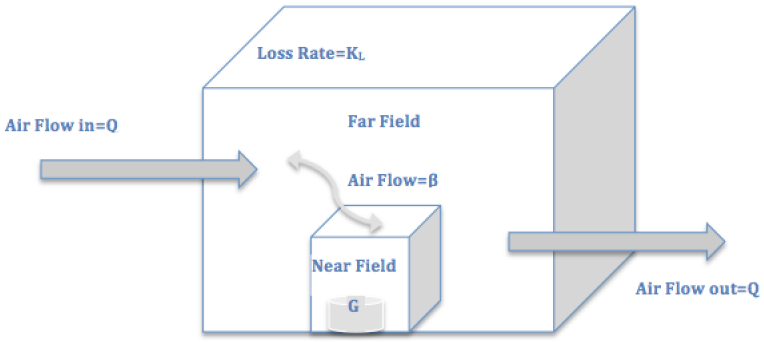

Figure 1.2: Two-zone model schematic showing key model parameters; generation rate G, ventilation rate Q airflow β and loss rate K_L

The following system of differential equations represents the two-zone model

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

Turbulent Eddy diffusion model [KBA09] is an example of models which, unlike the one-zone and the two-zone models, provides a concentration gradient from the source outward. It takes into account the worker's location relative to the source. The concentration $C_{t,s}$ is a function of location $s(x, y)$ in a two-dimensional Euclidean coordinates and time t. The parameter that is unique to this model is the turbulent eddy diffusion coefficient $D_T(\mathrm{m}^2/\mathrm{min})$. D_T describes how quickly the emission spreads with time (Figure 1.3) and is assumed to be constant over space and time. The following differential equation represents the change in concentration over time at location $s = (x, y)$

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

The one-zone and two-zone experimental datasets were generated in a controlled lab. [ASR17] conducted a series of chamber studies under controlled conditions in an exposure chamber of size 11.8 m^3. A solvent (toluene/ 2-butanone/ acetone) was released into the chamber at one of three known generation rates $G(\mathrm{mg/min})$ where ventilation rate

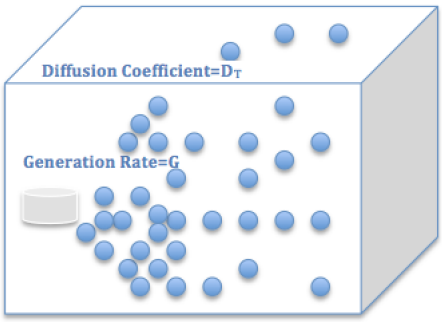

Figure 1.3: Eddy diffusion model schematic showing key model parameter; diffusion coefficient D_T

$Q(\mathrm{m}^3/\mathrm{min})$ was controlled. To achieve good mixing, two fans were placed in opposite corners of the chamber. Three ventilation rates representing residential and industrial settings of $0.04 - 0.07$ $\mathrm{m}^3/\mathrm{min}$, 0.23-0.27 $\mathrm{m}^3/\mathrm{min}$ and 0.47-0.77 $\mathrm{m}^3/\mathrm{min}$ were used. The two-zone model near field (0.105 m^3) was constructed from perforated wire mesh within the far filed (11.79 m^3). The airflow rate β cannot be measured directly but estimated according to the equation $\beta = 1/2 \times FSA \times S$ where FSA(m^2) is the free surface area of the near field which is the sum of the area across the six sides of the box and S(m/min) is the local air speed [Nic96]. Values of β were expected to be between 0.24 and 1.24 $\mathrm{m}^3/\mathrm{min}$. A study for each of the three solvents at the three ventilation and generation rates was conducted. Concentrations were measured every 90 seconds using gas monitors that require gas specific calibration. In our analysis, three different experimental datasets at three different ventilation rates were used for model comparison and assessment.

The eddy diffusion experimental datasets were generated in a controlled chamber as well. [SRA17] constructed an experimental chamber of size 11.9 m^3 with spatial concentration gradient away from the contaminant source at different airflow conditions. They used diffusers to promote eddy formation. A series of experiments were conducted where acetone or toluene was pumped into the chamber at known generation rates $G(\mathrm{mg}/\mathrm{min})$. Gas detectors were placed at two locations at distances 0.41m and 1.07 m away from the source and concentrations were measured every 120 seconds. Experiments were conducted at different values of air change per hour (ACH).

1.3 Contributions and Dissertation Outline

The focus of this dissertation is parameter inference and prediction of chemical exposure data using new approaches. The dissertation provides a framework for coastal kriging. It also provides a methodology to perform parameter estimation and state prediction in parametric and nonparametric state space models using MCMC methods. The nonparametric methods can be particularly useful when the model is misspecified. Chapter 2 presents a literature review of kriging, spatial processes for coastline measurements and discusses coastal kriging in details. Chapter 3 presents a literature review on state space models and Bayesian inference in state space models. It addresses inference of the latent state process that arises from different chemical-physical laws. The solution proposed utilizes the discretization of the physical model which, is expressed as a linear system of ordinary differential equations (ODEs) and uses a state space model as a representation of an unobserved state of interest that evolves over time and partial observations that are observed sequentially over discrete time. Chapter 4 expands upon the parametric state space models into a more flexible model that relaxes the distributional assumptions of the parametric model.

CHAPTER 2

Coastal Kriging: A Bayesian Approach

2.1 Introduction

Data observed over locations with known geographic coordinates are often referred to as point-referenced data and are commonly seen in environmental health. Recent applications consider such data measured along coastlines or shores. For example, to assess exposures of workers to chemicals along the coastline may require statistical interpolation of the chemical concentration at new locations along the coast. Statistical interpolation at new locations based upon a set of observed measurements at known locations is often referred to as "Kriging" in the geostatistical literature [Cre93]. Kriging customarily uses spatial analytic tools such as variograms or covariance functions to construct best linear unbiased predictors. When chemicals are sampled mostly along a coastline, interpolation is sought at new locations along the coast. Thus, all measurements are collected along a curve (approximating the coastline) and prediction is sought at new points on this curve. We call this "coastal kriging."

Models for waterway stream networks using moving averages have been developed [HP10]. They use stream distance rather than Euclidean distance. These models account for the volume and direction of flowing water in stream networks. They offer richness and flexibility, but are complicated and can be difficult to compute. Unlike networks, where we have complex structure of line-segments and joints, in simple coastal kriging we approximate the coastline with a single curve or a sequence of line segments. A simple parametrization of the coast will suffice and lead to easily implementable statistical models.

We will pursue Bayesian coastal kriging. Bayesian models offer easier interpretability for

parameter estimates, provide exact estimates of uncertainty without requiring assumptions of large sample sizes and independence of observations, and can incorporate prior information when available. Bayesian models can be easily executed using several software packages within the `R` statistical computing environment.

We will illustrate our models using a specific dataset extracted from the GuLF STUDY (Gulf Long-term Follow-up Study) database. In April, 2010 an explosion of the *Deepwater Horizon* oil rig resulted in an oil spill in the Gulf of Mexico. It was the largest oil spill in US history. Tens of thousands of workers were involved in stopping and in cleaning up the oil release. The GuLF STUDY is conducted by the National Institute of Environmental Health Sciences (NIEHS) and sponsored by the National Institute of Health (NIH) [KEM17]. It is collecting information to study potential adverse effects on the health of those workers. Among other activities, the workers capped the well, applied dispersants to break up the oil, skimmed or burned the oil on the Gulf waters, cleaned beaches, marshes and structures, decontaminated equipment, and provided support for these activities. Personal air measurements are available on many of these tasks. The highest portion of the STUDY participants were involved in cleaning the beaches, marshes and structures. One specific task in assessing exposures of workers cleaning the coastline is to statistically interpolate the chemical concentration at new locations along the coast.

Our contribution expands upon existing geostatistical models to allow for better prediction of quantities of interest at new locations over coastlines. The chapter is organized as follows. Section 2 provides a brief review of Bayesian methods for kriging. Section 3 discusses spatial processes for coastline measurements. Section 4 discusses our geostatistical models for interpolating point-referenced coastline data and simple algorithms for implementing Bayesian kriging. Section 5 discusses simulation results that help validate our method. Section 6 illustrates our model through applying it to the GuLF STUDY data. Section 7 provides conclusions and suggestions for some future work.

2.2 Model-based Kriging

2.2.1 Spatial process models with Euclidean coordinates

Point-referenced spatial modeling seeks to capture associations between observations geographically closer to each other and to predict the value of the response or outcome variable at arbitrary locations. This is achieved using a spatial regression model,

$$Z(s) = x(s)^\top \beta + \omega(s) + \epsilon(s) \ , \ \epsilon(s) \overset{iid}{\sim} N(0, \tau^2) \tag{2.1}$$

where $x(s)^\top$ is a $1 \times p$ vector of covariates (predictors) observed at location s, $\omega(s)$ is a latent (unobserved) spatial random effect at location s, and $\epsilon(s)$ accounts for measurement error. For any collections of locations, the measurement errors in (2.1) are normally distributed independently and identically, each with a zero mean and variance τ^2.

If $\omega(s) = 0$ for all locations, then (2.1) reduces to an ordinary linear model with independent outcomes. If the outcomes are spatially correlated, then $\omega(s)$ introduces dependence. There are several different mechanisms for specifying $\omega(s)$ [Cre93, BCG14], but we choose a fairly straightforward and interpretable model here. We assume that each $\omega(s)$ has mean 0 and the dependence at two points s and s' is modeled as

$$\text{Cov}\{\omega(s), \omega(s')\} = K_\theta(s, s') = \sigma^2 \exp(-\phi \|s - s'\|) \ , \tag{2.2}$$

where $\|s - s'\|$ is the distance between two locations s and s', $\theta = \{\sigma^2, \phi\}$, σ^2 captures the variation attributed to spatial effects (referred to as partial sill) and ϕ controls the rate at which the spatial correlation drops to zero. The *spatial range* is defined as the distance beyond which the spatial correlation becomes negligible. For the exponential covariance function in (2.2), the spatial range is given by approximately $3/\phi$ which is the distance where the correlation drops below 0.05.

We incorporate the covariance function (2.2) into a probability model. Let $\mathcal{S} = \{s_1, s_2, \ldots, s_n\}$ be the set of spatial locations. The $n \times 1$ vector ω, whose i-th entry is $\omega(s_i)$, follows a mul-

tivariate normal distribution $N(0, K_\theta)$, where K_θ is the $n \times n$ spatial covariance matrix with (i, j)th entry $K_\theta(s_i, s_j)$ in (2.2). The measurement errors are independent across locations, hence $\epsilon(s_i) \overset{iid}{\sim} N(0, \tau^2)$. This implies that the data vector Z, whose i-th element is $Z(s_i)$, is multivariate normal with mean vector $X\beta$, where $x(s_i)^\top$ are the rows of X, and variance-covariance matrix $K_\theta + \tau^2 I_n$, where I_n denotes the $n \times n$ identity matrix.

Spatial regression models, such as (2.1), are fitted by estimating geostatistical parameters σ^2, ϕ and τ^2 in addition to the regression coefficients β. We use (2.1) to predict the outcome at a new location after accounting for the uncertainty in parameter estimates. When all points lie on a region represented as a 2-D plane, the distance between s and s' in (2.2) is given by the standard Euclidean distance formula. Here, the correlation drops at the same rate for every direction, so the spatial range is a function of distance only. Also, the covariance function in (2.2) ensures that K_θ is always positive definite [BCG14].

In our current context, the points lie along a curve representing the coastline. There are two issues. First, the Euclidean distance is inappropriate for modeling spatial covariances because the effective spatial range will be the distance *along the coast* at which the correlation becomes negligible. Second, covariance functions that insure positive definiteness in Euclidean coordinates need not be valid for other domains [Ban05]. This means that we will need to construct valid covariance functions along the coastline. Subsequently, we describe a simple approach to construct models such as (2.1) using valid covariance functions for points along curves.

2.2.2 Spatial processes for coastline measurements

We now extend the model discussed in the previous section to the case where the data are observed over a coastline. Since all observations lie along the coastline, we will model spatial dependence along the coastline. The spatial range and variability will need to be interpreted in terms of distance along the coastline. Prediction is also sought at arbitrary points along the coast. We assume that any point s on the coast is given by $\gamma(t) = (\gamma_1(t), \gamma_2(t))$ for some $t \in \mathcal{T} \subset \Re^1$, where $\gamma_1(t) = f(t)$ and $\gamma_2(t) = g(t)$ are parametric equations for the

coordinates. Therefore, each value of t determines a coordinate on a plane and traces out a curve $\gamma(t)$ as t varies over a range $\mathcal{T}$. The coastline is now given by the set of all points on it: $\gamma(\mathcal{T}) = \{\gamma(t) : t \in \mathcal{T} \subset \Re^1\}$. For example, a simple curve could be approximated by line segments. For each line segment, $\gamma(t)$ is a straight line, $\gamma_i(t) = \{s_i + tu \,|\, t \in [0, \infty]\}$, originating at s_i and parallel to the direction vector u. Here, $s_i = \begin{bmatrix} \gamma_{1i} \\ \gamma_{2i} \end{bmatrix}$, $u = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$ and, hence,

$$\gamma_i(t) = \left\{ \begin{bmatrix} \gamma_{1i} + tu_1 \\ \gamma_{2i} + tu_2 \end{bmatrix} \,|\, t \in [0, \infty] \right\} \tag{2.3}$$

A customary choice for the parameter t is the arc length. As another example, consider a circular coast with radius r. The curve is defined as

$$\gamma(t) = \{\gamma_1(t) = r\cos t, \gamma_2(t) = r\sin t \,|\, t \in [0, \pi/2]\} \tag{2.4}$$

The point $\gamma(t) = (r\cos t, r\sin t)$ moves in a fixed orientation (e.g., clockwise) as t increases. If t is the length of an arc of the circle and λ is the angle in radians which the arc subtends at the center of the circle, then $t = r\lambda$.

A spatial regression model such as (2.1) can be defined over a coast by representing each point on the coast by $\gamma(t)$. Thus, we write $Y(t) = Z(\gamma(t))$ for every $t \in \mathcal{T}$. Therefore,

$$Y(t) = x^\top(t)\beta + \omega(t) + \epsilon(t);, \tag{2.5}$$

where $x(t)$ is the vector of covariates observed at the point $\gamma(t)$, $\omega(t)$ is now defined over $\mathcal{T}$ with covariance function

$$\mathrm{Cov}(\omega(t), \omega(t')) = K_\theta(t, t') = \sigma^2 \exp(-\phi|t - t'|) \,, \tag{2.6}$$

where $|t - t'|$ is the absolute difference between t and t', and $\epsilon(t) \overset{iid}{\sim} N(0, \tau^2)$. Coastal covariance is demonstrated in Figure 2.3. Here, the correlation between two points decreases as the coastal distance between them increases.

The choice of t depends on the parametric equation used to approximate the coast. If the coast can be well-represented in closed form using a parametric equation, then t as the arc-length is a reasonable and convenient choice. More generally, an arbitrary coastline can be well approximated using a series of small line segments. Each segment is then defined according to (2.3). For example, in our subsequent simulation experiments we present linear approximations for an elliptical coastline. In our real example we use a series of small linear segments to model the coast along Waveland Beach in Mississippi.

2.2.3 Coastal kriging

Exposure assessors may be interested in predicting the concentration of a toxicant at any arbitrary location on the coast. Let $Y(t_0)$ be the toxicant concentration measurement at the point $\gamma(t_0)$ on the coast. The posterior probability distribution of $Y(t_0)$, which is also referred to as the posterior predictive distribution, is computed in two steps. First, the unknown parameters in (2.5) are estimated by using Bayes' theorem to compute their posterior distributions. Thus, if $p(\theta, \beta, \tau^2)$ represents the prior distribution for unknown parameters and $p(y \,|\, \theta, \beta, \tau^2)$ represents the likelihood, then the posterior distribution is given by

$$p(\theta, \beta, \tau^2 \,|\, y) = \frac{p(\theta, \beta, \tau^2) \times p(y \,|\, \beta, \theta, \tau^2)}{p(y)} \propto p(\theta, \beta, \tau^2) \times p(y \,|\, \theta, \beta, \tau^2) \; . \tag{2.7}$$

The prior distribution can be informative or non-informative. Non-informative priors typically deliver inference consistent with classical methods. Even for weakly informative priors, the inference is often close to classical methods because the effect of the data typically overwhelms the prior. While often producing inference numerically very similar to classical inference, Bayesian inference will retain simpler interpretability.

Suppose we have toxin measurements at points $\gamma(t_1), \gamma(t_2), \ldots, \gamma(t_n)$ on the coast and have collected the $y(t_i)$'s in an $n \times 1$ vector y. Let X be the $n \times p$ matrix with i-th row $x^\top(t_i)$ and ω be the $n \times 1$ vector with elements ω_i. The posterior distribution of the model

parameters is

$$p(\beta, \sigma^2, \tau^2, \phi \,|\, y) \propto U(\phi \,|\, a_\phi, b_\phi) \times IG(\tau^2 \,|\, a_{\tau^2}, b_{\tau^2}) \times IG(\sigma^2 \,|\, a_{\sigma^2}, b_{\sigma^2}) \\ \times N(\beta \,|\, \mu_\beta, V_\beta) \times N(y \,|\, X\beta + \omega, \tau^2 I) \,, \tag{2.8}$$

where $U(\cdot, \cdot)$, $IG(\cdot, \cdot)$ and $N(\cdot, \cdot)$ represent the uniform, the inverse-Gamma and the Normal distributions, respectively, as expounded in [GCS13].

Posterior distributions, in general, are not available in simple closed-forms. Instead we sample $\{\beta, \omega, \theta, \tau^2\}$ from their posterior distribution, where $\theta = \{\sigma^2, \phi\}$, using Markov chain Monte Carlo (MCMC) methods [GCS13]; [BCG14]. Some simplifications are often made. One is to use a flat completely noninformative prior on β. Another is to integrate out ω from (2.8). The posterior samples for $\{\beta, \sigma^2, \tau^2, \phi\}$ are then obtained from

$$p(\beta, \theta, \tau^2 \,|\, y) \propto U(\phi \,|\, a_\phi, b_\phi) \times IG(\tau^2 \,|\, a_{\tau^2}, b_{\tau^2}) \times IG(\sigma^2 \,|\, a_{\sigma^2}, b_{\sigma^2}) \times N(y \,|\, X\beta, K_\theta + \tau^2 I) \,. \tag{2.9}$$

The posterior samples for ω are subsequently obtained by sampling one instance of ω from $N(\cdot, \cdot)$ for each sampled value of $\{\beta, \sigma^2, \tau^2, \phi\}$. This is called composition sampling [BCG14].

Suppose we have collected M post-convergence posterior samples for the model parameters, say $\{\beta_{(j)}, \theta_{(j)}, \tau^2_{(j)}\}$, for $j = 1, 2, \ldots, M$. Then the posterior samples for $Y(t_0)$ are obtained by composition sampling, i.e., for each j we draw $Y_{(j)}(t_0)$ from the conditional normal distribution, say $N(m_{(j)}, v^2_{(j)})$, where the mean and variance are

$$m_{(j)}(t_0) = x(t_0)^\top \beta_{(j)} + \tilde{K}_{\theta_{(j)}}(t_0, t) \tilde{K}^{-1}_{\theta_{(j)}}(t, t) \tilde{K}_{\theta_{(j)}}(t, t_0)(y - X\beta_{(j)}) \\ \text{and} \;\; v^2_{(j)}(t_0) = \tilde{K}_{\theta_{(j)}}(t_0, t_0) - \tilde{K}_{\theta_{(j)}}(t_0, t) \tilde{K}^{-1}_{\theta_{(j)}}(t, t) \tilde{K}_{\theta_{(j)}}(t, t_0)) \,, \tag{2.10}$$

where $\tilde{K}_{\theta_{(j)}}(\cdot, \cdot) = K_{\theta_{(j)}}(\cdot, \cdot) + \tau^2 I$. Note that $m_{(j)}(t_0)$ and $v^2_{(j)}(t_0)$ are precisely the classical kriging estimator and variance evaluated at $\{\beta_{(j)}, \theta_{(j)}, \tau^2_{(j)}\}$. Bayesian kriging, therefore, quantifies uncertainty in kriging by averaging the classical kriging estimator over the posterior distribution of the parameters. The resulting $Y_{(j)}(t_0)$ are samples from the posterior

predictive distribution. The mean of these samples yields a point estimate of the predicted value at t_0, while the variance of the posterior samples estimates the predictive variance.

One assumption to simplify matters is that ϕ and $\alpha = \frac{\tau^2}{\sigma^2}$ are fixed, say at values resulting from the empirical variogram [BCG14]. Hence, the posterior samples for the model parameters are obtained from the conjugate model

$$p(\beta, \sigma^2 \,|\, y) \propto IG(\sigma^2 \,|\, a_{\sigma^2}, b_{\sigma^2}) \times N(\beta \,|\, \mu_\beta, \sigma^2 V_\beta) \times N(y \,|\, X\beta, \sigma^2 V_y) \;, \tag{2.11}$$

where $V_y = R(\phi) + \alpha I$ and $R(\phi)$ is the spatial correlation matrix with elements $\exp(-\phi|t_i - t_j|)$. Here one can sample exactly from the posterior distribution in (2.11). For each $j = 1, 2, \ldots, M$ we first draw $\sigma^2_{(j)} \sim IG(a^*_{(j)}, b^*_{(j)})$ followed by $\beta_{(j)} \,|\, \sigma^2_{(j)}, y \sim N(Bb, B\sigma^2_{(j)})$, where $a^*_{(j)} = a_{\sigma^2} + n/2$ and $b^* = b_{\sigma^2} + (y^\top V_y y - b^\top B b)/2$, where $B = (X^\top V_y^{-1} X + V_\beta^{-1})^{-1}$ and $b = X^\top V_y^{-1} y$.

2.3 Simulation

The simulated data consists of $n = 100$ data points. The outcome $Y(t)$ values were generated on an ellipse. We first generated $l_i \sim Uni(0, 2\pi)$ for $i = 1, 2, \ldots, n$, where the corresponding parametric equations are $m = 2\cos(l)$ and $n = \sin(l)$. We then drew a multivariate normal random variable $\omega \sim N(0, K_\theta)$ and then $y(t_i) \sim N(\beta_0 + \omega(t_i), \tau^2)$, where t_i is the arc-length between points (m_{i-1}, n_{i-1}) and (m_i, n_i).

In the data generation step, we fixed $\tau^2 = 0.1$, $\beta = 0$ and $\theta = \{1, 1\}$. For assessing predictive performance, we used 75 observations for training the model and withheld 25 observations for testing the predictive validation.

We estimated the models in (2.9) and (2.11). To compare the performance of coastal kriging to kriging using Euclidean distance, we estimated the model in (2.1) as well using the covariance in (2.2). For all models, we assigned a noninformative prior to β_0 (i.e., $V_\beta^{-1} = O$ the matrix of zeroes) and an $IG(2, 2)$ prior to τ^2. In (2.9) σ^2 and ϕ were assigned $IG(2, 2)$ and $U(0.8, 30)$ priors. The $IG(2, b)$ prior provides a prior mean of b but has, in theory, an

infinite variance yielding a relatively vague prior but with a prior value centered around b. In (2.11), we fixed $\phi = 1.07$ and $\alpha = 0.25$ for the coastal kriging model and $\phi = 22009.68$ and $\alpha = 8.13 \times 10^{-5}$ for kriging with Euclidean distance. Starting values for σ^2, τ^2 and ϕ in (2.8) and the fixed values for ϕ and $\alpha = \tau^2/\sigma^2$ in (2.11) were provided using their estimates from the empirical variogram for the data [BCG14].

We also compared coastal kriging to universal kriging (UK). Universal kriging is kriging with a trend, where $E(Z(s))$ is a linear combination of the known functions $\{f_0(s), \ldots, f_p(s)\}$ [Cre93]. We assume that the mean $E(Z(s))$ is a function of the coordinates in a linear form, i.e $Z(s) = \beta_0 + \beta_1 x_1(s) + \beta_2 x_2(s) + \omega(s) + \epsilon(s)$, where $x_1(s)$ is the longitude at location s, $x_2(s)$ is the latitude at location s.

In practice we will not have an exact parametric formula for the coastline. This needs to be approximated by simple parametric curves. The easiest such option is a sequence of line segments, as described earlier. We used our simulated dataset to evaluate the performance of such linear approximations. Let $\Delta m_i = m_i - m_{i-1}$ and $\Delta n = n_i - n_{i-1}$, then the length of the straight line segment connecting the two points is $t^* = \sqrt{(\Delta m)^2 + (\Delta n)^2}$. For small Δm, the sum of the lengths of these line segments provides an approximation to the length of the curve. We will, therefore, consider four models for coastal kriging. The model in (2.9) with the exact parametrization for an ellipse will be called Model 1a , while that with linear approximation will be called Model 1b. Similarly, the exact and approximate parameterizations corresponding to the model in (2.11) will be referred to as Model 2a and Model 2b respectively.

Table 2.1 presents the posterior medians and 95% Bayesian credible intervals for the parameters in each of the above four models, the simple Euclidean distance kriging model and the UK model. The credible intervals from all models include the true values of β_0. Models 1a and 1b captured the true values of σ^2 and ϕ. Model 2b also captured the true value of σ^2 and Models 2a and 2b captured the true value of τ^2. To assess predictive performance across the six models, we used mean square prediction error (MSPE). Coastal kriging and UK models produced very similar MSPE values, and the highest MSPE was produced by the simple Euclidean distance kriging model. For model comparison we also

used the Kullback-Leibler (K-L) divergence criterion ($D_{KL}(M_0 \,|\, M_i)$), $i = 1, \ldots 5$, where M_0 is the true distribution and M_i is the distribution under model i. For multivariate normal distributions the Kullback-Leibler divergence ([BT94]) takes the form

$$\frac{1}{2}(tr(\Sigma_{M_i}^{-1}\Sigma_{M_0}) + [x\beta_{M_i} - x\beta_{M_0}]^\top \Sigma_{M_i}^{-1}[x\beta_{M_i} - x\beta_{M_0}] - n + \ln(\det(\Sigma_{M_i})) - \ln(\det(\Sigma_{M_0})))$$

where $\Sigma = K_\theta + \tau^2 \mathrm{I}$.

Figure 2.1: Map of interpolated total hydrocarbons (ppm) over Waveland Beach, Mississippi

Model 1a produced the lowest D_{KL} followed by Models 1b, 2b and 2a, and the highest values were produced by the UK model and the simple Euclidean distance kriging model. We also used deviance information criterion (DIC), which is commonly used in Bayesian model selection. Model 2a produced the lowest value followed by Models 2b, 1a and 1b, then

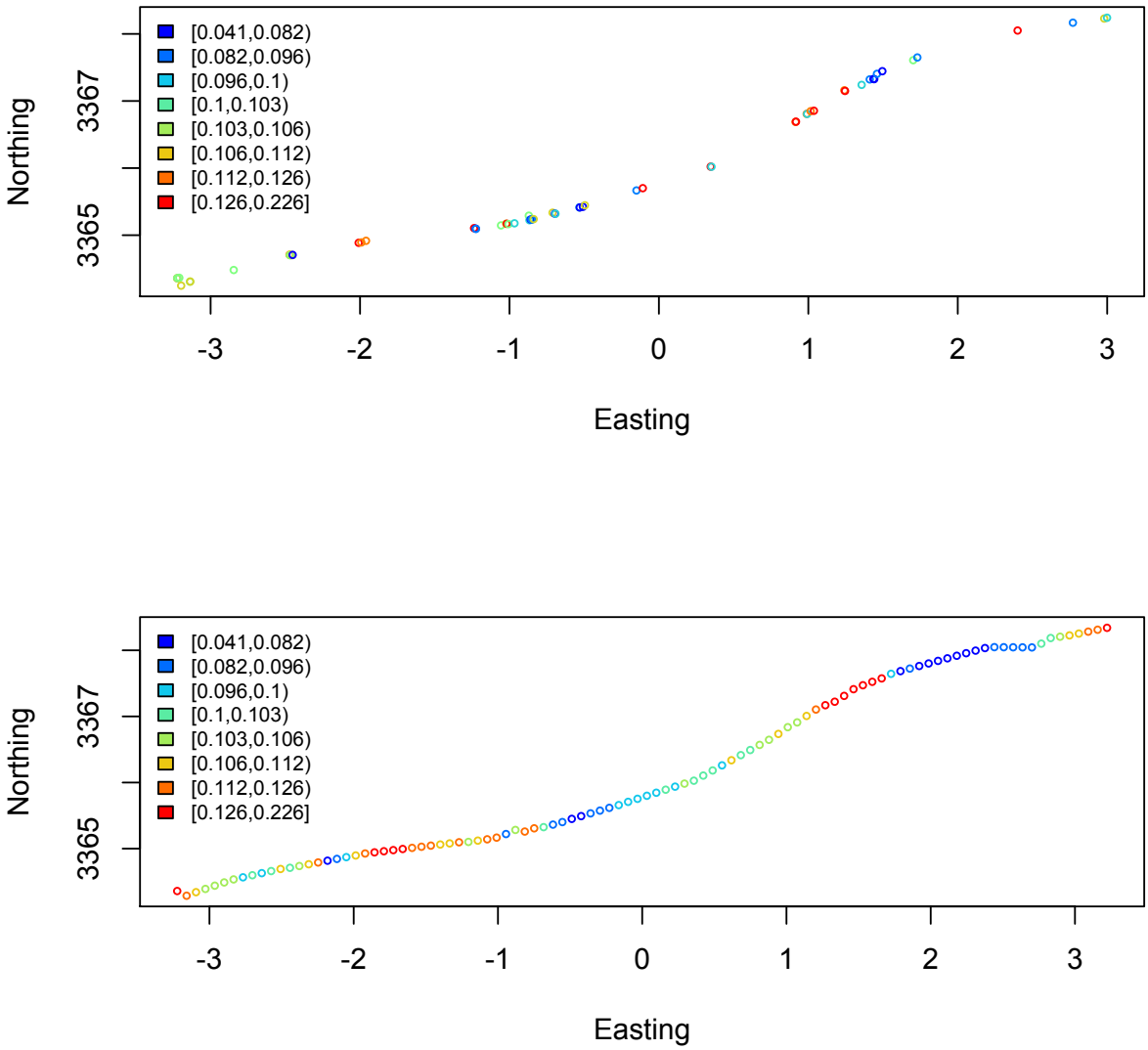

Figure 2.2: Observed total hydrocarbons (ppm) over Waveland Beach, Mississippi and Model 1b interpolated values

the UK model, and the highest value was produced by the simple Euclidean distance kriging model. Finally, ten-fold cross validation (CV(10)) was the lowest among coastal kriging models followed by the UK model then the simple Euclidean distance kriging model. We also used Bayesian 95% prediction intervals and the predicted mean values of the outcome from the 25 holdout locations and plotted them against the true values (Figure 2.4). For

Table 2.1: Medians, 2.5% and 97.5% quantiles of the posterior samples of the coefficient estimate, partial sill σ^2, nugget effect τ^2, decay parameter ϕ, MSPE, DIC, Kullback-Leibler and CV(10) for the fitted models to the simulated data

	True	Model 1a[1]	Model 1b[2]	Model 2a[3]	Model 2b[4]	Simple kriging[5]	Universal kriging
β_0	0	0.16(-0.41,0.79)	0.16 (-0.51,0.89)	0.19(-0.35, 0.71)	0.24(-0.49, 0.97)	0.006(-0.16, 0.18)	0.03(-0.24, 0.31)
σ^2	1	0.55(0.30,1.09)	0.56 (0.30, 1.14)	0.48(0.36,0.67)	0.74(0.55, 1.04)	0.58(0.43, 0.81)	0.18(0.12,0.26)
τ^2	0.1	0.18(0.12,0.28)	0.18 (0.12,0.27)	0.12(0.09, 0.17)	0.12(0.09, 0.17)	$2.8 \times 10^{-5}(2.1 \times 10^{-5}, 3.9 \times 10^{-5})$	0.17(0.12,0.25)
ϕ	1	1.20(0.85,2.81)	1.15 (0.71,3.58)	0.76	0.76	31773.42	0.32(0.16, 0.96)
MSPE		0.57	0.59	0.53	0.54	1.23	0.56
DIC		30.11	30.95	28.80	29.82	55.43	37.8
Kullback-Leibler		4.10	5.61	5.68	5.64	73.67	100.7
CV(10)		0.170	0.169	0.171	0.176	0.558	0.183

[1] Full hierarchical model using arc-length.
[2] Full hierarchical model using line segment approximation.
[3] Simplified hierarchical model using arc-length.
[4] Simplified hierarchical model using line segment approximation.
[5] Simplified hierarchical model using Euclidean distance.

the coastal kriging models, the intervals include the true values of the outcome variable in each of the holdout locations except for one location. The UK model provided improved prediction over simple Euclidean distance kriging model which produced the least accurate prediction with wider credible intervals.

These results indicate that Bayesian models using linear approximation to a parametric curve do not seem to adversely affect the inferential performance relative to models using the true form of the parametric curve. They also indicate that coastal kriging is better than classical kriging methods such as simple Euclidean distance kriging and UK when the source of variability in the data arises from a curve. Thus, Models 1b and 2b are good candidate models to be used in the data analysis.

2.4 Data Analysis

Coastal kriging of the concentration of chemicals inhaled by the clean-up workers following the oil spill in 2010 may be useful to assess the potential health effects associated with the spill for locations without measurements. The data set used here consists of air samples

collected on clean-up workers on Waveland beach, Mississippi which extends in an S-shape for seven or eight kilometers (Figure 2.1). The samples were collected for approximately 10 hours per day using passive dosimeters clipped to the workers' collars to measure breathing zone concentrations. The chemicals in the air diffused on to a charcoal pad inside the sampler. Five analytes were analyzed at the laboratory. They include total hydrocarbons (THC) which is a composite of the volatile chemicals in crude oil and is our main variable of interest. There were a total of 60 sample points (THC parts per million (ppm)) collected between September 19 and December 21, 2010 that were used in the analysis. Two exposure groups were considered, workers who cleaned jetties and other land-based structures and workers who cleaned beaches.

Candidate models include Models 1b and 2b where the curve is approximated by line segments and the parameterization in (2.3) is used. The fixed values of ϕ and α in (2.11) could be the estimated from the variogram. However in coastal kriging, the variogram may not provide accurate estimates. Hence, we will use Model 1b in the data analysis and compare the results to simple Euclidean distance kriging results. For both models, we assigned a noninformative prior to β_0 (i.e., $V_\beta^{-1} = O$ the matrix of zeroes) and an $IG(2, 2)$ prior to τ^2. In (2.9) σ^2 and ϕ were assigned $IG(2, 2)$ and $U(0.8, 30)$ priors. The prior on ϕ implies that the effective spatial range, i.e., the distance beyond which spatial correlation is negligible, is between 0.1 and 3.8 on a coastline with a distance of 7.6 kilometers. In addition, coastal kriging was compared to universal kriging (UK) with a linear trend.

Twelve observations acted as a holdout testing sample and the models were assessed based on their predictive performance at new locations using MSPE in addition to 10-fold cross validation (CV(10)) and on the goodness of fit measure DIC. All observations were log transformed to achieve normality.

Table 2.2 shows parameter estimates of the fitted models. MSPE is almost the same among the three models, and the highest CV(10) resulted by the UK model. Model 1b produced the lowest DIC value. Results show that coastal kriging proposed in (2.5) provides a better fit for coastal data compared to other classical kriging methods. Figures (2.1) and (2.2) show plot of 100 total hydrocarbon (ppm) interpolated values resulted from Model 1b fit at

randomly generated coordinates, and Figure (2.2) shows the true observed values as well.

Table 2.2: Medians, 2.5% and 97.5% quantiles of the posterior samples of the coefficient estimate, partial sill σ^2, nugget effect τ^2, decay parameter ϕ, and MSPE, DIC, and CV(10) for the fitted models of the log transformed total hydrocarbons

	Model 1[1]	Simple kriging[2]	Universal kriging
β_0	-2.29(-2.71, -1.83)	-2.23(-2.67,-1.73)	-71.2(-8663.9, 7497.1)
σ^2	0.59(0.29, 1.26)	0.59(0.28, 1.15)	0.52(0.34,0.89)
τ^2	0.46(0.25, 0.80)	0.46(0.27, 0.85)	0.17(0.12,0.23)
ϕ	9.08(1.26, 24.82)	7.43(1.78, 22.70)	0.29(0.29,6.48)
MSPE	0.06	0.06	0.05
DIC	34.4	38.6	65.05
CV(10)	0.06	0.06	0.13

[1] Full hierarchical model using line segment approximation.
[2] Full hierarchical model using Euclidean distance.

2.5 Discussion

We developed a flexible yet simple Bayesian framework for spatially-oriented data that can be used to assess exposures of workers by interpolating levels of chemicals along a coastline. The statistical models for coastal kriging exploit a simple representation of the coast as a parametric function of the coordinates of points along the coastline. We presented four models using two different parameterizations. We found that for a simple curve, kriging using line segment approximation performs better than spatial kriging using Euclidean distance. This could be a useful and practical approach for kriging over any simple curve. The model is relatively easy to fit since the covariance depends on parameters in $\Re^1$.

We remark that the current analysis only considers worker exposure assessment, not community-based exposure assessment. In the GuLF STUDY more than 28,000 samples of THC and several other chemicals were collected across the Gulf, along the coasts, and at ports and docks, providing sufficient data for the STUDY exposure estimates [SSR17]. These estimates were derived from groups of samples based on the tasks being performed.

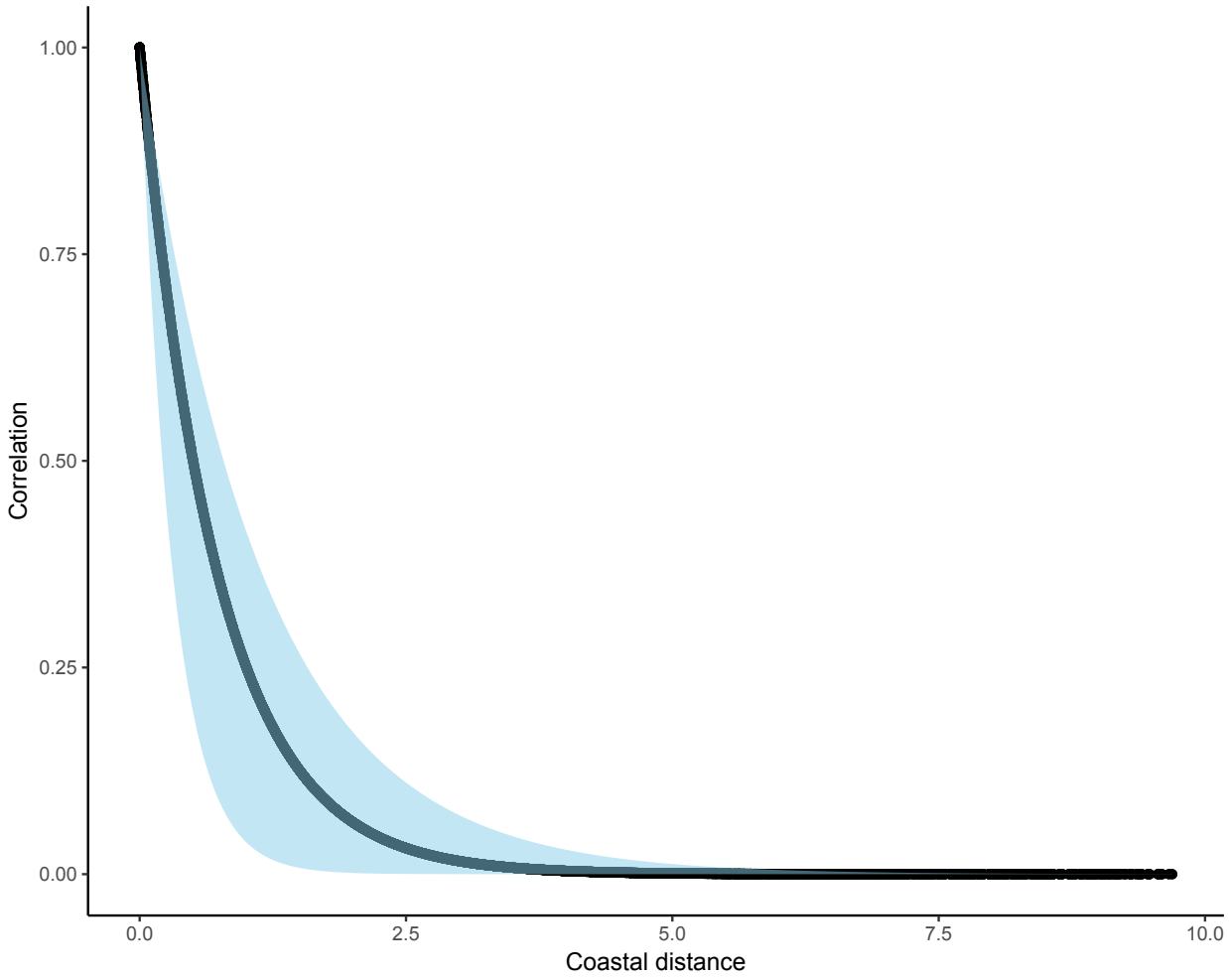

Figure 2.3: Coastal kriging estimated correlation versus coastal distance using the simulated data and Model 1a results with 95% C.I.

The concentrations generated by these tasks (i.e. cleaning the beaches of oil and tar) represent task-derived exposures and, to a lesser extent, ambient air exposures. Using such task-based measurements is not appropriate to impute general or community air concentrations because the task concentrations will be higher than ambient concentrations due to the workers being nearer to the source of the chemical emission than the community. With the data used here, however, the imputed concentrations from the methodology described above may represent workers' exposures performing those same tasks in unmeasured locations. To date, occupational assessment methodologies have focused primary on fairly localized exposure situations. The method described here may be useful in more geographically extended situations, such as workers building a highway or mitigating a chemical release in a river or

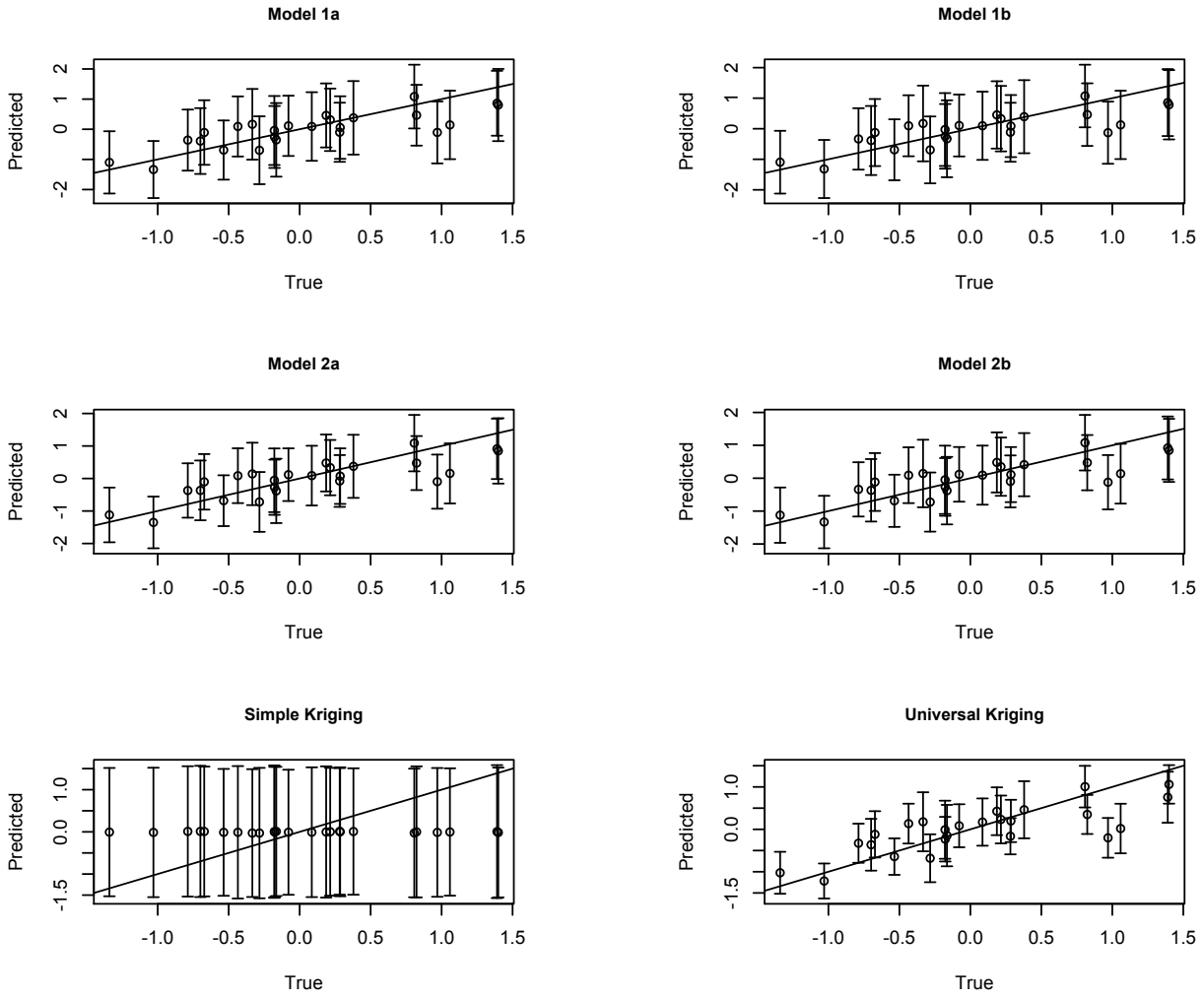

Figure 2.4: Simulated data true versus predicted values with 95% prediction intervals with 45° line of the four coastal kriging and simple kriging models

residents living along a fenceline adjacent to a manufacturing site.

Our study has some limitations within which our findings need to be interpreted carefully. First, the results are based on a total of 60 data points from which 48 were used in training the model and 12 were used in testing it. Second, the data points are distributed on a coast with little curvature which rendered the coastal kriging results slightly better than simple Euclidean distance kriging results. Last, but not least, the distribution of total hydrocarbons in the air is unknown and its source is not arising from the coast which may add some uncertainty in the fitted model, although in our data this uncertainty is assumed

to be minimal.

Building valid models for coastal kriging presents many new research opportunities. For instance, it would be of interest to develop a model for more complicated coastlines and, in particular, along closed curves such as the coasts of an island. Thus, future work will investigate potential problems such as, complexity of the curve, covariates inclusion, potential changes in the coastline and temporal changes. Future work will also consider the modeling and analysis of censored data, as is commonplace in exposure studies, due to measurements below the limits of detection. Finally, we will also consider extending this work to exposure assessment for communities rather than individuals.

CHAPTER 3

Bayesian State Space Modeling of Physical Processes in Industrial Hygiene

3.1 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 unobserved directly and partial noisy measurements are available. Exposure models aim at capturing the underlying physical processes generating chemical concentrations in the workplace. Exposure modeling through statistical and mathematical models may provide more accurate exposure estimates than monitoring [NJ02]. Industrial hygienists seek to infer these latent processes from the available measurements as well as quantification of uncertainty in parameter estimation. For example, generation and ventilation rates are crucial parameters that are difficult to obtain since most workplaces do not collect information routinely.

Traditional approaches involve using deterministic physical models that assign values to those parameters [KBA09]. These approaches however do not provide accurate representation in a real workplace environment as they ignore the model sources of uncertainty. For example, the uncertainties in the true values of the parameters, the numerical implementation, the adequacy of the physical model, the observations and the initial values, and uncertainties from physical processes that are not resolved at the temporal and spatial scales represented in the physical models. Bayesian methods combining professional judgment from experts and direct measurements [GCS13] were successful in different settings. For example, [ZBL09] introduced a nonlinear regression on the solution of the differential equations representing the underlying physical model within a Bayesian setting for the two-zone model 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. This severely limits the number of applicable physical models. [MBR11] introduced an R package (B2Z), which implements the Bayesian two-zone model proposed by [ZBL09]. [MBR14] demonstrated that straightforward Bayesian regression can be ineffective in predicting exposure concentrations in industrial workplaces since the information is limited to partial measurements and does not take into account the "bias" between the physical model and reality. They introduced a process-based Bayesian melding approach where measurements are related to the physical model through a stochastic process that captures 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, cumbersome computations and opaque interpretation. Bayesian formulation that utilized Gaussian process (GP) models was also provided by [HGW08] which allows for highly multivariate output.

We propose using a data assimilation approach in a Bayesian state space model, by discretizing the physical model differential equations that model the rate of change in concentrations, and incorporating information from observed measurements and experts prior knowledge. This approach will enrich the existing methods, as industrial hygienists will no longer be 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.

State space models provide filtered, more accurate state estimates using measurements, which contain some noise, along with the physical model. 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 Bayesian framework provides a natural approach for probabilistic forecasting [GK14], which helps quantify uncertainties in predicted or filtered states. [HGL13] used ensemble Kalman filter (EnKF) for computer model calibration from a Bayesian perspective. [HB18] used

a representation of uncertainty using a class of outer measures in filtering and smoothing, which do not require all sources of uncertainty to be described only by strict probability distributions. [HDZ18b] developed Bayesian inference in a nonlinear filtering framework for parameter uncertainty quantification, while in [HDZ18a], they used Bayes factors and the posterior model probability to quantify model uncertainty.

We focus on approaches for uncertainty quantification (UQ), data assimilation and model calibration that help estimate exposure levels and the model parameters. Recent developments in UQ have been mostly notable in climate and engineering applications. We offer a novel methodology for flexible modeling of chemical concentrations within a Bayesian state space model framework. Representing uncertainty when performing parameter estimation and inference is important for model assessment and selection. We offer a general framework using MCMC methods to sample from the posterior distributions for all the model parameters and the predictive distributions. Our method will be demonstrated on computer-simulated data and lab-generated data.

The chapter is organized as follows. In Section 3.2 we provide a brief review of state space models. Section 4.1.4 provides a brief review of three families of commonly referenced exposure physical models and the corresponding proposed statistical models. Section 3.4 describes the models implementations and assessment methods. Section 3.5 illustrates our models through applying it to simulated data and experinmental chamber data. Section 3.6 concludes the chapter.

3.2 State Space Models

Suppose that $\{X_t : t \in \mathcal{T}\}$, $X_t \in \Re^{n_x}$ is a stochastic process, where $\mathcal{T}$ is the index set and n_x is the dimension of the state vector. A state-space model (SSM) is a representation of that unobserved stochastic process that evolves over time, and sequential partial observations $\{Y_t : t \in \mathcal{T}\}$, $Y_t \in \Re^{n_y}$, where t can be taken as discrete time [Fea11] and n_y is the dimension of the observation vector. In a linear discrete SSM, the measurements Y_t can be represented

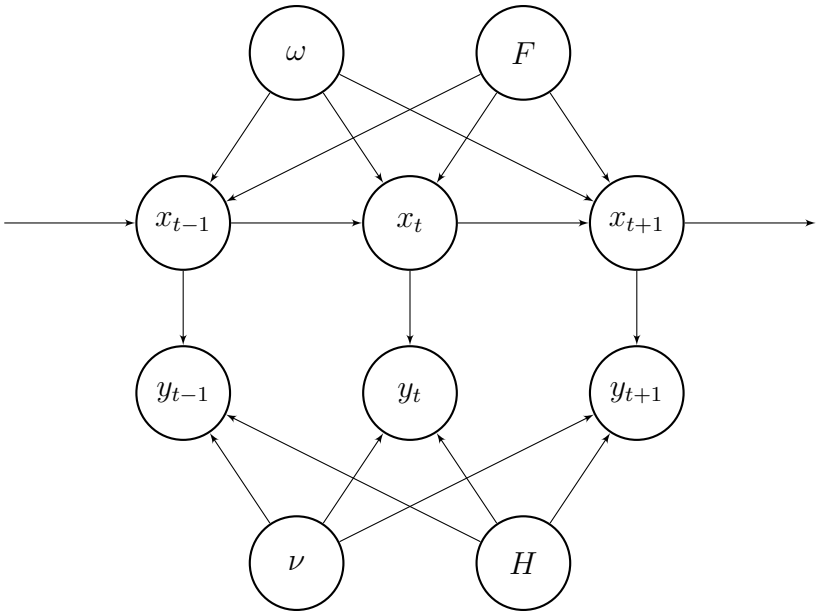

Figure 3.1: Graphical representation of state space model

as

$$Y_t = H_t X_t + \nu_t,$$

for unobserved state vectors X_t, $t \in \mathbb{N}$ and known transformation matrices that map the state vector into the measurement domain $H_t : \Re^{n_y} \times \Re^{n_x} \to \Re^{n_y}$ and zero mean i.i.d. random noise $\nu_t \in \Re^{n_y}$ with covariance R_t [Far12]. The state of a system at time t is assumed to evolve from the prior state at time $t-1$ according to the state equation

$$X_t = F_t X_{t-1} + g + \omega_t,$$

where F_t is a state transition matrix, g are control inputs and $\omega_t \in \Re^{n_x}$ is an i.i.d. process noise sequence with covariance Q_t [Eub05].

In a more general setting, the state vector $\{X_t : t \in \mathbb{N}\}$ is assumed to evolve according to

$$X_t = f_t(X_{t-1}, \omega_{t-1}) \tag{3.1}$$

where f_t is usually a known, possibly nonlinear function in X_{t-1} and ω_t is an i.i.d. process noise sequence. We want to find filtered or updated estimates of X_t based on the measure-

ments Y_t which are assumed to be related to the state vector according to

$$Y_t = h_t(X_t, \nu_t), \tag{3.2}$$

where h_t is the measurement function and ν_t is an i.i.d. measurement noise.

A continuous state space model can generally be expressed as

$$\frac{d}{dt}X_t = F_t X_t + g + \omega_t; \qquad Y_t = H_t X_t + \nu_t. \tag{3.3}$$

We want to transfer the continuous equation into a discrete one. The solution to the first equation in (4.3) when the eigenvalues of F_t are real and distinct can be obtained as follows

$$X_t = \exp(tF_t)X_0 + F_t^{-1} \times [\exp(tF_t) - I]g, \tag{3.4}$$

where $\exp(A)$ denotes the matrix exponential (see [BR14, p. 333-334]). The discretized latent state X_t follows an AR(1) transition model, i.e. for small steps of size δ_t, for $t = 1, \ldots, T$, the first equation in (4.3) can be approximated by

$$X_{t+\delta_t} \approx (I + \delta_t F_t)\, X_t + \delta_t g + \omega_t. \tag{3.5}$$

3.3 Physical models and their statistical counterparts

Bayesian state space representations for exposure assessment models combine direct measurements of the environmental exposure, physical models and prior information. There are several physical models varying in their level of complexity [Ram05]. Three commonly used families of physical models are the well-mixed compartment (one-zone) model, the two-zone model and the turbulent eddy diffusion model. We use discrete approximations to the deterministic physical models and introduce stochastic error terms to derive corresponding dynamic statistical models. 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 state space models (SSMs).

Dynamic models combine measurements with the true underlying state. They 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.

3.3.1 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 $V(\mathrm{m}^3)$ with ventilation rate $Q(\mathrm{m}^3/\mathrm{min})$. The room is assumed to be perfectly mixed, which means that there is a uniform concentration of the contaminant throughout the room (Figure 4.1). The loss term $K_L(\mathrm{mg/min})$ measures the loss rate of the contaminant due to other factors such as chemical reactions or the contaminant being absorbed by the room surfaces.

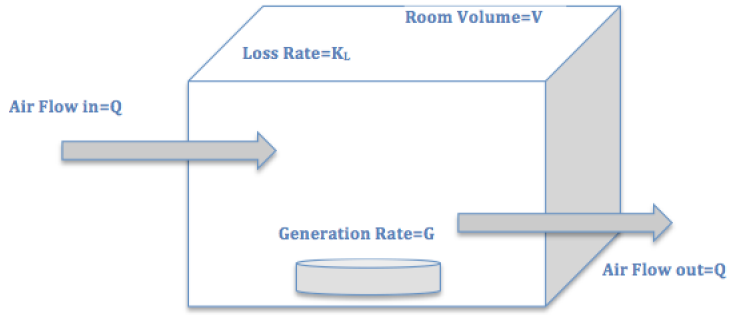

Figure 3.2: One-zone model schematic showing key model parameters; generation rate G, ventilation rate Q and loss rate K_L

The differential equation describing this model is

$$V\frac{d}{dt}C_t + (Q + K_L V)\, C_t = G\,. \tag{3.6}$$

The solution to the differential equation is

$$C_t = \exp\{-t(Q + K_L V)/V\}C(t_0) + ((Q + K_L V)/V)^{-1}\left[1 - \exp\{-t(Q + K_L V)/V\}\right] G/V \ . \tag{3.7}$$

Theoretically, the steady state concentration is the limit of C_t as $t \to \infty$ which is G/Q (mg/m^3). Details of the steady state solution are provided in the supplementary material. Further specifications yield the Bayesian SSM corresponding to (4.13). For example,

$$\begin{aligned}
&\text{Measurement: } \ Z_t = f(C_t) + \nu_t \ , \quad \nu_t \overset{iid}{\sim} P_{\nu,\theta_\nu} \ ; \\
&\text{Transition: } \ C_{t+\delta_t} = \left(1 - \delta_t \frac{Q + K_L V}{V}\right) C_t + \delta_t \frac{G}{V} + \omega_t \ , \quad \omega_t \overset{iid}{\sim} P_{\omega,\theta_\omega} \ . \\
&Q \sim Unif(a_Q, b_Q) \ ; \quad G \sim Unif(a_G, b_G) \ ; \quad K_L \sim Unif(a_{K_L}, b_{K_L}) \ ; \quad \sigma^2 \sim IG(a_\sigma, b_\sigma) \ ;
\end{aligned} \tag{3.8}$$

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.

3.3.2 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 *"far field"*, which completely encloses the near field. Both fields are assumed to be a well-mixed box, i.e., 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 4.2 illustrates the

dynamics of the system.

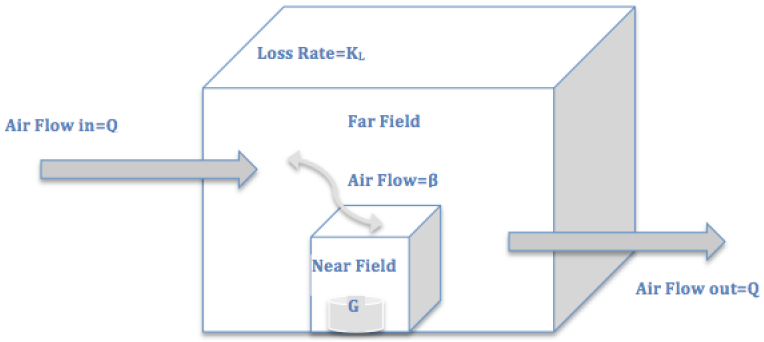

Figure 3.3: Two-zone model schematic showing key model parameters; generation rate G, ventilation rate Q, airflow β and loss rate K_L

The following system of differential equations represents the two-zone model

$$\overbrace{\frac{d}{dt}\begin{bmatrix} C_N(t) \\ C_F(t) \end{bmatrix}}^{\frac{d}{dt}C_t} = \overbrace{\begin{bmatrix} -\beta/V_N & \beta/V_N \\ \beta/V_F & -(\beta+Q)/V_F - K_L \end{bmatrix}}^{A} \overbrace{\begin{bmatrix} C_N(t) \\ C_F(t) \end{bmatrix}}^{C_t} + \overbrace{\begin{bmatrix} G/V_N \\ 0 \end{bmatrix}}^{g} . \tag{3.9}$$

The solution to the differential equations is

$$C_t = \exp(tA)C(t_0) + A^{-1}\left[\exp(tA) - I\right] g \ , \tag{3.10}$$

where $\exp(tA)$ is the matrix exponential. Theoretically, for large values of t, the steady state concentration in the near field is $G/Q + G/\beta$ (mg/m^3), and G/Q (mg/m^3) in the far field. We note that the matrix exponential may be numerically unstable to compute in general. For example, for non-diagonalizable matrices a Jordan decomposition (see, e.g.,[BR14]) may be required, which is very sensitive to small perturbations in the elements of A. Hence, we will avoid this approach.

Analogous to (3.8), the discrete counterpart of (4.16) can be

$$\text{Measurement: } Z_t = f(C_t) + \nu_t\,,\ \nu_t \overset{iid}{\sim} P_{\nu_t,\theta_\nu}\ ;$$

$$\text{Transition: } C_{t+\delta_t} = (\delta_t A(\theta_c; x) + I)\, C_t + \delta_t g(\theta_c; x) + \omega_t\ ; \quad \omega_t \overset{iid}{\sim} P_{\omega_t,\theta_\omega}\ ;$$

$$Q \sim Unif(a_Q, b_Q)\ ; \quad G \sim Unif(a_G, b_G)\ ; \quad K_L \sim Unif(a_{K_L}, b_{K_L})\ ; \quad \beta \sim Unif(a_\beta, b_\beta)\ ,$$

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 ν_t and for ω_t, where θ_ν and θ_ω are parameters in P_{ν,θ_ν} and P_{ω,θ_ω}, respectively.

3.3.3 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. This model accounts for a continuous concentration gradient from the source outward. It takes into account the worker's location relative to the source. The concentration $C_{t,s}$ is a function of the location $s = (x, y)$ in a two-dimensional Euclidean coordinate frame and time t. Without loss of generality, the source of the contaminant is assumed to be at coordinate $(0, 0)$. The parameter that is unique to this model is the turbulent eddy diffusion coefficient $D_T(\text{m}^2/\text{min})$. It describes how quickly the emission spreads with time (Figure 4.3) 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) [SRA17]. We will provide inference for this parameter.

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

$$C_{t,s} = \frac{G}{2\pi D_T \|s\|} \left\{ 1 - \text{erf}\left(\frac{\|s\|}{\sqrt{4 D_T t}} \right) \right\}, \tag{3.11}$$

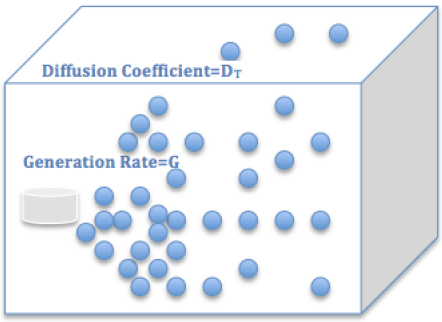

Figure 3.4: Eddy diffusion model schematic showing key model parameter; diffusion coefficient D_T

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

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

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

$$\begin{aligned}
&\text{Measurement:} \quad Z_{t,s} = f(C_{t,s}) + \nu_{t,s} + \eta_t\,,\ \nu_{t,s}\sim P_{\nu_{t,s},\theta_\nu}\,,\ \eta_t \sim P_{\eta_t,\theta_\eta}\ ;\\
&\text{Transition:} \quad C_{t+\delta_t,s} = C_{t,s} + \delta_t \frac{G}{4(D_T\pi t)^{3/2}} \exp\left(-||s||^2/4D_T t\right) + \omega_{t+\delta_t,s}\,,\ \omega_{t,s}\sim P_{\omega_{t,s},\theta_\omega}\ ;\\
&D_T \sim Unif(a_{D_T}, b_{D_T})\ ; \quad G \sim Unif(a_G, b_G)\,,
\end{aligned} \tag{3.12}$$

where $P_{\nu_{t,s},\theta_\nu}$ and $P_{\omega_{t,s},\theta_\omega}$ are spatial-temporal stochastic processes. Note that $\nu_{t,s}$ is a spatial-temporal process discrete in time and continuous in space. This is reasonable because the measurments are taken over discrete time intervals and the estimation for the latent concentration states are required at those intervals. On the other hand, $\omega_{t,s}$ 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.

3.4 Model Implementation and Assessment

For each physical model in Section 4.1.4 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

$$\begin{aligned} &p(\theta_c) \times IW(\Sigma_\omega \,|\, r_\omega, S_\omega) \times IW(\Sigma_\nu \,|\, r_\nu, S_\nu) \times N(C_{t_0} \,|\, m_0, \Sigma_0) \\ &\qquad \times \prod_{i=1}^{n} N(C_{t_i} \,|\, A_{t_i}(\theta_c) C_{t_{i-1}} + \delta_i g_{t_i}, \Sigma_\omega) \times \prod_{i=1}^{n} N(Z_{t_i} \,|\, B_{t_i} C_{t_i}, \Sigma_\nu) \,, \end{aligned} \tag{3.13}$$

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 [GCS13]. Gibbs updates are implemented using $p(C_{t_i} \,|\, \cdot) = N(C_{t_i} \,|\, M_{t_i} m_{t_i}, M_{t_i})$, where $m_{t_i} = \Sigma_\nu^{-1} Z_{t_i} + \Sigma_{t_i|t_{i-1}}^{-1} A_{t_i}(\theta_c) C_{t_{i-1}}$ and $M_{t_i} = (\Sigma_\nu^{-1} + \Sigma_{t_i|t_{i-1}}^{-1})^{-1}$, where $\Sigma_{t_i|t_{i-1}} = A_{t_i}(\theta_c) M_{t_{i-1}} A_{t_i}(\theta_c)^T + \Sigma_\omega$ and $M_{t_0} = \Sigma_0$, $p(\Sigma_\nu \,|\, \cdot) = IW(\Sigma_\nu \,|\, r_{\nu|\cdot}, S_{\nu|\cdot})$ and $p(\Sigma_\omega \,|\, \cdot) = IW(\Sigma_\omega \,|\, r_{\omega|\cdot}, S_{\omega|\cdot})$, where $r_{\nu|\cdot} = r_\nu + n$, $S_{\nu|\cdot} = S_\nu + \sum_{i=1}^{n} (Z_{t_i} - B_{t_i} C_{t_i})(Z_{t_i} - B_{t_i} C_{t_i})^T$, $r_{\omega|\cdot} = r_\omega + n$ and $S_{\omega|\cdot} = S_\omega + \sum_{i=1}^{n} (C_{t_i} - A_{t_i}(\theta_c) C_{t_{i-1}})(C_{t_i} - A_{t_i}(\theta_c) C_{t_{i-1}})^T$.

Note that 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 $IW(\cdot, \cdot)$ priors in (3.13) are replaced by $IG(\sigma^2 \,|\, a_\sigma, b_\sigma)$ and $IG(\tau^2 \,|\, a_\tau, b_\tau)$. The full conditionals now assume the form $p(C_{t_i} \,|\, \cdot) = N(C_{t_i} \,|\, M_{t_i} m_{t_i}, M_{t_i})$, where $m_{t_i} = \sigma^{-2} Z_{t_i} + \sigma_{t_i|t_{i-1}}^{-2} A_{t_i}(\theta_c) C_{t_{i-1}}$ and $M_{t_i} = 1/(\sigma^{-2} + \sigma_{t_i|t_{i-1}}^{-2})$, where $\sigma_{t_i|t_{i-1}}^2 = A_{t_i}(\theta_c)^2 M_{t_{i-1}} + \tau^2$, $p(\sigma^2 \,|\, \cdot) = IG(\sigma^2 \,|\, a_{\sigma|\cdot}, b_{\sigma|\cdot})$ and $p(\tau^2 \,|\, \cdot) = IG(\tau^2 \,|\, a_{\tau|\cdot}, b_{\tau|\cdot})$, where $a_{\sigma|\cdot} = a_\sigma + n/2$, $b_{\sigma|\cdot} = b_\sigma + \sum_{i=1}^{n} (Z_{t_i} - B_{t_i} C_{t_i})^2/2$, $a_{\tau|\cdot} = a_\tau + n/2$ and $b_{\tau|\cdot} = b_\tau + \sum_{i=1}^{n} (C_{t_i} - A_{t_i}(\theta_c) C_{t_{i-1}})^2/2$.

Although Gausian SSMs are very popular in dynamic modeling of physical systems, es-

pecially 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 with both 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 (3.12) 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 (3.12) 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, i.e., $\eta_t \overset{iid}{\sim} N(0, \tau^2)$ 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 $\nu_t \overset{ind}{\sim} N(0, \sigma_t^2 R_t(\phi_t))$, where $R_t(\phi_t)$ is an $m \times m$ matrix with (i, j)-th element $\exp(-\phi_t d_{ij})$ and $d_{ij} = \|s_i - s_j\|$.

Note that $P_{\nu_{t,s},\theta_\nu}$ can, in theory, be a continuous-time spatial-temporal process specified through a space-time covariance function (see, e.g.,[BCG14]). Alternatively, one 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 spatial processes independent across time (see, e.g.,[WC99, GBG]). One could continue to embellish the model in (3.12) using spatial-temporal structures that represent richer hypotheses and more flexible modeling. However, in realistic industrial hygiene applications such specifications will rarely lead to estimable models given the scarcity of data points. For example, most settings will provide measure-

ments from only a handful of locations (e.g., $m \sim 5$) and some moderate numbers of time points (e.g., $n \sim 100$). Therefore, 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 $\nu_t \overset{iid}{\sim} N(0, \sigma^2 R(\phi))$ so that each $m \times 1$ vector ν_t has the same m-variate Gaussian distribution.

Finally, we turn to smoothing and filtering. Smoothing is achieved by evaluating at each time point t_i the posterior expectation of the concentration value given the entire observed data $y = \{y_{t_i} : i = 1, 2, \ldots, n\}$, including observations before and after t_i. Thus, we sample from the posterior density $p(C_{t_i} \,|\, y)$ in posterior predictive fashion by sampling a C_{t_i} from its full conditional, $p(C_{t_i} \,|\, \cdot)$, for each sampled value of the parameters. For linear Gaussian SSM, Kalman smoother can be used where the smoothed distribution at time t also follows a Gaussian distribution. For the nonlinear non-Gaussian SSM, [BDM09] provided a discussion of the different smoothing approaches. This provides an idea about the structure of the smoothing distribution of the collection of states [GDW04]. Filtering, on the other hand, aims to estimate the posterior expectation of the concentration value C_{t_i}, given the data up to t_i, i.e., $\{y(t_j) : j = 1, 2, \ldots, i\}$. We have implemented both smoothing and filtering for all the physical models considered above.

To compare between models, we adopt a posterior predictive loss approach (see, e.g., [GG98]). We generate the posterior predictive distributions for each data point, $y_{rep,i}$ for $i = 1, 2, \ldots, n$ by sampling from $p(y_{rep} \,|\, y) = \int p(y_{rep} \,|\, \theta, \{C_t\}) p(\theta, \{C_t\} \,|\, y) d\theta$, where θ denotes the full collection of unknown parameters and $\{C_t\}$ is the collection of latent concentrations over the entire time frame. We will compute the posterior predictive mean, $\mu_{rep,i} = \mathrm{E}[y_{rep,i} \,|\, y]$, and dispersion, $\Sigma_{rep,i} = \mathrm{var}[y_{rep,i} \,|\, y]$, for each $y_{rep,i}$; these are easily calculated from the posterior samples for each $y_{rep,i}$. We will prefer models that will perform well under a decision-theoretic *balanced loss function* that penalizes departure of replicated means from the corresponding observed values (lack of fit), as well as the uncertainty in the replicated data. Using a squared error loss function, the measures for these two criteria are evaluated as $G = \sum_{i=1}^{n} \|y_i - \mu_{rep,i}\|^2$, where $\|\cdot\|$ denotes the Euclidean norm, and $P = \sum_{i=1}^{n} \mathrm{Tr}(\Sigma_{rep,i})$, where $\mathrm{Tr}(A)$ denotes the trace of the matrix A. We will use the score

$D = G + P$ as a model selection criteria, with lower values of D indicating better models.

We will also use the deviance information criterion (DIC) as a model comparison metric [SBC]. The DIC is a generalization of the AIC and is calculated by adding a measure of fit which is the posterior expected deviance $\bar{D} = E_{\theta|y}[-2\log p(\text{data}|\theta)]$ and a penalty $p_D = \bar{D} - D(\bar{\theta})$, where $\bar{\theta}$ refers to the posterior expectation of the model parameters.

3.5 Data Analysis

In this section we evaluate the performance of the models discussed in Section 3.4, for the three physical exposure models illustrated in Section 4.1.4, using computer-simulated datasets as well as experimental lab-generated data. In particular, we consider two models: a Gaussian linear model and a non-Gaussian nonlinear model, and they will be referred to as Gaussian SSM and non-Gaussian SSM respectively. The prior settings are based on physical knowledge and experience, and discussed in the following section.

The computer-simulated data was generated using R computing environment. The lab-generated data experiments were conducted in test chambers. [ASR17] examined parts of this data using the deterministic one-zone and two-zone models and showed that performance is highly reliable on the model assumptions and knowing the generation (G) and ventilation (Q) rates. [SRA17] 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.

3.5.1 Prior settings

In Bayesian exposure models, reasonable informative priors are usually used, based on expert knowledge and physical considerations [MBR14]. We assigned informative priors to 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 [MBR11]. The assigned parameter values simulate conditions similar to the real data set in [ZBL09]. In [ZBL09] a test chamber with length of 1.73 m, width of 1.27 m and hight of 1.73 m (volume of 3.8 m^3) was constructed, where a mixing fan was placed to maximize air mixing effect and toluene was released at a rate of $G = 351.5$ mg/min. 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 the difference between the volume of the room and the volume of the near field. A mixing fan was placed such that it maximizes the air-mixing effect. The measured average flow rate was $Q = 13.8$ m^3/min and toluene was generated at a rate of $G = 351.5$ mg/min. For the two zone model, the airflow flow rate was not measured directly but estimated using the steady-state solution, i.e $\beta \to \frac{G}{C_N - C_F} \approx 5$ m^3/min. For the eddy diffusion data, we assumed same test chamber with the same generation rate $G = 351.5$ mg/min, where $D_t = 1$ m^2/min. The assigned value of D_T agrees with the values in literature reported in [SRA17].

In the one-zone and two-zone models, we assume that $G \sim Unif(281, 482)$, $Q \sim Unif(11, 17)$, $K_L \sim Unif(0, 1)$, and $\beta \sim Unif(0, 10)$ in the two-zone model and $D_T \sim Unif(0, 3)$ 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 $\phi \sim Uni(0.5, 3)$ implies that the effective spatial range, i.e., the distance beyond which spatial correlation is negligible, is between 1 and 6 meters.

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 but rather a range. The ranges of the true values in the one-zone 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 $G \sim Unif(30, 150)$, $Q \sim Unif(0, 1)$, $K_L \sim Unif(0, 1)$ in the one-zone and two-zone models and $\beta \sim Unif(0, 5)$ in the two-zone model. For the eddy diffusion model,

the true value for G is 1318 (mg/sec) and from literature [SRA17] the range for D_T is (0.001-0.2) m^2/sec, hence we assigned priors of $G \sim Unif(1104, 1650)$ and $D_t \sim Unif(0, 1)$. Non informative priors were assigned to the variance covariance matrices using $IW(3, I)$ [GCS13].

3.5.2 Simulation results

Monte Carlo filtering methods were used to estimate the latent processes and the model parameters. The effectiveness of the model is assessed through checking whether the 95% C.I.s of the parameters include the true values, MSE, DIC and posterior predictive loss (D=G+P), in addition to graphical assessment.

3.5.2.1 One-zone model

We generated 100 exposure concentrations at equally spaced time points using the exact solution to the ODE in equation (3.7) and the measurements $(y_t, t = 1, \dots T)$ were generated by adding random noise to the true values. The initial concentration $C(0)$ was assigned a value of 1 mg/m^3. Theoretically, the steady state concentration is ≈ 25 mg/m^3. The models applied to the synthetic data and compared are: Gaussian SSM, non-Gaussian SSM in addition to the simple Bayesian nonlinear regression model (BNLR) proposed by [ZBL09]. The Gaussian SSM in (3.13) assumes linearity and Gaussian errors, where the Kalman filter equations are used, where

$$A_t(\theta_c) = \left(1 - \delta_t \frac{Q + K_L V}{V}\right) \quad \text{and} \quad g = \delta_t \frac{G}{V}.$$

Table 3.1 shows the medians and 95% credible intervals of the MCMC posterior samples of the model parameters, MSE, DIC and D=G+P for the three aforementioned models and Figure 3.5 shows the simulated concentrations, measurements and the mean of the posterior samples of the latent states conditional on the measurements, in addition to smoothed estimates obtained from the Non-Gaussian SSM filtered states for one of the simulations. Details of the performances are as follows:

- Non-Gaussian SSM: The 95% C.I.s 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 3.5.

- Gaussian SSM: Gaussian SSM: The 95% C.I.s 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, i.e. it produced noisy estimates for the state process.

- BNLR: The 95% C.I.s include the true values for all the parameters. The model estimates for the latent states are close to the true values.

The D=G+P scores show that the non-Gaussian SSM predictive ability is superior to the Gaussian and BNLR models. The DIC scores show that the Non-Gaussian model has the best model fit followed by the BNLR and then the Gaussian model.

Table 3.1: Posterior predictive loss (D=G+P), MSE, DIC, medians and 95% C.I. of the posterior samples of the one-zone model parameters for the simulated data

Parameter	Non-Gaussian SSM	Gaussian SSM	BNLR
G(351.5)	326.8 (283.3, 351.7)	363.5(314.2,413.8) 1	353.9(292.0,393.8)
Q(13.8)	12.9(11.1, 14.8)	12.8(11.4, 14.3)98	13.0(11.0,15.5)
K_L(0.1)	0.34(0.19,0.78)	0.30(0.28, 0.41)1	0.35(0.0,0.8)
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
DIC	-410	402.1	-184.3

3.5.2.2 Two-zone model

We generated 200 exposure concentrations each, at the near and far fields at equally spaced time points using the exact solution (3.10). 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 concentration

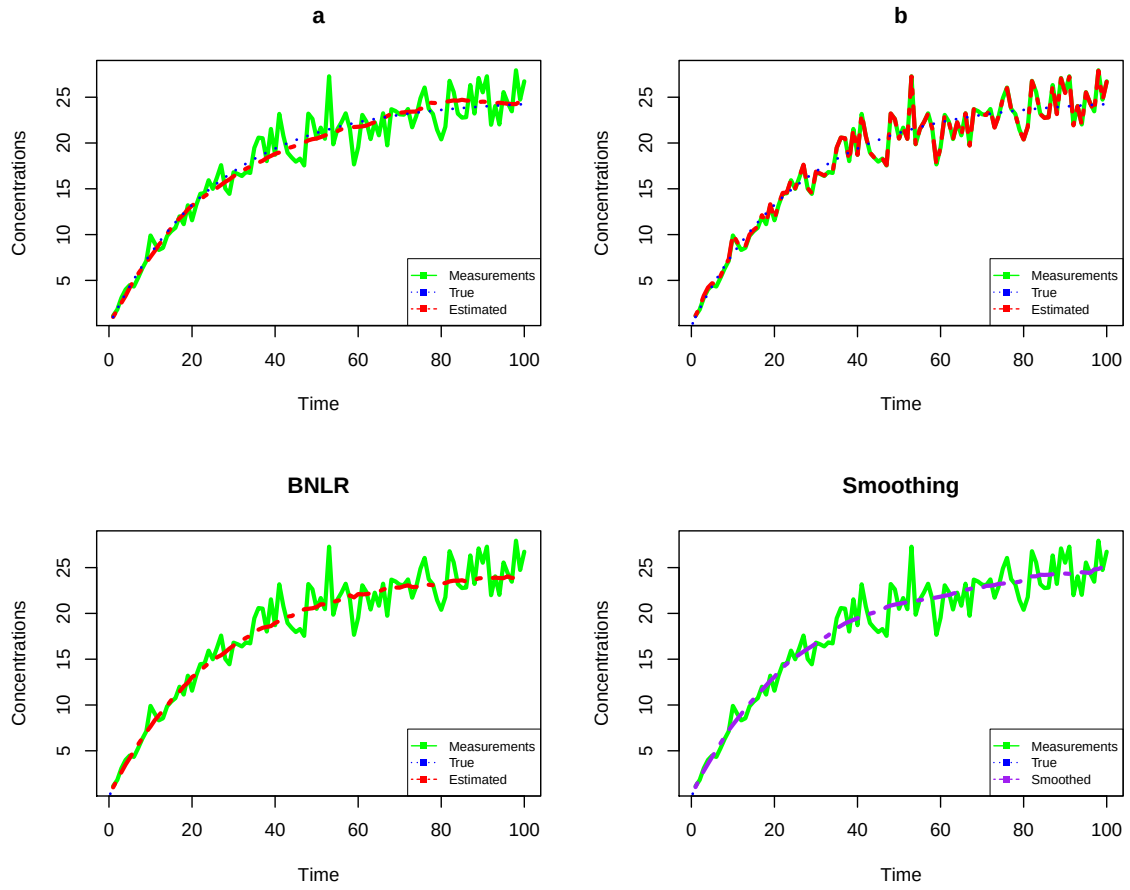

Figure 3.5: Plot of the simulated concentrations, measurements and the mean of the posterior samples of the latent states conditional on the measurements for:
a: Non-Gaussian SSM, b: Gaussian SSM and BNLR

at the near field is $\approx$ 95 mg/m^3, and $\approx$ 25 mg/m^3 at the far field. The Gaussian SSM in (3.13) assumes linearity and Gaussian errors, such that

$$A_t(\theta_c) = \delta_t A + I \quad \text{and} \quad g = \delta_t g.$$

Table 3.2 shows the medians and 95%C.Is of the MCMC posterior samples of the model parameters, MSE, DIC and D=G+P scores. Figure 3.6 shows the simulated concentrations, measurements and the mean of the posterior samples of the latent states conditional on the measurements at the near and the far fields in addition to smoothed estimates obtained from the non-Gaussian SSM filtered states for one of the simulations. We compared the

performance of the two SSMs and the BNLR. Details of the performances of the three models are as follows:

- Non-Gaussian SSM: The 95% C.I.s include the true values for all the parameters. The estimates of the latent states are close to the true values at both the near field and the far field as shown in Figure 3.6.

- Gaussian SSM: Gaussian SSM: The 95% C.I.s for all the parameters except the ventilation rate Q do not include the true values. The model estimates of the latent states are closer to the true values at the near field than the far field, yet all the estimates are noisy.

- BNLR: The 95% C.I.s include the true values for all the parameters. The model estimates of the latent states are closer to the true values at the near field than the far field

The D=G+P scores indicate that the non-Gaussian model provides better predictions and model fit than the BNLR and the Gaussian models. MSE, DIC and Figure 3.6 confirm these results.

Table 3.2: Posterior predictive loss (D=G+P), MSE, DIC, medians and 95% C.I. of the posterior samples of the two-zone model parameters for the simulated data

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)
Q(13.8)	14.7(12.1,16.8)	13.5(11.1, 16.7)	14.4(11.2, 15.8)
K_L(0.1)	0.38(0.02,0.78)	0.22(0.16,0.35)	-
β(5)	5.0(4.3,5.8)	0.40(0.23,1.2)	5.1(4.0, 6.8)
D=G+P	1049840= 1010905+38934.0	1118550= 1033428+85121.7	2504429= 1359016+ 1145413
MSE	15.3	116.1	54.9
DIC	167.2	1618.4	477.6

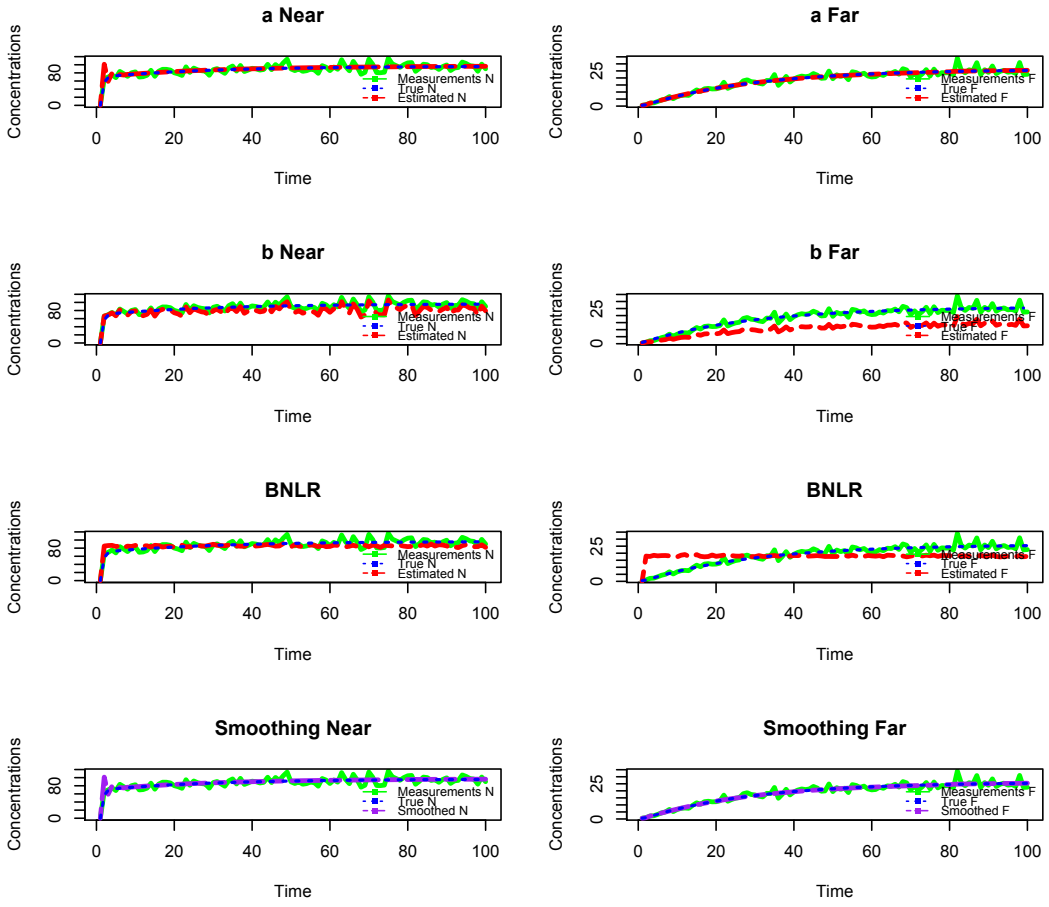

Figure 3.6: Plot of the simulated near and far fields concentrations, measurements and the mean of the posterior samples of the latent states conditional on the measurements for: a: Non-Gaussian SSM, b: Gaussian SSM and BNLR

3.5.2.3 Turbulent eddy diffusion model

We generated 500 exposure concentrations each, at 5 different locations over equally spaced 100 time points using the exact equation (3.11). Table 3.3 shows the medians and 95% C.I.s of the MCMC posterior samples of the model parameters, MSE, DIC and D=G+P. Figure 3.7 shows the simulated concentrations, measurements and the mean of the posterior samples of the latent states conditional on the measurements at three locations and the smoothed estimates obtained from the non-Gaussian SSM filtered states for one of the simulations. Figure 3.8 shows image plot of the posterior mean surface of the latent spatial process $\nu_{t,s}$. The plot indicates higher concentration values near the source of emission at the bottom-left

corner and lower values away from the source. Details of the performance of the two models are as follows:

- Non-Gaussian SSM: The 95% C.I.s include the true values for all the parameters. The estimates of the latent states are close to the true values at the five locations.
- Gaussian SSM: The 95% C.I.s 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 95% C.I.s 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 produced the most accurate parameters estimates. The D=G+P scores indicate that the non-Gaussian model provides better predictions and model fit than the BNLR and the Gaussian models. MSE, DIC and Figure 3.7 confirm these results.

Table 3.3: Posterior predictive loss (D=G+P), MSE, DIC, medians and 95% C.I of the posterior samples of the turbulent eddy diffusion model parameters for the simulated data

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)
D_T(1)	1.2(0.9,1.5)	1.4(1.3,1.6)	1.14(1.03, 1.8)
D=G+P	7062.4=1564.5+5497.9	22025.7=1112.5+20913.1	27719.1=14529.7+13189.5
MSE	3.11	5.55	20.6
DIC	-215.1	1583.2	1068.1

3.5.3 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.

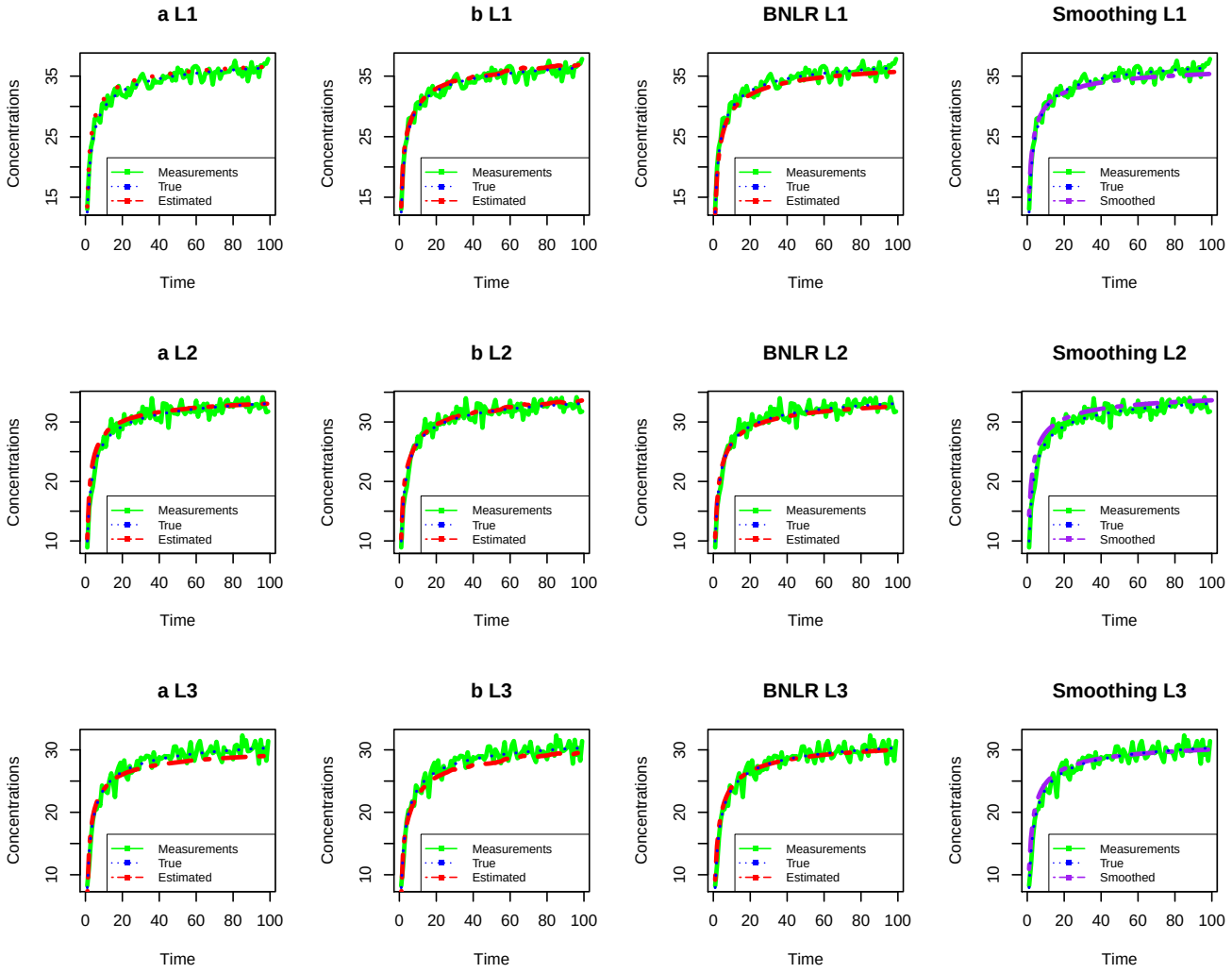

Figure 3.7: Plot of the simulated concentrations, measurements and the mean of the posterior samples of the latent states conditional on the measurements at three locations for:
a: Non-Gaussian SSM, b: Gaussian SSM and BNLR

3.5.3.1 One-zone model

A series of studies were conducted in an exposure chamber under different controlled conditions. [ASR17] 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 $G(\text{mg/min})$ and ventilation $Q(\text{m}^3/\text{min})$ 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

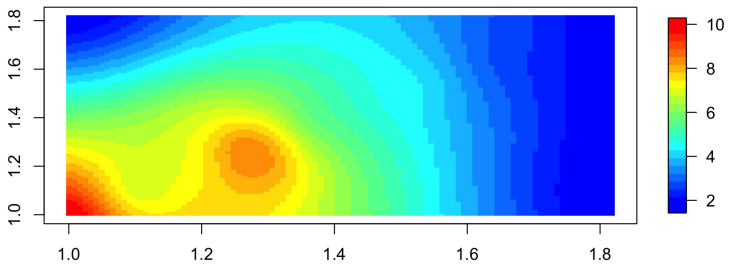

Figure 3.8: Interpolated surface of the mean of the random spatial effects posterior distribution

loss rate K_L was determined from empirical studies to be < 0.01. Solvent concentrations were measured every 1.5 minutes. Details of the experiments can be found in [ASR17].

Table 3.4 shows the medians and 95% C.I.s of the MCMC posterior samples in addition to DIC and D=G+P. The non-Gaussian SSM 95% C.I.s cover the true values for both G and Q, while Gaussian SSM 95% C.I.s include the true values for G at low and high ventilation levels. BNLR 95% C.I.s include the true values for G at high ventilation levels and Q at all levels. Posterior predictive loss (D=G+P) and DIC values indicate better fit for the non-Gaussian SSM model followed by the Gaussian model and finally the BNLR. Figure 3.9 confirms these results.

3.5.3.2 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 [ASR17]. The volume of the far field is 11.79 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 m^3/min. Similar to the one-zone model, three different experimental data sets at three different ventilation levels were used. Table 3.5 shows the medians and 95% C.I.s of the MCMC posterior samples, DIC and D=G+P. At medium and high ventilation rates, non-Gaussian SSM 95% C.I.s include the true values of Q but only at a medium ventilation rate, it includes the true value for G.

Table 3.4: Posterior predictive loss (D=G+P), DIC, medians and 95% C.I. of 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	52257.53= 36044.83+16212.71
	medium		9.8= 0.52+9.2	77.7= 0.20+77.1	16256.04= 3040.128+13215.91
	high		7.5= 1.0+6.5	38.2= 0.1+38.1	4345.8= 237.4+4108.4
DIC	low		-650.1	-640.5	55.8
	medium		-500.6	97.1	155.2
	high		-496	89.4	161.5

The Gaussian SSM 95% C.I.s cover the true value of Q at medium ventilation level but none of the generation rates G. The BNLR 95% C.I.s only cover the true value of Q at a high ventilation level. 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, non-Gaussian SSM 95% C.I.s for β are closer to those values.

DIC and D=G+P scores clearly indicate that non-Gaussian SSM produced better fit than the BNLR and the Gaussian SSM which is also confirmed in Figure 3.10.

3.5.3.3 Turbulent eddy diffusion model

[SRA17] constructed a chamber of size (2.8m $\times$ 2.15m $\times$ 2.0m = 11.9m^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 two minutes. Due to the limited spatial information from the two locations, an unstructured covariance for $\nu_{t,s}$ was used instead of the geostatistical exponential covariance that was considered in the simulation analysis. Non informative prior

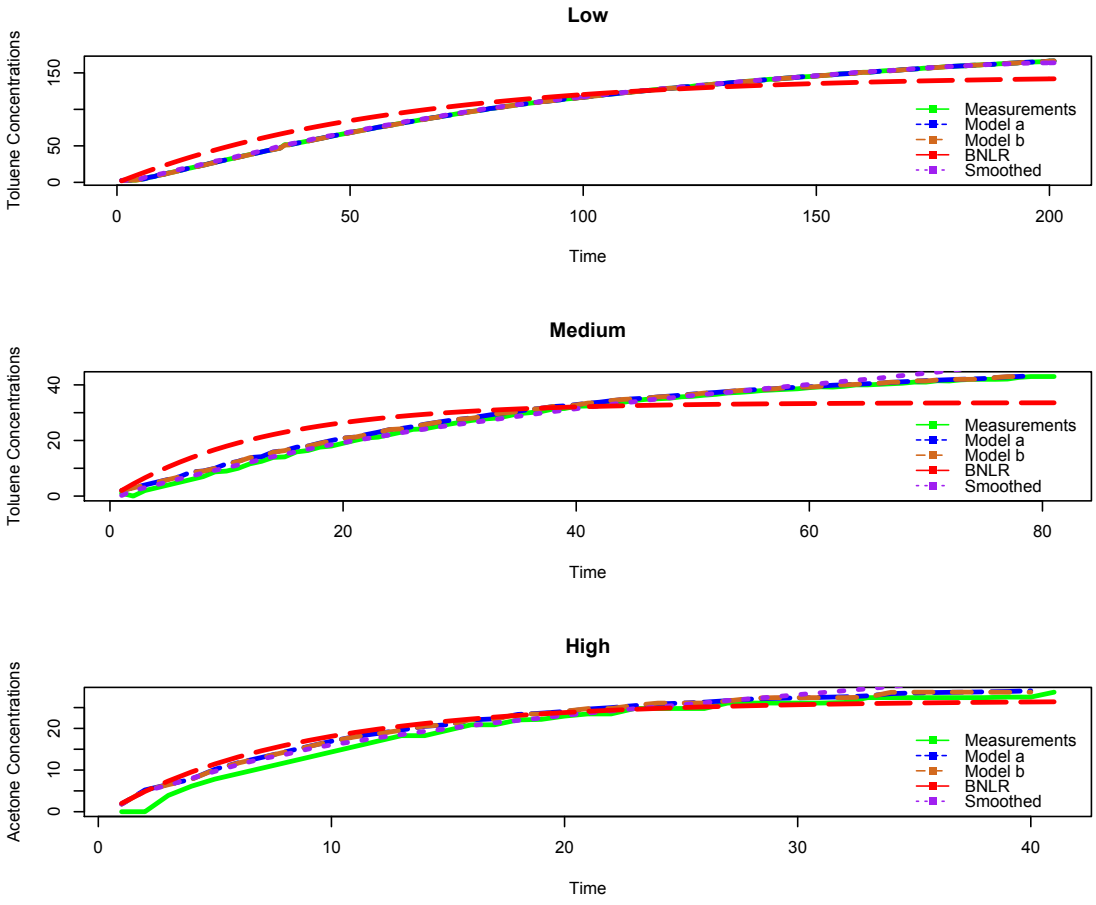

Figure 3.9: Plot of the measured concentrations and the mean of the posterior samples of the latent states conditional on the measurements for:
a: Non-Gaussian SSM, b: Gaussian SSM and BNLR

was assigned to the covariance matrix using $IW(3, I)$ [GCS13].

Table 3.6 shows the medians and 95% C.I.s of the MCMC posterior samples, DIC and the D=G+P. The value of D_T is difficult to measure; hence, the true value is unknown. However, [SRA17] demonstrated that most of the reported values of D_T in literature range from 0.001 to 0.01 m^2/sec. The 95%C.I.s for D_T in non-Gaussian SSM lie within that range. In addition, the 95%C.I.s of G include the true value. The 95%C.I.s of the Gaussian SSM do not include any of the true parameter values. The BNLR 95%C.I. of G does not include the true value and the range for D_t is very narrow. Figure 3.11 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

Table 3.5: Posterior predictive loss (D=G+P), DIC, medians and 95% C.I. of the posterior samples of 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)
	med	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)
	med	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)
	med	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	554650= 554234+416	248358= 73006+ 175352
	medium		22262= 10596+11666	850014= 424452+425562	93267= 16824+76443
	high		20941= 4345+16596	479098= 240278+238820	119212= 64968+54244
DIC	low		-116.0	490978	50.6
	medium		-152.7	3263	79.8
	high		-65.5	8089	-5.2

and in Figure 3.11. DIC and D=G+P scores show that non-Gaussian SSM provides a better fit.

Table 3.6: Posterior predictive loss (D=G+P), DIC, medians and 95% C.I. of the posterior samples of the turbulent eddy diffusion model parameters using toluene solvent

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		100877.8= 59369.9+41507.9	3664659= 3660710+3949.3	6458521= 6289785+168735.6
DIC		-31.6	1222	420.2

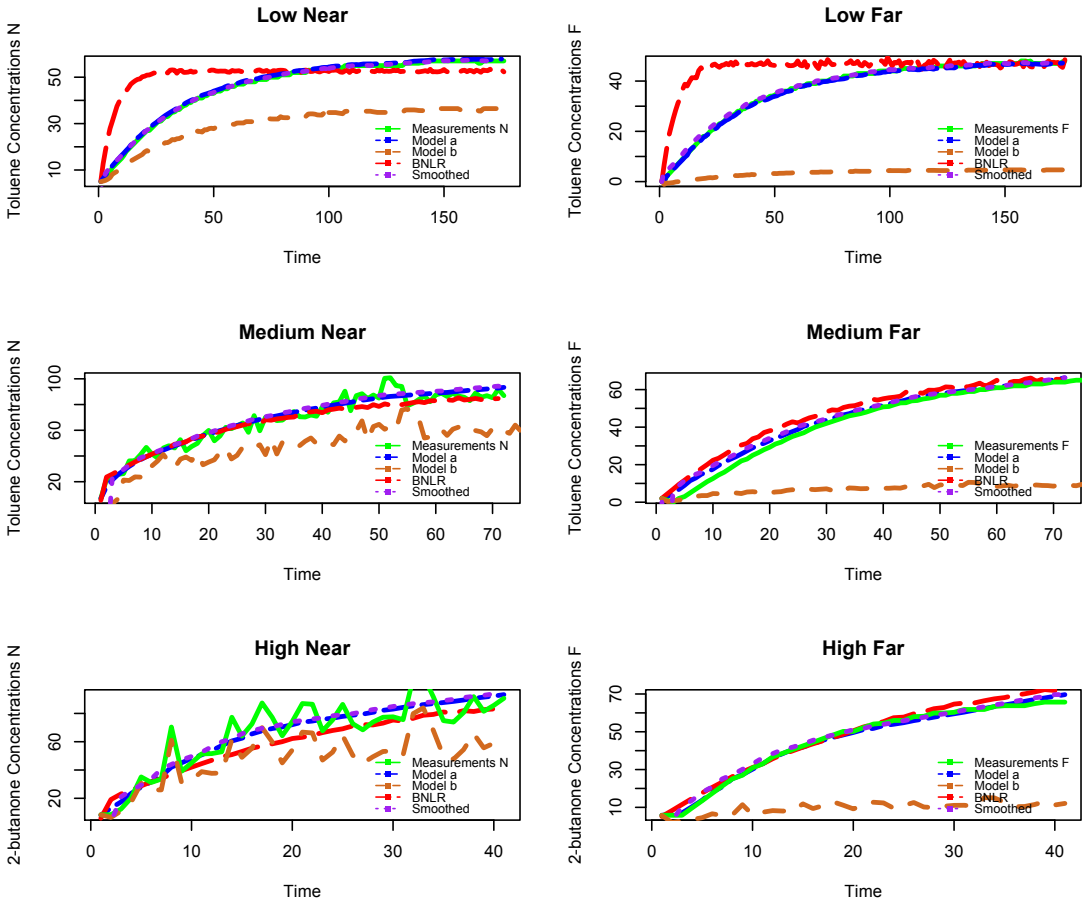

Figure 3.10: Plot of the measured concentrations and the mean of the posterior samples of the latent states conditional on the measurements in the near field and far field for: a: Non-Gaussian SSM, b: Gaussian SSM and BNLR

3.6 Discussion

We have proposed a framework of Bayesian SSMs for analyzing experimental exposure data specific to industrial hygiene. This approach combines information from physical models in industrial hygiene, observed data and prior information. We derive a likelihood by discretizing the physical models. It also expands upon the Gaussian noise assumptions, hence 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, [HYM08] allowed negative values in

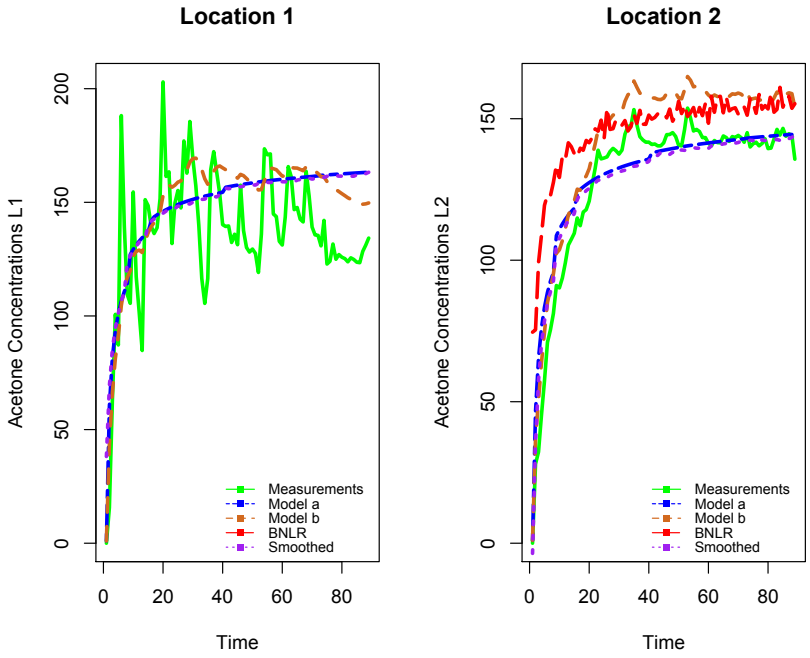

Figure 3.11: Plot of the measured concentrations and the mean of the posterior samples of the latent states conditional on the measurements at the two locations for:
a: Non-Gaussian SSM, b: Gaussian SSM and BNLR

estimating PM_{10} concentrations, while [LCC02] used Kalman filters to predict gas concentrations by using a tuning parameter to fix σ_ω^2 and σ_ν^2 in a one dimensional autoregressive exposure model, rather than pursuing full statistical inference. Our simulation experiments and results demonstrate that Gaussian SSM's may yield extremely poor fits when data are non-Gaussian. This was especially evident for the two-zone analysis. Our results will, we hope, inform the industrial hygiene community about some of the pitfalls of Gaussian SSMs.

Non-Gaussian SSM's tended to perform better than linear Gaussian SSM's, a result that appeared to be consistent across different exposure models and different experimental conditions. Moreover, our analysis revealed that the discretized models outperform the BNLR method proposed by [ZBL09]. 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. Finally, our proposed approach also enjoys better interpretation than the hierarchical Gaussian process models of [MBR14] as they provide greater precisions in

estimates because the random effects in the hierarchical models of [MBR14] tend to inflate variances.

For the experimental data, the performance of the models was better for simpler models. The one-zone model results were superior, followed by the two-zone model results then the eddy diffusion model results. This is not entirely surprising since simpler models imply simpler data and assumptions and possibly fewer parameters. We also believe that the one-zone model is superior because 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 (3.12) 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 SSM's in analyzing eddy diffusion experiments, our chamber data analysis had limited scope because of the very small number of spatial measurements. Another important consideration is misaligned data, such as was considered in [MBR14] 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 [MBR14]. 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.

CHAPTER 4

Nonparametric Bayesian State Space Modeling of Physical Processes in Industrial Hygiene

In exposure data modeling, a parametric model may not always capture the true latent process that generates the observed values. Relaxing the parametric assumptions (e.g. Gaussian noise) might render a more flexible and robust model. Physical models that describe emitted exposure concentrations in a workplace do not consider extraneous factors and may deviate from the true emitted values. It can be unrealistic to assume a distribution on the error terms which can lead to unsatisfactory inference and prediction.

Bayesian nonparametric or semi-parametric methods are highly flexible, but more complex. There is very little work on nonparametric Bayesian state space models. Recent work mostly focused on the functional forms in the transition and measurement equations assuming Gaussian errors. For example, [GMR14] assumed unknown functional forms for the transition and measurement functions, and modeled them as independent Gaussian processes. The approach proposed by [Lau14] to describe the unknown functional form, was based on using penalized splines. Another form of nonparametric modeling includes the use of kernel density approximation to the conditional probability density of the latent state in the update step [GV13]. In addition, some recent work considered Dirichlet processes (DP) in several contexts. For instance, [BC14] used a Poisson-Dirichlet prior on the coefficients in a state space model for clustering. Hierarchical DP hidden Markov model (HDP-HMM) was introduced by [TJB06] to address the problem of clustering of grouped data, where the number of clusters is unknown. [TJB06] considered a DP for each value of the current state, sharing a base measure which is itself a DP. [RVD12] introduced the use of DPM prior on the distribution of the transition error term in a nonlinear SSM in a global positioning system

(GPS) problem where they assumed that the distribution of the error term associated with the measurement equation is Gaussian. [CDD07] proposed Dirichlet process mixture (DPM) priors to provide a nonparametric specification for the distributions of the error terms in linear state space models. We extend upon the model proposed by [CDD07], where we impose constraints on the DPM and consider processes that are not constant in space and time.

We specify the distribution on the error terms as infinite mixture model using Dirichlet processes. [EW95] first introduced Bayesian inference in models for density estimation using Dirichlet processes. The use of a nonparametric Bayesian framework allows very flexible modeling of the physical processes. We offer a general framework using MCMC methods to sample from the posterior distributions for all the model parameters and the predictive distributions. Our method will be demonstrated on computer-simulated data and lab-generated data.

The chapter is organized as follows. In Section 4.1 we discuss the representation of physical models as state-space models with unknown distributions (Section 4.1.1). Specific modeling details are illustrated for common families of exposure models in Section 4.1.4. Section 4.2 describes the models implementations and assessment methods. Section 4.3 illustrates the proposed approach through applications to simulated data and experimental chamber data. We conclude with a critical discussion in Section 4.4.

4.1 Non-Parametric Bayesian Representation of Dynamic Physical Models

A hierarchical construction for a dynamic model of exposure starts by identifying the 'sure thing' relation, which as described by [WH97], arise due to physical and mathematical laws and constraints, which we refer to as the physical model. The physical model represents the structure of the underlying process $\{C_t : t \in \mathcal{T}\}$. In industrial hygiene, physical models are commonly expressed as linear systems of ordinary differential equations (ODEs), which include a source that emits the airborne contaminant at a fixed rate [Nic96, NJ02] and they describe the rate of change in the contaminant over time (dC_t/dt). The solution proposed

utilizes a discretization of the model [ABR18] and uses a state space model (SSM) as a representation of the unobserved state of interest that evolves over time and partial observations that are observed sequentially over discrete time. In the following section, we provide brief review of SSMs in general and illustrate how one can derive a state space representation using a discretized physical model.

4.1.1 Physical Models as State Space Models

Suppose that $\{C_t : t \in \mathcal{T}\}$, $C_t \in \Re^{n_C}$ is a stochastic process, where $\mathcal{T}$ is the index set and n_C is the dimension of the state vector. A state-space model (SSM) is a representation of that unobserved stochastic process that evolves over time, and sequential partial observations $\{Y_t : t \in \mathcal{T}\}$, $Y_t \in \Re^{n_y}$, where t can be taken as discrete time [Fea11] and n_y is the dimension of the observation vector.

In a general setting, the state vector $\{C_t : t \in \mathbb{N}\}$ is assumed to evolve according to the state or transition equation

$$C_t = f_t(C_{t-1}, \omega_{t-1}) \tag{4.1}$$

where f_t is usually a known, possibly nonlinear function in C_{t-1} and $\omega_t \sim P_{\omega_t}$ is an i.i.d. process noise sequence. We want to find filtered or updated estimates of C_t based on the measurements Y_t which are assumed to be related to the state vector according to

$$Y_t = h_t(C_t, \nu_t), \tag{4.2}$$

where h_t is the measurement function and $\nu_t \sim P_{\nu_t}$ is an i.i.d. measurement noise.

As previously mentioned, physical models in industrial hygiene describing the rate of change in the contaminant over time can be represented by differential equations, which gives rise to continuous state space models,

$$\frac{d}{dt}C_t = F_t C_t + g + \omega_t; \qquad Y_t = H_t C_t + \nu_t. \tag{4.3}$$

We want to transfer the continuous equation into a discrete one. The solution to the first

equation in (4.3) when the eigenvalues of F_t are real and distinct can be obtained as follows

$$C_t = \exp(tF_t)C_0 + F_t^{-1} \times [\exp(tF_t) - I]g, \tag{4.4}$$

where $\exp(A)$ denotes the matrix exponential (see [BR14]). The discretized latent state C_t follows an AR(1) transition model, i.e. for small steps of size δ_t, for $t = 1, \ldots, T$, the first equation in (4.3) can be approximated by

$$C_{t+\delta_t} \approx (I + \delta_t F_t)\, C_t + \delta_t g + \omega_t. \tag{4.5}$$

In parametric settings, the distributions of the error terms in the transition (4.1) and measurement (4.2) equations P_{ν_t} and P_{ω_t} are assumed to be known. Gaussian distributions are the most commonly used distributional assumptions as they offer optimal solution under linearity [Eub05]. However, such assumptions may not be valid in some situations. Here, we assume that the distributions P_{ν_t} and P_{ω_t} are unknown and modeled as Dirichlet process mixtures (DPM). Before introducing modeling details for these quantities, we provide a brief review of DPM in the following section.

4.1.2 Dirichelet Process Mixtures

Let $\mathbb{G} \sim \text{DP}(\alpha, \mathbb{G}_0)$ index a Dirichelet Process with precision $\alpha > 0$ and base measure $\mathbb{G}_0$. A DPM is defined as

the convolution of a positive normalized kernel $f_\theta(.)$ with the Dirichlet random probability [Fer83], so that

$$f_{\mathbb{G}}(x_i) = \int f_{\theta_i}(x_i) d\mathbb{G}(\theta_i). \tag{4.6}$$

An alternative representation of the DPM in 4.6 makes use of the so called stick breaking process [IJ01].

Specifically, for $1 \leq h < \infty$, let $w_h = \beta_h \prod_{l<h}(1 - \beta_l)$, $\beta_h \sim \text{Beta}(a_h, b_h)$ and m_h i.i.d draws from the centering measure G_0.

We can write Eq. (4.6) as

$$f_{\mathbb{G}}(x_i) = \sum_{h=1}^{\infty} w_h f_{m_h}(x_i). \tag{4.7}$$

In our work we consider finite dimensional approximations of the process in 4.6, by taking $1 \leq h \leq H$, with H finite [IJ02].

Specifically, we introduce a latent variable ζ, such that $\zeta_i = h$ if observation i came from group h, and define

$$X_i \,|\, \zeta_i = h \overset{ind}{\sim} f_{m_h}(.)\,; \quad \zeta_i \overset{ind}{\sim} Cat(H, w), \tag{4.8}$$

where $\zeta \sim Cat(H, w)$ is a categorical random variable with $P(\zeta = h) = w_h$, for $h = 1, \ldots, H$ and $w = (w_1, \ldots, w_H)'$. The main advantage of this construction shows in the reduced complexity of posterior simulation as we illustrate in Section 4.2.

4.1.3 State Space Models with Unknown Error Distributions

In the proposed model, we assume that the distributions P_{ν_t} and P_{ω_t} in the transition (4.1) and measurement (4.2) equations are modeled as DPMs. Specifically, the distributions P_{ν_t} and P_{ω_t} are assumed Normal with means μ_{ν_t} and μ_{ω_t} and covariances Σ_{ν_t} and Σ_{ω_t} respectively. Let $\theta_{\nu_t} = \{\mu_{\nu_t}, \Sigma_{\nu_t}\}$ and $\theta_{\omega_t} = \{\mu_{\omega_t}, \Sigma_{\omega_t}\}$, a DPM is constructed defining a DP as the base distribution of these parameter vectors, s.t. $\theta_{\nu_t} \sim \mathbb{G}_{0\nu}$, $\theta_{\omega_t} \sim \mathbb{G}_{0\omega}$.

Assuming $\mathbb{G}_\nu \sim DP(\alpha, \mathbb{G}_{0\nu})$, the joint density of Y_t given $\mathbb{G}_\nu$, and C_t, is

$$f(Y_t|\mathbb{G}_\nu, F_t, C_t) = \int N(F_t C_t + \mu_\nu, \Sigma_\nu)\mathbb{G}_\nu(d\theta_{\nu_t}). \tag{4.9}$$

Furthermore, assuming $\mathbb{G}_\omega \sim DP(\alpha, \mathbb{G}_{0\omega})$, the joint density of C_t given $\mathbb{G}_\omega$, and C_{t-1}, is

$$f(C_t|\mathbb{G}_\omega, H_t, C_{t-1}) = \int N(H_t C_{t-1} + \mu_\omega, \Sigma_\omega)\mathbb{G}_\omega(d\theta_{\omega_t}). \tag{4.10}$$

More generally, the proposed nonparametric SSM admits the following hierarchical representation:

$$
\begin{array}{llll}
Y_t \mid C_t, \nu_t & = & f(C_t,\, \nu_t), & C_t \mid C_{t-1}, \omega_t \;\; = \;\; h(C_{t-1},\, \omega_t); \\
\nu_t \mid \mu_{v_t}, \Sigma_{v_t} & \overset{i.i.d}{\sim} & N(\mu_{\nu_t}, \Sigma_{\nu_t}), & \omega_t \mid \mu_{\omega_t}, \Sigma_{\omega_t} \overset{i.i.d}{\sim} N(\mu_{\omega_t}, \Sigma_{\omega_t}); \\
\theta_{\nu_t} \mid \mathbb{G}_\nu & \overset{i.i.d}{\sim} & \mathbb{G}_\nu, & \theta_{\omega_t} \mid \mathbb{G}_\omega \overset{i.i.d}{\sim} \mathbb{G}_\omega; \\
\mathbb{G}_\nu \mid \mathbb{G}_{0\nu} & \sim & \mathrm{DP}(\alpha, \mathbb{G}_{0\nu}), & \mathbb{G}_\omega \mid \mathbb{G}_{0\omega} \sim \mathrm{DP}(\alpha, \mathbb{G}_{0\omega}).
\end{array}
\tag{4.11}
$$

In order to preserve the interpretation of C_t as the filtered process, we consider centering the random measure describing the distribution of ν_t. Specifically, we follow [YDB10] and assume ν_t follows a centered stick-breaking process (CSBP).

Moreover, because we model data in their original scale, non-negativity constraints are applied to the distribution of Y_t and C_t.

Some physical models aim to characterize the distribution of airborne toxicants in more than one location. For example, the turbulent eddy diffusion model model in Section 4.1.4.3 considers a monitoring configuration at multiple worker's locations. Following [GKM05], we readily extend the formulation presented in (4.11) to allow for point-referenced time-series. More precisely, for a finite set of locations $(s_1, \ldots, s_n)$, the joint density of $Y_t = (y_t(s_1), \ldots y_t(s_n))$ given $\mathbb{G}_\theta$, where $\mathbb{G}_\theta \sim DP(\alpha \mathbb{G}_{0\theta})$, and τ^2, is

$$
f(Y_t | \mathbb{G}_\theta, F_t, C_t, \tau^2) = \int N(F_t C_t + \theta_t, \tau^2) \mathbb{G}_\theta(d\theta_t), \tag{4.12}
$$

where $\theta_t = (\theta_t(s_1), \ldots, \theta_t(s_n))$ are realizations from $\mathbb{G}_\theta \sim DP(\alpha, \mathbb{G}_{0\theta})$, where $\mathbb{G}_{0\theta}$ is a random process (e.g. Gaussian process).

4.1.4 Physical Models in Industrial Hygiene

Physical models of airborne toxicant dynamics vary in their goals and levels of complexity [Ram05]. In what follows we apply the general framework introduced in Section 4.1.1 to three commonly used families of physical models; namely: the well-mixed compartment (one-zone) model (Section 4.1.4.1), the two-zone model (Section 4.1.4.2) and the turbulent eddy diffusion model (Section 4.1.4.3).

4.1.4.1 Well-mixed compartment (one-zone) model

The well-mixed compartment physical model assumes that a source is generating a pollutant at a rate G (mg/min) in a room of volume $V(\mathrm{m}^3)$ with ventilation rate $Q(\mathrm{m}^3/\mathrm{min})$. The room is assumed to be perfectly mixed, which means that there is a uniform concentration of the contaminant throughout the room (Figure 4.1). 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

$$V\frac{d}{dt}C_t + (Q + K_L V)\, C_t = G. \tag{4.13}$$

The solution to the differential equation using (B.1) is

$$\begin{aligned} C_t = \exp\{-t(Q + K_L V)/V\}C_{t_0} + ((Q + K_L V)/V)^{-1} \times \\ [1 - \exp\{-t(Q + K_L V)/V\}]\, G/V. \end{aligned} \tag{4.14}$$

The dynamic model after approximating equation (4.13) through discretization, can be expressed as the following measurement and transition equations describing the concentration change from time t to time $t + \delta_t$:

$$\begin{aligned} &\text{Transition:}\quad C_{t+\delta_t} = \left(1 - \delta_t \frac{Q + K_L V}{V}\right) C_t + \delta_t \frac{G}{V} + (\omega_t)\ , \quad \omega_t \overset{iid}{\sim} P_{\omega,\theta_\omega} \\ &\text{Measurement:}\quad Y_t = C_t + \nu_t\ , \quad \nu_t \overset{iid}{\sim} P_{\nu,\theta_\nu} \end{aligned}$$

The dynamic model relies on the unknown distributions P_{ν,θ_ν} and P_{ω,θ_ω}, which are assumed to be DPMs. As outlined in Section 4.1.3 the DPM is constructed through a location-scale

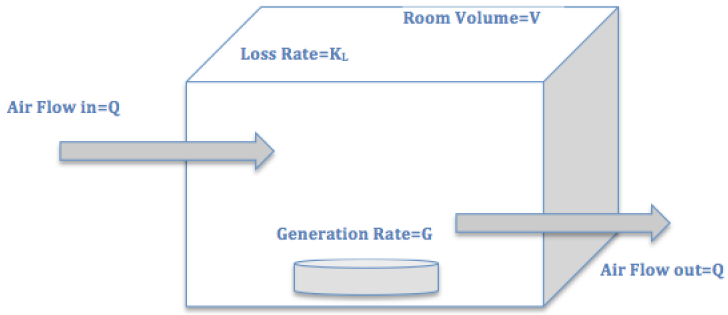

Figure 4.1: One-zone model schematic showing key model parameters; generation rate G, ventilation rate Q and loss rate K_L

mixture of Normal residuals, assuming:

$$\begin{aligned}
\omega_t \mid \mu_{\omega_t}, \sigma_{\omega_t} &\overset{iid}{\sim} N(\mu_{\omega_t}, \sigma_{\omega_t}), \quad \mu_{\omega_t} \mid \mathbb{G}_0^{\mu_\omega} \overset{iid}{\sim} \mathrm{DP}(\mathbb{G}_0^{\mu_\omega}, \alpha), \quad \mathbb{G}_0^{\mu_\omega} = N(\mu_0^\omega, \sigma_0^\omega), \\
&\qquad\qquad \sigma_{\omega_t} \mid \mathbb{G}_0^{\sigma_\omega} \overset{iid}{\sim} \mathrm{DP}(\mathbb{G}_0^{\sigma_\omega}, \alpha), \quad \mathbb{G}_0^{\sigma_\omega} = IG(a_0^\omega, b_0^\omega); \\
\nu_t \mid \mu_{\nu_t}, \sigma_{\nu_t} &\overset{iid}{\sim} N(\mu_{\nu_t}, \sigma_{\nu_t}), \quad \mu_{\nu_t} \mid \mathbb{G}_0^{\nu} \overset{iid}{\sim} \mathrm{DP}(\mathbb{G}_0^{\nu}, \alpha), \quad \mathbb{G}_0^{\nu} = N(\mu_0^\nu, \sigma_0^\nu), \\
&\qquad\qquad \sigma_{\nu_t} \mid \mathbb{G}_0^{\sigma_\nu} \overset{iid}{\sim} \mathrm{DP}(\mathbb{G}_0^{\sigma_\nu}, \alpha), \quad \mathbb{G}_0^{\sigma_\nu} = IG(a_0^\nu, b_0^\nu).
\end{aligned} \tag{4.15}$$

The full one-zone Bayesian SSM is completed by placing prior distributions on the parameters:

$$Q, G, K_L, \alpha \sim U(Q;\, a_Q, b_Q) \times U(G;\, a_G, b_G) \times U(K_L;\, a_{K_L}, b_{K_L}) \times Ga(\alpha;\, a_\alpha, b_\alpha).$$

Theoretically, the steady state concentration is the limit of C_t as $t \to \infty$ which is G/Q (mg/m^3). Details of the steady state solution are provided in [ABR18].

4.1.4.2 Two-zone model

The two zone model assumes the presence of a contaminant in the workplace and that the region closer to the source is called the *"near field"* while the rest of the room is called the far *"far field"*, which completely encloses the near field. Both fields are assumed to be

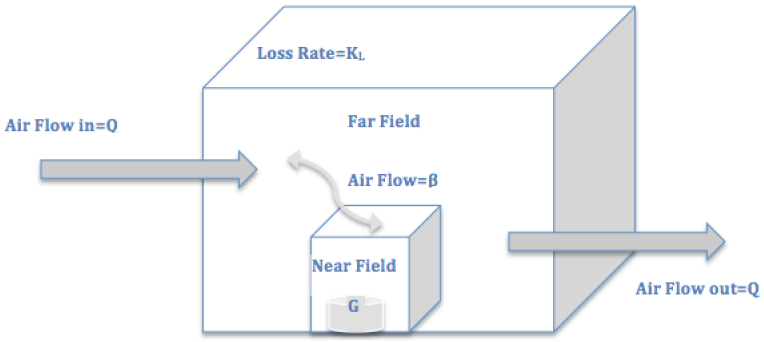

Figure 4.2: Two-zone model schematic showing key model parameters; generation rate G, ventilation rate Q, airflow β and loss rate K_L

a well-mixed box, i.e., 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 $G(\text{mg/min})$, in a room with supply and exhaust flow rates (ventilation rate) $Q(\text{m}^3/\text{min})$ and loss rate by other mechanisms $K_L(\text{mg/m}^3)$. This model includes one more parameter that indicates the airflow between the near and the far field $\beta(\text{m}^3/\text{min})$. The volume in the near field is denoted by $V_N(\text{m}^3)$ and the volume in the far field is denoted by $V_F(\text{m}^3)$. Figure 4.2 illustrates the dynamics of the system.

The following system of differential equations represents the two-zone model

$$\overbrace{\frac{d}{dt}\begin{bmatrix} C_N(t) \\ C_F(t) \end{bmatrix}}^{dC_t} = \overbrace{\begin{bmatrix} -\beta/V_N & \beta/V_N \\ \beta/V_F & -(\beta+Q)/V_F + K_L \end{bmatrix}}^{A} \overbrace{\begin{bmatrix} C_N(t) \\ C_F(t) \end{bmatrix}}^{C_t} + \overbrace{\begin{bmatrix} G/V_N \\ 0 \end{bmatrix}}^{g} \tag{4.16}$$

The solution to the differential equations using (B.1) is

$$C_t = \exp\{tA\}C_{t_0} + A^{-1}\left[\exp\{tA\} - I\right]g. \tag{4.17}$$

We approximate the differential equations in (4.16) by representing time as discrete points. So, for a δ_t unit change in time, the state space model equations are

$$\text{Transition: } C_{t+\delta_t} = (\delta_t A(\theta_c; x) + I)C_t + \delta_t g(\theta_c; x) + \omega_t\,,\ \omega_t \overset{iid}{\sim} P_{\omega_t,\theta_\omega},$$

$$\text{Measurement: } Y_t = C_t + \nu_t\,,\ \nu_t \overset{iid}{\sim} P_{\nu_t,\theta_\nu};\ \text{where,}$$

$$Y_t = \begin{bmatrix} Y_N(t) \\ Y_F(t) \end{bmatrix},\ C_t = \begin{bmatrix} C_N(t) \\ C_F(t) \end{bmatrix},\ A(\theta_c; x) = \begin{bmatrix} -\frac{\beta}{V_N} & \frac{\beta}{V_N} \\ \beta/V_F & -\frac{(\beta+Q)}{V_F} + K_L \end{bmatrix} \text{ and } g(\theta_c; x) = \begin{bmatrix} \frac{G}{V_N} \\ 0 \end{bmatrix}.$$

Similar to the one-zone model, we assume the distributions for ν_t and for ω_t to be DPMs, constructed as follows:

$$\begin{aligned} \omega_t \mid \mu_{\omega_t}, \Sigma_{\omega_t} &\overset{iid}{\sim} N(\mu_{\omega_t}, \Sigma_{\omega_t}), & \mu_{\omega_t} \mid \mathbb{G}_0^{\mu_\omega} &\overset{iid}{\sim} \mathrm{DP}(\mathbb{G}_0^{\mu_\omega}, \alpha), & \mathbb{G}_0^{\mu_\omega} &= N(\mu_0^\omega, \Sigma_0^\omega), \\ & & \Sigma_{\omega_t} \mid \mathbb{G}_0^{\Sigma_\omega} &\overset{iid}{\sim} \mathrm{DP}(\mathbb{G}_0^{\Sigma_\omega}, \alpha), & \mathbb{G}_0^{\Sigma_\omega} &= IW(r, S); \\ \nu_t \mid \mu_{\nu_t}, \Sigma_{\nu_t} &\overset{iid}{\sim} N(\mu_{\nu_t}, \Sigma_{\nu_t}), & \mu_{\nu_t} \mid \mathbb{G}_0^{\nu} &\overset{iid}{\sim} \mathrm{DP}(\mathbb{G}_0^{\nu}, \alpha), & \mathbb{G}_0^{\nu} &= N(\mu_0^\nu, \sigma_0^\nu), \\ & & \Sigma_{\nu_t} \mid \mathbb{G}_0^{\Sigma_\nu} &\overset{iid}{\sim} \mathrm{DP}(\mathbb{G}_0^{\Sigma_\nu}, \alpha), & \mathbb{G}_0^{\Sigma_\nu} &= IW(r, S); \end{aligned} \tag{4.18}$$

with priors

$$\begin{aligned} Q, G, K_L, \beta, \alpha \ \sim\ & U(Q;\, a_Q, b_Q) \times U(G\ a_G, b_G) \times U(K_L;\, a_{K_L}, b_{K_L}) \times \\ & \times U(\beta;\, a_\beta, b_\beta) \times Ga(\alpha,\ a_\alpha, b_\alpha). \end{aligned}$$

Theoretically, for large values of t, the steady state concentration in the near field is $G/Q + G/\beta$ (mg/m^3), and G/Q (mg/m^3) in the far field.

4.1.4.3 Turbulent eddy diffusion model

In real workplace settings, the room may neither be perfectly mixed nor consist of well-mixed zones. A popular model for such settings is the turbulent eddy diffusion model. This model accounts for a continuous concentration gradient from the source outward. It takes into

account 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 coordinates and time t. The parameter that is unique to this model is the turbulent eddy diffusion coefficient $D_T(\text{m}^2/\text{min})$. It describes how quickly the emission spreads with time (Figure 4.3) and is assumed to be constant over space and time. There has been very little research on the values of D_T due

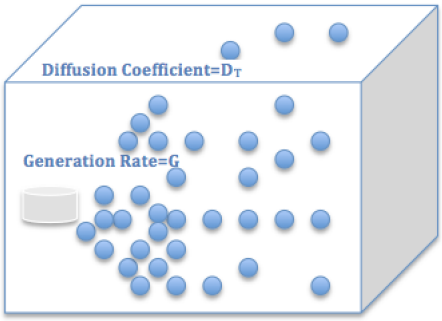

Figure 4.3: Eddy diffusion model schematic showing key model parameter; diffusion coefficient D_T

to the difficulty of measuring it. Some studies suggest a relationship between D_T and air change per hour (ACH) [SRA17].

Suppose we observe the contaminant at m different locations, the contaminant concentration at location s relative to the source of emission is

$$C_{s,t} = \frac{G}{2\pi D_T \|s\|} \left\{ 1 - \text{erf}\left(\frac{\|s\|}{\sqrt{4D_T t}} \right) \right\}, \tag{4.19}$$

where $\text{erf}(z) = \frac{2}{\pi} \int_0^z \exp(-u^2) du$. The following differential equation represents the change in concentration over time

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

We approximate the differential equation by representing time as discrete points. For a δ_t

unit change in time, the state space model equations are

Transition:

$$C_{s,t+\delta_t} = C_{s,t} + \delta_t \frac{G}{4(D_T \pi t)^{3/2}} \exp\left(-||s||^2/4D_T t\right) + \omega_{t,s}.$$

Measurement:

$$Y_{s,t} = C_{s,t} + \nu_{s,t} + \eta_t \ , \ \nu_{s,t} \sim P_{\nu_{s,t},\theta_\nu};$$

with $\omega_{t,s} \sim P_{\omega_{t,s},\theta_\omega}$ and $\eta_t \sim P_{\eta_t,\theta_\eta}$. The dynamic model relies on the distributions $P_{\omega_{t,s},\theta_\omega}$, $P_{\nu_{s,t},\theta_\nu}$ and P_{η_t,θ_η}. The full eddy diffusion Bayesian SSM is completed by placing prior distributions on the parameters;

$$\begin{aligned}
&C_{s,t+\delta_t} = C_{s,t} + \delta_t \frac{G}{4(D_T \pi t)^{3/2}} \exp\left(-||s||^2/4D_T t\right) + \omega_{s,t};\\
&Y_{s,t} = C_{s,t} + \nu_{s,t} + \eta_t;\\
&\omega_{s,t} \overset{iid}{\sim} N(\mu_\omega, \sigma_\omega);\ \mu_\omega|\mathbb{G}_{\mu_\omega} \overset{iid}{\sim} \mathbb{G}_{\mu_\omega},\ \mathbb{G}_{\mu_\omega} \sim \mathrm{DP}(\mathbb{G}_0^{\mu_\omega}, \alpha),\ \mathbb{G}_0^{\mu_\omega} = N(\mu_0^\omega, \sigma_0^\omega) \qquad (4.20)\\
&\sigma_\omega|\mathbb{G}_{\sigma_\omega} \overset{iid}{\sim} \mathbb{G}_{\sigma_\omega},\ \mathbb{G}_{\sigma_\omega} \sim \mathrm{DP}(\mathbb{G}_0^{\sigma_\omega}, \alpha),\ \mathbb{G}_0^{\sigma_\omega} = IG(a_0^\omega, b_0^\omega)\\
&\eta_t \overset{iid}{\sim} N(0, \sigma_\eta);\ \nu_{s,t}|\mathbb{G}_\nu \overset{iid}{\sim} \mathbb{G}_\nu,\ \mathbb{G}_\nu \sim \mathrm{DP}(\mathbb{G}_0^\nu, \alpha), \quad \mathbb{G}_0^\nu = GP(0, K_{\theta_\nu})\\
&Q, D_T, \sigma_\eta^2, \alpha \sim p(\theta) = Unif(a_Q, b_Q) \times Unif(a_{D_T}, b_{D_T}) \times\\
&\qquad\qquad IG(a_{\sigma_\eta}, b_{\sigma_\eta}) \times Ga(a_\alpha, b_\alpha)\ ,
\end{aligned}$$

Note that $\nu_{s,t}$ is a spatial-temporal process discrete in time and continuous in space, while $\omega_{s,t}$ ideally is a process continuous in both space and time since it models spatial-temporal associations between concentrations at arbitrary space-time coordinates. We extend upon the spatial DP introduced by [GKM05] by considering different structures for $\nu_{s,t}$. For example, one could treat time as discrete and evolving for each location s, representing it as an autoregressive process, such that $\nu_{s,t} = \gamma\nu_{t-1}(s) + \psi_t(s)$ with $\psi_t(s)$ being spatial processes independent across time (see, e.g.,[WC99, GBG]). In that case, we assume $\psi_t(s)|\mathbb{G}_\psi \overset{iid}{\sim} \mathbb{G}_\psi$,where $\mathbb{G}_\psi \sim \mathrm{DP}(\mathbb{G}_0^\psi, \alpha)$, and $\mathbb{G}_0^\psi = GP(0, K_{\theta_\psi})$ with geostatistical covariance K_{θ_ψ}, an

$m \times m$ spatial covariance matrix. Other spatial-temporal structures that represent richer hypotheses and more flexible modeling, where classes of non-separable (NS) covariances will also be considered [CH99]. Covariance functions of this type allow for more flexibility in modeling space-time interactions.

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).

4.2 Model Implementation and Calibration

4.2.1 Implementation

Let θ be a collection of all unknown parameters, and $x_{1:T}$ be the vector of hidden states. Bayesian inference for SSMs is based on the joint posterior distribution of the hidden states and the parameters $p(\theta, x_{1:T} \mid y_{1:T})$, where $y_{1:T}$ are the observed concentrations. We can write that posterior distribution as

$$p(\theta, x_{1:T} \mid y_{1:T}) \propto p(\theta)p(x_1 \mid \theta) \prod_{t=2}^{T} p(x_t \mid x_{t-1}, \theta) \prod_{t=1}^{T} p(y_t \mid x_t, \theta).$$

Monte Carlo samples from this distribution are often obtained through Markov transitions updating θ conditional on $x_{1:T}$, then updating $x_{1:T}$ conditional on θ.

For linear Gaussian models, efficient sampling strategies are implemented through the Kalman filter (KF) [Eub05]. However, inference under nonlinear and non-Gaussian assumptions can be challenging. In this setting, several algorithms have been proposed, including: the Extended Kalman filter [Jaz07], particle filtering strategies [GSS93], and Metropolis Hastings corrected versions [ADH10].

We consider updating $x_{1:T}$ a single component at a time [Fea11, WH97]. While not particularly efficient, this strategy can be implemented in standard generic samplers, and is readily extended to the nonparametric case. A `Jags` implementation based on the finite DP representation of [IJ02] is implemented under the `R` computational environment. Code

is provided in a supplementary document.

4.2.2 Calibration

We frame the question of uncertainty quantification from a predictive perspective. Specifically we evaluate model forecasts in terms of their calibration and sharpness [GFE07]. Model calibration refers to the statistical consistency between the model forecasts and the observations and is a property of both the observations and the forecasts, sharpness on the other hand, refers to the concentration of the predictive distribution, and hence is a property only of the forecasts.

Due to the existence of uncertainty in the forecasts, forecast distributions F in the form of Monte Carlo samples from the posterior predictive distribution rather than point forecasts are of main interest.

We consider graphical evaluation, summary measures and scoring rules to assess the different models. Following the definitions in [GFE07], we assume that G_t, $t = 1, \dots T$, is the true distribution generating the observations y_t, and F_t is the model's predictive CDF. The ideal forecaster will render $G_t = F_t$, $t = 1, \dots T$. Since the true distribution G_t is unknown, the predictive model would be assessed based on the predictive distribution F_t and the observations y_t, which motivates the use of the probability integral transform (PIT), first proposed by [Daw84]. PIT is used to asses probabilistic calibration (i.e $\frac{1}{T}\sum_{t=1}^{T} G_t \circ F_t^{-1}(p) \rightarrow p \quad \forall p \in (0,1)$), where uniformity of the PIT values $p_t = F_t(y_t)$, $t = 1, \dots, T$, implies an ideal forecaster.

Marginal calibration (i.e $\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} G_t(y) = \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} F_t(y)$, for all $y \in \mathbb{R}$), can be assessed by plotting the difference of the two CDFs, $\bar{F}(y) = \lim_{T\to\infty} \left\{ \frac{1}{T}\sum_{t=1}^{T} F_t(y) \right\}$ and $\hat{G}_T(y) = \frac{1}{T}\sum_{t=1}^{T} 1(y_t \leq y)$ and evaluating fluctuation patterns around 0.

Graphical evaluation of the width of the 50% and 90% prediction intervals are used to assess sharpness of the predictive distribution. These intervals are produced from the MCMC output of the posterior predictions.

In addition, we consider scoring rules denoted by $s(F, y)$ [GFE07], which are numerical measures that assess calibration and sharpness. For example, the continuous ranked probability score (crps) defined in [GR07] as

$$\text{crps}(F, y) = \int_{-\infty}^{\infty} \{F(s) - 1(y \geq s)\}^2 ds, \tag{4.21}$$

can be used. The average crps $\text{CRPS} = \frac{1}{T}\sum_{t=1}^{T} \text{crps}(F_t, y_t)$ corresponds to the total area between the CDF of the forecast and the CDF of the observation, hence a smaller value implies a better forecast.

Finally, let $\hat{C}_t$ and C_t be the estimated (posterior mean) and true values at time point t respectively, In our simulations, model performance is evaluated through the mean square error (MSE=$\sum_{t=1}^{T}(\hat{C}_t - C_t)^2/T$).

4.3 Data Analysis

In this section, we evaluate the performance of the models discussed in Section **??**, using computer-simulated datasets as well as experimental lab-generated data. Monte Carlo filtering methods were used to estimate the latent processes and the model parameters as discussed in Section 4.2.1. The effectiveness of the methods proposed are assessed through graphical evaluations, summary measures and scoring rules as discussed in Section 4.2.2. Moreover, we compare the performance of the proposed nonparamteric SSM to the parametric SSM discussed in the previous chapter. The prior settings are based on physical knowledge and experience, and discussed in the following section.

4.3.1 Simulation results

The computer-simulated data was generated using R computing environment. To investigate the performance of the proposed framework, we simulated concentrations at different signal-to-noise ratios under the one-zone, two-zone and eddy diffusion model scenarios as follows.

4.3.1.1 One-zone model

We conducted 100 simulations at different signal to noise ratios of 100 exposure concentrations each, at equally spaced time points using the exact solution to the ODE in equation (3.7). The initial concentration $C(0)$ was assigned a value of 1 mg/m^3. Theoretically, the steady state concentration is $G/Q \approx 25$ mg/m^3. Results of the nonparamteric SSM in Eq. 4.15 are evaluated and compared to the parametric SSM in Table 4.1.

Table 4.1: CRPS, empirical coverage of the forecasts, medians and 95% C.I. of the posterior samples of the one-zone model parameters for the simulated data (averaging over 100 simulations at different signal to noise ratios)

Parameter	DP	Parametric
G(351.5)	357.1 (322.5, 419.3)	326.8 (283.3, 351.7)
Q(13.8)	12.9(11.1, 15.5)	12.9(11.1, 14.8)
K_L(0.1)	0.29(0.02,0.70)	0.34(0.19,0.78)
CRPS from 100 simulations	1.4	2.1
MSE from 100 simulations	0.7(1.8)	0.6 (0.7)

4.3.1.2 Two-zone model

We conducted 100 simulations at different signal to noise ratios of 200 exposure concentrations each, at the near and far fields at equally spaced time points using the exact solution (3.10). 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 concentration at the near field is $G/Q + G/\beta \approx 95$ mg/m^3, and $G/Q \approx 25$ mg/m^3 at the far field. Results of the nonparamteric SSM in Eq. 4.18 are evaluated and compared to the parametric SSM in Table 4.2.

4.3.1.3 Turbulent eddy diffusion model

We conducted 50 simulations at different signal to noise ratios of of 500 exposure concentrations each, at 5 different locations over equally spaced 100 time points using the exact equation (3.11) at three different scenarios, where different covariance structures are used in

Table 4.2: CRPS, empirical coverage of the forecasts, medians and 95% C.I. of the posterior samples of the two-zone model parameters for the simulated data (averaging over 100 simulations at different signal to noise ratios)

Parameter	DP	Parametric
$G(351.5)$	297.8(281.5,363.1)	307.3(283.1,345.5)
$Q(13.8)$	12.8(11.1,16.2)	13.6(11.4,16.1)
$K_L(0.1)$	0.30(0.02,0.48)	0.38(0.02,0.78)
$\beta(5)$	4.2(3.9,5.2)	4.3(3.9,4.9)
CRPS	6.4	8.6
MSE	27.0	20.3

the simulation. The first scenario assumes that the error term ν_t is a random noise, hence the variation of y_t around C_t does not depend on time or location. The second scenario assumes an autoregressive process such that, for each location s, $\nu_{t,s} = \nu_{t-1}(s) + \psi_t(s)$ with $\psi_t(s)$ being a Gaussian process with a simple geostatistical covariance $K_{\theta_\psi} = \sigma^2 e^{-\phi\|s-s'\|}$, where $\|s - s'\|$ is the squared Euclidean distance between s and s', σ^2 and $1/\phi$ are the partial sill and the effective spatial range respectively. The third scenario representing more flexible modeling, where a NS covariance was used; $K_{\theta_\nu} = \frac{\sigma^2}{(a^2|t-t'|^2+1)^{d/2}} \exp\left\{-\frac{b^2\|s-s'\|^2}{a^2|t-t'|+1}\right\}$, where $a \geq 0$ is the scaling parameter of time, $b \geq 0$ is the scaling parameter of space and σ^2 is the covariance when $\|s - s'\|$ and $|t - t'|$ are equal to 0 [CH99]. Even though this class of NS covariances can be computationally demanding, it is appropriate for this application, since typically measurements are available over few distinct locations in a workplace chamber, hence estimating the processes will be feasible.

Two different models with different assumptions for $\nu_s(t)$ are considered in fitting the nonparamteric SSM in Eq. 4.20. The first is the additive AR model, where $\nu_{t,s} = \nu_{t-1}(s) + \psi_t(s)$, where $\psi_t(s)$ is a DP with GP as the base distribution and the second is the NS covariance model where $\nu_{t,s}$ is a DP with GP as the base distribution using a NS covariance. Results are evaluated and compared to the parametric SSM in Table 4.3.

Results were consistent across the three physical models. The 95% C.I.s include the true values for all the parameters and the latent state estimates are very close to the true simulated values. The values of CRPS indicate better calibration and sharpness among the

nonparametric models. Results of the eddy diffusion model varied by different assumptions, where the NS model consistently showed better calibration than the additive AR model. Details of the simulation results are reported in the supplementary material.

Table 4.3: CRPS, empirical coverage of the forecasts, medians and 95% C.I of the posterior samples of the turbulent eddy diffusion model parameters for three simulation scenarios

	Parameter	additive AR	NS1	Parametric
random	$G(351.5)$	376.9(285.0,476.1)	401.0(296.0, 474.8)	368.7(284.4,476.2)
	$D_T(1)$	1.2(0.9,1.5)	1.3(1.0,1.7)	1.3(1.0,1.6)
	CRPS	0.80	0.67	2.5
	MSE	387.9(26.9)	367.2(2.0)	361.7(8.4)
additive AR	$G(351.5)$	432.9(319.5,479.3)	407.0(288.0, 479.0)	370.9(285.0,475.6)
	$D_T(1)$	1.3(1.0,1.6)	1.5(1.0,2.9)	1.2(1.0,1.6)
	CRPS	0.8	0.7	2.2
	MSE	385.3(35.0)	367.4(2.7)	372.9(8.4)
NS	$G(351.5)$	405.2(289.0,477.2)	398.4(308.5,476.2)	365.5(284.8,474.0)
	$D_T(1)$	1.3(1.0,1.6)	1.4(1.0,2.6)	1.3(1.0,1.6)
	CRPS	0.64	0.70	2.1
	MSE	403.1(42.9)	368.7(7.0)	367.5(4.7)

4.3.2 Experimental Chamber Data Results

In this section, we study the performance of the models on controlled lab-generated data. The experiments were conducted in test chambers where solvent concentrations have been measured under different scenarios. [ASR17] examined parts of this data using the deterministic one-zone and two-zone models and showed that performance is highly reliable on the model assumptions and knowing the generation (G) and ventilation (Q) rates. [SRA17] studied the eddy diffusion data using a deterministic model and concluded that it is suitable for indoor spaces with persistent directional flow towards 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.

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. Details of the experimental chamber data results are as follows.

4.3.2.1 One-zone model

A series of studies were conducted in an exposure chamber under different controlled conditions. [ASR17] constructed a chamber of size (2.0m $\times$ 2.8m $\times$ 2.1m = 11.8m^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 minutes. Details of the experiments can be found in [ASR17]. Table 4.4 shows that estimates for the ventilation rate are more accurate using the non-parametric DP model, while estimates for the generation rate (G) are not accurate in either models. In both models, there is more uncertainty in the estimates when the ventilation level is high. The CRPS values show better calibration among the DP model throughout the three different ventilation level scenarios. Figure 4.4 shows that the estimated latent concentrations are close to the measurements. Results show better model calibration among the low and medium ventilation level scenarios which is reflected by the CRPS values. One reason might be the much lower number of observations in that particular data set.

Table 4.4: CRPS, empirical coverage of the forecasts, medians and 95% C.I. of the posterior samples of the one-zone model parameters using toluene and acetone solvents

Parameter	Ventilation level	True value	DP	Parametric
	low	43.2	100.5(93.1,103.7)	38.1(30.2,62.9)
G	medium	43.2	136.1(125.7,146.1)	141.6(130.6,149.7)
	high	39.55	82.4(54.0,142.4)	81.7(32.9,142.4)
	low	0.04-0.07	0.03(0.008, 0.04)	0.27(0.02, 0.41)
Q	medium	0.23-0.27	0.11(0.01,0.80)	0.50(0.02,0.97)
	high	0.47-0.77	0.45(0.02,0.98)	0.59(0.03,0.98)
	low		0.03	0.17
crps	medium		0.04	0.08
	high		0.05	0.10

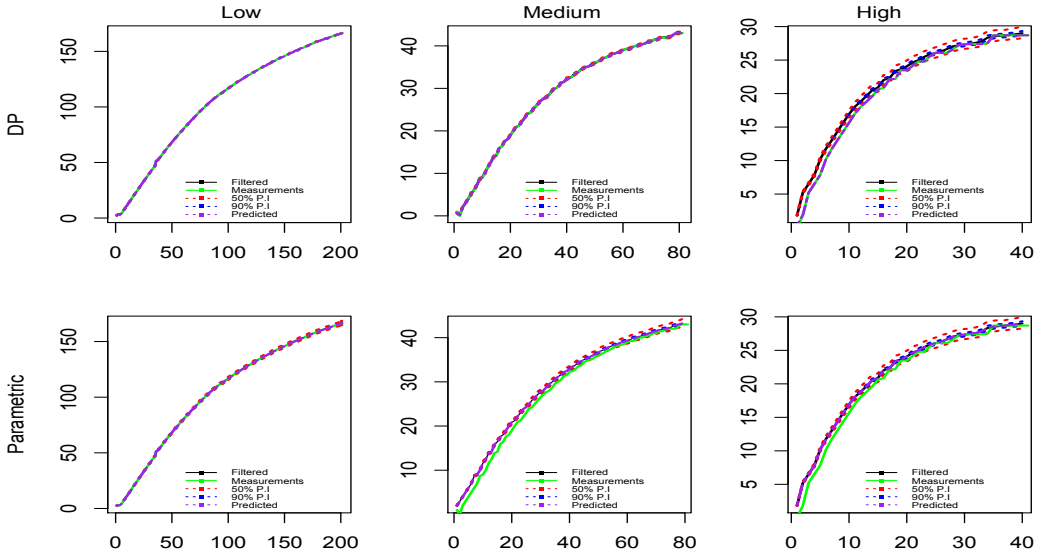

Figure 4.4: Plot of the measured concentrations and the mean of the posterior samples of the latent states conditional on the measurements

4.3.2.2 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 [ASR17]. The volume of the far field is 11.79 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 m^3/min. Table 4.5 shows that, while the point estimates in the DP model results are more accurate, the 95% C.I. of the posterior samples for both the generation rate (G) and the ventilation rate (Q) for the DP and the parametric models include the true values except for G at low ventilation level. Similar to the one-zone model, the CRPS values show better calibration among the DP model throughout the three different ventilation rate scenarios, the highest being at the high ventilation level, perhaps reflecting the higher noise among the lower number of observations as shown in Figure 4.5. Results show better calibration at the far field throughout.

Table 4.5: CRPS, empirical coverage of the forecasts, medians and 95% C.I. of the posterior samples of the two-zone model parameters using toluene and acetone solvents

Parameter	Ventilation level	True value	DP	Parametric
	low	43.2	30.4(30.0, 32.2)	30.4(30.0, 32.2)
G	med	86.4	83.5(76.1,114.6)	73.7(60.2,90.5)
	high	120.7	108.6(75.4,126.1)	49.8(33.9,68.3)
	low	0.04-0.07	0.15(0.03, 0.67)	0.68(0.09, 0.98)
Q	med	0.23-0.27	0.36(0.01,0.91)	0.38(0.11,0.50)
	high	0.47-0.77	0.42(0.03,0.93)	0.46(0.45,0.98)
	low	0.24-1.24	4.4(3.7,4.9)	3.0(2.3,3.7)
β	med	0.24-1.24	3.5(2.7,4.4)	2.9(2.5, 3.4)
	high	0.24-1.24	2.9(2.5, 3.4)	2.2(1.5, 2.8)
	low		0.17	6.2
CRPS	medium		1.6	5.5
	high		0.44	7.9

4.3.2.3 Turbulent eddy diffusion model

[SRA17] 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 0.41m and 1.07m away from the source every two minutes. Table 4.6 shows more parameter estimation uncertainty at the parametric model compared to the NS covariance and the additive AR nonparametric models. Despite the high noise at location 1, the NS covariance model was able to provide more smooth, accurate state estimates and better calibration compared to the additive AR model which is reflected in Figure 4.6.

Table 4.6: CRPS, empirical coverage of the forecasts, medians and 95% C.I. of the posterior samples of the turbulent eddy diffusion model parameters using toluene solvent

Parameter	True value	additive AR	NS cov	Parametric
G	1318.33	1367.2(1223.9,1506.3)	1174.8(1105.2,1373.5)	1503.7(1168.7,1645.3)
D_T	0.001-0.01	0.006(0.005,0.006)	0.008(0.007,0.008)	0.0014(0.001,0.002)
CRPS		18.4	8.9	26.4

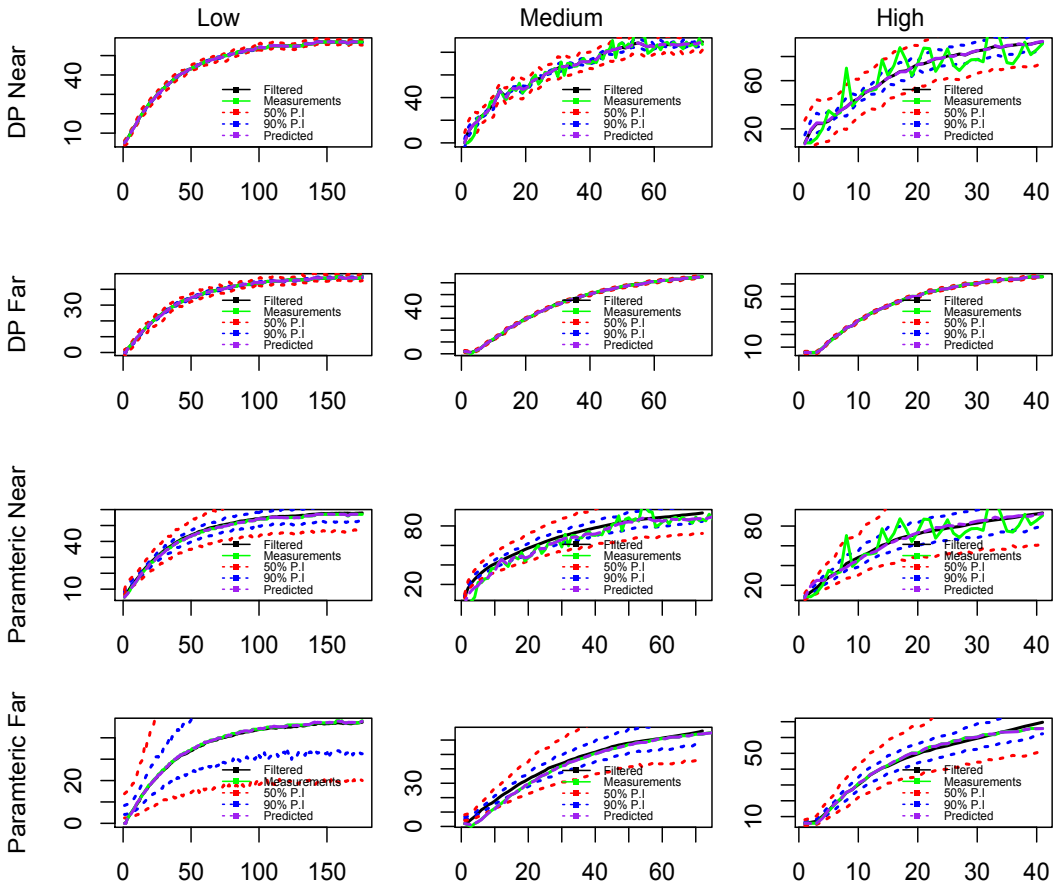

Figure 4.5: Plot of the measured concentrations and the mean of the posterior samples of the latent states conditional on the measurements in the near field and far field at low, medium and high ventilation levels

4.4 Discussion

In this chapter we developed a general Bayesian SSM based on centered Dirichlet processes for analyzing experimental exposure data specific to industrial hygiene. The unknown distributions of the error terms in the measurement and transition equations at any time point are constructed using Dirichlet process mixtures of Gaussian distributions with constraints to prevent negative concentration values. The proposed method allows for very flexible yet robust modeling of exposure data, such that any physical model in theory can be accommodated. In addition, the Bayesian framework provides a natural approach for probabilistic forecasting where we can study calibration and sharpness in predictions. One more advan-

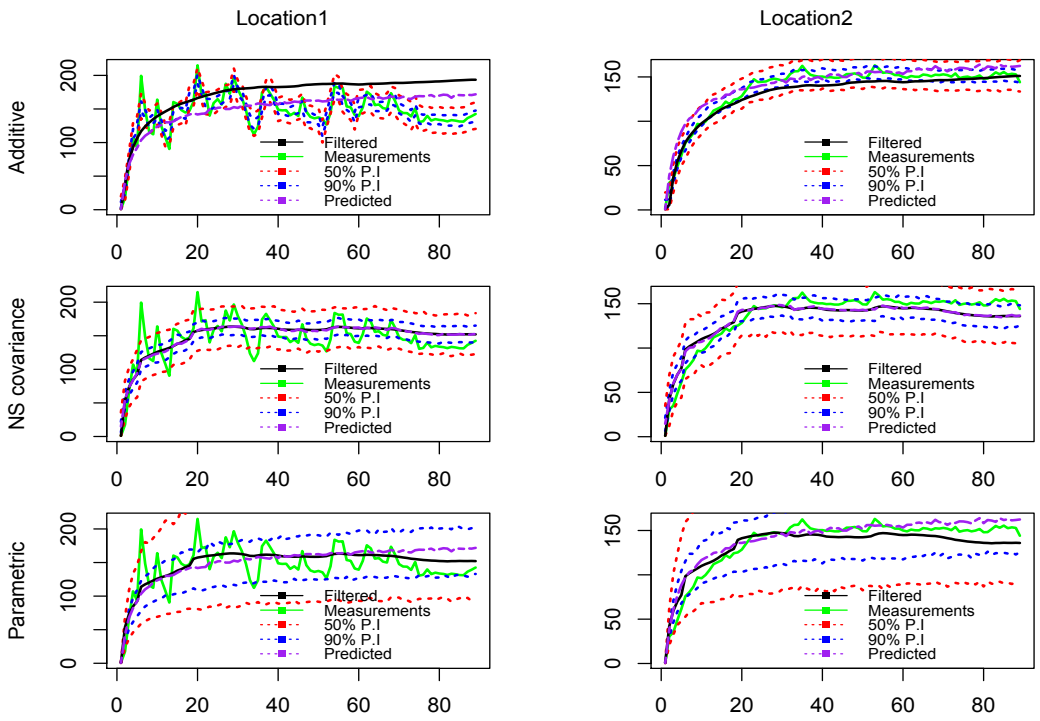

Figure 4.6: Plot of the measured concentrations and the mean of the posterior samples of the latent states conditional on the measurements at the two locations using add cov

tage of the proposed method, is the simplicity in implementing the MCMC sampler, which uses an extension to the Gibbs Sampling in [IJ01] and the Gaussian dynamic linear models (DLMs) in [WH97]. We have showed the effectiveness of our methodology with simulation experiments and lab-generated data under several experimental settings where different physical models and different experimental conditions were considered.

Results show that inference and prediction are robust to different physical models and different experimental scenarios. Nonparametric SSMs tended to perform better than parametric SSMs, a result that appeared to be consistent across different exposure models and different experimental conditions, in particular regarding calibration and sharpness. Our simulation experiments showed the promise of using spatiotemporal SSMs in analyzing eddy diffusion experiments, and even though our chamber data analysis had limited scope because of the very small number of spatial measurements, the model that used non-separable covariance was more robust and was able to characterize the data well.

Although we have focused on three classes of physical models, the proposed methodology is more general and in theory, can accommodate any physical model. We conclude with some indicators for future research. First, as suggested earlier, the use of non-separable covariance is the most auspicious which encourages exploring more general classes of non-separable covariances proposed by [Gne02]. A limitation of using SSMs in general, is that conditional on the state vector, observations are assumed to be independent, which in more complex physical processes that exhibit strong temporal dependence may be unrealistic.

CHAPTER 5

Discussion

5.1 Exposure Modeling Challenges and Applications Addressed

In this dissertation, I outlined challenges commonly arise in the filed of exposure modeling and parameter inference. The proposed frameworks are likely to be useful to industrial hygienists for exposure assessment and management. The main challenges that motivated the work arised from important research questions that cannot be addressed by the current common methodologies due to data limitations.

The GuLF STUDY coastal data at Waveland beach Mississippi motivated our first methodology. The available data consisted of breathing concentration measurements from clean-up coastal workers which extended in an S-shape for seven or eight kilometers. Available methods for point-referenced data modeling use Geostatistical covariances that rely on the Euclidean distance between the measurements in capturing the spatial correlation between them. The problem with this approach is that the Euclidean distance is inappropriate for modeling spatial covariances because the effective spatial range will be the distance *along the coast* at which the correlation becomes negligible. Second, covariance functions that insure positive definiteness in Euclidean coordinates need not be valid for other domains [Ban05]. This means that we will need to construct valid covariance functions along the coastline. Subsequently, we describe a simple approach to construct models such as (2.1) using valid covariance functions for points along curves. We developed a flexible yet simple Bayesian framework for spatially-oriented data that can be used to assess exposures of workers by interpolating levels of chemicals along a coastline. The statistical models for coastal kriging exploit a simple representation of the coast as a parametric function of the

coordinates of points along the coastline. We presented four models using two different parameterizations. We found that for a simple curve, kriging using line segment approximation performs better than spatial kriging using Euclidean distance. This could be a useful and practical approach for kriging over any simple curve. The model is relatively easy to fit since the covariance depends on parameters in $\Re^1$.

Another motivating problem that commonly arise in exposure assessment in industrial hygiene, is the complexity of the workplace and the lack of physical models that deliver a complete representation of the underlying processes generating chemical concentrations. An accurate representation will produce better concentration estimates and facilitate decision-making in exposure management. Therefore, accounting for parameter and model uncertainty is crucial and a synergy of physical and statistical models is needed to better estimate the processes in the workplace. Traditionally, one needs the solutions for the nonlinear differential equations representing the physical model and they need to be evaluated through several iterations for convergence. This precludes fitting computationally demanding but richer physical models that could well have yielded better estimation of physical parameters and concentrations.

We offered a principled Bayesian approach to efficiently and effectively synergize information from the three sources of information, (a) professional judgment from experts, and (b) direct measurements of the environment exposure in the workplace and (c) scientific physical models representing the state in the workplace in theoretically ideal conditions. Furthermore, the approach we proposed will completely obviate the need to solve the nonlinear equations governing the physical model. We achieved this by deriving a dynamic statistical model by discretizing the deterministic physical model and incorporating stochastic measurement error. This is then extended to a Bayesian framework by assigning prior distributions (using information from (a)) to the parameters and the model parameters (including variance components attributed to measurement error and model approximations) are estimated by sampling from the posterior distribution.

Our proposed framework enriches and expands upon existing methods. IH practitioners will no longer be restricted to fitting a confined selection of physical models amenable to an-

alytic solutions. Any conceivable physical model, in theory, can be accommodated. Neither will IH practitioners be restricted to Gaussian or transformed Gaussian data, an assumption that most practitioners will agree is rarely tenable, especially given the small to moderate number of measurements they have to deal with. Our Bayesian framework also allowed statistically sound model evaluation and assessment in terms of how well it fits the data. We also introduced Bayesian computation methods and algorithms to efficiently implement the proposed synergistic Bayesian modeling framework.

The adaptation of Bayesian Kalman filters to IH settings was novel. We also innovated to expand upon Bayesian Kalman filters with Gaussian noise. We explored classes of skewed error distributions and the results showed that non-Gaussian state space models tended to perform better than linear Gaussian state space models, a result that appeared to be consistent across different exposure models and different experimental conditions.

To expand upon the proposed approach for exposure concentration modeling, we considered an extension that allows for more flexibility in the model. We developed classes of nonparametric Bayesian Kalman filters and designed algorithms for their implementation based on centered Dirichlet processes for analyzing experimental exposure data specific to industrial hygiene. The unknown distributions of the error terms in the measurement and transition equations at any time point are constructed using Dirichlet process mixtures of Gaussian distributions with constraints to prevent negative concentration values. We demonstrated the performance of the proposed extension using simulations and real data, in which we compared the performance of the nonparametric method to the one previously presented that impose distributional assumptions on the error terms. Results showed that the nonparametric models tended to perform better than the parametric ones, a result that appeared to be consistent across different exposure models and different experimental conditions, in particular regarding calibration and sharpness. Our simulation experiments showed the promise of using spatiotemporal state space models in analyzing eddy diffusion experiments, and even though our chamber data analysis had limited scope because of the very small number of spatial measurements, the model that used non-separable covariance was more robust and was able to characterize the data well.

5.2 Future Work

This dissertation provides a new direction to exposure modeling in industrial hygiene. First, building valid models for coastal kriging presents many new research opportunities, such as developing a model for more complicated coastlines or along closed curves such as the coasts of an island. Thus, future work will investigate potential problems such as, complexity of the curve, covariates inclusion, potential changes in the coastline and temporal changes. Future work can also consider the modeling and analysis of censored data, which is common in exposure studies, due to measurements below the limits of detection.

The area of parameter estimation and exposure prediction using state space models is still also an open area of research. As mentioned previously, a much more comprehensive spatiotemporal analysis for eddy diffusion experiments is needed. Our simulation experiments showed the promise of spatiotemporal state space models in analyzing eddy diffusion experiments which encourages exploring broader classes of non-separable covariances that reflects different types of interaction between space and time. Another area of interest for future work is related to physical processes that exhibit strong temporal dependence. A limitation of using state space models in general, is that conditional on the state vector, observations are assumed to be independent, which may be unrealistic in some scenarios. Another important consideration is misaligned data, where not all measurements at different locations come from the same set of time points. An advantage of the Bayesian paradigm is that we can handle missing data, hence misaligned data, very easily and indeed our Bayesian state space models should be able to handle them.

A third area for future work may integrate coastal kriging into state space models in what we call *coastal state space models* that will enable us to analyze chemical concentrations from underlying physical processes along a coastline. A simple re-parametrization of the physical model can produce accurate parameter inference and prediction. A scenario for coastline concentrations assuming a single source of emission situated in the water could be a malfunctioning discharge pipe in an oil rig. If a uniform concentration of the contaminant throughout the coast is assumed, a one-zone model can be used. If different physical behavior

near the source of emission *"near field"* from that far from the source *"far field"* is assumed, then the two-zone model can be used.

5.3 Final Remarks and Conclusion

Exposure modeling and parameter estimation are important as they enable industrial hygienists to manage and assess exposure in the workplace. It is also crucial to have efficient representation of uncertainty, and account for its different sources such as errors, complex physical models and others, and hence choosing the appropriate uncertainty quantification is an important question. There are many different applications of the approaches presented in this dissertation and we anticipate more use of the presented models in industrial hygiene.

Appendices

Appendix A

Supplementary details for Chapter 3

This document provides technical details of the solution to the linear ODE, discretization of the SSMs and steady state solution.

Solution to the linear system of ODE: The simulated data was generated from the exact solution to the ODE [BR14]. In order to obtain that exact solution, we assume an m dimensional system

$$\frac{d}{dt}x_t = F_t x_t + g;, \tag{A.1}$$

where x_t is an $m \times 1$ vector at time t, F_t is an $m \times m$ matrix and g is an $m \times 1$ vector. The solution when the eigenvalues of F_t are real and distinct can be obtained as follows. The eigen decomposition of $F_t = L\Lambda L^{-1}$ where L is the matrix of linearly independent eigenvectors and Λ is the diagonal matrix of eigenvalues λ_i, $i = 1, \dots m$. By definition of the matrix exponential and the fact that $F_t^n = L\Lambda^n L^{-1}$ we can write $e^{F_t} = Le^{\Lambda}L^{-1}$. Now, let $G_i = u_i v_i^T$, where u_i is the i-th column of L and v_i^T is the i-th row of L^{-1}. It easily follows that $e^{F_t} = \sum_{i=1}^{m} e^{\lambda_i} G_i$. Consequently,

$$\frac{d}{dt}e^{tF_t} = \sum_{i=1}^{m} \lambda_i e^{\lambda_i t} G_i = \sum_{i=1}^{m} \lambda_i e^{\lambda_i t} u_i v_i^T = L\Lambda e^{t\Lambda} L^{-1} = L\Lambda L^{-1} L e^{t\Lambda} L^{-1} = F_t e^{tF_t}$$

$$\text{and} \quad \int e^{tF_t} dt = \sum_{i=1}^{m} \frac{1}{\lambda_i} e^{\lambda_i t} G_i = L\Lambda^{-1} e^{t\Lambda} L^{-1} = L\Lambda^{-1} L^{-1} L e^{t\Lambda} L^{-1} = F_t^{-1} e^{tF_t}.$$

Multiplying both sides of (A.1) by e^{-tF_t} from the left yields:

$$e^{-tF_t}\left[\frac{d}{dt}x_t - F_t x_t\right] = e^{-tF_t} g \Longrightarrow \frac{d}{dt}\left[e^{-tF_t} x_t\right] = e^{-tF_t} g\,. \tag{A.2}$$

Integrating out both sides of (A.2), we obtain $e^{-tF_t}x_t = -F_t^{-1}e^{-tF_t}g + k$, where k is a constant vector. The initial condition at $t = 0$ yields $x_0 = -F_t^{-1}g + k$, so $k = x_0 + F_t^{-1}g$. Consequently, $x_t = e^{tF_t}x_0 + F_t^{-1}\left[e^{tF_t} - I_m\right]g$ which is the solution to (A.1).

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

$$C(t^* + \delta_t) = \sum_{n=0}^{\infty} \frac{C^{(n)}(t^*)}{n!}(\delta_t)^n = C(t^*) + \frac{C'(t^*)}{1!}\delta_t + o(\delta_t), \tag{A.3}$$

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

$$C'(t^*) = \frac{C(t^* + \delta_t) - C(t^*)}{\delta_t} + o(\delta_t). \tag{A.4}$$

In the applications to the three physical models we replace the first order derivative $\frac{d}{dt}C_t$ at $t = t^*$ with equation (A.4) 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.

$$\lim_{t\to\infty} \exp\{tF_t\}C(t_0) + F_t^{-1}[\exp\{tF_t\} - I]g. \tag{A.5}$$

For the one zone model $F_t = -(Q + K_L V)/V$ and $g = G/V$ so $A.5 = F_t^{-1}[-I]g = G/(Q + K_L V)$. Since K_L is usually small, it can be approximated by G/Q. Hence as $t \to \infty$ $C_t \approx G/Q$.

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 $g = \begin{bmatrix} G/V_N \\ 0 \end{bmatrix}$. Since K_L is usually small it can be ignored for simplicity. The term $\exp(tF_t)$, where

exp() is the matrix exponential, can be written as $\exp(tL\Lambda L^{-1}) = \sum e^{t\lambda} G_i$ where $G_i = u_i v_i^T$, u_i is the i-th column of L and v_i^T is the i-th 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 [ZBL09] as

$$\begin{aligned} \lambda_1 &= \tfrac{1}{2}\left[-\left(\tfrac{\beta V_F + (\beta+Q)V_N}{V_N V_F}\right) + \sqrt{\left(\tfrac{\beta V_F + (\beta+Q)V_N}{V_N V_F}\right)^2 - 4\left(\tfrac{\beta Q}{V_N V_F}\right)} \right], \\ \lambda_2 &= \tfrac{1}{2}\left[-\left(\tfrac{\beta V_F + (\beta+Q)V_N}{V_N V_F}\right) - \sqrt{\left(\tfrac{\beta V_F + (\beta+Q)V_N}{V_N V_F}\right)^2 - 4\left(\tfrac{\beta Q}{V_N V_F}\right)} \right]. \end{aligned} \tag{A.6}$$

As long as β and Q are positive, 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 $A^{-1}[-I]g$. The determinant of A is $\det(A) = Q\beta/V_N V_F$, and $A^{-1} = \begin{bmatrix} -((\beta+Q)/V_F)(V_N V_F/\beta Q) & -(\beta/V_N)(V_N V_F/\beta Q) \\ -(\beta/V_F)(V_N V_F/\beta Q) & -((\beta)/V_N)(V_N V_F/\beta Q) \end{bmatrix}$. So the steady state is a 2×1 vector equal to $A^{-1}[-I]g = \begin{bmatrix} \frac{G}{Q} + \frac{G}{\beta} \\ \frac{G}{Q} \end{bmatrix}$. So as $t \to \infty$ $C_N(t) \approx \frac{G}{Q} + \frac{G}{\beta}$ and $C_F(t) \approx \frac{G}{Q}$.

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

Appendix B

Further results for Chapter 4

B.1 Simulations

We studied the performance of the proposed framework through simulations at different physical model scanarios and different model assumptions. We generated the latent concentrations from the exact solution to the differential equation corresponding to the physical model in use,

$$C_t = \exp(tF_t)C_0 + F_t^{-1} \times [\exp(tF_t) - I]g, \tag{B.1}$$

where F_t is a state transition matrix, g are control inputs and C_0 is the initial concentration value. The observed concentrations are generated from $Y_t = H_t C_t + \nu_t$, at different signal-to-noise ratios. Then we evaluated the performance of the proposed non-parametric SSM framework and compared the results to the parametric SSM proposed in Chapter 3.

B.1.1 One-zone model

Setup: We generated 100 datasets at different signal to noise ratios of 100 exposure concentrations each, at equally spaced time points. The latent concentrations were generated from

$$C_t = \exp\{-t(Q + K_L V)/V\}C(t_0) + ((Q + K_L V)/V)^{-1} \left[1 - \exp\{-t(Q + K_L V)/V\}\right] G/V, \tag{B.2}$$

where, $Q = 13.8$ m^3/min, $G = 351.5$ mg/min, $V = 3.8$ m^3, $K_L = 0.1$ mg/min and $C(0) = 1$ mg/m^3 [ZBL09]. The observed measurements were generated from $Y_t = C_t + \nu_t$, $\nu_t \overset{iid}{\sim} N(0, \sigma_{\nu_t})$, where σ_{ν_t} varied between 1 and 5.

Results: Table B.1 presents the medians and 95% credible intervals of the MCMC posterior samples of the model parameters averaged over all simulations. The 95% credible intervals include the true values for all the parameters values. The CRPS values indicate better calibration and sharpness among the nonparametric model. Figure B.1 shows the simulated concentrations, measurements and the mean of the posterior samples of the latent states conditional on the measurements for one simulated dataset, in addition to the 50% and 90% prediction intervals which shows that the latent state estimates are close to the true simulated values. Figure B.3 and Figure B.2 show that both the sample PIT histogram and the marginal calibration plot do not indicate a deficiency in the forecasts.

Table B.1: CRPS, empirical coverage of the forecasts, medians and 95% C.I. of the posterior samples of the one-zone model parameters for the simulated data (averaging over 100 simulations at different signal to noise ratios)

Parameter	DP	Parametric
$G(351.5)$	357.1 (322.5, 419.3)	326.8 (283.3, 351.7)
$Q(13.8)$	12.9(11.1, 15.5)	12.9(11.1, 14.8)
$K_L(0.1)$	0.29(0.02,0.70)	0.34(0.19,0.78)
CRPS from 100 simulations	1.4	2.1
MSE from 100 simulations	0.7(1.8)	0.6 (0.7)

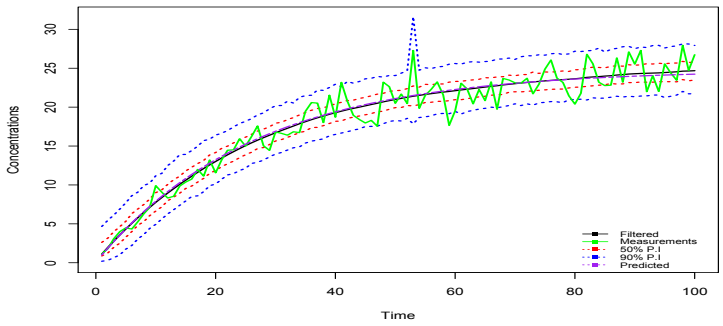

Figure B.1: Plot of the simulated concentrations, measurements and the mean of the posterior samples of the latent states conditional on the measurements

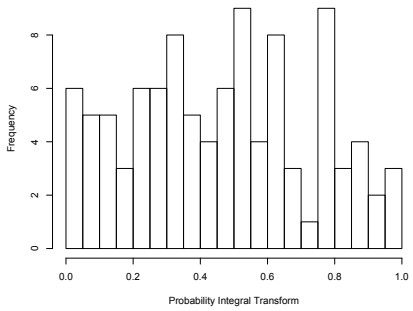

Figure B.2: PIT histograms and marginal calibration plot for one-zone model simulations

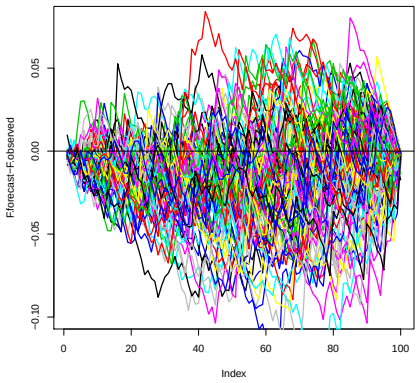

Figure B.3: Marginal calibration plot for one-zone model 100 simulations

B.1.2 Two-zone model

Setup: We generated 100 datasets at different signal to noise ratios of 100 exposure concentrations each, at equally spaced time points. The latent concentrations were generated from

$$C_t = \exp\{tA\}C(t_0) + A^{-1}\left[\exp\{tA\} - I\right]g, \tag{B.3}$$

where, $A(\theta_c; x) = \begin{bmatrix} -\beta/V_N & \beta/V_N \\ \beta/V_F & -(\beta+Q)/V_F + K_L \end{bmatrix}$, $g(\theta_c; x) = \begin{bmatrix} G/V_N \\ 0 \end{bmatrix}$, $Q = 13.8\ \text{m}^3/\text{min}$, $G = 351.5\ \text{mg/min}$, $V = 3.8\ \text{m}^3$, $K_L = 0.1$, $\beta = 5\ \text{m}^3/\text{min}$, $V_N = \pi \times 10^{-3}\ \text{m}^3$ and $V_F = 3.8\ \text{m}^3\ \text{mg/min}$ and $C_N(0)$ and $C_F(0)$ were assigned values 0 and $0.5\ \text{mg/m}^3$ respectively [ZBL09]. The observed measurements were generated from $Y(t) = C_t + \nu_t$, where $Y(t) = \begin{bmatrix} Y_N(t) \\ Y_F(t) \end{bmatrix}$, $C_t = \begin{bmatrix} C_N(t) \\ C_F(t) \end{bmatrix}$ $\nu_t \overset{iid}{\sim} N_2(0, \Sigma_{\nu_t})$.

Results: Table B.2 presents the medians and 95% credible intervals of the MCMC posterior samples of the model parameters averaged over all simulated datasets. The 95% credible intervals include the true values for all the parameters values. The CRPS values indicate better calibration and sharpness among the nonparametric model. Figure B.4 shows the simulated concentrations, measurements, the mean of the posterior samples of the latent

states conditional on the measurements and the 50% and 90% prediction intervals at the near and far fields for one of the generated datasets. Figure B.21 and Figure B.6 show that both the PIT histogram and the marginal calibration plot do not indicate a deficiency in the forecasts.

Table B.2: CRPS, empirical coverage of the forecasts, medians and 95% C.I. of the posterior samples of the two-zone model parameters for the simulated data (averaging over 100 simulations at different signal to noise ratios)

Parameter	DP	Parametric
$G(351.5)$	297.8(281.5,363.1)	307.3(283.1,345.5)
$Q(13.8)$	12.8(11.1,16.2)	13.6(11.4,16.1)
$K_L(0.1)$	0.30(0.02,0.48)	0.38(0.02,0.78)
$\beta(5)$	4.2(3.9,5.2)	4.3(3.9,4.9)
CRPS	6.4	8.6
MSE	27.0(18.0)	20.3(14.5)

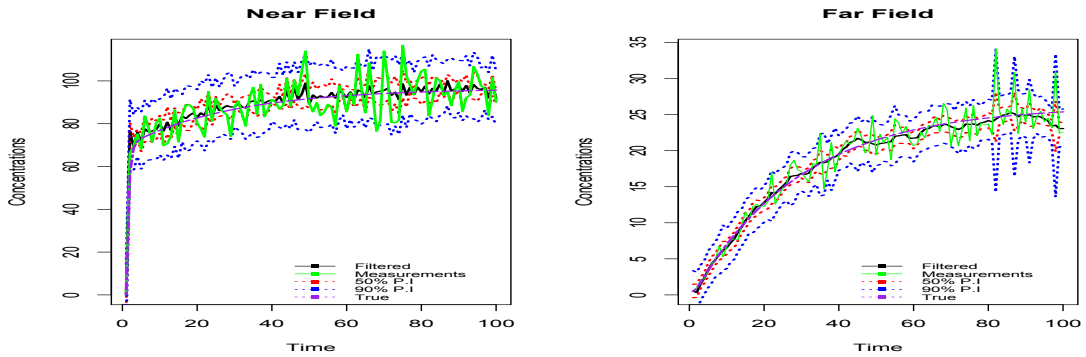

Figure B.4: Plot of the simulated near field and far field concentrations, measurements and the mean of the posterior samples of the latent states conditional on the measurements, and 50% and 90% quantiles

B.1.2.1 Turbulent eddy diffusion model

Setup: We conducted 50 simulations at different signal to noise ratios of of 500 exposure concentrations each, at 5 different locations over equally spaced 100 time points. The latent

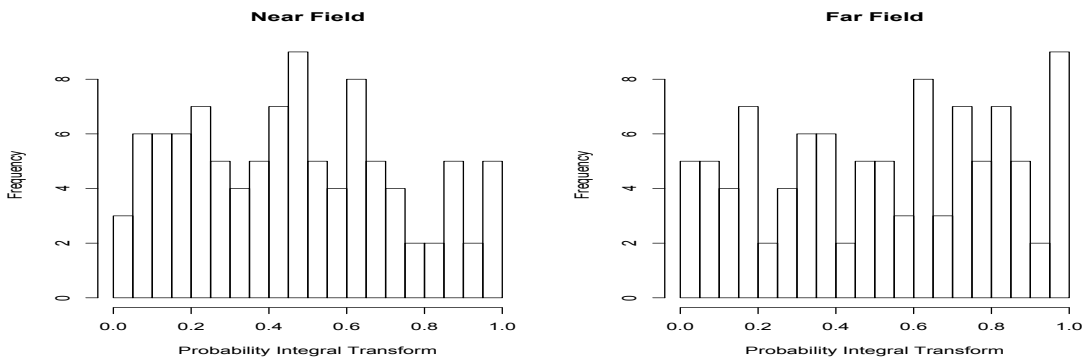

Figure B.5: PIT histograms for two-zone model simulations

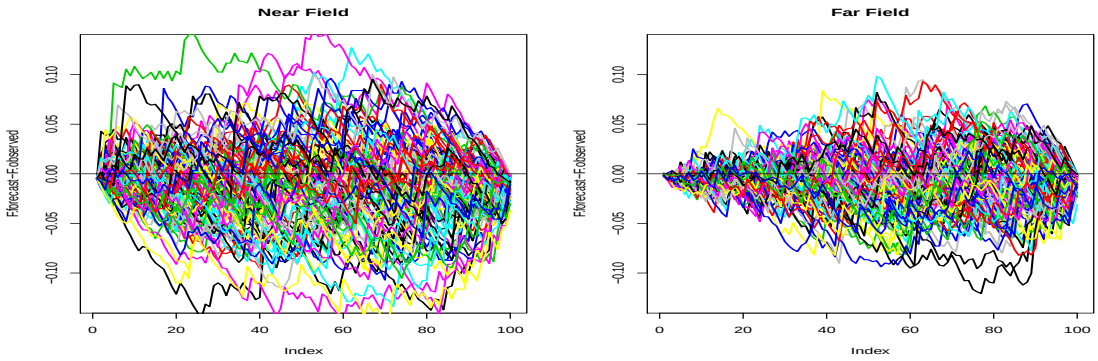

Figure B.6: Marginal calibration plot for two-zone model simulations

concentrations were generated from using the exact equation

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

where $\text{erf}(z) = \frac{2}{\pi}\int_0^z \exp(-u^2)du$, such that $G = 351.5$ mg/min and $D_t = 1$ m^2/min and the observations were generated from $Y_{t,s} = C_{t,s} + \nu_{t,s} + \eta_t$ at three different scenarios, where different covariance structures are used in the simulation. The first scenario assumes that $\eta_t = 0$ and $\nu_{t,s}$ is a random noise. The second scenario assumes an autoregressive process such that, for each location s, $\nu_{t,s} = \nu_{t-1}(s) + \psi_t(s)$ with $\psi_t(s)$ being a Gaussian process with a simple geostatistical covariance $K_{\theta_\psi} = \sigma^2 e^{-\phi\|s-s'\|}$, where $\sigma = 0.1$ and $1/\phi = 1$. The third scenario representing more flexible modeling, where a NS covariance was used; $K_{\theta_\nu} = \frac{\sigma^2}{(a^2|t-t'|^2+1)^{d/2}}\exp\left\{-\frac{b^2\|s-s'\|^2}{a^2|t-t'|+1}\right\}$, where $a = b = 1$ and $\sigma = 0.1$.

Results: Two non-parametric SSMs were considered, the first assumes an additive AR

model and the second assumes a NS covariance model. Table B.3 presents the results of the three simulation scenarios and the three models. The 95% credible intervals include the true values for all the parameters values. The CRPS values indicate better calibration and sharpness among the nonparametric models with a slightly better calibration results among the NS covariance model. Figure B.15 shows the simulated concentrations, measurements, the mean of the posterior samples of the latent states conditional on the measurements and the 50% and 90% prediction intervals at the five locations for one of the generated datasets at the random, additive AR and NS covariance simulations for the NS covariance model. The PIT histogram and the marginal calibration plots show much better calibration and sharpness among the non-parametric models which was also reflected in the values of the CRPS.

Table B.3: CRPS, empirical coverage of the forecasts, medians and 95% C.I of the posterior samples of the turbulent eddy diffusion model parameters for three simulation scenarios

	Parameter	additive AR	NS1	Parametric
random	$G(351.5)$	376.9(285.0,476.1)	401.0(296.0, 474.8)	368.7(284.4,476.2)
	$D_T(1)$	1.2(0.9,1.5)	1.3(1.0,1.7)	1.3(1.0,1.6)
	CRPS	0.80	0.67	2.5
	MSE	387.9(26.9)	367.2(2.0)	361.7(8.4)
additive AR	$G(351.5)$	432.9(319.5,479.3)	407.0(288.0, 479.0)	370.9(285.0,475.6)
	$D_T(1)$	1.3(1.0,1.6)	1.5(1.0,2.9)	1.2(1.0,1.6)
	CRPS	0.8	0.7	2.2
	MSE	385.3(35.0)	367.4(2.7)	372.9(8.4)
NS	$G(351.5)$	405.2(289.0,477.2)	398.4(308.5,476.2)	365.5(284.8,474.0)
	$D_T(1)$	1.3(1.0,1.6)	1.4(1.0,2.6)	1.3(1.0,1.6)
	CRPS	0.64	0.70	2.1
	MSE	403.1(42.9)	368.7(7.0)	367.5(4.7)

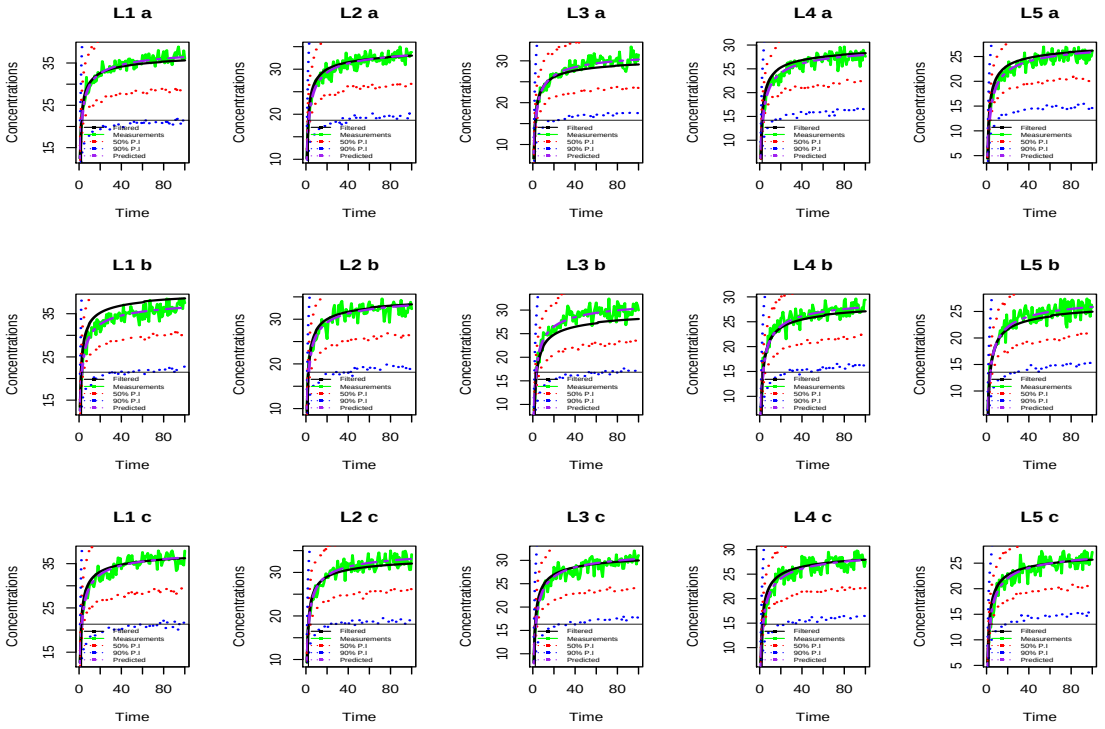

Figure B.7: Parametric model plot of the simulated concentrations at five locations, measurements and the mean of the posterior samples of the latent states conditional on the measurements, and 50% and 90% quantiles using a: simulations from random error, b: simulations from additive AR error and c: simulations from error with NS covariance

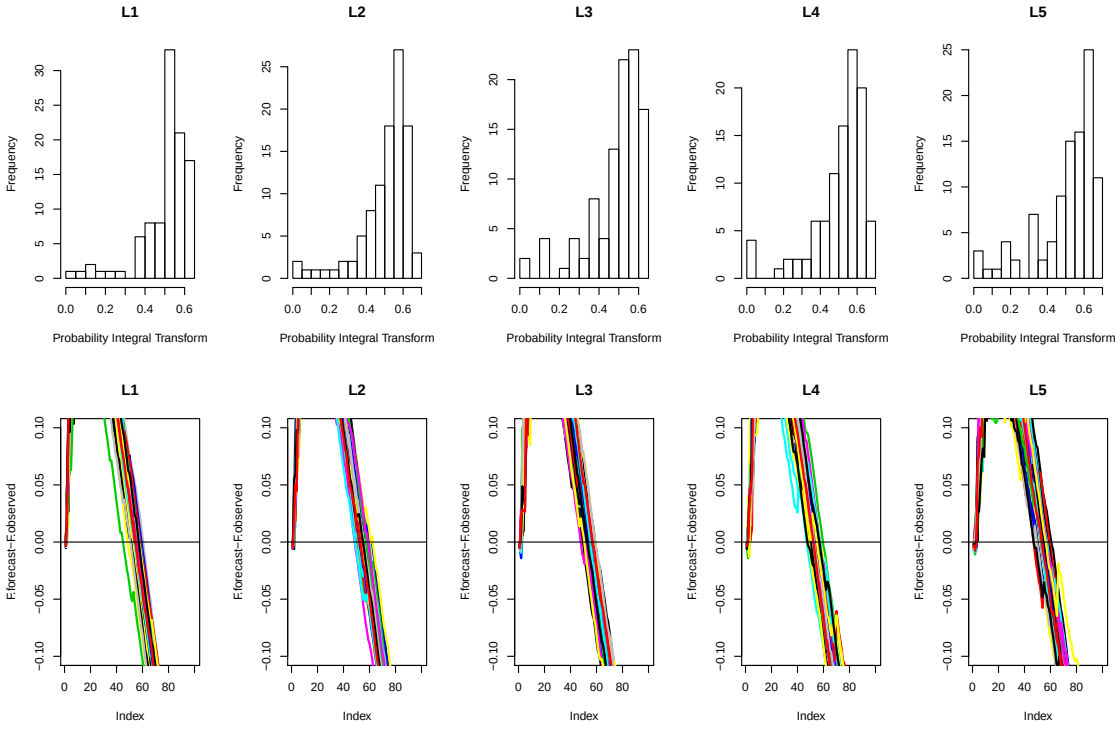

Figure B.8: PIT histograms and marginal calibration plot for eddy diffusion model random error simulation using the parametric model

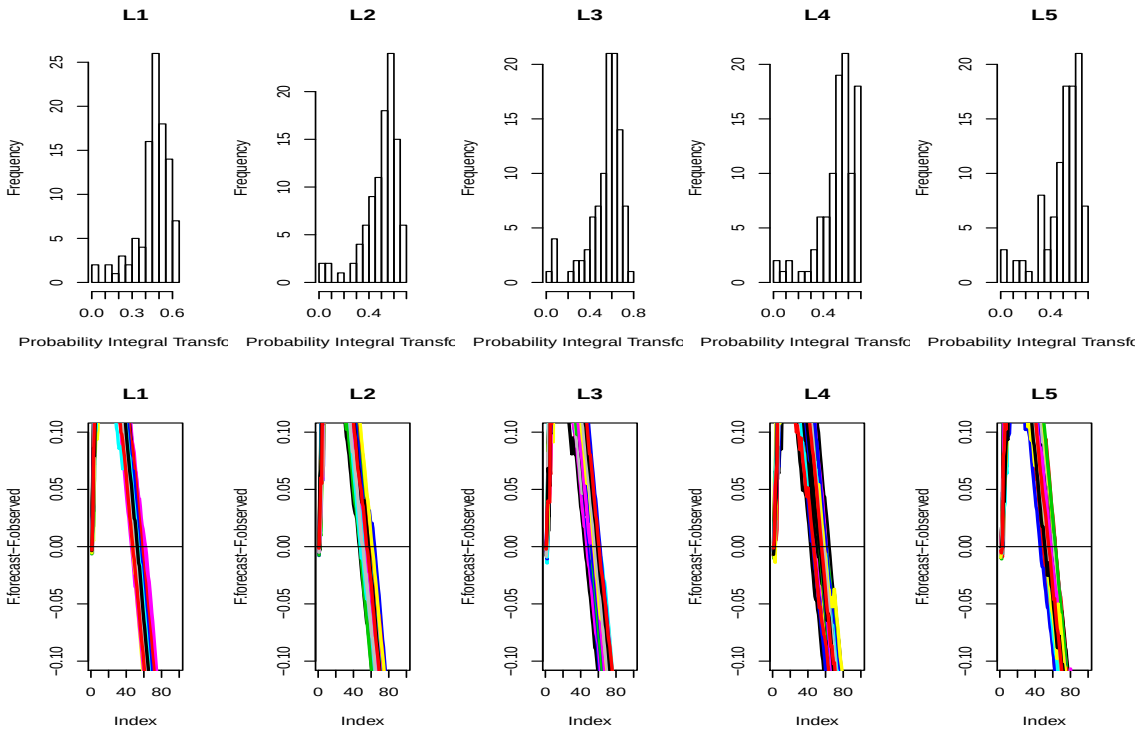

Figure B.9: PIT histograms and marginal calibration plot for eddy diffusion model additive AR additive simulation using the parametric model

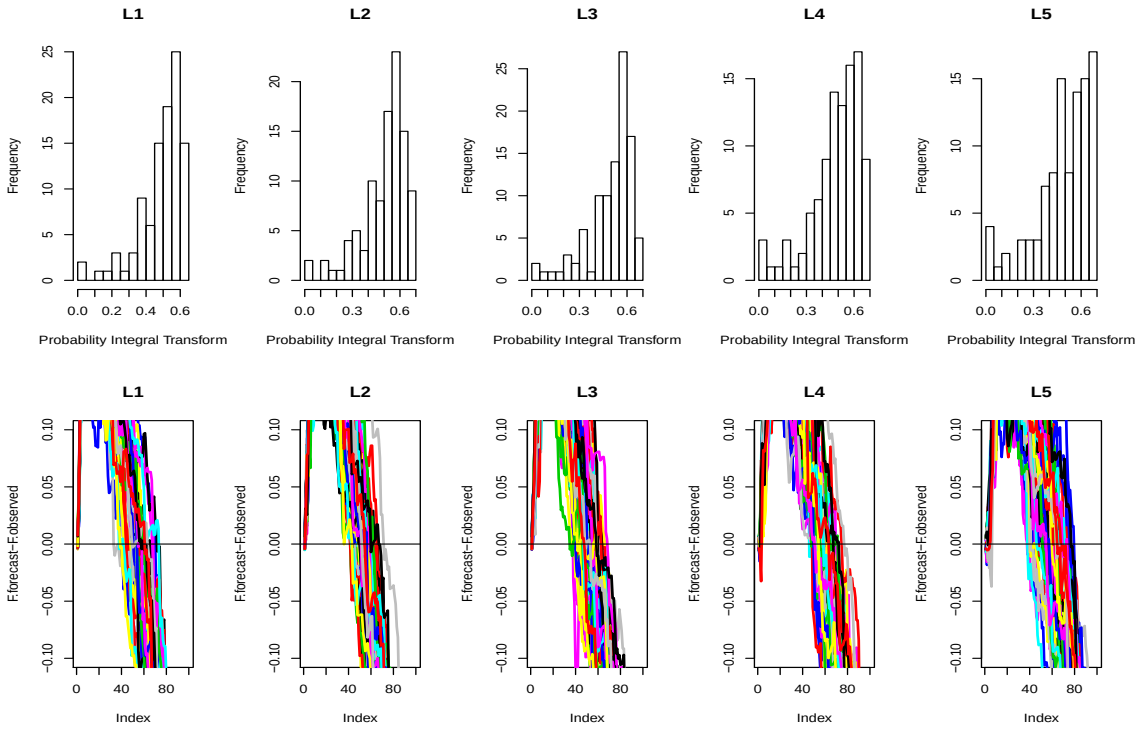

Figure B.10: PIT histograms and marginal calibration plot for eddy diffusion model additive NS covariance simulation using the parametric model

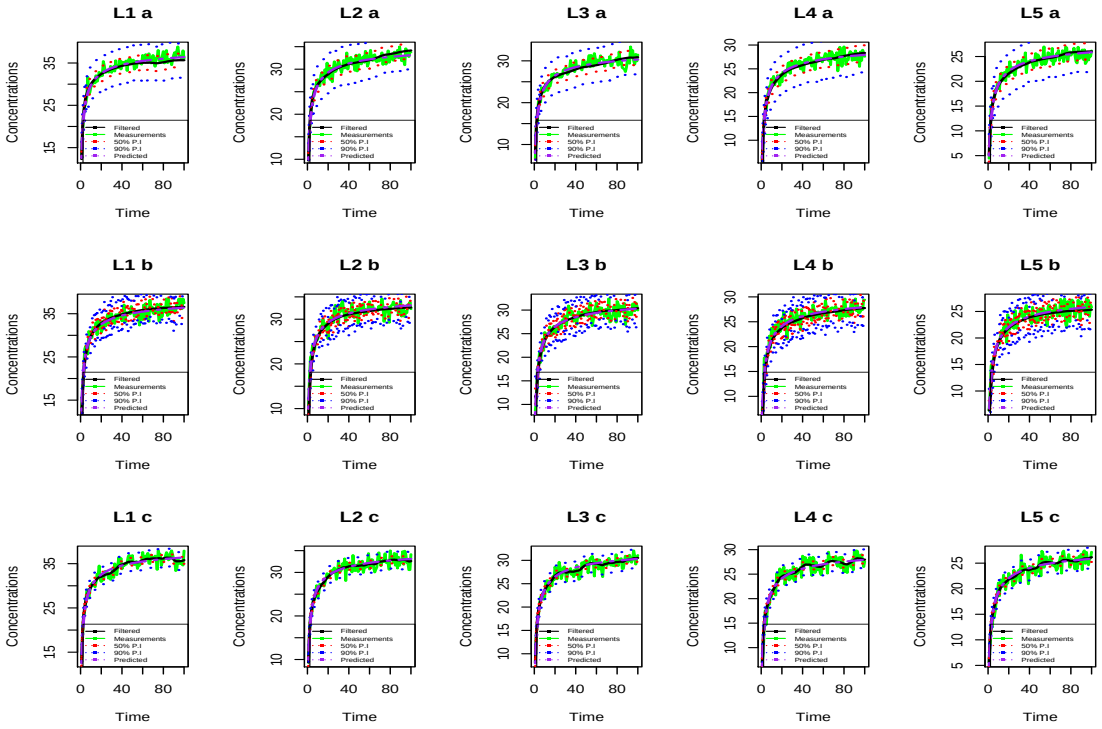

Figure B.11: Additive AR model plot of the simulated concentrations at five locations, measurements and the mean of the posterior samples of the latent states conditional on the measurements, and 50% and 90% quantiles using a: simulations from random error, b: simulations from additive AR error and c: simulations from error with NS covariance

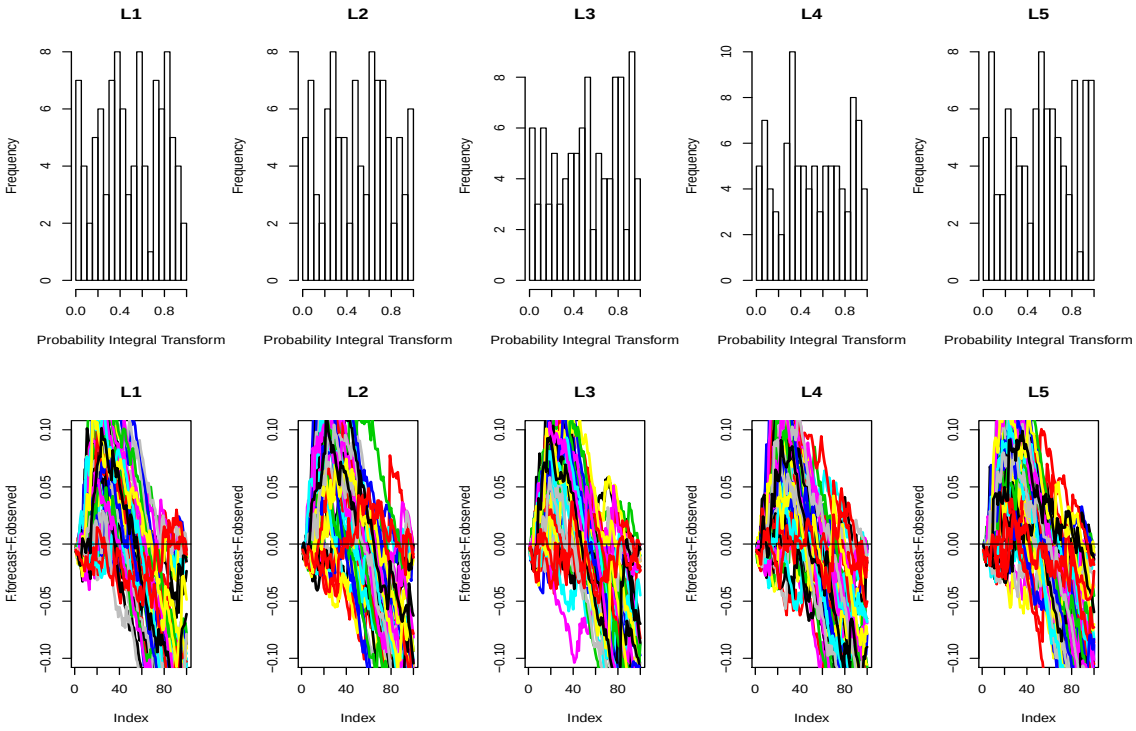

Figure B.12: PIT histograms and marginal calibration plot for eddy diffusion model random error simulation using the additive AR model

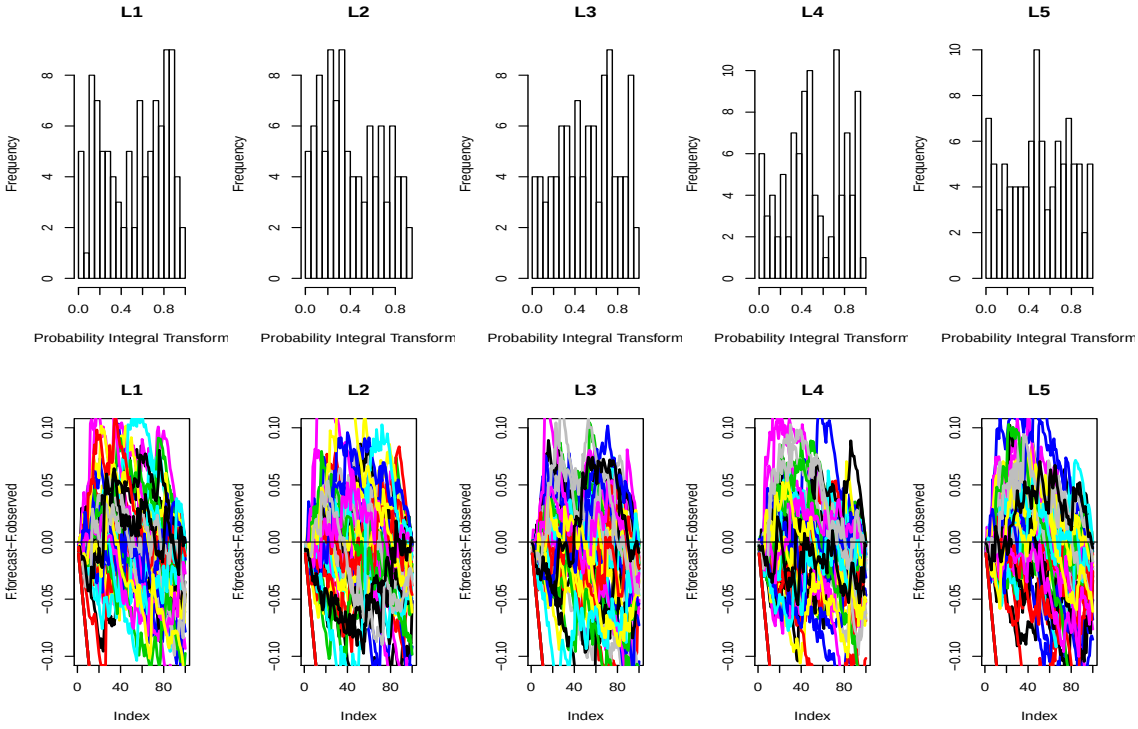

Figure B.13: PIT histograms and marginal calibration plot for eddy diffusion model additive AR additive simulation using the additive AR model

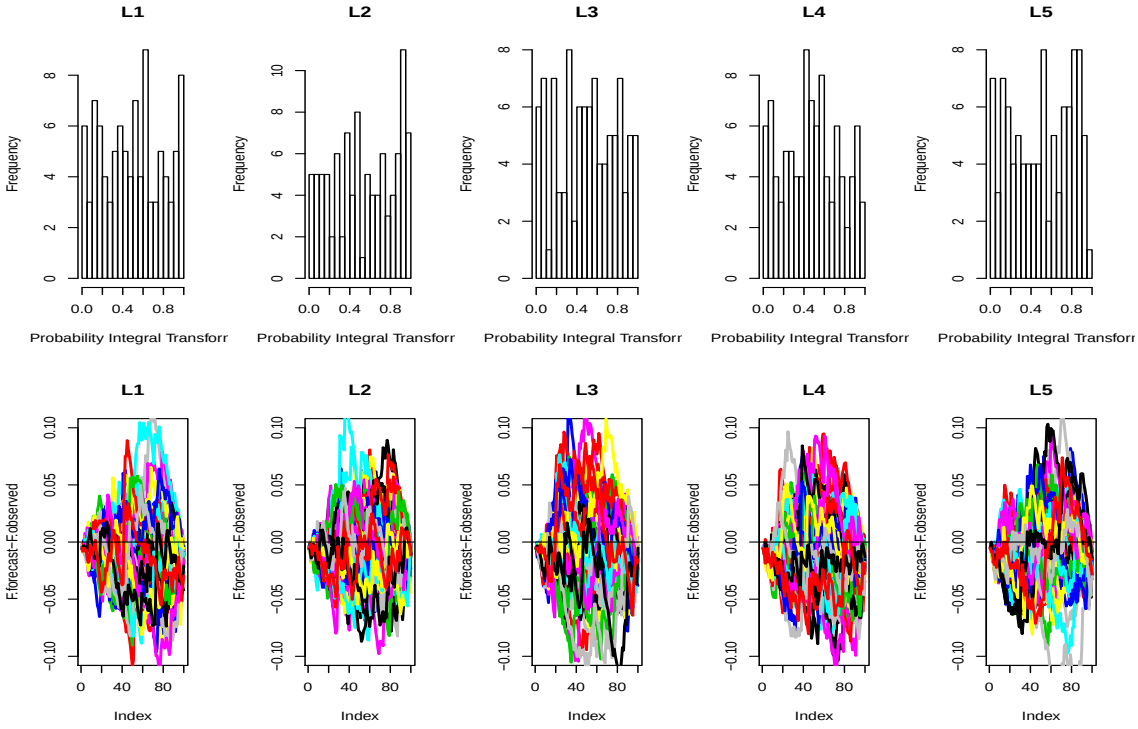

Figure B.14: PIT histograms and marginal calibration plot for eddy diffusion model additive NS covariance simulation using the additive AR model

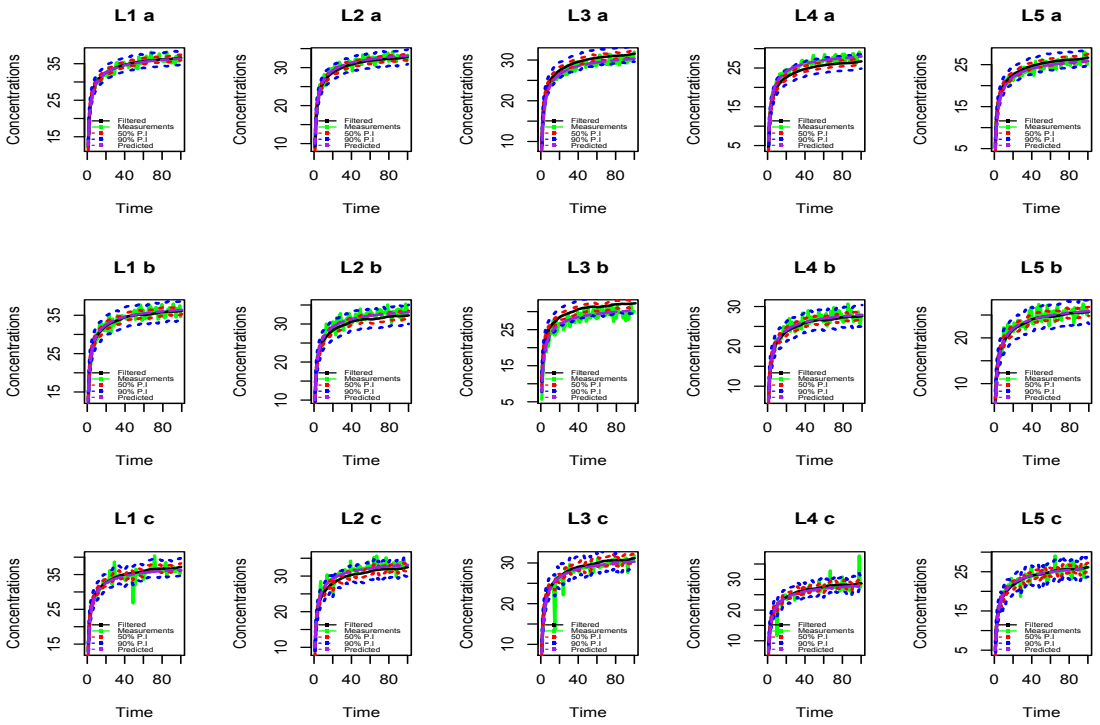

Figure B.15: NS covariance model plot of the simulated concentrations at five locations, measurements and the mean of the posterior samples of the latent states conditional on the measurements, and 50% and 90% quantiles using a: simulations from random error, b: simulations from additive AR error and c: simulations from error with NS covariance

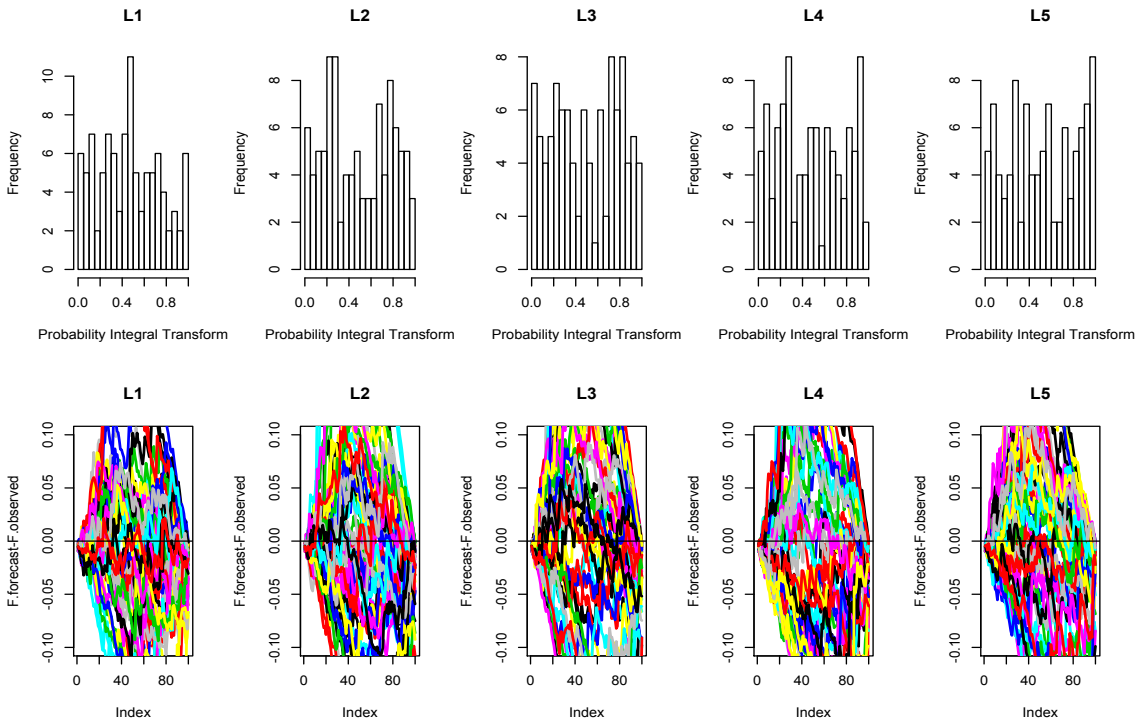

Figure B.16: PIT histograms and marginal calibration plot for eddy diffusion model random error simulation using NS covariance model

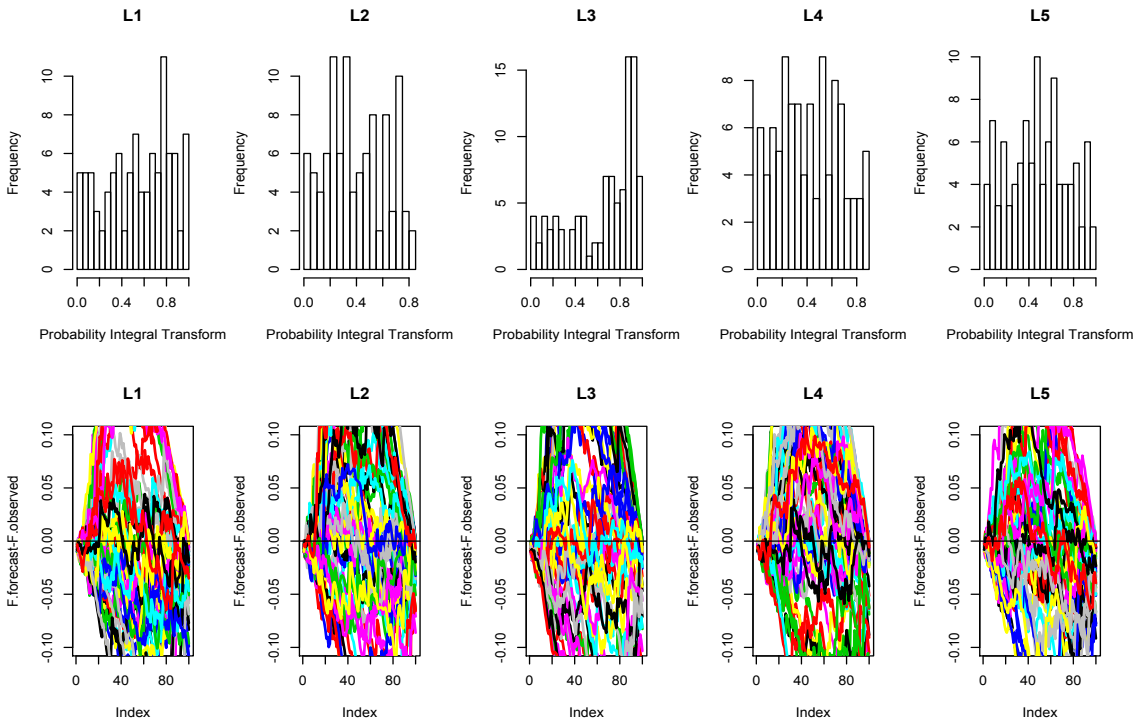

Figure B.17: PIT histograms and marginal calibration plot for eddy diffusion model additive AR additive simulation using NS covariance model

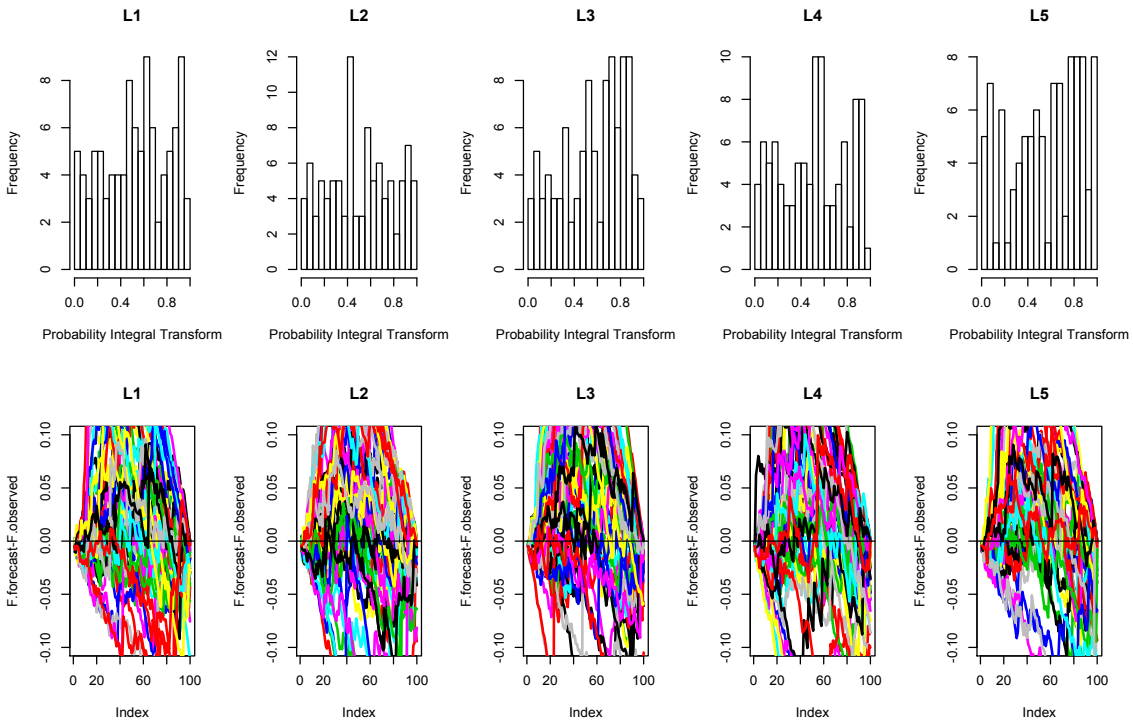

Figure B.18: PIT histograms and marginal calibration plot for eddy diffusion model additive NS covariance simulation using NS covariance model

B.2 Data Analysis

B.2.1 One-zone model

The CRPS values show better calibration among the DP model throughout the three different ventilation level scenarios which is also reflected in Figure B.19 and Figure B.20. Results show better model calibration among the low and medium ventilation level scenarios. One reason might be the much lower number of observations in that particular data set.

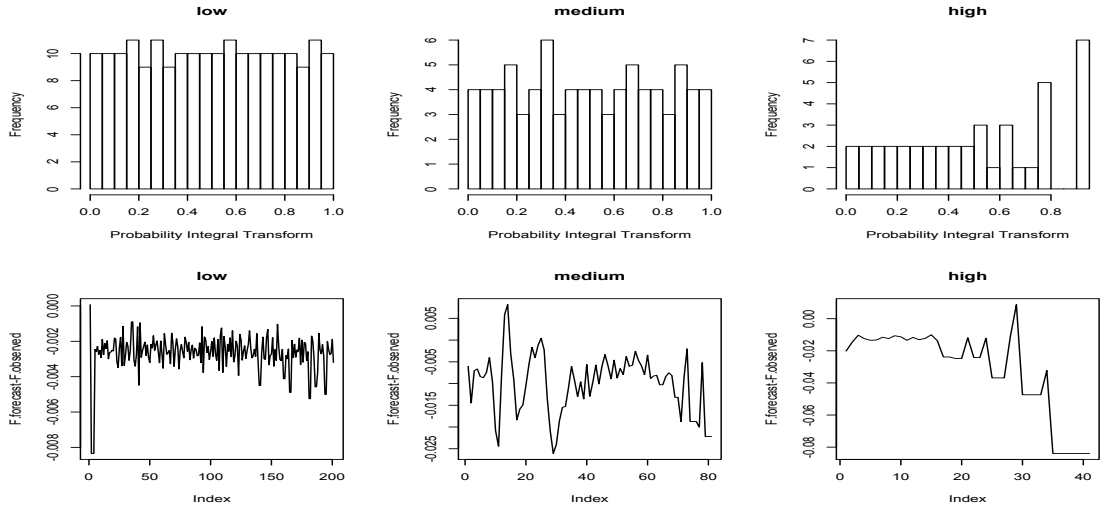

Figure B.19: PIT histograms and marginal calibration plot for one-zone at low, medium and high ventilation levels

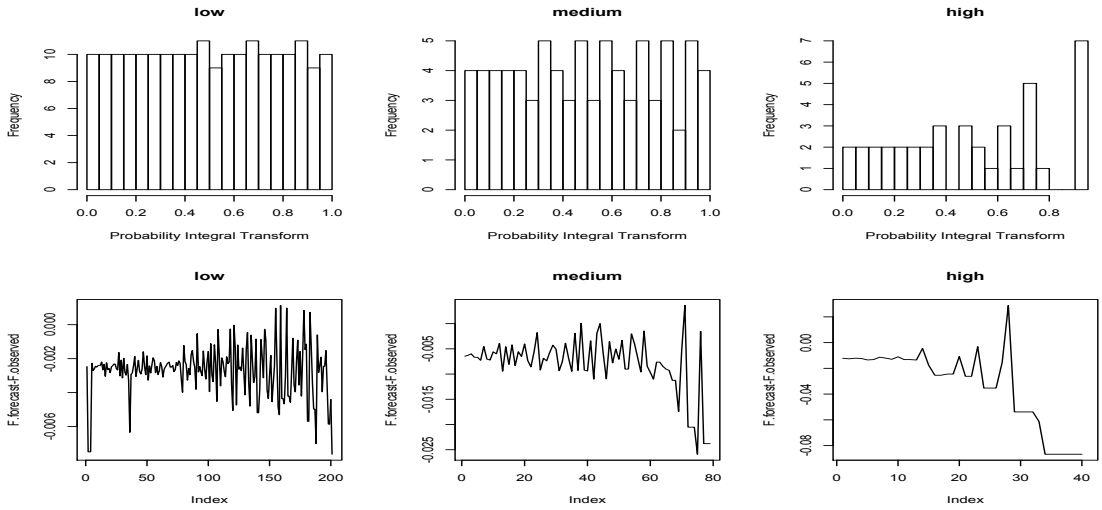

Figure B.20: PIT histograms and marginal calibration plot for one-zone at low, medium and high ventilation levels using parametric model

B.2.2 Two-zone model

The calibration plots do not show better calibration among the DP models as shown in Figure B.21 and Figure B.22. Results show better calibration at the far field throughout.

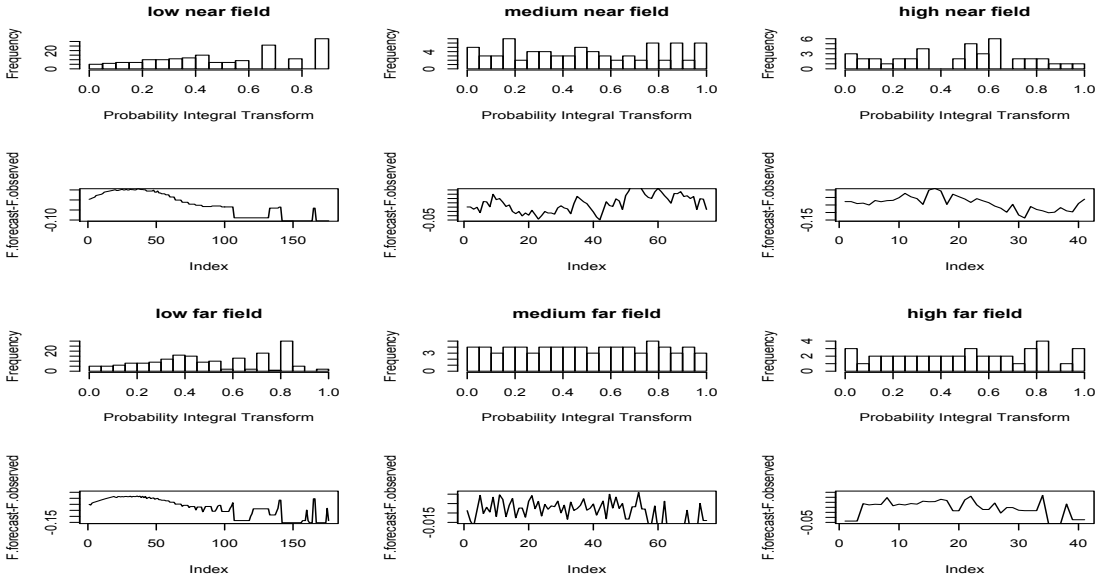

Figure B.21: PIT histograms and marginal calibration plot for two-zone at low, medium and high ventilation levels

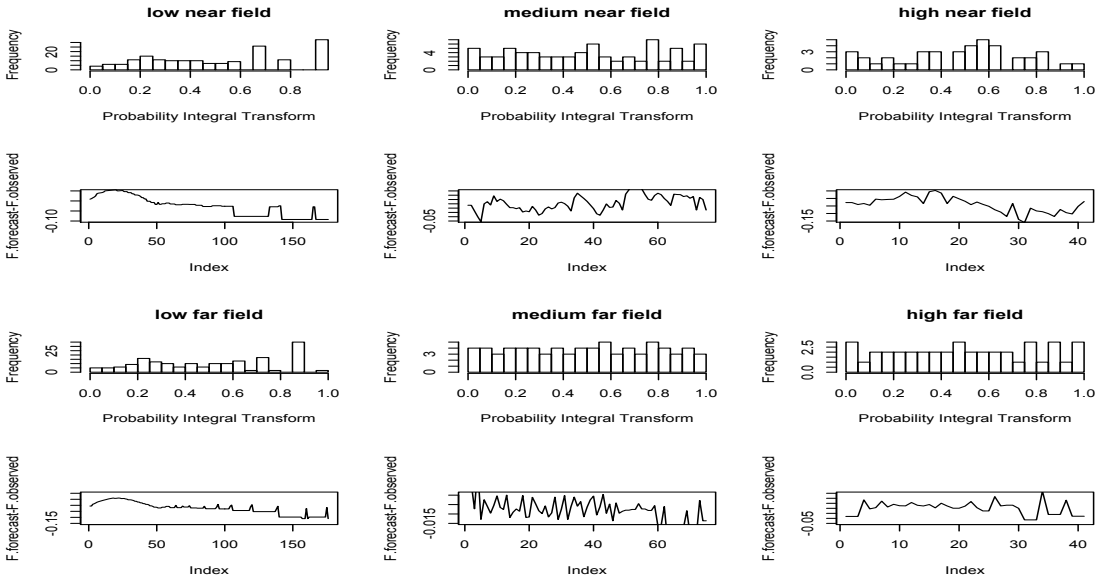

Figure B.22: PIT histograms and marginal calibration plot for two-zone at low, medium and high ventilation levels using parametric model

B.2.3 Eddy diffusion model

Despite the high noise at location 1, the NS covariance model was able to provide more smooth, accurate state estimates and better calibration compared to the additive AR model and to the parametric model which is reflected in Figure B.24 and Figure B.23.

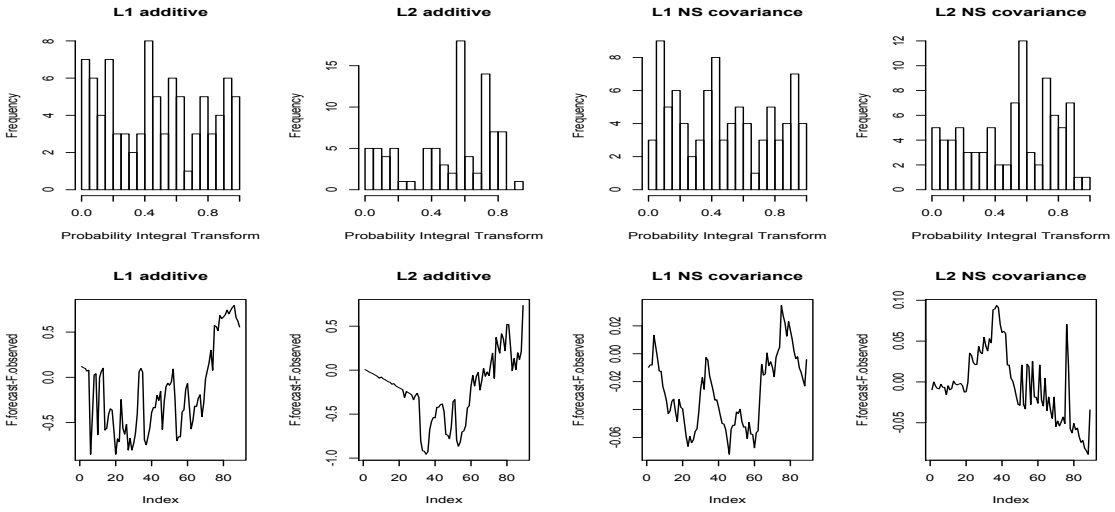

Figure B.23: Marginal calibration plot for eddy diffusion data

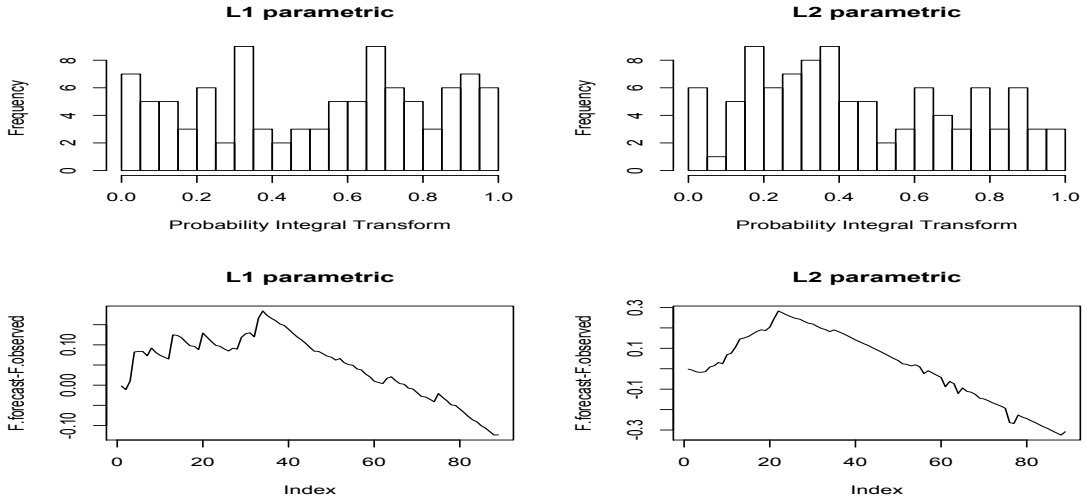

Figure B.24: Marginal calibration plot for eddy diffusion data using parametric model

REFERENCES

[ABR18] Nada Abdalla, Sudipto Banerjee, Gurumurthy Ramachandran, and Susan Arnold. "Bayesian State Space Modeling of Physical Processes in Industrial Hygiene." *ArXiv e-prints*, 2018. `http://adsabs.harvard.edu/abs/2018arXiv180702228A`.

[ADH10] Christophe Andrieu, Arnaud Doucet, and Roman Holenstein. "Particle Markov Chain Monte Carlo Methods." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **72**(3):269–342, 2010.

[ASR17] Susan Arnold, Yuan Shao, and Gurumurthy Ramachandran. "Evaluating Well-Mixed Room and Near-Field-Far-Field Model Performance under Highly Controlled Conditions." *Journal of Occupational and Environmental Hygiene*, **14**(6):427–437, 2017.

[Ban05] Sudipto Banerjee. "On Geodetic Distance Computations in Spatial Modeling." *Biometrics*, **61**:617–625, 2005.

[BC14] Luis E Nieto Barajas and Alberto Contreras Cristan. "A Bayesian Nonparametric Approach for Time Series Clustering." *Bayesian Analysis*, **9**(1):147–170, 2014.

[BCG14] Sudipto Banerjee, Bradley P. Carlin, and Alan E. Gelfand. *Hierarchical Modeling and Analysis for Spatial Data*. Chapman and Hall/CRC, 2014.

[BDM09] Mark Briers, Arnaud Doucet, and Simon Maskell. "Smoothing Algorithms for State-Space Models." *Annals of the Institute of Statistical Mathematics*, **62**:61–89, 2009.

[BR14] Sudipto Banerjee and Anindya Roy. *Linear Algebra and Matrix Analysis for Statistics*. Chapman and Hall/CRC, 2014.

[BRV14] Sudipto Banerjee, Gurumurthy Ramachandran, Monika Vadali, and Jennifer Sahmel. "Bayesian Hierarchical Framework for Occupational Hygiene Decision Making." *The Annals of Occupational Hygiene*, **58**(9):1079–1093, 2014.

[BT94] Edward J. Bedrick and Chih-Ling Tsai. "Model Selection for Multivariate Regression in Small Samples." *Biometrics*, **50**(1):226–231, 1994.

[CDD07] Franois Caron, Manuel Davy, Arnaud Doucet, Emmanuel Duflos, and Philippe Vanheeghe. "Bayesian Inference for Linear Dynamic Models with Dirichlet Process Mixtures." *IEEE Transactions on Signal Processing*, **56**(1):71–84, 2007.

[CH99] Noel Cressie and Hsin-Cheng Huang. "Classes of Nonseparable, Spatio-Temporal Stationary Covariance Functions." *Journal of the American Statistical Association*, **94**(448):1330–1340, 1999.

[Cre93] Noel A.C. Cressie. *Geostatistics, in Statistics for Spatial Data, Revised Edition*. John Wiley & Sons, Inc., Hoboken, NJ, USA, 1993.

[Daw84] A. P. Dawid. "Present Position and Potential Developments: Some Personal Views: Statistical Theory: The Prequential Approach." *Journal of the Royal Statistical Society. Series A (General)*, **147**(2):278–292, 1984.

[Eub05] Randall L Eubank. *A Kalman Filter Primer*. Chapman and Hall/CRC, 2005.

[EW95] Michael D. Escobar and Mike West. "Bayesian Density Estimation and Inference using Mixtures." *Journal of the American Statistical Association*, **90**(430):577–588, 1995.

[Far12] Ramsey Faragher. "Understanding the Basis of the Kalman Filter Via a Simple and Intuitive Derivation." *IEEE Signal Processing Magazine*, **29**:128–132, 2012.

[Fea11] Paul Fearnhead. *MCMC for state-space models*, pp. 513–529. Chapman and Hall, 2011.

[Fer83] Thomas S. Ferguson. "Bayesian Density Estimation by Mixtures of Normal Distributions." *Recent advances in statistics: Papers in honor of Herman Chernoff on his sixtieth birthday*, p. 287?302, 1983.

[GBG] Alan E. Gelfand, Sudipto Banerjee, and Dani Gamerman. "Spatial process modelling for univariate and multivariate dynamic spatial data." *Environmetrics*, **16**(5):465–479.

[GCS13] Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2013.

[GDW04] Simon J Godsill, Arnaud Doucet, and Mike West. "Monte Carlo Smoothing for Nonlinear Time Series." *Journal of the American Statistical Association*, **99**(465):156–168, 2004.

[GFE07] Tilmann Gneiting, Balabdaoui Fadoua, and Raftery Adrian E. "Probabilistic forecasts, calibration and sharpness." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **69**(2):243–268, 2007.

[GG98] Alan E. Gelfand and Sujit K. Ghosh. "Model choice: A minimum posterior predictive loss approach." *Biometrika*, **85**(1):1–11, 1998.

[GK14] Tilmann Gneiting and Matthias Katzfuss. "Probabilistic Forecasting." *Annual Review of Statistics and Its Application*, **1**(1):125–151, 2014.

[GKM05] Alan E Gelfand, Athanasios Kottas, and Steven N MacEachern. "Bayesian Nonparametric Spatial Modeling With Dirichlet Process Mixing." *Journal of the American Statistical Association*, **100**(471):1021–1035, 2005.

[GMR14] Anurag Ghosh, Soumalya Mukhopadhya, Sandipan Roy, and Sourabh Bhattacharya. "Bayesian Inference in Nonparametric Dynamic State-Space Models." *Statistical Methodology*, **21**:35–48, 2014.

[Gne02] Tilmann Gneiting. "Nonseparable, Stationary Covariance Functions for Space-Time Data." *Journal of the American Statistical Association*, **97**(458):590–600, 2002.

[GR07] Tilmann Gneiting and Adrian E Raftery. "Strictly Proper Scoring Rules, Prediction, and Estimation." *Journal of the American Statistical Association*, **102**(477):359–378, 2007.

[GSS93] N.J. Gordon, D.J. Salmond, and A.F.M. Smith. "Novel Approach to Nonlinear/Non-Gaussian Bayesian State Estimation." **140**:107 – 113, 05 1993.

[GV13] Jean Pierre Gauchi and Jean Pierre Vila. "Nonparametric Particle Filtering Approaches for Identification and Inference in Nonlinear State-Space Dynamic Systems." *Statistical Computation*, **23**:523–533, 2013.

[HB18] Jeremie Houssineau and Adrian N Bishop. "Smoothing and Filtering with a Class of Outer Measures." *SIAM/ASA Journal on Uncertainty Quantification*, **6**(2):845–866, 2018.

[HDZ18a] Grace X Hu, D. David R Kuipers, and Yong Zeng. "Bayesian Inference via Filtering Equations for Ultrahigh Frequency Data (I): Model and Estimation." *SIAM/ASA Journal on Uncertainty Quantification*, **6**(1):34–60, 2018.

[HDZ18b] Grace X Hu, D. David R Kuipers, and Yong Zeng. "Bayesian Inference via Filtering Equations for Ultrahigh Frequency Data (II): Model Selection." *SIAM/ASA Journal on Uncertainty Quantification*, **6**(1):61–86, 2018.

[HGL13] Dave Higdon, Jim Gattiker, Earl Lawrence, Charles Jackson, Michael Tobis, Matt Pratola, Salman Habib, Katrin Heitmann, and Steve Price. "Computer Model Calibration Using the Ensemble Kalman Filter." *Technometrics*, **55**(4):488–500, 2013.

[HGW08] Dave Higdon, James Gattiker, Brian Williams, and Maria Rightley. "Computer Model Calibration Using High-Dimensional Output." *Journal of the American Statistical Association*, **103**(482):570–583, 2008.

[HP10] Jay M. Ver Hoef and Erin E. Peterson. "A Moving Average Approach for Spatial Statistical Models of Stream Networks." *Journal Of The American Statistical Association*, **105**:6–18, 2010.

[HYM08] K.I Hoi, K.V. Yuen, and K.M. Mok. "Kalman filter based prediction system for wintertime PM10 concentrations in Macau." *Global NEST Journal*, **10**(2):140–150, 2008.

[IJ01] Hemant Ishwaran and Lancelot F James. "Gibbs Sampling Methods for Stick-Breaking Priors." *Journal of the American Statistical Association*, **96**(453):161–173, 2001.

[IJ02] Hemant Ishwaran and Lancelot F. James. "Approximate Dirichlet Process Computing in Finite Normal Mixtures: Smoothing and Prior Information." *Journal of Computational and Graphical Statistics*, **11**(3):508–532, 2002.

[Jaz07] Andrew H Jazwinski. *Stochastic Processes and Filtering Theory*. Dover Books on Electrical Engineering Series. Dover Publications, 2007.

[KBA09] Charles B Keil, Wil F Ten Berge, and AIHA. *Mathematical models for estimating occupational exposure to chemicals*. AIHA Press, 2009.

[KEM17] RK Kwok, LS Engel, AK Miller, A Blair, MD Curry, Jackson II, W.B., PA Stewart, MR Stenzel, LS Birnbaum, DP Sandler, and the GuLF STUDY Research Team. "(in press) The GuLF STUDY: A prospective study of persons involved in the Deepwater Horizon oil spill response and clean-up." *Environmental Health Perspectives.* doi: 10.1289/EHP715, 2017.

[Lau14] Marcio Poletti Laurini. "Dynamic Functional Data Analysis with Non-parametric State Space Models." *Journal of Applied Statistics*, **41**(1):142–163, 2014.

[LCC02] D.P. Leleux, R. Claps, W. Chen, F.K.Tittel, and T.L. Harman. "Applications of Kalman filtering to real-time trace gas concentration measurements." *Applied Physics B*, **74**(1):85–93, 2002.

[MBR11] J V. D. Monteiro, Sudipto Banerjee, and Gurumurthy Ramachandran. "B2Z: An R Package for Bayesian Two-Zone Models." *Journal of Statistical Software*, **43**(2), 2011.

[MBR14] J V. D. Monteiro, Sudipto Banerjee, and Gurumurthy Ramachandran. "Bayesian Modeling for Physical Processes in Industrial Hygiene Using Misaligned Workplace Data." *Technometrics*, **56**(2):238–247, 2014.

[Nic96] Mark Nicas. "Estimating Exposure Intensity in an Imperfectly Mixed Room." *AIHA Journal*, **57**(6):542–550, 1996.

[NJ02] Mark Nicas and Michael Jayjock. "Uncertainty in Exposure Estimates Made by Modeling Versus Monitoring." *AIHA Journal*, **63**(3):275–283, 2002.

[Ram05] Gurumurthy Ramachandran. *Occupational Exposure Assessment for Air Contaminants*. CRC Press, 2005.

[Ram08] Gurumurthy Ramachandran. "Toward Better Exposure Assessment Strategies?The New NIOSH Initiative." *The Annals of Occupational Hygiene*, **52**(5):297–301, 2008.

[RVD12] Asma Rabaoui, Nicolas Viandier, Emmanuel Duflos, Juliette Marais, and Philippe Vanheeghe. "Dirichlet Process Mixtures for Density Estimation in Dynamic Nonlinear Modeling: Application to GPS Positioning in Urban Canyons." *IEEE Transactions on Signal Processing*, **60**(4):1638–1655, 2012.

[SBC] David J. Spiegelhalter, Nicola G. Best, Bradley P. Carlin, and Angelika Linde. "The deviance information criterion: 12 years on." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **76**(3):485–493.

[SRA17] Yuan Shao, Sandhya Ramachandran, Susan Arnold, and Gurumurthy Ramachandran. "Turbulent Eddy Diffusion Models in Exposure Assessment - Determination of the Eddy Diffusion Coefficient." *Journal of Occupational and Environmental Hygiene*, **14**(3):195–206, 2017.

[SSR17] PA Stewart, MR Stenzel, G Ramachandran, S Banerjee, T Huynh, C Groth, R Kwok, A Blair, LS Engel, and DP Sandler. "(in press) Development of a Total Hydrocarbon Ordinal Job-Exposure Matrix for Workers Responding to the Deepwater Horizon Disaster: The GuLF STUDY." *Exposure Science & Environmental Epidemiology*, 2017.

[TJB06] Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. "Hierarchical Dirichlet Processes." *Journal of the American Statistical Association*, **101**(476):1566–1581, 2006.

[WC99] Christopher K Wikle and Noel Cressie. "A dimension-reduced approach to space-time Kalman filtering." *Biometrika*, **86**(4):815–829, 1999.

[WH97] Mike West and Jeff Harrison. *Bayesian Forecasting and Dynamic Models*. Springer, 1997.

[YDB10] Mingan Yang, David Dunson, and Donna Baird. "Semiparametric Bayes hierarchical models with mean and variance constraints." **54**:2172–2186, 2010.

[ZBL09] Yufen Zhang, Sudipto Banerjee, Claudia Lungu, and Gurumurthy Ramachandran. "Bayesian Modeling of Exposure and Airflow using Two-Zone Models." *Annals of Occupational Hygiene*, **53**(4):409–424, 2009.