Spatio-temporal methods for applications in biological and environmental studies

by

Shiyu He

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Statistics

Waterloo, Ontario, Canada, 2025

© Shiyu He 2025

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner:	Farouk Nathoo Professor, Dept. of Mathematics and Statistics University of Victoria
Supervisor(s):	Samuel W.K. Wong Associate Professor, Dept. of Statistics & Actuarial Science, University of Waterloo
Internal Member:	Charmaine Dean Professor, Dept. of Statistics & Actuarial Science, University of Waterloo
	Alex Stringer Assistant Professor, Dept. of Statistics & Actuarial Science, University of Waterloo
Internal-External Member:	Jane Law Associate Professor, School of Public Health Sciences & School of Planning, University of Waterloo

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

The chapters of this thesis contain Shiyu He's work, under the supervision of Professor Samuel W.K. Wong, which has been published in the following peer-reviewed journals.

Chapter 2: He, S., & Wong, S. W. (2021). Statistical Challenges in the Analysis of Sequence and Structure Data for the COVID-19 Spike Protein. *Journal of Data Science*, 19(2), 314-333.

Chapter 3: He, S., & Wong, S. W. (2024). Spatio-temporal data fusion for the analysis of in situ and remote sensing data using the INLA-SPDE approach. *Spatial Statistics*, 64, 100863.

The entirety of the related code is written in R and is solely authored by Shiyu He for these manuscripts.

Abstract

With advances in technology over the past few decades, the availability and resolution of spatial and spatio-temporal data have increased dramatically. As a result, many research fields now face new analytical and computational challenges in effectively utilizing such complex data. This thesis develops spatio-temporal models within the Bayesian framework to address practical problems in epidemiology, engineering, and environmental science.

In Chapter 2, we develop a Bayesian hierarchical model to investigate the temporal and spatial evolution of spike protein sequences of SARS-CoV-2, which serve as critical targets for vaccines and neutralizing antibodies. To reduce dimensionality and facilitate interpretation, the sequences are grouped into representative clusters based on their similarity. The robustness of the model is demonstrated through simulation studies. We then apply the model to real-world sequence data and the results uncover clear geographical differences in the spread of SARS-CoV-2 variants.

In Chapter 3, we propose a Bayesian hierarchical model to address the challenge of spatial misalignment in spatio-temporal data using the INLA-SPDE approach. The research is motivated by the study of harmful algal bloom (HAB) events in environmental science, where the convention is to conduct separate analyses based on either in situ samples or satellite images. Our methodology combines the different data sources in a "fusion" model via the construction of projection matrices in both spatial and temporal domains. Simulation studies demonstrate that the proposed fusion model generally outperforms standalone models in both parameter estimation and predictive accuracy. This fusion framework thus represents an important step toward a unified characterization of bloom dynamics.

In Chapter 4, we investigate the effects of the spatial arrangement of knots on the flexural properties of lumber crossarms using Bayesian methods. The proposed frameworks integrate a deterministic analysis of induced mechanical stress with a Bayesian modeling approach that captures uncertainties in the failure mechanisms. To further enhance model flexibility, we extend the framework to a Bayesian mixture model, which enables the identification and quantification of multiple underlying failure mechanisms observed in lumber testing.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisor, Samuel W.K. Wong. His mentorship has been instrumental in shaping my academic and personal growth throughout my PhD. His encouragement motivated me to persevere during challenging times, and I am truly thankful for his invaluable guidance, patience, and unwavering support across every stage of this journey.

I am sincerely grateful to Dr. Charmaine Dean, Dr. Alex Stringer, Dr. Farouk Nathoo, and Dr. Jane Law for generously serving as committee members for my dissertation. Their valuable insights and feedback have significantly enriched this work.

I would extend my heartfelt appreciation to all the faculty, staff, and fellow graduate students in our department for their invaluable assistance and support. Most importantly, I am deeply grateful to Mary Lou for her constant help and support. Her compassion and patience provided not only practical assistance but also emotional comfort—something I will always remember with gratitude.

I would love to to sincerely thank Dr. Charleton, Dr. Molckovsky, Dr. Kobeleva, and all the nurse practitioners at Grand River Hospital. Their unwavering care, compassion, and support have made a profound difference during one of the most challenging periods of my life. Whether through medical expertise, care, or kind words, each of them contributed to making this journey more bearable and hopeful. I will always carry with me the memory of their dedication, humanity and warmth.

Lastly, I am deeply grateful to my family and friends for their wholehearted support and encouragement. I am especially thankful to my mom and my partner, whose love, strength, and companionship have carried me through the most challenging moments of this journey. Thank you for your presence and standing by my side.

Dedication

To my mom, with all my love.

Table of Contents

Examining Committee ii

Author's Declaration iii

Statement of Contributions iv

Abstract v

Acknowledgements vi

Dedication vii

List of Figures xi

List of Tables xv

1 Introduction 1

1.1 Dynamic spatio-temporal models 2

1.2 Approximate Bayesian methods in large scale spatio-temporal models . . . 5

1.3 Outline of the thesis 7

2 Statistical challenges in the analysis of sequence and structure data for the COVID-19 spike protein 9

2.1 Introduction 9

2.2 Data description and exploratory analysis . . . 11
2.3 Methods and models . . . 14
2.4 Results . . . 18
2.5 Discussion . . . 22

3 Spatio-temporal data fusion for the analysis of in situ and remote sensing data using the INLA-SPDE approach 24
3.1 Introduction . . . 24
3.2 Method . . . 27
3.2.1 Spatial data fusion model . . . 27
3.2.2 Spatio-temporal data fusion model . . . 28
3.3 Inference . . . 29
3.3.1 INLA-SPDE . . . 29
3.3.2 Missing values in response . . . 31
3.3.3 Approximation of integrals . . . 31
3.4 Simulation Study . . . 33
3.4.1 Simulation Scenarios . . . 33
3.4.2 Evaluation Metric . . . 36
3.4.3 Simulation Results . . . 37
3.5 Application to the cyanobacteria HAB data . . . 46
3.6 Discussion . . . 48
3.7 Supplementary Tables and Figures . . . 52

4 Characterizing the spatial effect of knots on flexural properties of wood crossarms through Bayesian methods 65
4.1 Introduction . . . 65
4.2 Background on mechanical properties . . . 67
4.3 Data description . . . 70
4.4 Methods . . . 71

4.4.1 A hierarchical ICAR model 71
4.4.2 A Geostatistical Gaussian process model 76
4.4.3 A mixture of geostatistical Gaussian process model 78
4.5 Applications to real dataset 84
4.5.1 Tensile stress model 84
4.5.2 A mixture of tensile and shear stress model 88
4.6 Discussion 90
4.7 Supplementary Materials 92
4.7.1 Stress analysis of zones 92
4.7.2 Sample pieces of Douglas-fir crossarms 93

5 Discussion and Future Work **94**

References **97**

List of Figures

2.1 Pie charts of cluster proportions for 9 countries, based on all complete sequences accumulated in GISAID. For each country, the proportions of cluster I to cluster V are represented by the 5 different colors as indicated. 14

2.2 Temporal trend of cluster counts (upper panels) and cluster proportions (lower panels) in the US and UK. The points indicate the observed proportions and counts on each day, while the solid lines show the corresponding smoothed LOESS curves. 15

2.3 Posterior means and 95% credible intervals of inferred cluster proportions from the Bayesian hierarchical model for 9 countries. For each day, the points show the observed proportions, the middle solid lines indicate the posterior means of cluster proportions, and the bands indicate the 95% credible interval of cluster proportions. 19

3.1 A simulated realization of log chl-a concentration in Western Lake Erie on days 1, 5, 10, 15, and 19. The black lines represent the mesh triangulation. The colors superimposed on the mesh indicate the relative levels of log chl-a concentration. 35

3.2 Boxplots for relative bias of parameters $\beta_0, a, \beta_1, \beta_2, \tau_1, \tau_2, \kappa, \sigma_\omega^2$ for the in situ, satellite, and fusion models. k from 1 to 12 represents the simulation scenario number in Table 3.1. The boxplot colors represent the three different maximum inner lengths of the triangulation mesh. For visualization clarity, the parameters κ, σ_ω^2 are displayed on a cubic root scale, and τ_1, τ_2 are displayed on a ninth root scale. 39

3.3 Boxplots for RMSE of parameters $\beta_0, a, \beta_1, \beta_2, \tau_1, \tau_2, \kappa, \sigma_\omega^2$ for the in situ, satellite, and fusion models. k from 1 to 12 represents the simulation scenario number in Table 3.1. The boxplot colors represent the three different maximum inner lengths of the triangulation mesh. For visualization clarity, the parameters $\tau_1, \tau_2, \kappa, \sigma_\omega^2$ are displayed on a logarithmic scale. 40

3.4 Prediction RMSE over unobserved locations on day 1-19 (with test days 15-19 shaded in grey; the training period is days 1-14). Panels 1-12 denote the different scenarios in Table 3.1. The lines colored in red, green, and blue correspond to the in situ, satellite, and fusion models, respectively. 43

3.5 Posterior mean of the latent spatio-temporal field fitted by the in situ, satellite, and fusion model on day 1, 5, 10, 15 and 19 respectively for scenario 12 (30 in situ samples, 80% missingness). The color scale represents the posterior mean level of the log chl-a concentration. 44

3.6 Posterior standard deviation of the latent spatio-temporal field fitted by the in situ, satellite, and fusion model on day 1, 5, 10, 15 and 19 respectively for scenario 12 (30 in situ samples, 80% missingness). The color scale represents the posterior standard deviation of the log chl-a concentration. 45

3.7 In situ monitoring stations and a satellite image example for the study of cyanobacteria HABs in Western Lake Erie. 47

3.8 Comparison between in situ and satellite log chl-a concentration on a matched spatio-temporal scale. The colors of the points represent different stations where in situ data are collected. The estimated regression line is $y = 0.9611([0.7638, 1.1584])x + 0.1463([-0.3231, 0.6156])$. 52

3.9 Missingness in satellite images on different days 53

3.10 An example of simulated in situ samples on days 1, 5, 10, 15, and 19 (5 each day). 53

3.11 Prediction RMSE (calculated using (3.18)) on day 1, 5, 10, 15, and 19. k from 1 to 12 represents the simulation scenario number in Table 3.1. The color in the boxplot represents the three different maximum inner lengths of the triangulation. 54

3.12 Prediction RMSE comparing spatial model with spatio-temporal model . . 55

3.13 Boxplots for relative bias of parameters for the in situ, satellite, and fusion models averaged for factor "No. in situ samples". For visualization clarity, the parameters κ, σ_ω^2 are displayed on a cubic root scale, and τ_1, τ_2 are displayed on a ninth root scale. 58

3.14 Boxplots for RMSE of parameters for the in situ, satellite, and fusion models averaged for factor "No. in situ samples". For visualization clarity, the parameters $\tau_1, \tau_2, \kappa, \sigma^2_\omega$ are displayed on a logarithmic scale. 59

3.15 Boxplots for relative bias of parameters for the in situ, satellite, and fusion models averaged for factor "% missingness". For visualization clarity, the parameters κ, σ^2_ω are displayed on a cubic root scale, and τ_1, τ_2 are displayed on a ninth root scale. 60

3.16 Boxplots for relative bias of parameters for the in situ, satellite, and fusion models averaged for factor "% missingness". For visualization clarity, the parameters $\tau_1, \tau_2, \kappa, \sigma^2_\omega$ are displayed on a logarithmic scale. 61

3.17 Boxplots for relative bias of parameters for the in situ, satellite, and fusion models averaged for factor mesh size. For visualization clarity, the parameters κ, σ^2_ω are displayed on a cubic root scale, and τ_1, τ_2 are displayed on a ninth root scale. 62

3.18 Boxplots for relative bias of parameters for the in situ, satellite, and fusion models averaged for factor mesh size. For visualization clarity, the parameters $\tau_1, \tau_2, \kappa, \sigma^2_\omega$ are displayed on a logarithmic scale. 63

3.19 Parameter RMSE (fusion model) and Prediction RMSE for scenarios with mesh max. inner length of 0.05 vs.0.04. Scenario 1: 5 in situ samples, 50% missingness, Scenario 2: 30 in situ samples, 50% missingness, Scenario 3: 5 in situ samples, 80% missingness, Scenario 4: 30 in situ samples, 80% missingness. 64

4.1 Sample image of knots on the surface of a piece of lumber. 66

4.2 Cantilever stress experiment mimicking the real mechanical loading of utility pole crossarms. Reprinted from "Ability to predict flexural properties of Douglas-fir crossarms," by Anderson et al., *Wood Material Science & Engineering*, © 2021, reprinted by permission of Informa UK Limited, trading as Taylor & Francis Group, https://www.tandfonline.com. 68

4.3 Partitioning of the Douglas-fir crossarm into seven zones to examine failure locations during testing. Reprinted from"Ability to predict flexural properties of Douglas-fir crossarms," by Anderson et al., *Wood Material Science & Engineering*, © 2021, reprinted by permission of Informa UK Limited, trading as Taylor & Francis Group, https://www.tandfonline.com. . . . 71

4.4 Hierarchical ICAR model diagnostics 86

4.5 Geostatistical GP model diagnostics 86
4.6 Mixture model diagnostics 90
4.7 A sample of the Douglas-fir crossarms sent to the test. Facets A and C represent the narrow face of the crossarm, and B and D represent the wide face of the crossarm. Black circles representing knots were regenerated based on the diameter and the location of the knots documented in the data. . . 93

List of Tables

2.1 Relative frequencies of the most common mutations present in each cluster. The top 9 mutations in descending order and their corresponding relative frequencies are shown in the columns for each of clusters I–V. For example, the D614G mutation is present in 89% of the sequences belonging to cluster I. 12

2.2 Top three unique sequences in each cluster, ranked by frequency. For each unique sequence, its specific mutation positions are shown in the right column. For example, 58,271 sequences belonging to cluster I had exactly the one mutation D614G, while 777 sequences in cluster I had exactly the two mutations R21I and D614G. 13

2.3 Posterior means and 95% credible intervals (represented by the 2.5% and 97.5% percentiles) for the parameters in the Bayesian hierarchical model. $\boldsymbol{\alpha}$ represents the daily growth rate parameters for clusters II to V relative to cluster I. σ_C is the random noise standard deviation for countries within continent C, and $\boldsymbol{p}_{\cdot,1}$ represents the initial proportions on Jan 7th, 2020 for all countries. Σ_C for each continent together forms the diagonal block matrix Σ that represents the covariance matrix for the random noise vector. 20

2.4 Sensitivity analysis of the posterior parameters with different priors. Scenario 1 sets $\boldsymbol{p}_{i1} \sim Dirichlet(2.5, 38.5, 0.5, 0.5, 0.5)$. Scenario 2 sets the growth parameter $\alpha_c \sim N(0,1), c = 2, \ldots, 5$. Scenario 3 combines both changes to the priors from scenario 1 and 2. 22

3.1 Summary of the 12 simulation scenarios. The scenarios vary by three changing factors: the number of in situ samples collected each day, the percentage of missingness in satellite images, and the maximum inner length of the triangulation. 36

3.2 Results for the parameter estimates obtained from the log chl-a concentration data during May 2014 - May 2016. The columns show the posterior mean and the 95% CI for each parameter for the fusion model, satellite model and in situ model. 49

3.3 Parameter relative bias for the in situ, satellite, and fusion model. k from 1 to 12 represents the simulation scenario number in Table 3.1. 56

3.4 Average parameter RMSE across 100 simulation runs for the in situ, satellite, and fusion models. k from 1 to 12 represents the simulation scenario number in Table 3.1. 57

4.1 Number of crossarms by zone of failures in the destructive test for 250 kiln-dried Douglas-fir crossarms. 71

4.2 Summary of the true parameter values, posterior mean, standard deviation, and 95% credible intervals of the parameters of the simulated ICAR model. 75

4.3 Summary of the true parameter values, posterior mean, standard deviation, and 95% credible intervals of the parameters of the simulated Gaussian process model. 78

4.4 Summary of the true parameter values, posterior mean, standard deviation, and 95% credible intervals of the parameters of the simulated mixture model. 83

4.5 True vs. estimated latent stress type for the simulation of mixture model. $Z = 1$ indicates the specimen is failed by tensile stress and $Z = 2$ indicates the specimen is failed by shear stress. 84

4.6 Posterior summaries (mean, standard deviation, median, and 95% credible intervals) for parameters under the hierarchical ICAR model. 85

4.7 Posterior summaries (mean, standard deviation, median, and 95% credible intervals) for parameters under the geostatistical GP model. 85

4.8 Comparison of model fit using PSIS-LOO. 88

4.9 Posterior mean, standard deviation, median, and 95% credible intervals for the parameters of the GP mixture model. 89

Chapter 1

Introduction

In the last few decades, the fields of spatial data and spatio-temporal data have expanded greatly, mainly due to technological advances and the capacity to collect and store spatial and spatio-temporal data. Advances in technologies such as satellite images, Global Positioning System (GPS), and harmonic radar tracking have led to a vast influx of data, prior to which direct observation of these data was impossible. Consequently, spatial statistics has received considerable attention, and numerous spatial and spatio-temporal models have been developed and applied in a wide array of research fields, including epidemiology (Beale et al., 2008; Auchincloss et al., 2012), ecology (Griffith and Peres-Neto, 2006; Patterson et al., 2008), climatology (Blois et al., 2013), economics (Pace et al., 2000; Wilhelmsson, 2002) and social science (Wong et al., 2006). The objective of the spatial analysis varies depending on the research question, for example, researchers might be interested in the underlying process of generating the observations in space (spatial point process) or estimating the variables at unobserved locations based on the values at observed locations (spatial interpolation). On the other hand, spatial-temporal models are typically developed for filtering, smoothing, and forecasting in space and time, while providing inference of parameters.

There has been a growing interest in using Bayesian models to deal with spatial and spatio-temporal data starting around the year 2000, with the development of Markov chain Monte Carlo (MCMC) methods (Metropolis et al., 1953; Hastings, 1970). Prior to it, the Bayesian approach had few applications in real studies due to the lack of numerical or analytical tools to compute posterior distributions. The main advantage of the Bayesian approach resides in its flexibility to incorporate multiple levels of randomness, and its capability of dealing with issues such as small sample size, missing data, and hierarchical data structure. As a result, it allows one to take into account a much wider class of

conceptual models than the frequentist approach. In addition, Bayesian approaches allow the specification of prior distributions to incorporate information from preceding studies.

Bayesian spatial models have been developed for various fields of applications. For example, in disease mapping, Bayesian hierarchical models were developed to model the risk surface as a set of spatial random effects, together with any potentially available covariate information (Riebler et al., 2016; Aswi et al., 2019). In the meta-analysis of brain imagery, regions of consistent activation in the brain can be located for diagnosis and treatment (Derado et al., 2013). In ecology, the Bayesian approach has been adopted to identify and map fish species richness and abundant hotspots for the management and conservation of fish stocks (Fonseca et al., 2017). Many other examples can be found in the fields of geostatistics, materials science, agriculture, environmental science, etc.

1.1 Dynamic spatio-temporal models

Dynamic linear models, and more generally state-space models were first developed as a method for Bayesian time series forecasting, e.g., see the method in Harrison and Stevens (1976). Such a framework allows one not only to make predictions about the future, but also to describe the dynamic evolutional aspect of the real-world phenomena. These models were naturally adapted for spatio-temporal analysis, as a spatio-temporal process could be thought of as a time series of spatial processes. The spatio-temporal extension of dynamic linear models, often referred to as dynamic spatio-temporal models (DSTMs), provides a flexible framework for modeling a wide range of spatio-temporal processes. Many scientifically motivated models, e.g., physical, biological, or ecological models, inform the dynamic component of these frameworks, as they are grounded in the understanding of physical or biological mechanisms that drive spatio-temporal variability and interactions. These types of models are well suited for Bayesian analysis due to their associated computational difficulties.

Denote the data from any temporal and spatial domain by $\{Y_t(\boldsymbol{s}) : \boldsymbol{s} \in D_s, t = 1, \ldots, T\}$ which might be partially observed, and denote the latent process by $\{X_t(\boldsymbol{s}) : \boldsymbol{s} \in D_s, t = 1, \ldots, T\}$. In general, we could write the dynamic model in the form of

$$Y_t(\cdot) = \mathcal{H}_t(X_t(\cdot), \boldsymbol{\theta}, \epsilon_t(\cdot)), \quad t = 1, \ldots, T, \tag{1.1}$$

where $Y_t(\cdot)$ represents the data at time t at any arbitrary location, with $X_t(\cdot)$ corresponding to the latent spatio-temporal process. $\mathcal{H}_t(\cdot)$ refers to the linear or non-linear function that maps from the latent process to the observations. The error term specified by $\epsilon_t(\cdot)$ typically refers to measurement error and/or small-scale spatio-temporal variability. Finally,

$\boldsymbol{\theta}$ defines the parameters of the latent process. Typically, interest lies in making inference about $X_{t_0}(\boldsymbol{s})$ for $1 < t_0 \leq T, \boldsymbol{s_0} \in D_s$, even if the observation $Y_{t_0}(\boldsymbol{s_0})$ is not available. Other interests include estimating the parameters $\boldsymbol{\theta}$, as well as predicting $X_{t_1}(\boldsymbol{s_0})$ and $Y_{t_1}(\boldsymbol{s_0})$ for some future time $t_1 > T$, where $\boldsymbol{s_0} \in D_s$.

An important assumption in the hierarchical formulation of dynamic linear models is that the observations $Y_t(\cdot)$ are conditionally independent over time given the latent process $X_t(\cdot)$ and parameters $\boldsymbol{\theta}$. Under this conditional independence assumption, the distribution of the data given the latent process and parameters can be factorized as a product of individual conditional distributions:

$$p(\{Y_t(\cdot)\}_{t=1}^T | \{X_t(\cdot)\}_{t=1}^T, \boldsymbol{\theta}) = \prod_{t=1}^T p(Y_t(\cdot)|X_t(\cdot), \boldsymbol{\theta}). \tag{1.2}$$

Another common modeling assumption is that the latent process satisfies the Markov property. Under the first-order Markov assumption, the latent process at time t, conditional on all past values, depends only on the process at time $t-1$. This assumption simplifies the conditional distribution of $X_t(\cdot)$ given the full history of the latent process:

$$p(X_t(\cdot)|X_{t-1}(\cdot), \ldots, X_1(\cdot), \boldsymbol{\theta}) = p(X_t(\cdot)|X_{t-1}(\cdot), \boldsymbol{\theta}). \tag{1.3}$$

The above assumptions simplify the joint distribution of $\{Y_t(\cdot)\}_{t=1}^T, \{X_t(\cdot)\}_{t=1}^T, \boldsymbol{\theta}$ by allowing it to be factorized into a product of conditional distributions. This factorization reduces computational complexity and enhances the scalability of inference procedures.

Gaussian linear state-space models represent the first important class of DSTM, in which the latent process is assumed to follow additive Gaussian error distributions. Denote the observation vector $\boldsymbol{Y}_t = (Y_t(\boldsymbol{s_1}), \ldots, Y_t(\boldsymbol{s_n}))$, and the latent process by $\boldsymbol{X}_t = (X_t(\boldsymbol{s_1}), \ldots, X_t(\boldsymbol{s_n}))$, the model can be written in vector–matrix form as a linear first-order autoregressive DSTM,

$$\boldsymbol{Y}_t = \boldsymbol{H}_t \boldsymbol{X}_t + \boldsymbol{\epsilon}_t, \quad \boldsymbol{\epsilon}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{V}_t) \tag{1.4}$$

$$\boldsymbol{X}_t = \boldsymbol{G}_t \boldsymbol{X}_{t-1} + \boldsymbol{w}_t, \quad \boldsymbol{w}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{W}_t) \tag{1.5}$$

where $\boldsymbol{H}_t$ and $\boldsymbol{G}_t$ are some known matrices, and the $\boldsymbol{v}_t$ and $\boldsymbol{w}_t$ are two Gaussian error terms, with mean zero and known covariance matrices $\boldsymbol{V}_t$ and $\boldsymbol{W}_t$ respectively. It is assumed that $\boldsymbol{X}_1$ also follows a Gaussian distribution. The measurement error covariance matrix $\boldsymbol{V}_t$ is often taken to be diagonal, assuming independence among observations with constant variance, and the latent process covariance matrix $\boldsymbol{W}_t$ generally incorporates dependence in space or time. Linear DSTMs can be extended to accommodate non-Gaussian and

nonlinear data models, including those from the general exponential family as well as nonparametric formulations. A detailed review of such methods can be found in Wikle et al. (2019).

Although linear Gaussian DSTMs are widely used, many mechanistic processes are more appropriately modeled in a nonlinear fashion, at least at certain spatial and temporal scales of variability. A general nonlinear DSTM can be specified as

$$Y_t(\cdot) = \mathcal{M}(Y_{t-1}(\cdot), \eta_t(\cdot); \boldsymbol{\theta}), \tag{1.6}$$

where $\mathcal{M}$ is a nonlinear function, $\eta_t(\cdot)$ is an error process, and $\boldsymbol{\theta}$ are parameters. In such settings, subject-matter expertise is often incorporated to embed prior scientific knowledge of the data-generating mechanism. Examples include state-dependent DSTMs, where the transition matrix evolves with the process value at each time, and quadratic nonlinear DSTMs (e.g., those involving diffusion and advection) that explicitly capture mechanistic spatiotemporal dynamics. Recent nonlinear DSTMs in the statistical literature include regime-switching models (Wu et al., 2013), agent-based models (Hooten and Wikle, 2010), analog models (McDermott and Wikle, 2016), among others.

With the advent of high-dimensional, complex datasets, it has become increasingly difficult to specify parsimonious parameterizations for nonlinear DSTMs that are scientifically meaningful. Deep hierarchical DSTMs are frequently employed to model complex spatiotemporal data applications. These models incorporate multiple layers, including the data model, conditional mean, process model, dynamic process model, non-dynamic process model, prior distributions, and hyper-prior distributions (Wikle, 2019). Nonetheless, implementing such models typically relies on MCMC algorithms, which are computationally intensive and require substantial prior scientific knowledge.

Recently, a growing body of work has adopted hybrid approaches for modeling complex spatio-temporal data. These approaches combine statistical modeling principles with deep neural network architectures and have produced notable advances in spatio-temporal modeling. For instance, spatio-temporal echo state networks have been applied to generate long-lead forecasts of nonlinear spatio-temporal processes (McDermott and Wikle, 2017, 2019b). Similarly, convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have been integrated into hierarchical statistical frameworks to enable scalable forecasting of diverse spatio-temporal processes (McDermott and Wikle, 2019a; Zammit-Mangion and Wikle, 2020).

1.2 Approximate Bayesian methods in large scale spatio-temporal models

For Bayesian methods developed to model large real-time datasets, such as those from GPS, satellite imagery, and other remote sensing technologies, a major limitation is the substantial computational burden of running MCMC algorithms. This challenge is particularly important because spatio-temporal datasets often feature high spatial and temporal resolution, as well as integration of data from multiple sources. Applying MCMC methods to such complex structures can result in prohibitively long computing times.

In recent years, several classes of methods have been developed to enable approximate inference in Bayesian analysis. These methods offer computationally feasible solutions to statistical problems involving intractable likelihoods, high-dimensional models, or large datasets. Approximation techniques can be broadly categorized into simulation-based approaches, such as approximate Bayesian computation (ABC) and Bayesian synthetic likelihood (BSL), and optimization-based approaches, including variational Bayes (VB) and integrated nested Laplace approximation (INLA). For a comprehensive review, see Martin et al. (2024).

In the spatial and spatio-temporal domain, Bayesian hierarchical models offer considerable flexibility for capturing complex hierarchical structures and stochastic behavior. Leveraging the sparse precision structure of Gaussian Markov random fields (GMRFs), the INLA method enables fast and accurate approximations to posterior marginals in high-dimensional spatial and spatio-temporal models with latent Gaussian processes, as demonstrated by Rue et al. (2009). The class of latent Gaussian models includes a broad spectrum of empirically relevant models, such as generalized linear models, state-space models, and spatial or spatio-temporal models.

We outline the basic structure of model specification in INLA as follows. The full set of unknown variables is partitioned into a vector of latent Gaussian variables $\mathbf{x}$ and a vector of hyper-parameters $\boldsymbol{\theta}$ that define the latent Gaussian model. Each observation y_i, for $i = 1, \ldots, n$, is assumed to belong to an exponential family and to be conditionally independent given a linear predictor η_i. Typically, the dimension of the latent Gaussian field exceeds that of the observations. The model is formulated as:

$$\boldsymbol{y}|\mathbf{x}, \boldsymbol{\theta} \sim \prod_i p(y_i|\eta_i(\mathbf{x}), \boldsymbol{\theta}) \qquad \text{Observations} \tag{1.7}$$

$$\mathbf{x}|\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{Q}^{-1}(\boldsymbol{\theta})) \qquad \text{Latent Gaussian field} \tag{1.8}$$

$$\boldsymbol{\theta} \sim p(\boldsymbol{\theta}) \qquad \text{Hyper-parameters} \tag{1.9}$$

The structured additive predictor accounts for the effects of various covariates in an additive way:

$$\eta_i = \alpha + \sum_{j=1}^{n_f} f^{(j)}(u_{ji}) + \sum_{k=1}^{n_\beta} \beta_k z_{ki} + \epsilon_i, \tag{1.10}$$

where $\{f^{(j)}(\cdot)\}s$ are unknown functions of the covariates u, $\{\beta_k\}s$ represent the linear effect of covariates z, and ϵ_i are unstructured terms. This predictor structure can incorporate a wide class of models, as the unknown functions $f^{(j)}(\cdot)$ can take on very different forms, e.g., random effects, autoregressive structure, smoothing splines, etc.

The latent Gaussian field $\mathbf{x}$ is defined through a precision matrix $\boldsymbol{Q}(\boldsymbol{\theta})$ (i.e., the inverse of the covariance matrix), which exhibits a Gauss–Markov structure due to the conditional independence of the latent variables. This structure results in a sparse precision matrix, which allows for efficient numerical computation. For example, the computational complexity of matrix inversion can be reduced from $O(n^3)$ to approximately $O(n^{3/2})$ by leveraging the sparsity of the precision matrix. The resulting joint posterior distribution of the latent variables $\mathbf{x}$ and hyper-parameters $\boldsymbol{\theta}$ is expressed as:

$$p(\mathbf{x}, \boldsymbol{\theta}|\boldsymbol{y}) \propto \exp\big\{ -\frac{1}{2}\mathbf{x}^{\mathbf{T}}\boldsymbol{Q}(\boldsymbol{\theta})\mathbf{x} + \sum_i \log p(y_i|\eta_i, \boldsymbol{\theta}) + \sum \log p(\boldsymbol{\theta})\big\}. \tag{1.11}$$

The main aim is to approximate the posterior marginals $p(\boldsymbol{\theta}|\boldsymbol{y}), p(\mathbf{x}_j|\boldsymbol{y})$ and $p(\theta_k|\boldsymbol{y})$. The marginal posterior of $\boldsymbol{\theta}$ is obtained by the approximation:

$$\tilde{p}(\boldsymbol{\theta}|\boldsymbol{y}) \propto \frac{p(\mathbf{x}, \boldsymbol{\theta}, \boldsymbol{y})}{\tilde{p}_G(\mathbf{x}|\boldsymbol{\theta}, \boldsymbol{y})}\bigg|_{\mathbf{x}=\boldsymbol{\mu}(\boldsymbol{\theta})}, \tag{1.12}$$

where $\tilde{p}_G(\mathbf{x}|\boldsymbol{\theta}, \boldsymbol{y})$ is the Gaussian approximation to the true $p(\mathbf{x}|\boldsymbol{\theta}, \boldsymbol{y})$ by matching the mode and curvature around the mode of $p(\mathbf{x}|\boldsymbol{\theta}, \boldsymbol{y})$. The marginal posteriors of latent variables $\mathbf{x}_j$, and the marginal posteriors of hyper-parameters $\boldsymbol{\theta}_k$ are approximated by integrating with respect to $\mathbf{x}$ and $\boldsymbol{\theta}$:

$$p(\mathbf{x}_j|\boldsymbol{y}) = \int p(\mathbf{x}_j|\boldsymbol{\theta}, \boldsymbol{y}) p(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta}, \tag{1.13}$$

$$p(\boldsymbol{\theta}_k|\boldsymbol{y}) = \int p(\boldsymbol{\theta}|\boldsymbol{y}) d\boldsymbol{\theta}_{-k}. \tag{1.14}$$

The approximations to $p(\theta_k|\boldsymbol{y})$ are computed by integrating out $\boldsymbol{\theta}_{-k}$ from the approximated $p(\boldsymbol{\theta}|\boldsymbol{y})$. Since direct numerical integration is computationally demanding, previously computed points are instead used to construct an interpolant to $\log \tilde{p}(\boldsymbol{\theta}|\boldsymbol{y})$, from

which the marginals can then be obtained through numerical integration. The approximation to $p(\mathbf{x}_j|\boldsymbol{\theta}, \boldsymbol{y})$ can be obtained in two ways, either from the Gaussian approximation $p_G(\mathbf{x}|\boldsymbol{\theta}, \boldsymbol{y})$, or from another Laplace approximation:

$$\tilde{p}(\mathbf{x}_j|\boldsymbol{\theta}, \boldsymbol{y}) \propto \frac{p(\mathbf{x}, \boldsymbol{\theta}|\boldsymbol{y})}{\tilde{p}_G(\mathbf{x}_{-j}|\mathbf{x}_j, \boldsymbol{\theta}, \boldsymbol{y})}\bigg|_{\mathbf{x}_{-j}=\boldsymbol{\mu}_{-j}(\theta)}, \tag{1.15}$$

where $\boldsymbol{\mu}_{-j}(\theta)$ is the mode of $p_G(\mathbf{x}_{-j}|\mathbf{x}_j, \boldsymbol{\theta}, \boldsymbol{y})$. While the Gaussian approximation is computationally efficient and often yields reasonable results, it can be inaccurate for non-Gaussian likelihoods due to errors in the location and/or the lack of skewness (Rue and Martino, 2007). In contrast, incorporating the additional nested Laplace approximation, although computationally more demanding, yields inference of sufficient accuracy to ensure reliable results. Details of these approximation methods can be found in Rue et al. (2009).

Recently, new formulations of the internal model representation have been developed that achieve faster inference alongside improved numerical stability and scalability (Van Niekerk et al., 2023). In this modern approach, the linear predictors are no longer augmented within the latent field $\mathbf{x}$; instead, their marginal posteriors are computed after obtaining $\tilde{p}(\mathbf{x}_j|\boldsymbol{y})$ and $\tilde{p}(\boldsymbol{\theta}|\boldsymbol{y})$. This reformulation yields computational advantages in data-rich settings, particularly when the data greatly exceeds the size of the model or involves higher-dimensional structures.

As a result of the relatively accurate and fast analytical approximations to posterior quantities of interest, many space-time case studies have been conducted using INLA; see, for example, Schrödle and Held (2011); Schrödle et al. (2012) in disease mapping, Cameletti et al. (2013) in air pollution risk assessment, and Lindgren et al. (2011) in global climate modeling.

1.3 Outline of the thesis

In this thesis, we propose novel spatial and spatio-temporal modeling frameworks, supported by simulation studies and applications to biological and environmental datasets. Our work emphasizes methodological development, simulation-based evaluation, and the analysis of real-world data to extract meaningful scientific insights. In Chapter 2, we develop methods for the analysis of sequence data of S-protein in the SARS-CoV-2. The chapter is adapted from the paper "Statistical challenges in the analysis of sequence and structure data for the COVID-19 spike protein" that was published in the *Journal of Data Science* (He and Wong, 2021). In Chapter 3, we develop a spatio-temporal framework to address the spatial misalignment between satellite imagery and in situ samples,

with applications in the analysis of harmful algal blooms. The chapter is published as "Spatio-temporal data fusion for the analysis of in situ and remote sensing data using the INLA-SPDE approach" in *Spatial Statistics* (He and Wong, 2024). In Chapter 4, we develop novel Bayesian frameworks to quantify the effects of knot spatial arrangement in wood crossarms. Finally, in Chapter 5, we conclude the thesis with a brief summary and discussion of the preceding chapters, followed by some potential future work.

Chapter 2

Statistical challenges in the analysis of sequence and structure data for the COVID-19 spike protein

2.1 Introduction

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), a strain of novel coronavirus that caused the COVID-19 outbreak in Wuhan, China in December 2019, has quickly spread across the world and has been characterized as a global pandemic (Zhou et al., 2020a; Wu et al., 2020). As of December 19, 2020, there have been over 74 million probable or confirmed cases of COVID-19, and the illness has been associated with 1.66 million deaths around the world (WHO, 2020a). The development of vaccines and antibody-based therapeutic agents has been initiated since the beginning of the pandemic and several have moved into phase III trials (Krammer, 2020; WHO, 2020b). Results concerning the long-term immunogenicity and efficacy of these vaccine candidates are a subject of continued research. Meanwhile, the virus has been found to mutate in human-to-human transmissions over time, and these changes can potentially alter the efficacy of these interventions. Therefore, it is also of vital importance to identify and study mutations with possible fitness advantages and increased infectiousness.

SARS-CoV-2 is a single-stranded RNA virus, and RNA viruses are known to have high mutation rates and genetic diversity compared to DNA viruses (Duffy, 2018; Lauring and Andino, 2010). Their ability to evolve underlies why they can adapt to novel hosts and develop resistance to either vaccine or infection-induced immunity. Often, the most

consequential mutations in terms of viral functions and resistance to neutralizing antibodies are those that alter the surface proteins of the virus. For instance, the mutation A82V in the Ebola virus glycoprotein was confirmed to have enhanced infectivity and increased the severity of the EVD epidemic (Diehl et al., 2016). Further, co-occurring mutations of A143V and R148K in the influenza H7N9 surface protein led to a 10-fold reduction in its sensitivity to neutralizing antibodies (Ning et al., 2019). As a result, mutations in the SARS-CoV-2 genome are being continuously monitored over time, and a major public repository for sequenced genomes is GISAID (`https://gisaid.org`). In addition to collecting viral genome data, GISAID also provides tools for visualizing the spread of various mutations, organized into phylogenetic clusters (also known as *clades*) over space and time.

Four structural proteins – spike (S), envelope (E), membrane (M), and nucleocapsid (N) – are the building blocks for the SARS-CoV-2 virus particle (Phan, 2020). Out of these four proteins, the S-protein plays the most critical role in attachment and entry into host cells, through its binding with the human ACE2 receptor (Wan et al., 2020). For this reason, the S-protein is the major target of many vaccines and neutralizing antibodies against SARS-CoV-2 (Amanat and Krammer, 2020). In the event of infection, these antibodies can disrupt the spike protein's ability to bind with the ACE2 receptor, thereby blocking its entry into host cells. The analyses in this paper focus on mutations in the amino acid sequence of the S-protein due to its particular importance.

Among all currently known S-protein sequence variants resulting from mutations in the underlying genome, D614G has been the most extensively studied due to its relatively early emergence and subsequent prevalence. The notation "D614G" means that the amino acid D (aspartic acid) in position 614 of the original (or *reference*) sequence has mutated to G (glycine), where the letters are used to denote the 20 different amino acid types. A rapid increase in D614G was observed in many regions after its initial appearance, which suggested fitness advantages and the hypothesis that variants with D614G are likely more infectious (Korber et al., 2020). This was later corroborated by experimental evidence that D614G, either by itself or in conjunction with other mutations, is significantly more infectious than the reference S-protein sequence (Li et al., 2020). Overall, the continued evolution of the virus has resulted in thousands of distinct S-protein sequence variants recorded in GISAID, although many of these only differ by a few mutated sequence positions. While clinical or laboratory experiments can test the infectivity of specific mutations, it is challenging to analyze large numbers of sequence variants.

Computational researchers have thus used clustering as a means to gain interpretable insight into the effect of S-protein mutations across different geographical regions. Temporal changes in the prevalence of S-protein mutations have also been studied in related

research. For instance, Chen et al. (2020b) clustered mutations occurring in the receptor binding domain (RBD) of the S-protein and studied binding affinity changes for each cluster. Based on common amino acid mutations, Toyoshima et al. (2020) classified 28 countries into three clusters and studied correlations between fatality rate and S-protein D614G variants. In addition, the hypotheses of monotonic trends for D614G and various other mutations have been tested using isotonic regression by Korber et al. (2020) and in the COVID-19 pipelines of the Los Alamos National Laboratory (`https://cov.lanl.gov`). However, to the best of our knowledge, few authors have built comprehensive statistical models for the evolution of S-protein mutant clusters (i.e., groups of closely related sequence variants) over space and time. Such models can have an important practical value in providing forecasts and early warnings for countries where S-protein sequence variants with potential fitness advantages or higher infectiousness are actively being transmitted. To that end, this paper presents one such Bayesian hierarchical model for multinomial time series that can help tackle this problem.

Overall then, our goal in this paper to illustrate statistical ideas for tackling the aforementioned challenges associated with the analysis of S-protein sequence data. Specifically based on presently available data, we study temporal and spatial changes in the mutations of S-protein sequences and their structural impact, with the aim to better understand the ongoing evolution of the disease. Specifically, we develop a Bayesian hierarchical model to study the evolution of mutant clusters, and discuss our findings and relate them to other recent work reported in the literature.

2.2 Data description and exploratory analysis

The S-protein sequence dataset for SARS-CoV-2 was obtained from GISAID on Oct 14th, 2020, with the number of sequences totaling 98,699 after incomplete sequences were removed. The full S-protein, based on the first discovered reference sequence, is 1273 amino acids long. Out of all complete sequences, 3,205 of them are unique, indicating that viral evolution has resulted in substantial genetic diversity. Our analysis of complete sequences shows that D614G is the most frequent mutation, appearing in 86.5% of recorded sequences, followed by S477N (6.3%), A222V (3.6%), L18F (1.9%), and L5F (0.99%), R21I (0.98%), and D936Y (0.74%). Many of these mutations are also mentioned in the recent literature where mutation analysis was considered (Korber et al., 2020; Chen et al., 2020a; Hodcroft et al., 2020). The sequences from GISAID are indexed by country and date of deposition, which allows us to conveniently group them for subsequent analysis. Sequences are separately deposited by local laboratories, therefore sequence counts vary widely by

country and may not be well-correlated with actual case counts. Due to this concern, we instead focus on the relative prevalence, i.e., proportions of counts, throughout this paper.

To analyze the large numbers of sequence variants, we first implemented hierarchical clustering to group the unique sequences according to their similarities. The distance matrix for hierarchical clusters was based on the number of pairwise mismatched letters and the Ward-D linkage criterion, which creates groups such that variance is minimized within clusters (Ward, 1963). For our illustrative analysis, we chose to use a total of five clusters, which aims to achieve a balance between separability of the different clusters and interpretability of the results. Table 2.1 shows the most common mutations in each cluster and their frequencies, where it can be seen that these clusters all have identifiable patterns. For example, D614G is dominant in cluster I and present in 89% of sequences within that cluster, while cluster II has the lowest frequency of mutations, indicating that it is composed of sequences with a high level of similarity to the original reference sequence. In clusters III, IV, and V, D614G frequently occurred together with L5F, S477N, A222V respectively, and these paired mutations are observed in almost all sequences within those clusters, evidencing a high probability for the co-occurrence of some common mutations. Evidence for these co-occurrences can also be seen in Table 2.2, which displays the top three unique sequences in each cluster (as ranked by frequency within that cluster) and their specific mutation positions.

	Cluster I		Cluster II		Cluster III		Cluster IV		Cluster V	
	Mutant	Freq	Mutant	Freq	Mutant	Freq	Mutant	Freq	Mutant	Freq
1	D614G	89%	P863H	3%	L5F	98%	D614G	100%	D614G	99%
2	S477N	3%	A262T	2%	D614G	87%	S477N	100%	A222V	98%
3	L5F	2%	Y453F	2%	H655Y	15%	T632N	8%	L18F	54%
4	D936Y	2%	T572I	1%	A222V	13%	L822F	4%	A262S	13%
5	A222V	2%	V615I	1%	V3G	10%	S939F	4%	P272L	9%
6	R21I	2%	K77M	1%	D574Y	10%	W258L	4%	D1163Y	9%
7	L54F	1%	A845S	1%	S459Y	6%	E1144Q	4%	L5F	7%
8	P1263L	1%	L8V	1%	M1229I	4%	G566S	4%	G1167V	6%
9	Q677H	1%	H655Y	1%	T859I	4%	P330A	2%	L176F	4%

Table 2.1: Relative frequencies of the most common mutations present in each cluster. The top 9 mutations in descending order and their corresponding relative frequencies are shown in the columns for each of clusters I–V. For example, the D614G mutation is present in 89% of the sequences belonging to cluster I.

We selected 9 countries from the GISAID database to study based on the larger numbers

Cluster	Frequency	Mutation Positions
I	58,271	D614G
	777	R21I, D614G
	602	D614G, D936Y
II	11,617	Reference Sequence
	117	A829T
	43	L8V
III	486	L5F, D614G
	121	L5F
	43	L5F, A222V, D574Y, D614G, H655Y
IV	5,472	S477N, D614G
	85	S477N, D614G, T632N
	51	S477N, D614G, A930V
V	1,388	A222V, D614G
	1,369	L18F, A222V, D614G
	150	A222V, A262S, P272L, D614G

Table 2.2: Top three unique sequences in each cluster, ranked by frequency. For each unique sequence, its specific mutation positions are shown in the right column. For example, 58,271 sequences belonging to cluster I had exactly the one mutation D614G, while 777 sequences in cluster I had exactly the two mutations R21I and D614G.

of sequences deposited, which are United States (US), Canada (CA), United Kingdom (UK), Netherlands (NL), France (FR), Spain (SP), China (CN), India (IN), and Australia (AU). The pie charts in Figure 2.1 show the composition of clusters for each country for sequences accumulated since the outbreak. Countries with distinctly different compositions are China and Australia, where China has the majority of its sequences from cluster II, and Australia has cluster IV as its major cluster. Cluster I is the largest cluster for the rest of the countries, followed by cluster II, while cluster V has a noticeable presence in Europe, especially the UK.

The composition of clusters also changes over time, and as examples we show the temporal trend of the daily cluster proportions (left panels) and counts (right panels) for the United States (US) and the United Kingdom (UK) in Figure 2.2. The graphs show a major difference between the US and UK composition trends over time: cluster I in the UK peaked around May to July and cluster V saw a surge since late August or early September, while cluster I remains dominant in the US. Similarly, most European countries

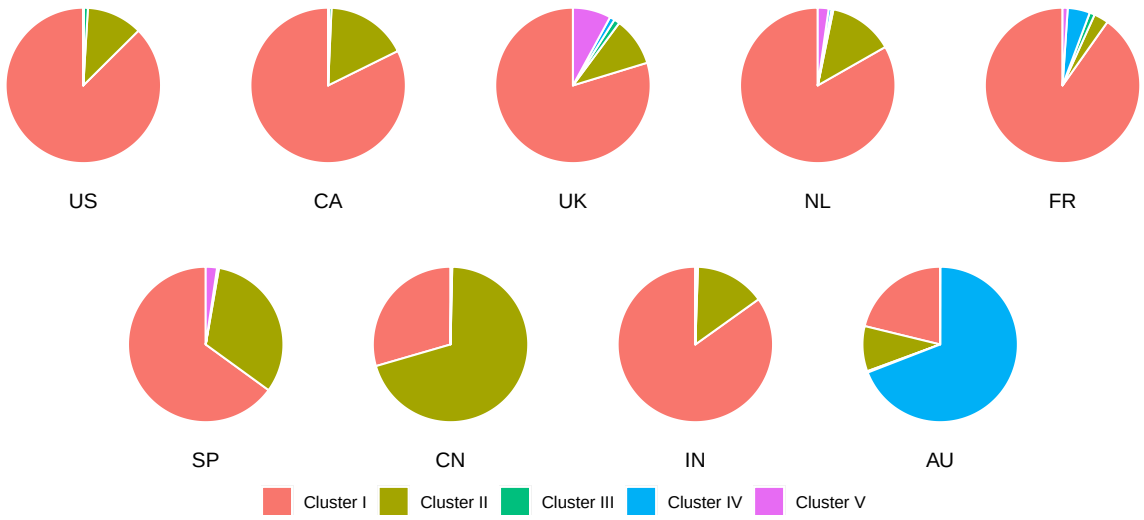


Figure 2.1: Pie charts of cluster proportions for 9 countries, based on all complete sequences accumulated in GISAID. For each country, the proportions of cluster I to cluster V are represented by the 5 different colors as indicated.

have cluster V surging during this period, which suggests the most common mutations in cluster V, D614G and A222V, might be related to the rise of infections across Europe during the late summer and early autumn of 2020 (Dong et al., 2020). In addition to this, the majority of countries have cluster II proportions decreasing over time in direct correspondence to the emergence of cluster I infections over time. It can be seen that the GISAID sequence counts in Figure 2.2 panels (b) and (d) vary widely over time and do not necessarily correspond well with the actual case counts in the US and UK during this period.

2.3 Methods and models

In this section we present a Bayesian hierarchical model for the observed sequence counts, by day and country, in each of the five clusters identified via our exploratory analysis. In motivating the model, we recall that the number of deposited GISAID sequences varies widely over time and by country. Thus, it is sensible to focus inference on the fraction of sequences belonging to each cluster (i.e., cluster proportions) in each country, as in

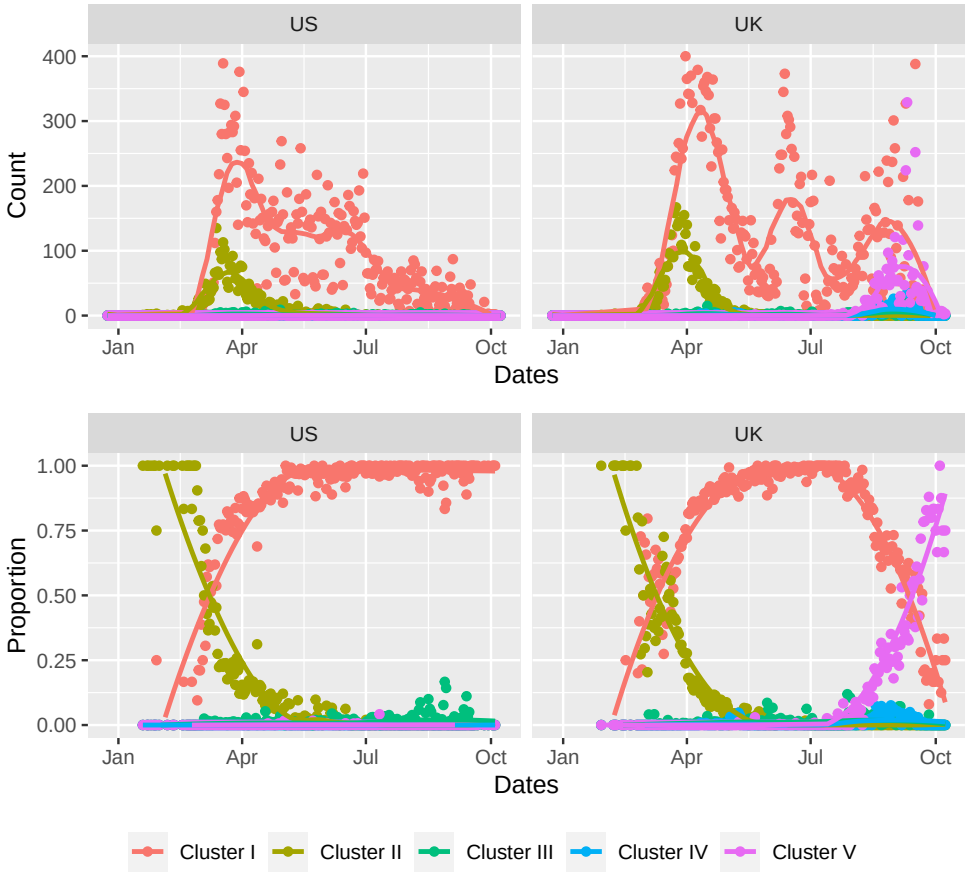


Figure 2.2: Temporal trend of cluster counts (upper panels) and cluster proportions (lower panels) in the US and UK. The points indicate the observed proportions and counts on each day, while the solid lines show the corresponding smoothed LOESS curves.

compositional data analysis (Aitchison, 1982). Also, while LOESS identified some temporal trends for these cluster proportions in each country, it cannot be used for prediction or for comparing the overall growth rates of different clusters. Thus, our model contains the following key features. First, it provides estimates of the cluster proportions for each country on any given day. Second, it assumes a growth rate parameter for each cluster that is common across all countries, through which differences in prevalence among the S-protein mutant clusters can be quantified. Third, the temporal evolution of the cluster proportions is allowed to be dependent across countries via correlated errors.

Denote the observed sequence counts by the vectors $\boldsymbol{y}_{it} = (y_{it1}, \ldots, y_{it5})$, where y_{itc} is the number of sequences observed in GISAID for country $i = 1, \ldots, K$ on day $t \geq 1$ belonging to cluster $c = 1, \ldots 5$ (corresponding to clusters I–V in section 2.2). We then let

$$\boldsymbol{y}_{it} \sim Multinomial\left(n_{it}, \frac{1}{14}\sum_{j=t-13}^{t} \boldsymbol{p}_{ij}\right) \text{ for } t \geq 14 \tag{2.1}$$

where $n_{it} = \sum_{c=1}^{5} y_{itc}$ is the total number of GISAID sequences recorded for country i on day t, and $\boldsymbol{p}_{it} = (p_{it1}, \ldots, p_{it5})$ is the vector of probabilities representing the true underlying cluster proportions for country i on day t such that $\sum_{c=1}^{5} p_{itc} = 1$. Thus the model assumes that the observed sequences represent a random sample from the population of infected individuals in a country, with a reporting delay uniformly at random over the commonly assumed 14 day incubation period for the virus (Lauer et al., 2020). The sampling fraction (i.e., number of reported GISAID sequences out of the number of infected individuals in the country) may vary over time, and this setup permits inference on $\boldsymbol{p}_{it}$ in a manner that accommodates that sampling variability. Prior to Jan 20th, 2020 (i.e., when $t < 14$ in the model), we simply take the multinomial probability to be the average of the days modeled thus far, namely $\frac{1}{t}\sum_{j=1}^{t} \boldsymbol{p}_{ij}$.

We apply a log-ratio transformation on the cluster proportions, as commonly used in compositional data analysis (Aitchison, 1999). Treating cluster I as the baseline, we define

$$\tilde{\boldsymbol{p}}_{it} = \left[\log\left(\frac{p_{it2}}{p_{it1}}\right), \log\left(\frac{p_{it3}}{p_{it1}}\right), \log\left(\frac{p_{it4}}{p_{it1}}\right), \log\left(\frac{p_{it5}}{p_{it1}}\right)\right] \tag{2.2}$$

and model these transformed $\tilde{\boldsymbol{p}}_{it}$ values according to

$$\tilde{\boldsymbol{p}}_{it} = \tilde{\boldsymbol{p}}_{i,t-1} + \boldsymbol{\alpha} + \boldsymbol{\epsilon}_{it} \text{ for } t > 1 \tag{2.3}$$

where $\boldsymbol{\alpha} = (\alpha_2, \alpha_3, \alpha_4, \alpha_5)$ is a vector of growth rate parameters for clusters II to V, and $\boldsymbol{\epsilon}_{it} = (\epsilon_{it2}, \epsilon_{it3}, \epsilon_{it4}, \epsilon_{it5})$ represents a random noise vector with mean zero that governs daily

fluctuations. Note that when $\boldsymbol{\epsilon}_{it} = \mathbf{0}$, equation (2.3) implies

$$p_{itc} = \frac{p_{i,t-1,c} \exp(\alpha_c)}{\sum_{c=1}^{5} p_{i,t-1,c} \exp(\alpha_c)} \propto p_{i,t-1,c} \exp(\alpha_c) \tag{2.4}$$

for $c = 1, \ldots, 5$ with α_1 defined to be 0, so that $\exp(\boldsymbol{\alpha})$ can be interpreted as the multiplicative daily growth rates of clusters II to V relative to cluster I, with the proportions normalized to sum to 1, similar to a model recently used to study the prevalence of different flu strains (Huddleston et al., 2020). The $\boldsymbol{\alpha}$ are assumed to be the same for all countries, to reflect the intrinsic fitness of each mutant cluster. Define $\boldsymbol{\epsilon}_{1:K,t,c} = (\epsilon_{1tc}, \ldots, \epsilon_{Ktc})$, namely the random noise vector on day t across all K countries for cluster c, where $c \in \{2, 3, 4, 5\}$. Then we let $\boldsymbol{\epsilon}_{1:K,t,c} \sim N_K(\mathbf{0}, \Sigma)$ independently for each day and cluster, where N_K denotes a K-variate Normal distribution and Σ is a covariance matrix. This formulation allows for spatial dependence in the sense that the noise term can be correlated among countries; intuitively, if a certain cluster experiences faster than expected growth in one country for a period of time, nearby or geographically linked countries may experience similar changes. In practice, we could set Σ to be of block diagonal form, for example, with each block as countries within the same continent, assuming correlations between different continents are likely negligible. Overall, our model setup follows the general structure of Gaussian dynamic models for multinomial time series described in Cargnoni et al. (1997).

We complete the model specification with the choice of priors. During the early outbreak period from Dec 24th, 2019 to Jan 6th, 2020, a total of 40 sequences worldwide were deposited in GISAID, with two that we would now classify to be in cluster I and 38 in cluster II. Thus we set a Dirichlet prior for the cluster proportions in the model for the starting date of the model (Jan 7th, 2020), $\boldsymbol{p}_{i1} \sim \mathit{Dirichlet}(3, 39, 1, 1, 1)$ independently for all $i = 1, \ldots K$ countries, obtained by adding the cluster I and II observations as pseudocounts to a uniform Dirichlet distribution. Weakly informative Cauchy priors with scale 0.5 are assigned to the each unique diagonal (variance) element of Σ. A weakly informative LKJ correlation distribution with shape parameter 2 is assigned as the prior for the correlation matrices corresponding to the blocks of Σ. Finally, uniform priors are assigned for $\boldsymbol{\alpha}$.

To obtain the samples for the posterior distribution of the parameters, Markov chain Monte Carlo sampling for the model was carried out via Stan (Carpenter et al., 2017). Four parallel chains were run, with 5000 iterations each and the first half discarded as burn-in.

2.4 Results

The posterior distribution of the parameters in the Bayesian hierarchical model (Table 2.3) show that the daily growth rate parameters of clusters II to V relative to cluster I, estimated via their posterior means, are -0.05 (-0.07, -0.03), 0.00 (-0.01, 0.02), 0.02 (-0.01, 0.04), 0.03 (0.00, 0.05), with 95% credible intervals in brackets. These estimates show that during the study period, the sequences from clusters IV and V tend to have higher growth relative to clusters I, II, and III, of which mutant cluster II clearly has the weakest growth, as seen via its posterior interval that does not overlap the others. Referring to the mutation positions in Table 2.2, this indicates that sequences with amino acid D in position 614 (as in the reference sequence) will tend to decrease in prevalence over time in the presence of the other mutant clusters. In addition, cluster I, which is mostly characterized by the lone D614G mutation, may have a growth disadvantage if clusters IV or V are also spreading in the country. The clusters with the strongest growth (IV and V) are primarily composed of variants with the co-occurrence of D614G with S477N or A222V.

The covariance matrix Σ for the random noise vector is set up in block diagonal form: each block represents countries within the same continent and parameterized as $\Sigma_C = diag(\sigma_C) \times \Omega_C \times diag(\sigma_C)$, where σ_C is the standard deviation of the daily noise term for countries within continent C, and Ω_C is the corresponding correlation matrix. Specifically, we defined four continents: North America (NA) as (US, CA); Europe (EU) as (UK, NL, FR, SP); Asia (AS) as (CN, IN); and Australia (AU) as its own continent. The posterior means for σ_{EU} and σ_{AS} are relatively larger than σ_{NA} and σ_{AU}, indicating that overall cluster growth trends are somewhat more predictable in North America and Australia than Europe or Asia. On the other hand, spatial dependence in the noise terms in general is low for countries studied, with no estimated correlations exceeding 0.4. That said, based on the posterior means, European countries still have relatively larger spatial correlations in their daily fluctuations, e.g., UK and Netherlands have a correlation of 0.33, UK and France have a correlation of 0.28, and Netherlands and Spain have a correlation of 0.22. The estimates show these European countries may experience more similar day-to-day changes, while countries in North America and Asia do not, as the correlation appears to be negligible between US and Canada (-0.087) and between China and India (-0.045), both of which are close to 0. The relatively low correlations between countries suggest that zero noise correlation between continents (i.e., block diagonal Σ) is a reasonable simplifying assumption.

Figure 2.3 shows the posterior means and 95% credible intervals for the inferred cluster proportions in the different countries on each day t from Jan 7th, 2020 to Oct 14th, 2020. The posterior means of $\boldsymbol{p}_{\cdot,1}$ indicate that on Jan 7th, 2020, cluster I accounts for

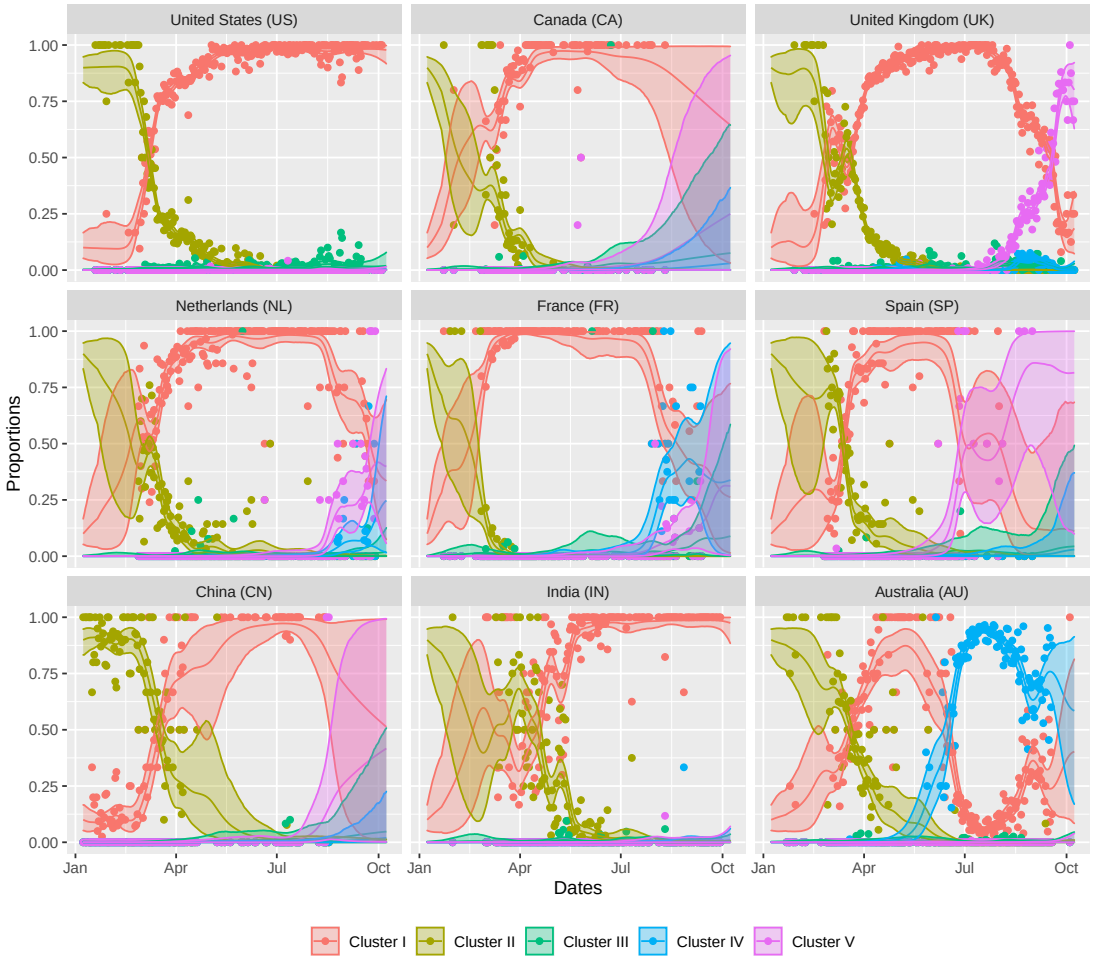


Figure 2.3: Posterior means and 95% credible intervals of inferred cluster proportions from the Bayesian hierarchical model for 9 countries. For each day, the points show the observed proportions, the middle solid lines indicate the posterior means of cluster proportions, and the bands indicate the 95% credible interval of cluster proportions.

Parameter	Mean	2.5%	97.5%	Parameter	Mean	2.5%	97.5%
α_2	-0.0495	-0.0658	-0.0344	$\Sigma_{\text{EU1,3}}$	0.0431	-0.0309	0.1313
α_3	0.0022	-0.0142	0.0178	$\Sigma_{\text{EU1,4}}$	-0.0133	-0.1055	0.0841
α_4	0.0172	-0.0073	0.0398	$\Sigma_{\text{EU2,1}}$	0.0486	-0.0353	0.1429
α_5	0.0267	0.0047	0.0478	$\Sigma_{\text{EU2,2}}$	0.1468	0.0850	0.2276
σ_{NA}	0.2432	0.1664	0.3820	$\Sigma_{\text{EU2,3}}$	0.0178	-0.0610	0.1048
σ_{EU}	0.3802	0.2915	0.4771	$\Sigma_{\text{EU2,4}}$	0.0320	-0.0926	0.1294
σ_{AS}	0.3578	0.1885	0.6661	$\Sigma_{\text{EU3,1}}$	0.0431	-0.0309	0.1313
σ_{AU}	0.2768	0.1989	0.3926	$\Sigma_{\text{EU3,2}}$	0.0178	-0.0610	0.1048
$p_{\cdot,1,1}$	0.0999	0.0524	0.1660	$\Sigma_{\text{EU3,3}}$	0.1468	0.0850	0.2276
$p_{\cdot,1,2}$	0.8993	0.8327	0.9473	$\Sigma_{\text{EU3,4}}$	0.0153	-0.0802	0.1114
$p_{\cdot,1,3}$	0.0008	0.0001	0.0030	$\Sigma_{\text{EU4,1}}$	-0.0133	-0.1055	0.0841
$p_{\cdot,1,4}$	0.0000	0.0000	0.0001	$\Sigma_{\text{EU4,2}}$	0.0320	-0.0926	0.1294
$p_{\cdot,1,5}$	0.0000	0.0000	0.0001	$\Sigma_{\text{EU4,3}}$	0.0153	-0.0802	0.1114
$\Sigma_{\text{NA1,1}}$	0.0618	0.0277	0.1459	$\Sigma_{\text{EU4,4}}$	0.1468	0.0850	0.2276
$\Sigma_{\text{NA1,2}}$	-0.0059	-0.0475	0.0344	$\Sigma_{\text{AS1,1}}$	0.1422	0.0355	0.4436
$\Sigma_{\text{NA2,1}}$	-0.0059	-0.0475	0.0344	$\Sigma_{\text{AS1,2}}$	0.0051	-0.1227	0.1803
$\Sigma_{\text{NA2,2}}$	0.0618	0.0277	0.1459	$\Sigma_{\text{AS2,1}}$	0.0051	-0.1227	0.1803
$\Sigma_{\text{EU1,1}}$	0.1468	0.0850	0.2276	$\Sigma_{\text{AS2,2}}$	0.1422	0.0355	0.4436
$\Sigma_{\text{EU1,2}}$	0.0486	-0.0353	0.1429				

Table 2.3: Posterior means and 95% credible intervals (represented by the 2.5% and 97.5% percentiles) for the parameters in the Bayesian hierarchical model. $\boldsymbol{\alpha}$ represents the daily growth rate parameters for clusters II to V relative to cluster I. σ_C is the random noise standard deviation for countries within continent C, and $\boldsymbol{p}_{\cdot,1}$ represents the initial proportions on Jan 7th, 2020 for all countries. Σ_C for each continent together forms the diagonal block matrix Σ that represents the covariance matrix for the random noise vector.

around 10% and cluster II accounts for around 90% of sequences. Nonetheless these initial proportions are quite uncertain due to the limited number of early cases, as seen in the wide credible intervals on the plot. The credible intervals narrow as we reach periods where a larger number of sequences are deposited. Overall, we see growth for cluster I accelerates during January to June but appears to rapidly fall off in July for many countries, while maintaining a substantive presence in the US, India and Australia. Cluster II clearly diminishes over time worldwide, and cluster III is fairly small but stable for all countries. Cluster IV estimates show small fluctuations for most countries except Australia, where it expands to over 90% from July to August and may remain as the dominant cluster. Cluster V is estimated to first appear in August and thereafter shows a rapid growth in all European countries.

Our Bayesian hierarchical model also allows prediction of changes in cluster proportions for countries with missing or very sparse GISAID sequence data. Canada, Spain and China only have deposited sequences up to August, while France only has sequence data deposited until mid-September; our model is nonetheless able to provide the point and interval estimates for their cluster proportions over the entire period. As suggested in Figure 2.3, both Canada and China are projected to have cluster I gradually decrease together with a possible rise in cluster V; the expanding credible intervals reflect the increasing uncertainty associated with the increasing number of days with missing data. Meanwhile, the main clusters present in France by mid-October might be IV or V (or their combination), while Spain is projected to be dominated by cluster V much like the rest of Europe.

To ensure that our results are robust, we performed a sensitivity analysis on the posterior parameters given different sets of priors, with a special focus on the sensitivity of $\boldsymbol{\alpha}$ and $\boldsymbol{p}_{i1}$. We created the following three scenarios to compare with our base scenario prior choices in Section 2.3, and the results are shown in Table 2.4. In scenario 1, we set the cluster proportions on the starting date as $\boldsymbol{p}_{i1} \sim Dirichlet(2.5, 38.5, 0.5, 0.5, 0.5)$, obtained by adding the pseudocounts to Dirichlet parameters corresponding to the Jeffreys prior, that is, $Dirichlet(0.5, 0.5, 0.5, 0.5, 0.5)$. In scenario 2, we set the priors for each growth parameter $\alpha_c \sim N(0, 1)$ independently for $c = 2, \ldots, 5$, instead of the uniform priors in our base scenario. In scenario 3, we combined both changes to the priors made in scenario 1 and 2. Compared with the main results in Table 2.3, all three scenarios in Table 2.4 show comparable posterior means and 95% credible intervals with the base scenario, which indicates our posterior parameters are fairly stable and robust to the different choices of priors.

	Scenario 1			Scenario 2			Scenario 3		
Param	Mean	2.5%	97.5%	Mean	2.5%	97.5%	Mean	2.5%	97.5%
α_2	-0.0492	-0.0655	-0.0351	-0.0494	-0.0656	-0.0345	-0.0503	-0.0664	-0.0352
α_3	0.0040	-0.0110	0.0188	0.0020	-0.0137	0.0173	0.0040	-0.0119	0.0195
α_4	0.0240	0.0007	0.0467	0.0177	-0.0069	0.0407	0.0242	-0.0000	0.0468
α_5	0.0341	0.0127	0.0552	0.0266	0.0050	0.0470	0.0338	0.0116	0.0548
$p_{\cdot,1,1}$	0.0839	0.0423	0.1418	0.1006	0.0528	0.1668	0.0850	0.0435	0.1447
$p_{\cdot,1,2}$	0.9157	0.8573	0.9575	0.8986	0.8319	0.9465	0.9145	0.8546	0.9563
$p_{\cdot,1,3}$	0.0004	0.0000	0.0016	0.0008	0.0001	0.0030	0.0004	0.0000	0.0017
$p_{\cdot,1,4}$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0000	0.0000	0.0000
$p_{\cdot,1,5}$	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0000	0.0000	0.0000

Table 2.4: Sensitivity analysis of the posterior parameters with different priors. Scenario 1 sets $\boldsymbol{p}_{i1} \sim Dirichlet(2.5, 38.5, 0.5, 0.5, 0.5)$. Scenario 2 sets the growth parameter $\alpha_c \sim N(0,1), c = 2, \ldots, 5$. Scenario 3 combines both changes to the priors from scenario 1 and 2.

2.5 Discussion

In this paper, we presented statistical approaches to tackle the challenges associated with the analysis of S-protein sequence data. To better understand the evolution of S-protein sequences, we grouped the S-protein sequences into hierarchical clusters, and studied the spatial and temporal trend of these mutant clusters using a Bayesian hierarchical model.

Based on our model estimates, we found that on average the reference sequence and its closely-related variants will diminish, while variants with the co-occurring mutations of D614G together with S477N or A222V tend to increase most strongly in prevalence over time. Our estimates of trend not only examined individual mutations as was analyzed in Korber et al. (2020), but also captured the prevalence of some co-occurring mutations that have so far received limited attention in the literature. Nonetheless, our findings on the reference sequence do align with Korber et al. (2020), where the authors showed that a transition of position 614 from D to G occurred in many regions around the world with varying levels of statistical significance. Our estimates of S477N and A222V are in agreement with the trends observed by the Los Alamos National Laboratory, and A222V in particular is also consistent with Hodcroft et al. (2020) where the authors reported its presence in the majority of sequences in Europe by the fall of 2020. In addition, we found spatial dependence in COVID-19 transmission across country boundaries to be low in general, but higher within Europe. This could be related to their relatively loose travel

policies within EU members during COVID-19 (European Commission, 2020). Finally, a useful feature of our Bayesian approach is the ability to make projections of cluster proportions in countries where data is scarce or missing.

The result of the sequence analysis suggests potential fitness advantages or higher infectiousness for the co-occurring mutations D614G + S477N and D614G + A222V. In reality, while higher infectiousness may fully explain their growth in prevalence, other epidemiological factors may also play a role, for example, the characteristics of the infected population and the founder effect (Korber et al., 2020). Although Li et al. (2020) confirmed that D614G combined with other mutations (e.g., L5F, V341I, K458R, etc.) are more infectious than the reference sequence, the infectivities of D614G + S477N or D614G + A222V have not yet been mentioned and examined. Therefore, further experimental evidence is needed to confirm the increased infectivity of these co-occurrent mutations.

Overall, S-protein sequence and structure datasets are a rich source of information that further research efforts can leverage for better understanding COVID-19, and we list some examples. First, the sequence data could be expanded to include other data sources; the sequences used in this paper were collected from GISAID only where the sequence deposition rates from some countries is low. Second, the sequence data could be combined with data on COVID-19 testing and case counts to estimate the actual prevalence of mutant clusters in different countries, in addition to their proportions. Third, many laboratories have separately contributed S-protein structures to the PDB and these could be further analyzed to quantify uncertainties associated with structure determination efforts.

Chapter 3

Spatio-temporal data fusion for the analysis of in situ and remote sensing data using the INLA-SPDE approach

3.1 Introduction

Spatial misalignment has become a challenge along with the technological advances in the collection and storage of spatio-temporal data, such as the use of GPS, satellites, etc. While convenient for undertaking large-scale spatio-temporal analyses, it can be difficult to integrate or compare data derived from different sources, when each has its own spatial and temporal reference frames. These data might be obtained from a network of monitoring stations, satellites, or numerical outputs of computer simulations. While data collected from monitoring networks give measurements at points in space, remote sensing observations and outputs from computer simulations provide predictions at the level of grid cells (Bruno et al., 2016). In spatial statistics, the challenge brought by data measured at different scales is commonly known as the change of support problem (COSP) (Cressie, 2015; Banerjee et al., 2003; Gotway and Young, 2002). Data from monitoring networks are classified as point-referenced or geostatistical data, and often tend to be sparsely located in both spatial and temporal dimensions. When available, these data are generally considered to accurately represent the true conditions or phenomena being measured (Berrocal et al., 2010). On the other hand, measurements from satellites provide high-resolution data on grid cells at regular time intervals, are readily accessible, and it is desirable to leverage such massive data for analyses. However, care must be taken to ensure seamless compati-

bility across different data sources and to ensure the reliability of analyses. The associated methodological challenges are particularly significant in fields that include epidemiology, urban planning, and environmental science.

In the realm of environmental science, the integration of data from sources such as satellites, monitoring stations, and field measurements is pivotal for the understanding of ecological changes and climatic trends. In the spatio-temporal modeling of harmful algae bloom (HAB) dynamics, analyses tend to rely heavily on either in situ samples or satellite images in isolation. Within this field, common practices involve constructing an algae bloom index followed by categorization as "in situ" or "remote sensing" (Fang et al., 2019; Manning et al., 2019; Stumpf et al., 2016b). However, these practices have contributed to disparities in the characterization of bloom dynamics, leading to divergent findings regarding the main drivers of HABs (Ho and Michalak, 2015). Indeed, integrating different types of observations was recognized as a key research area to progress towards a better predictive knowledge of HAB dynamics (Bullerjahn et al., 2016). Spatial misalignment also occurs in epidemiology studies for understanding the relationship between environmental factors and human health (Fuentes et al., 2006; Zhu et al., 2003), as well as environmental prediction and model calibration (Lawson, 2012). Thus, to ensure reliable and accurate analyses in such application areas, spatial misalignment needs to be addressed in a modeling framework.

Different methods have been proposed to address the spatial misalignment problem, including block kriging (Banerjee et al., 2003; Cressie, 2015; Gotway and Young, 2007), hierarchical downscaler (Berrocal et al., 2010), and Bayesian melding (Fuentes and Raftery, 2005). The utilization of Bayesian models, particularly those developed through latent point-level processes has garnered significant attention, as demonstrated by the work of Fuentes and Raftery (2005). This approach assumes that there exists an underlying unobserved spatial process driving both the observational data and the numerical model output, and can be implemented via Markov chain Monte Carlo (MCMC) methods. However, this type of state space framework requires stochastic integration of the latent spatial process over the area of interest, which is a computationally demanding task. To mitigate the computational bottleneck, fast and approximate methods, including the integrated nested Laplace approximation (INLA) (Rue et al., 2009) and stochastic partial differential equation (SPDE) (Lindgren et al., 2011) approaches, have been proposed. To address COSP, Moraga et al. (2017) proposed a method for the combined analysis of point and areal data utilizing the INLA-SPDE framework.

Spatio-temporal data fusion is an important research area in environmental science; however, it is much more challenging to tackle due to the higher dimensionality introduced by time. Following Fuentes and Raftery (2005)'s Bayesian melding approach, a spatio-

temporal fusion model has been proposed to study the association between mortality and pollution exposure to daily fine particulate matter, and addresses the COSP between point-referenced data and model output (Choi et al., 2009). Further, McMillan et al. (2010) presents a spatio-temporal fusion model at the grid-cell level by assuming that the data process at the point level is linked to the same underlying true process at the grid-cell level. These models rely on MCMC methods for inference, so do not scale well to much higher dimensions due to the computational requirements of MCMC. While the INLA-SPDE-based approach has been proposed to model spatio-temporal processes (Blangiardo and Cameletti, 2015; Cameletti et al., 2013), its application to spatio-temporal data fusion is limited. Villejo et al. (2023) proposed a two-stage estimation approach to address spatial misalignment in linking pollutant exposures and health outcomes; however, the COSP was not specifically addressed in their first stage model for the pollutants. In their work, satellite data were not treated as block data; instead, they used the centroid of the image pixels as point data. Thus, to the best of our knowledge, there has been no previously proposed spatio-temporal data fusion approach that explicitly addresses the COSP through the INLA-SPDE framework.

The motivation of this paper is to address the COSP between different data sources, with application to better quantifying the disparities in characterizing the dynamics of HABs. To achieve this goal, we propose a Bayesian hierarchical model to analyze spatially misaligned data with a latent spatio-temporal dynamic process. The model considers the COSP and coherently combines data with different spatio-temporal resolutions and potential missingness. We show that via the construction of appropriate projection matrices, the different types of data can be spatially aligned and fitted using the INLA-SPDE approach in a computationally efficient manner. A real data analysis of HABs is presented, and the results illustrate the necessity of the fusion model for accurately characterizing their spatio-temporal dynamics.

The structure of the paper is presented as follows. In Section 3.2, we describe our proposed spatio-temporal fusion model to address the spatially misaligned data. Section 3.3 discusses model inference and the application of INLA and the SPDE approach for fitting the proposed model. Section 3.4 provides a simulation study inspired by the cyanobacteria HAB data to evaluate model performance. In Section 3.5, we present a real-world example analyzing cyanobacteria HAB data over Western Lake Erie. Finally, Section 3.6 concludes the paper and outlines directions for future work.

3.2 Method

3.2.1 Spatial data fusion model

We first describe the spatial data fusion framework known as the Bayesian melding model; such a model structure can be found in Fuentes and Raftery (2005). The model is designed to integrate both point and areal-level data by assuming a common spatial random field $\{y(\boldsymbol{s}) : \boldsymbol{s} \in D \subset \mathbb{R}^2\}$ underlying all observations on the spatial domain of interest D. This spatial process is, in part, explained by large-scale explanatory variables $\boldsymbol{x}(\boldsymbol{s}) = (\boldsymbol{x}_0(\boldsymbol{s}), \ldots, \boldsymbol{x}_p(\boldsymbol{s}))^T$ (i.e., a vector of p covariates and an intercept), with $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^T \in \mathbb{R}^{p+1}$ being the vector of coefficients:

$$y(\boldsymbol{s}) = \boldsymbol{x}(\boldsymbol{s})\boldsymbol{\beta} + \xi(\boldsymbol{s}), \tag{3.1}$$

where the residual component $\xi(\boldsymbol{s})$ is modeled as a Gaussian random field with mean $E(\xi(\boldsymbol{s})) = 0$ and a stationary and isotropic Matérn covariance function.

On the observation level, denote the remote sensing observations as $z_1(\boldsymbol{B}_j)$ for block area $\boldsymbol{B}_j \in D \subset \mathbb{R}^2, j = 1, \ldots, J$, and denote the in situ observations as $z_2(\boldsymbol{s}_i)$ for $\boldsymbol{s}_i \in D \subset \mathbb{R}^2, i = 1, \ldots, I$. More generally, $z_1(\boldsymbol{B}_j)$ can be extended to model any areal data, and $z_2(\boldsymbol{s}_i)$ can be extended to model any point-referenced data. The two sets of observations are assumed to be generated with the following framework:

$$z_1(\boldsymbol{B}_j) = a(\boldsymbol{B}_j) + \frac{b}{|\boldsymbol{B}_j|}\int_{\boldsymbol{B}_j} y(\boldsymbol{s})d\boldsymbol{s} + e_1(\boldsymbol{B}_j), \tag{3.2}$$

$$z_2(\boldsymbol{s}_i) = y(\boldsymbol{s}_i) + e_2(\boldsymbol{s}_i), \tag{3.3}$$

where $e_1(\boldsymbol{B}_j) \sim N(0, \tau_1)$ and $e_2(\boldsymbol{s}_i) \sim N(0, \tau_2)$ are all independent. The parameter $a(\boldsymbol{B}_j)$ is the additive bias for the satellite image in area $\boldsymbol{B}_j$ and b is the multiplicative bias across space. $|\boldsymbol{B}_j|$ represents the area of block j and hence $\frac{1}{|\boldsymbol{B}_j|}\int_{\boldsymbol{B}_j} y(\boldsymbol{s})d\boldsymbol{s}$ represents the block average of the spatial random process within block j. Since satellite data tend to be more prone to noise and less associated with the true latent field compared to in situ data collected directly from ground monitoring stations, the introduced bias parameters allow the model to properly calibrate for such a discrepancy when present. Additive bias parameters are assumed to be either constant or spatially varying, although they are typically treated as constant $a(\boldsymbol{s}) \equiv a$ across space to avoid identifiability issues. Note that this model set up allows for addressing the COSP, as satellite data are represented as block averages. Moraga et al. (2017) in their paper proposed a simplification of the original melding model that assumes no multiplicative bias (i.e., by setting $b = 1$) and reformulate

the above structure using $\boldsymbol{\mu}(\boldsymbol{s}) = \boldsymbol{x}(\boldsymbol{s})\boldsymbol{\beta}$, such that $z_2(\boldsymbol{s}_i) = \mu(\boldsymbol{s}_i) + \boldsymbol{\xi}(\boldsymbol{s}_i) + e_2(\boldsymbol{s}_i)$, and $z_1(\boldsymbol{B}_j) = \frac{1}{|\boldsymbol{B}_j|}\int_{\boldsymbol{B}_j}(a + \mu(\boldsymbol{s}) + \boldsymbol{\xi}(\boldsymbol{s}))d\boldsymbol{s} + e_1(\boldsymbol{B}_j) = \frac{1}{|\boldsymbol{B}_j|}\int_{\boldsymbol{B}_j}(\mu^*(\boldsymbol{s}) + \boldsymbol{\xi}(\boldsymbol{s}))d\boldsymbol{s} + e_1(\boldsymbol{B}_j)$ where $\mu^*(\boldsymbol{s}) = a + \mu(\boldsymbol{s})$.

3.2.2 Spatio-temporal data fusion model

We propose an extension of the spatial data fusion framework to the spatio-temporal domain. The spatio-temporal extension to the data fusion framework is designed to borrow information and conduct inference across both space and time. The model configuration is explained as follows for completely-observed data. The remote sensing observations $z_1(\boldsymbol{B}_j, t)$ for each block area $\boldsymbol{B}_j \in D \subset \mathbb{R}^2, j = 1, \ldots, J$, at discrete time $t, t = 1, \ldots, T$, are modeled as

$$z_1(\boldsymbol{B}_j, t) = a + \frac{1}{|\boldsymbol{B}_j|}\int_{\boldsymbol{B}_j} y(\boldsymbol{s}, t)d\boldsymbol{s} + e_1(\boldsymbol{B}_j, t), \tag{3.4}$$

where a is the bias term assumed to be constant across space and time, $y(\boldsymbol{s}, t)$ is the hidden spatio-temporal process that varies continuously over space and time, and $e_1(\boldsymbol{B}_j, t) \sim N(0, \tau_1)$ is the measurement error for remote sensing. A constant additive bias simplifies the processing and analysis and can serve as a baseline correction for systematic errors in satellite images, e.g., where in situ data are sparse and/or the environmental conditions are relatively homogeneous. We further assume that there is no multiplicative bias, as in Moraga et al. (2017). On the other hand, the in situ data $z_2(\boldsymbol{s}, t)$ are modeled as

$$z_2(\boldsymbol{s}_i, t) = y(\boldsymbol{s}_i, t) + e_2(\boldsymbol{s}_i, t), \tag{3.5}$$

for $\boldsymbol{s}_i \in D \subset \mathbb{R}^2, i = 1, \ldots, I, t = 1, \ldots, T$, where $e_2(\boldsymbol{s}_i, t) \sim N(0, \tau_2)$ is the measurement error for the in situ data.

Similar to the spatial setup, the spatio-temporal process is assumed to be a Gaussian random field underlying both sets of observations. The hidden process is inspired by models developed in air quality literature for modeling pollutants (Cameletti et al., 2011, 2013; Cocchi et al., 2007), which exhibits flexibility in capturing various factors influencing spatio-temporal dynamics. The hidden spatio-temporal process $y(\boldsymbol{s}, t)$ is assigned the following structure:

$$y(\boldsymbol{s}, t) = \boldsymbol{x}(\boldsymbol{s}, t)\boldsymbol{\beta} + \xi(\boldsymbol{s}, t), \tag{3.6}$$

where $\boldsymbol{x}(\boldsymbol{s}, t)$ is the set of large-scale spatially and temporally varying covariates, such as meteorological and geographical variables, although some of these may only vary on the spatial domain and others on the temporal domain. The latent process $\xi(\boldsymbol{s}, t)$ accounts for

spatial and temporal dependencies and is assigned an AR(1) (i.e., first-order autoregressive) Gaussian process:

$$\xi(\boldsymbol{s}, t) = \rho\xi(\boldsymbol{s}, t-1) + \omega(\boldsymbol{s}, t), \tag{3.7}$$

for $t = 2, \ldots, T$, which is stationary with $|\rho| < 1$, and $\xi(\boldsymbol{s}, 1) \sim N(0, \sigma_\omega^2/(1-\rho^2))$. Further, $\omega(\boldsymbol{s}, t)$ is a spatially correlated innovation term, following a Gaussian distribution with a Matérn covariance structure that is independent of time:

$$Cov(\omega(\boldsymbol{s}_i, t), \omega(\boldsymbol{s}_j, t')) = \begin{cases} 0, & t \neq t', \\ \frac{\sigma_\omega^2}{\Gamma(\nu)2^{\nu-1}}(\kappa||\boldsymbol{s}_i - \boldsymbol{s}_j||)^\nu K_\nu(\kappa||\boldsymbol{s}_i - \boldsymbol{s}_j||), & t = t'. \end{cases} \tag{3.8}$$

The above specification assumes space-time separability and indicates when $t \neq t'$, the innovation term has zero covariance, i.e., $\omega(\cdot, \cdot)$ does not contribute to the spatial dependence between locations for measurements taken at different times. When $t = t'$, the covariance has a Matérn structure that depends on $||\boldsymbol{s}_i - \boldsymbol{s}_j||$, i.e., the Euclidean distance between two generic locations $\boldsymbol{s}_i, \boldsymbol{s}_j \in \mathbb{R}^2$. The covariance function assumes the Gaussian field is second-order stationary and isotropic between locations. The term K_ν denotes the modified Bessel function of the second kind with order $\nu > 0$, which is usually kept fixed and measures the degree of the smoothness of the process. The parameter σ_ω^2 is the marginal variance, and $\kappa > 0$ is a scaling parameter associated with the range r, i.e. the distance at which the spatial correlation becomes very small. Typically, the empirically derived definition for the range is $r = \sqrt{8\nu}/\kappa$, with r corresponding to the distance at which the spatial correlation is close to 0.1, for each $\nu \geq 0.5$.

3.3 Inference

3.3.1 INLA-SPDE

The INLA method, proposed by Rue et al. (2009), provides a computationally efficient approach for estimating posterior distributions in a broad class of latent Gaussian models. Unlike MCMC, INLA employs a deterministic approximation technique, including Laplacian approximations and numerical integration techniques, that allows for fast approximate Bayesian inference. INLA-SPDE extends the capabilities of INLA by drawing the link between the Gaussian field (GF) and Gaussian Markov random field (GMRF), and has been shown to be effective for handling complex spatial processes (Lindgren et al., 2011).

In equation (3.7), the latent process $\omega(\boldsymbol{s}, t)$ with a Matérn covariance of form (3.8) is a solution to a certain SPDE; this SPDE can be represented as a discretely indexed

GMRF by means of a finite basis function defined on a triangulation of the region of study. The finite element representation of the SPDE establishes a link between the GF and the GMRF characterized by the Gaussian weights with Markov dependencies. The SPDE that can be used to construct the GMRF representation is

$$(\kappa^2 - \Delta)^{\alpha/2}(\tau_\omega \omega(\boldsymbol{s}, t)) = \mathcal{W}(\boldsymbol{s}, t), \quad \alpha = \nu + d/2, \quad \kappa > 0, \quad \nu > 0, \tag{3.9}$$

where $d = 2$ for $\boldsymbol{s} \in \mathbb{R}^2$, $(\kappa^2 - \Delta)^{\alpha/2}$ is a pseudo-differential operator, and $\Delta = \sum_{i=1}^{d} \frac{\partial^2}{\partial \xi_i^2}$ is the Laplacian operator. The parameter $\kappa > 0$ controls the scale, and τ_ω controls the variance of the GF respectively. The parameter α is the order of the differential operator and controls the smoothness of the GF. The marginal variance σ_ω^2 of the GF can be derived as $\sigma_\omega^2 = 1/(4\pi)\kappa^2\tau_\omega^2$. In addition, the innovation term $\mathcal{W}(\boldsymbol{s}, t)$ is the Gaussian spatial white noise process with unit variance.

The solution to the SPDE in (3.9) is a finite element representation with piecewise linear basis functions defined on a triangulation of D, which is given by

$$\omega(\boldsymbol{s}, t) = \sum_{g=1}^{G} \varphi_g(\boldsymbol{s}) \tilde{\omega}_{gt}, \tag{3.10}$$

where G is the total number of vertices of the triangulation, $\{\varphi_g(\cdot)\}_{g=1}^{G}$ is the set of basis functions, and $\{\tilde{\omega}_{gt}\}_{g=1,\ldots,G,t=1,\ldots,T}$ are zero-mean Gaussian distributed weights. The weights $\tilde{\omega}_{gt}$ are updated over time in tandem with the temporal evolution of the GMRF, however, the set of basis functions remains constant over time, as the mesh associated with the triangulation remains fixed.

Consequently, the Gaussian weights $\tilde{\boldsymbol{\omega}}_t := \boldsymbol{\omega}_{\cdot t}$ for each t is a GMRF with distribution $N(0, \boldsymbol{Q}_S^{-1})$ and represents the approximated solution to the SPDE. Denote $\tilde{\boldsymbol{\omega}} := (\tilde{\boldsymbol{\omega}}_1, \ldots, \tilde{\boldsymbol{\omega}}_T)$, then given the modern formulation of the INLA methodology (Van Niekerk et al., 2023), the latent GF of the model is $\boldsymbol{\mathcal{X}} := (\tilde{\boldsymbol{\omega}}, a, \boldsymbol{\beta})$ and the hyper-parameters are $\boldsymbol{\Theta} := (\kappa, \tau_\omega, \rho, \tau_1, \tau_2)$. Denote $\boldsymbol{\omega}_t := \boldsymbol{\omega}(\cdot, t)$, $\boldsymbol{\xi}_t := \boldsymbol{\xi}(\cdot, t)$ and $\boldsymbol{\xi} := (\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_T)^T$, then (3.7) can be rewritten as $\boldsymbol{\xi}_t = \rho\boldsymbol{\xi}_{t-1} + \boldsymbol{\omega}_t$. Denote $\boldsymbol{y}_t := \boldsymbol{y}(\cdot, t)$ and $\boldsymbol{x}_t := \boldsymbol{x}(\cdot, t)$, then (3.6) can be rewritten as $\boldsymbol{y}_t = \boldsymbol{x}_t\boldsymbol{\beta} + \boldsymbol{B}\boldsymbol{\xi}_t$, where $\boldsymbol{B}$ is a sparse matrix that connects $\boldsymbol{\xi}_t$ to each observational vector $\boldsymbol{y}_t$, and $\boldsymbol{\xi}_t = \boldsymbol{M}_t\tilde{\boldsymbol{\omega}}$ where $\boldsymbol{M}_t$ is a sparse matrix that connects the Gaussian weights $\tilde{\boldsymbol{\omega}}$ to $\boldsymbol{\xi}_t$. Denote $\boldsymbol{z}_{1t} := \boldsymbol{z}_1(\cdot, t)$ and $\boldsymbol{z}_{2t} := \boldsymbol{z}_2(\cdot, t)$, then the joint posterior of $\boldsymbol{\mathcal{X}}$ and $\boldsymbol{\Theta}$, given complete observations $\boldsymbol{z}$, can be written as

$$p(\boldsymbol{\mathcal{X}}, \boldsymbol{\Theta}|\boldsymbol{z}) = p(\boldsymbol{\Theta})p(\boldsymbol{\mathcal{X}}|\boldsymbol{\Theta}) \prod_{t=1}^{T} p(\boldsymbol{z}_{1t}|\boldsymbol{\mathcal{X}}, \boldsymbol{\Theta}) \prod_{t=1}^{T} p(\boldsymbol{z}_{2t}|\boldsymbol{\mathcal{X}}, \boldsymbol{\Theta}), \tag{3.11}$$

where $\boldsymbol{z}_{1t}|\boldsymbol{\mathcal{X}}, \boldsymbol{\Theta} \sim N(a+\frac{1}{|\boldsymbol{B}_j|}\int_{\boldsymbol{B}_j}(\boldsymbol{x}_t\boldsymbol{\beta}+\boldsymbol{B}\boldsymbol{M}_t\tilde{\boldsymbol{\omega}})d\boldsymbol{s}, \tau_1^{-1}\boldsymbol{I})$, $\boldsymbol{z}_{2t}|\boldsymbol{\mathcal{X}}, \boldsymbol{\Theta} \sim N(\boldsymbol{x}_t\boldsymbol{\beta}+\boldsymbol{B}\boldsymbol{M}_t\tilde{\boldsymbol{\omega}}, \tau_2^{-1}\boldsymbol{I})$, and $p(\boldsymbol{\Theta})$ is the prior distribution of the hyper-parameters. Finally, given this construction of $p(\boldsymbol{\mathcal{X}}, \boldsymbol{\Theta}|\boldsymbol{z})$, the marginal posterior distributions of $p(\boldsymbol{\mathcal{X}}|\boldsymbol{z})$ and $p(\boldsymbol{\Theta}|\boldsymbol{z})$ can be approximated through INLA. Given the Gaussian weights $\tilde{\boldsymbol{\omega}}$ and the hyper-parameters $\boldsymbol{\Theta}$, the distribution of $\boldsymbol{\xi}$ can be reconstructed with a sparse design matrix that links $\boldsymbol{\xi}$ to the latent field $\boldsymbol{\mathcal{X}}$.

3.3.2 Missing values in response

In the context of satellite observations, missing values in the response (i.e., unobserved blocks) are prevalent within the daily satellite images due to factors such as the presence of cloud cover. The extent of missing pixels varies each day and there is no specific pattern of missingness, as illustrated by the real data in Figure 3.9. We assume a missing at random (MAR) missingness mechanism, i.e., the probability of a missing satellite pixel is plausibly independent of HAB levels given the observed covariates (e.g., weather and climate variables). Denote the complete satellite data as $\boldsymbol{z}_2 = (\boldsymbol{z}_{2,obs}, \boldsymbol{z}_{2,mis})$, where $\boldsymbol{z}_{2,obs}$ and $\boldsymbol{z}_{2,mis}$ are the observed and missing components of $\boldsymbol{z}_2$, respectively. Since the MAR missingness mechanism depends on factors apart from the response itself, we may treat the missingness mechanism of the satellite data as ignorable, i.e., it does not affect the specification of the posterior distribution (Little and Rubin, 2019). Specifically, the term $p(\boldsymbol{z}_{2t}|\boldsymbol{\mathcal{X}}, \boldsymbol{\Theta})$ in the posterior distribution (3.11) may have missing responses for a different subset of blocks at each time t. INLA handles these missing values by computing their predictive distribution during inference (Gómez-Rubio, 2020), which can be useful for a post-hoc analysis; in this context, the predictive distribution of an individual missing response z_{mis} in the satellite data is

$$p(z_{mis}|\boldsymbol{z}_1, \boldsymbol{z}_{2,obs}) = \int p(z_{mis}, \boldsymbol{\Theta}|\boldsymbol{z}_1, \boldsymbol{z}_{2,obs})d\boldsymbol{\Theta} = \int p(z_{mis}|\boldsymbol{\Theta}, \boldsymbol{z}_1, \boldsymbol{z}_{2,obs})p(\boldsymbol{\Theta}|\boldsymbol{z}_1, \boldsymbol{z}_{2,obs})d\boldsymbol{\Theta}. \tag{3.12}$$

3.3.3 Approximation of integrals

In the spatial fusion method using INLA-SPDE (Moraga et al., 2017), the integral of the spatial process in the areal data is approximated through:

$$\int_{B_j} y(\boldsymbol{s})d\boldsymbol{s} \approx \sum_{g=1}^{G} A_{jg}^{a}\varphi_g = \boldsymbol{A}_{\boldsymbol{j}\cdot}^{\boldsymbol{a}}\boldsymbol{\varphi}, \tag{3.13}$$

where G is the number of vertices in the triangulation, $\boldsymbol{\varphi}$ is the vector of finite element basis functions, and $\boldsymbol{A^a} \in \mathbb{R}^{J\times G}$ is a projection matrix that maps the GMRF from the J areal observation locations to the G triangulation vertices. Moraga et al. (2017) suggests an algorithm to construct the weight matrix $\boldsymbol{A^a}$ with each area represented as an average of the vertices inside the area.

We extend such an approximation to the spatio-temporal integration, with the integral of the areal data being approximated as:

$$\int_{B_j} y(\boldsymbol{s},t)d\boldsymbol{s} \approx \sum_{g=1}^{G}[\boldsymbol{A_t^a}]_{jg}\varphi_g = [\boldsymbol{A_t^a}]_{\boldsymbol{j}\cdot}\boldsymbol{\varphi}, \tag{3.14}$$

where for each $t, t = 1,\ldots,T$, there exists a projection matrix $\boldsymbol{A_t^a} \in \mathbb{R}^{J\times G}$ that maps from the spatio-temporal GMRF to the areal observations. Each weight matrix $\boldsymbol{A_t^a}$ is a sparse matrix with row-wise entries summing to 1. For vertex g outside the area j, the entry $[\boldsymbol{A_t^a}]_{jg} = 0$. For vertices $g_1,\ldots,g_m$ inside the area j, each of them is assigned an equal weight of $1/m$ for mapping from the areal region to each triangulation vertex, i.e., $[\boldsymbol{A_t^a}]_{g_m,j} = 1/m$. Since the triangulation vertices at time t are only relevant for mapping GMRF at time t, $\boldsymbol{A^a}$ for the spatio-temporal integral approximation is a sparse block diagonal matrix of the form

$$\boldsymbol{A^a} = \begin{bmatrix} \boldsymbol{A_1^a} & 0 & \cdots & 0 \\ 0 & \boldsymbol{A_2^a} & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \boldsymbol{A_T^a} \end{bmatrix}. \tag{3.15}$$

The construction of this projection matrix enables the fitting of the fusion model using the INLA-SPDE approach, thereby bypassing the need for expensive MCMC computations. Note that the mesh size plays a crucial role in defining the precision of the spatial field. A finer mesh provides a more accurate representation of the spatial domain but comes at the cost of increased computational complexity. On the other hand, a coarser mesh reduces computational demands but may sacrifice accuracy in representing spatial features. The practical effects of mesh size are explored in the subsequent simulation study.

We use non-informative/weakly informative priors for both fixed effects and the hyper-parameters. The priors for fixed effects a and $\boldsymbol{\beta}$ are chosen to be normal with mean zero and large variance. The prior for the precision hyper-parameters are $\log\tau_1, \log\tau_2 \sim \log\Gamma(0.01, 0.01)$, and the prior for ρ is $\log(\frac{1+\rho}{1-\rho}) \sim N(0, 0.15)$. The SPDE hyper-parameters are represented as $\log(\tau_\omega) = \theta_1$, and $\log(\kappa) = \theta_2$ in INLA, with $\theta_1, \theta_2 \sim N(0,1)$.

3.4 Simulation Study

3.4.1 Simulation Scenarios

This section presents a simulation study designed to evaluate the performance of our methodology. The study serves two primary objectives: (1) to evaluate the accuracy of the parameter estimates, and (2) to assess the performance of predictions. It is crucial to ensure that the proposed model performs well not only in making accurate predictions on unobserved data, but also in parameter estimation. The simulation is designed to mimic the actual concentration levels of chlorophyll-a (chl-a), which are a key indicator of cyanobacteria abundance. In practice, measurements of chl-a may be directly collected at specific points near monitoring stations, or alternatively acquired from satellite pixels over larger geographical areas. We use Western Lake Erie as our study region, as this area is also the focus of our subsequent real data application. Western Lake Erie is the western basin of Lake Erie, part of the North American Great Lakes system, covering approximately 4,000 km^2. This region is known for being prone to algal blooms, making it an important area for monitoring and environmental studies. Other details of Western Lake Erie can be found in section 3.5.

Simulation of the spatio-temporal GMRF

First, we create an SPDE mesh on the spatial domain with a fine inner extension of 0.04 and outer extension of 0.2 to ensure the precision of the resulting field. Following the mesh creation, we simulate a zero-mean Gaussian Markov Random Field $\boldsymbol{\xi_1}$ on the SPDE mesh. The spatial correlations within this field are governed by the Matérn covariance function, with parameters set to $\kappa = 7$ and $\sigma_\omega^2 = 0.25$. In addition, we adopt the default $\alpha = 2$ as the order of the differential operator in the SPDE. The configuration dictates a moderate level of smoothness and variability (with spatial range $\sim$0.40 in degrees or 33.85–44.86 km) inherent in the spatial structure of the Gaussian field.

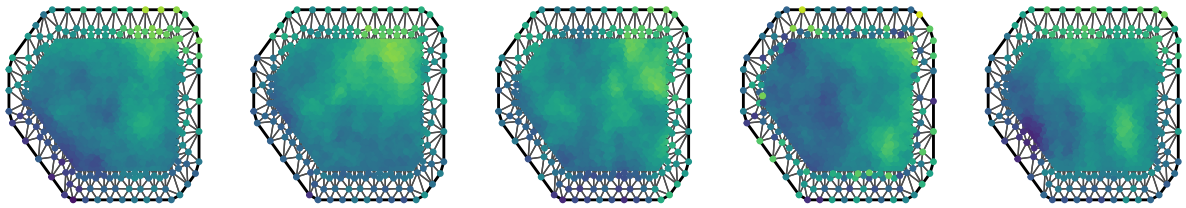
We opt for a simulation duration of $T = 19$ days, comprised of 14 days for training and 5 days for testing. This selection is made to strike a balance by ensuring a period of sufficient length for model training while maintaining a fast computation speed. Based on the selected T, we generate a realization of the spatio-temporal process sequentially according to (3.7), with $\rho = 0.7$. Note that $\boldsymbol{\omega}_t = \sqrt{1-\rho}\boldsymbol{\xi_t}$ is also a GMRF with mean zero and variance σ_ω^2, and this ensures the spatial process is stationary (Rue and Held, 2005). Moreover, to introduce a geographical component, we include the demeaned longitude and latitude as covariates; their regression coefficients are then set as $\beta_1 = -1$ and $\beta_2 = -1$,

respectively. Consequently, the mean of the resulting simulated spatio-temporal process reflects the actual spatial trends characterized by an increase in chl-a concentration along the South and West of the basin. Figure 3.1a shows a realization of the simulated spatio-temporal GMRF on day 1, 5, 10, 15 and 19.

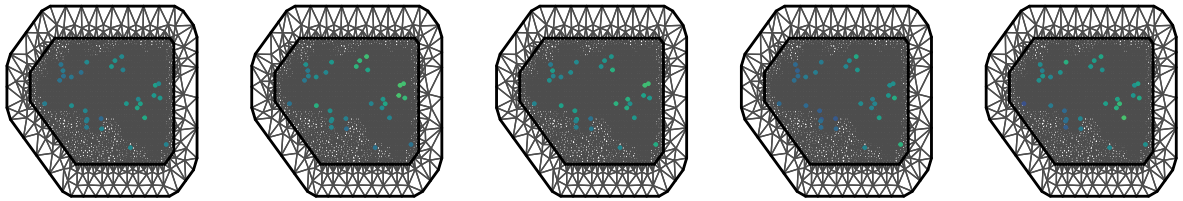
Simulation of the in situ and satellite data

We explore two settings for the sampling of in situ data: 5 vs. 30 sampling locations. The first scenario represents an extremely sparse sampling strategy, inspired by the sparsity of our real data application. In the second scenario, while the sampling strategy is less sparse, it is designed to represent the upper limit of samples that might be collected in the field each day. This variation allows us to assess the impact of sampling intensity on the representation of the latent process at observed locations. The sampling locations could be treated as different or fixed over the T days, and we adopt the latter as it is more likely to mimic the data generating mechanism of in situ samples. Once the locations are sampled, we interpolate the mesh nodes to align with the observation locations and subsequently generate in situ data according to (3.5), with $\tau_2 = 20$ to introduce some small variability via the random noise term. Figure 3.1b shows an example of the simulated in situ data with 30 sampling locations, and Figure 3.10 shows an example with 5 sampling locations.

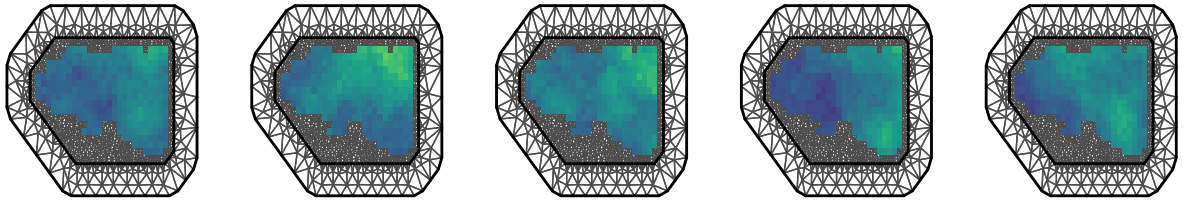
The satellite data are generated based on the weighted average of the vertices inside each rectangular grid, which have a resolution of 4 km $\times$ 4 km to mimic the real data. We construct the mapping matrix $\boldsymbol{A}^{\boldsymbol{a}}$ outlined in (3.15) and generate satellite data according to (3.4), with $\tau_1 = 50$ and a constant bias $a = 0.5$ across space. τ_1 is chosen to be larger than τ_2, as suggested by the real data application. Figure 3.1c shows an example of simulated satellite data, assuming all pixels are fully observed. It is important to note that a common challenge in satellite imagery is the presence of missing data, often caused by cloud cover. Consequently, we consider two settings for the simulation of satellite data with a different level of missingness per day: 50% and 80%. The locations of the missing satellite pixels are randomly sampled every day, resulting in potentially different missing areas for different days to mimic the impact of cloud cover in the real data.



(a) The latent spatio-temporal process $y(\boldsymbol{s}, t)$ on days 1, 5, 10, 15, and 19.



(b) In situ samples on days 1, 5, 10, 15, and 19 (30 each day). The locations were selected randomly over the spatial domain.



(c) Satellite images with full observations on days 1, 5, 10, 15, and 19. Each pixel represents the average log chl-a level inside the block area.

Figure 3.1: A simulated realization of log chl-a concentration in Western Lake Erie on days 1, 5, 10, 15, and 19. The black lines represent the mesh triangulation. The colors superimposed on the mesh indicate the relative levels of log chl-a concentration.

In addition, we consider three settings for the mesh used in model fitting, obtained by varying the lengths of the inner and outer extensions[1]. This variation is designed to assess the effect of mesh size on the parameter estimation and final prediction results. Since we are addressing COSP, the effect of mesh can be important as it affects how the weights are assigned among the triangulation vertices for approximating the integral of the areal data; e.g., the mesh size required for accurate estimation may depend on the smoothness of the field. Table 3.1 summarizes the 12 simulation scenarios obtained as combinations of the three factors described above.

Scenario	No. in situ samples	% missingness	Mesh maximum inner length
1	5	50%	0.15
2	30	50%	0.15
3	5	80%	0.15
4	30	80%	0.15
5	5	50%	0.1
6	30	50%	0.1
7	5	80%	0.1
8	30	80%	0.1
9	5	50%	0.05
10	30	50%	0.05
11	5	80%	0.05
12	30	80%	0.05

Table 3.1: Summary of the 12 simulation scenarios. The scenarios vary by three changing factors: the number of in situ samples collected each day, the percentage of missingness in satellite images, and the maximum inner length of the triangulation.

3.4.2 Evaluation Metric

The models were assessed on both parameter estimates and the predictive performance at unobserved locations. To evaluate the estimation performance for the model parameters,

[1]We kept other mesh parameters fixed, including the cut-off distance, minimum triangle angle, and boundary offset. These settings can affect model smoothness, numerical accuracy, and computational cost. Our decision to retain the default values reflects computational and tuning constraints rather than an assumption that they are universally optimal. A more thorough sensitivity analysis could vary these parameters alongside the inner and outer edge lengths to assess the robustness of parameter estimates and predictive accuracy.

we measure the relative bias and the root mean squared error (RMSE) from the estimated posterior distribution of each parameter $\theta \in \{a, \boldsymbol{\beta}, \Theta\}$. We run n_{sim} replications of the simulation, and denote each run (i.e., an independent realization of the spatio-temporal process) with the index j, where $j = 1 \ldots, n_{sim}$. For simulation run j, we denote each sample of θ as $\hat{\theta}_{jk}$, where $k = 1, \ldots, n_{samp}$. Each sample is drawn from the corresponding posterior distribution in the j-th simulation run. We chose $n_{sim} = 100$ and $n_{samp} = 100$ for the simulation study. Then the expressions for the relative bias and RMSE of a given parameter θ for each individual simulation run j is:

$$\text{Relative Bias}_\theta = \frac{1}{n_{samp}} \sum_{k=1}^{n_{samp}} (\hat{\theta}_{jk} - \theta)/\theta, \tag{3.16}$$

$$\text{RMSE}_\theta = \sqrt{\frac{1}{n_{samp}} \sum_{k=1}^{n_{samp}} (\hat{\theta}_{jk} - \theta)^2}. \tag{3.17}$$

We also use RMSE to evaluate the predictive performance of the model. In each simulation run $j, j = 1, \ldots, n_{sim}$, denote $y_j(\boldsymbol{s}_l, t), \boldsymbol{s}_l \subset D, l = 1, \ldots, n_{pred}$ as the value of the simulated surface at location $\boldsymbol{s}_l$ and time t, and $\hat{y}_j(\boldsymbol{s}_l, t)$ is the prediction of $y_j(\boldsymbol{s}_l, t)$, i.e., the posterior mean from $p(\boldsymbol{y}(\boldsymbol{s}_l, t)|\boldsymbol{z}_1, \boldsymbol{z}_{2,obs})$ for $t \in \{1, \ldots, 14\}$ and the posterior predictive mean from $p(\boldsymbol{y}(\boldsymbol{s}_l, t)|\boldsymbol{z}_1, \boldsymbol{z}_{2,obs})$ for $t \in \{15, \ldots, 19\}$. Since the size of the model fit scales with the number of predictions required, we chose $n_{sim} = 100$ and $n_{pred} = 20$ to be able to complete the simulations within a reasonable computational budget. The n_{pred} locations were randomly sampled over the spatial domain D for each simulation run, so that in aggregate we have representative coverage of the entire domain D. For each simulation run j, the RMSE of the prediction at time t over all n_{pred} locations is calculated as follows:

$$\text{RMSE}_{\boldsymbol{y}_t} = \sqrt{\frac{1}{n_{pred}} \sum_{l=1}^{n_{pred}} (\hat{y}_j(\boldsymbol{s}_l, t) - y_j(\boldsymbol{s}_l, t))^2}. \tag{3.18}$$

3.4.3 Simulation Results

Since the fusion model is proposed to reconcile spatial misalignment between in situ and satellite data, we assess its performance and effectiveness by comparing it with the standalone in situ and satellite models. This comparative analysis involves the estimation of parameters and prediction using each dataset in isolation. Therefore, we obtain the estimates and predictions for the standalone in situ model using (3.5), (3.6), (3.7) and for

the standalone satellite model using (3.4), (3.6), (3.7). Note that the bias parameter a is not estimable in both standalone models. This is because in the in situ model (3.5), a is absent, and in the satellite model (3.4), a is subsumed into the parameter β_0.

Parameter estimation results

Our results indicate that all three simulation factors have an overall effect on the relative bias and RMSE of each parameter (see Figures 3.13– 3.18, which aggregate the results by simulation factor). More specifically, the in situ model and fusion model display improved estimates with an increasing number of in situ samples, while the satellite and fusion model provide better estimates when the percentage of missingness decreases from 80% to 50%. The estimates also improve, particularly for the satellite and fusion models, as the mesh becomes more refined. With a spatial range of approximately 0.40 for $\kappa = 7$, mesh sizes of 0.15 and 0.1 are too coarse to adequately represent this range, while 0.05 works well. We also found that the estimates may not exhibit much further improvement when the mesh size becomes too fine relative to the smoothness of the latent field, and thus using an overly fine mesh is computationally inefficient; e.g., Figure 3.19 shows results with an even finer 0.04 mesh size for computation and the predictions are substantively identical to the 0.05 mesh considered in our main simulation scenarios. While the missingness in satellite data is typically beyond control, the findings suggest the importance of choosing a suitably small mesh size for computation (and if possible, increasing the number of in situ samples collected each day).

For all 12 scenarios considered, Figures 3.2 and 3.3 show boxplots of relative bias and RMSE of the parameter estimates for each individual simulation run calculated using (3.16) and (3.17). While the relative biases do not necessarily show a clear pattern over all the scenarios, the RMSEs indicate that the fusion model provides more accurate parameter estimates than in situ and satellite models for the majority of scenarios. Notably, the fusion model exhibits the lowest RMSE for parameters β_0, ρ, κ, and σ^2_ω across all scenarios. It also outperforms the in situ model for all parameters except τ_2, where it requires a finer mesh to yield better results. The fusion model may also not consistently outperform the satellite model for τ_1, β_1 and β_2; the differences between the fusion and standalone models in these scenarios are nonetheless relatively minor, as their boxplots largely overlap. The average relative biases and RMSEs of parameter estimates across the 100 simulation runs are shown in Tables 3.3 and 3.4 respectively.

Further, it is evident from the boxplots that all three models show a relatively small relative bias and RMSE for the fixed effects $\beta_0, a, \beta_1, \beta_2$ as well as for the hyper-parameters ρ and σ^2_ω. On the other hand, the relative bias and RMSE show a larger discrepancy across

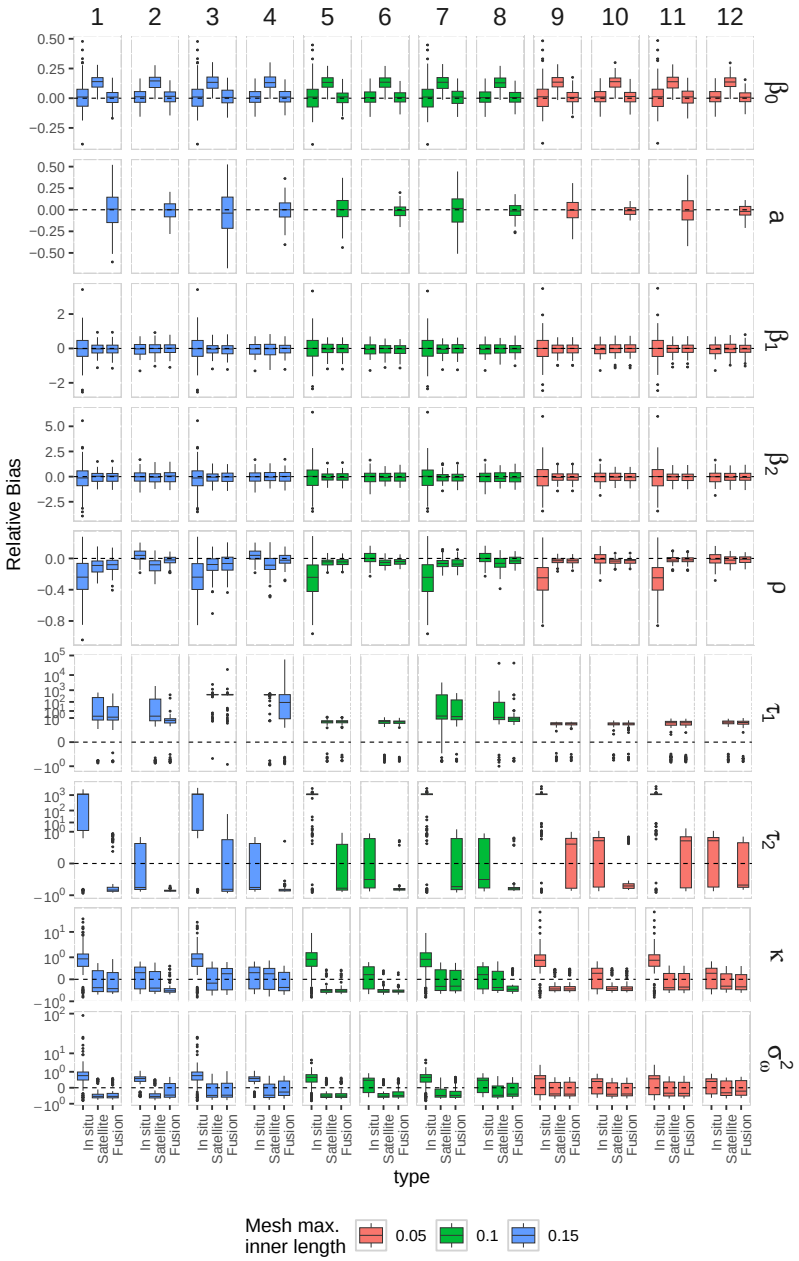


Figure 3.2: Boxplots for relative bias of parameters $\beta_0, a, \beta_1, \beta_2, \tau_1, \tau_2, \kappa, \sigma_\omega^2$ for the in situ, satellite, and fusion models. k from 1 to 12 represents the simulation scenario number in Table 3.1. The boxplot colors represent the three different maximum inner lengths of the triangulation mesh. For visualization clarity, the parameters κ, σ_ω^2 are displayed on a cubic root scale, and τ_1, τ_2 are displayed on a ninth root scale.

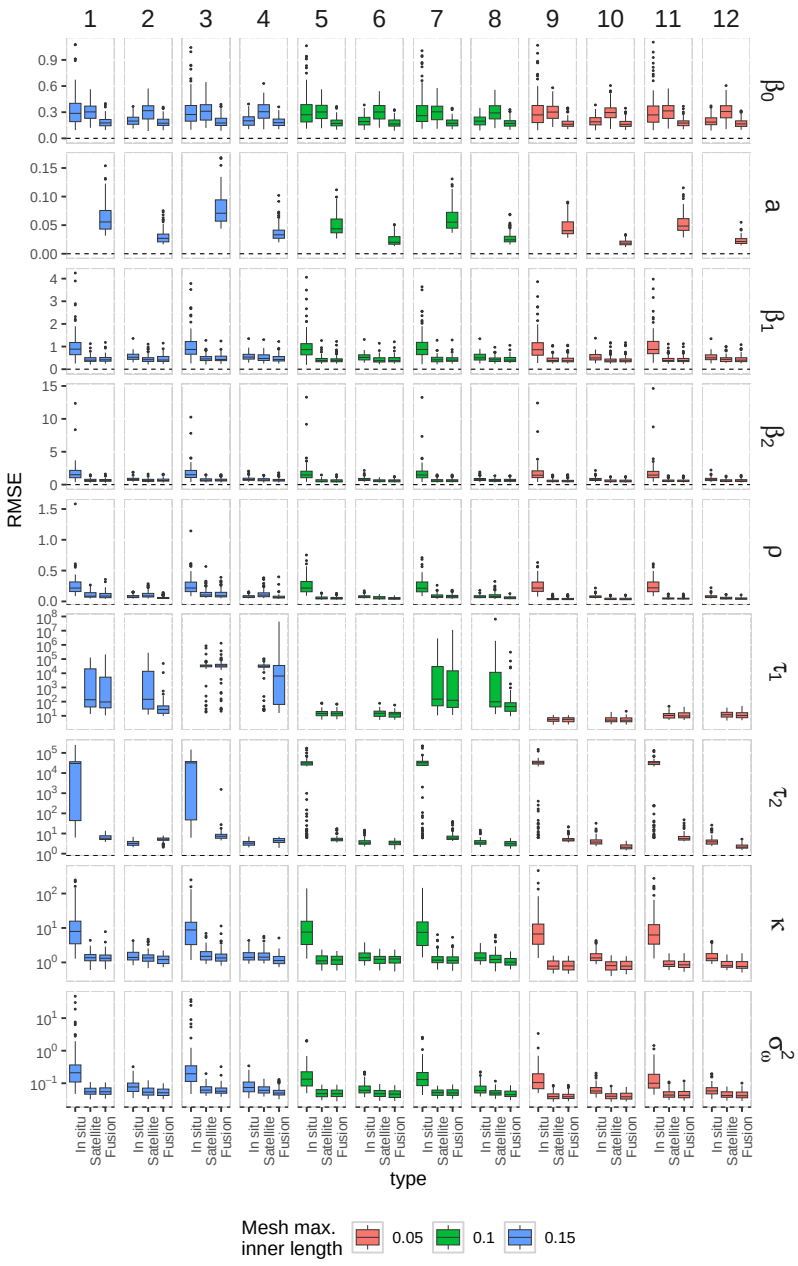


Figure 3.3: Boxplots for RMSE of parameters $\beta_0, a, \beta_1, \beta_2, \tau_1, \tau_2, \kappa, \sigma_\omega^2$ for the in situ, satellite, and fusion models. k from 1 to 12 represents the simulation scenario number in Table 3.1. The boxplot colors represent the three different maximum inner lengths of the triangulation mesh. For visualization clarity, the parameters $\tau_1, \tau_2, \kappa, \sigma_\omega^2$ are displayed on a logarithmic scale.

scenarios for the hyper-parameters τ_1, τ_2, and κ, especially in the in situ model with only 5 observations, where extreme outliers tend to be present among the estimates. The results suggest $\beta_0, a, \beta_1, \beta_2$, ρ, and σ_ω^2 are less sensitive to changes in simulation settings, while τ_1, τ_2, and κ are more sensitive and should be interpreted with care when data is sparse.

Overall, the fusion model provides better parameter estimates compared to both the in situ and satellite models across various scenarios. The results also indicate the importance of selecting a mesh size that is suitable for the smoothness of the latent field.

Prediction results

Figure 3.4 shows the prediction results over unobserved locations on each day $t = 1, \ldots, 19$, averaged across the 100 simulation runs. Overall, the prediction accuracy from the fusion model outperforms those obtained from the in situ or satellite models in all scenarios. The smallest prediction RMSE occurs in scenario 10, i.e., when the number of in situ samples collected is the highest (30 per day), the percentage of missingness is the lowest (50%), and the mesh size configuration is the smallest.

Similar to parameter estimation, the size of the mesh configuration plays a role in the prediction RMSE. The smallest mesh is observed to yield the lowest RMSE across the evaluated days (scenario 9 - 12), which indicates a sufficiently small mesh is necessary to generate accurate predictions. Time-wise, the RMSE is expected to increase when transitioning from the training period (day 1 - day 14) to the test period (day 15 - day 19). This is because as we progress into the test period, predictions become less reliant on observations and more on noise terms as the correlation decays. Moreover, as we transit to the test period, the distinctions in RMSE between the different models and scenarios become less apparent, as illustrated further in Figure 3.11.

Figure 3.4 also suggests that predictions based on in situ data are not guaranteed to be more accurate than those based on satellite data, and vice versa. Utilizing only 5 in situ samples collected each day may not provide sufficient information for accurate estimation at unobserved locations, as the RMSE is observed to be highest for the in situ model when sampling is sparse (scenarios 1, 3, 5, 7, 9, and 11). However, as the sample size increases to 30, the prediction performance of in situ data has the potential to outperform satellite data, driven by the bias in the latter. The implications could be valuable for institutions heavily reliant on satellite data. Nevertheless, it is noteworthy that predictions based on the fusion model consistently demonstrate better performance, with a substantial margin on training days and a reduced margin on test days. This indicates that both in situ and satellite data

are needed to generate more reliable predictions and draw accurate conclusions in such environmental studies.

To visualize the results for an individual simulation run, Figures 3.5 and 3.6 respectively show the posterior mean and standard deviation for the estimated latent spatio-temporal field on day 1, 5, 10, 15, and 19, where the in situ, satellite, and fusion models are fitted to the simulated data depicted in Figure 3.1. The posterior mean panels, presented on the same color scale as Figure 3.1, are seen to closely reflect the original simulation output in Figure 3.1a and the increasing chl-a concentration observed along the Southern and Western basin. Among all three models, the fusion model generally has the smallest posterior standard deviation in Figure 3.6. The posterior standard deviation panels also reflect the increasing variability on test days 15-19, when compared with the training days 1-14. This emphasizes the fact that future predictions (i.e., during the test period) are more challenging and associated with greater uncertainty.

Spatial only vs. spatial-temporal model

Since data fusion was previously applied only to the spatial model using INLA-SPDE, we provide a comparison of prediction RMSE between the spatial-only and spatio-temporal models in Figure 3.12. The spatial model used to fit the data for this comparison follows the algorithm described in Moraga et al. (2017). The figure illustrates that using a spatio-temporal approach for interpolation in the 14-day period is always better than using a spatial-only approach, when autocorrelation between days is present. The extent of improvement depends on the specific scenario. For example, the improvement from adding a temporal dimension is high when the missingness in satellite data is high and in situ samples are scarce. Conversely, the improvement is relatively small when the missingness in satellite data is low and more in situ samples have been collected.

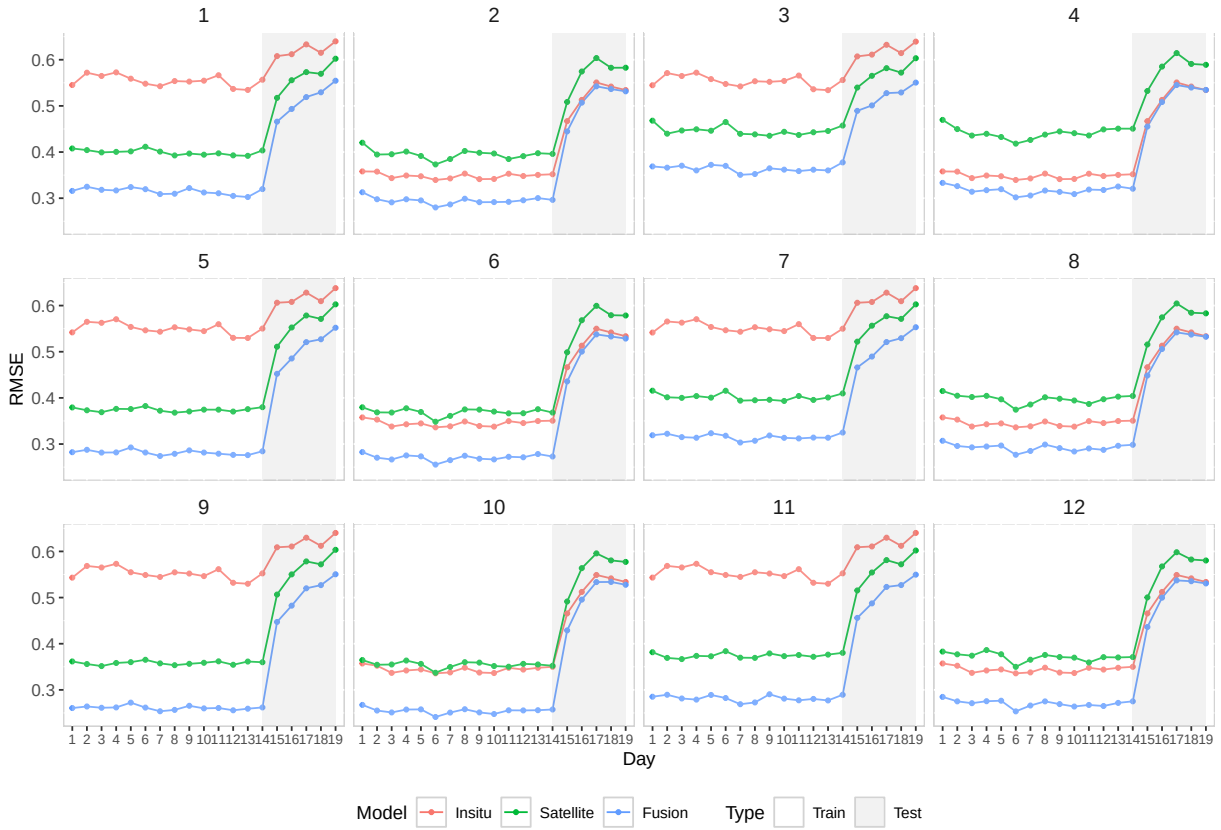
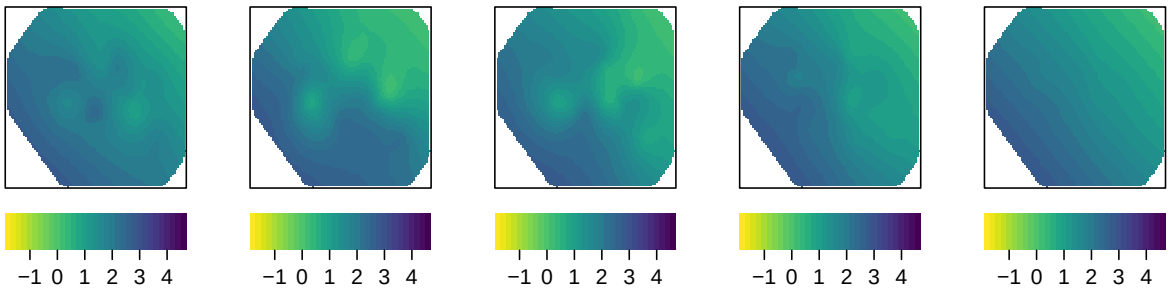
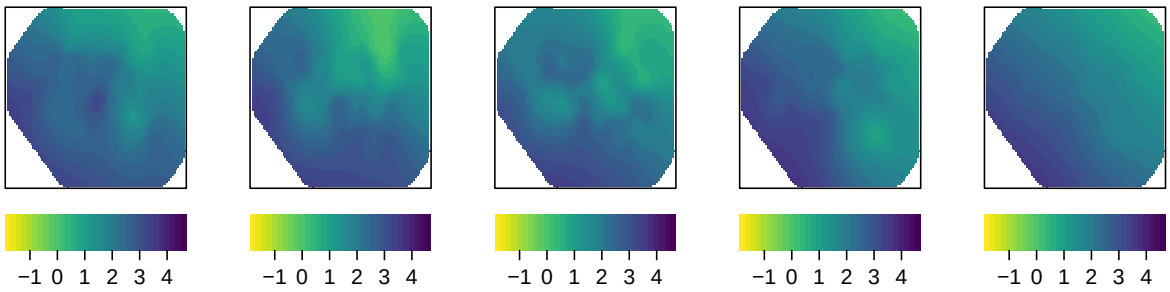


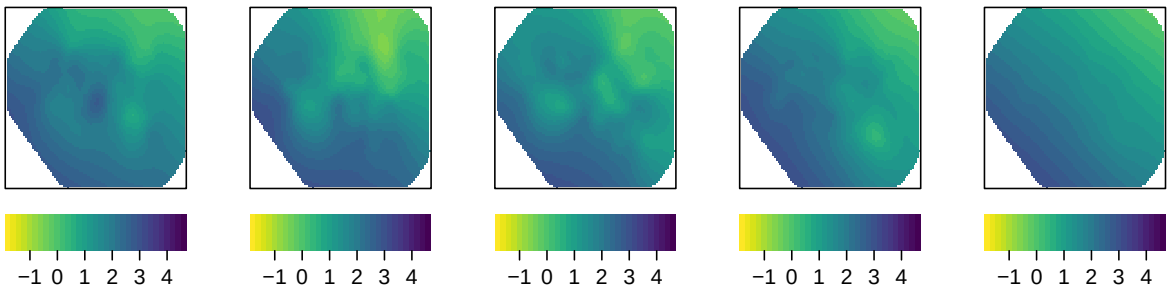
Figure 3.4: Prediction RMSE over unobserved locations on day 1-19 (with test days 15-19 shaded in grey; the training period is days 1-14). Panels 1-12 denote the different scenarios in Table 3.1. The lines colored in red, green, and blue correspond to the in situ, satellite, and fusion models, respectively.



(a) In situ model

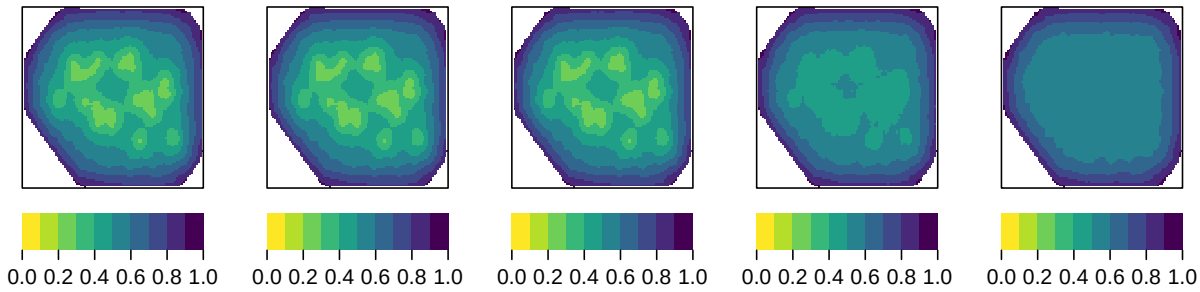


(b) Satellite model

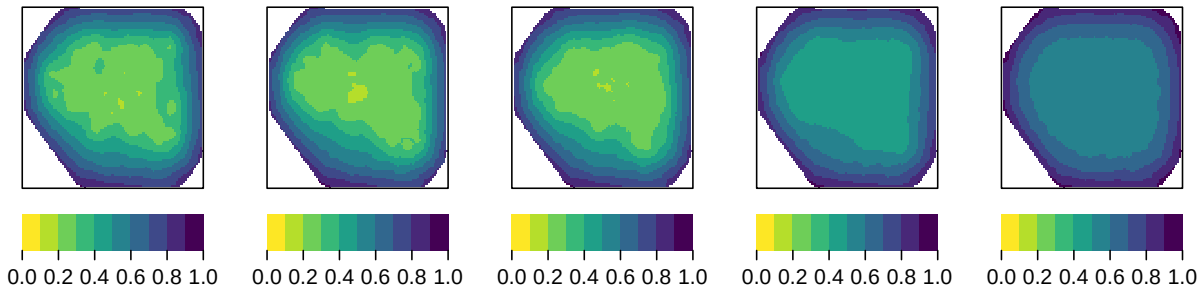


(c) Fusion model

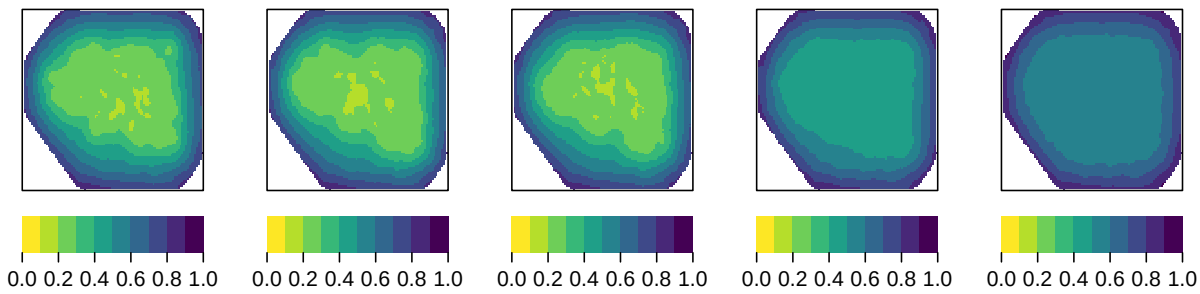
Figure 3.5: Posterior mean of the latent spatio-temporal field fitted by the in situ, satellite, and fusion model on day 1, 5, 10, 15 and 19 respectively for scenario 12 (30 in situ samples, 80% missingness). The color scale represents the posterior mean level of the log chl-a concentration.



(a) In situ model



(b) Satellite model



(c) Fusion model

Figure 3.6: Posterior standard deviation of the latent spatio-temporal field fitted by the in situ, satellite, and fusion model on day 1, 5, 10, 15 and 19 respectively for scenario 12 (30 in situ samples, 80% missingness). The color scale represents the posterior standard deviation of the log chl-a concentration.

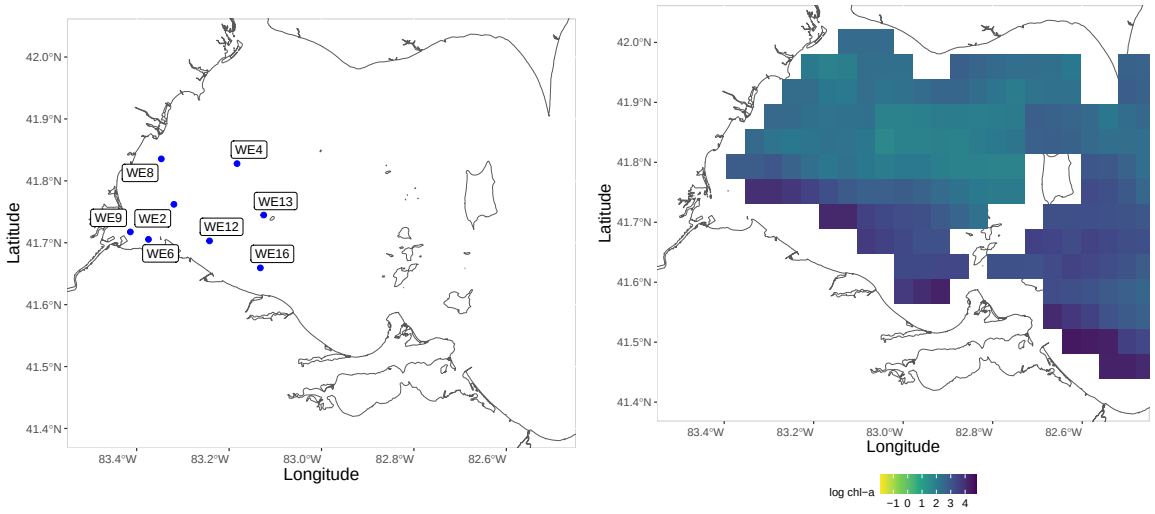
3.5 Application to the cyanobacteria HAB data

We showcase an application of the method with a study focused on cyanobacteria HAB in Western Lake Erie. Lake Erie is one of the largest freshwater lakes in the world, ranking 9^{th} by area and 15^{th} by volume (Herdendorf, 1990, 1992). It is the southernmost of the North American Great Lakes, located between 41°21′N and 42°50′N latitude and 78°50′W and 83°30′W longitude. Based on bathymetry, Lake Erie is divided into western, central, and eastern basins. The western basin, lying west of a line from the tip of Pelee Point, Ontario, to Cedar Point, Ohio, is the smallest and the shallowest basin.

Lake Erie has been observed to experience an increase in the frequency and intensity of cyanobacterial HAB events over the last decade (Bertani et al., 2017; Michalak et al., 2013). The most severe and intense lake-wide cyanobacteria bloom in Lake Erie occurred in 2011 and 2015, with a peak coverage of over 5000 km^2 affecting both the U.S. and Canadian coastlines (Bertani et al., 2016; Stumpf et al., 2012). Moreover, a bloom in 2014 resulted in a loss of safe drinking water for almost 500,000 people living in the Toledo, OH area (Bullerjahn et al., 2016). To deepen our understanding of and mitigate such events, accurate estimation of the underlying drivers and precise predictions are crucial. Successfully achieving this objective hinges on the effective utilization and integration of both in situ and satellite data.

We obtained the in situ data from the in situ monitoring programs at NOAA (National Oceanic and Atmosphere Administration, Great Lakes Environmental Research Laboratory). Figure 3.7a indicates the locations of in situ samples collected over the bloom season. The satellite image data in our study were obtained from NASA Moderate-Resolution Imaging Spectroradiometer (MODIS) Level-3 chl-a products. They were aggregated and projected onto a spatial grid, with each pixel representing the average chl-a in a resolution of 4 km $\times$ 4 km in Lake Erie. Figure 3.7b shows an example of the satellite image over Western Lake Erie on Aug 13th, 2015. On average, these images are limited by 80% cloud coverage, which suggests considerable uncertainties associated with remote sensing measurements.

Two simplifying assumptions of the proposed model might be examined based on the data. First, the adequacy of the constant additive bias a can be assessed via a scatter plot of the matched in situ and satellite data (i.e., by taking the satellite pixel that contains the location of the in situ sample, if observed). Figure 3.8 presents this plot for our dataset, which shows the in situ data (on average) have higher values than satellite observations and indicates evidence of an additive bias. The magnitude of that additive bias possibly depends on the locations of the monitoring stations, though assuming a constant additive



(a) NOAA monitoring stations in Western Lake Erie.

(b) NASA MODIS level-3 chl-a products for Western Lake Erie on Aug 13th, 2015.

Figure 3.7: In situ monitoring stations and a satellite image example for the study of cyanobacteria HABs in Western Lake Erie.

bias appears to be an adequate simplification given the sparsity of the in situ data. The superimposed regression line further suggests the absence of multiplicative bias (i.e., a regression slope of 0.9611 that is close to 1) and the regression intercept (0.1463) provides a rough estimate for the magnitude of the additive bias. Second, a test for the assumption of space-time separability could be considered (e.g., Cappello et al., 2020); however, individual pixels are rarely observed more than 4-5 consecutive days at a time with the high fraction of missingness in this dataset. Hence, such a test cannot be reliably conducted in this setting, and the separable covariance structure should be an adequate simplification.

We fit the combined datasets from May 2014 to May 2016 to the fusion model presented in this paper, incorporating relevant covariates that include longitude, latitude, basin-wide wind speed, air temperature, and surface temperatures as covariates. We used a mesh size of 0.05 for computation, which the simulation study suggests is sufficiently refined for the resolution of these satellite images. The runtime of INLA-SPDE for this dataset is 2-3 hours for the fusion model. For comparison, we also fitted standalone satellite and in situ models to the same dataset. The parameter estimates from all three model fits are summarized in Table 3.2. The fusion model estimates tend to be closer to those of the

satellite model, while the in situ model parameters are more different and have much wider credible intervals (CIs). The discrepancies between the satellite and in situ models may be attributed to the sparsity of the in situ data, in particular the absence of monitoring stations on the Canadian side of the lake, where we would encourage more sampling and public data sources to be available. The fusion model indicates a notable misalignment between in situ and satellite observations, as the 95% CI for the bias term associated with satellite images does not contain zero (as suggested by the scatter plot in Figure 3.8). The presence of bias necessitates the consideration of such a fusion model for an accurate analysis of harmful algae bloom events. Specifically, while the satellite model may yield predictions that are relatively closer to those of the fusion model, they will be biased; meanwhile, in situ observations are more accurate but are too sparse to fit a reliable model for the whole basin. Finally, we find that the covariates longitude, latitude, basin-wide wind speed, air temperature, and surface temperatures are significant predictors for an increase in chl-a concentration.

3.6 Discussion

In this work, we propose a data fusion method for in situ and satellite data on a spatio-temporal scale, which provides an extension to the work in Moraga et al. (2017). Such an extension is necessary as most environmental data has both spatial and temporal dimensions, and relying solely on a snapshot of the spatial data often fails to provide a comprehensive understanding of the underlying dynamics. The fusion method assumes a latent spatio-temporal process that captures the spatial and temporal dependencies while also accommodating large-scale covariates. The model is fitted using the INLA-SPDE approach to allow for fast and scalable computation. The combined in situ and satellite data are mapped via projection matrices that link the satellite and in situ data to the triangulation nodes of the underlying GMRF. Since the INLA-SPDE approach reduces the computational burden compared to traditional MCMC methods, we can effectively handle large spatio-temporal datasets with dimensions up to $10^5 - 10^6$ for the latent Gaussian field $\boldsymbol{\mathcal{X}}$.

Through simulations, we show the fusion model tends to provide better parameter estimates and predictions compared to those derived solely from the in situ and satellite models. Both the parameter estimation and prediction accuracy were shown to be affected by the mesh size, the number of in situ samples, and the percentage of missingness in the satellite data. For parameter estimations, fixed effects, ρ and σ_ω^2 show less sensitivity to changes in simulation settings, while τ_1, τ_2, and κ are more sensitive and should be

Parameter	Fusion Mean (95% CI)	Satellite Mean (95% CI)	In Situ Mean (95% CI)
β_0 (Intercept)	1.55	1.09	0.72
	[1.34, 1.76]	[0.93, 1.25]	[-4.97, 6.14]
a (Bias for satellite images)	-0.47	–	–
	[-0.61, -0.34]		
β_1 (Longitude)	-0.34	-0.30	-0.87
	[-0.46, -0.22]	[-0.43,-0.18]	[-2.23, 0.50]
β_2 (Latitude)	-1.15	-1.13	-1.47
	[-1.33, -0.97]	[-1.31, -0.95]	[-2.84, -0.10]
β_3 (Wind Speed)	-0.02	-0.02	-0.22
	[-0.05, 0.00]	[-0.05, 0.00]	[-0.68, 0.24]
β_4 (Air Temperature)	-0.01	-0.01	-0.00
	[-0.02, 0.00]	[-0.02, 0.00]	[-0.32, 0.31]
β_5 (Surface Temperature)	0.05	0.05	0.10
	[0.04, 0.07]	[0.04, 0.07]	[-0.20, 0.43]
τ_1	36198.12	47598.53	–
	[28964.45, 43902.10]	[18308.94, 140170.08]	
τ_2	1.43	–	22999.16
	[1.12, 1.72]		[1408.88, 95053.89]
ρ	0.43	0.43	0.77
	[0.41, 0.45]	[0.41, 0.44]	[0.23, 0.96]
κ	5.89	5.92	21.88
	[5.68, 6.10]	[5.73, 6.13]	[10.98, 57.71]
σ^2_ω	0.28	0.28	1.26
	[0.27, 0.30]	[0.26, 0.29]	[0.68, 2.15]

Table 3.2: Results for the parameter estimates obtained from the log chl-a concentration data during May 2014 - May 2016. The columns show the posterior mean and the 95% CI for each parameter for the fusion model, satellite model and in situ model.

interpreted with care when data is sparse. Regarding prediction accuracy, the simulation study shows that the fusion model outperforms the in situ and satellite models across all scenarios, and that the relative improvement is more pronounced during the training period than the test period. In general, the gain in prediction accuracy will likely depend on ρ and the lead time of predictions.

We apply the method to the address the COSP of chl-a concentration obtained from in situ and satellite data to better quantify the HAB events. By integrating data from NOAA monitoring stations and NASA chl-a level-3 products, we show strong evidence for a bias in satellite data, and that the inclusion of both datasets in the analysis of HAB events is necessary. Within this field, we believe the proposed fusion method can contribute to the construction of a more robust algae bloom index, going beyond the limitations

of simplistic categorizations as "in situ" or "remote sensing". Furthermore, it has the potential to enhance the unified characterization of bloom dynamics and aid in identifying the key drivers of HAB events.

In this work, we considered a fixed satellite resolution where each pixel represented a 4 km $\times$ 4km area, as informed by our real data application. Higher resolution satellite data, if available, would capture finer spatial details and therefore are expected to reduce the prediction RMSE for the fusion model; such a pattern is observed in the spatial-only fusion model of Moraga et al. (2017). The increased granularity could also improve the precision for some parameter estimates. In related applications, the areal data might be the output of a computer model that represents a smoothed version of reality (i.e., with lower variance). Satellite data could also be considered smoothed in this regard, since satellite pixels represent averages over each pixel area and are therefore smoother than the true spatial field. In a sense, a decrease in areal resolution is associated with a lower variance in the areal data. Investigating this interplay between areal resolution and variance, and how it affects parameter estimates and prediction accuracy, could be interesting in other future applications.

An assumption of constant additive bias was made in this work, which was adequate and improved upon standalone models, but might be restrictive for broader real data applications. The discrepancies between satellite and in situ data may vary across different locations, as suggested by Figure 3.8. To address this, future work could explore more flexible models that allow for spatially (or spatio-temporally) varying additive biases. A possible extension would be to replace the constant bias a with a bias that varies over space and time, namely $a(\boldsymbol{s}, t)$, and treating it as another latent process within the model. For such a spatio-temporally varying additive bias to be identifiable, the in situ data would need to be relatively dense, i.e., collected at a multitude of locations and times to capture local variations. Furthermore, we did not need to consider a multiplicative bias. Thus, another extension could introduce a multiplicative bias into the model, e.g., by letting $z_1(\boldsymbol{B}_j, t) = a + \frac{b}{|\boldsymbol{B}_j|}\int_{\boldsymbol{B}_j} y(\boldsymbol{s}, t)d\boldsymbol{s} + e_1(\boldsymbol{B}_j, t)$ where a is an additive bias and b is a multiplicative bias. Following the discussion in Muff et al. (2015), the product of an unknown parameter b with an integral over the latent process $y(\boldsymbol{s}, t)$ is not directly supported in INLA, as their product is not Gaussian. However, this problem could still be handled with INLA by treating b as a hyper-parameter, so that the latent process conditional on b is Gaussian.

The satellite data often had missing pixels due to cloud cover, which we assumed to be MAR. In other applications, the probability of missingness may depend on the outcome, i.e., a missing not at random (MNAR) mechanism. Then, in contrast to the MAR setting, the missingness mechanism must be included in the model to obtain valid inference. The

recent work of Gomez-Rubio et al. (2022) presents a possible approach to handle MNAR variables directly within the INLA framework, which explicitly adds a binomial likelihood for the missingness indicators with parameters that are estimated together with the main model. Such an approach could also be adopted in the data fusion framework as a future extension.

Future work could address misalignment in time between different data sources. This requires the model to be set up in continuous time or on a set of time knots (Blangiardo and Cameletti, 2015), where satellite data – often representing daily averages or spanning several hours – could be formulated as $z_1(\boldsymbol{B}_j, t) = a(\boldsymbol{B}_j, t) + \frac{1}{|u|}\frac{b}{|\boldsymbol{B}_j|}\int_u \int_{\boldsymbol{B}_j} y(\boldsymbol{s}, u)d\boldsymbol{s}du + e_1(\boldsymbol{B}_j, t), |u| > 0, t = 1, \ldots, T$. The consideration of misalignment is crucial when the measurements have high fluctuations during a day, as seen in variables like temperature. Another direction of future work could be extensions to more complicated space-time models, e.g., non-separable models if the data exhibit a clear violation of the separability assumption. For example, Clarotto et al. (2024) proposed computationally efficient methods that leverage an advection–diffusion SPDE to define a large class of nonseparable spatio-temporal models. These models could offer more accurate representations of asymmetrical space–time processes, such as those found in atmospheric and geophysical phenomena. Recent work by Lindgren et al. (2024) has extended Gaussian Matern fields to a family of diffusion-based spatio-temporal stochastic processes (DEMFs). The DEMF family provides a new class of models, with parameters governing the degree of non-separability, that are compatible with the INLA framework. While choosing appropriate non-separable models can still be a challenging problem, the DEMF family is an accessible starting point for future study in this direction. Other types of data can be considered in such a framework, such as non-stationary, non-isotropic and non-Gaussian data. These extensions are non-trivial, due to the computational complexities of large-scale spatio-temporal models. Generally speaking, methodologies for the integration of point-level and area-level referenced data remains a significant and partially unresolved challenge within the field of spatial statistics.

3.7 Supplementary Tables and Figures

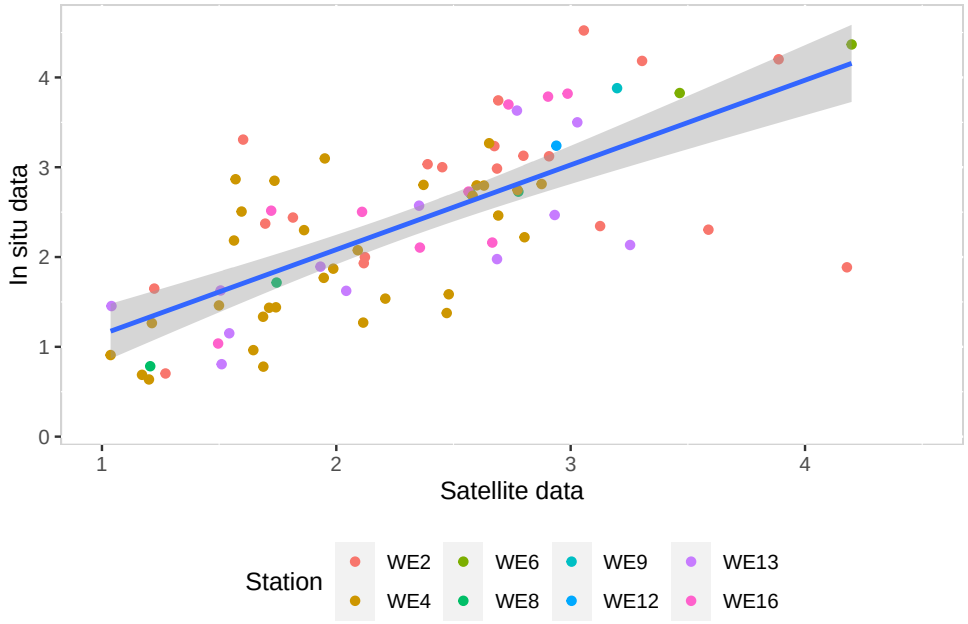


Figure 3.8: Comparison between in situ and satellite log chl-a concentration on a matched spatio-temporal scale. The colors of the points represent different stations where in situ data are collected. The estimated regression line is $y = 0.9611([0.7638, 1.1584])x + 0.1463([-0.3231, 0.6156])$.

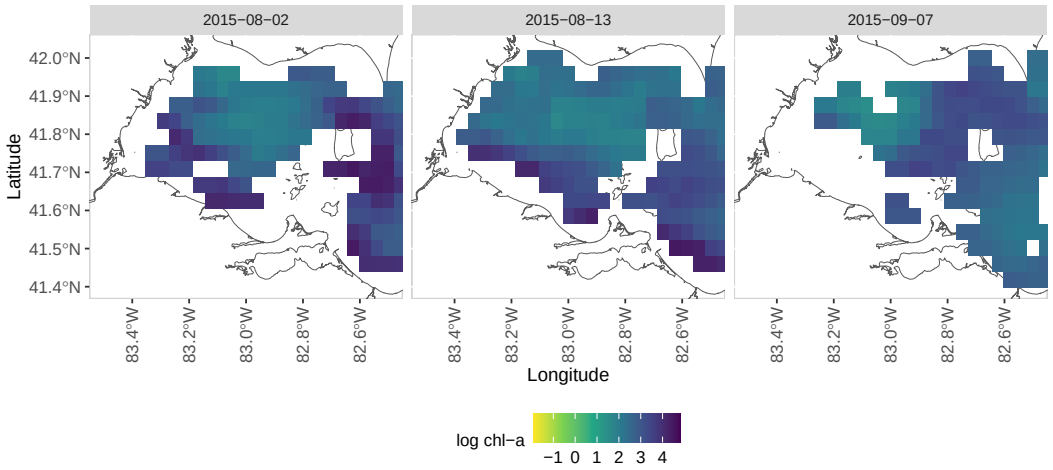


Figure 3.9: Missingness in satellite images on different days

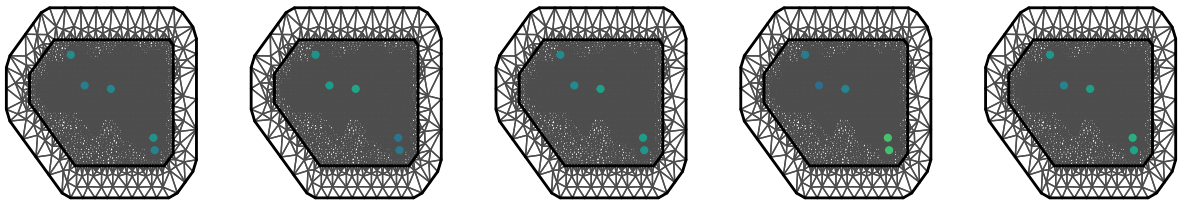


Figure 3.10: An example of simulated in situ samples on days 1, 5, 10, 15, and 19 (5 each day).

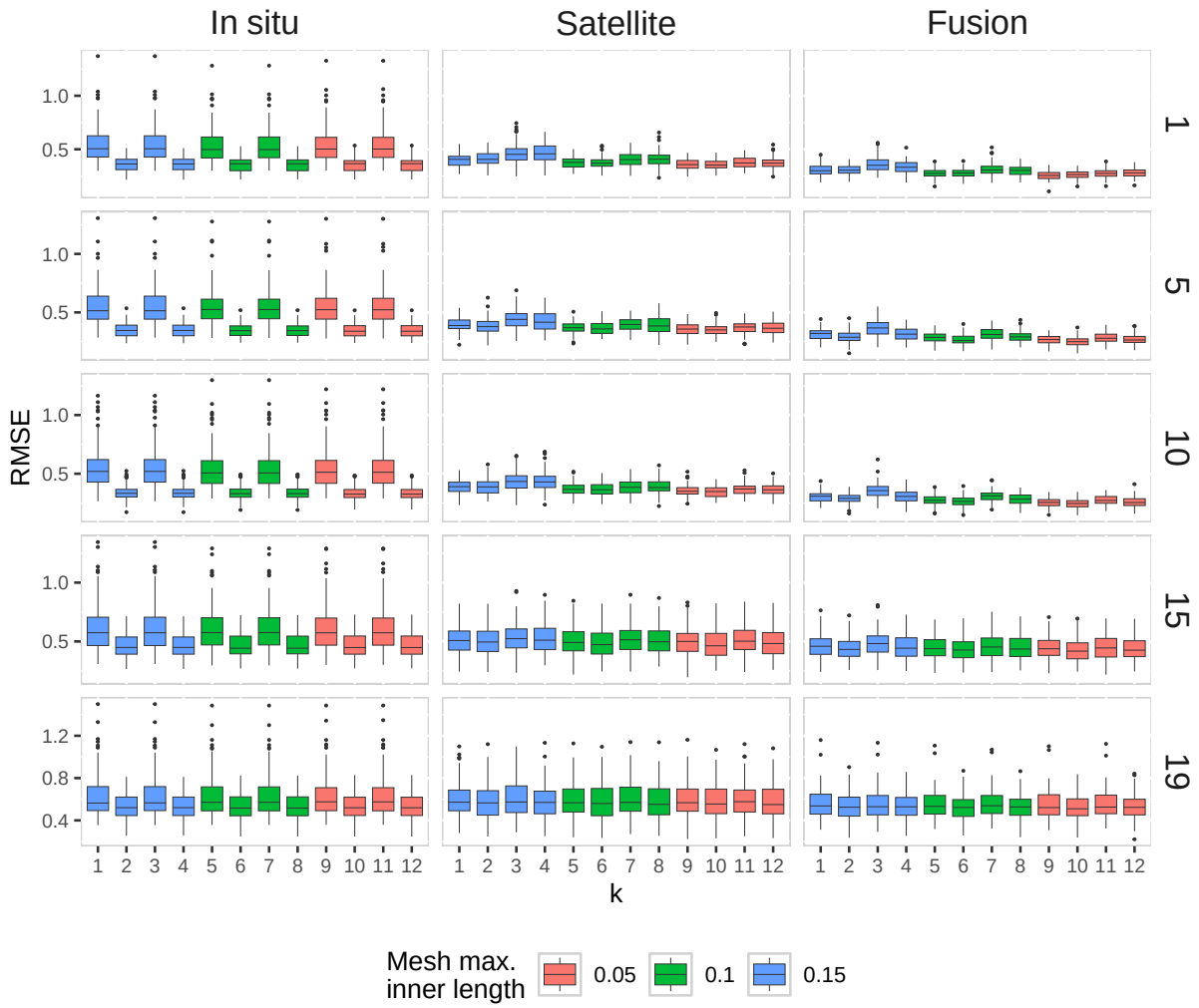


Figure 3.11: Prediction RMSE (calculated using (3.18)) on day 1, 5, 10, 15, and 19. k from 1 to 12 represents the simulation scenario number in Table 3.1. The color in the boxplot represents the three different maximum inner lengths of the triangulation.

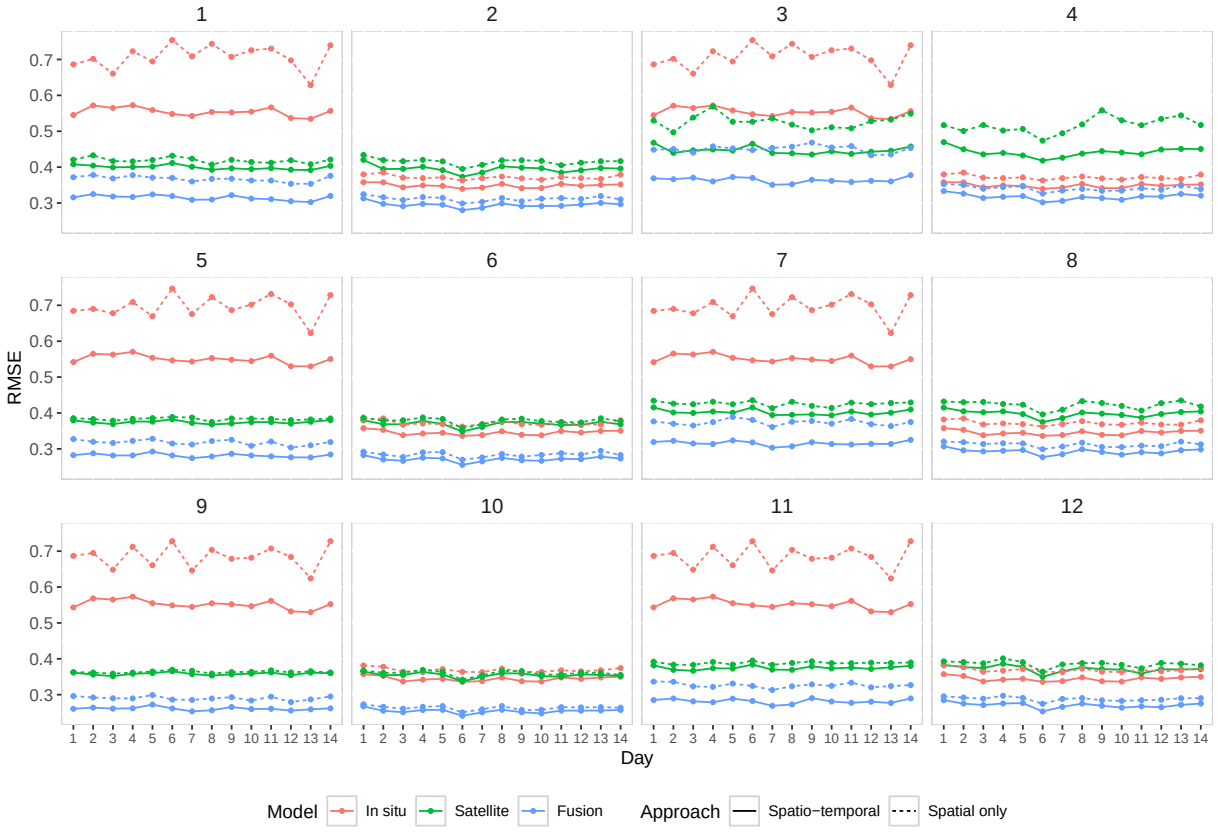


Figure 3.12: Prediction RMSE comparing spatial model with spatio-temporal model

param	k	Insitu	Satellite	Fusion	Parameter	k	In situ	Satellite	Fusion
β_0	1	0.0209	0.1325	0.0081	ρ	7	-0.2391	-0.0678	-0.0627
β_0	2	0.0104	0.1344	0.011	ρ	8	0.0094	-0.0665	-0.0269
β_0	3	0.0209	0.1351	0.0126	ρ	9	-0.2429	-0.0316	-0.0316
β_0	4	0.0104	0.1336	0.0092	ρ	10	-0.0055	-0.0365	-0.0339
β_0	5	0.0155	0.1301	0.0049	ρ	11	-0.2431	-0.0169	-0.0181
β_0	6	0.0067	0.1297	0.0071	ρ	12	-0.0055	-0.0237	-0.0164
β_0	7	0.0155	0.1299	0.0057	τ_1	1	NA	127.9371	97.8812
β_0	8	0.0067	0.1295	0.0061	τ_1	2	NA	134.4736	7.2436
β_0	9	0.0183	0.1334	0.0095	τ_1	3	NA	442.6567	608.8766
β_0	10	0.0096	0.1333	0.0106	τ_1	4	NA	395.8604	1808.5439
β_0	11	0.0184	0.1338	0.0099	τ_1	5	NA	0.2663	0.2598
β_0	12	0.0096	0.1345	0.0112	τ_1	6	NA	0.2522	0.2241
a	1	NA	NA	-0.0045	τ_1	7	NA	209.9192	110.524
a	2	NA	NA	-0.011	τ_1	8	NA	530.5691	449.506
a	3	NA	NA	-0.0318	τ_1	9	NA	0.0922	0.0919
a	4	NA	NA	-0.0025	τ_1	10	NA	0.0906	0.0893
a	5	NA	NA	0.0033	τ_1	11	NA	0.1659	0.1678
a	6	NA	NA	-0.0134	τ_1	12	NA	0.2034	0.1894
a	7	NA	NA	-0.0083	τ_2	1	764.2974	NA	-0.1642
a	8	NA	NA	-0.0091	τ_2	2	-0.0727	NA	-0.2467
a	9	NA	NA	-0.0086	τ_2	3	789.9248	NA	0.4531
a	10	NA	NA	-0.0137	τ_2	4	-0.0727	NA	-0.1999
a	11	NA	NA	-0.0096	τ_2	5	890.8517	NA	-0.0619
a	12	NA	NA	-0.0127	τ_2	6	0.0234	NA	-0.1579
β_1	1	-0.0099	-0.0288	-0.0302	τ_2	7	902.1796	NA	0.0277
β_1	2	-0.0423	-0.0299	-0.0214	τ_2	8	0.0234	NA	-0.1239
β_1	3	-0.0099	-0.0531	-0.0446	τ_2	9	930.557	NA	0.0446
β_1	4	-0.0423	-0.0473	-0.031	τ_2	10	0.0712	NA	-0.0506
β_1	5	-0.0012	-0.0274	-0.03	τ_2	11	925.0311	NA	0.1138
β_1	6	-0.0561	-0.0476	-0.056	τ_2	12	0.0712	NA	-0.022
β_1	7	-0.0013	-0.0441	-0.0408	κ	1	2.0305	-0.0433	-0.0564
β_1	8	-0.0561	-0.0338	-0.0428	κ	2	0.0543	-0.049	-0.1339
β_1	9	0.0075	-0.0358	-0.0366	κ	3	1.821	0.0177	0.0156
β_1	10	-0.0554	-0.0308	-0.0251	κ	4	0.0543	0.0258	-0.0465
β_1	11	0.0078	-0.0175	-0.0173	κ	5	1.2531	-0.1434	-0.1437
β_1	12	-0.0554	-0.0345	-0.0284	κ	6	0.0459	-0.1478	-0.154
β_2	1	-0.0462	-0.0329	-0.017	κ	7	1.2552	-0.0181	-0.0232
β_2	2	-0.0096	-0.05	-0.0129	κ	8	0.0459	-0.0251	-0.0905
β_2	3	-0.0399	-0.0542	-0.0354	κ	9	1.6741	-0.0809	-0.0808
β_2	4	-0.0096	-0.0224	0.0156	κ	10	0.0411	-0.0799	-0.0812
β_2	5	-0.009	-0.0724	-0.0697	κ	11	1.6728	-0.0422	-0.0394
β_2	6	-0.0221	-0.0774	-0.0642	κ	12	0.0411	-0.0371	-0.0441
β_2	7	-0.0077	-0.071	-0.0711	σ^2_ξ	1	2.7177	-0.1546	-0.1504
β_2	8	-0.0221	-0.0738	-0.0457	σ^2_ξ	2	0.2098	-0.1635	-0.0903
β_2	9	-0.013	-0.0408	-0.0394	σ^2_ξ	3	2.039	-0.0869	-0.0745
β_2	10	-0.0265	-0.0499	-0.0422	σ^2_ξ	4	0.2098	-0.1001	-0.006
β_2	11	-0.0129	-0.0622	-0.0598	σ^2_ξ	5	0.406	-0.1319	-0.1303
β_2	12	-0.0265	-0.0386	-0.0326	σ^2_ξ	6	0.1075	-0.1294	-0.1102
ρ	1	-0.2281	-0.097	-0.0847	σ^2_ξ	7	0.4087	-0.0922	-0.0933
ρ	2	0.0415	-0.0951	-0.0241	σ^2_ξ	8	0.1075	-0.0887	-0.0532
ρ	3	-0.2255	-0.0884	-0.0818	σ^2_ξ	9	0.2773	-0.0526	-0.052
ρ	4	0.0415	-0.0892	-0.0164	σ^2_ξ	10	0.0627	-0.0575	-0.0513
ρ	5	-0.2383	-0.049	-0.0479	σ^2_ξ	11	0.2778	-0.0318	-0.0337
ρ	6	0.0094	-0.055	-0.0451	σ^2_ξ	12	0.0627	-0.0226	-0.0075

Table 3.3: Parameter relative bias for the in situ, satellite, and fusion model. k from 1 to 12 represents the simulation scenario number in Table 3.1.

Parameter	k	In situ	Satellite	Fusion	Parameter	k	In situ	Satellite	Fusion
β_0	1	0.3277	0.3048	0.1869	ρ	7	0.2503	0.0898	0.0846
β_0	2	0.2091	0.3083	0.1874	ρ	8	0.079	0.0937	0.0602
β_0	3	0.3229	0.3154	0.1939	ρ	9	0.2466	0.0404	0.0402
β_0	4	0.2088	0.3099	0.1893	ρ	10	0.0803	0.0431	0.0413
β_0	5	0.3187	0.3031	0.183	ρ	11	0.2465	0.05	0.049
β_0	6	0.1997	0.3026	0.1788	ρ	12	0.0804	0.0541	0.0481
β_0	7	0.318	0.2996	0.1827	τ_1	1	NA	12721.6238	11931.2771
β_0	8	0.2001	0.3015	0.1773	τ_1	2	NA	12758.8789	650.5663
β_0	9	0.3167	0.3059	0.1759	τ_1	3	NA	48129.7098	53430.5742
β_0	10	0.198	0.3005	0.1711	τ_1	4	NA	32170.7138	677183.7517
β_0	11	0.3209	0.3058	0.1809	τ_1	5	NA	17.6039	16.5921
β_0	12	0.1968	0.3083	0.1733	τ_1	6	NA	16.8586	14.5407
a	1	NA	NA	0.063	τ_1	7	NA	53460.0442	126278.072
a	2	NA	NA	0.03	τ_1	8	NA	687269.7137	4367.7672
a	3	NA	NA	0.0784	τ_1	9	NA	5.9181	5.9577
a	4	NA	NA	0.0367	τ_1	10	NA	5.8251	5.6901
a	5	NA	NA	0.0504	τ_1	11	NA	12.6236	12.6465
a	6	NA	NA	0.0239	τ_1	12	NA	13.8086	12.9054
a	7	NA	NA	0.061	τ_2	1	27827.1796	NA	6.3757
a	8	NA	NA	0.0276	τ_2	2	3.4645	NA	5.1891
a	9	NA	NA	0.0463	τ_2	3	28767.7894	NA	23.2354
a	10	NA	NA	0.0197	τ_2	4	3.4725	NA	4.4725
a	11	NA	NA	0.0534	τ_2	5	31491.0275	NA	5.5356
a	12	NA	NA	0.0229	τ_2	6	4.0709	NA	3.5522
β_1	1	1.0355	0.4448	0.4485	τ_2	7	33325.0227	NA	7.6994
β_1	2	0.5534	0.4584	0.4583	τ_2	8	4.0714	NA	3.2575
β_1	3	1.0352	0.4924	0.4829	τ_2	9	30529.604	NA	5.6596
β_1	4	0.5594	0.5114	0.4731	τ_2	10	4.6747	NA	2.262
β_1	5	1.012	0.4248	0.4301	τ_2	11	29807.253	NA	7.3065
β_1	6	0.5363	0.4242	0.4255	τ_2	12	4.678	NA	2.4194
β_1	7	1.0136	0.4387	0.4403	κ	1	24.7347	1.468	1.4614
β_1	8	0.5344	0.4361	0.4446	κ	2	1.6911	1.4838	1.2417
β_1	9	1.0111	0.4263	0.4247	κ	3	19.9366	1.7384	1.6784
β_1	10	0.5321	0.423	0.4218	κ	4	1.7062	1.6859	1.2781
β_1	11	1.0266	0.4322	0.4318	κ	5	14.8972	1.2149	1.209
β_1	12	0.5306	0.4476	0.4436	κ	6	1.6262	1.2308	1.2305
β_2	1	1.8208	0.6584	0.6604	κ	7	14.7073	1.3604	1.2945
β_2	2	0.8338	0.6779	0.6911	κ	8	1.6214	1.4372	1.0945
β_2	3	1.8155	0.7343	0.7269	κ	9	19.2096	0.8662	0.863
β_2	4	0.8446	0.7382	0.7153	κ	10	1.5946	0.8499	0.8401
β_2	5	1.8035	0.5985	0.5875	κ	11	17.7331	0.9485	0.9206
β_2	6	0.8139	0.6063	0.6061	κ	12	1.5845	0.9083	0.8733
β_2	7	1.7796	0.6238	0.6272	σ^2_ε	1	1.5005	0.0596	0.0596
β_2	8	0.8124	0.6542	0.6656	σ^2_ε	2	0.0886	0.0597	0.0545
β_2	9	1.7856	0.5913	0.583	σ^2_ε	3	1.4209	0.0666	0.0646
β_2	10	0.8066	0.5807	0.5766	σ^2_ε	4	0.0889	0.0657	0.0563
β_2	11	1.816	0.6216	0.6088	σ^2_ε	5	0.2177	0.0513	0.0511
β_2	12	0.8057	0.6192	0.6241	σ^2_ε	6	0.0715	0.0511	0.0484
ρ	1	0.2577	0.1079	0.1004	σ^2_ε	7	0.2294	0.0534	0.0532
ρ	2	0.0813	0.1102	0.0619	σ^2_ε	8	0.071	0.0534	0.0487
ρ	3	0.2548	0.126	0.1217	σ^2_ε	9	0.1921	0.0421	0.0420
ρ	4	0.0809	0.1211	0.0759	σ^2_ε	10	0.0657	0.0427	0.0421
ρ	5	0.2502	0.057	0.0546	σ^2_ε	11	0.168	0.0485	0.0481
ρ	6	0.0792	0.0601	0.0516	σ^2_ε	12	0.0662	0.0467	0.046

Table 3.4: Average parameter RMSE across 100 simulation runs for the in situ, satellite, and fusion models. k from 1 to 12 represents the simulation scenario number in Table 3.1.

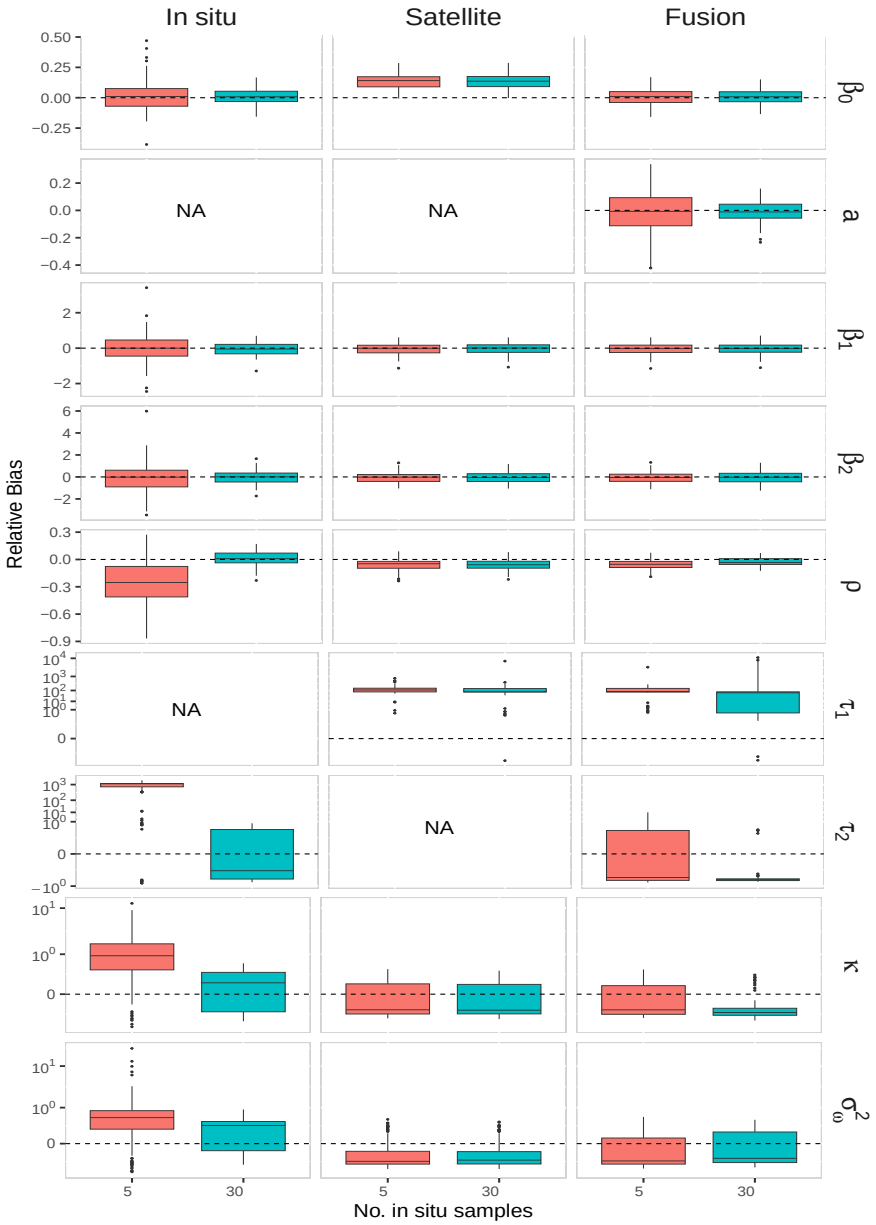


Figure 3.13: Boxplots for relative bias of parameters for the in situ, satellite, and fusion models averaged for factor "No. in situ samples". For visualization clarity, the parameters κ, σ^2_ω are displayed on a cubic root scale, and τ_1, τ_2 are displayed on a ninth root scale.

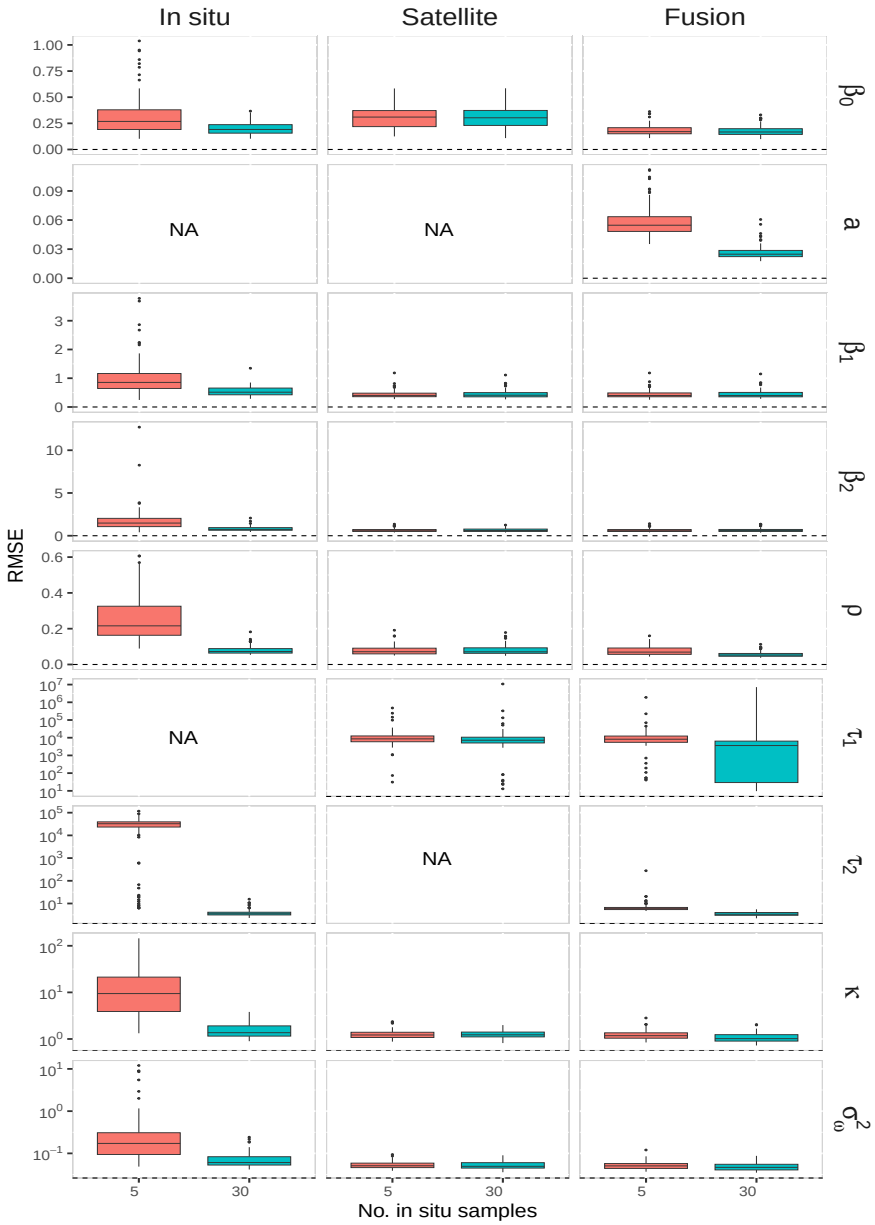


Figure 3.14: Boxplots for RMSE of parameters for the in situ, satellite, and fusion models averaged for factor "No. in situ samples". For visualization clarity, the parameters $\tau_1, \tau_2, \kappa, \sigma_\omega^2$ are displayed on a logarithmic scale.

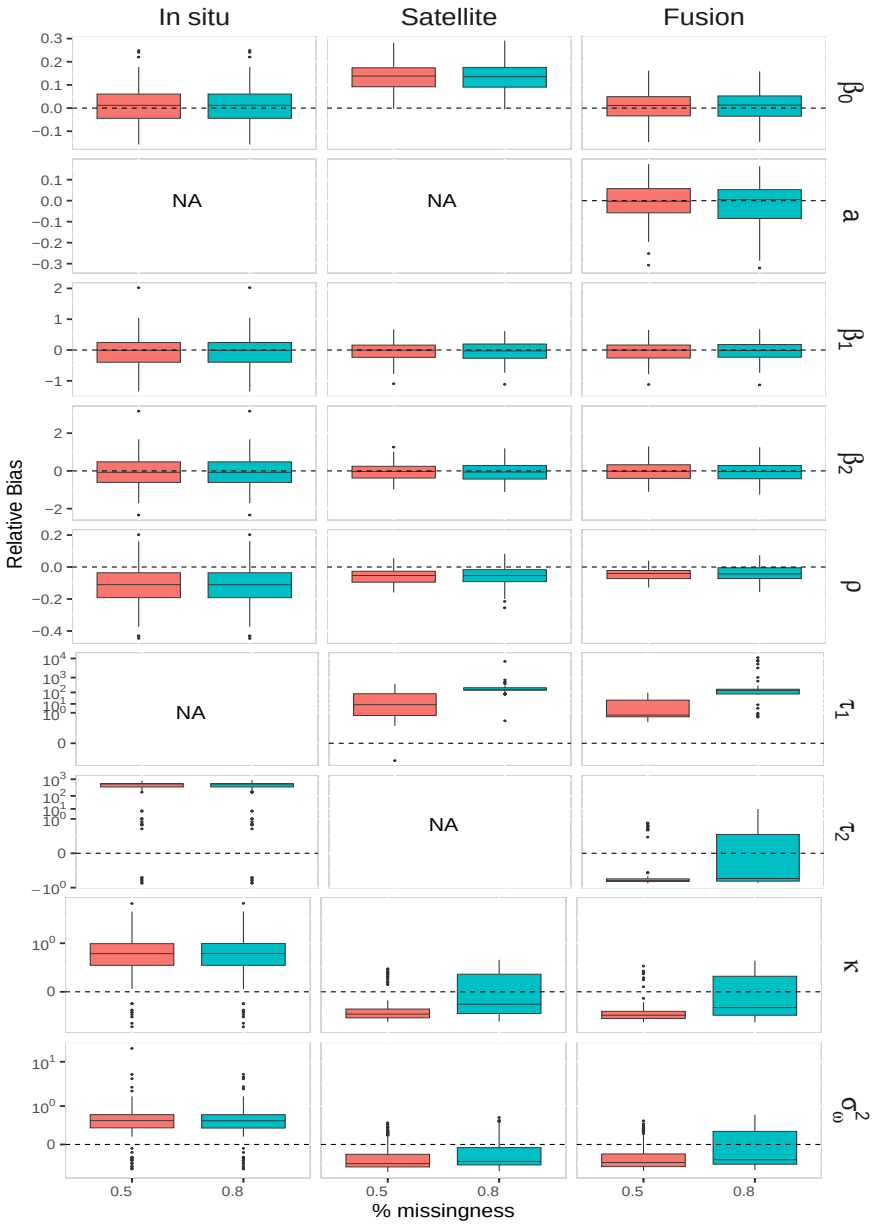


Figure 3.15: Boxplots for relative bias of parameters for the in situ, satellite, and fusion models averaged for factor "% missingness". For visualization clarity, the parameters κ, σ_ω^2 are displayed on a cubic root scale, and τ_1, τ_2 are displayed on a ninth root scale.

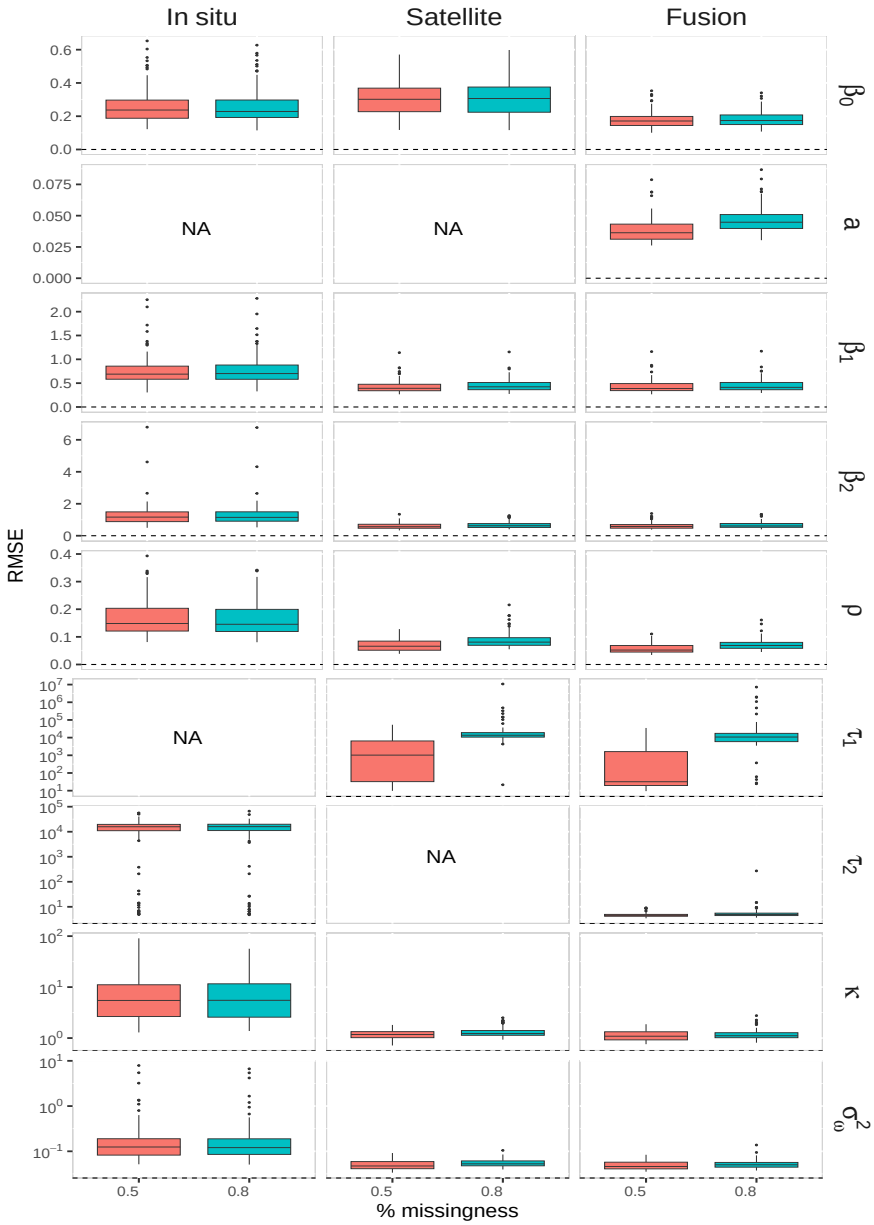


Figure 3.16: Boxplots for relative bias of parameters for the in situ, satellite, and fusion models averaged for factor "% missingness". For visualization clarity, the parameters $\tau_1, \tau_2, \kappa, \sigma^2_\omega$ are displayed on a logarithmic scale.

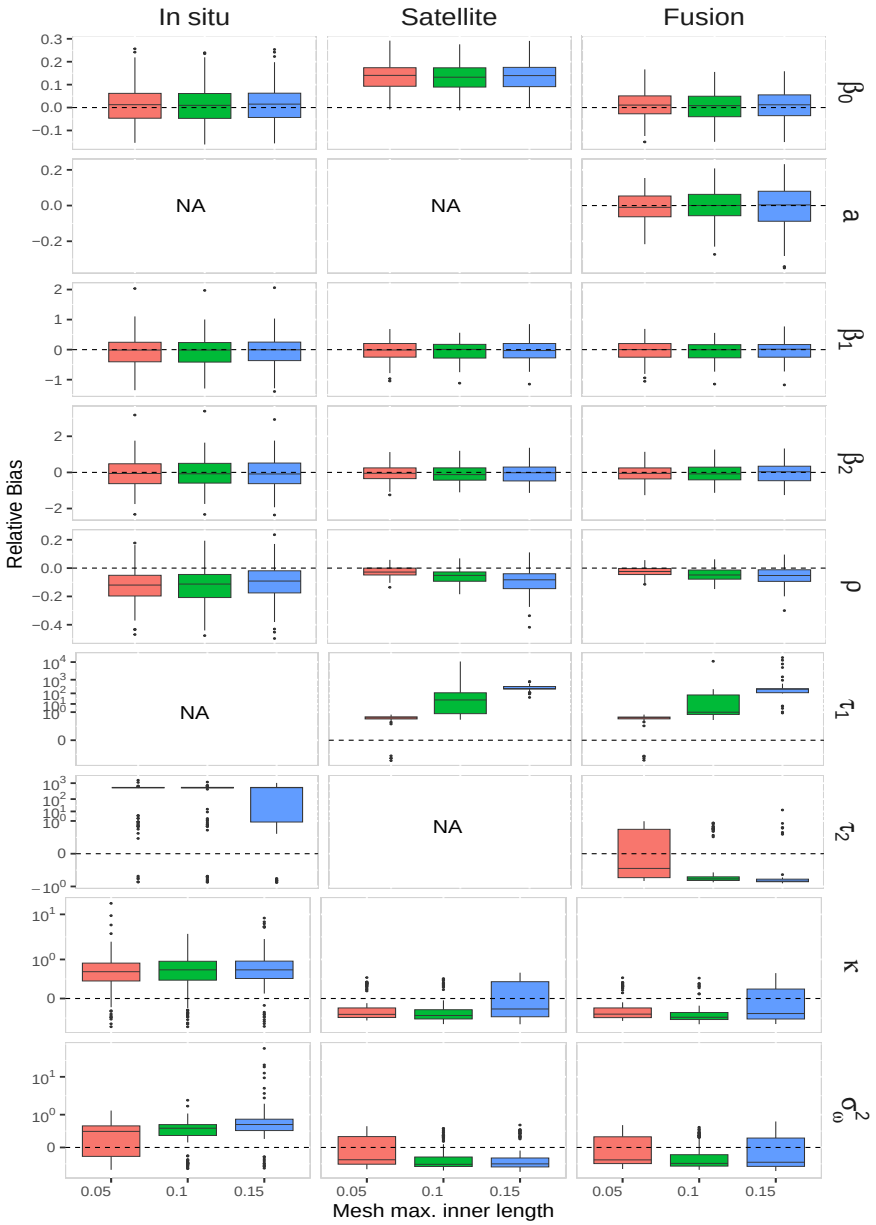


Figure 3.17: Boxplots for relative bias of parameters for the in situ, satellite, and fusion models averaged for factor mesh size. For visualization clarity, the parameters κ, σ_ω^2 are displayed on a cubic root scale, and τ_1, τ_2 are displayed on a ninth root scale.

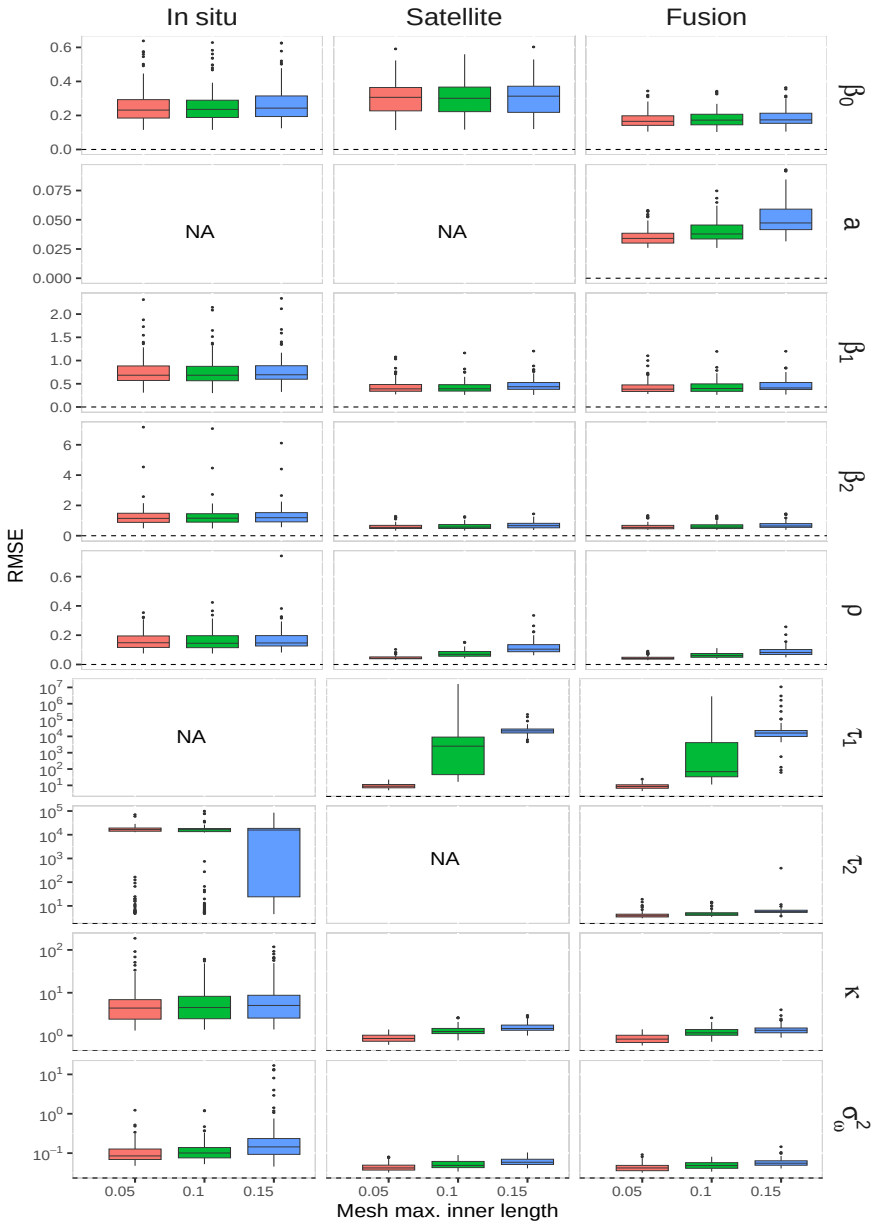


Figure 3.18: Boxplots for relative bias of parameters for the in situ, satellite, and fusion models averaged for factor mesh size. For visualization clarity, the parameters $\tau_1, \tau_2, \kappa, \sigma_\omega^2$ are displayed on a logarithmic scale.

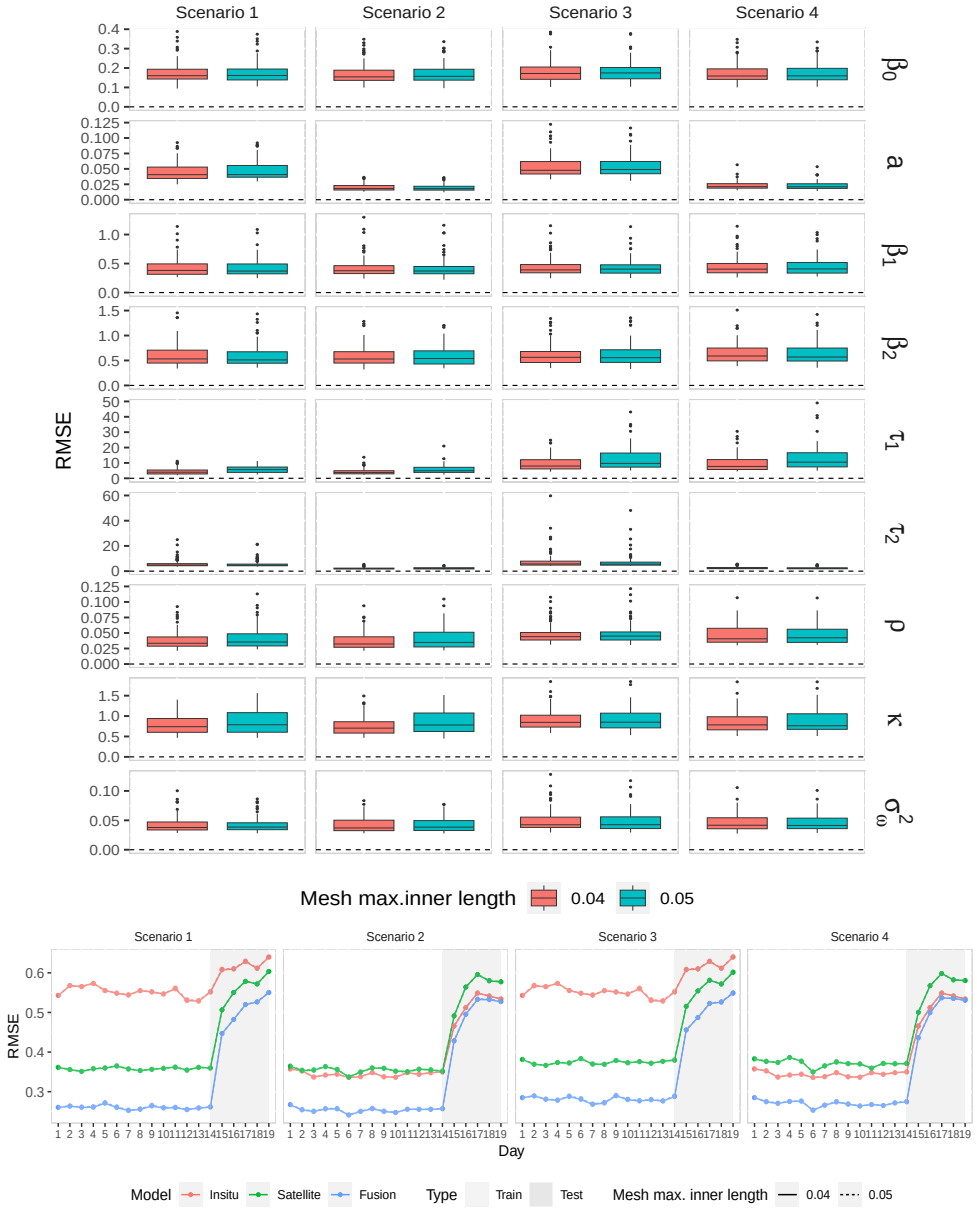


Figure 3.19: Parameter RMSE (fusion model) and Prediction RMSE for scenarios with mesh max. inner length of 0.05 vs.0.04. Scenario 1: 5 in situ samples, 50% missingness, Scenario 2: 30 in situ samples, 50% missingness, Scenario 3: 5 in situ samples, 80% missingness, Scenario 4: 30 in situ samples, 80% missingness.

Chapter 4

Characterizing the spatial effect of knots on flexural properties of wood crossarms through Bayesian methods

4.1 Introduction

Wood-based utility products, such as wood poles and components, have long been favored for supporting overhead electrical power systems because of their wide availability and low-cost (Goodman et al., 1983). In addition to economic advantages, wood-based products offer greater environmental benefits compared to alternative materials. With an estimated 150 million poles in place across the U.S. and Canada (North American Wood Pole Council, 2020), wood has become an integral part of North America's electrical and communication infrastructure. Utility companies deliver electricity from power plants to homes utilizing vast networks of power lines, which are primarily supported by wood crossarms mounted on wood utility poles. Although utility poles are primarily accredited for supporting the transmission lines, crossarms play a critical role in supporting not only the wires but also components such as transformers, fuses, and insulators. Accurately predicting the maximum loads that crossarms may encounter is challenging, due to the variations in loading patterns from wires and components, particularly under adverse weather or biological decay (Goodman et al., 1983). Failures of crossarms in the utility networks can lead to severe consequences, including expensive service interruptions and potential public backlash. As a result, stringent selection criteria are imposed for crossarm materials, which significantly raise costs due to the limited availability of timber meeting these performance requirements

(Liebel and Mueller, 1994).

The natural growing process and the natural growth defects of wood led to high variations in the inherent strength and other mechanical properties of timber (Goodman et al., 1983). During the growth of a tree, knots are formed from branches or limbs, forcing the bole wood to deviate from its relatively straight path, also known as grain deviation. Knots present a weakened zone in lumber that causes higher stresses in adjacent material, interrupts the continuity and change the direction of wood fibers associated with the knot (Zhou and Smith, 1991). Therefore, the presence of knots and grain deviation around it highly impact the structural properties and lead to high variations in the strength. For example, in a bending test on 1,215 pieces of white spruce lumber, 76% of the failure mechanisms observed were caused by knots and 10% were caused by grain variation (Rochester, 1938). Therefore, knots are considered crucial defects in structural timber.

The current lumber grades have incorporated knots into their grading criteria (WWPA, 2021; WCLIB, 2022), based on the visual assessment of lumber characteristics or defects. For example, as shown in Figure 4.1, the criteria used to assess different grades would specify the allowable sizes and locations of the largest single knot or knot cluster for each grade level. This means the complete spatial arrangement of knots is not taken into account, neither does the collective effect of knots in strength reduction. However, knot clusters as well as other adjacent knots with small knot spacing, are found to have additional strength reduction of lumber (Oh et al., 2010; Fink and Kohler, 2014). The effect of knots is also found to be linked to many characteristics of knots, such as size, shape, location, and the attendant local slopes of grain (Cramer and Goodman, 1983; Cramer et al., 1988). In addition, the impact of knots highly depends on the stress state of the wood, e.g., knots in tension have a much stronger impact than knots in compression, and the effect is more serious when the lumber is subjected to forces perpendicular to the grain than along the grain (Baño et al., 2013).



Figure 4.1: Sample image of knots on the surface of a piece of lumber.

The relationship between visual characteristics and wood strength has been explored in the literature. For instance, regression models have been used to characterize lumber strength based on features such as modulus of elasticity, knot diameter ratios, and the maximum diameter of strength-reducing knots (Divos and Tanaka, 1997; França et al., 2018). More recently, Bayesian hierarchical frameworks have been developed to model lumber strength while incorporating uncertainties derived from observable grading characteristics

(Wong et al., 2016). In particular, Fan et al. (2023) studied the strength-reducing effects of knots on the maximum tensile strength of Douglas Fir lumber using a Bayesian approach. Stochastic finite element methods, incorporating random process models, have also been applied to describe the spatial variability in the strength distribution of timber boards (Kandler et al., 2015). Despite these advances, the evaluation of wood strength remains subject to significant uncertainty, and the influence of spatial interactions between knots on strength properties is not yet fully understood.

This project aims to build on the work of Fan et al. (2023) by modeling the impact of knot spatial arrangements and developing a strength prediction framework based on mechanical experiments of crossarms. Section 4.2 provides background on the experimental setup and mechanical analysis of crossarms. Section 4.3 introduces the dataset used in this study. Section 4.4 presents three methods to model the behavior of wood crossarms, which incorporates both a deterministic analysis of induced mechanical stress and a Bayesian framework to capture uncertainties in the failure mechanism. Simulation studies are presented for each proposed method. Section 4.5 demonstrates the application of these methods to real data. Finally, section 4.6 provides concluding remarks and discusses future research directions.

4.2 Background on mechanical properties

To evaluate the effect of knots on the structural strength of crossarms, a researcher at Oregon State University designed an experiment to simulate the real mechanical loading of crossarms (Anderson et al., 2021). The crossarms were subjected to a complex cantilever stress state: each arm was pinned at its center and one end to a steel beam, while loads were applied at the free end using an actuator. Spacers were used to ensure the arms could freely deflect between the pinned supports. See Figure 4.2 for an illustration of the experimental set up. During testing, the arm was placed horizontally and the load was applied at an angle of 17.5° perpendicular to the arm via the actuator. The experimental setup enables the analysis of internal stresses that develop within the crossarms during loading.

When a lateral load is applied, internal stresses develop within the structure in response to the externally applied force. These internal stresses can be categorized into two types: bending stress and shear stress. Bending stress arises when the arm experiences a positive bending moment, causing it to curve concave-upward. This bending induces compressive stresses near the top surface and tensile stresses near the bottom surface. As their names imply, compressive stress acts to compress the material, while tensile stress stretches it



Figure 4.2: Cantilever stress experiment mimicking the real mechanical loading of utility pole crossarms. Reprinted from "Ability to predict flexural properties of Douglas-fir crossarms," by Anderson et al., *Wood Material Science & Engineering*, © 2021, reprinted by permission of Informa UK Limited, trading as Taylor & Francis Group, https://www.tandfonline.com.

apart. Theoretically, the bending stress distribution across the cross-section is linear (Craig et al., 2020). It transitions from a maximum tensile stress at the underside to zero at the neutral axis (the arm's midplane), and then to a maximum compressive stress at the overside. These stresses can lead to lumber failure when they exceed the material's strength.

Lumber failure in tension typically occurs through the initiation and growth of cracks originating from the lower tensile surface. In contrast, lumber failure in compression often happens due to the buckling of outer fibers at the upper compressive surface. The tensile or compressive stress at a specific location s on the top or bottom surface can be calculated using the equation:

$$\sigma(s) = \frac{M(s)X}{I}, \tag{4.1}$$

where M denotes the bending moment, X denotes the distance from the neutral axis to location s and I is the moment of Inertia (Mott, 2007). Since the load was applied at an angle of 17.5° to the crossarm, bending stress was induced in both the length-wise and width-wise directions. According to the principle of superposition, the combined effect of multiple loads acting simultaneously is equivalent to the algebraic sum of the effects of each load acting individually in a linear elastic structure (Courtney, 2005). Consequently, the combined bending stress σ at location s on the top or bottom surface of the crossarm can be expressed as:

$$\sigma(s) = \frac{M_X(s)Y}{I_X} + \frac{M_Y(s)X}{I_Y}, \tag{4.2}$$

where M_X and I_X represent the bending moment and moment of inertia in the lengthwise direction, and M_Y and I_Y represent the bending moment and moment of inertia in the width-wise direction (Boresi and Schmidt, 2003). X denotes the distance from the length-wise neutral axis to location s and Y denotes the distance from the width-wise neutral axis to location s. Using the above equation, the tensile and compressive stresses can be calculated for each crossarm.

Shear stress also develops in bending, parallel to the lateral loads, and tends to slide vertical planes tangentially against one another. While shear stress is often negligible compared to bending stress in cases where the arm's length-to-height ratio is large, it can still contribute to structural failures. Shear stress depends on the shear force, the cross-sectional area of the arm, and other geometric properties. It can be calculated using the equation:

$$\tau(s) = \frac{V(s)Q}{IB}, \tag{4.3}$$

where V is the shear force of the location, Q is the first moment of area, I is the moment of inertia and B is the width of the arm perpendicular to the direction of the shear. Using this equation, the shear stress can be computed for each crossarm.

4.3 Data description

Destructive testing was performed on 250 kiln-dried Douglas-fir crossarms, each with dimension $3.75'' \times 4.75'' \times 96''$. Out of these 250 pieces, 200 had been rejected under current visual grading standards, while the remaining 50 had been accepted. In the course of testing, several characteristics of the crossarms were recorded, including the maximum flexural strength (MOR), the zone of failure, and the location and size of knots. Figure 4.7 shows an example R plot of a tested crossarm, where facets A and C represent the narrow faces and facets B and D represent the wide faces. The black circles indicate knots, which were reconstructed based on their documented diameters and locations.

To analyze failure patterns, each crossarm was partitioned into seven zones, as illustrated in Figure 4.3. The figure also shows the load was applied in zone 1 at an angle of 17.5° relative to the lengthwise direction. Due to the bending curvature during loading, zones $\{1, 3, 4, 6, 7\}$ were subjected to maximum tensile stress along the bottom surface, whereas zones $\{2, 5\}$ experienced maximum compressive stress along the top. This setup also created increased stress concentration around the pin hole and introduced additional shear stress in zone 1. Table 4.1 summarizes the failure distribution by zone across all 250 crossarms. More than half of the crossarms failed in zone 4, followed by approximately 20% in zone 3, and 12% in zone 1.

To account for the varying failure rates across zones, we conducted a stress analysis using equations (4.2) and (4.3), based on the MOR values for each crossarm. Full details are provided in section 4.7.1. The calculated stresses confirm that zone 4 experiences the highest tensile stress, followed by zone 3, which is consistent with the observed failure trends. In contrast, zones 2 and 5 are primarily under compressive stress, which exhibited significantly fewer failures. The relatively high failure rate in zone 1 can be attributed to the extra shear stress introduced by the pin holes. Overall, the analysis shows that tensile stress is the dominant failure mechanism in crossarms, followed by shear.

For subsequent modeling, we excluded crossarms with missing measurements or invalid (e.g., negative) knot locations. This leads to a final dataset of 198 crossarms.

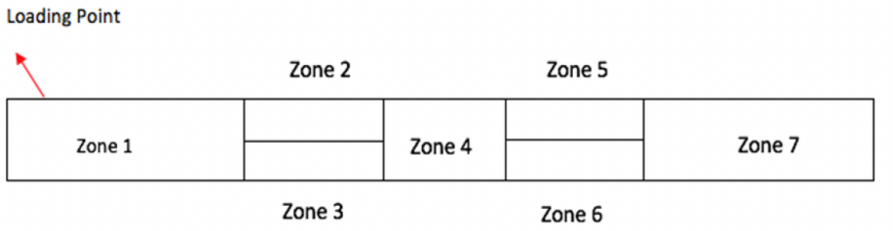


Figure 4.3: Partitioning of the Douglas-fir crossarm into seven zones to examine failure locations during testing. Reprinted from"Ability to predict flexural properties of Douglas-fir crossarms," by Anderson et al., *Wood Material Science & Engineering*, © 2021, reprinted by permission of Informa UK Limited, trading as Taylor & Francis Group, https://www.tandfonline.com.

Zone of Failure	1	2	3	4	5	6	7
Count	30	2	49	137	5	14	6

Table 4.1: Number of crossarms by zone of failures in the destructive test for 250 kiln-dried Douglas-fir crossarms.

4.4 Methods

4.4.1 A hierarchical ICAR model

As indicated by the stress analysis in section 4.3, failures predominantly occur due to tensile stress, with fractures typically forming on the tension side of the crossarm. During data collection, each specimen was partitioned into $r = 7$ zones to facilitate analysis.

Let $\boldsymbol{y_i} = (y_{i1}, \ldots, y_{ir})^\top \in \mathbb{R}^r$ represent the tensile strength for the i-th specimen, where each element $y_{ij}, j = 1, \ldots, r$ corresponds to the zone-specific tensile strength. For clear (knot-free) wood, tensile strength is assumed to be uniform across zones. Accordingly, we model the baseline strength using a constant mean value μ shared across all zones.

To capture the weakening effect of knots, we incorporate zone-specific covariates, with their structure inspired by Fan et al. (2023). Let $\mathbf{x}_{ij}$ denote features describing the knot effects in zone j of specimen i (e.g., location and size), and let their effect on strength be represented by a function $h(\mathbf{x}_{ij})$. Suppose specimen i contains K_i knots. For each knot k, let a_{ik} denote its area, and define d_{ijk} as its distances to the j-th zone along the length of the specimen. The influence of knot k on the strength of zone j depends on both its size and proximity to that zone. We model this using a exponential decay

function: $\beta \exp\{-\gamma d_{ijk}\} \cdot \sqrt{a_{ik}}$, which reflects the intuition that larger knots, and knots closer to the zone of maximum tensile stress exert a stronger weakening effect. Here, β is a scaling factor that controls the magnitude of strength reduction, while γ governs the rate at which the influence diminishes with distance. The knot's area contributes via a square root transformation, which is proportional to its radius. Assuming the effects of multiple knots are additive, the cumulative contribution of all knots in zone j of specimen i is captured through the summation of their individual contributions:

$$h(\mathbf{x}_{ij}) = \sum_{k=1}^{K_i} \sqrt{a_{ik}} \cdot \beta \exp\{-\gamma d_{ijk}\} = \mathbf{D}_{ij}\mathbf{A}_i, \tag{4.4}$$

where $\mathbf{D}_{ij} = \beta \exp(-\gamma \mathbf{d}_{ij})$ is the vector of distance decay terms, $\mathbf{d}_{ij} = (d_{ij1}, \ldots, d_{ijK_i}) \in \mathbb{R}^{1\times K_i}$, and $\mathbf{A}_i = (a_{i1}^{1/2}, \ldots, a_{iK_i}^{1/2})^\top \in \mathbb{R}^{K_i \times 1}$ is the vector of transformed knot areas.

In addition to the effects of knots, tensile strength across zones within a specimen tends to exhibit spatial correlation (Taylor and Bender, 1991; M. Bulleit and A. Chapman, 2004). To account for this, we incorporate an intrinsic conditional autoregressive (ICAR) component to capture the spatial structure across discrete zones. We also incorporate an unstructured random effect term to capture residual heterogeneity. This combined specification resembles the Besag-York-Mollié (BYM) model, which is commonly used in epidemiology and disease mapping (Besag et al., 1991).

Let y_{ij} represent the tensile strength of zone j in crossarm i. The strength measurements are assumed to follow:

$$y_{ij} \sim \mathcal{N}(\eta_{ij}, \sigma_y^2), \quad i = 1, \ldots, n, \quad j = 1, \ldots, r, \tag{4.5}$$

$$\eta_{ij} = \mu + h(\boldsymbol{x_{ij}}) + \phi_j + v_j, \tag{4.6}$$

where ϕ_j and v_j are the structured and unstructured spatial effects. ϕ_j is designed to capture unmeasured wood properties that exhibit spatial correlation, such as variations in grain patterns or deviations that persist across neighboring zones. The unstructured effect v_j accounts for random variability unique to each zone. The parameter σ_y represents residual variability across individual zones and specimens beyond what is explained by the structured and unstructured effects. The distribution of ϕ_j for $j = 1, \ldots, r$ is conditioned on the average values of its neighbors:

$$\phi_j | \phi_{k,k\neq j} \sim \mathcal{N}\left(\frac{\sum_{j\sim k} \phi_j}{d_{j,j}}, \frac{1}{\tau_\phi d_{j,j}}\right), \tag{4.7}$$

where $j \sim k$ represents that region j and region k are neighbors based on the spatial adjacency structure, $d_{j,j}$ represents the number of neighbors directly adjacent to region j,

and τ_ϕ is the precision parameter. This local specification of the conditional distributions leads to a Gaussian Markov random field for $\boldsymbol{\phi}$ that centered at 0:

$$\boldsymbol{\phi} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{Q}^{-1}\right), \tag{4.8}$$

where $\boldsymbol{\phi} = (\phi_1, \phi_2, \ldots, \phi_r)^\top$, and the precision matrix can be constructed from the adjacency matrix given by $\mathbf{Q} = \tau_\phi(\mathbf{D} - \mathbf{W})$. Here, $\mathbf{W}$ is the adjacency matrix, where $\mathbf{W}_{jk} = 1$ if zones j and k are neighbors, and $\mathbf{W}_{jk} = 0$ otherwise. $\mathbf{D}$ is the diagonal degree matrix of row sums, i.e., $\mathbf{D}_{ii} = \sum_j \mathbf{W}_{ij}$, and the precision parameter τ_ϕ controls the spatial variability. Such a precision matrix $\mathbf{Q}$ is known to be symmetric and positive definite.

Let f denote the failure zone, and $G = \{1, \ldots, r\} \setminus \{f\}$ the set of non-failure zones. The observed stress in the failure zone is ξ_{if}, and the non-failure zones is $\xi_{ig}, \forall g \in G$. A specimen is considered to fail when the tensile stress in one of its zones first exceeds the corresponding tensile strength. Accordingly, we characterize the stress-strength relationship across zones as:

$$y_{if} = \xi_{if}, \quad y_{i,g} > \xi_{i,g}, \quad \forall g \in G. \tag{4.9}$$

This formulation allows us to model the likelihood of failure. The parameters of the model are $\boldsymbol{\Theta} = (\boldsymbol{\phi}, \boldsymbol{v}, \mu, \beta, \gamma, \tau_\phi, \tau_v)$. The posterior distribution of parameters can be written as:

$$\begin{aligned}
p(\boldsymbol{\Theta}|\boldsymbol{y}) &\propto p(\boldsymbol{y}|\boldsymbol{\phi}, \boldsymbol{v}, \mu, \beta, \gamma, \tau_\phi, \tau_v) p(\boldsymbol{\phi}|\tau_\phi) p(\boldsymbol{v}|\tau_v) p(\mu) p(\beta) p(\gamma) p(\sigma_y) p(\tau_\phi) p(\tau_v) \\
&\propto \prod_{i=1}^{n} p(y_{if} = \xi_{if}, y_{i,g} > \xi_{i,g}, \forall g \in G | \boldsymbol{\phi}, \boldsymbol{v}, \mu, \beta, \gamma, \sigma_y, \tau_\phi, \tau_v) \\
&\qquad \times p(\boldsymbol{\phi}|\tau_\phi) p(\boldsymbol{v}|\tau_v) p(\mu) p(\beta) p(\gamma) p(\sigma_y) p(\tau_\phi) p(\tau_v) \\
&\propto \prod_{i=1}^{n} \int_{\xi_{i,g}}^{\infty} p(y_{i,g}, \forall g \in G | \boldsymbol{\phi}, \boldsymbol{v}, \mu, \beta, \gamma, \sigma_y) dy_{i,g} \times p(y_{if} | \boldsymbol{\phi}, \boldsymbol{v}, \mu, \beta, \gamma, \sigma_y) \\
&\qquad \times p(\boldsymbol{\phi}|\tau_\phi) p(\boldsymbol{v}|\tau_v) p(\mu) p(\beta) p(\gamma) p(\sigma_y) p(\tau_\phi) p(\tau_v) \\
&\propto \prod_{i=1}^{n} \int \cdots \int_{\xi_{i,g}}^{\infty} \prod_{g \in G} \sigma_y^{-1} \exp\left(-\frac{1}{2\sigma_y^2} (y_{ig} - \eta_{ig})^2 \right) \prod_{g \in G} dy_{i,g} \\
&\qquad \times \sigma_y^{-1} \exp\left(-\frac{1}{2\sigma_y^2} (y_{if} - \eta_{if})^2 \right) \times \tau_v^{1/2} \exp\left(-\frac{\tau_v v_j^2}{2} \right) \\
&\qquad \times |\mathbf{Q}|^{1/2} \exp\left(-\frac{1}{2} \boldsymbol{\phi}^\top \mathbf{Q} \boldsymbol{\phi} \right) \times p(\mu) p(\beta) p(\gamma) p(\sigma_y) p(\tau_\phi) p(\tau_v).
\end{aligned} \tag{4.10}$$

Simulation for the hierarchical model

To assess the model's performance in capturing the spatial effects of knots, we conduct a simulation study. To generate synthetic data, we first construct an adjacency matrix $\mathbf{W}$ that defines the spatial relationships among the zones 1 to 7 depicted in Fig. 4.3. This adjacency matrix is constructed as follows:

$$\mathbf{W} = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 \end{bmatrix}. \tag{4.11}$$

The diagonal degree matrix $\mathbf{D}$ is given by:

$$\mathbf{D} = \begin{bmatrix} 2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 2 \end{bmatrix}. \tag{4.12}$$

Therefore, the precision matrix $\mathbf{Q}$ of $\boldsymbol{\phi}$ according to equation (4.8) is given by:

$$\mathbf{Q} = \tau_\phi(\mathbf{D} - \mathbf{W}) = \tau_\phi \begin{bmatrix} 2 & -1 & -1 & 0 & 0 & 0 & 0 \\ -1 & 3 & -1 & -1 & 0 & 0 & 0 \\ -1 & -1 & 3 & -1 & 0 & 0 & 0 \\ 0 & -1 & -1 & 4 & -1 & -1 & 0 \\ 0 & 0 & 0 & -1 & 3 & -1 & -1 \\ 0 & 0 & 0 & -1 & -1 & 3 & -1 \\ 0 & 0 & 0 & 0 & -1 & -1 & 2 \end{bmatrix}. \tag{4.13}$$

The structured spatial effect $\boldsymbol{\phi}$ is then generated from an ICAR prior with $\tau_\phi = 2$. In addition, the unstructured component v_j is independently sampled from a normal distribution $\mathcal{N}(0, \tau_v^{-1})$ with $\tau_v = 1$. The latent mean η_{ij} for each zone j in specimen i is generated according to equation (4.6) with $\mu = 12, \beta = -0.5$ and $\gamma = 0.5$. Finally, the tensile strength values y_{ij} are generated by adding Gaussian noise with $\sigma_y = 2$. To generate stress from the

simulated strength, we first generate MOR using a scale function $\psi(j)$, where the failure zone for each crossarm is determined by:

$$f = \arg\min_j \{y_{ij}\psi(j)\}, \tag{4.14}$$

and the MOR is determined by:

$$\text{MOR} = y_{ij} \times \frac{\psi(f)}{\psi(4)}. \tag{4.15}$$

That is, we scale tensile strength in each zone by the ratio of the maximum stress across all zones to the stress within the zone. The zone with the minimum scaled strength will be the failure zone, and the minimum scaled strength will be the MOR recorded in the experiment. The stress for each zone will be MOR $\times\, \psi(j)$ for $j = 1, \ldots, r$.

	True	Mean	SD	2.5%	97.5%
μ	12.0000	12.1187	0.4866	11.1882	13.1134
β	-0.5000	-0.4563	0.3433	-1.3091	-0.0226
γ	0.5000	0.4818	0.2749	0.0960	1.1525
σ_y	2.0000	1.9909	0.1023	1.8035	2.2026
τ_ϕ	2.0000	1.4103	1.0640	0.1904	4.1307
τ_v	1.0000	1.7521	1.1443	0.3219	4.6770

Table 4.2: Summary of the true parameter values, posterior mean, standard deviation, and 95% credible intervals of the parameters of the simulated ICAR model.

We fit the hierarchical ICAR model in Stan with weakly informative priors to the hyper-parameters, i.e., $\mu \sim \mathcal{U}(8, 15)$, $\beta \sim$ half-$\mathcal{N}(0, 1)$ for $\beta < 0$, $\gamma \sim$ half-$\mathcal{N}(0, 0.5)$ for $\gamma > 0$, $\tau_\phi \sim \Gamma(1, 1)$, $\tau_v \sim \Gamma(1, 1)$, and $\sigma_y \sim$ half-$\mathcal{N}(0, 1)$ for $\sigma_y > 0$. The simulation was run for 5000 iterations with the first 2500 iterations discarded as burn-in.

Table 4.2 presents a summary of the parameter estimates for the simulated model. The posterior estimates for most parameters closely align with their true values. The fixed parameters μ, β and γ are well recovered, though β and γ exhibit wider credible interval, reflecting more uncertainties. The standard deviation σ_y of the observation noise is estimated with high precision. In contrast, the precision parameters τ_ϕ and τ_v show higher uncertainty, as indicated by their wider credible intervals. These results suggest that while the fixed parameters are well-recovered, the spatial precision parameters are more difficult to estimate accurately. This is expected due to the identifiability challenges

commonly associated with separating structured and unstructured spatial components in spatial models. Overall, the results show that the proposed model is capable of recovering the key parameters, but with increased uncertainty in estimating spatial parameters.

4.4.2 A Geostatistical Gaussian process model

In this section, we propose a geostatistical Gaussian process (GP) model to capture spatial correlation in continuous space. Unlike ICAR models, which lack a valid covariance structure, Gaussian processes enable the specification of spatial dependence through covariance functions. Let $y_i(\boldsymbol{s})$ represent the tensile strength at location $\boldsymbol{s}$ in crossarm i, where $\boldsymbol{s} \in D \subset \mathbb{R}^2$. The strength measurements are assumed to follow:

$$y_i(\boldsymbol{s}) = \mu + h(\mathbf{x}_i(\boldsymbol{s})) + e_i(\boldsymbol{s}), \tag{4.16}$$

where μ is the baseline strength shared across all zones, and $h(\mathbf{x}_i(\boldsymbol{s}))$ represents the nonlinear effect of knots at location $\boldsymbol{s}$ in specimen i. The Gaussian process $e_i(\boldsymbol{s})$ captures smooth variations in tensile strength across zones that are not explained by knots alone, and is specified with mean 0 and covariance function $\mathcal{K}(\mathbf{s}, \mathbf{s}')$:

$$e_i(\mathbf{s}) \sim \mathcal{GP}(0, \mathcal{K}(\mathbf{s}, \mathbf{s}')),$$

where $\mathcal{K}(\mathbf{s}, \mathbf{s}')$ is specified using the squared exponential kernel (Williams and Rasmussen, 2006):

$$\mathcal{K}(\boldsymbol{s}_i, \boldsymbol{s}_j) = \alpha^2 \exp\left(-\frac{(s_i - s_j)^2}{2\rho^2}\right). \tag{4.17}$$

Here, α controls the overall variance and ρ is the length-scale parameter governing the degree of smoothness of the function. A small ρ yields rapid variation, while a large ρ leads to a smoother function.

We evaluate the process at a finite set of spatial locations using the centroids of each zone within a crossarm. Let $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_r\} \subset D$ denote the set of centroid locations within each crossarm. The continuous Gaussian process $e_i(\boldsymbol{s}) \sim \mathcal{GP}(0, \mathcal{K}(\boldsymbol{s}, \boldsymbol{s}'))$, when evaluated at these r locations, induces an r-dimensional multivariate Gaussian distribution:

$$\mathbf{e}_i = (e_{i1}, \ldots, e_{ir})^\top = (e_i(\boldsymbol{s}_1), \ldots, e_i(\boldsymbol{s}_r))^\top \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}),$$

where the covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{r \times r}$ has entries $\boldsymbol{\Sigma}_{jk} = \mathcal{K}(\boldsymbol{s}_j, \boldsymbol{s}_k)$ for $j, k \in \{1, \ldots, r\}$.

Let $\boldsymbol{y}_i = (y_i(\boldsymbol{s}_1), \ldots, y_i(\boldsymbol{s}_r))^\top = (y_{i1}, \ldots, y_{ir})^\top$ denote the vector of tensile strength measurements across the r zones in crossarm i, for $i = 1, \ldots, n$. The observed strength measurements are modeled as:

$$y_{ij} = \mu + h(\mathbf{x}_{ij}) + e_{ij}, \quad j = 1, \ldots, r, \quad i = 1, \ldots, n, \tag{4.18}$$

where $h(\mathbf{x}_{ij})$ follows the structure in (4.4). Denote $\eta_{ij} = \mathbf{E}(y_{ij}) = \mu + h(\boldsymbol{x_{ij}})$ and $\boldsymbol{\eta_i} = (\eta_{i1}, \ldots, \eta_{ir})^\top$. Given the strength measurements for each zone, segregation of the failure zone from the non-failure zones follows the specification in equation (4.9). The likelihood of the data can be expressed as follows:

$$p(y_{i,f} = \xi_{i,f}, y_{i,g} > \xi_{i,g}, \forall g \in G) = (1 - \Phi(\boldsymbol{\xi_{i,-f}} | \boldsymbol{\eta_i^*}, \boldsymbol{\Sigma^*}))\phi(\xi_{ij} | \eta_{ij}, \boldsymbol{\Sigma}), \tag{4.19}$$

where Φ and ϕ represent the CDF and PDF of a multivariate Gaussian distribution. In addition, $\boldsymbol{\eta_i^*}$ and $\boldsymbol{\Sigma^*}$ are the conditional mean and covariance of $\boldsymbol{y_{i,-f}}$ given $y_{i,f} = \xi_{i,f}$. They are computed as:

$$\boldsymbol{\eta_i^*} = \boldsymbol{\eta_{i,-f}} + \boldsymbol{\Sigma_{12}\Sigma_{22}^{-1}}(\xi_{ij} - \eta_{ij}), \tag{4.20}$$

$$\boldsymbol{\Sigma^*} = \boldsymbol{\Sigma_{11}} - \boldsymbol{\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}}, \tag{4.21}$$

and $\boldsymbol{\Sigma_{11}} \in \mathbb{R}^{(r-1)\times(r-1)}, \boldsymbol{\Sigma_{12}} \in \mathbb{R}^{(r-1)\times 1}, \boldsymbol{\Sigma_{21}} \in \mathbb{R}^{1\times(r-1)}, \boldsymbol{\Sigma_{22}} \in \mathbb{R}$ are the partitions of the full covariance matrix Σ. Let $\boldsymbol{\Theta} = (\mu, \beta, \gamma, \alpha, \rho)$ denote the full set of model parameters. The posterior distribution of $\boldsymbol{\Theta}$ is given by:

$$\begin{aligned}
p(\boldsymbol{\Theta}|\boldsymbol{y}) &\propto p(\boldsymbol{y}|\mu, \beta, \gamma, \alpha, \rho)p(\mu)p(\beta)p(\gamma)p(\alpha)p(\rho) \\
&\propto \prod_{i=1}^{n}(1 - \Phi(\boldsymbol{\xi_{i,-f}}|\boldsymbol{\eta_i^*}, \boldsymbol{\Sigma^*}))\phi(\xi_{if}|\eta_{if}, \boldsymbol{\Sigma})p(\mu)p(\beta)p(\gamma)p(\alpha)p(\rho) \\
&\propto \prod_{i=1}^{n}\int_{\boldsymbol{\xi_{i,-f}}}^{\infty} \phi(\boldsymbol{y_{i,-f}}|\boldsymbol{\eta_i^*}, \boldsymbol{\Sigma^*})d\boldsymbol{y_{i,-f}} \times \phi(\xi_{if}|\eta_{if}, \boldsymbol{\Sigma}) \times p(\mu)p(\beta)p(\gamma)p(\alpha)p(\rho) \\
&\propto \prod_{i=1}^{n}\int \cdots \int_{\xi_{i,g}}^{\infty} |\boldsymbol{\Sigma^*}|^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{y_{i,-f}} - \boldsymbol{\eta_i^*})^\top(\boldsymbol{\Sigma^*})^{-1}(\boldsymbol{y_{i,-f}} - \boldsymbol{\eta_i^*})\right) \prod_{g\in G} dy_{i,g} \\
&\qquad \times \Sigma_{22}^{-1/2} \exp\left(-\frac{1}{2\Sigma_{22}}(\xi_{if} - \eta_{if})^2\right) \times p(\mu)p(\beta)p(\gamma)p(\alpha)p(\rho).
\end{aligned} \tag{4.22}$$

Simulation for the geostatistical GP model

We assess the model's performance through a simulation study. The parameters used in the simulation study are selected to closely mimic the observed structural variations of

the crossarms in our dataset. To simulate the GP, we sample from the GP prior following equations (4.17) using hyper-parameters $\alpha = 2$ and $\rho = 1.5$. The fixed effect parameters are set to $\mu = 12, \beta = -0.5$, and $\gamma = 0.5$ to generate the latent mean η_{ij} for each zone j in specimen i. Finally, the tensile strength values $\boldsymbol{y_i}$ are simulated with mean $\boldsymbol{\eta_i}$ and covariance matrix $\boldsymbol{\Sigma}$. The failure zone and MOR for each specimen are determined by equations (4.14) and (4.15).

The model is fitted using Metropolis-Hastings samplers with the following weekly informative priors: $\mu \sim \mathcal{U}(8, 15), \beta \sim$ half-$\mathcal{N}(0, 1)$ for $\beta < 0$, $\gamma \sim$ half-$\mathcal{N}(0, 0.5)$ for $\gamma > 0, \alpha \sim$ half-$\mathcal{N}(0, 2)$ for $\alpha > 0, \rho \sim$ half-$\mathcal{N}(0, 2)$ for $\rho > 0$. The simulation was run for 5000 iterations with the first 2500 iterations discarded as burn-in. The acceptance rate of each Metropolis-Hastings sampler was tuned to fall between 40% and 50%. After model fitting, the estimated parameters are compared with true values.

	True	Mean	SD	2.5%	97.5%
μ	12.0000	12.2091	0.1588	11.9086	12.5392
β	-0.5000	-0.4815	0.2946	-1.1523	-0.0458
γ	0.5000	0.5350	0.3322	0.1407	1.4218
α	2.0000	2.0924	0.0905	1.9323	2.2898
ρ	1.5000	1.2313	0.7910	0.0475	2.8999

Table 4.3: Summary of the true parameter values, posterior mean, standard deviation, and 95% credible intervals of the parameters of the simulated Gaussian process model.

Table 4.3 summarizes the posterior estimates for parameters in the simulation. The posterior means are generally close to the true values, indicating good parameter recovery by the model. The results indicate μ and α are estimated with high accuracy and relatively low uncertainty. The knot effect parameters β and γ also showed reasonable alignment with their true values, although they exhibited slightly higher uncertainty, as reflected by their wider credible interval. Among all parameters, the length-scale parameter ρ shows the greatest uncertainty, which suggests the model is less certain about the smoothness of the spatial process. These results suggest the model performs well overall in recovering parameters, with increased uncertainty in estimating spatial correlation.

4.4.3 A mixture of geostatistical Gaussian process model

In practice, specimens experience multiple types of stress, which collectively contribute to failure during testing. As shown in the stress analysis in section 4.3, such failures may arise

from tensile, compressive, or shear stress. To build a more comprehensive failure prediction model, it is essential to account for the individual and combined effects of these different stress components. To this end, we propose a generalized modeling framework, based on the geostatistical GP model, that accounts for failures arising from multiple stress types.

Let $\boldsymbol{y_i}(\boldsymbol{s}) = (y_{i1}(\boldsymbol{s}), y_{i2}(\boldsymbol{s}), y_{i3}(\boldsymbol{s}))^\top$ denote the tensile, compressive, and shear strengths, respectively, at location $\boldsymbol{s} \in D \subset \mathbb{R}^2$, within crossarm i. Following the model specification outlined in section 4.4.2, we assume all zones share a common mean μ_l for each strength type $l \in \{1, 2, 3\}$. The strength measurements for each type l are assumed to follow:

$$y_{il}(\boldsymbol{s}) = \mu_l + h_l(\mathbf{x}_i(\boldsymbol{s})) + e_{il}(\boldsymbol{s}), \tag{4.23}$$

where $h_l(\cdot)$ captures the non-linear effect of knots on the l-th strength component. Each error term $e_{il}(\boldsymbol{s})$ is modeled as a zero-mean Gaussian process: $e_{il}(\mathbf{s}) \sim \mathcal{GP}(0, \mathcal{K}_l(\mathbf{s}, \mathbf{s}'))$, with covariance function $\mathcal{K}_l(\boldsymbol{s}_i, \boldsymbol{s}_j) = \alpha_l^2 \exp\left(-\frac{(s_i - s_j)^2}{2\rho_l^2}\right)$, where α_l controls the marginal variance and ρ_l governs the spatial correlation range for strength type l.

Similar to the univariate model, we evaluate each process at a finite set of spatial locations $\{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_r\} \subset D$, corresponding to the centroids of the zones within each crossarm. For each strength type $l \in \{1, \ldots, 3\}$, the Gaussian process $e_{il}(\boldsymbol{s}) \sim \mathcal{GP}(0, \mathcal{K}(\boldsymbol{s}, \boldsymbol{s}'))$, when evaluated at these r locations, induces an r-dimensional multivariate Gaussian distribution:

$$\mathbf{e}_{il} = (e_{i1l}, \ldots, e_{irl})^\top = (e_{il}(\boldsymbol{s}_1), \ldots, e_{il}(\boldsymbol{s}_r))^\top \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_l),$$

where the covariance matrix $\boldsymbol{\Sigma}_l \in \mathbb{R}^{r \times r}$ has entries $[\boldsymbol{\Sigma}_l]_{jk} = \mathcal{K}_l(\boldsymbol{s}_j, \boldsymbol{s}_k)$ for $j, k \in \{1, \ldots, r\}$. Let $\boldsymbol{y_{ij}} = (y_{ij1}, y_{ij2}, y_{ij3})^\top$ represent the tensile, compressive, and shear strengths, respectively, for zone j in crossarm i. Then, each strength component is modeled using the Gaussian process formulation:

$$y_{ijl} = \mu_l + h_l(\boldsymbol{x_{ij}}) + e_{ijl}, \quad i = 1, \ldots, n; \quad j = 1, \ldots, r; \quad l = 1, \ldots, 3, \tag{4.24}$$

where $h_l(\boldsymbol{x_{ij}})$ is specified as in equation (4.4), i.e., $h_l(\mathbf{x}_{ij}) = \sum_{k=1}^{K_i} \sqrt{a_{ik}} \cdot \beta_l \exp\{-\gamma_l d_{ijk}\}$.

Since the arm will induce tensile, compressive, and shear stress in response to the external load, we denote the corresponding stress as $\boldsymbol{\xi_{ij}} = (\xi_{ij1}, \xi_{ij2}, \xi_{ij3})^\top$ for specimen i zone j. For specimens that failed without a clear cause, all three stress types must be considered, as the failure could potentially result from any of them. To account for this uncertainty, we model the failure mechanism using a three-component finite mixture model. Let $Z_i \in \{1, 2, 3\}$ denote the type of stress causing the failure of crossarm i, i.e., $Z_i = 1$ indicates the specimen is failed by tensile stress, $Z_i = 2$ indicates the specimen is

failed by compressive stress, and $Z_i = 3$ indicates the specimen is failed by shear stress. We assume

$$Z_i \sim Categorical(\boldsymbol{\pi}),$$

where $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3)$ represents the probability of failure caused by each type of stress. The random variable $Z_i, i = 1, \ldots, n$ are treated as nuisance parameters in the posterior distribution of interest. The distribution of each auxiliary variables Z_i can be stated as:

$$p(Z_i = l|\boldsymbol{\pi}) = \pi_l, \quad l \in \{1, 2, 3\}.$$

Denote $\eta_{ijl} = \mathbf{E}(y_{ijl}) = \mu_l + h_l(\boldsymbol{x_{ij}})$ for $i = 1, \ldots, n, j = 1, \ldots, r, l = 1, \ldots, 3$, and f as the failure zone. Given the failure mechanism is Z_i, the following conditions can be derived. For the failure-causing stress component, we have:

$$y_{i,f,Z_i} = \xi_{i,f,Z_i}, y_{i,g,Z_i} > \xi_{i,g,Z_i}, \forall g \in G,$$

where $G = \{1, \ldots, r\} \setminus \{f\}$. For all other stress components, we have:

$$y_{i,j,m} > \xi_{i,j,m}, \forall j \in \{1, \ldots, r\}, \forall m \neq Z_i.$$

Given the mixture component Z_i, the likelihood of the data can be expressed as follows:

$$\begin{aligned} p(y_{i,f,Z_i} = \xi_{i,f,Z_i}, y_{i,g,Z_i} > \xi_{i,g,Z_i}, \forall g \in G, y_{i,j,m} > \xi_{i,j,m}, \forall j \in \{1, \ldots, r\}, \forall m \neq Z_i|Z_i) = \\ (1 - \Phi(\boldsymbol{\xi_{i,-f,Z_i}}|\boldsymbol{\eta^*_{i,\cdot,Z_i}}, \boldsymbol{\Sigma^*_{Z_i}}))\phi(\xi_{ijZ_i}|\eta_{ijZ_i}, \boldsymbol{\Sigma}_{Z_i}) \prod_{m \neq Z_i} (1 - \Phi(\boldsymbol{\xi_{i,\cdot,m}}|\boldsymbol{\eta_{i,\cdot,m}}, \boldsymbol{\Sigma}_m)), \end{aligned} \tag{4.25}$$

where Φ and ϕ represent the CDF and PDF of a multivariate Gaussian distribution. In addition, $\boldsymbol{\eta^*_{i,\cdot,Z_i}}$ and $\boldsymbol{\Sigma^*_{Z_i}}$ are the conditional mean and covariance of $\boldsymbol{y_{i,-f,Z_i}}$ given $y_{i,f,Z_i} = \xi_{i,f,Z_i}$. They are computed as:

$$\boldsymbol{\eta^*_{i,\cdot,Z_i}} = \boldsymbol{\eta_{i,-f,Z_i}} + \boldsymbol{\Sigma_{12Z_i}}\boldsymbol{\Sigma^{-1}_{22Z_i}}(\xi_{ijZ_i} - \eta_{ijZ_i}), \tag{4.26}$$

$$\boldsymbol{\Sigma^*_{Z_i}} = \boldsymbol{\Sigma_{11Z_i}} - \boldsymbol{\Sigma_{12Z_i}}\boldsymbol{\Sigma^{-1}_{22Z_i}}\boldsymbol{\Sigma_{21Z_i}}, \tag{4.27}$$

and $\boldsymbol{\Sigma_{11Z_i}} \in \mathbb{R}^{(r-1)\times(r-1)}, \boldsymbol{\Sigma_{12Z_i}} \in \mathbb{R}^{(r-1)\times 1}, \boldsymbol{\Sigma_{21Z_i}} \in \mathbb{R}^{1\times(r-1)}, \boldsymbol{\Sigma_{22Z_i}} \in \mathbb{R}$ are the partitions of the full covariance matrix $\boldsymbol{\Sigma}_{Z_i}$. Denote $\boldsymbol{\Theta} = (\boldsymbol{\eta}, \boldsymbol{\mu}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{\rho})$, the posterior distribution

of the complete data model can be derived as follows:

$$
\begin{aligned}
p(\boldsymbol{\Theta}, \boldsymbol{\pi}, \boldsymbol{Z}|\boldsymbol{y}) &\propto p(\boldsymbol{y}|\boldsymbol{Z}, \boldsymbol{\Theta}) p(\boldsymbol{Z}|\boldsymbol{\pi}) p(\boldsymbol{\pi}) p(\boldsymbol{\Theta}) \\
&\propto \prod_{i=1}^{n} \Bigg\{ (1 - \Phi(\boldsymbol{\xi_{i,-f,Z_i}} | \boldsymbol{\eta^*_{i,\cdot,Z_i}}, \boldsymbol{\Sigma^*_{Z_i}})) \phi(\xi_{ifZ_i} | \eta_{ifZ_i}, \boldsymbol{\Sigma}_{Z_i}) \\
&\qquad \times \prod_{m \neq Z_i} (1 - \Phi(\boldsymbol{\xi_{i,\cdot,m}} | \boldsymbol{\eta_{i,\cdot,m}}, \boldsymbol{\Sigma}_m) \times \pi_{Z_i} \Bigg\} p(\boldsymbol{\pi}) p(\boldsymbol{\Theta}) \\
&\propto \prod_{i=1}^{n} \int \cdots \int_{\xi_{igZ_i}}^{\infty} \phi(\boldsymbol{y_{i,-f,Z_i}} | \boldsymbol{\eta^*_{i,\cdot,Z_i}}, \boldsymbol{\Sigma^*_{Z_i}}) \prod_{g \in G} dy_{igZ_i} \times \phi(\xi_{ifZ_i} | \eta_{ifZ_i}, \boldsymbol{\Sigma}_{Z_i}) \\
&\qquad \times \prod_{m \neq Z_i} \int \cdots \int_{\xi_{ijm}}^{\infty} \phi(\boldsymbol{y_{i,\cdot,m}} | \boldsymbol{\eta_{i,\cdot,m}}, \boldsymbol{\Sigma}_m) \prod_{j \in \{1,\ldots,r\}} dy_{ijm} \times \pi_{Z_i} p(\boldsymbol{\pi}) p(\boldsymbol{\Theta})
\end{aligned}
\tag{4.28}
$$

The mixture model requires sampling the latent variable Z_i at each iteration. We show the detailed sampling steps in Algorithm 1. Since we use conjugate priors for $\boldsymbol{\pi}$ and $\boldsymbol{Z}$, they can be sampled through Gibbs samplers. Multivariate Metropolis-Hastings samplers are used for the other parameters as they lack conjugate priors.

Simulation for the mixture of stress model

We demonstrate the validity of the mixture model through the following simulation study. Since few specimens in the experiment failed due to compressive stress, we only simulate failure cases caused by tensile and shear stress.

We set the parameters $\boldsymbol{\Theta}$ as follows: $\boldsymbol{\mu} = (12, 0.6), \boldsymbol{\beta} = (-0.5, -0.025), \boldsymbol{\gamma} = (0.3, 1), \boldsymbol{\alpha} = (0.5, 0.02)$, and $\boldsymbol{\rho} = (3, 5)$. Gaussian processes for each strength type are then simulated for each specimen using $\boldsymbol{\alpha}$ and $\boldsymbol{\rho}$. The strength values are generated accordingly using equation (4.24).

Next, for each specimen $i = 1, \ldots, n$, we sample the failure mode $Z_i \sim \text{Categorical}(\pi_1, \pi_2)$ with $\pi_1 = 0.7$ and $\pi_2 = 0.3$. Given Z_i, the failure zone is determined as the zone with the minimum scaled strength of the corresponding stress type. This minimum value is taken as the MOR recorded in the experiment. Stress values for each zone are computed based on the MOR, and for non-failure stress types, stress values are simulated to remain below their corresponding strengths.

Weakly informative priors are used for the parameters: $\mu_1 \sim \mathcal{U}(8, 15), \beta_1 \sim \text{half-}\mathcal{N}(0, 1)$ for $\beta_1 < 0, \gamma_1 \sim \text{half-}\mathcal{N}(0, 0.5)$ for $\gamma_1 > 0, \alpha_1 \sim \text{half-}\mathcal{N}(0, 2)$ for $\alpha_1 > 0, \rho_1 \sim \text{half-}\mathcal{N}(0, 2)$ for $\rho_1 >$

Algorithm 1 Sampling algorithm for the GP mixture model

1. Update

   $Z_1|\boldsymbol{y}, \boldsymbol{\rho}, \boldsymbol{\alpha}, \boldsymbol{\mu}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\pi} \sim$ categorical$(\pi_{11}, \pi_{12}, \pi_{13})$,

   $\vdots$

   $Z_n|\boldsymbol{y}, \boldsymbol{\rho}, \boldsymbol{\alpha}, \boldsymbol{\mu}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\pi} \sim$ categorical$(\pi_{n1}, \pi_{n2}, \pi_{n3})$,

   where $\pi_{il} = \dfrac{\pi_l(1-\Phi(\boldsymbol{\xi_{i,-f,l}}|\boldsymbol{\eta^*_{i,\cdot,l}}, \boldsymbol{\Sigma^*_l}))\phi(\xi_{ijl}|\eta_{ijl}, \boldsymbol{\Sigma}_l)\prod_{m\neq l}(1-\Phi(\boldsymbol{\xi_{i,\cdot,m}}|\boldsymbol{\eta_{i,\cdot,m}}, \boldsymbol{\Sigma}_m))}{\sum_l \pi_l(1-\Phi(\boldsymbol{\xi_{i,-f,l}}|\boldsymbol{\eta^*_{i,\cdot,l}}, \boldsymbol{\Sigma^*_l}))\phi(\xi_{ijl}|\eta_{ijl}, \boldsymbol{\Sigma}_l)\prod_{m\neq l}(1-\Phi(\boldsymbol{\xi_{i,\cdot,m}}|\boldsymbol{\eta_{i,\cdot,m}}, \boldsymbol{\Sigma}_m))}$

   for $l \in \{1, 2, 3\}$.

2. Update $\boldsymbol{\pi}|\boldsymbol{y}, \boldsymbol{Z}, \boldsymbol{\rho}, \boldsymbol{\alpha}, \boldsymbol{\mu}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \sim Dir(\alpha_1 + n_1, \alpha_2 + n_2, \alpha_3 + n_3)$,
   where $n_l = \sum_i 1_{\{Z_l=1\}}, \quad l = 1, \ldots, 3.$

3. For $l \in \{1, 2, 3\}$, update block l: α_l and ρ_l with a bi-variate random walk proposal $\mathcal{N}((\alpha_{l,t-1}, \rho_{l,t-1})^\top,\ \text{diag}((\sigma_{\alpha_l}, \sigma_{\rho_l})^\top))$

4. For $l \in \{1, 2, 3\}$, update block l: $\{\mu_l, \beta_l, \gamma_l\}$ with a multivariate random walk proposal $\mathcal{N}([\mu_{2,t-1}, \beta_{2,t-1}, \gamma_{2,t-1}]^\top,\ \text{diag}((\sigma_{\mu_l}, \sigma_{\beta_l}, \sigma_{\gamma_l})^\top))$

0. $\mu_2 \sim \mathcal{U}(0,1), \beta_2 \sim$ half-$\mathcal{N}(0,0.1)$ for $\beta_2 < 0$, $\gamma_2 \sim$ half-$\mathcal{N}(0,0.5)$ for $\gamma_2 > 0, \alpha_2 \sim$ half-$\mathcal{N}(0,0.1)$ for $\alpha_2 > 0, \rho_2 \sim$ half-$\mathcal{N}(0,2)$ for $\rho_2 > 0$. The simulation was run for 5000 iterations with the first 2500 iterations discarded as burn-in. The acceptance rate for each Metropolis-Hastings sampler was tuned to fall between 20% and 30%.

	True	Mean	SD	2.5%	97.5%
μ_1	12.0000	11.9463	0.0677	11.8566	12.1067
μ_2	0.6000	0.6060	0.0026	0.6014	0.6112
β_1	-0.5000	-0.3675	0.1473	-0.6364	-0.1585
β_2	-0.0250	-0.0155	0.0139	-0.0465	-0.0018
γ_1	0.3000	0.3215	0.0900	0.1551	0.4696
γ_2	1.0000	0.9398	0.2161	0.4537	1.2375
α_1	0.5000	0.5313	0.0285	0.4785	0.5974
α_2	0.0200	0.0213	0.0023	0.0177	0.0275
ρ_1	3.0000	1.6934	0.4176	0.1758	1.6655
ρ_2	5.0000	4.0753	0.4360	3.3249	4.8533
π_1	0.7000	0.6711	0.0324	0.6096	0.7329
π_2	0.3000	0.3289	0.0324	0.2671	0.3904

Table 4.4: Summary of the true parameter values, posterior mean, standard deviation, and 95% credible intervals of the parameters of the simulated mixture model.

Table 4.4 summarizes the result of the simulation study for the mixture model. Overall, the posterior estimates are able to recover most true values. Parameters $\boldsymbol{\mu}, \boldsymbol{\gamma}, \boldsymbol{\alpha}$, and mixture weights $\boldsymbol{\pi}$ are estimated with more precision, with narrow credible intervals that contain the true values. The estimates of $\boldsymbol{\beta}$ exhibit greater uncertainty and some bias, though they remain within the 95% credible intervals. The model shows difficulty in recovering the length-scale parameters, particularly ρ_1, which deviates substantially from its true value. This suggests that the model has challenges in accurately estimating the spatial smoothness. Overall, the 95% credible intervals capture the true values in nearly all cases, except for the spatial smoothness parameters, which are not well recovered.

Table 4.5 presents a comparison between the true latent stress types and their estimated classifications from the mixture model simulation. The true stress type $Z = 1$ represents specimens failed due to tensile stress, while $Z = 2$ corresponds to specimens failed due to shear stress. Out of 137 specimens truly failed by tensile stress, 131 were correctly estimated as $\hat{Z} = 1$, with 6 misclassified as $\hat{Z} = 2$. For the 61 specimens truly failed by shear stress, all were correctly estimated as $\hat{Z} = 2$ with no misclassifications. Overall, the model is able to identify the failure mode for the majority of specimens.

	Est. $\hat{Z}=1$	Est. $\hat{Z}=2$	Total
True $Z=1$	131	6	137
True $Z=2$	0	61	61
Total	131	67	198

Table 4.5: True vs. estimated latent stress type for the simulation of mixture model. $Z=1$ indicates the specimen is failed by tensile stress and $Z=2$ indicates the specimen is failed by shear stress.

4.5 Applications to real dataset

4.5.1 Tensile stress model

In this section, we present and compare the results obtained from the hierarchical ICAR and the geostatistical GP models for analyzing the tensile strength of crossarms. We assume that all 198 crossarms in the dataset failed due to tensile stress. Both spatial models are employed to estimate the effects of knots and other covariates on tensile strength. The priors used in this application are consistent with those specified in the simulation study.

Posterior summaries for the hierarchical ICAR and geostatistical GP models are presented in Table 4.6 and Table 4.7, respectively. Comparing the parameter estimates from the two models reveals several notable differences. The posterior mean of the baseline strength parameter μ is higher in the GP model (11.53) than in the ICAR model (10.66), and the associated credible interval is narrower, suggesting greater uncertainty under the ICAR specification. Both models yield similar estimates for the knot effect parameters β and γ, suggesting a consistent reduction in strength in the presence of nearby knots. Specifically, the ICAR model estimates the mean knot effect as $0.77\exp(-0.14d)$, while the GP model yields $0.70\exp(-0.16d)$, where d denotes the distance from a knot to the edge experiencing maximum stress. The ICAR model exhibits slightly less posterior uncertainty around the knot effect parameters.

The GP model captures spatial dependence through the spatial covariance parameters α and ρ. Notably, the estimate for ρ exhibits high uncertainty, reflecting greater variability in spatial correlation length across crossarms. The ICAR model also show high uncertainty in estimating the precision parameters τ_ϕ and τ_v, indicating challenges in precisely estimating spatial precision components. In addition, the relatively large posterior standard deviations for σ_y in the ICAR model and α in the GP model imply that neither model fully accounts for the reduction in strength due to knots. Ideally, lower variability in the observation

noise and spatial components would indicate that knot effects explain the majority of the strength variation.

Parameter	Mean	SD	2.5%	Median	97.5%
μ	10.6578	0.7738	9.1207	10.6379	12.3082
β	-0.7707	0.2167	-1.2345	-0.7571	-0.3894
γ	0.1424	0.0461	0.0627	0.1386	0.2451
σ_y	2.3804	0.1196	2.1630	2.3744	2.6241
τ_ϕ	0.8511	0.9422	0.0453	0.5109	3.4290
τ_v	0.5553	0.6350	0.0787	0.3483	2.4134

Table 4.6: Posterior summaries (mean, standard deviation, median, and 95% credible intervals) for parameters under the hierarchical ICAR model.

Parameter	Mean	SD	2.5%	Median	97.5%
μ	11.5294	0.3194	10.9069	11.5286	12.1287
β	-0.7006	0.2748	-1.3585	-0.6683	-0.2504
γ	0.1607	0.0676	0.0553	0.1532	0.3196
α	3.2932	0.1474	3.0156	3.2889	3.5868
ρ	1.4841	1.1410	0.0581	1.1831	4.3083

Table 4.7: Posterior summaries (mean, standard deviation, median, and 95% credible intervals) for parameters under the geostatistical GP model.

Model Diagnostics

We use the posterior predictive check (PPC) to assess the adequacy of both models. In a PPC, we simulate new data from the posterior predictive distribution (Gelman et al., 2013):

$$p(\tilde{y} \mid y) = \int p(\tilde{y} \mid \theta) p(\theta \mid y) d\theta,$$

where $\tilde{y}$ represents the simulated data and y is the observed data. In this analysis, y and $\tilde{y}$ represent MOR from the original dataset and the posterior predictive simulations, respectively.

Figures 4.4 and 4.5 present posterior predictive diagnostic plots for the hierarchical ICAR and geostatistical GP models, respectively. Each figure includes a density plot, and

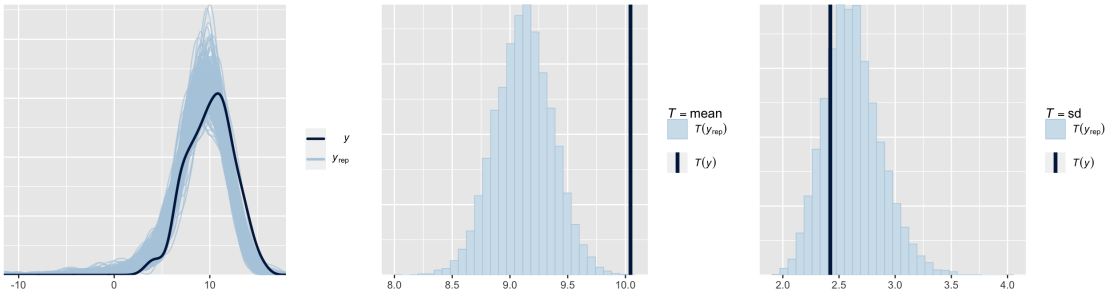


Figure 4.4: Hierarchical ICAR model diagnostics

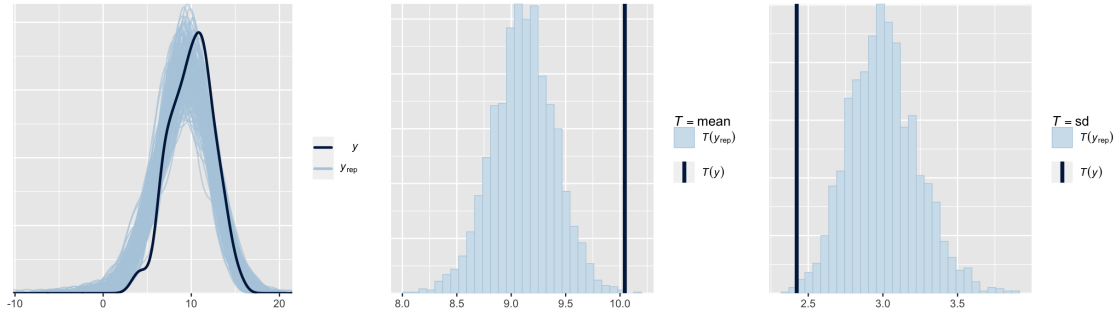


Figure 4.5: Geostatistical GP model diagnostics

plots comparing predicted and observed means and standard deviations. Predicted means are computed as $\hat{\mu}_i = \int \tilde{y}_i p(\tilde{y}_i \mid \theta) d\tilde{y}_i$ and are plotted against the observed means $\hat{\mu}_{obs}$. Predicted standard deviations are calculated as $\hat{\sigma}_i = \text{SD}(\tilde{y}_i \mid y)$ and compared with $\hat{\sigma}_{obs}$.

In both models, the density plots show that the predicted data distributions align well with the observed data, suggesting that each model can adequately reproduce the overall shape of the strength distribution. However, the plots comparing the predicted and observed means reveal systematic deviations, pointing to potential limitations in the current model structures. The standard deviation plot for the ICAR model indicates good agreement between the predicted and observed variability. In contrast, the GP model tends to overestimate the standard deviation, as the replicated data exhibit greater variability than observed in the original dataset. These discrepancies suggest areas where model improvements are needed. One possible direction for refinement is to incorporate stochastic, rather than deterministic, stress calculations to better capture the physical stress distribution.

Model Comparison

We use Leave-One-Out Cross-Validation (LOO-CV) to compare the two models. LOO-CV is a model validation technique used to estimate out-of-sample predictive accuracy of Bayesian models. Given a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$, the LOO-CV estimate of model fit is based on the leave-one-out predictive density:

$$p(y_i|\mathcal{D}_{-i}) = \int p(y_i|\theta) p(\theta|\mathcal{D}_{-i}) d\theta \tag{4.29}$$

where $p(y_i|\theta)$ is the likelihood of y_i given parameters θ, $p(\theta|\mathcal{D}_{-i})$ is the posterior distribution of θ given the dataset excluding (x_i, y_i). The LOO log predictive density (ELPD) is then:

$$\text{ELPD}_{\text{LOO}} = \sum_{i=1}^{N} \log p(y_i|\mathcal{D}_{-i}), \tag{4.30}$$

which measures how well the model predicts unseen data points. However, this requires refitting the model N times and is computationally expensive. To address this, approximate leave-one-out predictive density can be computed using importance sampling (Gelfand et al., 1992; Gelfand, 1995):

$$p(y_i|\mathcal{D}_{-i}) \approx \frac{\sum_{s=1}^{S} w_i^{(s)} p(y_i|\theta^{(s)})}{\sum_{s=1}^{S} w_i^{(s)}}, \tag{4.31}$$

where $\theta^{(s)}$ are posterior samples from $p(\theta|\mathcal{D})$, and the importance weights are $w_i^{(s)} = \frac{1}{p(y_i|\theta^{(s)})}$. However, standard importance sampling can lead to unstable estimates when individual weights vary greatly. To stabilize importance weights, Pareto-smoothed importance sampling (PSIS) applies a transformation to reduce extreme weight values (Vehtari et al., 2017):

$$\widehat{\text{ELPD}}_{\text{loo}} = \sum_{i=1}^{n} \log \left(\frac{1}{S} \sum_{s=1}^{S} w_i^s p(y_i \mid \theta^s) \right), \tag{4.32}$$

where S is the number of posterior draws, w_i^s are the stabilized importance weights, $p(y_i \mid \theta^s)$ is the predictive density for observation y_i given posterior sample θ^s.

	$\widehat{\text{ELPD}}_{\text{diff}}$	$\text{SE}(\widehat{\text{ELPD}}_{\text{diff}})$	$\widehat{\text{ELPD}}_{\text{loo}}$	$\text{SE}(\widehat{\text{ELPD}}_{\text{loo}})$
Hierarchical ICAR model	0.00	0.00	-642.32	16.83
Geostatistical GP model	-75.55	13.02	-717.87	18.32

Table 4.8: Comparison of model fit using PSIS-LOO.

We implemented PSIS-LOO using the loo package (Vehtari et al., 2020), and the results are presented in Table 4.8. To compare models, $\widehat{\text{ELPD}}_{\text{diff}}$ is calculated to represent the difference in $\widehat{\text{ELPD}}_{\text{loo}}$ between the two models. The associated $\text{SE}(\widehat{\text{ELPD}}_{\text{diff}})$ measures the uncertainty in this comparison. The comparison indicate that the hierarchical ICAR model provides a better fit to the data, with a higher $\widehat{\text{ELPD}}_{\text{loo}}$ of –642.32 compared to –717.87 for the geostatistical GP model. The estimated difference, $\widehat{\text{ELPD}}_{\text{diff}} = -75.55$, with a standard error of 13.02, suggests a significant difference in predictive performance between the two models. These results favor the hierarchical ICAR model for this application.

4.5.2 A mixture of tensile and shear stress model

In this section, we present the results obtained from the geostatistical GP mixture model for analyzing both the tensile and shear strengths of crossarms. The priors used in this application are consistent with those used in the simulation study. To ensure that shear failures occur only in zone 1, we incorporate an additional condition into the density function used in Algorithm 1. Specifically, we modify the density to be close to zero when the failure zone is 1 and $Z_i = 1$, thereby enforcing the constraint that shear failures are restricted to zone 1.

Posterior summaries for the mixture GP model are presented in Table 4.9. The parameter $\boldsymbol{\pi}$ indicates tensile failure accounts for around 88% pieces while shear failure accounts

for around 12% pieces. The posterior mean of μ_1 suggests that the baseline tensile strength, before accounting for knot effects, is approximately 11.83 ksi. The estimated knot effect on tensile strength, represented by β_1 and γ_1, indicates a mean exponential decay at a rate of $0.89\exp(-0.16d)$, where d is the distance from a knot to the edge experiencing maximum stress. The posterior mean shear strength is 0.25 ksi, before accounting for knot effects. The posterior mean knot effect on shear strength, represented by β_2 and γ_2, is $0.02\exp(-0.53d)$. Finally, the estimated variation, represented by $\boldsymbol{\alpha}$, is 2.91 and 0.05, respectively.

Parameter	Mean	SD	2.5%	Median	97.5%
μ_1	11.8306	0.3074	11.2641	11.8318	12.4473
μ_2	0.2462	0.0105	0.2280	0.2457	0.2701
β_1	-0.8908	0.3644	-1.7212	-0.8660	-0.2567
β_2	-0.0193	0.0173	-0.0719	-0.0153	-0.0009
γ_1	0.1588	0.0706	0.0450	0.1536	0.3040
γ_2	0.5346	0.2656	0.1449	0.5002	1.2319
α_1	2.9061	0.1379	2.6441	2.9037	3.2341
α_2	0.0503	0.0064	0.0385	0.0502	0.0633
ρ_1	0.6920	0.4796	0.0252	0.6285	1.8793
ρ_2	0.8085	0.6089	0.0525	0.6697	2.1606
π_1	0.8844	0.0218	0.8400	0.8890	0.9275
π_2	0.1156	0.0218	0.0725	0.1110	0.1600

Table 4.9: Posterior mean, standard deviation, median, and 95% credible intervals for the parameters of the GP mixture model.

Model Diagnostics

The diagnostic plots for the mixture GP model is shown in Figure 4.6. The density plots suggest that the overall distribution of the predicted data aligns reasonably well with the observed data for the model. However, comparisons of predicted versus observed means and standard deviations reveal systematic deviations. While the deviation in the mean improved compared to that observed in the tensile stress model, the discrepancy in standard deviation remains substantial. These differences suggest potential shortcomings in the current model structure, which may not adequately capture the underlying variability. To address these issues, future work could explore the use of stochastic rather than deterministic stress calculations, which may yield a better model fit during the transformation from

strength components to the final MOR. Additionally, integrating larger and higher-quality datasets will be crucial for improving model accuracy and overall performance.

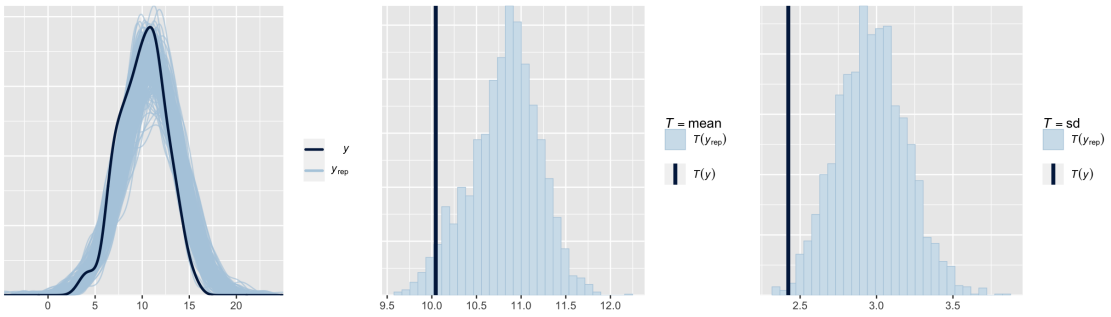


Figure 4.6: Mixture model diagnostics

4.6 Discussion

In this work, we developed three modeling approaches to analyze the effect of knot arrangements based on mechanical experiments of crossarms. Given that tensile stress is the most plausible failure mode in the experiment, we first proposed two Bayesian hierarchical models focused exclusively on tensile failure. These models combine mechanical analysis with Bayesian methods to better understand the effects of knots on structural failures. Both models incorporate knot effects and zone-specific heterogeneity, but they differ fundamentally in their hierarchical structure and treatment of spatial dependence. The ICAR model introduces spatial structure through zone-level random effects, while the GP model captures spatial variation by assigning a Gaussian process to each realization of the spatial field. Simulation studies demonstrate that both models perform well in recovering the fixed effect parameters, although they show greater uncertainty in spatial smoothness parameters.

We applied both models to experimental data, and the results reveal several differences between the GP and ICAR approaches. While both models capture a similar exponential decay in strength due to the presence of knots, they differ in their estimates of baseline strength and spatial parameters. Model diagnostics indicate that the hierarchical ICAR model provides better predictive performance, as reflected by higher PSIS-LOO values. In terms of goodness-of-fit, posterior predictive density plots show good agreement between the predicted and observed data for both models. However, posterior predictive means exhibit systematic deviations, suggesting that the current model structures may not fully

capture all aspects of the underlying process. This means that while the proposed models are helpful for understanding the baseline tensile strength and effect of knot arrangements, there's room for improvement on model specifications. One possible way is to incorporate stochastic terms for tensile stress to better capture the complexity of crossarm behavior in addition to modeling tensile strengths.

To address the fact that crossarms can fail under different types of mechanical stress, we developed a finite mixture model to analyze the underlying failure mechanisms of test specimens. This model introduces latent variables to represent the primary cause of failure (tensile, compressive, or shear stress), while incorporating fixed effects to account for the influence of knots, along with a Gaussian process to capture spatial dependence across zones. Simulation studies show that the model can recover the true parameter values with reasonable accuracy, even in the presence of complex dependencies and small sample sizes. As with the tensile-stress-only models, estimating spatial smoothness parameters remains challenging. When applied to the experimental data, the posterior results indicate that tensile failure is the dominant mode, accounting for approximately 88% of failures, while shear failure accounts for around 12%. The model differentiates the impact of knots on tensile and shear strength using separate fixed effect terms. Although the model performs reasonably well overall, diagnostic checks reveal discrepancies in predicted means and variances, highlighting areas where further model refinement may be needed.

Future work could explore more flexible modeling approaches to better capture the complexities of wood failure. For instance, rather than relying on deterministic stress calculations for each zone, incorporating stochastic stress fields could better reflect underlying uncertainties and improve the model's ability to match observed data. This may lead to more accurate predictions of MOR by comparing stress with the corresponding strengths. The model could also be extended to incorporate non-Gaussian distributions, such as the Weibull distribution, which may provide a better fit given that material strength values are strictly positive. Additionally, neural network-based methods could offer advantages for detecting wood defects and predicting wood properties, especially given the complexity of natural features like grain patterns and knots (Feng et al., 2024). Recent advances in neural networks can be a potentially powerful alternative to traditional statistical modeling approaches for prediction (Fathi et al., 2020; Wang et al., 2024). Nonetheless, accurately predicting wood performance remains a challenging task due to its natural variability and inherently heterogeneous properties. As a result, both experimental studies and advanced statistical modeling are essential to improve our understanding and predictive capabilities in this field.

4.7 Supplementary Materials

4.7.1 Stress analysis of zones

The maximum bending (tensile and compressive) stress for each zone is shown below:

$$\sigma_{bending_{1,3,4,6,7}} = \begin{cases} \frac{6P_y x}{bh^2} + \frac{6P_x y}{hb^2} & \text{if } x \leq 48, \\ \\ \frac{6(96P_y - P_y x)}{bh^2} + \frac{6(96P_x - P_x y)}{hb^2} & \text{if } x > 48. \end{cases} \tag{4.33}$$

$$\sigma_{bending_{2,5}} = \begin{cases} \frac{6P_y x}{hb^2} - \frac{6P_x y}{hb^2} & \text{if } x \leq 48, \\ \\ \frac{6(96P_y - P_y x)}{hb^2} - \frac{6(96P_x - P_x y)}{hb^2} & \text{if } x > 48. \end{cases} \tag{4.34}$$

In the above equations, P_y and P_x denote the applied load P decomposed in the width-wise (y) and height-wise axis (x) and can be obtained from σ (MOR) for each specimen as follows:

$$P_y = P \sin\theta = \frac{\sigma b^2 h^2 \sin\theta}{6l(b\sin\theta + h\cos\theta)}, \tag{4.35}$$

$$P_x = P \cos\theta = \frac{\sigma b^2 h^2 \cos\theta}{6l(b\sin\theta + h\cos\theta)}, \tag{4.36}$$

where $l = 96'', \theta = 17.5°, b = 3.5'', h = 4.5''$. The maximum shear stress for each zone is shown below:

$$\tau_{shear} = \frac{3P_y}{2bh} + \frac{3P_x}{2h^2}. \tag{4.37}$$

The maximum shear stress for zone 1 is adjusted for extra shear stress introduced by the pin holes.

4.7.2 Sample pieces of Douglas-fir crossarms

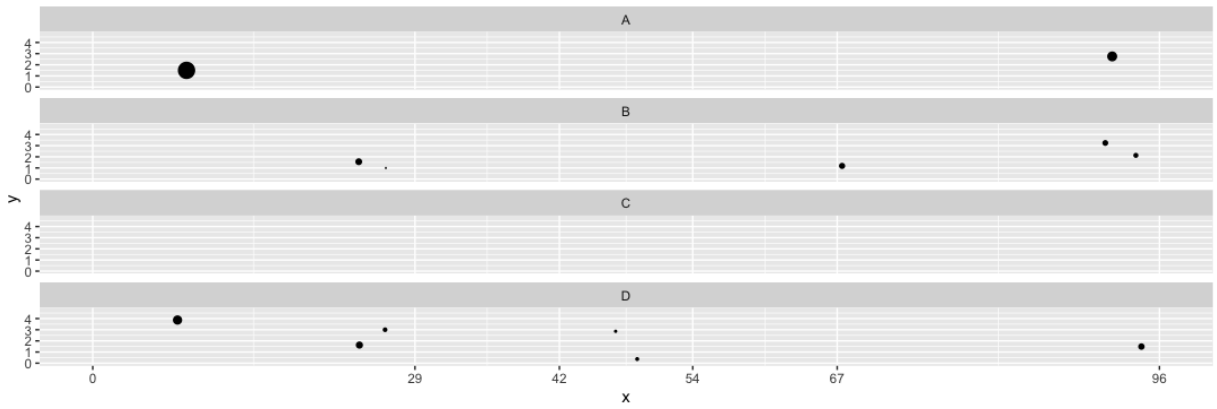


Figure 4.7: A sample of the Douglas-fir crossarms sent to the test. Facets A and C represent the narrow face of the crossarm, and B and D represent the wide face of the crossarm. Black circles representing knots were regenerated based on the diameter and the location of the knots documented in the data.

Chapter 5

Discussion and Future Work

In this chapter, we provide a summary for each of the preceding chapters. In addition, we provide some potential extensions and future work before closing the thesis.

In Chapter 2, we developed a Bayesian hierarchical model to study the spatial and temporal trend of S-protein sequences in SARS-CoV-2, after grouping them into mutant clusters. Instead of the sequence counts in each country, we focus inference on the cluster proportions. Through a log-ratio transformation of the cluster proportions, the model can quantify the differences in prevalence among the S-protein mutant clusters through the growth rate parameter for each cluster and the dependence across countries via correlated error terms. Based on our model estimates, we found that on average the reference sequence and its closely-related variants will diminish, while variants with double mutations of D614G + S477N or D614G + A222V tend to increase most strongly in prevalence over time, suggesting potentially higher infectiousness for these double mutations. The model is also able to make projections of cluster proportions in countries with missing data, thereby providing early warnings of potentially increasing mutations. Although the paper was published before the emergence of the Delta variant in 2021, we are among the few early publications that found the potential fitness advantages of double mutations compared to the reference sequence and single mutation. Since the acute phase of the Covid-19 pandemic has concluded, we do not expect to further extend the methodology.

In Chapter 3, we proposed a Bayesian hierarchical model to address the challenge of spatial misalignment in spatio-temporal data obtained from in situ and satellite sources. The model was fit using the INLA-SPDE approach, which provides efficient computation. Our methodology combines the different data sources in a "fusion" model via the construction of projection matrices in both spatial and temporal domains. Through simulation studies,

we demonstrated that the fusion model has superior performance in prediction accuracy across space and time compared to standalone "in situ" and "satellite" models based on only in situ or satellite data, respectively. The fusion model also generally outperforms the standalone models in terms of parameter inference. Such a modeling approach is motivated by environmental problems, and our specific focus is on the analysis and prediction of harmful algae bloom (HAB) events, where the convention is to conduct separate analyses based on either in situ samples or satellite images. A real data analysis showed that the proposed model is a necessary step towards a unified characterization of bloom dynamics and identifying the key drivers of HAB events.

Future work could extend the alignment to address temporal misalignment between data sources. This allows satellite data—often daily or hourly averages—to be expressed as smoothed integrals over time and space. Alignment in time is especially important for variables with high variability, such as temperature. Another direction is extending to non-separable space-time models when separability is violated. Recent approaches, such as the diffusion-based spatio-temporal stochastic processes, offer flexible and computationally efficient non-separable frameworks compatible with INLA. Further research may consider extensions to non-stationary, non-isotropic, or non-Gaussian data, though these remain computationally challenging.

In Chapter 4, we developed three Bayesian modeling approaches to analyze the effects of knot arrangements on the mechanical failure of crossarms. The models incorporated mechanical analysis and accounted for the presence of knots through hierarchical Bayesian frameworks. The first two models focused exclusively on tensile failure, differing primarily in their hierarchical structure and how they represent spatial correlation: the ICAR model used a discrete spatial structure with zone-level random effects, while the GP model employed a continuous spatial correlation through a Gaussian process prior. When applied to the experimental data, both models produced similar estimates for the knot effect parameters. Model diagnostics indicated that the ICAR model outperformed the GP model in both goodness-of-fit and predictive performance. To extend the analysis beyond tensile failure, we proposed a finite mixture model to handle multiple failure types—tensile, compressive, and shear. When fitted to real data, the model identified tensile failure as the predominant mode, while also distinguishing the influence of knots across the two different failure mechanisms.

Despite these promising results, several limitations remain, suggesting avenues for future work. First, both the tensile-only and mixture models showed systematic discrepancies between posterior predictive means and observed values, suggesting that the models do not fully capture the complexity of wood behavior. One improvement would be to incorporate stochastic stress fields rather than rely on deterministic stress estimates, as this would allow

better representation of local uncertainties in material stress and strength. Given the natural complexity of wood, including features such as grain patterns and knots, deep learning approaches may provide greater flexibility and predictive power compared to traditional statistical models. Continued methodological advances and integration of experimental data will be essential to improve predictive modeling in this domain.

References

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):139–160.

Aitchison, J. (1999). Logratios and natural laws in compositional data analysis. *Mathematical Geology*, 31(5):563–580.

Amanat, F. and Krammer, F. (2020). Sars-cov-2 vaccines: status report. *Immunity*, 52(4):583–589.

Anderson, C. H., Sinha, A., Konkler, M. J., and Morrell, J. J. (2021). Ability to predict flexural properties of douglas-fir crossarms. *Wood Material Science & Engineering*, 16(6):366–374.

Anderson, D. M. (2009). Approaches to monitoring, control and management of harmful algal blooms (habs). *Ocean & coastal management*, 52(7):342–347.

Aswi, A., Cramb, S., Moraga, P., and Mengersen, K. (2019). Bayesian spatial and spatio-temporal approaches to modelling dengue fever: a systematic review. *Epidemiology & Infection*, 147.

Auchincloss, A. H., Gebreab, S. Y., Mair, C., and Diez Roux, A. V. (2012). A review of spatial methods in epidemiology, 2000–2010. *Annual review of public health*, 33:107–122.

Banerjee, A. K., Begum, F., and Ray, U. (2020). Mutation hot spots in spike protein of covid-19. *Preprints*, 2020:2020040281.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2003). *Hierarchical modeling and analysis for spatial data*. Chapman and Hall/CRC.

Baño, V., Arriaga, F., Soilán, A., and Guaita, M. (2011). Prediction of bending load capacity of timber beams using a finite element method simulation of knots and grain deviation. *Biosystems engineering*, 109(4):241–249.

Baño, V., Arriaga, F., and Guaita, M. (2013). Determination of the influence of size and position of knots on load capacity and stress distribution in timber beams of pinus sylvestris using finite element model. *Biosystems engineering*, 114(3):214–222.

Beale, L., Abellan, J. J., Hodgson, S., and Jarup, L. (2008). Methodologic issues and approaches to spatial epidemiology. *Environmental health perspectives*, 116(8):1105–1110.

Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., and Tasumi, M. (1977). The protein data bank. *European Journal of Biochemistry*, 80(2):319–324.

Berrocal, V. J., Gelfand, A. E., and Holland, D. M. (2010). A spatio-temporal downscaler for output from numerical models. *Journal of agricultural, biological, and environmental statistics*, 15(2):176.

Berrocal, V. J., Gelfand, A. E., and Holland, D. M. (2012). Space-time data fusion under error in computer model output: an application to modeling air quality. *Biometrics*, 68(3):837–848.

Bertani, I., Obenour, D. R., Steger, C. E., Stow, C. A., Gronewold, A. D., and Scavia, D. (2016). Probabilistically assessing the role of nutrient loading in harmful algal bloom formation in western lake erie. *Journal of Great Lakes Research*, 42(6):1184–1192.

Bertani, I., Steger, C. E., Obenour, D. R., Fahnenstiel, G. L., Bridgeman, T. B., Johengen, T. H., Sayers, M. J., Shuchman, R. A., and Scavia, D. (2017). Tracking cyanobacteria blooms: Do different monitoring approaches tell the same story? *Science of the Total Environment*, 575:294–308.

Besag, J., York, J., and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics*, 43(1):1–20.

Bisanzio, D., Giacobini, M., Bertolotti, L., Mosca, A., Balbo, L., Kitron, U., and Vazquez-Prokopec, G. M. (2011). Spatio-temporal patterns of distribution of west nile virus vectors in eastern piedmont region, italy. *Parasites & vectors*, 4(1):1–11.

Blangiardo, M. and Cameletti, M. (2015). *Spatial and spatio-temporal Bayesian models with R-INLA*. John Wiley & Sons.

Blangiardo, M., Cameletti, M., Baio, G., and Rue, H. (2013). Spatial and spatio-temporal models with r-inla. *Spatial and spatio-temporal epidemiology*, 7:39–55.

Blois, J. L., Williams, J. W., Fitzpatrick, M. C., Jackson, S. T., and Ferrier, S. (2013). Space can substitute for time in predicting climate-change effects on biodiversity. *Proceedings of the National Academy of Sciences*, 110(23):9374–9379.

Boresi, A. P. and Schmidt, R. J. (2003). *Advanced Mechanics of Materials.* Wiley.

Botev, Z. I., Grotowski, J. F., Kroese, D. P., et al. (2010). Kernel density estimation via diffusion. *The Annals of Statistics*, 38(5):2916–2957.

Bruno, F., Cameletti, M., Franco-Villoria, M., Greco, F., Ignaccolo, R., Ippoliti, L., Valentini, P., and Ventrucci, M. (2016). A survey on ecological regression for health hazard associated with air pollution. *Spatial statistics*, 18:276–299.

Brunsdon, C., Fotheringham, S., and Charlton, M. (1998). Geographically weighted regression. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 47(3):431–443.

Bullerjahn, G. S., McKay, R. M., Davis, T. W., Baker, D. B., Boyer, G. L., D'Anglada, L. V., Doucette, G. J., Ho, J. C., Irwin, E. G., Kling, C. L., et al. (2016). Global solutions to regional problems: Collecting global expertise to address the problem of harmful cyanobacterial blooms. a lake erie case study. *Harmful Algae*, 54:223–238.

Cameletti, M., Ignaccolo, R., and Bande, S. (2011). Comparing spatio-temporal models for particulate matter in piemonte. *Environmetrics*, 22(8):985–996.

Cameletti, M., Lindgren, F., Simpson, D., and Rue, H. (2013). Spatio-temporal modeling of particulate matter concentration through the spde approach. *AStA Advances in Statistical Analysis*, 97(2):109–131.

Cappello, C., De Iaco, S., and Posa, D. (2020). covatest: An R Package for Selecting a Class of Space-Time Covariance Functions. *Journal of Statistical Software*, 94(1):1–42.

Cargnoni, C., Müller, P., and West, M. (1997). Bayesian forecasting of multinomial time series through conditionally gaussian dynamic models. *Journal of the American Statistical Association*, 92(438):640–647.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of statistical software*, 76(1):1–32.

Cha, Y., Park, S. S., Kim, K., Byeon, M., and Stow, C. A. (2014). Probabilistic prediction of cyanobacteria abundance in a korean reservoir using a bayesian poisson model. *Water Resources Research*, 50(3):2518–2532.

Chen, A. T., Altschuler, K., Zhan, S. H., Chan, Y. A., and Deverman, B. E. (2020a). COVID-19 CG: Tracking SARS-CoV-2 mutations by locations and dates of interest. *bioRxiv Prepr. Serv. Biol.* bioRxiv:2020.09.23.310565.

Chen, J., Wang, R., Wang, M., and Wei, G.-W. (2020b). Mutations strengthened sars-cov-2 infectivity. *Journal of Molecular Biology*, 432(19):5212–5226.

Choi, J., Fuentes, M., and Reich, B. J. (2009). Spatial–temporal association between fine particulate matter and daily mortality. *Computational statistics & data analysis*, 53(8):2989–3000.

Chorus, I. and Welker, M. (2021). *Toxic Cyanobacteria in Water: A Guide to Their Public Health Consequences, Monitoring and Management*. CRC Press.

Clarotto, L., Allard, D., Romary, T., and Desassis, N. (2024). The SPDE approach for spatio-temporal datasets with advection and diffusion. *Spatial Statistics*, 62:100847.

Clayton, D. G. (1996). Generalized linear mixed models. *Markov chain Monte Carlo in practice*, 1:275–302.

Cocchi, D., Greco, F., and Trivisano, C. (2007). Hierarchical space-time modelling of pm10 pollution. *Atmospheric environment*, 41(3):532–542.

Courtney, T. H. (2005). *Mechanical behavior of materials*. Waveland Press.

Craig, J., Roy, R., and Taleff, E. M. (2020). *Mechanics of materials*. John Wiley & Sons.

Cramer, S., Fohrell, W., Mcdonald, K., and Stahl, D. (1988). Exploring the relationship between local slope of grain angle and initial fracture in lumber subject to tensile load. In *Proceedings of 1988 International conference on timber engineering, Seattle*, volume 2, pages 566–575.

Cramer, S. and Goodman, J. (1983). Model for stress analysis and strength prediction of lumber. *Wood and Fiber Science*, pages 338–349.

Cressie, N. (2015). *Statistics for spatial data*. John Wiley & Sons.

Derado, G., Bowman, F. D., Zhang, L., and Initiative, A. D. N. (2013). Predicting brain activity using a bayesian spatial model. *Statistical methods in medical research*, 22(4):382–397.

Diehl, W. E., Lin, A. E., Grubaugh, N. D., Carvalho, L. M., Kim, K., Kyawe, P. P., et al. (2016). Ebola virus glycoprotein with increased infectivity dominated the 2013–2016 epidemic. *Cell*, 167(4):1088–1098.

Divos, F. and Tanaka, T. (1997). Lumber strength estimation by multiple regression. *Holzforschung*, 51(5):467–471.

Dong, E., Du, H., and Gardner, L. (2020). An interactive web-based dashboard to track covid-19 in real time. *The Lancet infectious diseases*, 20(5):533–534.

Duffy, S. (2018). Why are rna virus mutation rates so damn high? *PLoS biology*, 16(8):e3000003.

European Commission (2020). Coronavirus: Commission proposes more clarity and predictability of any measures restricting free movement in the european union. https://ec.europa.eu/commission/presscorner/detail/en/ip_20_1555. Last checked on Dec 20, 2020.

Fan, S., Wong, S. W., and Zidek, J. V. (2023). Knots and their effect on the tensile strength of lumber: A case study. *Journal of Quality Technology*, 55(4):510–522.

Fang, S., Del Giudice, D., Scavia, D., Binding, C. E., Bridgeman, T. B., Chaffin, J. D., Evans, M. A., Guinness, J., Johengen, T. H., and Obenour, D. R. (2019). A space-time geostatistical model for probabilistic estimation of harmful algal bloom biomass and areal extent. *Science of the Total Environment*, 695:133776.

Fathi, H., Nasir, V., and Kazemirad, S. (2020). Prediction of the mechanical properties of wood using guided wave propagation and machine learning. *Construction and Building Materials*, 262:120848.

Feng, Y., Mekhilef, S., Hui, D., Chow, C. L., and Lau, D. (2024). Machine learning-assisted wood materials: Applications and future prospects. *Extreme Mechanics Letters*, 71:102209.

Figley, W., Pyle, B., and Halgren, B. (1976). Socioeconomic impacts. *Oxygen depletion and associated benthic mortalities in New York Bight*, pages 315–322.

Fink, G. and Kohler, J. (2014). Model for the prediction of the tensile strength and tensile stiffness of knot clusters within structural timber. *European Journal of Wood and Wood Products*, 72(3):331–341.

Fonseca, V. P., Pennino, M. G., de Nóbrega, M. F., Oliveira, J. E. L., and de Figueiredo Mendes, L. (2017). Identifying fish diversity hot-spots in data-poor situations. *Marine environmental research*, 129:365–373.

Fotheringham, A. S., Brunsdon, C., and Charlton, M. (2003). *Geographically weighted regression: the analysis of spatially varying relationships*. John Wiley & Sons.

França, F., Seale, R. D., Shmulsky, R., and França, T. (2018). Modeling mechanical properties of 2 by 4 and 2 by 6 southern pine lumber using longitudinal vibration and visual characteristics. *Forest Products Journal*, 68(3):286–294.

Fuentes, M. and Raftery, A. E. (2005). Model evaluation and spatial interpolation by bayesian combination of observations with outputs from numerical models. *Biometrics*, 61(1):36–45.

Fuentes, M., Song, H.-R., Ghosh, S. K., Holland, D. M., and Davis, J. M. (2006). Spatial association between speciated fine particles and mortality. *Biometrics*, 62(3):855–863.

Gelfand, A. E. (1995). Model determination using. *Markov chain Monte Carlo in practice*, page 145.

Gelfand, A. E., Dey, D. K., and Chang, H. (1992). Model determination using predictive distributions with implementation via sampling-based methods. *Bayesian statistics 4*, pages 147–168.

Gelfand, A. E., Kim, H.-J., Sirmans, C., and Banerjee, S. (2003). Spatial modeling with spatially varying coefficient processes. *Journal of the American Statistical Association*, 98(462):387–396.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. CRC Press.

Gómez-Rubio, V. (2020). *Bayesian inference with INLA*. CRC Press.

Gomez-Rubio, V., Cameletti, M., and Blangiardo, M. (2022). Missing data analysis and imputation via latent gaussian markov random fields. *SORT-Statistics and Operations Research Transactions*, 46(2):217–244.

Goodman, J. R., Vanderbilt, M. D., and Criswell, M. E. (1983). Reliability-based design of wood transmission line structures. *Journal of Structural Engineering*, 109(3):690–704.

Goossens, M., Mittelbach, F., and Samarin, A. (1994). *The LaTeX Companion*. Addison-Wesley, Reading, Massachusetts.

Gotway, C. A. and Young, L. J. (2002). Combining incompatible spatial data. *Journal of the American Statistical Association*, 97(458):632–648.

Gotway, C. A. and Young, L. J. (2007). A geostatistical approach to linking geographically aggregated data from different sources. *Journal of Computational and Graphical Statistics*, 16(1):115–135.

Griffith, D. A. and Peres-Neto, P. R. (2006). Spatial modeling in ecology: the flexibility of eigenfunction spatial analyses. *Ecology*, 87(10):2603–2613.

Hallegraeff, G. (2003). Harmful algal blooms: a global overview. *Manual on harmful marine microalgae*, 33:1–22.

Hallegraeff, G. M. (1993). A review of harmful algal blooms and their apparent global increase. *Phycologia*, 32(2):79–99.

Harrison, P. J. and Stevens, C. F. (1976). Bayesian forecasting. *Journal of the Royal Statistical Society: Series B (Methodological)*, 38(3):205–228.

Hastings, W. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109.

Hauer, F. R. and Lamberti, G. A. (2011). *Methods in stream ecology*. Academic Press.

He, S. and Wong, S. W. (2021). Statistical challenges in the analysis of sequence and structure data for the covid-19 spike protein. *Journal of Data Science*, 19(2).

He, S. and Wong, S. W. (2024). Spatio-temporal data fusion for the analysis of in situ and remote sensing data using the inla-spde approach. *Spatial Statistics*, 64:100863.

Herdendorf, C. and Braidech, L. (1972). A study of the physical characteristics of the major reef areas in western lake erie. *Ohio Dept of Natural Resources, Div of Geological Survey, Project AFCS-1*.

Herdendorf, C. E. (1990). Distribution of the world's large lakes. In *Large lakes*, pages 3–38. Springer.

Herdendorf, C. E. (1992). Lake erie coastal wetlands: an overview. *Journal of Great Lakes Research*, 18(4):533–551.

Herfst, S., Schrauwen, E. J., Linster, M., Chutinimitkul, S., de Wit, E., Munster, V. J., et al. (2012). Airborne transmission of influenza a/h5n1 virus between ferrets. *Science*, 336(6088):1534–1541.

Ho, J. C. and Michalak, A. M. (2015). Challenges in tracking harmful algal blooms: A synthesis of evidence from lake erie. *Journal of Great Lakes Research*, 41(2):317–325.

Hodcroft, E. B., Zuber, M., Nadeau, S., Comas, I., Candelas, F. G., Stadler, T., and Neher, R. A. (2020). Emergence and spread of a SARS-CoV-2 variant through Europe in the summer of 2020. *medRxiv*. medRxiv:2020.10.25.20219063.

Hooten, M. B. and Wikle, C. K. (2010). Statistical agent-based models for discrete spatio-temporal systems. *Journal of the American Statistical Association*, 105(489):236–248.

Huddleston, J., Barnes, J. R., Rowe, T., Xu, X., Kondor, R., Wentworth, D. E., Whittaker, L., Ermetal, B., Daniels, R. S., McCauley, J. W., et al. (2020). Integrating genotypes and phenotypes improves long-term forecasts of seasonal influenza a/h3n2 evolution. *Elife*, 9:e60067.

Jurek, M. and Katzfuss, M. (2021). Multi-resolution filters for massive spatio-temporal data. *Journal of Computational and Graphical Statistics*, 30(4):1095–1110.

Kandler, G., Füssl, J., and Eberhardsteiner, J. (2015). Stochastic finite element approaches for wood-based products: theoretical framework and review of methods. *Wood science and technology*, 49(5):1055–1097.

Knorr-Held, L. (2000). Bayesian modelling of inseparable space-time variation in disease risk. *Statistics in medicine*, 19(17-18):2555–2567.

Knuth, D. (1986). *The TEXbook*. Addison-Wesley, Reading, Massachusetts.

Korber, B., Fischer, W. M., Gnanakaran, S., Yoon, H., Theiler, J., Abfalterer, W., Hengartner, N., Giorgi, E. E., Bhattacharya, T., Foley, B., et al. (2020). Tracking changes in sars-cov-2 spike: evidence that d614g increases infectivity of the covid-19 virus. *Cell*, 182(4):812–827.

Krammer, F. (2020). Sars-cov-2 vaccines in development. *Nature*, 586(7830):516–527.

Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K., and Moult, J. (2019). Critical assessment of methods of protein structure prediction (casp)—round xiii. *Proteins: Structure, Function, and Bioinformatics*, 87(12):1011–1020.

Kutser, T., Metsamaa, L., Vahtmäe, E., and Strömbeck, N. (2006). Suitability of modis 250 m resolution band data for quantitative mapping of cyanobacterial blooms. *Proc. Estonian Acad. Sci. Biol. Ecol*, 55(4):318œ328.

Lam, F. and Varoglu, E. (1991). Variation of tensile strength along the length of lumber: Part 1: Experimental. *Wood science and technology*, 25(5):351–359.

Lamport, L. (1994). *LaTeX — A Document Preparation System*. Addison-Wesley, Reading, Massachusetts, second edition.

Lauer, S. A., Grantz, K. H., Bi, Q., Jones, F. K., Zheng, Q., Meredith, H. R., Azman, A. S., Reich, N. G., and Lessler, J. (2020). The incubation period of coronavirus disease 2019 (covid-19) from publicly reported confirmed cases: estimation and application. *Annals of internal medicine*, 172(9):577–582.

Lauring, A. S. and Andino, R. (2010). Quasispecies theory and the behavior of rna viruses. *PLoS Pathog*, 6(7):e1001005.

Lawson, A. B. (2012). Bayesian point event modeling in spatial and environmental epidemiology. *Statistical Methods in Medical Research*, 21(5):509–529.

Leeds, W. B., Wikle, C. K., and Fiechter, J. (2014). Emulator-assisted reduced-rank ecological data assimilation for nonlinear multivariate dynamical spatio-temporal processes. *Statistical Methodology*, 17:126–138.

Li, Q., Wu, J., Nie, J., Zhang, L., Hao, H., Liu, S., et al. (2020). The impact of mutations in sars-cov-2 spike on viral infectivity and antigenicity. *Cell*, 182(5):1284–1294.

Liebel, S. A. and Mueller, R. E. (1994). Douglas fir crossarms: solid sawn vs. laminated comparison. In *Proceedings of IEEE/PES Transmission and Distribution Conference*, pages 581–586. IEEE.

Lindgren, F., Bakka, H., Bolin, D., Krainski, E., and Rue, H. (2024). A diffusion-based spatio-temporal extension of gaussian matérn fields. *SORT-Statistics and Operations Research Transactions*, 48:3–66.

Lindgren, F., Rue, H., and Lindström, J. (2011). An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498.

Little, R. J. and Rubin, D. B. (2019). *Statistical analysis with missing data*, volume 793. John Wiley & Sons.

Lopez, C., Jewett, E., Dortch, Q., Walton, B., and Hudnell, H. (2008). Scientific assessment of freshwater harmful algal blooms. Technical report, Interagency Working Group on Harmful Algal Blooms, Hypoxia, and Human Health.

M. Bulleit, W. and A. Chapman, R. (2004). Characterization of the correlation structure of lumber strength properties. *Wood Science and Technology*, 38(4):285–296.

Ma, P. and Kang, E. L. (2020). Spatio-temporal data fusion for massive sea surface temperature data from modis and amsr-e instruments. *Environmetrics*, 31(2):e2594.

Manning, N. F., Wang, Y.-C., Long, C. M., Bertani, I., Sayers, M. J., Bosse, K. R., Shuchman, R. A., and Scavia, D. (2019). Extending the forecast model: Predicting western lake erie harmful algal blooms at multiple spatial scales. *Journal of Great Lakes Research*, 45(3):587–595.

Marks, C., Nowak, J., Klostermann, S., Georges, G., Dunbar, J., Shi, J., Kelm, S., and Deane, C. M. (2017). Sphinx: merging knowledge-based and ab initio approaches to improve protein loop prediction. *Bioinformatics*, 33(9):1346–1353.

Martin, G. M., Frazier, D. T., and Robert, C. P. (2024). Approximating bayes in the 21st century. *Statistical Science*, 39(1):20–45.

McDermott, P. L. and Wikle, C. K. (2016). A model-based approach for analog spatio-temporal dynamic forecasting. *Environmetrics*, 27(2):70–82.

McDermott, P. L. and Wikle, C. K. (2017). An ensemble quadratic echo state network for non-linear spatio-temporal forecasting. *Stat*, 6(1):315–330.

McDermott, P. L. and Wikle, C. K. (2019a). Bayesian recurrent neural network models for forecasting and quantifying uncertainty in spatial-temporal data. *Entropy*, 21(2):184.

McDermott, P. L. and Wikle, C. K. (2019b). Deep echo state networks with uncertainty quantification for spatio-temporal forecasting. *Environmetrics*, 30(3):e2553.

McMillan, N. J., Holland, D. M., Morara, M., and Feng, J. (2010). Combining numerical model output and particulate data using bayesian space–time modeling. *Environmetrics: The official journal of the International Environmetrics Society*, 21(1):48–65.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092.

Michalak, A. M., Anderson, E. J., Beletsky, D., Boland, S., Bosch, N. S., Bridgeman, T. B., Chaffin, J. D., Cho, K., Confesor, R., Daloğlu, I., et al. (2013). Record-setting algal bloom in lake erie caused by agricultural and meteorological trends consistent with expected future conditions. *Proceedings of the National Academy of Sciences*, 110(16):6448–6452.

Millie, D. F., Weckman, G. R., Fahnenstiel, G. L., Carrick, H. J., Ardjmand, E., Young, W. A., Sayers, M. J., and Shuchman, R. A. (2014). Using artificial intelligence for cyanohab niche modeling: discovery and visualization of microcystis–environmental associations within western lake erie. *Canadian journal of fisheries and aquatic sciences*, 71(11):1642–1654.

Moraga, P., Cramb, S. M., Mengersen, K. L., and Pagano, M. (2017). A geostatistical model for combined analysis of point-level and area-level data using inla and spde. *Spatial Statistics*, 21:27–41.

Mott, R. L. (2007). *Applied strength of materials*. CRC Press.

Muff, S., Riebler, A., Held, L., Rue, H., and Saner, P. (2015). Bayesian analysis of measurement error models using integrated nested laplace approximations. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 64(2):231–252.

NASA MODIS (2014). Modis chlorophyll-a concentration. https://modis.gsfc.nasa.gov/data/dataprod/chlor_a.php. Accessed: 2022/11/20.

Nguyen, H., Katzfuss, M., Cressie, N., and Braverman, A. (2014). Spatio-temporal data fusion for very large remote sensing datasets. *Technometrics*, 56(2):174–185.

Ning, T., Nie, J., Huang, W., Li, C., Li, X., Liu, Q., et al. (2019). Antigenic drift of influenza a (h7n9) virus hemagglutinin. *The Journal of infectious diseases*, 219(1):19–25.

North American Wood Pole Council (2020). Preserved wood utility poles and the environment.

Obenour, D. R., Gronewold, A. D., Stow, C. A., and Scavia, D. (2014). Using a b ayesian hierarchical model to improve l ake e rie cyanobacteria bloom forecasts. *Water Resources Research*, 50(10):7847–7860.

Oh, J.-K., Kim, K.-M., and Lee, J.-J. (2010). Use of adjacent knot data in predicting bending strength of dimension lumber by x-ray. *Wood and fiber science*, pages 10–20.

Onuchic, J. N., Luthey-Schulten, Z., and Wolynes, P. G. (1997). Theory of protein folding: the energy landscape perspective. *Annual review of physical chemistry*, 48(1):545–600.

Pace, R. K., Barry, R., Gilley, O. W., and Sirmans, C. (2000). A method for spatial–temporal forecasting with an application to real estate prices. *International Journal of Forecasting*, 16(2):229–246.

Patterson, T. A., Thomas, L., Wilcox, C., Ovaskainen, O., and Matthiopoulos, J. (2008). State–space models of individual animal movement. *Trends in ecology & evolution*, 23(2):87–94.

Phan, T. (2020). Novel coronavirus: From discovery to clinical diagnostics. *Infection, Genetics and Evolution,*, 79:104211.

Páez, A., Uchida, T., and Miyamoto, K. (2002). A general framework for estimation and inference of geographically weighted regression models: 1. location-specific kernel bandwidths and a test for locational heterogeneity. *Environment and Planning A: Economy and Space*, 34(4):733–754.

Pörtner et al. (2019). Ipcc special report on the ocean and cryosphere in a changing climate. https://www.ipcc.ch/srocc.

Riebler, A., Sørbye, S. H., Simpson, D., and Rue, H. (2016). An intuitive bayesian spatial model for disease mapping that accounts for scaling. *Statistical methods in medical research*, 25(4):1145–1165.

Rochester, G. (1938). *The strength of eastern Canadian spruce timbers in sizes shipped to the United Kingdom*. Department of Mines and Resources, Canada, Lands, Parks and Forests Branch.

Rue, H. and Held, L. (2005). *Gaussian Markov random fields: theory and applications*. CRC press.

Rue, H. and Martino, S. (2007). Approximate bayesian inference for hierarchical gaussian markov random field models. *Journal of statistical planning and inference*, 137(10):3177–3192.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, 71(2):319–392.

Sahu, S. K. (2012). Hierarchical bayesian models for space-time air pollution data. In *Handbook of Statistics*, volume 30, pages 477–495. Elsevier.

Sahu, S. K., Gelfand, A. E., and Holland, D. M. (2010). Fusing point and areal level space–time data with application to wet deposition. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 59(1):77–103.

Schaefer, C. and Rost, B. (2012). Predict impact of single amino acid change upon protein structure. *BMC genomics*, 13(S4):1–10.

Schrödle, B. and Held, L. (2011). Spatio-temporal disease mapping using inla. *Environmetrics*, 22(6):725–734.

Schrödle, B., Held, L., and Rue, H. (2012). Assessing the impact of a movement network on the spatiotemporal spread of infectious diseases. *Biometrics*, 68(3):736–744.

Sedova, M., Jaroszewski, L., Alisoltani, A., and Godzik, A. (2020). Coronavirus3d: 3d structural visualization of covid-19 genomic divergence. *Bioinformatics*, 36(15):4360–4362.

Smayda, T. J. (1990). " novel and nuisance phytoplankton blooms in the sea" evidence for a global epidemic. *Toxic marine phytoplankton*, pages 29–40.

Soto, C. S., Fasnacht, M., Zhu, J., Forrest, L., and Honig, B. (2008). Loop modeling: Sampling, filtering, and scoring. *Proteins: Structure, Function, and Bioinformatics*, 70(3):834–843.

Stroud, J. R., Stein, M. L., Lesht, B. M., Schwab, D. J., and Beletsky, D. (2010). An ensemble kalman filter and smoother for satellite data assimilation. *Journal of the american statistical association*, 105(491):978–990.

Stumpf, R., Culver, M., Tester, P., Tomlinson, M., Kirkpatrick, G., Pederson, B., Truby, E., Ransibrahmanakul, V., and Soracco, M. (2003). Monitoring karenia brevis blooms in the gulf of mexico using satellite ocean color imagery and other data. *Harmful Algae*, 2(2):147–160.

Stumpf, R. P., Davis, T. W., Wynne, T. T., Graham, J. L., Loftin, K. A., Johengen, T. H., Gossiaux, D., Palladino, D., and Burtner, A. (2016a). Challenges for mapping cyanotoxin patterns from remote sensing of cyanobacteria. *Harmful algae*, 54:160–173.

Stumpf, R. P., Johnson, L. T., Wynne, T. T., and Baker, D. B. (2016b). Forecasting annual cyanobacterial bloom biomass to inform management decisions in lake erie. *Journal of Great Lakes Research*, 42(6):1174–1183.

Stumpf, R. P., Wynne, T. T., Baker, D. B., and Fahnenstiel, G. L. (2012). Interannual variability of cyanobacterial blooms in lake erie. *PLoS ONE*, 7:e42444.

Tang, K., Zhang, J., and Liang, J. (2014). Fast protein loop sampling and structure prediction using distance-guided sequential chain-growth monte carlo method. *PLoS computational biology*, 10:e1003539.

Taylor, S. E. and Bender, D. A. (1991). Stochastic model for localized tensile strength and modulus of elasticity in lumber. *Wood and fiber science*, pages 501–519.

Tobler, W. R. (1970). A computer movie simulating urban growth in the detroit region. *Economic geography*, 46(sup1):234–240.

Toyoshima, Y., Nemoto, K., Matsumoto, S., Nakamura, Y., and Kiyotani, K. (2020). Sars-cov-2 genomic variations associated with mortality rate of covid-19. *Journal of human genetics*, 65:1075–1082.

UNESCO's Intergovernmental Oceanographic Commission (2021). An unprecedented analysis on global harmful algal blooms launched by ioc. https://ioc.unesco.org/index.php/news/unprecedented-analysis-global-harmful-algal-blooms-launched-ioc. Accessed: 2022/11/20.

United States and Canada (2012). The 2012 great lakes water quality agreement. https://www.canada.ca/en/environment-climate-change/corporate/international-affairs/partnerships-countries-regions/north-america/great-lakes-water-quality-agreement.html. Accessed: 2022/11/20.

USDA (2002). *RUS Bulletin 1724E-151*. Piscataway, NJ.

USDA (2003). Rus bulletin 1724e-152.

Van Niekerk, J., Krainski, E., Rustand, D., and Rue, H. (2023). A new avenue for bayesian inference with inla. *Computational Statistics & Data Analysis*, 181:107692.

Vehtari, A., Gabry, J., Magnusson, M., Yao, Y., Bürkner, P.-C., Paananen, T., and Gelman, A. (2020). loo: Efficient leave-one-out cross-validation and WAIC for bayesian models.

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and computing*, 27:1413–1432.

Villejo, S. J., Illian, J. B., and Swallow, B. (2023). Data fusion in a two-stage spatio-temporal model using the inla-spde approach. *Spatial Statistics*, 54:100744.

Wan, Y., Shang, J., Graham, R., Baric, R. S., and Li, F. (2020). Receptor recognition by the novel coronavirus from wuhan: an analysis based on decade-long structural studies of sars coronavirus. *Journal of virology*, 94(7):e00127–20.

Wang, B., Nong, S., Pan, L., You, G., Li, Z., Sun, J., and Shi, S. (2024). Machine learning-based non-destructive testing model for high precision and stable evaluation of mechanical properties in bamboo-wood composites. *European Journal of Wood and Wood Products*, 82(3):621–633.

Wang, C., Liu, Z., Chen, Z., Huang, X., Xu, M., He, T., and Zhang, Z. (2020a). The establishment of reference sequence for sars-cov-2 and variation analysis. *Journal of medical virology*, 92(6):667–674.

Wang, R., Hozumi, Y., Yin, C., and Wei, G. (2020b). Decoding sars-cov-2 transmission, evolution and ramification on covid-19 diagnosis, vaccine, and medicine. *Journal of Chemical Information and Modeling*.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244.

WCLIB (2022). Grading rules for west coast lumber. https://www.plib.org/. Accessed: 2022/12/10.

Webb, B. and Sali, A. (2017). Protein structure modeling with modeller. In *Functional Genomics*, pages 39–54. Springer.

Wells, M. L., Karlson, B., Wulff, A., Kudela, R., Trick, C., Asnaghi, V., Berdalet, E., Cochlan, W., Davidson, K., De Rijcke, M., et al. (2020). Future hab science: Directions and challenges in a changing climate. *Harmful algae*, 91:101632.

WHO (2020a). Coronavirus disease (covid-19) situation dashboard. https://who.sprinklr.com/. Last checked on Dec 19, 2020.

WHO (2020b). Draft landscape of covid-19 candidate vaccines. https://www.who.int/publications/m/item/draft-landscape-of-covid-19-candidate-vaccines. Last checked on Dec 20, 2020.

Wikle, C. K. (2019). Comparison of deep neural networks and deep hierarchical models for spatio-temporal data. *Journal of Agricultural, Biological and Environmental Statistics*, 24(2):175–203.

Wikle, C. K., Zammit-Mangion, A., and Cressie, N. (2019). *Spatio-temporal Statistics with R*. Chapman and Hall/CRC.

Wilhelmsson, M. (2002). Spatial models in real estate economics. *Housing, theory and society*, 19(2):92–101.

Williams, C. K. and Rasmussen, C. E. (2006). *Gaussian processes for machine learning*. MIT press Cambridge, MA.

Wong, L. H., Pattison, P., and Robins, G. (2006). A spatial model for social networks. *Physica A: Statistical Mechanics and its Applications*, 360(1):99–120.

Wong, S. W. (2020). Assessing the impacts of mutations to the structure of covid-19 spike protein via sequential monte carlo. *Journal of Data Science*, 18(3):511–525.

Wong, S. W., Liu, J. S., and Kou, S. (2017). Fast de novo discovery of low-energy protein loop conformations. *Proteins: Structure, Function, and Bioinformatics*, 85(8):1402–1412.

Wong, S. W., Liu, J. S., and Kou, S. (2018). Exploring the conformational space for protein folding with sequential monte carlo. *The Annals of Applied Statistics*, 12(3):1628–1654.

Wong, S. W., Lum, C., Wu, L., and Zidek, J. V. (2016). Quantifying uncertainty in lumber grading and strength prediction: a bayesian approach. *Technometrics*, 58(2):236–243.

Wrapp, D., Wang, N., Corbett, K. S., Goldsmith, J. A., Hsieh, C.-L., Abiona, O., Graham, B. S., and McLellan, J. S. (2020). Cryo-em structure of the 2019-ncov spike in the prefusion conformation. *Science*, 367(6483):1260–1263.

Wu, F., Zhao, S., Yu, B., Chen, Y. M., Wang, W., Song, Z. G., et al. (2020). A new coronavirus associated with human respiratory disease in china. *Nature*, 579(7798):265–269.

Wu, G., Holan, S. H., and Wikle, C. K. (2013). Hierarchical bayesian spatio-temporal conway–maxwell poisson models with dynamic dispersion. *Journal of Agricultural, Biological, and Environmental Statistics*, 18(3):335–356.

WWPA (2021). Western lumber grading rules 2021. https://www.wwpa.org/. Accessed: 2022/12/10.

Wynne, T., Stumpf, R., and Briggs, T. (2013). Comparing modis and meris spectral shapes for cyanobacterial bloom detection. *International journal of remote sensing*, 34(19):6668–6678.

Wynne, T. T., Stumpf, R. P., Tomlinson, M. C., and Dyble, J. (2010). Characterizing a cyanobacterial bloom in western lake erie using satellite imagery and meteorological data. *Limnology and Oceanography*, 55(5):2025–2036.

Wynne, T. T., Stumpf, R. P., Tomlinson, M. C., Schwab, D. J., Watabayashi, G. Y., and Christensen, J. D. (2011). Estimating cyanobacterial bloom transport by coupling remotely sensed imagery and a hydrodynamic model. *Ecological Applications*, 21(7):2709–2721.

Yurkovetskiy, L., Wang, X., Pascal, K. E., Tomkins-Tinch, C., Nyalile, T. P., Wang, Y., Baum, A., Diehl, W. E., Dauphin, A., Carbone, C., et al. (2020). Structural and functional analysis of the d614g sars-cov-2 spike protein variant. *Cell*, 183(3):739–751.

Zammit-Mangion, A. and Wikle, C. K. (2020). Deep integro-difference equation models for spatio-temporal forecasting. *Spatial Statistics*, 37:100408.

Zhang, J., Kou, S. C., and Liu, J. S. (2007a). Biopolymer structure simulation and optimization via fragment regrowth monte carlo. *The Journal of Chemical Physics*, 126(22):06B605.

Zhang, J., Lin, M., Chen, R., Liang, J., and Liu, J. S. (2007b). Monte carlo sampling of near-native structures of proteins with applications. *Proteins: Structure, Function, and Bioinformatics*, 66(1):61–68.

Zhang, L., Jackson, C. B., Mou, H., Ojha, A., Peng, H., Quinlan, B. D., ..., and Li, W. (2020). Sars-cov-2 spike-protein d614g mutation increases virion spike density and infectivity. *Nature communications*, 11(1):1–9.

Zhou, H. and Smith, I. (1991). Factors influencing bending properties of white spruce lumber. *Wood and Fiber Science*, 23(4):483–500.

Zhou, P., Yang, X. L., Wang, X. G., Hu, B., Zhang, L., Zhang, W., et al. (2020a). A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*, 579(7798):270–273.

Zhou, T., Tsybovsky, Y., Olia, A. S., Gorman, J., Rapp, M., Cerutti, G., Chuang, G.-Y., Katsamba, P. S., Nazzari, A., Sampson, J. M., et al. (2020b). Cryo-em structures delineate a ph-dependent switch that mediates endosomal positioning of sars-cov-2 spike receptor-binding domains. *bioRxiv*.

Zhu, L., Carlin, B. P., and Gelfand, A. E. (2003). Hierarchical regression with misaligned spatial data: relating ambient ozone and pediatric asthma er visits in atlanta. *Environmetrics: The official journal of the International Environmetrics Society*, 14(5):537–557.