Spatial Modelling of Soil and Air Pollution Extremes

M. Daniela Cuba

Submitted in fulfilment of the requirements for the
Degree of Doctor of Philosophy

School of Engineering
College of Science and Engineering
University of Glasgow

September 2024

Declaration of Authorship

I, Miriam Daniela Cuba, declare that this thesis titled 'Spatial Modelling of Soil and Air Pollution Extremes' and the work presented in it are entirely my own. I confirm that:

- The work presented in this thesis was developed wholly while under candidature for a PhD at the University of Glasgow.
- External work is cited.
- All main sources of help have been appropriately acknowledged.

The work for the application in Chapters 4 and 5 was carried out in collaboration with Dr. Benjamin Marchant, at the British Geological Survey. Results from Chapter 4 were presented as a poster at the Royal Statistical Society conference in 2022, in Aberdeen, UK. The results from Chapter 5 were presented in an invited talk at the Royal Statistical Society conference in 2023, in Harrogate, UK, and as a poster at the Extreme Value Analysis (EVA) conference in Milan, Italy, in 2023. A manuscript of the work has also been submitted for publication in the Royal Statistics Society Journal - Series C. The work in Chapter 6 was done in collaboration with Dr. Craig Wilkie, at the University of Glasgow. A discussion piece related to the work in Chapter 6 has been accepted by the Royal Statistical Society Journal - Series C. Additionally, a manuscript of the work is in preparation for submission to a special issue of the Spatial Statistics journal. Finally, Chapter 7 covers only my contribution to the EVA 2023 Data Challenge, as part of the "Wee Extremes" team entry. My work focused on challenges C2 and C4, while the remaining challenges, C1 and C3, were authored by other members of the team and are not featured in this thesis. A manuscript of the complete team entry was accepted for publication in a special issue of the Extremes journal.

Abstract

In this thesis, novel spatial statistical methods for unreplicated bivariate heavy metal soil contamination and replicated $PM_{2.5}$ air pollution are developed by combining existing statistical approaches with extreme value theory. An introduction to the motivation behind this research is given in Chapter 1 while the necessary statistical and applied background for this research is given in Chapters 2 and 3, respectively.

In Chapter 4, the extremal dependence between threshold exceedances of heavy metal contaminants in the Glasgow Conurbation is investigated using two extreme value models with different extremal dependence structures that ignore the spatial dimension of the contaminants. The results show that for most contaminant pairs, moderately low quantile thresholds ($u < 0.95$) exhibit constant dependence, which can be modelled using a rigid dependence model, while exceedances of extreme quantile thresholds ($u > 0.95$) almost always display decaying dependence, requiring the flexible dependence of subasymptotic models. More specifically, the results show that chromium has a different migration behaviour than other elements, resulting in strongly decaying extremal dependence. Further evidence of this difference in behaviour is provided in the literature, as chromium is less likely to migrate regardless of conditions and persists in the soil longer than other heavy metals. In Chapter 5, a spatial model for the application is developed, which uses a bivariate mixture model approach to model the body and tail of the heavy metal distributions. Our approach is tailored to the case of unreplicated observations, which is non-standard in the extreme value theory literature. The body of the contaminant distributions was modelled using a Gaussian distribution, while the tails were modelled using a Gaussian-generalised Pareto composition. The body-tail components of both contaminants were modelled jointly under a coregionalisation framework, allowing the tail components to share a scaled spatial random effect, effectively accounting for the dependence in the tails. The model showed that the probability of exceeding a safety threshold was high in the south banks of the river Clyde in urban Glasgow and some villages to the east - all areas of historical industrial activity and mining legacies.

In Chapter 6, we present an approach for data fusion of $PM_{2.5}$ air pollution extremes in the Greater London region. Data fusion models are generally motivated by the need to integrate information from different data sources to obtain a better description of the

underlying phenomenon. In this case, we fuse remote-sensing data (modelled data), which enjoy complete spatial and temporal records, and in-situ measurements from observation stations. The model proposed is a tailored approach for threshold exceedances, representing extreme concentrations of $PM_{2.5}$, which enhances observations of the remote-sensing data (EAC4 dataset) to better represent threshold exceedances observed using data from the observation stations of the AURN - a high-quality air quality monitoring network in the UK. Results from the model show that the extremes data fusion model improves threshold exceedances reported by the EAC4 model, in the sense that it better approximates in-situ measurements. The extremes data fusion model also outperforms a competitive data fusion approach based on the Gaussian distribution. A map of fused observations shows different spatial patterns than the modelled observations, assigning higher concentrations to locations on the coast - a claim which is further corroborated by air pollution literature.

Finally, Chapter 7 presents my contribution to the challenges C2 and C4 of the EVA 2023 Data Challenge, a competition organised for the EVA 2023 conference in Milan, Italy. In C2, organisers ask for an extrapolated value that minimises an arbitrary loss function. To address the question, we propose a novel approach to extrapolate high quantiles under an application-specific loss function using an extreme-weighted bootstrap. C4 asks to estimate the probability of joint exceedance in a high-dimensional setting, for which we propose using a probabilistic principal component analysis model (PPCA). The methods were found to have mixed success, and we discuss the limitations and potential presented by these models.

Acknowledgements

First, I would like to thank my supervisors, Dr. Daniela Castro-Camilo and Prof. Marian Scott. You not only gave me an opportunity but committed to supporting me with everything you could, always fighting my corner with my best interest at heart. You have taught me as much about how to be a good researcher as you have about how to be a kind, caring, and generous mentor. I could never have wished for better (or more patient!) supervisors. I will always be grateful for everything you did for me as your student, and grateful too, for the opportunity to be your friend.

I would also like to thank Dr. Benjamin Marchant and Dr. Craig Wilkie, who were vital collaborators in this work. A special thank you to Dr. Benjamin Marchant, who once heard of my dream to get a PhD and endeavoured to help me make it happen. A big thank you to the School of Mathematics and Statistics at the University of Glasgow, who funded my PhD and have always gone above and beyond to show me kindness.

I would like to thank the 12:30pm lunch group: Stephen, Robin, Iain, Alba, Gabriel (+ Renata!), William, Toby, Vanessa, and Giovanni (and everyone else!). You have filled my days with laughter and adventure. A special thank you to my cohort, Stephen Villejo and Robin Muegge, who started this journey with me during COVID and have remained my close companions. My life is richer because all of you are in it, and for that, I am grateful.

Finally, I am grateful to my parents, Arturo and Patricia, and my brothers Matias and Juan Esteban. Though we may live continents away, your love, support, and constant presence in my life have preserved my sanity and fought off my loneliness. Everything I do is an ode to you and a small attempt to love you the way you love me — unconditionally and to a fault.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Statistical models have proven to be an invaluable tool in a world that is increasingly eager to promote data-driven decision-making. While applications in modern society have benefited from the development of these statistical tools, we consider that few present a more pressing issue than environmental pollution. Evidence of this crisis brought on by environmental pollution is present in every aspect of life on earth, from climate change and the extinction events of modern times (Kaiho, 2023) to the smallest microorganisms affected via various exposure pathways. Even we, humans, who are responsible for this crisis, are also victims of it. In 2015, an estimated 9 million people suffered premature deaths due to environmental pollution, making it the most significant environmental risk factor for disease and premature death (Fuller *et al.*, 2022).

The efforts to counter the adverse effects of environmental pollution are called pollution control measures. These measures aim to prevent the occurrence of pollution and remediate existing pollution by identifying sources, promoting government regulation to reduce pollution and promote accountability, using technology to reduce emissions, remediating damage, and reducing population exposure (Boccaccio, 2023). Statistical models are the bridge that turns data into useful information, enabling data-driven decisions for effective planning, data collection, analysis, modelling and interpretation. In this thesis, we develop novel spatial statistical methods to address specific problems encountered in two environmental pollution applications - soil contamination and air pollution - representing our contribution to the statistical literature and pollution control efforts. The remainder of this introductory chapter is as follows. The motivation and background for these projects are given in Section 1.1. Specific aims and objectives of the research are given in Section 1.3. Finally, Section 1.4 provides the structure for the remainder of the thesis.

1.1 Statistical Research in Environmental Pollution

In recent decades, developed countries have begun promoting and enforcing policies to mitigate the effects of environmental pollution through management and remediation. Effectively wielding policy as an agent of change, however, is conditioned on having access to reliable information on the extent of the pollution at useful scales in space and time. While this requirement is seemingly simple, environmental data are notoriously complex, with each application presenting a unique set of challenges. The research in this thesis addresses the challenges of two specific environmental pollution applications: spatial modelling of unreplicated bivariate heavy metal soil contaminants and spatiotemporal data fusion for extreme concentrations of particle matter ($<2.5\,\mu m$, $PM_{2.5}$) air pollution.

1.1.1 Heavy Metal Soil Contamination

Modelling heavy metal soil contamination typically refers to the interpolation of contaminant concentrations in space. The statistical models used for this purpose commonly follow classical approaches. The Gaussian framework is arguably the most important, underpinning a vast section of the spatial and geostatistical literature. A useful tool in the Gaussian toolbox are Gaussian processes, which naturally arise in spatial settings where any number of finite observations can be described using a multivariate Gaussian distribution. These models are widely used in soil sciences (see Webster and Oliver 2007), including geochemical mapping (Tóth *et al.*, 2016). However, soil contamination refers to the values above the baseline - the high and extremely high values that constitute the tail of the distribution. These extreme values at the tail are often ill-posed for modelling under Gaussian frameworks, which assume a lighter tail and result in underestimated extreme values. Furthermore, sources of heavy metal contamination often produce more than one contaminant, eliciting multivariate approaches. When modelling more than one contaminant, the problem of unsuitable modelling of the tails under Gaussian frameworks is further exacerbated. While most geostatistical models can account for the dependence between contaminants, the dependence between extreme values, that is, high concentrations of the contaminants, differs from the dependence in the rest of the distribution (Coles *et al.*, 1999). As a result, Gaussian models can underestimate both extreme values and the dependence between contaminants in the tail.

We propose to model the tail of the distribution of heavy metal contaminants using an extreme value approach while maintaining a Gaussian model for the body. The use of extreme value theory in this application is novel, as extreme value models require replicate observations at each location because of its asymptotic nature, which are generally unavailable in soil surveys. We present a workaround to this limitation, resulting in a body-tail approach to model the heavy-tailed metal contaminant distributions and cap-

ture extreme values. Moreover, we extend this framework to the bivariate setting, where two contaminants are modelled simultaneously in a manner that accounts for the extremal dependence between them. The first practical output of the model is the interpolation of marginal contaminant concentrations in space. The second one is a measure of risk by providing the probability of both contaminants jointly exceeding regulatory safety thresholds at any given location. These maps can be used in risk assessment for improved identification of sources of this joint pollution and inform pollution control measures. The research was undertaken in collaboration with Dr. Ben Marchant at the British Geological Survey (BGS), who kindly provided the data for the application. Chapter 5 develops the bivariate mixture model, while an in-depth investigation of extremal dependence between contaminants is given in Chapter 4.

1.1.2 $PM_{2.5}$ Air Pollution

The second research project in this thesis consists of the development of a data fusion model to improve the prediction of extreme observations of $PM_{2.5}$ using remote-sensing data and in-situ measurements from observation stations. Extreme events of particle matter pollution pose a global and significant risk to public health. Zhang *et al.* (2021) show that longer periods of heavy $PM_{2.5}$ pollution events increase cardiovascular mortality and morbidity. Policy to mitigate the effect of air pollution has largely focused on reducing emissions and public exposure. Monitoring networks, consisting of observation stations sparsely distributed in space due to costs, can be set up to obtain accurate measurements of pollution at specific locations, help quantify potential risk, and assess the effectiveness of pollution control measures. For locations where data from an observation station is available, these in-situ measurements of $PM_{2.5}$ are considered the best representation of the process at that location. However, due to the sparse spatial coverage of these stations, access to in-situ measurements is limited. Data from other sources, such as remote-sensing and modelled data, are generally available at regular intervals in space and time, providing information for locations where no observation station is found. However, these alternative data sources are known to be smoother than in-situ measurements and result in an underestimation of extreme values, as shown in Pendergrass *et al.* (2021). Inaccurate representation of the extreme values can negatively affect the accuracy of risk assessment and exposure quantification (Becker *et al.*, 2021). Data fusion models have been widely used to bridge this gap (Carnevale *et al.*, 2020) by correcting the bias in remote-sensing observations to better approximate in-situ measurements at locations with no observation station while preserving the remote-sensed data's spatial and temporal coverage.

Statistical data fusion models exist for this purpose (Berrocal *et al.*, 2010; Wilkie *et al.*, 2019), but are primarily based on Gaussian assumptions. Alternative models using extreme value theory also exist (Friederichs and Thorarinsdottir, 2012; Amaral Turkman

et al., 2021), but they are centred on matching the distributional properties between remote-sensing data and in-situ measurements. While this is a valuable approach when modelling marginal distributions, it does not result in a time series that mirrors in-situ measurements (Engelke *et al.*, 2019). The research in Chapter 6 consists of developing a data fusion model that targets extreme values. The model fuses remote-sensing data and in-situ observations to produce a complete spatial and temporal coverage dataset that better approximates in-situ measurements. This enhanced dataset can provide more accurate data for risk assessment and exposure modelling.

1.2 EVA Data Challenge

The EVA data challenge was organised as part of the Extreme Value Analysis (EVA) 2023 conference in Milan, Italy. The challenge consisted of 4 sub-challenges, which represented common problems in the application of extreme value analysis to real-world problems. No methodological novelty was required, rather, teams were encouraged to use existing methodology in novel and creative ways. Students at the University of Glasgow banded together as "The Wee Extremes Team" to submit an entry. The part of the challenge for which the author of this thesis is responsible is covered in Chapter 7.

1.3 Aims and Objectives

The aim of this thesis is to provide novel statistical approaches that build on existing spatial models and extreme value theory to better inform pollution risks at local scales. Given that the applications in this thesis are distinct, we consider project-specific aims. The first two objectives were set for the case of heavy metal soil contamination, while the third is specific to the data fusion for extremes project.

- ***Unreplicated spatial data.*** Develop a model for the body and tail of contaminant distributions by using a Gaussian distribution for the body and adapting an extreme value distribution for the unreplicated setting (the case where a single temporal replicate is available at each sampled location) for the tail.

- ***Bivariate extremal dependence.*** Model bivariate unreplicated spatial contaminants by accounting for the extremal dependence between components. A coregionalised framework using a mixture model is explored where the construction of components is specified to account for dependence at extreme concentrations.

- ***Spatiotemporal data fusion.*** To perform the fusion of two data sources for the improved assessment of extreme values of $PM_{2.5}$ at a local scale. Extend existing

data fusion models for this purpose through the inclusion of a modified generalised Pareto likelihood that accounts for the occurrence of non-extreme observations.

1.4 Structure of the Thesis

The remainder of this thesis is made of six chapters. Chapter 2 presents the statistical background necessary to understand the research in this thesis. It provides a summary of the spatial and data fusion models commonly used in the applications presented here. The foundations of extreme value theory are provided in this chapter, including spatial extremes and bivariate dependence. Finally, the necessary methodology for Bayesian inference is also covered. Chapter 3 gives the context of both applications. It includes the sources of contamination and pollution, as well as its impacts on society and public health. The chapter also provides descriptions of the data and an exploratory analysis of each dataset used. Chapters 4 and 5 cover the heavy metal soil contamination application. Chapter 4 investigates the extremal dependence of heavy metal contaminants in the soil by comparing two models with different extremal dependence structures. Chapter 5 covers the development of the bivariate mixture model that incorporates Gaussian and extreme value distributions for accurate modelling of the body and tail of each contaminant in space while accounting for the extremal dependence between them. The application for data fusion for extremes is presented in Chapter 6. The chapter includes a comparison of data sources and a detailed description of the inference methodology. Chapter 7 differs from the rest of this thesis as it describes the work undertaken for the Extreme Value Analysis Conference (EVA) in 2023. Final remarks on the methodologies and potential future developments of the work in this thesis are given in Chapter 8.

Chapter 2

Statistical Background

2.1 Geostatistical Models and Framework

Geostatistics is the statistical field that aims to quantify phenomena distributed in space or space and time by exploiting information from observations at close distances. It is commonly associated with environmental applications such as hydrology, geology, mineral exploration, agriculture, forestry, and ecology, among others (Chilès and Delfiner, 2012), where the phenomena are often too complex to be described by simplistic mathematical functions. Furthermore, the cost and difficulties associated with data collection are often restrictive and result in sparse spatial coverage, which introduces uncertainty into the modelling process. Accurately quantifying this uncertainty is a central aim of geostatistics.

Data in the spatial and spatiotemporal dimensions can be collected in various formats, but they usually come in the form of spatial point processes, areal data, and point-referenced or geostatistical data (Davison and Gholamrezaee, 2012). The models for analysing such data are chosen based on the use case. However, common modelling aims include structural analysis, survey optimisation, interpolation, quantification of polynomial drift, integration of multiparameter information, data fusion, spatiotemporal modelling, indicator estimation and classification, selection and change-of-support problems, and spatial point patterns, among others (Chilès and Delfiner, 2012). This section provides an overview of geostatistical methodology for spatial and spatiotemporal interpolation. Section 2.1.1 gives an overview of spatial processes and their characteristics as the underpinning of geostatistical methodology. Section 2.1.2 describes univariate geostatistical methodology, mainly Kriging and its variants. Finally, Section 2.1.3 summarises methodological extensions of geostatistical methodology for the multivariate setting.

2.1.1 Spatial Processes

In order to describe a geostatistical phenomenon in space, it is first necessary to find an appropriate mathematical definition that is spatially continuous, defining the behaviour of the phenomenon at every location in space. Stochastic processes are a natural fit for this requirement. Defined as a random function (RF), they can be continuous in space and have flexible constructions, allowing for different spatial dependence structures. The most common of these processes are Gaussian Processes (GPs), which underpin a vast portion of geostatistical theory.

The mathematical definition of RFs is straightforward and flexible. Given a domain $\mathcal{S} \subset \mathbb{R}^d$ and a probability space $(\Omega, \mathcal{A}, P)$, a RF is a function of two variables $Z(s, \omega)$ such that for each location $s \in \mathcal{S}$, $Z(s, \cdot)$ is a random variable on $(\Omega, \mathcal{A}, P)$. Each function $Z(\cdot, \omega)$ defined on $\mathcal{S}$ is a realisation of the process $Z(s)$ with an observation defined as $z(s)$ (Chilès and Delfiner, 2012, Ch. 1). The extension of a random function to s to the spatial dimension, $d = 2$, is called a random field. A GP arises as a special case where n samples of the random field are multivariate Gaussian, a GP (or Gaussian random field) can be defined using an n-dimensional mean vector, $\boldsymbol{\mu}$, and a $n \times n$ positive-definite covariance matrix, Σ.

The behaviour of the process throughout $\mathcal{S} \subset \mathbb{R}^2$ can be modelled based on underlying assumptions. In this section, we only consider two: stationary and intrinsically random, where stationary processes can be further subdivided into strictly stationary and second-order stationary. Strict stationarity is a strong assumption of simplicity. A strictly stationary process is invariant under translation for any vector h so that

$$\Pr(Z(s_1) < z_1, \cdots, Z(s_n) < z_n) = \Pr(Z(s_1 + h) < z_1, \cdots, Z(s_n + h) < z_n),$$

where h is a measure of distance known as the lag and s_i are locations in $\mathcal{S}$ for $i = 1, ..., n$ (Webster and Oliver, 2007, Ch. 4). Second-order stationarity, or weak stationarity, describes a process characterised by its covariance matrix Σ. The entries of the matrix, which represent the covariance between observations, are obtained using a covariance function

$$C(h) = \mathrm{E}[Z(s) - \mu][Z(s + h) - \mu],$$

where $\mu = \mathrm{E}[Z(s)]$ is the mean, and the covariance function $C(\cdot)$ is a function of the lag between observations. The covariance between $Z(s)$ and $Z(s+h)$ can also be linked to the correlation between the two observations through the *correlogram* defined as $\rho(h) = \frac{C(h)}{C(0)}$, where $C(0) = \sigma^2$ is the covariance at lag 0. Unless stated otherwise, the work in this dissertation will use stationarity to refer only to second-order stationary.

An intrinsically random function (IRF) is defined as that with second-order stationary increments (Chilès and Delfiner, 2012, Ch. 1), making stationary random functions a

subset of IRFs. Just as in the stationary case, it is characterised by the mean, defined as a function of the distance, h, between observations

$$\mu(h) = \mathrm{E}[Z(s+h) - Z(s)],$$

and the covariance matrix, with elements also defined as a function of the lag h as

$$2\gamma(h) = \mathrm{Var}[Z(s+h) - Z(s)]. \tag{2.1}$$

The function in (2.1) has been known by various terms. Kolmogorov (1941); Gandin (1966); Yaglom (1987) referred to it as the structure function, but the more enduring name was given by Matheron (1963) who defined $\gamma(\cdot)$ as the semivariogram. Equation (2.1) shows that the difference between $Z(s)$ and $Z(s+h)$ is a function of the distance h between them. In practice, using 2γ over the covariance is preferable, as it does not require the mean μ to be known. Additionally, given that second-order stationary random functions are a subset of IRFs, the variogram is a generalised characterisation of the variance structure. The two, however, are linked through $\gamma(h) = C(0) - C(h)$.

Given that information about the process is restricted to that provided by the observations, the true variogram is unknown and must be estimated from the sampled observations. For this reason, data collection schemes must be appropriate for variogram estimation. Chiles and Delfiner (1999) define the minimum number of samples necessary as $n = 50$, with a small number of samples being linked to bias in variogram estimation (Kravchenko, 2003). Furthermore, since the variogram estimation is defined by the distance between observations, h, the sampling strategy can be used to effectively optimise the variogram estimation by optimising three variables: location, distance between observations, and directionality. The location of the observations is typically defined by field experts, who are best placed to assess a representative sampling strategy for the phenomenon under study. The distance between observations must also be optimised, as it may affect the continuity of the estimated variogram. In theory, the variogram is continuous everywhere in $\mathcal{S}$, including at the origin, as the spatial process is also continuous at the origin. In practice, this is often not the case, as the true variogram is unknown, and samples are collected at limited distances, typically omitting distances at a microscale, preventing the definition of the estimated variogram at the origin. When $\gamma(h) \to c_0 > 0$, as $h \to 0$, then c_0 is known as the *nugget effect* (Matheron, 1963). Obtaining a good representation of distances between observations is important in mitigating the nugget effect and appropriately capturing the spatial dependence of the data. Finally, if the process is invariant in the direction of the lag h, the variogram is known as *isotropic*; if it is not and the direction affects $2\gamma(\cdot)$, it is known as *anisotropic*. The sampling strategy must reflect the underlying process and its dimensionality. Webster and Oliver (1992) proposed

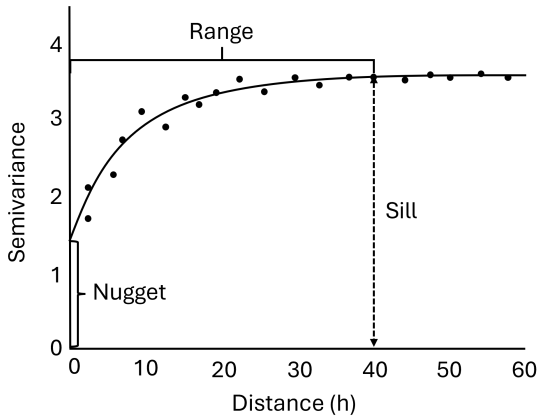

Figure 2.1: Diagram displaying the elements of a semivariogram: the nugget, range, and sill.

optimisation of sampling locations using the method of moments, and Lark (2002) suggests a maximum likelihood approach, but many other strategies have been proposed in the literature. For more details, see Lawrence *et al.* (2020).

Earlier attempts at the estimation of $2\gamma(\cdot)$ include Matheron (1963) who proposed the method of moments, now known as the *classical* estimator, as a natural estimator of the variogram such that

$$2\hat{\gamma}(h) = \frac{1}{|N(h)|}\sum_{N(h)}(Z(s_i) - Z(s_j))^2, \text{ for } h \in R^D,$$

where $N(h) = \{(s_i, s_j) : s_i - s_j = h\}$, is the number of distinct pairs at lag h. Cressie (1993) remarked that the classical estimator is sensitive to outliers and proposed a *robust* variogram estimator to mitigate outlier effects. It is defined as

$$2\overline{\gamma}(h) = \left(\frac{1}{|N(h)|}\sum_{N(h)}|Z(s_i) - Z(s_j)|^{1/2}\right)^4 /(0.457 + 0.494/|N(h)|).$$

Other variogram estimators have been proposed for different cases. For more details, see Lark (2000).

Given the relationship between the covariance and the variogram, a covariance or variogram function can be fitted to model the empirical variogram. Mathematically, a valid covariance function needs to represent monotonic increases with lag increase, have a constant maximum (sill), and have a positive intercept on the ordinate to represent the discontinuity at the origin (nugget), as shown in the diagram in Figure 2.1. Additional features of the variogram, such as periodic fluctuation (hole) and anisotropy, can be accommodated using specialised covariance functions.

One of the simplest variogram models (covariance functions) is the bounded linear

model

$$\gamma(h) = \begin{cases} c\left(\frac{h}{a}\right) & \text{for } h \leq a \\ c & \text{for } h > a, \end{cases}$$

where c is the sill, and a is the range - the lag at which the variogram reaches its sill.

Many other parametric models have been proposed for isotropic variograms (see Chilès and Delfiner 2012, Ch. 1). However, the Matern covariance function is arguably the most important, and it is defined as

$$\gamma(h)_{\text{Matern}} = c\left\{1 - \frac{1}{2^{\nu-1}\Gamma(\nu)}\left(\frac{h}{r}\right)^{\nu} K_{\nu}\left(\frac{h}{r}\right)\right\}, \tag{2.2}$$

where c is the sill, r is the range parameter, K_{ν} is the modified Bessel function of the second kind, and ν is the smoothness parameter. The special cases of the Matern model are the exponential model when $\nu = 0.5$, the Wittle function when $\nu = 1$, and converges to the squared exponential function as $\nu \to \infty$.

Fitting a model to the variogram is done through model comparison using appropriate selection criteria. Once the variogram has been estimated, models with approximately similar shapes can be chosen as candidates. Then, each can be fitted by weighted least squares - minimisation of the sum of squares - and optimised. The results can be plotted alongside the sample variogram, and the best model can be chosen visually and by the smallest residual sum of squares or mean square error (Cressie, 1993; Webster and Oliver, 2007).

2.1.2 Univariate Geostatistical Models: Kriging

Kriging, also known as the best linear unbiased predictor (BLUP), is likely the most important of the probabilistic spatial interpolation methods. It was proposed by D.G. Krige, who developed the method for mineral ore prediction in the 1950s (Krige, 1951). Kriging is often defined as a weighted average, with weights chosen to minimise prediction variance (Matheron, 1963). Simply put, it is a linear model to predict a spatial process Z and quantify prediction uncertainty. An underpinning assumption of Kriging is that the process $Z(\mathbf{s})$ is a realization of a GP at locations $\mathbf{s} = s_1, ..., s_n \subset \mathcal{S}$ and $\mathcal{S} \in \mathbb{R}^2$. Generally, the objective of Kriging is to estimate Z at an unsampled location s_0, meaning

$$Z(s_0) = \int Z(s) p_0(ds), \tag{2.3}$$

where p_0 is an integrable measure that corresponds to a Dirac measure $p_0(\mathrm{d}s) = \delta(s - s_0)$ (Chilès and Delfiner, 2012, Ch. 2). The prediction is easily extendable to estimate the

average over a block B as

$$Z(s_0) = \frac{1}{|B|}\int_B Z(s)\mathrm{d}s,$$

where $p_0(\mathrm{d}s) = (1/|B|)\mathbf{1}_B(s)\mathrm{d}s$, and $\mathbf{1}_B(\cdot)$ is the indicator function over the block B centred at s_0 and is known as *block Kriging*. The prediction is accomplished as a weighted mean of observations, defined as

$$\hat{Z}(s_0) = \sum_{i=1}^{n} \lambda_i(s_0)Z(s_i) + \lambda_0(s_0),$$

where s_0 is centred at B, $\lambda_i(s_0)$ is a weight placed on $Z(s_i)$ and $\lambda_0(s_0)$ is a constant that depends solely on s_0.

Simple Kriging (SK)

Simple kriging (Journel and Huijbregts, 1978) is the simplest form of Kriging. It represents the case where the mean, μ, is fixed and known. When the purpose is to produce a point estimate at location s_0, as in (2.3), only the variance needs to be estimated as a weighted average where the weights λ_i are optimised to minimise the mean square error (MSE) defined as

$$\mathrm{E}(\hat{Z}(s_0) - Z(s_0))^2 = \mathrm{Var}(\hat{Z}(s_0) - Z(s_0)) + [\mathrm{E}(\hat{Z}(s_0) - Z(s_0))]^2.$$

This estimation is the same as assuming a zero mean for Z and adding the constant μ, only leaving the estimation of λ_i.

The MSE can be expanded as

$$\mathrm{E}(\hat{Z}(s_0) - Z(s_0))^2 = \sum_i \sum j\lambda_i\lambda_j\sigma_{ij} - 2\sum_i \lambda_i\sigma_{i0} + \sigma_{00}.$$

The minimum of the function can then be obtained using the partial derivative with respect to λ_i:

$$\frac{\partial}{\partial\lambda_i}\mathrm{E}(\hat{Z}(s_0) - Z(s_0))^2 = 2\sum_j \lambda_j\sigma_{ij} - 2\sigma_{i0} = 0,$$

where λ_i are solutions to the linear system of n equations (also known as Yule-Walker equations) $\sum_j \lambda_j\sigma_{ij} = \sigma_{i0}$, also defined in matrix notation as $\mathbf{\Sigma\lambda} = \boldsymbol{\sigma}_0$, where $\mathbf{\Sigma} = [\sigma_{ij}]$ is an $n \times n$ matrix of covariances and $\boldsymbol{\sigma}_0$ is a vector of covariances with the target at s_0.

The final estimate is

$$\mathrm{E}(\hat{Z}(s_0) - Z(s_0))^2 = \sigma_{00} - \sum_i \lambda_i\sigma_{i0} = \sigma^2_{SK}, \tag{2.4}$$

where σ^2_{SK} is known as the *kriging variance* associated with $\hat{Z}$ and represents the uncertainty at s_0.

Ordinary Kriging (OK)

Ordinary Kriging arises when the mean, μ, is unknown but constant $\mu(s) = a_0$. The MSE for the point estimator in (2.3) can be written as

$$\mathrm{E}(\hat{Z}_0 - Z_0)^2 = \mathrm{Var}(\hat{Z}_0 - Z_0) + \left[\lambda_0 + \left(\sum_{i=1} -1\right) a_0\right]^2.$$

Under the restriction that $\sum_i \lambda_i = 1$, the variance of the error is

$$\mathrm{Var}(\hat{Z}(s_0) - Z(s_0)) = \sum_i \sum_j \lambda_i \lambda_j \sigma_{ij} - 2\sigma_i \lambda_i \sigma_{i0} + \sigma_{00},$$

which depend on covariances and weights λ. It is possible to solve for λ_i using Lagrange multipliers with

$$Q = \mathrm{Var}(\hat{Z}(s_0) - Z(s_0)) = +2\lambda_\mu \left(\sum_i \lambda_i - 1\right),$$

where λ_μ is the Lagrange multiplier. The solution to the system is then

$$\mathrm{E}(\hat{Z} - Z(s_0))^2 = \sigma_{00} - \sum_i \lambda i \sigma_{i0} - \lambda_\mu = \sigma^2_{OK}$$

Universal Kriging (UK) and Kriging with External Drift (KED)

In universal Kriging (UK), the target process, Z, is decomposed as

$$Z(s) + \mu(s) + Z_0(s)$$

where Z_0 is a zero mean random function referred to as the residuals, and the mean, $\mu(s)$, is unknown but is assumed to be a function of covariates as

$$\mu(s) = \sum_{j=0}^{J} \beta_j f_j(s), \tag{2.5}$$

where $f_j(s)$ are functions and β_j are unknown coefficients. External drift is used in stochastic processes to refer to a trend. Kriging with external drift (KED) is an extension of UK where the trend is given by external variables, and the process is defined as

$$Z(s) + \beta_0 + \beta_1 T(s) + Z_0(s),$$

where $T(s)$ is a deterministic function or set of j functions, $f_j(s)$. Parameter estimation works in a similar manner as simpler variants by minimising $\mathrm{E}(\hat{Z}(s_0) - Z(s_0))^2$ using Lagrange multipliers. The UK system is defined as

$$\begin{cases} \sum_\beta \lambda_\beta \sigma_{i\beta} + \sum_j \beta j f_i^j = \sigma_{i0}, & i = 1, \ldots, n \\ \sum_i \lambda_i f_i^j = f_0^j, & j = 0, \ldots, J \end{cases}. \tag{2.6}$$

In matrix notation, the system can be simplified as $\boldsymbol{A}\boldsymbol{w} = \boldsymbol{b}$ with the structure

$$\underbrace{\begin{bmatrix} \boldsymbol{\Sigma} & \mathbf{F} \\ \mathbf{F}' & \mathbf{0} \end{bmatrix}}_{\boldsymbol{A}} \underbrace{\begin{bmatrix} \lambda \\ \boldsymbol{\beta} \end{bmatrix}}_{\boldsymbol{W}} = \underbrace{\begin{bmatrix} \boldsymbol{\sigma}_0 \\ \boldsymbol{f}_0, \end{bmatrix}}_{\boldsymbol{b}}$$

for $\boldsymbol{\Sigma}, \boldsymbol{\lambda}, \boldsymbol{\sigma}_0$ defined as in simple kriging and where

$$\mathbf{F} = \begin{bmatrix} 1 & f_1^1 & . & f_1^L \\ 1 & f_2^1 & . & f_2^L \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ 1 & f_N^1 & . & f_N^L \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \mu_0 \\ \mu_1 \\ . \\ . \\ . \\ \mu_L \end{bmatrix}, \quad \boldsymbol{f}_0 = \begin{bmatrix} 1 \\ f_0^1 \\ . \\ . \\ . \\ f_0^L \end{bmatrix}.$$

Under the conditions that $\boldsymbol{A}$ is not singular, $\boldsymbol{\Sigma}$ is strictly positive definite, and $\mathbf{F}$ is of full rank, the system is solved to yield a UK variance:

$$\sigma^2_{\mathrm{UK}} = \sigma_{00} - \boldsymbol{\lambda}'\boldsymbol{\sigma}_0 - \boldsymbol{\beta}'\boldsymbol{f}_0. \tag{2.7}$$

In the special (yet common) case that a constant function 1 is in the basis of drift functions $f^j(s)$, σ can be replaced by the variogram $-\gamma$ in the system in (2.6). The variogram matrix, $\boldsymbol{\Gamma}$, is used instead of $\boldsymbol{\Sigma}$, as the true covariance is not known but rather is estimated through the variogram. Given that Z has a Gaussian distribution and is centred at zero, uncertainties are easily quantified as

$$\Pr(|\hat{Z} - Z| > 1.96\sigma_{\mathrm{K}}) = 0.05,$$

which produces the 95% confidence interval

$$[\hat{Z} - 1.96\sigma_{\mathrm{K}}, \hat{Z} + 1.96\sigma_{\mathrm{K}}].$$

The estimation of the mean $\mu(s)$ is done in the usual way to a generalised linear model (GLM) or generalised mixed linear model (GLMM) in case that the drift includes random

effects. The model in matrix notation is expressed as

$$\boldsymbol{Z} = \boldsymbol{F}\boldsymbol{\beta} + \boldsymbol{Z}_0,$$

where estimates of $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}$ arise by minimising

$$(\boldsymbol{Z} - \boldsymbol{F}\hat{\boldsymbol{\beta}})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{Z} - \boldsymbol{F}\hat{\boldsymbol{\beta}}).$$

Other Variants

Variants of the classical Kriging models have been proposed in the literature for a myriad of applications. Some notable examples are Transgaussian Kriging, which is the Kriging process performed on data after a transformation into a Gaussian distribution; Fixed-ranked Kriging, which is kriging using a non-stationary covariance function, and spatiotemporal Kriging which is kriging extension for the spatiotemporal case. For more details see (Cressie, 1993).

2.1.3 Multivariate Geostatistical Models

While modelling the spatial variability of a single spatial variable motivated the development of geostatistical models, real-world applications often involve two or more variables. In these multivariate scenarios, whether variables are related or the purpose is joint modelling, geostatistical models have been extended to capture marginal and joint spatial patterns while accounting for the spatial interdependence between variables. This section will cover the extension of Kriging to the multivariate setting, known as Cokriging (CK).

When considering p simultaneous response functions defined as $\{Z_i(s) = s \in D \subset \mathbb{R}^n\}$. Each variable can be sampled over a different set of locations $\mathcal{S}_i$ so that $\mathcal{S}_i = \{s_\alpha \in D : Z_i(s_\alpha)\}$ for $\alpha = \{1, ..., N_\alpha\}$. Assumptions are similar to UK in that $\mu_i(s)$ is a linear combination of covariates. The extension of the UK to the multivariate setting has to do with the incorporation of the cross-covariance between $Z_i(s_\alpha)$ and $Z_j(s_\beta)$, which is denoted as $\sigma_{ij}(s_\alpha, s_\beta)$, and is defined as

$$\sigma_{ij}(s_\alpha, s_\beta) = \mathrm{Cov}[Z_i(s_\alpha), Z_j(s_\beta)] = \mathrm{E}[Z_i(s_\alpha), Z_j(s_\beta)] - \mu_i(s_\alpha)\mu_j(s_\beta).$$

As in UK, the cross-covariance is unknown and must be estimated empirically. The variogram is extended to the multivariate case and referred to as the cross-variogram (Cressie, 1993), defined as

$$2\gamma_{jj'}(\boldsymbol{h}) = \mathrm{Var}\left[Z_j(\boldsymbol{s}+\boldsymbol{h}) - Z_{j'}(\boldsymbol{s})\right] = \mathrm{E}\left[Z_j(\boldsymbol{s}+\boldsymbol{h}) - Z_{j'}(\boldsymbol{s})\right]^2 - \left(\boldsymbol{\mu}_j - \boldsymbol{\mu}_{j'}\right)^2. \quad (2.8)$$

The modelling of the cross-variogram is done similarly to the univariate case with

various standard models proposed (Cressie and Helterbrand, 1994).

CK is a multivariate extension of UK where information on a primary response variable is provided by auxiliary data (secondary variables) not sampled at the same locations as the primary variable (Webster and Oliver, 2007). UK arises as a special case of CK where the secondary variables are independent from the primary. However, if the primary and secondary variables are dependent, CK is preferred over UK. The CK estimator for the process of interest Z^* has the form

$$\hat{Z}^* = \sum_i \boldsymbol{\lambda}_i' \boldsymbol{Z}_i. \tag{2.9}$$

Under the conditions that $\sum_{i=1}^{N_i} \boldsymbol{\lambda}_{1i} = 1$ and $\sum_{i=1}^{N_i} \boldsymbol{\lambda}_{ji} = 0$, the CK system arises:

$$\begin{cases} \sum_j \mathbf{C}_{ij}\boldsymbol{\lambda}_j + \mathbf{F}_i\boldsymbol{\eta}_i = \mathbf{c}_{i0}, & i = 1, \ldots, p, \\ \mathbf{F}_i'\boldsymbol{\lambda}_i = \mathbf{f}_{10}\delta_{i1}, & i = 1, \ldots, p. \end{cases} \tag{2.10}$$

The resulting CK variance is then

$$\sigma^2_{\mathrm{CK}} = \mathrm{E}\left(\hat{Z}^* - Z(s_0)\right)^2 = c_{00} - \sum_i \boldsymbol{\lambda}_i'\mathbf{c}_{i0} - \boldsymbol{\eta}_1'\mathbf{f}_{10},$$

where $\boldsymbol{\eta}$ are Lagrange parameters. Unlike UK, the kriging weights in CK depend on the relative dispersion of the variations and are estimated using least squares estimation. Parameter estimation is done by optimising the parameters in (2.10) to minimise the prediction errors between the co-kriging estimate in (2.9) and the observed value. As in the univariate case, this minimisation is performed using either the Lagrange multiplier or multivariate numerical optimisation techniques.

2.2 Extreme Value Theory

Extreme value theory (EVT) is the branch of statistics that deals with the modelling and analysis of extreme values or extreme events. It represents a rigorous mathematical foundation for characterising the distribution of extreme values and estimatin the probability of rare events and their magnitude. This section provides a review of the foundations of EVT and its important developments in multivariate and spatiotemporal settings.

2.2.1 Classical Extreme Distributions

Generalised Extreme Value Distribution (GEVD)

Extreme values are usually classified as the maximum value in a block (block maxima), the r-largest values in a block, or the values exceeding a sufficiently high threshold (threshold exceedances). The r-largest values approach falls outside the scope of this dissertation and will not be considered in the remainder of the document.

To understand the theory behind the block-maxima approach, let $M_n = \max\{X_1, ..., X_n\}$ be block-maxima, where $X_1, ..., X_n$ is an i.i.d. sequence with a common (but unknown) cumulative distribution function (cdf) F. The block from which each maximum is taken is typically a temporal measure such as months, seasons, or years. The extremal types theorem states that if there exist normalising sequences $a_n > 0$ and b_n such that

$$\Pr\left\{\frac{(M_n - b_n)}{a_n} \leq x\right\} \to G(x) \quad \text{as } n \to \infty,$$

where G is a non-degenerate function, G is in the family of Generalized Extreme Value distributions (GEVDs). The general form of the GEVD is

$$G(x) = \exp\left\{-\left[1 + \xi\Big(\frac{x-\mu}{\sigma}\Big)\right]\right\}^{-1/\xi}, \tag{2.11}$$

defined on the set $\{x : 1 + \xi(x-\mu)/\sigma > 0\}$, where μ, $\xi \in \mathbb{R}$ and $\sigma > 0$. The three parameters of the model are location μ, scale σ, and shape ξ. Notable special cases of the GEVD family include the Fréchet distribution when $\xi > 0$ (Frechét, 1927); the Weibull distribution when $\xi < 0$; and the Gumbel distribution as $\xi \to 0$ (Gumbel, 1935), defined as a special case of (2.11):

$$G(x) = \exp\Big\{-\exp\Big(\frac{x-\mu}{\sigma}\Big)\Big\}.$$

The GEVD family possess the property of max-stability, meaning the distribution is invariant to the process of sampling maxima, only changing location and scale but maintaining the same shape, meaning that for every $n = 2, 3, ...$, there are constants α_n and β_n for which

$$G^n(\alpha_n z + \beta_n) = G(z),$$

where the distribution G^n is the distribution function of M_n, where each X_i is independent and G-distributed.

The distribution in (2.11) enables extrapolation into unobserved extreme quantiles

through inversion:

$$r_p = \begin{cases} \mu - \frac{\sigma}{\xi}\left[1 - \{-\log(1-p)\}^{-\xi}\right], & \text{for } \xi \neq 0, \\ \mu - \sigma \log\{-\log(1-p)\}, & \text{for } \xi = 0. \end{cases} \tag{2.12}$$

Here, r_p is considered the return level and is associated with the probability of occurrence, referred to as the return period, $1/p$.

Parameter estimation is generally straightforward but is subject to inherent limitations set by the shape parameter ξ. First, the k-th moment of the distribution is lost when $\xi > 1/k$, which can serve as a practical limitation. Second, the distribution has an upper bound of $\mu - \sigma/\xi$ in the case that $\xi < 0$. In the case that $\xi > 0$, there is no upper bound but the lower bound is also μ. Finally, when $\xi > -0.5$, maximum likelihood estimators (MLEs) with full regular and asymptotic properties exist, which is not the case when $\xi < -0.5$ (Coles, 2001, Ch. 3).

Generalised Pareto Distribution (GPD)

The generalised Pareto distribution (GPD) is the limiting distribution of threshold exceedances, just as the GEVD is to block-maxima. The two are also closely related and are said to be associated, where the GPD can be derived by using the point-process representation of the GEVD (Coles, 2001, Ch. 4). If $\Pr\{M_n \leq x\} \approx G(x)$ as defined in (2.11) then, for a large enough threshold u, the distribution of $(X - u)$, conditional on $(X > u)$ is asymptotically approximated by

$$H(x) = 1 - \left(1 + \frac{\xi x}{\tilde{\sigma}}\right)^{-1/\xi}, \tag{2.13}$$

defined on $\{x : x > 0 \text{ and } (1 + \xi x/\tilde{\sigma}) > 0\}$ due to the conditioning of $X > u$, and where

$$\tilde{\sigma} = \sigma + \xi(u - \mu),$$

where $\tilde{\sigma}$ is the scale parameter and ξ is the shape parameter. Associated distributions share the same shape parameter but differ in scale and location parameters. Similarly to the GEVD, the shape parameter of the GPD is dominant, defines the tail behaviour, and is often a practical consideration during inference. When $\xi < 0$, it has an upper bound, as does the GEVD. In the case that $\xi > 0$, the upper tail is unbounded. The case of $\xi = 0$ is obtained by taking $\xi \to 0$, resulting in

$$H(x) = 1 - \exp\left(-\frac{x}{\tilde{\sigma}}\right), \quad x > 0,$$

which is equal to an exponential distribution with rate $1/\tilde{\sigma}$.

The only conditions set on the threshold u are that it be sufficiently high and positive ($u > 0$). However, the choice of threshold is similar to the choice of block size and invokes a bias-variance tradeoff. More data points, available at lower values of u and smaller block sizes, reduce variance but increase bias. The alternative is to increase u, or block sizes in the GEVD, which results in high variance but lower bias. However, inherent properties of the GPD can guide threshold selection. For an appropriately high value of u_0, H will remain invariant to higher thresholds ($u > u_0$). This property is referred to as the threshold-stability property and is analogous to the max-stability property of the GEVD. The threshold-stability property of the GPD means there is a minimum threshold u_0 for which all thresholds $u > u_0$ produce a constant mean of excesses above a threshold $E(X - u|X > u)$. These estimates are expected to change linearly with u when $u > u_0$, enabling the estimation of

$$\left\{ \left(u, \frac{1}{n_u} \sum_{i=1}^{n_u} (x_{(i)} - u) : u < x_{\max} \right) \right\}, \tag{2.14}$$

where $x_{(1)}, ..., x_{(n_u)}$ consist of the n_u observations that exceed u. When plotted, it is useful for finding an appropriate threshold u_0, the smallest threshold for which exceedances can be appropriately modelled by the GPD.

Finally, return levels for the GPD can also be estimated by noticing that

$$\Pr(X > x) = \zeta_u \left[1 + \xi \left(\frac{x - u}{\sigma} \right) \right]^{-1/\xi},$$

where $\zeta_u = \Pr(X > u)$. The closed-form solution for the return value is then obtained as

$$x_m = u + \frac{\sigma}{\xi}[(m\zeta_u)^{\xi} - 1],$$

where m indicates the return level, meaning x will exceed x_m once every m observations on average.

2.2.2 Multivariate and Spatial Extreme Models

When the aim of the analysis is to model two or more variables of interest, the much more challenging task of modelling extremal dependence arises. Extremal dependence can be classified into two classes; asymptotic dependence and asymptotic independence, and is a central consideration in the selection of multivariate extreme value models (Resnick, 1987; Beirlant, 2004; Haan and Ferreira, 2006). In this section, we provide a more detailed description of extremal dependence, covering asymptotic block-maxima and threshold exceedance approaches for the multivariate case, spatial extreme value models, and

subasymptotic models for flexible dependence structures in spatial extremes.

Extremal Dependence

Extremal dependence, or tail dependence, refers to the dependence between the extreme values of two variables. The definitions presented below are easily extendable beyond the bivariate case, but dimensions above $d = 2$ are outside the scope of this thesis.

A widely used measure of extremal dependence is the coefficient of tail dependence χ_{ij} (Coles, 2001, Ch. 8) which, for two components X_i and X_j with cdf F_i and F_j respectively, is defined as

$$\chi_{ij} = \lim_{u \to 1} \chi_{ij}(u) = \lim_{u \to 1} \Pr[F_i(X_i) > u, F_j(X_j) > u]/(1-u), \qquad \text{for } u \in [0,1]. \tag{2.15}$$

Intuitively, χ_{ij} represents the conditional probability X_i is large, given that X_j is also large, meaning the probability that extreme values occur in X_i and X_j simultaneously. Note that χ_{ij} is a limiting measure on the uniform scale, where 0 denotes asymptotic independence, and where 1 represents perfect asymptotic dependence. In general, χ_{ij} quantifies dependence between components when they are asymptotically dependent. In the case that the components are asymptotically independent, a measure of their dependence can be measured using $\bar{\chi}_{ij}$ (Coles, 2001, Ch. 8), defined as

$$\begin{aligned} \bar{\chi}_{ij} = \lim_{u \to 1} \bar{\chi}_{ij}(u) &= \frac{2 \log \Pr\{F_i(X_i) > u\}}{\log \Pr\{F_i(X_i) > u, F_j(X_j) > u\}} - 1 \\ &= \frac{2 \log(1-u)}{\log \Pr\{F_i(X_i) > u, F_j(X_j) > u\}} - 1, \end{aligned}$$

where $-1 \leq \bar{\chi}_{ij} \leq 1$. Components are considered asymptotically dependent when $\bar{\chi}_{ij} = 1$, and asymptotically independent when $\bar{\chi}_{ij} = 0$. $\bar{\chi}_{ij}$ is considered a measure of extremal dependence for a pair of variables that are already known to be asymptotically independent. Specifically, the dependence between asymptotically independent variables increases with increasing $\bar{\chi}_{ij}$.

An alternative dependence measure for the asymptotically independent case, i.e. when $\chi_{ij} = 0$, is the residual tail dependence coefficient (Ledford and Tawn, 1997) defined as

$$\eta_{ij} = \Pr(F_i(X_i) > q, F_j(X_j) > q) = (1-q)^{1/\eta_{ij}} \ell(1-q), \tag{2.16}$$

where the function $\ell : [0,1] \to \mathbb{R}$ is slowly varying at zero (Engelke and Ivanovs, 2021). η_{ij} effectively describes the rate of convergence of the joint exceedance probability to zero. Because it is also on the uniform scale, it is considered a measure of how dependent the variables are given that they are asymptotically independent, with 1 denoting total dependence and 0 independence.

Multivariate Block-Maxima Models

Multivariate block-maxima approaches naturally extend the univariate definition in 2.2.1. In the bivariate case, suppose $(X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n)$ is a bivariate sequence of random vectors with common distribution function F. For $M_{x,n} = \max_{i=1,...,n}\{X_i\}$ and $M_{y,n} = \max_{i=1,...,n}\{Y_i\}$, let $\boldsymbol{M}_n = (M_{x,n}, M_{y,n})$ be the vector of componentwise maxima. Then for $\boldsymbol{z} = (z_1, z_2)$

$$\Pr(\boldsymbol{M}_n \leq \boldsymbol{z}) = F(\boldsymbol{z})^n.$$

Let $\boldsymbol{a}_n > 0$, and $\boldsymbol{b}_n \in \mathbb{R}$ be normalising vectors such that $\boldsymbol{a}^{-1}(\boldsymbol{M}_n - \boldsymbol{b}_n)$ converges to a non-degenerate limit G, i.e.,

$$\Pr(\boldsymbol{a}^{-1}(\boldsymbol{M}_n - \boldsymbol{b}_n) \leq \boldsymbol{z}) = F^n(\boldsymbol{a}_n\boldsymbol{z} + \boldsymbol{b}_n) \to G(\boldsymbol{z}), \quad n \to \infty. \tag{2.17}$$

If (2.17) holds, F is said to be in the maximum domain (max-domain) of attraction of G, and G is called the bivariate extreme value distribution function (Beirlant 2004, BGEVD). The limiting distribution of the margins of G is GEVD.

If we assume that X_i and Y_i have standard Fréchet marginal distributions, then G takes the form

$$G(x, y) = \exp\{-V(x, y)\}, \quad x, y > 0, \tag{2.18}$$

where $V(x, y)$ is the exponent measure (Resnick, 1987), satisfying

$$V(x, y) = 2\int_0^1 \max\left(\frac{\omega}{x}, \frac{1-\omega}{y}\right) dH(\omega),$$

where H is a distribution in $[0, 1]$ with constraint

$$\int_0^1 \omega dH(\omega) = \frac{1}{2}.$$

We can see that, unlike the univariate case, the BGEVD has no unique parametric form because it depends on the distribution H. Different models can arise with different characterisations of H. For example, when H is

$$H(\omega) = \begin{cases} \frac{1}{2} & \text{if } \omega = 0 \text{ or } 1 \\ 0 & \text{else,} \end{cases} \tag{2.19}$$

then the resulting BGEVD is asymptotically independent with form

$$G(x, y) = \exp\left\{-\left(x^{-1} + y^{-1}\right)\right\}.$$

A perfectly dependent case of G arises if H places mass equal to 1 at 0.5, defined as

$$G(x, y) = \exp\left(-\max\left\{x^{-1}, y^{-1}\right\}\right).$$

Any parametric family of H that satisfies (2.19) results in a different BGEVD with specific properties and dependence structures (Toulemonde *et al.*, 2015). For additional parametric forms for H and G, see (Coles, 2001, Ch. 8).

Multivariate Threshold Exceedance Models

The generalised Pareto distribution can be extended to the multivariate case, as with the GEVD, where the bivariate case of the multivariate GPD is restricted to the asymptotically dependent case. Additionally, it has no single representation in the case of more than one dimension. It is most commonly presented using one of the four representations - R, S, T and U - proposed by Rootzén *et al.* (2018b,a). Here, we will focus on the so-called U representation, which is typically preferred over the others for simulation and can be used to model the dependence of data in the unit scale.

Let $\mathbf{U}$ be a random vector in $\mathbb{R}^d$ with density f_U under the condition that $0 < \mathrm{E}(e^{U_j}) < \infty \quad \forall j \in \{1, ..., d\}$. Then, a GPD density h_U can be constructed as

$$h_U(\mathbf{x};\mathbf{1},\mathbf{0}) = \frac{\mathbb{1}\{\max(\mathbf{x}) > 0\}}{\mathrm{E}[e^{\max(\mathbf{U})}]} \int_0^\infty f_U(\mathbf{x} + \log(t))\mathrm{d}t, \tag{2.20}$$

where $\mathrm{E}[e^{\max(\mathbf{U})}] = \int_0^\infty \Pr(\max(\mathbf{U}) > \log(t))\mathrm{d}t$. Kiriliouk *et al.* (2019) proposed a definition for f_U which allows for a more flexible construction. Let $\mathbf{V} \in \mathbb{R}^d$ be a random vector of independent components such that its joint density is the product of the independent marginal densities, $f_v(\boldsymbol{\nu}) = \prod_{j=1}^d f_j(\nu_j)$. No restrictions are placed on f_v, allowing almost any density to be used. Some distributions, however, result in simpler integrals of closed-form and thus are preferred to others. The reverse exponential distribution was chosen for this reason. It is defined as

$$f_j(\nu_j) = \alpha_j e^{\alpha_j(\nu_j + \beta_j)}, \quad -\infty < \nu_j < -\beta_j,$$

where $\alpha_j > 0$ is the scale parameter and $\beta_j \in \mathbb{R}$ is the location parameter. Substituting the product of these densities in (2.20) with $f_U = f_V$, yields the h_U density in closed form

$$h_U(\mathbf{x};\mathbf{1},\mathbf{0}) = \frac{(e^{-\max(\mathbf{x}+\boldsymbol{\beta})})^{\sum_{j=1}^d \alpha_j + 1}}{E[e^{\max(\mathbf{U})}]} \frac{1}{1 + \sum_{j=1}^d \alpha_j} \prod_{j=1}^d \alpha_j (e^{x_j + \beta_j})^{\alpha_j}. \tag{2.21}$$

Because (2.21) is fitted to standardised components in the unit scale, the model captures the dependence between components, with $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ as dependence parameters.

Kiriliouk *et al.* (2019) explores the possibility of common parameters, $\alpha = \alpha_1 = \alpha_2$ and/or $\beta = \beta_1 = \beta_2$, on standardised data. To ensure the identifiability of the location parameter, the parameter for the last component, β_2, was always fixed to 0 (Kiriliouk *et al.*, 2019).

Spatial Extreme Models

Spatial extreme models were first proposed by Coles and Casson (1998), who used a hierarchical Bayes model with a stationary correlation function to fit a point process likelihood to extreme wind speed data in the US Eastern seaboard. Rigorous inference for the spatial counterpart of max-stable distributions was provided by Padoan *et al.* (2010), opening the door to a rich assortment of methodologies for spatial extremes. As extreme value analysis is inherently temporal, the spatial extremes methodology is spatiotemporal, with replicates assumed available at every observation location.

The simplest spatial extreme model for block-maxima (or threshold exceedances) can be constructed by assuming that every location available has a different GEVD defined by a location-specific set of parameters so that for locations $s \in \mathcal{S} \subset \mathbb{R}^2$, every location has the distribution GEVD$(\mu(s), \sigma(s), \xi(s))$, where parameters $\mu(s), \sigma(s)$, and $\xi(s)$ vary smoothly in space. Youngman (2019) proposed using a generalised additive model (GAM) to model the parameters in space as functions of latitude and longitude. Other variants of the linear additive framework have also been proposed and are typically fitted using Bayesian inference (Davison *et al.*, 2012).

A common approach in spatial extremes is the assumption of a max-stable process $Z(s)$ (Haan, 1984), which can be asymptotically dependent or perfectly asymptotically independent (Richards and Wadsworth, 2021). The process is defined as having unit Fréchet marginal distributions as $\Pr(Z(s) \leq u) = \exp(-1/u)$ for any $u > 0$. The joint distribution is imposed as max-stable, meaning

$$\Pr(Z(s_1) \leq tu_1, ..., Z(s_r) \leq tu_r)^t = \Pr(Z(s_1) \leq u_1, ..., Z(s_r) \leq u_r),$$

for any $t > 0$, $r \geq 1$, $s_i \in \mathbb{R}^2$, $u_i > 0$, for $i = 1, ..., r$, which can be rewritten as

$$\Pr[Z(s) \leq u(s), \text{ for all } s \in \mathbb{R}^2] = \exp\left[-\int \max_{s \in \mathbb{R}^2} \left\{\frac{g(c,s)}{u(s)}\right\} \delta(\mathrm{d}s)\right], \tag{2.22}$$

where $g(\cdot)$ is a non-negative function that satisfies $\int g(c,s)\delta(\mathrm{d}s) = 1$.

For the spectral representation of a max-stable process, let $\{P_j^{-1}\}_{j=1}^{\infty}$ be realisations of a homogeneous Poisson point process of unit rate with intensity dp/p^2 and $\{W_j(x)\}_{j=1}^{\infty}$ be independent replicates of a stationary process $W(s)$ on $\mathbb{R}^p$ satisfying $\mathrm{E}[\max\{0, W_j(s_0)\}] =$

1, where s_0 is the origin. Then, the max-stable process $Z(s)$ is defined as

$$Z(s) = \max_j P_j \max\{0, W_j(s)\},$$

with unit Fréchet margins. Different choices of $W(s)$ lead to different max-stable models. The most common defines $W(s)$ as a stationary standard GP known as the Schlather model (Schlather, 2003). For more details, see Davison *et al.* (2012).

Copulas have also been proposed for spatial extremes. Sklar's theorem (Sklar, 1959) established that a D-dimensional joint distribution F of a random vector Y can be written as

$$F(y_1, ..., y_D) = C(\{F_1(y_1), ..., F_D(y_D)\}),$$

where $F_1, ..., F_D$ are the univariate marginal distributions of $X_1, .., X_D$ and C is a copula. If the copula satisfies the relationship

$$C(u_1, .., u_d) = C(u_1^{1/m}, ..., u_1^{1/m})^m \tag{2.23}$$

for every integer $m \geq 1$ and all $(u_1, ..., u_d) \in [0, 1]^d$, then the copula is max-stable. A copula C is an extreme-value copula when defined as

$$C(u_1, .., u_d) = \exp(\ell(-\log u_1, .., \log u_d)), \quad (u_1, ..., u_d) \in [0, 1],$$

if and only if there exists a finite Borel measure H on Δ_{d-1} such that the tail dependence function $\ell : [0, \infty)^d \to [0, \infty)$ is defined as

$$\ell(x_1, \ldots, x_d) = \int_{\Delta_{d-1}} \bigvee_{j=1}^{d} (w_j x_j)\, dH(w_1, \ldots, w_d), \quad (x_1, \ldots, x_d) \in [0, \infty)^d.$$

Various parametric models for extreme-value copulas have been proposed by specifying H, similarly to the bivariate GEVD, to define a dependence structure. Some notable mentions are the logistic model (or Gumbel-Hougaard copula), the negative logistic model, and the Hüstler-Reiss model. For more details, see Gudendorf and Segers (2010).

Subasymptotic and Flexible Dependence Models

While the classical spatial extremes models for the asymptotic dependence class is myriad, applications arise where data exhibit non-stationary dependence structures which might converge into asymptotic independence or display a weakening dependence structure (Huser and Wadsworth, 2019). Additionally, classical models are prohibitive at large dimensions, limiting their usefulness in large applications (Huser and Wadsworth, 2022; Huser *et al.*, 2024).

Subasymptotic models have been proposed as an alternative to the limiting but rigid max-stable models. Motivated by decaying dependence as the level of extremes of the event increases, Huser and Wadsworth (2022) suggests estimating the rate at which the sub-asymptotic χ-measure in (2.15) decays as $u \to 1$, Common models that capture this flexible, slow-convergence cases are inverted max-stable processes (Wadsworth and Tawn, 2012), mixture models such as those in Huser and Wadsworth (2019), factor copula models (Krupskii *et al.*, 2018; Castro-Camilo and Huser, 2020), or the max-infinitely divisible processes proposed by Huser and Wadsworth (2022). For more details on these approaches, see Huser and Wadsworth (2022).

2.3 Approaches for Statistical Data Fusion

Castanedo (2013) define data fusion as "the integration of data and knowledge from several sources". The concept was originally developed in the late 1970s by the US Department of Defence (DoD) for war-related purposes (Hall and Llinas, 1997). Since then, technological advances, data collection, and data storage have motivated the development of data fusion techniques. Today, they are a valuable tool in fields such as military and defense industry (Benaskeur and Rhéaume, 2007; Farina *et al.*, 2014; Chmielewski *et al.*, 2020; Noh, 2020; Vallikannu *et al.*, 2023), healthcare (Shoaib *et al.*, 2014; Dautov *et al.*, 2019; Issa *et al.*, 2022; Hassani *et al.*, 2024), environmental monitoring (Larios *et al.*, 2012; Beauchamp *et al.*, 2017; Okafor *et al.*, 2020; Dudek and Baranowski, 2023), finance and economics (Zhang *et al.*, 2013; Guo *et al.*, 2014; Li *et al.*, 2021; Yuan and Zhan, 2022), transportation and traffic management (Anand *et al.*, 2014; Neumann *et al.*, 2016; Zhao *et al.*, 2021; Zißner *et al.*, 2023), among many others.

Several schemes to define and classify the conceptual models of data fusion have been proposed in the literature (Durrant-Whyte, 1988; Luo and Kay, 1989; Dasarathy, 1997). However, specific definitions and classifications highly depend on the context of the application. Morabito *et al.* (2008) provides an initial categorisation of data fusion methodology as phenomenological or non-phenomenological. Phenomenological models are deterministic and rely on the physical properties of the underlying process to fuse data sources. Non-phenomenological models, also called statistical data fusion models, represent a probabilistic approach to the problem. Braverman (2014) defines statistical data fusion as "the process of combining statistically heterogeneous samples from marginal distributions to construct a new sample that can be regarded as having come from the unobserved joint distribution of interest".

In the context of environmental research, Castrignanò *et al.* (2017) proposes a division of data fusion into three further categories: information fusion (merging information from different sources), sensor fusion (simultaneous information from different sensors), and im-

age fusion (fusion of two or more images). However, the three categories are not mutually exclusive. In this literature review, we will cover the main methods commonly used to perform the aforementioned types of statistical data fusion. We will also provide a more tailored literature review on data fusion models commonly used in air quality monitoring and data fusion of extreme values.

The relative definitions of data fusion by application occur because of the diversity of data types and sources available in different applications. Castanedo (2013) offers to divide data fusion techniques into three non-exclusive or exhaustive categories: data association, state estimation, and decision fusion. Data association is mainly concerned with identifying the set of observations or measurements from one or various sources that have the same target over time. On the other hand, state estimation seeks to infer the true state of the target by using observations from one or various sources. Finally, decision fusion, which fuses data at a higher level, means the fusion of data inferred from the perceived situation. For each one of these data fusion types, the sources of data can be sensors (Ran *et al.*, 2018), images in the form of pixels or features (Agyeman *et al.*, 2023), or decisions, which include maps or other graphic derivatives (Castanedo, 2013).

Further considerations are the type of data, such as numeric or categorical. Multimodal data fusion, the process of fusing data of multiple types, is a useful technique that falls outside the scope of this work and will be excluded from the literature review below. This section covers the most important techniques in data fusion. Section 2.3.1 covers data fusion using geostatistical methods. Section 2.3.2 is about data fusion using Bayesian methods. Section 2.3.3 covers data fusion specifically used in air quality monitoring. Finally, Section 2.3.4 gives an overview of the methods available for extreme values.

2.3.1 Geostatistical Models

In the data fusion context, the process of interest, $Z(\boldsymbol{s})$, observed at particular locations $\boldsymbol{s} \in \mathcal{S} \subset \mathbb{R}^2$, can be considered a faithful measurement of the process of interest measured at ground level. As such, common properties of the data are often high-quality measurements but sparse spatial coverage. The predictor variable, $Y(\boldsymbol{s})$, can be a dataset measuring the same process but obtained from a different source, resulting in different properties such as high spatial coverage but larger measurement error. In this way, data fusion models aim to adjust the dataset with high spatial coverage, $Y(s_0)$, to match the ground observations of process $Z(s_0)$ at a location where Z has not been previously observed (s_0).

While universal kriging (UK; see 2.1.2) is a valuable and intuitive method for data fusion, several variants exist to accommodate different challenges in data fusion. In the case of fusing multiple data sources at different spatial resolutions, kriging with external drift (KED; see 2.1.2) was used by Ribeiro Sales *et al.* (2013) to fuse multiple satellite spectral

images for forest monitoring by treating the coarse and fine-resolution images as joint realizations of two autocorrelated and cross-correlated random fields where pixels of each resolution are assumed to be weighted averages of the underlying process. Variations have also been proposed for the cases where data size prohibits the inversion of the covariance matrix and, thus, the use of UK. Nguyen *et al.* (2012) proposed a method for the fusion of large (complementary) satellite spectral images based on fixed rank kriging (Cressie and Johannesson, 2008) that parametrises the covariance matrix through a spatial random effects model, allowing for fast inversion of large matrices. Jinnagara Puttaswamy *et al.* (2014) proposed an extension of universal Kriging to use sensor aerosol optical depth (AOD) using large satellite data. They partition the region and apply a Kriging model locally, linking the covariance functions of each partition using a low-rank linear model. Manzione and Castrignanò (2019) used various data sources with different support to map water table depth. They regularised covariate information from the sources using a linear model of coregionalization at block support to obtain block cokriging (CK) predictions of water table depth.

2.3.2 Fusion Models using the Bayesian Framework

Bayesian methods are a common technique for spatial data fusion, with a general formulation provided by Bogaert and Fasbender (2007). It is originally based on the idea that a spatial phenomenon of interest, $Z(s) \in \mathbb{R},\ s \in \mathcal{S}$, is not observed directly. Observations collected at location s, $Y(s)$, are observed with some error, $E(s)$, giving rise to the model

$$\mathbf{Y} = \mathbf{Z} + \mathbf{E},$$

which can be generalised as

$$\mathbf{Y} = g(\mathbf{Z}) + \mathbf{E},$$

where $g(\cdot)$ is some function denoting the relationship between the latent process Z and the observations Y.

A fusion model arises when the relationship between observations $Y_{n,j}$ and the latent process Z_n at location s_n is defined as

$$Y_{n,j} = g_j(Z_n) + E_{n,j} \quad \text{for } j = 1, ..., m, \tag{2.24}$$

where $E_{n,j}$ is a random error term. Bayes' theorem can be used to obtain the conditional pdf $f(\mathbf{z}|\mathbf{y}_a, \mathbf{y}_n)$, where $\mathbf{y}_n$ are realisations of $\mathbf{Y}_n = (Y_{n,1}, ..., Y_{n,m})'$ and $\mathbf{y}_a = (y_0, ..., y_n - 1)$ (where the subscript a always refers to the first $n-1$ elements of the corresponding vector), resulting in

$$f(\mathbf{z}|\mathbf{y}_a, \mathbf{y}_n) = \frac{f(\mathbf{y}_a, \mathbf{y}_n|\mathbf{z})f(\mathbf{z})}{\int_{\mathbb{R}^n} f(\mathbf{y}_a, \mathbf{y}_n|\mathbf{z})f(\mathbf{z})\mathrm{d}\mathbf{z}} = \frac{1}{A} f(\mathbf{y}_a, \mathbf{y}_n|\mathbf{z})f(\mathbf{z}),$$

where A is a normalising constant. The assumptions of independence between the process Z and the corresponding error terms $\mathbf{E}_a \perp \mathbf{Z}$ and $\mathbf{E}_n \perp \mathbf{Z}$, implies that

$$f(\mathbf{y}_a|\mathbf{z}) = f_{\mathbf{E}_a}(\mathbf{y}_a - \mathbf{g}(\mathbf{z}_a)); \quad f(\mathbf{y}_n|\mathbf{z}) = f_{\mathbf{E}_n}(\mathbf{y}_n - \mathbf{g}(z_n)).$$

Further assuming that $\mathbf{E}_a \perp \mathbf{E}_n$, leads to

$$f(\mathbf{y}_a, \mathbf{y}_n|\mathbf{z}) = f_{\mathbf{E}_a}(\mathbf{y}_a - \mathbf{g}(\mathbf{z}_a)) f_{\mathbf{E}_n}(\mathbf{y}_n - \mathbf{g}(z_n)). \tag{2.25}$$

This definition enables us to make a prediction of Z at a location s_0, using

$$f(z_0|\mathbf{y}_a, \mathbf{y}_n) = \frac{1}{A} \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} f(\mathbf{z}) f_{\mathbf{E}_a}(\mathbf{y}_a - \mathbf{g}(\mathbf{z}_a)) f_{\mathbf{E}_n}(\mathbf{y}_n - \mathbf{g}(z_n)) \mathrm{d}z_1 \cdots \mathrm{d}z_n. \tag{2.26}$$

Simplifying (2.25) leads to

$$f_{E_{n,j}}(y_{n,j} - g_j(z_n)) \propto \frac{f(z_n|y_{n,j})}{f(z_n)},$$

which allows for a simplification of (2.26) resulting in

$$f(z_0|\mathbf{y}) \propto \frac{f(z_0|y_0)}{f(z_0)} \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} f(\mathbf{z}) \phi(z_n|\mathbf{y}_n) \prod_{i=1}^{n-1} \frac{f(z_i|y_i)}{f(z_i)} \mathrm{d}z_1 \cdots \mathrm{d}z_n,$$

where $\phi(z_n|\mathbf{y}_n)$ is the posterior pdf of the fusion operator (Bogaert and Fasbender, 2007), i.e.

$$\phi(z_n|\mathbf{y}_n) \propto f(z_n)^{-m} \prod_{j=1}^{m} f(z_n|y_{n,j}). \tag{2.27}$$

In the case where spatial information is ignored, and all locations are independent, $f(z_0|\mathbf{y}_0) \propto \phi(z_0|\mathbf{y}_0)$ corresponds to the denominated *naive Bayes'* (or Idiot's Bayes) fusion rule. Another useful case is the Gaussian case, where all $f(z_n|y_{n,j})$ are the pdfs of $N(\mu_j, \sigma_j^2)$ variables, as we only require the conditional expectations, $\mathrm{E}[Z_n|y_{n,j}] = \mu_j$, and the conditional variance $\mathrm{Var}[Z_n|y_{n,j}] = \sigma_j^2$, for estimation.

This flexible construction gives way to an extensive family of models. Bogaert and Fasbender (2007) use the Gaussian case above to induce spatial dependence in the model, where the resulting Bayesian data fusion model is identical to the ordinary kriging (OK) model presented in Section 2.1.2, where the mean is constant but unknown (Cressie, 1993) and can be referred to as Bayesian kriging (Banerjee *et al.*, 2015). In the context of image-fusion, Fasbender *et al.* (2007) used the model to improve the spatial resolution of target phenomenon by using spectral imaging with known relationships for $g(\cdot)$ and placing uninformative priors on $f(\mathbf{z})$. Xue *et al.* (2017a) proposed using the Maximum A Posterior (MAP) estimator as the predictor of z_0 but placed a further spatiotempo-

ral Gaussian structure on E. Gengler and Bogaert (2014) propose an extension of the model for categorical variables, and Gengler and Bogaert (2016) applied this extension to update landcover data by fusing remove-sensed images of land cover and crowdsourced observations.

A noteworthy special case of this approach is Bayesian downscaling, which refers to relating coarse-scale data to fine-scale data without assuming a latent "true" process. It is accomplished by using ground-level observations such as those from observation stations to improve the scale of coarse-scale data and is also known as "data calibration". (Gelfand *et al.*, 2003), and later (Berrocal *et al.*, 2010) proposed a spatial downscaling model with spatially varying coefficient defined as

$$Y(\mathbf{s}) = \mu(\mathbf{s}) + W(\mathbf{s}) + \epsilon(\mathbf{s}),$$

where μ is a function of a covariate x at the same location (collocated), such that $\mu(\mathbf{s}) = \alpha(\mathbf{s}) + \beta(\mathbf{s})x(\mathbf{s})$, $\epsilon(\mathbf{s})$ are independent, normally distributed error terms centred at 0 and with τ^2 variance, and $W(\mathbf{s})$ is a second-order stationary mean 0 process that is independent of ϵ. Then, it is intuitive to see that $W(\mathbf{s}) = \alpha(\mathbf{s})$ results in a model that is well-poised for a hierarchical structure in the Bayesian framework. Data fusion occurs in the case that x are data for the same phenomenon from a different source, such as remote-sensing, and $Y(\mathbf{s})$ represent in-situ measurements.

Bayesian Melding

Bayesian melding is a Bayesian data fusion approach that considers the process of interest, $Z(\mathbf{s})$, to be a common latent spatial process of the different data sources (Fuentes and Raftery, 2005). This common latent Gaussian process is assumed to follow the model

$$Z(\mathbf{s}) = \mu(\mathbf{s}) + \epsilon(\mathbf{s}),$$

where $\mu(\mathbf{s})$ can be a function of explanatory variables and $\epsilon(\mathbf{s})$ are zero-mean correlated errors. The point observations taken at ground level are not the "true" observations, but they can be defined as

$$\hat{Z}(\mathbf{s}) = Z(\mathbf{s}) + e(\mathbf{s}),$$

where e is a measurement error independent of Z and $e(\cdot) \sim N(0, \sigma_e^2)$. Data from a second source with more spatial coverage is defined as a more flexible function of Z:

$$\tilde{Z}(\mathbf{s}) = a(\mathbf{s}) + b(\mathbf{s})Z(\mathbf{s}) + \delta(\mathbf{s}),$$

where $a(\mathbf{s})$ and $b(\mathbf{s})$ are parameter functions. The process $\delta(\mathbf{s})$ is once again a random deviation term as $\delta(\mathbf{s}) \sim N(0, \sigma_\delta^2)$, and is independent of Z and e. In the case that the data

are not point-estimates but areal estimations over subregions $B_1, ..., B_m$ of the domain D, then the model becomes

$$\tilde{Z}(\mathbf{s}) = \int_{B_i} a(\mathbf{s})\mathrm{d}\mathbf{s} + b(\mathbf{s}) \int_{B_i} Z(\mathbf{s})\mathrm{d}\mathbf{s} + \int_{B_i} \delta(\mathbf{s})\mathrm{d}\mathbf{s}, \quad \text{for } \mathbf{s} \in B_i.$$

A spatial prediction then follows naturally as $P(Z|\hat{Z}, \tilde{Z})$.

Alkema *et al.* (2007) used Bayesian melding to fuse HIV antenatal clinic prevalence to population prevalence by inserting random effects to account for different clinics. Villejo *et al.* (2023) used Bayesian melding to link NO_2 pollution exposure to various health outcomes by assuming a latent Gaussian process and fitting the melding modelling using the integrated nested Laplace approximation (INLA) and the stochastic partial differential equations (Lindgren *et al.* 2011, SPDE) approaches. Liu *et al.* (2016) used an approximated likelihood to perform Bayesian melding to model marine mammal movement by fusing a sparse set of direct observations and high-resolution but high-bias modelled tracks.

2.3.3 Data Fusion in Air Quality Monitoring

Data fusion in application to air quality mostly comprises multiple sensor fusion and multi-source downscaling, where fine-scale information is obtained from a coarse-scale source with the aid of ground observations or other ancillary data, often including epidemiological variables such as exposure. Characteristic challenges of air quality monitoring data fusion are multi-source heterogeneity, dynamic mutability, spatio-temporal correlation, and sometimes large data volumes (Huang *et al.*, 2021).

Geostatistical Models

Geostatistical tools have been a popular approach to data fusion in air quality monitoring. Li *et al.* (2014) modelled remote-sensed aerosol optical thickness (AOD) observations in eastern China by using a kriging with external drift model fuse data from different remote-sensed AOD sources using other meteorological variables as explanatory variables. Schneider *et al.* (2017) used universal kriging to fuse air quality data from low-cost sensors with an urban-scale air quality map of Oslo, Norway, to produce a map of nitrogen oxide at urban scale.

In the case that the spatial and the temporal domains are fused separately, Friberg *et al.* (2016) and Liang *et al.* (2017) used a weighted average of an ordinary kriging interpolation of the data that captured the spatial correlation and a scale adjustment of the annual mean to fuse chemical transport model data and ground observations in the state of Georgia and northern China respectively. Chatterjee *et al.* (2010) used the spatio-temporal extension of universal kriging to fuse remote sensing and ground-based AOD.

Munir *et al.* (2021) used a simple universal kriging model with spatiotemporal covariance function to fuse NO_2 data from a dispersion model, land cover, and concentrations from low-cost sensors in Sheffield, United Kingdom.

Other kriging variants have been widely used. Lin *et al.* (2020) fused ground-level observations of $PM_{2.5}$ from high-quality observation stations with low-quality sensors with higher spatial coverage. They used multi-step Kriging, which fits a kriging model at every time slice at $t + 1$ by utilising the fitted model at t. Beloconi *et al.* (2016) approached spatio-temporal data fusion using 3-D Kriging, which incorporates a spatio-temporal covariance function. They performed the fusion of remote-sensed Sky-viewing Factor (SVF) with Land Cover to predict the PM concentrations at ground level. They later upscaled the resulting daily concentrations of the pollutant by using block kriging. Ghigo *et al.* (2018) pursued a similar approach to map common pollutants (PM_{10} and NO_2) to regional municipalities in Spain. They performed kriging with external drift to fuse ground-level observations with remote-sensing estimates. They later performed a weighted mean to upscale the concentrations to the scale of desired municipalities.

A more complex process of data fusion using Kriging was proposed by Xue *et al.* (2017b), by using a three-step procedure to fuse ground-level observations of $PM_{2.5}$, data from a multi-scale air quality model, and other explanatory variables in mainland China. In the first steps, they used linear mixed models to derive $PM_{2.5}$ data using explanatory data and ground-level observations and to calibrate modelled observations using the ground-level $PM_{2.5}$ data. In the second step, they fuse the derived-$PM_{2.5}$ and the calibrated data using inverse deviation weighted averages. Finally, they interpolate the fused observations using spatiotemporal Kriging.

Bayesian Approaches

As mentioned in Section 2.3.2, the Bayesian paradigm offers a convenient framework for data fusion models in spatial and spatiotemporal dimensions.

For instance Beloconi *et al.* (2016) applied the Bayesian geostatistical regression model in Bogaert and Fasbender (2007) to fuse NO_2 concentration data from observation stations, satellites, and large physical models to obtain reliable, high-resolution maps.

McMillan *et al.* (2010) fused log-$PM_{2.5}$ concentrations in the eastern USA. The data were from ground-level observations and modelled high temporal and spatial coverage observations. Specifically, they used the model in (2.24) with $g(\cdot)$ chosen as a a non-linear function and performed inference using MCMC. Chang *et al.* (2014) proposed a Bayesian spatio-temporal downscaling model using satellite-collected aerosol optical depth (AOD) measurements to model $PM_{2.5}$ concentrations using the model:

$$PM(s,t) = \alpha_0(s,t) + \alpha_1(s,t)\text{AOD}(s,t) + \epsilon(s,t),$$

where $\alpha_0(s,t)$ and $\alpha_1(s,t)$ are the spatiotemporal additive and multiplicative bias respectively. The hierarchical setup of the model allows for the introduction of spatiotemporal trends via:

$$\begin{aligned}\alpha_0(s,t) &= \beta_0(s) + \beta_0(t) + \gamma_0 X_0 \\ \alpha_1(s,t) &= \beta_1(s) + \beta_1(t) + \gamma_1 X_1,\end{aligned}$$

where $\beta_i(s)$ and $\beta_i(t)$ are unobserved correlated spatial and temporal random effects, X_0 and X_1 represent predictor variables and γ_0 and γ_1 are fixed effect coefficients. Wang *et al.* (2018) extended the model for the same application to include discrete regions of interest (political regions sharing a common climate), meaning

$$PM(s,t) = \alpha_0(s,t) + \alpha_1(s,t)\text{AOD}(s,t) + \gamma_{\text{reg,tem}} X(s,t) + \epsilon(s,t), \tag{2.28}$$

where {reg,tem} refers to regional (discrete) and temporal terms, $X(s,t)$ is a spatiotemporal predictor at a discrete spatial scale and $\gamma_{\text{reg,tem}}$ is its corresponding fixed effect.

Forlani *et al.* (2020) used a coregionalisation framework (Krainski *et al.*, 2018) to perform Bayesian melding to fuse NO_2 concentrations from observation stations and data from two different air quality models in the London region. They allowed for each of the different data sources to provide different spatial or spatio-temporal information:

$$\begin{aligned}y_1(s,t) \sim& N(\eta_1(s), \sigma^2_{\epsilon_1}), \\ y_2(s,t) \sim& N(\eta_2(s,t), \sigma^2_{\epsilon_2}), \\ y_3(s,t) \sim& N(\eta_3(s,t), \sigma^2_{\epsilon_3}),\end{aligned}$$

where y_1 and y_2 are data from remote sensing sources, i.e., the pollution climate mapping (PCM) model and the Air Quality Unified Model (AQUM) sources, respectively, y_3 represents the ground-level observations, η_i are the mean, and $\sigma^2_{\epsilon_i}$ are measurements of error variance. The data sources are woven together through coregionalisation as

$$\begin{aligned}\eta_1(s) =& \alpha_1 + z_1(s), \\ \eta_2(s,t) =& \alpha_2 + \lambda_{1,2} z_1(s) + z_2(t), \\ \eta_3(s,t) =& \alpha_3 + \beta_k + \lambda_{2,3} z_2(t) + z_3(t, k_s),\end{aligned}$$

where α_i are intercepts, $\lambda_{i,j}$ are scaling parameters for the shared components, z_i are shared random spatial, temporal, or spatiotemporal components, and β_k are fixed effects to account for information on the location. Model inference was carried out using INLA (Rue *et al.*, 2009), under the assumption that the latent process is Gaussian.

2.3.4 Data Fusion Approaches for Extreme Values

Data fusion for extremes is not a common approach and has largely centred around downscaling extreme precipitation. Downscaling refers to the process of obtaining information at a finer scale from data at a coarse scale. Although it can be done doing non-data fusion methods, here we will only refer to downscaling as a data fusion approach where in situ measurements are used to downscale remote-sensing data at coarse resolutions. In this section, we give an overview of data fusion and downscaling approaches in the literature that have been proposed for extremes.

Foufoula-Georgiou *et al.* (2014) proposed a framework for downscaling spatial satellite precipitation observations by defining the downscaling process as a discrete inverse problem and solving it using a variational regularisation approach. In this way, the model imposes constraints on the smoothness of the precipitation field while preserving large gradients. To illustrate the approach, assume $f(t)$ is the true signal observed at a given location. The measurement device introduces an error and thus smooths the original state, returning the observation $g(s)$. The two are related as

$$\int_0^1 K(s,t)f(t)\mathrm{d}t = g(s), \quad t \leq 1, \tag{2.29}$$

where $K(s,t)$ is a known kernel responsible for the downgrade in resolution, effectively smoothing the true signal. The discretisation of (2.29) can be translated into the regression problem

$$\mathbf{y} = \boldsymbol{H}\mathbf{x} + \boldsymbol{\epsilon}, \tag{2.30}$$

where $\boldsymbol{\epsilon} \sim N(0, \Sigma)$, $\boldsymbol{H}$ matrix operator, and $\mathbf{x}$ is a vector of observations at high resolution. The nature of downscaling means $\boldsymbol{H}$ is a rectangular matrix with more columns than rows and solving for $\mathbf{x}$ is ill-posed.

Foufoula-Georgiou *et al.* (2014) proposed a regularisation of the problem. They define the distance between the observations and the true state using a residual Euclidean norm as

$$R(f) = \left|\left| \int_0^1 K(s,t)f(t)\mathrm{d}t - g(s) \right|\right|_2 .$$

A unique and stable solution of the inverse can then be posed as a variational minimisation problem as

$$f(t) = \underset{f}{\operatorname{argmin}} \left\{ R(f)^2 + \lambda^2 S(f) \right\},$$

where λ is a regularisation parameter to balance precision and smoothing. Combining all the above, the high-resolution vector $\mathbf{x}$ can be obtained by solving the minimisation problem

$$\hat{\mathbf{x}} = \underset{x}{\operatorname{argmin}} \left\{ \frac{1}{2} ||\mathbf{y} - \boldsymbol{H}\mathbf{x}||^2_{\mathbf{R}^{-1}} + \lambda S(\mathbf{x}) \right\}.$$

Although the estimation of $\mathbf{x}$ is possible from $\mathbf{y}$ in a way that guarantees the preservation of extremes, it is important to note that the downscaling operator $\mathbf{H}$ is not known in most applications. Even after careful estimation, the method is limited to stationary fields and does not easily scale to larger datasets.

Engelke *et al.* (2019) highlight that in some cases the marginal tail behaviour is different from the spatially aggregated data from some process Y, aggregated using some functional ℓ. They show that for sufficiently large n,

$$\Pr\left\{\frac{\ell(Y)-\ell[b(n)]}{\ell[a(n)]}>y\right\}\approx\theta^{\ell}\Pr\left\{\frac{Ys_0-b_{s_0}(n)}{a_{s_0}(n)}>y\right\},\quad y\in\mathbb{R},\quad s_0\in S,$$

meaning that θ^{ℓ}, called the ℓ-extremal coefficient, quantifies this difference between the marginal tail behaviour at s_0, and the spatially aggregated $\ell(Y)$. Effectively, this coefficient provides a link between the aggregated data and the underlying process, a central aim of downscaling, which is a type of data fusion which they use for downscaling daily temperature maxima in the south of France by maximising the censored log-likelihood. The assumption of a constant θ^{ℓ} means the approach is useful under stationary assumptions. Moreover, it assumes that the difference between the tail behaviour of the aggregated data and the point data are the same throughout the region, which Maraun and Widmann (2018) show is not true in most applications.

Another example of data fusion for extremes found in the literature is the calibration model based on the extended generalised Pareto distribution (EGPD, Papastathopoulos and Tawn 2013) initially proposed by Pereira *et al.* (2019) and extended by Amaral Turkman *et al.* (2021) to the spatiotemporal case. Both methods are based on quantile matching calibration where the observed data Y can be obtained from a simulated data set of lower rank X by "calibrating" X to match Y such as

$$x_i^*=F_Y^{-1}(F_X(x_i)),\quad i=1,...,n, \tag{2.31}$$

where x_i^* is the new calibrated data point. Pereira *et al.* (2019) proposed an adjustment where the distribution of both X and Y change with a covariate. The introduction of the EGPD (Naveau *et al.*, 2016) allows for modelling extreme values without the need to set a threshold and thus can effectively model the bulk and the tail of the data. It is defined as

$$F_Y(y|\theta)=G\left(H(y|\xi,\sigma)\right),$$

where H is the cummulative distribution function of a GPD and G is a function obeying some general assumptions to ensure a Pareto-type tail and a bulk driven by the carrier G. While Naveau *et al.* (2016) proposes various options for G, Pereira *et al.* (2019) and Amaral Turkman *et al.* (2021) choose $G(u)=u^{\kappa}$, where κ is a parameter controlling the shape

of the lower tail, making the EGPD a three-parameter distribution as $EGPD(\kappa, \xi, \sigma)$. Under the assumption that both X and Y follow an EGPD distribution, the calibration method in (2.31) can be re-written for $\xi \neq 0$ and $\delta = -\frac{\sigma}{\xi}$ as

$$F_{X(s,t)}\left(x(s,t) \mid \delta_x(s,t), \xi_x, \kappa_x\right) = \left(1 - \left(1 - \frac{1}{\delta_x(s,t)} x(s,t)\right)_+^{-1/\xi_x}\right)^{\kappa_x},$$

for $x > 0$ if $\xi_x > 0$ and $x < \delta_x$ if $\xi_x < 0$. And

$$F_{Y(s,t)}\left(y(s,t) \mid \delta_y(s,t), \xi_y, \kappa_y\right) = \left(1 - \left(1 - \frac{1}{\delta_y(s,t)} y(s,t)\right)_+^{-1/\xi_y}\right)^{\kappa_y}$$

under the same respective assumptions. Amaral Turkman *et al.* (2021) then introduces a spatiotemporal dependence structure by modelling δ_x and δ_y as a function of a common latent spatiotemporal process, meaning $\delta_y(i,j) \sim \text{Exp}(\lambda_y(i,j))$, $\delta_y(i,j) > \max(y)$ follows a shifted exponential distribution with

$$\log(\lambda_y(i,j)) = \beta_y + W(s_i) + Z(t_j), \tag{2.32}$$

where $Z(t_j)$ is a temporal random walk process or order 1, and $W \sim MVN(0, \tau_W \Sigma_W)$ follows a Multivariate Gaussian process with precision τ_W and the matrix Σ_W with unit diagonals and non-diagonal elements with spatiotemporal structure as $\Sigma_{i\ell} = f(d_{i\ell}; \alpha)$ where $d_{i\ell}$ represents the centroid of every two stations s_i and s_ℓ, and α is a parameter representing the radius of the "disc" centred at each s, controlling the rate of correlation decline with distance. The data $X(s,t)$ then shares the same latent processes W and Z in a similar manner. Inference on the model is carried out using MCMC, and the authors present a case study for extreme wind speed data where they show the model has a good prediction coverage of observations in both the bulk and the tail. However, bias is present in the bulk when data is very extreme and uncertainty bands are wide, limiting the reliability of the results. Furthermore, extension to large spatial extents is not trivial and additional assumptions must be placed on the spatiotemporal structure.

In more applied settings, Kallache *et al.* (2011) proposed a quantile-matching approach to downscaling extreme precipitation. For observation station data Y and modelled data X, a calibration period C is defined, whereby observations within this period, Y_C and X_C, are considered as the training set. F_{Y_C} and F_{X_C} are the cumulative distribution functions (cdfs) of Y and X, respectively, fitted for observations in the calibration period. The downscaling model proposed is

$$F_{Y_P}(x) = T(F_{X_P}(x)) = F_{Y_C}(F_{X_C}^{-1}(F_{X_P}(x))),$$

where $T(\cdot)$ is a transfer function, and Y_P and X_P refer to the observed data and modelled data at a prediction period P. The transfer function is a linear regression with a wide variety of pertinent meteorological and geograhical covariates. The results show mixed results, with prediction periods correctly matched at some locations and no improvement at others.

Friederichs (2010) utilised ERA40 data to model conditional 95−th quantiles of precipitation in Germany and fitted a GPD to the exceedances of the conditional threshold quantile using MLE. The results show that the approach has improved uncertainty estimates over the non-parametric Quantile Regression, particularly far into the upper tail. Finally, Ebtehaj and Foufoula-Georgiou (2010) proposed using a mixture of Gaussian distributions to preserve the extremes during the fusion of multi-sensor precipitation data. The work was never formalised, and attempts to contact the authors have not yielded more information; consequently, the remaining section will not include this approach.

2.4 Methods for Bayesian Inference

Bayesian inference is a framework for statistical inference that updates prior beliefs using information the data provides. It represents an alternative to frequentist inference, and it has proven a viable way to perform inference for complex models such as hierarchical models or multi-stage models requiring uncertainty propagation.

Gelman *et al.* (2015) summarises the process of Bayesian data analysis as: 1) Setting up a full probability model, 2) Conditioning the model on the observed data, i.e., calculating the posterior distribution by conditioning the unobserved quantities of interest on the observed data, and 3) Evaluating the model fit. The conclusions of Bayesian inference are probabilistic statements about a parameter θ or about the unobserved data $\tilde{y}$, conditioned on the observations y, written as $p(\theta|y)$ or $p(\tilde{y}|y)$.

The first step for making probabilistic statements about θ given y is constructing a joint model. It is obtained as the product of the distribution of θ, known as the *prior distribution*, and a likelihood $p(y|\theta)$, resulting in

$$p(\theta, y) = p(\theta)p(y|\theta).$$

Bayes' rule can then be used to obtain a *posterior* density defined as

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)} = \frac{p(\theta)p(y|\theta)}{p(y)}, \tag{2.33}$$

where $p(y) = \int_\theta p(\theta)p(y|\theta)$, meaning $p(y)$ is the integral over all possible values of θ (Gelman *et al.*, 2015). The normalising denominator $p(y)$ can be removed from (2.33),

resulting in an unnormalised posterior density defined as:

$$p(\theta|y) \propto p(\theta)p(y|\theta). \tag{2.34}$$

Obtaining $p(\theta|y)$ in (2.34) is step 2, following naturally from the definition of $p(\theta, y)$ in (2.33).

If the desired target is some unobserved quantity, $\tilde{y}$, predictive inference is performed similarly. First, we defined the distribution of y, also known are the marginal distribution of y or the prior predictive distribution, as

$$p(y) = \int p(y, \theta)d\theta = \int p(\theta)p(y|\theta)d\theta.$$

The distribution of the unobserved value that is predicted, $\tilde{y}$, is called the posterior predictive distribution and is defined as

$$\begin{aligned} p(\tilde{y}|y) &= \int p(\tilde{y}, \theta|y)d\theta \\ &= \int p(\tilde{y}|\theta, y)p(\theta|y)d\theta \\ &= \int p(\tilde{y}|\theta)p(\theta|y)d\theta. \end{aligned} \tag{2.35}$$

The simple probability principles above are the basis of Bayesian inference and represent a flexible statistical inference framework. In the remainder of this section, we will cover the basics of the two central methodologies to perform Bayesian inference: Markov Chain Monte Carlo (MCMC, Gelfand and Smith (1990)) and integrated nested Laplace approximations (INLA, Rue *et al.* (2009)).

2.4.1 Markov Chain Monte Carlo (MCMC)

In order to make probabilistic statements of θ and or $\tilde{y}$, samples are taken from the posteriors in (2.33) and (2.35) respectively. Markov chain simulation, also known as Markov chain Monte Carlo (MCMC), is a method to sample from the posteriors by drawing samples from approximate distributions and progressively correcting the draws to better approximate the target distribution. The draws are taken sequentially following a Markov chain, where new draws depend on the last value drawn (Gelman *et al.*, 2015). In this section, we will only cover the Gibbs sampler and Metropolis-Hastings algorithms; however, many variants of MCMC have been proposed for specific settings; notable variants are Hamiltonian Monte Carlo, slice sampling, sequential Monte Carlo, delayed-rejection Monte Carlo, importance sampling, and adaptive MCMC. For more details, see Brooks *et al.* (2011).

Gibbs Sampler

The Gibbs sampler is a useful method for multidimensional problems where $\boldsymbol{\theta}$ is a d-dimensional vector defined as $\boldsymbol{\theta} = (\theta_1, ..., \theta_d)$. As the sampling works sequentially, each draw of θ_j is conditional on all the others, meaning there are d steps in every iteration t. For each iteration t, θ_j^t is sampled from the conditional distribution defined as

$$p(\theta_j | \theta_{-j}^{t-1}, y),$$

where θ_{-j}^{t-1} represent all the components of $\boldsymbol{\theta}$ minus θ_j at their current values, meaning $\theta_{-j}^{t-1} = (\theta_1^t, ..., \theta_{j-1}^t, \theta_{j+1}^{t-1}, ..., \theta_d^{t-1})$. In summary, each θ_j is updated conditional on the data y and latest values of all the other components of $\boldsymbol{\theta}$, which contain updated components at t and components not yet updated at $t-1$. The Gibbs sampler is particularly well suited for hierarchical models where sequential parameter updating is natural. Additionally, it is appropriate for the case where the posterior has a closed-form solution.

Metropolis-Hastings Algorithm

The Metropolis-Hastings algorithm (MH) is a general term for a family of Markov chain simulations. It has the Gibbs sampler as a special case, but it represents a more flexible approach to posterior sampling. It is based on the Metropolis algorithm, which adapts a random walk with an acceptance/rejection rule to guide convergence to the target distribution. The steps of the algorithm are as follows:

1. Obtain a starting value of θ, θ^0, for which $p(\theta^0|y) > 0$.

2. For $t = 1, 2, ...$:

   - Sample θ^* from a proposal distribution at time t, $J_t(\theta^*|\theta^{t-1})$. The proposal distribution should resemble the target as much as possible to speed convergence.
   - Calculate the ratio of densities

   $$r = \frac{p(\theta^*|y)}{p(\theta^{t-1}|y)}. \tag{2.36}$$

   - Set

   $$\theta^t = \begin{cases} \theta^* & \text{with probability } \min\{r, 1\} \\ \theta^{t-1} & \text{otherwise.} \end{cases}$$

In the case that the proposal distribution is not symmetric or that there are imposed bounds on the possible values of θ resulting in truncated sampling, then an adjustment is

made to the ratio of densities as

$$r = \frac{p(\theta^*|y)J(\theta^{t-1}|\theta^*)}{p(\theta^{t-1}|y)J(\theta^*|\theta^{t-1})}.$$

The acceptance/rejection, therefore, ensures that progressively better and better values of θ are chosen under the proposed posterior. It does not require a closed-form posterior, and the product of prior and likelihood can be used instead. The Gibbs sampler, therefore, arises as a special case of the MH algorithm where the exact posterior is known and the proposed value is accepted with a probability of 1.

2.4.2 Integrated Nested Laplace Approximation (INLA)

The major limitation to performing Bayesian statistics is the computational feasibility of doing Bayesian inference. Although MCMC and the introduction of simulation-based inference represent a significant step towards accessible Bayesian inference, performing inference with MCMC is computationally expensive, time-consuming, and usually does not scale well. Within the general class of latent Gaussian models (LGMs), an alternative to MCMC is the integrated nested Laplace approximation (INLA). In INLA, posterior distributions are numerically approximated using the Laplace approximation and variational Bayes, constituting a computationally appealing method for Bayesian inference.

As mentioned above, INLA applies to the general class of LGMs, which can be defined using a three-stage hierarchical model formulation (Rue *et al.*, 2017). Here, let $\boldsymbol{y}$ represent observations that are conditionally independent given the latent Gaussian random field $\boldsymbol{x}$ and hyperparameters $\boldsymbol{\theta}$, i.e.,

$$\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}_1 \sim \prod_{i \in \mathcal{I}} p(y_i|x_i, \boldsymbol{\theta}_1),$$

where x_i is the i-th component of the latent Gaussian field $\boldsymbol{x}$, defined as

$$\boldsymbol{x}|\boldsymbol{\theta}_2 \sim N(\boldsymbol{\mu}(\boldsymbol{\theta}_2), \boldsymbol{Q}^{-1}(\boldsymbol{\theta}_2)).$$

Following Bayes' rule, the posterior becomes

$$p(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{\theta})p(\boldsymbol{x}|\boldsymbol{\theta}) \prod_{i \in \mathcal{I}} p(y_i|x_i, \boldsymbol{\theta}),$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$.

Rue *et al.* (2017) show three critical assumptions placed on latent Gaussian fields in the context of INLA. First, the number of hyperparameters $|\boldsymbol{\theta}|$ is relatively small (< 20). Second, the distribution of the latent field $\boldsymbol{x}|\boldsymbol{\theta}$ is Gaussian and makes up a sparse Gaussian

Markov random field (GMRF). And finally, the data $\boldsymbol{y}$ are conditionally independent of both $\boldsymbol{x}$ and $\boldsymbol{\theta}$.

In this context, the response variable $\boldsymbol{y}$ with density function $p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta})$ is related to the covariates $\boldsymbol{Z} = (\boldsymbol{M}, \boldsymbol{U})$ through a linear predictor defined as

$$\boldsymbol{\eta} = \beta_0 + \boldsymbol{\beta}\boldsymbol{M} + \sum_{k=1}^{K} f^k(\boldsymbol{u}_k), \tag{2.37}$$

where $\boldsymbol{f}$ are flexible functions of $\boldsymbol{U} = (\boldsymbol{u}_1, ..., \boldsymbol{u}_k)$, $\boldsymbol{\beta}$ are linear coefficients for the deterministic effect of the covariates $\boldsymbol{M}$. INLA uses Laplace approximations to estimate $\boldsymbol{x} = \{\beta_0, \boldsymbol{\beta}, \boldsymbol{f}\}$. This additive formulation enables generalised mixed models, generalised additive models, splines, and many other linear formulations to be considered in the modelling structure.

In the classical formulation of INLA (Rue *et al.*, 2009), the posterior of the hyperparameters is approximated using the Laplace method as

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \propto \left. \frac{p(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y})}{p_G(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})} \right|_{\boldsymbol{x}=\boldsymbol{\mu}(\boldsymbol{\theta})},$$

where $p_G(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})$ is a Gaussian approximation to $p(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{y})$. In the modern formulation of Van Niekerk *et al.* (2023), the latent field $\boldsymbol{x} = \{\beta_0, \boldsymbol{\beta}, \boldsymbol{f}\}$ has Gaussian prior $\boldsymbol{x}|\boldsymbol{\theta} \sim N(\boldsymbol{0}, \boldsymbol{Q}^{-1}(\boldsymbol{\theta}))$, and the n linear predictors are defined as

$$\boldsymbol{\eta} = \boldsymbol{A}\boldsymbol{x},$$

where $\boldsymbol{A}$ is a sparse design matrix that links the linear predictors to the latent field. The joint density then becomes

$$p(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{\theta})p(\boldsymbol{x}|\boldsymbol{\theta}) \prod_{i=1}^{n} (y_i|(\boldsymbol{A}\boldsymbol{x})_i, \boldsymbol{\theta}).$$

The results, however, could be skewed with a presumptive Gaussian approximation. For this reason, Van Niekerk *et al.* (2023) apply a low-rank variational Bayes correction (Niekerk and Rue, 2024) to improve the mean of the marginal posteriors of the linear predictors and the latent field.

Chapter 3

Environmental Pollution

In this section, we provide the motivation and context for the bivariate heavy metal soil contamination application in Chapters 4 and 5, and the $PM_{2.5}$ air pollution case study in Chapter 6. The chapter is as follows. Section 3.1 introduces heavy metal soil contamination and provides an overview of the sources, impacts, and management of soil contamination. Section 3.2 gives an exploratory analysis of the data used in Chapters 4 and 5. Section 3.3 provides background, including sources, impacts, and relevant policy, of the $PM_{2.5}$ air pollution application in Chapter 6. Finally, Section 3.4 provides an exploratory analysis of the two datasets used in the application.

3.1 Heavy Metal Soil Contamination

3.1.1 Definition of HM Soil Contamination

Heavy metals (HM) are defined as metallic elements with atomic mass greater than 20 and specific gravity greater than 5. The most common HM contaminants are mercury (Hg), cadmium (Cd), lead (Pb), chromium (Cr), arsenic (As), zinc (Zn), copper (Cu), nickel (Ni), stannum (Sn), and vanadium (V).

HM soil contamination refers to the excessive accumulation of toxic HM elements in the soil (Su *et al.*, 2014; Tang *et al.*, 2019; Mishra *et al.*, 2019). It is characterised by its wide spatial distribution, strong latency, irreversibility, and complex multivariate nature (Su *et al.*, 2014). Unlike organic types of contamination, HM contamination persists in the pedosphere and is difficult to remediate, as remediation techniques require large financial investments over long periods (Su *et al.*, 2014). HMs are also biologically toxic, meaning they play a significant role in the degradation of the quality of the soil, water bodies, the atmosphere, ecological and plant health, and, ultimately, public health. More details about the impacts of HM contamination on both the environment and public health are given in Section 3.1.3.

3.1.2 Sources of Contamination

The total HM soil content is the sum of the content produced by all possible sources. At any given moment, the HM content in the soil, M_{total}, can be obtained in the unit of parts per million (ppm) by the sum

$$M_{total} = (M_{pm} + M_{atm} + M_{sed} + M_{fert} + M_{ac} + M_{tm} + M_{om} + M_{ic}) - (M_{cr} + M_e + M_l + M_v),$$

where pm is the content from parent material, atm is atmospheric deposition, sed is deposited sediment, $fert$ is fertilisers, ac are agricultural chemicals, tm are technogenic materials, om are organic materials, ic are inorganic contaminants, cr is crop removal, e is soil erosion, l is leaching, and v is volatilisation.

Lithogenic Sources

The lithogenic sources, natural geological processes that produce soil parent materials, are the dominant factor in determining the M_{total}. At a global scale, 99% of the total element content of the earth's crust is comprised of oxygen (O), silicon (Si), aluminium (Al), iron (Fe), calcium (Ca), sodium (Na), potassium (K), magnesium (Mg), phosphorus (P), and titanium (Ti). The elements that make up the remaining 1%, such as HMs, are considered *trace* elements and are naturally found in the soil at small concentrations but can be found at higher levels depending on parent material. For example, igneous and sedimentary rocks, which make up 95% and 5% of the earth's surface, tend to have high concentrations of Zn, Cu, Pb, Ni, Cd, and Ag. For more details on specific HM content of different rock species, see Alloway (2013).

The composition of the soil in a natural state is determined by its parent material (PM), which represents a rock or unconsolidated drift material that undergoes dismantling, weathering, or pedogenesis to form the mineral skeleton of the soil (Alloway, 2013). Weathering is the chemical decomposition of the parent material whereby the constituent elements in soluble form are slowly released to interact (physically and chemically) with the environment and the weathering byproducts of the materials around it. The weathering process determines the texture of the resulting soil, represented as percentages of sand, silt, and clay particles. It can have a major influence on the physical and chemical properties of the soil, such as the soil's ability to absorb cations and anions, which can directly influence the retention and spread of the bioavailability of HMs.

Another lithogenic source of HM soil contamination is volcanic activity. During eruptions, pyroclastic material and ash are released into the environment. Both have high concentrations of micro and macronutrients such as N, P, K, Ca Na, and Mg, which at small concentrations are beneficial to the environment and can result in fertile soils but can also contain high concentrations of toxic elements, including HMs such as Ni, Zn, Cd,

Ag, Sn, Hb, and Pb among others (Ermolin *et al.*, 2018). Although volcanic eruptions are short-lived, the ash and pyroclastic material produced can endure long periods in the environment, shaping the environment around it (Ruggieri *et al.*, 2010). While non-toxic elements tend to be soluble, the non-degradable and persistent nature of HMs means they can pose a risk to the environment and living organisms in the region for long periods after eruptions.

Mobilisation: Atmospheric Deposition and Runoff

Atmospheric deposition and runoff are non-localised sources of HM contamination that refer to HMs being transported from the source to soil in another location (Alloway, 2013), which can be natural or anthropogenic.

The main natural mobilisation path is atmospheric deposition, which is when the HMs in soil particles, dust, aerosol particles, and gaseous metals, such as Hg, are moved by natural air currents (wind). It is the most extensive form of contaminant transportation, as it can transport contaminating particles thousands of kilometres from the source. However, it is the least effective, as contaminants are diluted in the atmosphere and the deposition ratio is large. Moreover, it is only responsible for the contamination of the surface layer of the soil profile. It can result in a smooth spatial spread, with higher contaminant concentrations closer to the sources and decreased concentrations at increasing distances.

HM particles can also be transported by moving bodies or water. HMs can be found suspended in rivers, where the suspended sediment is deposited on alluvial soils in the event of a flood. Large HM particles enter the water system at contaminating concentrations through metalliferous mines, where particles are broken down into fine particles, known as "tailings", and were historically dumped into rivers (Alloway, 2013). HM can also be deposited on bodies of water via atmospheric deposition. Water in the form of glaciers can also transport trapped soil and dust by depositing it where the glacier melts.

Anthropogenic activities can also significantly contribute to HM transportation as they can be responsible for the movement of contaminated soil. Human-operated machinery, such as tractors, sprayers, and manure spreaders, can also move heavy materials, including HM-contaminated soil. Finally, the movement of soils can also occur during large landslides, which can transport substantial quantities of contaminated soil downslope, especially in the presence of heavy rain or other extreme hydrological events.

Industrial Activity

Industrial activity is the primary source of anthropogenic HM contamination in industrial and urban areas but can also affect rural regions through deposition and mobilisation pathways. The occurrence of HM contamination in soils at industrial sites varies and is specific to the region's industrial history. It can arise from dust, spillages, raw or

processed materials, wastes, final products, fuel ash, gaseous emissions from furnaces, or other high-temperature processes. This relationship between industrial activity, HM-rich waste, and byproducts results in a heterogeneous and non-stationary spatial distribution of multivariate contamination.

Steel and iron processing have long been associated with environmental pollution. The industry is a large producer of technogenic particles, which accumulate in shallow soils through different mobility pathways. HMs such as Cd, Pb, and Hg are commonly found as impurities in iron ore. Coke, a coal product, and other products are used in blast furnaces to remove impurities from the iron ore, resulting in Zn, Pb, Cd, and As in gaseous emissions and solid waste. Making steel requires oxygen furnaces and electric arc furnaces, which generate steel slag and dust, which are high in Cr, Ni, and Zn. Yang *et al.* (2020) show that steel and iron plant workers contain toxic levels of HM contaminants in their blood and urine, mostly absorbed through inhalation.

Non-ferrous metalliferous mining and smelting, meaning mining and smelting of non-magnetic metals such as Al, Cd, Cu, Zn, Pb, Ti, and Mg, are significant current and historical sources of HM contamination. They extract and process metal ores and gangue minerals. The process involves converting sulphide ore minerals to oxides by roasting them in air and reducing them in a furnace, which allows for the separation of different molten metals (Angon *et al.*, 2024). The historical inefficiency of the process is still responsible for HM contamination today. While smelting and mining industries are not often located in urban areas, they have been major contaminants of arable soil and potable water, posing a significant risk to rural populations in the surroundings (Sterckeman *et al.*, 2002).

Other manufacturing processes have also been linked to HM soil contamination. Pb is a common byproduct of the combustion of fossil fuels, manufacturing of paints and pigments, incineration of industrial waste, ceramic and dishware manufacturing, lead battery manufacturing and recycling, and plastics manufacturing. Ni is commonly found in industrial dust as a byproduct of electroplating but is also common in food processing industries. Cr can be found around textile manufacturing plants but has historically been helpful in manufacturing paints, dyes, pigments, photographic film, and tanning, which can be among the world's oldest industrial activities. Hg can also be found in all those mentioned above but is a particular concern in manufacturing fluorescent bulbs, chlor-alkali, scientific instruments, waste incineration, electrical switches, thermometers, cellulose, and rubber production. Cu can be found in some of the above but is particularly important in the production of explosives and textile fabrics such as rayon.

Agriculture

Many agricultural enhancement practices have been linked to HM soil contamination. Although these practices directly impact arable soils and soil quality, they can significantly

affect public health as they present a pathway for contamination to enter the food chain via produce and are a source of other environmental contamination via atmospheric deposition.

Fertilisers, both organic and inorganic, are one of the major sources of contamination in the agricultural industry today. For example, using livestock manure as organic fertiliser is a ubiquitous practice that can introduce contamination levels of HM to the soil depending on the source animal, their feed, and manure processing. For example, pigs and poultry in Europe have historically been fed more than their nutritional need of Cu and Zn to encourage growth, resulting in high Cu and Zn manure (Eckel *et al.*, 2005), and ultimately, high concentrations of Cu and Zn in the soil. In the USA, a growth promoter containing As was fed to chickens since the 1960s to speed maturation (Bellows, 2005).

Sewage sludge, known as biosolids, has also been widely used as agricultural fertiliser. It is typically obtained from wastewater at sewage treatment plants (STP) and can be high in N and P, which are essential for plant health at moderate concentrations. However, biosolids can also be high in HM, as they represent all domestic and urban discharges, and can vary significantly by STP (Alloway, 2013).

Inorganic fertilisers, on the other hand, are highly common in industrial agriculture worldwide. They focus on providing the primary macronutrients, such as N, P, and K. Secondary macronutrients are Ca, Mn, and S. Essential trace elements are also added to ensure plant health, including B, Cu, Co, Fe, Mn, Mo, Ni, and Zn. Phosphatic fertilisers contain the highest concentrations of potential HM contaminants, including As, Cd, U, Th, and Zn. While trace concentrations of these elements are considered essential micronutrients, the excessive historical and current use of these elements is a source of soil contamination.

Other enhancements of industrial agriculture can also be sources of contamination, such as insecticides, fungicides, and herbicides, especially when used in excess. Fungicides, for example, are often made from organo-metallic compounds such as Pb arsenate ($AsHO_4Pb$), Cu acetoarsenate, Cu oxychloride and phenyl mercury chloride. They are mainly responsible for Cu and As accumulation in the soil and contamination of nearby water sources through runoff (Angon *et al.*, 2024). The World Health Organisation has presented a classification of these substances based on toxicity levels, ranging from extremely hazardous to slightly hazardous. For more details, see Sharma *et al.* (2019).

Waste Disposal

While industrial and agricultural waste are significant sources of HM contamination, other forms of waste, such as landfills, electronic waste, recycling plants, waste incinerators, and sewage treatment facilities, can also be significant contributors.

Landfills are sites that contain municipal solid waste, which generally refers to waste objects made from plastic, glass, food, metals, and paper, among others, and are typically

used in commercial, office, domestic, and industrial settings. It is estimated that 2 billion tonnes of municipal solid waste are produced each year (Maalouf and Mavropoulos, 2023). At their simplest, landfills consist of holes in the ground where the waste is deposited and later compacted and covered. In modern landfills, contamination is contained by lining the hole with a clay lining and placing a plastic liner on top. In these cases, leachate, the liquid from solid waste produced by physical and chemical reactions inside the landfill, is contained. When these measures are not in place, leachate, which contains inorganic and organic pollutants and heavy metals, can leach into soil and nearby groundwater sources (Hosseini Beinabaj *et al.*, 2023). Since the capability of proper contamination mitigation measures is often a matter of economy, developing countries are more vulnerable to leachate pollution.

Electronic waste (e-waste) is rapidly emerging as a major public health risk and is generated by discarded electronic products. It is primarily composed of large household appliances (49%), but other e-waste includes small household appliances, information and communication electronics, entertainment equipment, electrical tools, toys and leisure equipment, and medical devices (Chakraborty *et al.*, 2022). The disposal of these objects differs depending on governing policy, as they could end in general landfill sites or have specific disposal sites. In the case where these objects are disposed of in specific sites, contaminating concentrations of Cd, Hg, As, Pb, and Cr in both surrounding soils, water, and sediment have been found (Hosseini Beinabaj *et al.*, 2023). For more details on specific contaminants and e-waste categories, see (Chakraborty *et al.*, 2022).

3.1.3 Impacts on Public Health

Human exposure to heavy metals occurs via three main pathways: ingestion, inhalation, and direct dermal contact (Adamo *et al.*, 2014). Ingestion exposure is the consumption of produce grown on contaminated soils and accounts for 90% of HM intake, while only 10% of the exposure is due to direct skin contact or inhalation of polluted dust (Mitra *et al.*, 2022). Even though trace concentrations of most HMs are essential for the human body, excessive concentrations are toxic to humans, pose significant risks to various systems in the body, and can even lead to death. In this section, we describe the specific toxic properties of HM on the human body, including neurotoxicity, nephrotoxicity, carcinogenicity, hepatotoxicity, immunological toxicity, cardiovascular toxicity, dermal toxicity, and reproductive and developmental toxicity.

Neurotoxicity

The main neurotoxic HM elements are Mn, As, and Cd. There is robust evidence of their adverse effects on neurological health by affecting neurotransmitter receptors, the synaptic

cytoskeleton, and scaffolding proteins, all of which result in lowered neurological function (Carmona *et al.*, 2021). Exposure to high concentrations of Mn has been shown to increase neurological complications due to apoptotic cell death (programmed cell death), presenting as Alzheimer's and Parkinson's disease (Goldhaber, 2003). On the other hand, exposure to As through ingestion has been shown to result in cognitive impairments of the central nervous system. It has been linked to neurological diseases, mainly of neurodevelopmental and neurodegenerative natures. Additionally, As poisoning also causes changes in synaptic transmission and the neurotransmitter balance and can even lead to death (Garza-Lombó *et al.*, 2019). Cd, which enters the body mainly through ingestion, affects cell proliferation, differentiation, apoptosis, and other cellular activities. As a result, it is strongly linked to neurodegenerative diseases such as amyotrophic lateral sclerosis (ALS), Parkinson's disease, Alzheimer's disease, and multiple sclerosis (MS) (Branca *et al.*, 2018).

Other HMs and metalloids are also known to have neurotoxic consequences. Large concentrations of Cu, Zn, and Fe can accumulate in the brain and impede neurodevelopment, while excess retention of Cu causes Wilson's disease, which has similar symptoms to schizophrenia (Mitra *et al.*, 2022).

Nephrotoxicity

The main nephrotoxic elements (those adversely affecting kidney health and function) are Cd, Pb, and Hg. Excessive accumulation of Cd results in symptoms like glucosuria (glucose in the urine), Fanconi-like syndrome (essential substances being excreted through urine), phosphaturia (phosphorus in urine), and aminoaciduria (abnormally high amino acids in urine) (Reyes *et al.*, 2013), which can eventually lead to renal tubular acidosis, renal failure, and hypercalciuria (Charkiewicz *et al.*, 2023).

Pb exposure is directly responsible for proximal tubular dysfunction, resulting in Fanconi syndrome. Chronic lead exposure is characterised by hyperplasia (excess cells in organ tissue), interstitial fibrosis (thickening of the kidney walls), atrophy of the tubules, renal failure, and glomerulonephritis (acute inflammation of the kidney) (Mitra *et al.*, 2022).

The toxicity of Hg affects multiple body systems. However, acute exposure causes acute tubular necrosis, presenting as acute dyspnea (shortness of breath), altered mental status, abdominal pain, profuse salivation, tremors, vomiting, chills, and hypotension. Chronic exposure to Hg causes injury to the epithelium and necrosis in the pars recta of the proximal tubule, presenting as a tubular failure, higher urine excretion of albumin and retinol protein, and a nephritic state, all leading to renal failure and potentially fatal (Lentini *et al.*, 2017).

Carcinogenicity

Although chronic exposure to many HMs is carcinogenic, those exposed to As, Pb, Hg, and Ni are at the highest risk. Chronic and acute exposure to As has been shown to cause epigenetic alterations, DNA damage, changes in protein expression and DNA methylation, among others (Martinez *et al.*, 2011). It significantly increases the risk of cancer by attaching to DNA-binding proteins and hindering the DNA-repair process (Mitra *et al.*, 2022).

Pb and Hg exposure causes cancer by damaging the DNA repair mechanism, cellular tumour regulation genes, and the chromosomal structure and sequence by releasing Reactive Oxygen Species (ROS). ROS, in turn, are highly carcinogenic and aid protumorigenic signalling by damaging cellular proteins, lipids, and DNA, resulting in cell damage (Pizzimenti *et al.*, 2010; Reczek and Chandel, 2017; Zefferino *et al.*, 2017). Pb exposure has been directly linked to a higher risk of lung, stomach, and bladder cancer (Rousseau *et al.*, 2007).

Ni has strong carcinogenic properties by affecting mechanisms such as gene regulation, transcription factor management, and free radical generation, which contribute significantly to carcinogenesis in human beings (Zambelli *et al.*, 2016).

Hepatotoxicity

Pb is also highly toxic to liver cells (hepatotoxic). Exposure, both acute and chronic, increases oxidative stress, causing liver damage by glycogen depletion and cellular infiltration, and can result in chronic cirrhosis of the liver (Hegazy and Fouad, 2014). Acute Cd exposure also causes oxidative stress and hepatocellular damage and can result in liver failure and increase the risk of liver cancer (Hyder *et al.*, 2013). The accumulation of Cu leads to Wilson's disease but can also lead to cholestatic liver diseases through oxidative stress mechanisms (Yu *et al.*, 2019). Finally, Cr affects the liver by elevating ROS levels, lipid peroxidation, suppression of DNA, RNA, and protein synthesis, DNA damage, decrease of antioxidant enzyme activity, mitochondrial dysfunction, cell growth arrest, and apoptosis (Hasanein and Emamjomeh, 2019).

Immunological Toxicity

Both acute and chronic exposure to Pb have a toxic effect on the immune system and result in the rise of allergies, infectious diseases, autoimmune diseases, and cancer (Rousseau *et al.*, 2007; Hsiao *et al.*, 2011). Cr also elicits a harmful immune response by reducing the phagocytic action of alveolar macrophages and hindering the immune response when exposure is through inhalation. Additionally, a link between Cr and contact dermatitis has also been well-established (Mitra *et al.*, 2022).

Cardiovascular Toxicity

Important cardiovascular toxic elements include Cd, Hg, and Pb. Population exposure to Cd is known to increase cardiovascular mortality (Tellez-Plaza *et al.*, 2013). Even low to moderate exposure results in hypertension, diabetes, carotid atherosclerosis, peripheral arterial disease, myocardial infarction, stroke and heart failure (Everett and Frithsen, 2008). Hg has been directly linked to atherosclerosis and increased risk of coronary heart disease, cardiovascular disease, acute myocardial infarction, coronary heart disease, and carotid artery stenosis (Kulka, 2016).

Chronic exposure to Pb can lead to arteriosclerosis and hypertension, thrombosis, atherosclerosis, and cardiac diseases through an increase of OS, reducing NO availability, altering vasoconstrictor and vasodilator prostaglandin balance, and raising blood pressure (Vaziri, 2008; Mitra *et al.*, 2022).

Dermal Toxicity

The main HMs with adverse effects on the skin are As, Cr, and Hg. Chronic exposure to As can cause skin diseases such as hyperkeratosis, hyperpigmentation, Bowen's disease and skin cancer (Huang *et al.*, 2019). Chronic exposure to Cr, on the other hand, can result in contact dermatitis, systemic contact dermatitis, and skin cancer. Finally, direct exposure to Hg and Hg-containing compounds can result in many skin infections, including acrodynia (Mitra *et al.*, 2022).

Reproductive and Developmental Toxicity

HM toxicity has been shown to affect reproductive and developmental health. In males, As is known to reduce the weight of the testes, negatively affecting sperm production, testosterone and gonadotropin levels, disturbing the steroidogenesis process and decreasing fertility as a result (Kim and Kim, 2015). In females, arsenic exposure increases the risk of endometrial problems and impairs endometrial angiogenesis, resulting in lowered fertility, prematurity, sterility, and spontaneous abortions (Milton *et al.*, 2017).

3.1.4 Environmental and Economic Impacts

Environmental

Soil is a complex structure consisting of five major components: mineral matter, water, air, organic matter, and living organisms (Chopra *et al.*, 2009), all susceptible to HM contamination. Changes in soil properties such as pH, porosity, conductivity, and natural chemistry can lower soil quality. Microbial and enzymatic activity, which is essential for the breakdown of organic matter and minerals for plant intake, are also significantly

hindered in contaminated soils, threatening ecosystem health, negatively affecting agricultural production, threatening food security, and increasing exposure through ingestion of food when grown on contaminated soils (Xin *et al.*, 2022) for both human beings and dependent organisms. Emissions from urbanisation and industrialisation can mobilise in runoff or direct waste disposal and enter surface and groundwater and form contaminated sediment or remain suspended in solution (Briffa *et al.*, 2020). As such, contaminated water also plays an important role in the exposure pathways of the ecosystem. Many sources and emitters of HMs into the environment can do so by releasing emissions as suspended particles in the air. These particles can remain in the atmosphere for a long time, be deposited in other soils and water, or be inhaled by living organisms. Section 3.3 gives more details on air pollution.

Economic

Globally, there are over 5 million polluted sites, covering 20 million ha of land, in which the soils are contaminated by various heavy metals or metalloids (Liu *et al.*, 2018). HM pollution in the soil has a combined worldwide economic impact estimated to be in excess of US$10 billion per year (He *et al.*, 2015) with poorer, less educated households being at higher risk of pollution injuries (Levasseur *et al.*, 2022). Pollution injuries include the economic and social burden of healthcare due to pollution and the cost of remediation. In the US, phytoremediation, a common remediation technique, has a cost of US$37.7m^3 (Wan *et al.*, 2016). For more details on the costs incurred by the healthcare burden of HM soil contamination, see Xu *et al.* (2023), and Khalid *et al.* (2017) for the cost of different remediation techniques.

3.1.5 Management of Contaminated Soils

It is estimated that there is 20 million ha of HM-contaminated soil globally, with As, Cd, Cr, Hg, Pb, Co, Cu, Ni, Zn, and Se as the most common contaminants (Liu *et al.*, 2018). Given the long-term persistence and the harmful effects of HM soil contamination, managing contaminated soils consists of limiting or decreasing exposure through imposing urban policy for landuse or, where exposure cannot be managed, establishing guidelines for the remediation of contaminated soils.

Remediation

Remediation techniques are numerous and occur in-situ or ex-situ, referring to treatment at the site or removal of contaminated material, respectively. General remediation techniques can be categorised as physical, chemical, electrical, thermal, and biological remediation. A Specific techniques may surface capping, soil flushing, electrokinetic extraction,

solidification, vitrification, and phytoremediation.

In containment-based techniques, such as surface capping, the contaminated soils are contained by covering them with a layer of waterproof material to form a protective surface. The cover forms an impermeable barrier to rain or surface water interaction and prevents contamination from diffusing into groundwater sources. The disadvantage is that the remediated soils lose their environmental functions, such as supporting plant growth, and are consequently only used for civil purposes (Liu *et al.*, 2018).

In electrokinetic extraction, HMs are removed from the contaminated soils via electrical absorption. Low-density direct current is applied to the soil via electrodes inserted in the ground. In the solution phase, cations migrate to the cathode while anions migrate to the anode. While this is an effective and cost-friendly approach to remediation, its success is highly dependent on specific soil conditions such as soil type, pH, water saturation, and organic content (Figueroa *et al.*, 2016). Soil flushing passes an extraction fluid through the soil to remove contaminants. It is costly and challenging, as extensive infrastructure is necessary to recover the flushing elutriate (Liu *et al.*, 2018).

Phytoremediation is one of the most common and accessible forms of soil remediation as it is operationally simple, aesthetically preferable, economically viable, and widely accepted by surrounding communities (Liu *et al.*, 2018). It consists of growing plants in contaminated soils and using their natural processes to remove heavy metals (phytoextraction and phytovolatilisation) or to stabilise them into harmless substances known as phytostabilsation (Mahmood *et al.*, 2015). Over 721 species of plants are considered hyperaccumulators due to their ability to accumulate heavy metals without suffering phytotoxic damage (Reeves *et al.*, 2018). However, phytoremediation is ineffective, as it is slow to reduce contamination, and the plants are subject to nutrient depletion or pests.

Ex-situ remediation techniques move contaminated soils from one location to another for treatment or safe storage and include techniques like landfilling, soil washing, solidification, and vitrification. These techniques have high costs, infrastructure, and management expenses and are therefore reserved for particularly suitable cases.

The choice of remediation technique must be made on a case-by-case basis, as every contaminated site is highly heterogeneous and depends on the local sources of contamination and landuse. Overall, remediation techniques are chosen by the geography of the contaminated site, contamination properties, time required, remediation goals, cost-effectiveness, financial budget, implementation readiness, and public acceptability (Khalid *et al.*, 2017).

Policy

The United Kingdom has a history of regulating heavy metal waste and emission discharge that could eventually enter the soil ecosystem, that dates back to the Industrial Revolu-

tion, such as the Rivers Pollution Prevention Act of 1876 or the Clean Air Act of 1956. However, specific policy targeting heavy metal soil contamination was only passed in the Environmental Protection Act of 1990: Part 2A (DEFRA, 1995). It defines contaminated land as "any land which, by condition or reason of substances, can cause significant harm, or there is a possibility of significant harm can be caused, or one which poses the threat of polluting waters". It grants specific provisions for local authorities to assess risk and places the liability on the original polluter, following the "polluter pays" principle. If the responsible party cannot be found, the landlord is responsible for risk assessment and appropriate remediation.

While policy documents do not directly state guidance values, the Contaminated Land Exposure Assessment (CLEA, Environment Agency 2009) documents by the Environment Agency have provided direct soil guideline values (SGV) according to the function of the contaminated land. Table 3.1.5) shows SGV for allotment soils, residential soils with home-grown produce, residential soil without home-grown produce, and industrial grounds. The table shows As, Cd, Cr, Hg, Ni, and Pb. Cu and Zn are omitted because their recommended values depend on the soil's pH levels.

3.2 HM Contamination: Exploratory Analysis

3.2.1 Data Description

The British Geological Survey (BGS) performed the geochemical baseline survey of the environment (G-BASE) starting in the 1960s and ending in 2014 (Johnson *et al.*, 2005). Although the survey spanned multiple decades, it only collected a single sample for every location and therefore cannot inform as to changes in time. Initially commissioned for mineral exploration, it is now a valuable tool for the systematic assessment of the geochemical baseline of the UK environment. While the original plans of the survey were to collect data in England, the Glasgow Conurbation was added later (collected in 2014) as an important representation of historical urban soils. The data produced by this survey provides a single multivariate observation - representing the geochemical profile of the soil - for the locations sampled for the time when the sample was collected.

In this application, only the Glasgow Conurbation region in the Clyde River Basin, west of Scotland, is considered. It consists of approximately 2745 topsoil (5-20cm deep) samples taken at approximately 4 observations per km^2 in the urban areas and 1 per 1 km^2 in rural areas. Each sample was decomposed chemically using X-ray fluorescence spectrometry (XRFS), and the concentration of each element in the sample was in parts per million (ppm). Only the concentration of uranium (U) was measured using the delayed neutron method (DNM). Soil properties such as conductivity and pH were measured using portable field equipment. As is common practice in soil surveys, a single sample was

Table 3.1: UK CLEA soil guidance values by land use function for the major HM contaminants As, Cd, Cr, Hg, Ni, and Pb. Cu and Zn content are excluded from this table because their concentrations are highly dependent on soil pH.

Contaminant	**Function of Land Use**	**SGV (mg/kg)**
Arsenic (As)	Allotment	49
	Residential with home grown produce	37
	Residential without home grown produce	70
	Industrial	640
Cadmium (Cd)	Allotment	3.9
	Residential with home grown produce	22
	Residential without home grown produce	150
	Industrial	410
Chromium (Cr)	Allotment	
	Residential with home grown produce	130
	Residential without home grown produce	200
	Industrial	5000
Mercury (Hg)	Allotment	26
	Residential with home grown produce	10
	Residential without home grown produce	10
	Industrial	26
Nickel (Ni)	Allotment	230
	Residential with home grown produce	230
	Residential without home grown produce	230
	Industrial	1800
Lead (Pb)	Allotment	80
	Residential with home grown produce	200
	Residential without home grown produce	310
	Industrial	2300

collected at each location, rendering the data "unreplicated", meaning a single temporal replication is available per location. Further details can be found in Johnson *et al.* (2005). Some of the most common HM contaminants in the soil are As, Cr, Cu, Ni, Pb, and Zn; therefore, these elements will remain the focus of the work presented in this dissertation.

Skewness and heavy-tails are common features of HM contaminant distributions (Marchant *et al.*, 2010). Figure 3.1 shows histograms of each individual contaminant in the original scale, showing that the contaminants display strong right skewness and heavy tails. Figure 3.2 shows histograms for the individual contaminants after the log-transformation, already exhibiting less-skewed behaviour, improved symmetry, and lighter tails.

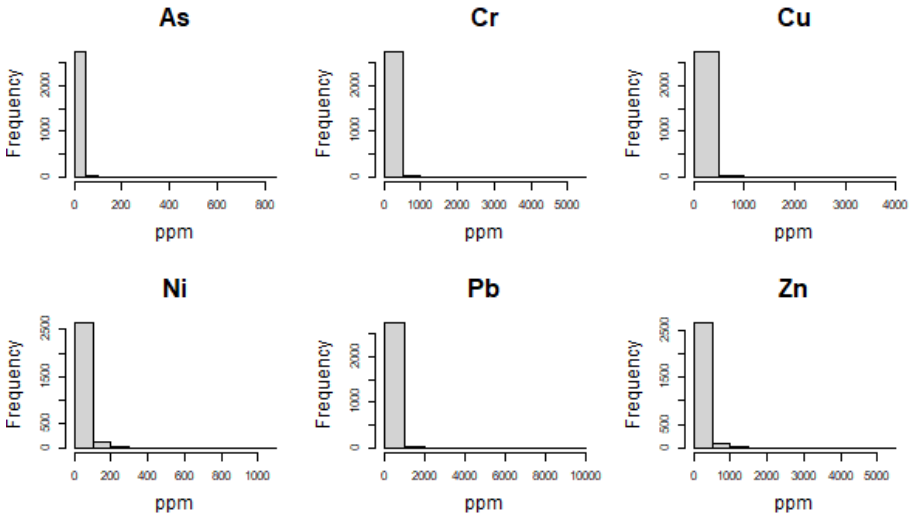

Figure 3.1: Histograms of As, Cr, Cu, Ni, Pb, and Zn in their original scale (ppm) for samples taken in the Glasgow Conurbation.

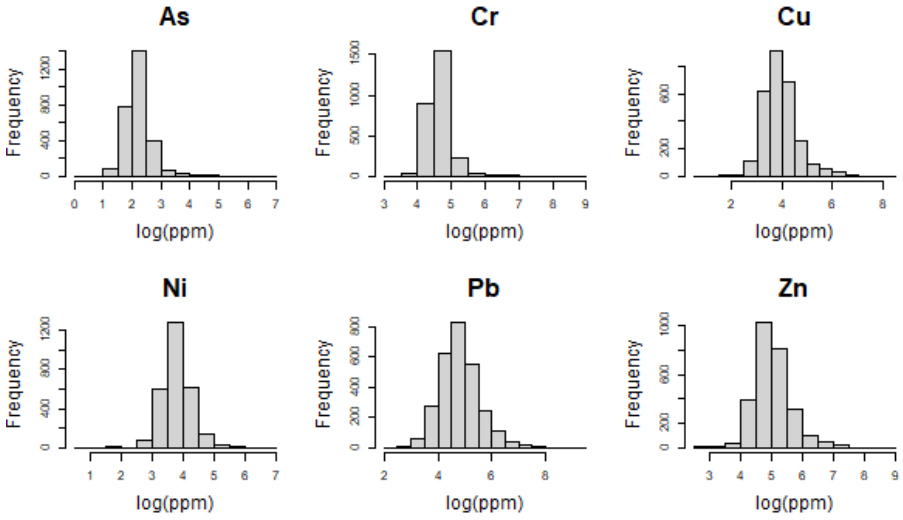

Figure 3.2: Histograms of As, Cr, Cu, Ni, Pb, and Zn after a log-transformation for samples taken in the Glasgow Conurbation.

Although the transformed data appear approximately normal after a log transformation, kurtosis estimates are 13.51 for As, 18.32 for Cr, 5.64 for Cu, 6.55 for Ni, 4.69 for Pb, and 5.57 for Zn. The kurtosis provides a measurement for the heaviness of the tail of a distribution, with Gaussian tails having a kurtosis of 3. In this application, all contaminants exceed a kurtosis of 3, with Cr and As having the heavier tails, and Pb and Cu having the lighter ones. This is indicative of non-Gaussian behaviour in the tail and a warning of the potential underestimation of the extremes by Gaussian models, suggesting non-Gaussian modelling alternatives might be prudent.

3.2.2 Spatial Distribution of Individual Contaminants

Maps of individual contaminants are given in Figures 3.3 to 3.8. The spatial heterogeneity and overall behaviour are visible from the different spatial patterns for each contaminant. For example, the maps of arsenic (As) concentrations in Figure 3.3 show that there is no clear region of high concentrations. However, observations surpassing the 95th percentile are mostly located in the west half of the city. They tend to follow roads with high traffic, like the road to the northwest of the city near the Trossachs National Park.

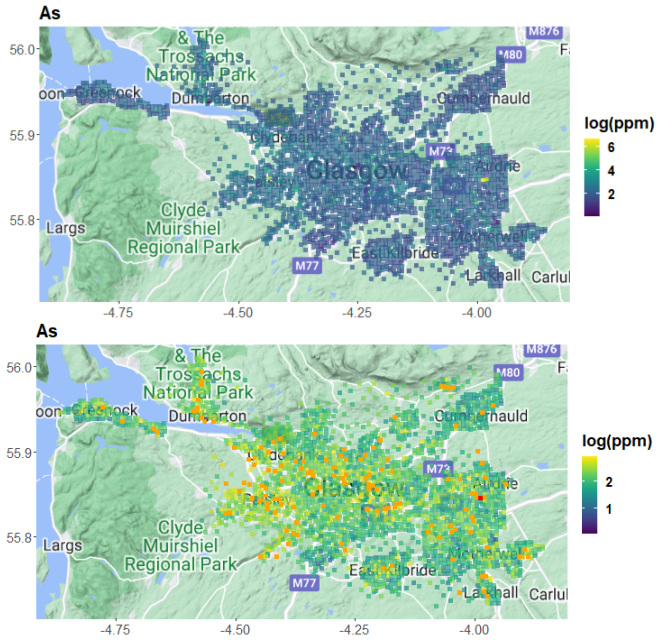

Figure 3.3: Maps of the log concentration of As. The map on the top shows the entire range of values, while the map on the bottom is censored at 2.93 log(ppm), the 95th percentile. Concentrations above 2.93 log(ppm) are shown in orange, and the maximum value, 6.75 log(ppm), is shown in red.

The map of Cr concentrations in Figure 3.4 shows a clear region of high concentrations south of the Glasgow city centre, on the southern banks of the River Clyde - an area of significant industrial history. The maximum concentration, however, is found to the

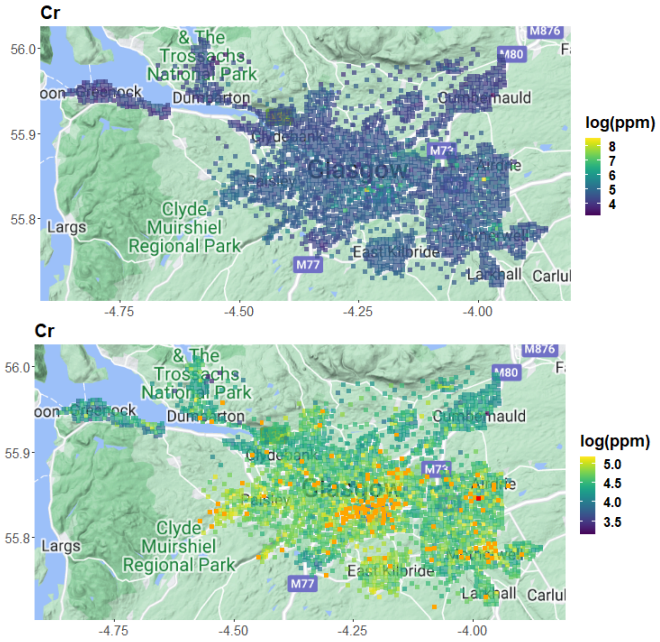

Figure 3.4: Maps of the log concentration of Cr. The map on the top shows the entire range of values, while the map on the bottom is censored at 5.2 log(ppm), the 95th percentile. Concentrations above 5.2 log(ppm) are shown in orange, and the maximum value, 8.58 log(ppm), is shown in red.

west of the city in a suburb called Coatbridge, which is also the location of the maximum concentration of As. Maps for the concentrations of Cu and Ni are given in Figures 3.5 and 3.6, respectively. While neither contaminant has a clear spatial pattern, concentrations of Cu display more spatial smoothness, with areas of larger concentrations to the west of the city, near Paisley and the Clyde Muirshiel Regional Park, and to the southeast, including the villages of East Kilbride (south) and Wishaw (southeast). Both contaminants share the same location for maximum value, which is west of the city between the Clyde River (a historically industrial area) and the Glasgow International Airport. The maps of concentrations of Pb and Zn are given in Figures 3.7 and 3.8, respectively. While there is no single area of large concentrations of either contaminant, higher concentrations of Pb are found in the western half of the city, along major roads, and alongside the river banks. Zn displays a similar pattern, with higher concentrations south of the river and along the banks. The maximum concentration of Zn is found in Coatbridge to the east of the city, while Pb has a maximum concentration east of the city on the banks of the River Clyde.

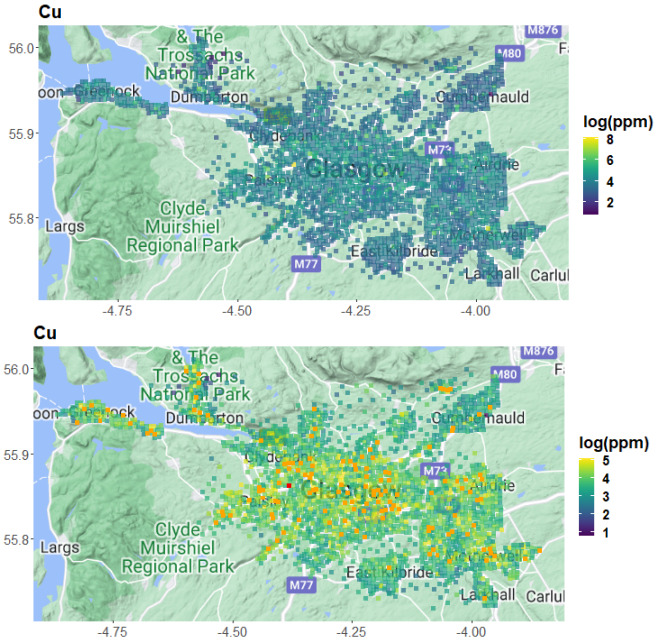

Figure 3.5: Maps of the log concentration of Cu. The map on the top shows the entire range of values, while the map on the bottom is censored at 5.11 log(ppm), the 95th percentile. Concentrations above 5.11 log(ppm) are shown in orange, and the maximum value, 8.21 log(ppm), is shown in red.

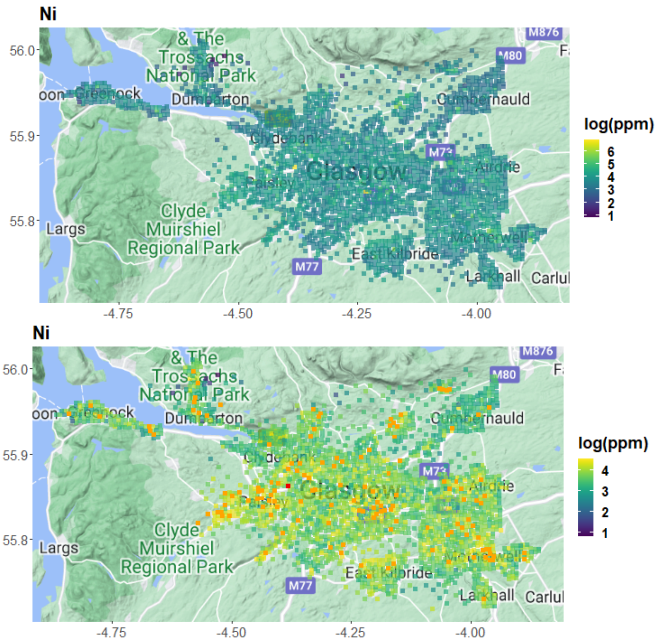

Figure 3.6: Maps of the log concentration of Ni. The map on the top shows the entire range of values, while the map on the bottom is censored at 4.59 log(ppm), the 95th percentile. Concentrations above 4.59 log(ppm) are shown in orange, and the maximum value, 6.95 log(ppm), is shown in red.

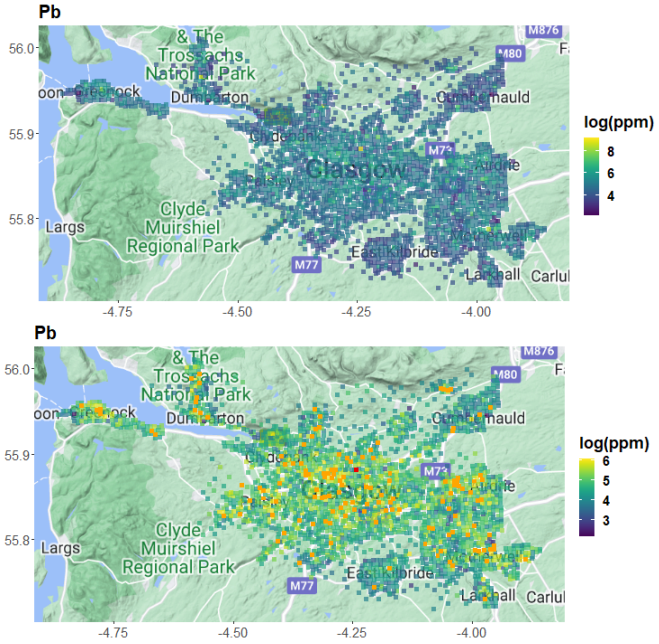

Figure 3.7: Maps of the log concentration of Pb. The map on the top shows the entire range of values, while the map on the bottom is censored at 6.1 log(ppm), the 95th percentile. Concentrations above 6.1 log(ppm) are shown in orange, and the maximum value, 9.2 log(ppm), is shown in red.

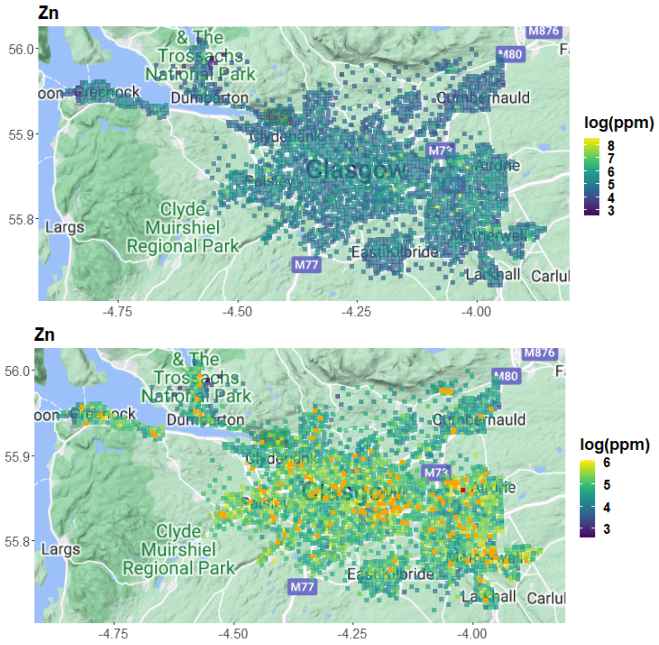

Figure 3.8: Maps of the log concentration of Zn. The map on the top shows the entire range of values, while the map on the bottom is censored at 6.1 log(ppm), the 95th percentile. Concentrations above 6.1 log(ppm) are shown in orange, and the maximum value, 8.53 log(ppm), is shown in red.

3.3 Air Pollution: Particle Matter 2.5

3.3.1 Definition of $PM_{2.5}$ Air Pollution

Particle matter (PM) of less than 2.5 µm in diameter is called $PM_{2.5}$. It is one of the six major air pollutants linked to significant health risks (Cheng *et al.*, 2024). Snider *et al.* (2016) used the Surface PARTiculate mAtter Network (SPARTAN), a long-term global network working on the characterisation of chemical and physical attributes of aerosols, and found that the major constituents (relative contribution ± SD) of $PM_{2.5}$ are ammoniated sulfate (20%± 11%), crustal material (13.4%±9.9%), equivalent black carbon (11.9%±8.4%), ammonium nitrate (4.7%±3.0%), sea salt (2.3%±1.6%), trace element oxides (1.0%±1.1%), water (7.2%±3.3%), and residual matter (40%±24%) from a variety of emission sources and atmospheric processes. However, these fractions are highly localised, reflecting emission sources in the area. Identifying the precise composition of the $PM_{2.5}$ pollution at a local scale is a current area of research, as it can determine the specific public health risks posed by the exposed population (Cheng *et al.*, 2024).

3.3.2 Sources of Air Pollution

The sources and composition of $PM_{2.5}$ pollution display spatial and temporal heterogeneity and are therefore considered at a local or regional scale. They are typically identified using source inventories in an area to match unique chemical profiles to known sources of pollution. Generally, sources of $PM_{2.5}$ pollution can be classified into six different categories (Ryou *et al.*, 2018): motor vehicle, secondary aerosol, soil dust, industrial combustions, and natural sources.

Motor Vehicles

Motor vehicle emissions of gasoline or diesel are major contributors to the denominated "vehicle emissions" or "road dust". The literature makes no distinction between the emissions from gasoline or diesel combustion (Ryou *et al.*, 2018). Vehicle-related sources of these emissions include a mixture of tailpipe emissions from gasoline and diesel engines and road dust components of lubricating oil combustion, brake, and tyre abrasion products (Viana *et al.*, 2008). This source is primarily responsible for the carbon content in $PM_{2.5}$, including total, organic, and elemental carbon (Yi and Hwang, 2014). They are also responsible for inorganic ion content, including sulfate (SO_4^{2-}), nitrate (NO_3^-), and ammonium (NH_4^+) (Moon *et al.*, 2008). Non-tailpipe emissions were characterised by the existence of Cu, Zn, and Ba (Ryou *et al.*, 2018). Motor vehicle emissions are mainly diurnal and higher on weekdays than on weekends.

Secondary Aerosol

Primary PM refers to the direct emission of particles, whereas secondary PM refers to the particles formed once in atmospheric suspension between polluting emissions and atmospheric content. Secondary aerosol includes "secondary nitrate" and "secondary sulfate" sources. They are considered to be aerosol-related summer sources of SO_4^{2-}, NO_3^-, and NH_4^+, which are produced by gas-to-particle conversion processes, sulfur dioxide oxidation, and ammonium neutralisation. Secondary aerosol emissions show strong seasonal variation with high levels of SO_4^{2-} in the summer and NO_3^- in the winter, enhanced by increased photochemical reactions related to temperature (Ryou *et al.*, 2018).

Soil Dust

Soil dust encompasses sources such as "soil", "soil-related", "geological", "urban dust", and "soil dust" and is exclusive of "road dust" or any other traffic-related dust emission. Soil dust can reflect sand, city dust, local or regional re-suspension, and wind-blown dust in the air (Viana *et al.*, 2008). Chemically, it is characterised by the presence of Al, Ca, Fe, and K (Ryou *et al.*, 2018). Soil dust emissions can be higher during planting and harvest seasons in the spring and autumn, as seasons of higher agricultural activity increase suspended dust content.

The category can also include biomass and field burning. Biomass burning refers to the chimney emissions from burning wood for domestic purposes. Field burning is burning crop fields to clear them of bugs or unwanted plants before sowing. Similar components characterise emissions from these sources as soil dust but with a large proportion of black carbon (BC; Yi and Hwang 2014).

Industrial Combustion

This source encompasses emissions known as "oil combustion", "fossil combustion", "municipal and waste incineration", and "cement and construction". It is characterised by high contents of Ca, Cl, Zn, Antimony (Sb), Fe, and K. Oil combustion can be separated from the others for its Ni and V content, which is typically related to ships and industrial plants (Ryou *et al.*, 2018).

Natural Sources

Natural sources of $PM_{2.5}$ contamination are aged sea salt, marine aerosol, and volcanic emissions. They are characterised by Na and Cl content (Ryou *et al.*, 2018). Although forest fires release large amounts of $PM_{2.5}$ into the atmosphere, it is high in BC content and typically categorised with soil dust.

3.3.3 Impacts of Air Pollution

$PM_{2.5}$ is a pernicious presence in the urban atmosphere and poses a major threat to public health (Martenies *et al.*, 2015), with studies such as the Global Burden of Disease Study (GBD;GBD 2016) in 2015, ranking $PM_{2.5}$ as the fifth leading risk factor for death. Exposure to $PM_{2.5}$ can endanger multiple organ systems and lead to systemic adverse effects. Robust associations have been made between long-term exposure to ambient and indoor $PM_{2.5}$ exposure and increased mortality due to heart disease, stroke, chronic respiratory disease, and lung cancer, among others (Sharma and Mujumdar, 2022).

Respiratory Effects

The small size of $PM_{2.5}$ pollution enables its toxic components to penetrate deep into the lungs and deposit in the terminal pulmonary bronchioli and alveoli with each breath, resulting in increased oxidative stress, inflammation of tissue and cells (Davel *et al.*, 2012; Habre *et al.*, 2014), and altering the immune response (Feng *et al.*, 2016). Together, these adverse effects are responsible for the decline in lung function, incidence, exacerbation, and maintenance of asthma, chronic obstructive pulmonary disease (COPD; Don D. Sin *et al.* 2023), and for increasing the lung's vulnerability to infection (Duan *et al.*, 2013; Jedrychowski *et al.*, 2013).

The deposition of $PM_{2.5}$ particles in the pulmonary bronchioli and alveoli results in the internalisation of toxins into lung cells (Gualtieri *et al.*, 2011), eliciting oxidative stress and triggering impairments to normal cellular function and can even cause cellular death by ways of apoptosis, autophagy or others (Gualtieri *et al.*, 2011). The oxidative stress by $PM_{2.5}$ exposure elicits an inflammatory response, which has also been shown to worsen previous pulmonary injuries and may lead to alveolar collapse (Duan *et al.*, 2013). The immune system is triggered simultaneously, causing bronchial remodelling that may result in the thickening of the bronchial walls and tissue fibrosis, further decreasing lung function (Zaiss *et al.*, 2015). $PM_{2.5}$ exposure can also alter the immune response in the lung and render it susceptible to infections by decreasing bacterial clearance (Duan *et al.*, 2013), triggering the death of lung epithelial cells, and impeding antimicrobial activities in the lower airways (Feng *et al.*, 2016).

Cardiovascular Effects

$PM_{2.5}$ particles enter the cardiovascular and circulatory system through the gas-blood barrier in the alveoli (Schulze *et al.*, 2017). The entrance of $PM_{2.5}$ toxins into the circulatory system triggers cardiovascular events through similar mechanisms to those in the respiratory system: oxidative stress and inflammation. These two are responsible for an increase in cardiovascular diseases, including atherosclerosis, coagulation, hypertension,

myocardial remodelling, and thrombotic and non-thrombotic acute cardiovascular events such as heart failure, endothelial dysfunction, and arrhythmias Basith *et al.* (2022). The impacts have been measured as significant, with $PM_{2.5}$ exposure being considered a risk factor for cardiovascular morbidity and mortality (Basith *et al.*, 2022).

Cerebrovascular Effects

Strokes are the second leading cause of premature mortality in the world (GBD, 2016). $PM_{2.5}$ air pollution is the third significant contributor to the global stroke burden after preventable risk factors, accounting for 29.2% of the burden of stroke (Feigin *et al.*, 2016). Strokes can be categorised as ischemic or hemorrhagic. Ischemic strokes are characterised by a blockage of the circulation system, leading to decreased blood flow and tissue necrosis. Hemorrhagic strokes are indicative of bleeding in the brain through a damaged blood vessel. While the effects of chronic $PM_{2.5}$ exposure are the same in the cerebrovascular system, inflammation and oxidative stress, Lamorie-Foote *et al.* (2023) show that $PM_{2.5}$ is strongly related to increased risk of ischemic stroke, but their effect on hemorrhagic strokes is variable and may be influenced by other factors. They also show that $PM_{2.5}$ exposure increases the probability of ischemic and hemorrhagic strokes in older patients with pre-existing diabetes.

Other Public Health Impacts

Recent studies suggest a link between prenatal $PM_{2.5}$ exposure and hindered neurodevelopment. Xu *et al.* (2022) showed that $PM_{2.5}$ exposure increases the risk of non-optimal gross motor development by 31% for every 10 µg increase in the average $PM_{2.5}$ with SO_4^{2-} concentrations considered to be the most significant toxin. Lertxundi *et al.* (2019) showed that the NO_2 is linked to lower global cognition and language development in children up to 1 year old. Hurtado-Díaz *et al.* (2021) also showed that $PM_{2.5}$ exposure during pregnancy hindered language development in children up to 24 months old. Additionally, He *et al.* (2017) show the association of levels of atmospheric $PM_{2.5}$ and type 2 diabetes and gestational diabetes.

Environmental Impacts

The environmental impacts of $PM_{2.5}$ are severe and can result in a significant deterioration of environmental quality. Soil dust, the black carbon in soot and a type of $PM_{2.5}$, is the dominant absorber of visible solar radiation in the atmosphere (Ramanathan and Carmichael, 2008). Its high absorption properties and strong regional spatial distribution make black carbon $PM_{2.5}$ the second strongest contributor to global warming after carbon dioxide emissions (Ramanathan and Carmichael, 2008).

A significant effect of $PM_{2.5}$ and PM_{10} in the environment is a reduction in visibility. PM_{10} and $PM_{2.5}$ are hygroscopic, meaning they can absorb water. When atmospheric humidity is high, $PM_{2.5}$ and PM_{10} absorb water vapour and enlarge, participating in fog formation and lowering visibility for long periods (Khanna *et al.*, 2018). The suspended $PM_{2.5}$ in high humidity conditions also results in acidic rain. In turn, acidic rain lowers soil quality and decreases the rate of leaf and compost breakdown, slowing the natural reintegration of essential micronutrients to the soil (Wu and Zhang, 2018).

Finally, the deposition of $PM_{2.5}$ on wet areas and arable soils increases the acidification of water and soil and is a major contributor to contamination. Moreover, nitrogen (N) deposition on water bodies encourages plant growth inside water bodies, playing a significant role in eutrophication and increasing biodiversity loss (Erisman *et al.*, 2013).

3.3.4 Monitoring and Policy

The UK has a long history of air pollution, starting with the Industrial Revolution when the country became increasingly reliant on the burning of fossil fuels to meet energy needs. Consequently, large urban smogs developed, with acute and chronic exposure causing thousands of premature deaths, such as in the Great London Smog incident of 1952 (Laskin, 2006). As a result, the Clean Air Act was passed in 1956, giving power to local authorities to control emissions of smoke, grit, dust, and fumes by banning the sources of these emissions. To meet the stipulations of the Clean Air Act, the National Survey was established in 1961 as a nationwide air pollution monitoring network. In 1992, the Department for Environment, Food and Rural Affairs (DEFRA) established an Enhanced Urban Network (EUN) for air pollution monitoring, and in 1995, consolidated all urban monitoring under one comprehensive program, including the London Air Quality Monitoring Network sites. In 1998, urban and rural automatic networks, which had previously been separate, were combined and brought under the Automatic Urban and Rural Network (AURN). Today, AURN comprises over 170 sites across the UK and is responsible for $PM_{2.5}$ and PM_{10} monitoring.

The Air Quality Standards Regulations (AQSR, 2010) passed in 2010 states that PM concentrations must not exceed

- PM_{10}: An annual average of $40\,\mu m^3$.
- PM_{10}: Any 24-hour average of $50\,\mu m^3$ more than 35 times in a single year.
- $PM_{2.5}$: An annual average of $20\,\mu m^3$.

The targets for both $PM_{2.5}$ and PM_{10} were updated in 2023 for England in the Environmental Targets Regulations (ETR, 2023). The 2040 targets for England are:

- $PM_{2.5}$: An annual average of $10\,\mu m^3$ is not exceeded at any monitoring station (known as the Annual Mean Concentration Target).
- $PM_{2.5}$: Population exposure is at least 35% less than that in 2018 (known as the Population Exposure Reduction Target).

The Environmental Improvement Plan of 2023 for England sets the following interim targets for the end of January 2028:

- $PM_{2.5}$: An annual average of $12\,\mu m^3$ is not exceeded at any monitoring station.
- $PM_{2.5}$: Population exposure is at least 22% less than that in 2018.

3.4 Air Quality: EAC4 and AURN Datasets

In this Section, we explore the data used for the case study in Chapter 6, which focuses on $PM_{2.5}$ pollution in the Greater London area shown in Figure 3.9, for daily averages in the year 2022. The data come from two sources, one representing remote-sensing data, and the other in-situ measurements. The CAMS global renanalysis dataset, known as EAC4 (http://www.atmos-chem-phys.net/19/3515/2019/), is obtained from the Centre for Atmosphere Monitoring Service (CAMS), a subsidiary of the European Centre for Medium-Range Weather Forecasts (ECMWF). The EAC4 data is obtained by merging satellite observations of atmospheric composition from the Copernicus satellite with computer simulations of the atmosphere in a process known as data assimilation (Inness *et al.* 2019; https://ads.atmosphere.copernicus.eu).

The second data source is the AURN monitoring network, administered by the Environment Agency (https://uk-air.defra.gov.uk/networks/network-info?view=aurn). Both datasets are freely accessible online and can be downloaded at various time scales and periods.

3.4.1 CAMS Global Reanalysis (EAC4)

The CAMS global reanalysis data, formally known as ECMWF Atmospheric Composition Reanalysis 4 (EAC4), is a global reanalysis dataset of atmospheric composition (see Inness *et al.* 2019 for details). Analysis combines dynamic models based on the physical and chemical processes in the atmosphere with satellite observations to formulate the initial conditions of 12-hour forecasts. Reanalysis, on the other hand, performs retrospective analysis of past periods and can use data from observation stations to enhance forecasts. Observations from analysis and reanalysis models are different from in-situ stations. Significant discrepancies between the two sources can be due to biases introduced through the construction of the model (Sheridan *et al.*, 2020), the smoothing nature of a coarse

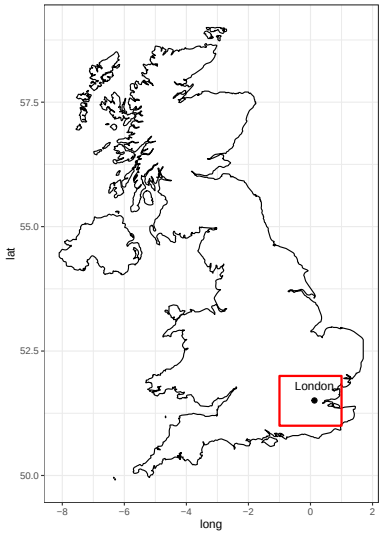

Figure 3.9: Greater London region in the case study of data fusion for $PM_{2.5}$ extremes model proposed in Chapter 6.

spatial resolution such as those commonly used in global reanalysis datasets, and the spatial smoothness of the phenomenon. Discontinuous variables, such as precipitation or windspeed, have been shown to display large discrepancies between reanalysis data and measurements from observation stations (Essou *et al.*, 2016). Additionally, Sheridan *et al.* (2020) showed that the differences between data from both sources increase for extreme temperature events, especially away from the central latitudes and close to the coast, showing that significant bias is found in reanalysis datasets even after assimilation.

Because air pollution guidelines (see Section 3.3.4) are on a 24-hour average scale, we aggregated data from the sub-daily scale to the daily average. To maximise the amount of data available in the AURN network, we chose to focus on the Greater London area for the year 2022, which represents the area inside the coordinates $(-1°, 51°)$ and $(1°, 52°)$ (for degrees East and North), shown in Figure 3.9. This area has a diverse geographical setting, containing urban, suburban, and rural regions. At a spatial scale of $0.1° \times 0.1°$, data at each cell centroid is defined as X_i, where $i = \{1, ..., 220\}$ represent the 220 available cells shown in Figure 3.10.

Figure 3.10 also shows sites 1 to 5, marked for demonstration, representing suburban, urban, and rural areas. Site 1 is Crawley to the south of London, which is a suburban location. Sites 2 and 3 are urban areas inside metropolitan London, Croydon and London City Centre, respectively. Finally, sites 4 and 5 are rural areas by the coast and inland, respectively.

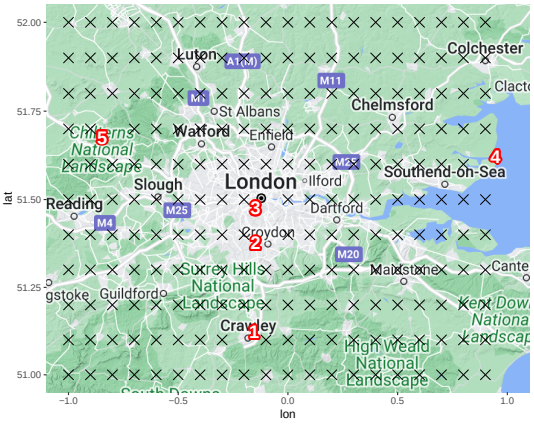

Figure 3.10: Map of the Greater London region and the 11×20 grid of the EAC4 dataset. Sites 1 to 5 are marked for further demonstration purposes representing suburban, urban, and rural settings.

Temporal Patterns

Cell centroids across the region behave similarly, following consistent temporal patterns as seen in the box plots in Figures 3.11 and 3.12. The months of January and March (Figure 3.11) exhibit the highest $PM_{2.5}$ concentrations across all cells, with the overall maximum observed in March. These months also experience the highest daily variability, potentially due to a combination of the holiday season (between December and January) and the meteorological features of the winter season. Higher concentrations are also seen in April, May (Figure 3.11), and November (Figure 3.12). The remaining months, February and June to October, show limited variability, with mostly constant concentrations and small peaks. During most of the year, site 3, the London city centre, experiences higher values of $PM_{2.5}$ while site 4 on the rural coast (Southend-on-Sea) generally experiences the smallest concentrations with brief relative peaks such as the first peak in March and the middle of May. As seen in the Figures, no clear pattern is discernible between yearly or weekly patterns, such as summer-winter differences. Moreover, the possible existence of weekday-weekend patterns or differences was investigated, but no patterns were found, suggesting the presence of more influential factors in $PM_{2.5}$contamination in the London region or providing evidence to the "smoothed" surface provided by remote-sensing datasets.

Spatial Patterns

The spatial patterns of 5 descriptive statistics are shown in Figure 3.13, specifically, the pointwise minimum, maximum, median, and the width of the range of observations at each location. The figure shows a consistent, and perhaps expected, pattern: the London city centre has higher concentrations of $PM_{2.5}$ than the rest of the region. The top left corner figure shows that the London city centre and the south-east region have the highest

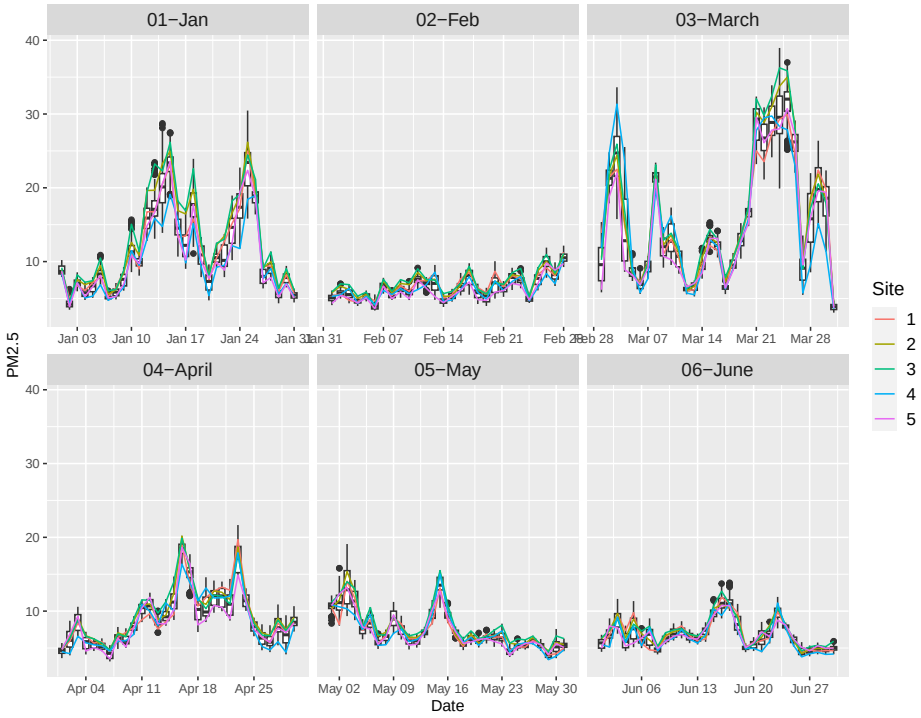

Figure 3.11: Box-plots for daily mean concentrations of $PM_{2.5}$ for the months January to June. The daily concentration of the Sites 1 to 5 are given in coloured lines.

Figure 3.12: Box-plots for daily mean concentrations of $PM_{2.5}$ for the months July to December. The daily concentration of the Sites 1 to 5 are given in coloured lines.

minimum concentrations, while the north, west, and southwest areas have lower concentrations. It is particularly noticeable that cells over the coast have lower concentrations than those on land. While the same behaviour is present in the median (top right corner), the map of maxima in the bottom row of the figure shows that the city centre has a maximum value that is $8\,\mu g/m^3$ higher than the surrounding areas. Additionally, it is evident that the region north of the city has higher maxima than the south, which could be explained by commuting patterns in the city. As expected from modelled data, the EAC4 observations have smooth patterns across space and are expected to have a strong spatial dependence. Finally, the width of the range has an almost identical pattern to the maxima, showing a higher range in the London city centre and the region to the east.

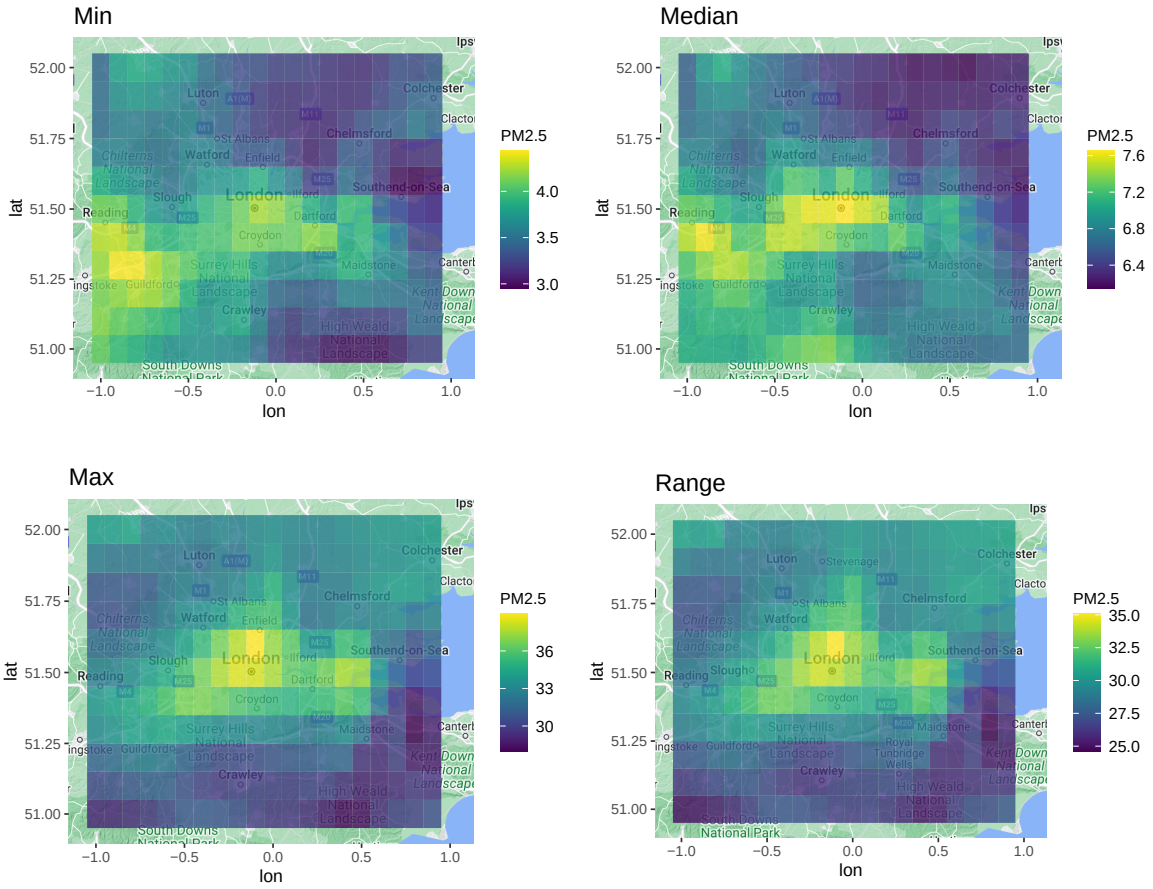

Figure 3.13: Spatial distribution of the EAC4 data in the Greater London area for 2022. The plots show the minimum, median, maximum, and width of the range of values at each location.

3.4.2 Automatic Urban and Rural Network (AURN)

The AURN is the largest automatic monitoring network in the UK, and it is used to measure compliance with the Ambient Quality Directives in the UK (https://uk-air.

defra.gov.uk/networks/network-info?view=aurn). There are 281 historical sites, with 174 currently active sites. Each site is unique, but most measure the major air pollutants, namely, oxides of nitrogen (NO_x), sulphur dioxide (SO_2), ozone (O_3), carbon monoxide (CO), and particle matter ($PM_{2.5}$ and PM_{10}) at hourly intervals. In the Greater London region, there are 26 active monitoring sites; however, only 12 sites contain records for more than 75% of the days in 2022, meaning the remaining 14 stations have more than 25% missing daily observations. For this reason, only the 12 most complete sites shown in Figure 3.14 were considered in Chapter 6. Like the EAC4 data, the AURN data was also aggregated to daily mean.

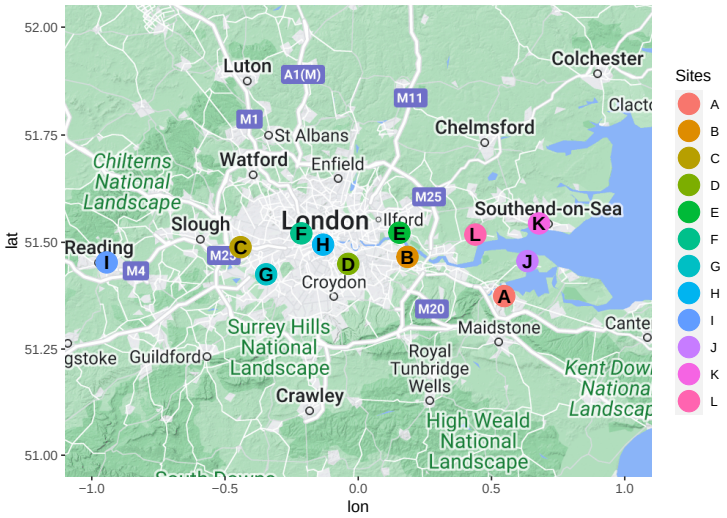

Figure 3.14: Circles denoting the 12 stations with at least 75% complete coverage in the year 2022.

Temporal Patterns

Figure 3.15 shows time series of the 12 selected sites in the region. As with the EAC4 modelled data in Section 3.4.1, higher concentrations of $PM_{2.5}$ are found between December and May, with the largest concentrations of $PM_{2.5}$ at the end of March and beginning of April. The period between April and December is relatively constant, with no large peaks in pollution or any other noticeable trends.

Spatial Patterns

Five points in the distribution at each location were mapped and are shown in Figure 3.16 - the minima, median, maxima, and the width of the range ($\max - \min$). The minimum value shown in the top left corner of the figure shows that site I, to the west of the city,

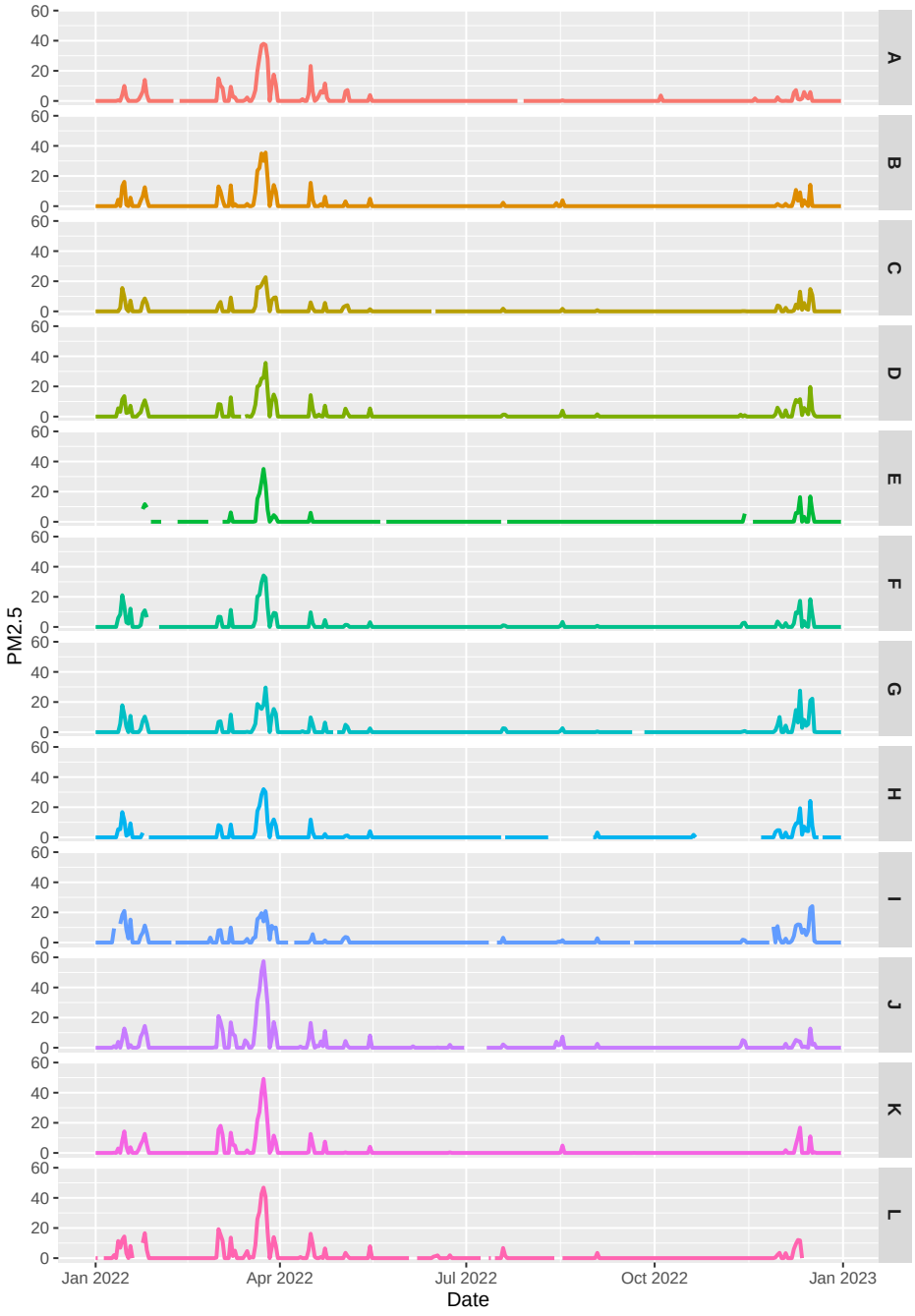

Figure 3.15: Time series for the 12 observation stations of the AURN in the Greater London area with at least 75% complete temporal records for the days in 2022.

has a minimum of $0\,\mu g/m^3$ while site E in urban London has a minimum of $4\,\mu g/m^3$. The pattern is visible in the median (top right), where locations to the west of the city have lower concentrations than those to the east. In the maximum value map in the bottom row, the large discrepancy between sites is clear. Site J on the east coast has a maximum of $71\,\mu g/m^3$ while site I has a maximum of $39\,\mu g/m^3$. A possible explanation for this behaviour is the pollution generated by marine transportation moving through the river Thames as well as other pollution-favourable geographical and meteorological conditions. Finally, as with the EAC4 data, the width of the range at each observation location is strongly determined by the maxima, and thus shows similar spatial patterns.

Overall, spatial patterns are less smooth than those in the EAC4 data. Nonetheless, the divide between east and west of the London city centre is still visible. Unlike the EAC4 data, differences between locations are large, highlighting the difference between measurements taken in-situ and those modelled from remote-sensing data.

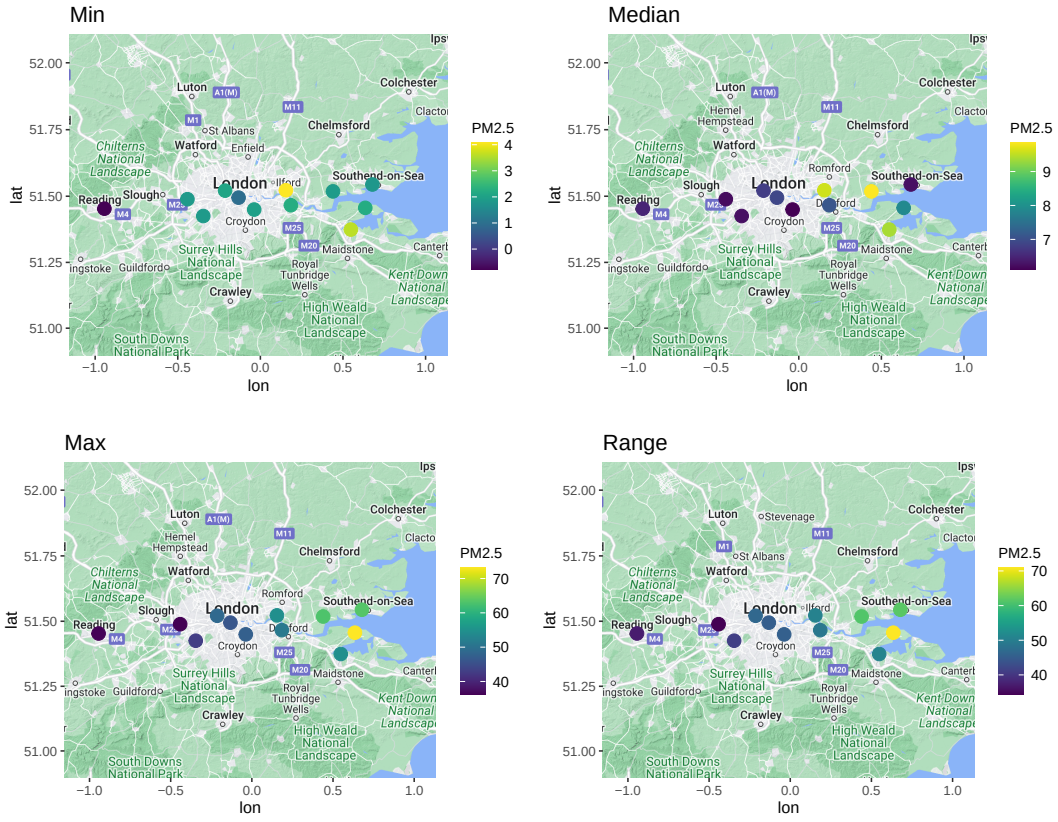

Figure 3.16: Spatial distribution of the AURN observation stations in the Greater London area for 2022. The plots show the minimum, median, maximum, and width of the range of values at each observation station.

Chapter 4

Dependence in Unreplicated Extremes

4.1 The Importance of Extremal Dependence

Heavy metal (HM) soil contamination, i.e., the presence of high and extremely high concentrations of HM in the soil, poses a significant risk to living organisms. As such, an accurate understanding of the spatial distribution and magnitude of the contamination is central to public health and remediation efforts.

Undertaking extreme value analysis on such complex applications is not without its challenges. Due to the high cost of collection and the slow-changing nature of soil concentrations, the data are unreplicated, meaning they consist of a single sample at each location. Moreover, considering more than one contaminant at a time entails assessing extremal dependence between contaminants, which can be diverse (see Section 2.2.2). While some pairs exhibit asymptotic dependence, many display decaying dependence at increasingly extreme observations, a common occurrence in environmental applications (Huser and Wadsworth, 2022). Capturing the appropriate dependence structure is a central consideration of multivariate extreme value models due to several reasons. First, it results in a more accurate understanding of the joint extremes, e.g., all contaminants displaying extreme values simultaneously. Increased accuracy of the relationship of contaminants at extreme levels helps identify contaminants. As a result, it is possible to enforce pollution control measures at a localised scale, increasing effectiveness. Finally, improving the accuracy of models of extreme values in soil contamination helps assess risk and manage exposure.

In this chapter, we discuss the challenge of diversity in extremal dependence structures in heavy metal soil contaminants by exploring and modelling the extremal dependence between all possible contaminant pairs. To overcome the problem of lack of replicates at each location, the concentrations of each individual contaminant are pooled together, discarding their spatial information and considering each observation as a replicate at a univariate random variable. The spatial aspect of the data is considered in the modelling

approach covered in Chapter 5, where we propose an approach for the spatial modelling of unreplicated bivariate extreme observations.

In this Chapter we present a comparison of multivariate extreme value models to model the dependence between extreme heavy metal concentrations in soil survey samples. In Section 4.2, we review a specific methodology for the multivariate generalised Pareto distribution, a model for asymptotically dependent data which results in a rigid, constant dependence between components. Section 4.3 provides an overview of the Exponential Factor Copula model, which is a subasymptotic model that displays flexible dependence between components. This section also includes a simulation study to explore model behaviour under different dependence specifications. A comparison of both models using the G-BASE data set is provided in Section 4.4. Finally, Section 4.5 provides conclusions and a discussion of our findings.

4.2 Multivariate Generalized Pareto Distribution (MGPD)

The multivariate extension of the Generalized Pareto distribution (MGPD) is mentioned briefly in Section 2.2.2 but covered in more detail in this section.

In the univariate case, the GPD is well-defined and can be easily fitted to threshold exceedances using frequentist or Bayesian inference so long as the shape parameter permits it (see Section 2.2.1). This is not the case in the multivariate ($d \geq 2$) setting for two reasons. First, there is no single definition for what constitutes a threshold exceedance in the multivariate case. Second, the family of limiting distributions that arise from any definition of a multivariate threshold exceedance is not parametric (Rootzén and Nader, 2006).

Rootzén and Nader (2006) define a threshold exceedance in the multivariate case as any point $\boldsymbol{y} \in \mathbb{R}^d$ where $d \geq 2$ that has at least one element exceeding its corresponding threshold, i.e., $y_j > u_j$ for $j \in \{1, ..., d\}$.

The MGPD is derived from a max-stable distribution, inheriting its marginal parameter and dependence structure after a transformation. Consider G from (2.17) to be the multivariate generalised extreme value distribution (MGEVD), which is defined as

$$\Pr(\boldsymbol{a}^{-1}(\boldsymbol{M}_n - \boldsymbol{b}_n) \leq \boldsymbol{z}) = F^n(\boldsymbol{a}_n \boldsymbol{z} + \boldsymbol{b}_n) \to G(\boldsymbol{z}), \quad n \to \infty,$$

where $M_{x,n} = \max_{i=1,...,n} \{X_i\}$ and $M_{y,n} = \max_{i=1,...,n} \{Y_i\}$ are component-wise block maxima so that $\boldsymbol{M}_n = (M_{x,n}, M_{y,n})$, and $\boldsymbol{a}_n > 0$ and $\boldsymbol{b}_n \in \mathbb{R}$ are normalising vectors.

Here, we let $\boldsymbol{x} = \boldsymbol{y} - \boldsymbol{u}$, where $\boldsymbol{u}$ is a d-dimensional vector of threshold values. As in (2.17), $\boldsymbol{X}$ has cdf F. The existence of $\boldsymbol{a}_n$ and $\boldsymbol{b}_n$ ensure that $0 < G_j(0) < 1$ for

$j \in \{1, ..., d\}$. This implies that

$$\lim_{n\to\infty} n\{1 - F(\boldsymbol{a}_n\boldsymbol{x} + \boldsymbol{b}_n)\} = -\ln G(\boldsymbol{x}).$$

It then follows that for all $\boldsymbol{x} \in \mathbb{R}^d$ such that $G_j(x_j)$ for all $j \in \{1, ..., d\}$,

$$\lim_{n\to\infty} \Pr(\boldsymbol{a}_n^{-1}(\boldsymbol{X} - \boldsymbol{b}_n)|\boldsymbol{X} > \boldsymbol{b}_n) = \frac{\ln G(\min\{\boldsymbol{x}, \boldsymbol{0}\}) - G(\boldsymbol{x})}{\ln G(\boldsymbol{0})}.$$

Let $\boldsymbol{\eta} \in [-\infty, 0)^d$ denote a vector of lower endpoints of the marginals of G, $G_1, ..., G_d$, then it follows that as $n \to \infty$

$$\mathcal{L}[\max\{\boldsymbol{a}^{-1}(\boldsymbol{X} - \boldsymbol{b}_n), \boldsymbol{\eta}\}|\boldsymbol{X} > \boldsymbol{b}_n] \to H,$$

where $\mathcal{L}$ is the law of the random variable, H is a MGPD. H is said to be associated with G through

$$H(\boldsymbol{x}) = \frac{\ln G(\min\{\boldsymbol{x}, \boldsymbol{0}\}) - \ln G(\boldsymbol{x})}{\ln G(\boldsymbol{0})}.$$

Some properties of H should be noted. For example, its marginal distributions are not univariate GPD because not all components of the vector $\boldsymbol{y}$ exceed their respective thresholds $\boldsymbol{u}$. But if a component exceeds its threshold, its conditional marginal distribution is GPD. The non-conditional marginal distributions of the MGPD are still able to place density on negative threshold exceedances (non-exceedances) using the natural lower endpoint of the univariate GEVD. Just like the max-stable property of the MGEVD, the MGPD has the property of threshold stability. This means if $\boldsymbol{X} \sim H$ and if at least one element of $\boldsymbol{w}$ is positive $w_j > 0$, with $H(\boldsymbol{w}) < 1$ and $\boldsymbol{\sigma} + \boldsymbol{\xi}\boldsymbol{w} > 0$, then $\boldsymbol{X} - \boldsymbol{w}|\boldsymbol{X} > \boldsymbol{w}$ is MGPD with parameters $\boldsymbol{\sigma} + \boldsymbol{\xi}\boldsymbol{w}$ as the scale parameter and $\boldsymbol{\xi}$ as the shape parameter where $\boldsymbol{\sigma}$ and $\boldsymbol{\xi}$ are the scale and shape parameters of the associated MGEV G. As in the univariate case, increasing the thresholds will result in a different scale but the same shape parameter.

As mentioned in 2.2.2, the MGPD has no unique parametric form, but can be represented using one of the four representations proposed by Rootzén *et al.* (2018b,a). In order to use these representations, the data must first be transformed from $\boldsymbol{X}$ to a standardised version $\boldsymbol{X}_0$, through the transformation

$$\boldsymbol{X} = \boldsymbol{\sigma}\frac{x^{\boldsymbol{\xi}\boldsymbol{X}_0} - \boldsymbol{1}}{\boldsymbol{\xi}},$$

where $\boldsymbol{\sigma}$ and $\boldsymbol{\xi}$ are the parameters of the marginal G distributions of $\boldsymbol{X}$. The parameters of $\boldsymbol{X}_0$ after the transformation become $\boldsymbol{\sigma} = \boldsymbol{1}$ and $\boldsymbol{\xi} = \boldsymbol{0}$. While it is possible to fit the standardised $\boldsymbol{X}_0$ using various representations of the MGPD, the work presented in

this chapter uses only the "U" representation proposed by (Rootzén and Nader, 2006), as it is most appropriate for the purpose of modelling and simulation. To obtain this parametrisation, let $\boldsymbol{U}$ be a random vector in $\mathbb{R}^d$ with density f_U under the condition that $0 < \mathrm{E}(e^{U_j}) < \infty \quad \forall j \in \{1, ..., d\}$. The resulting density is

$$h_U(\boldsymbol{x}; \mathbf{1}, \mathbf{0}) = \frac{\mathbb{1}\{\max(\boldsymbol{x}) > 0\}}{\mathrm{E}[e^{\max(\boldsymbol{U})}]} \int_0^\infty f_U(\boldsymbol{x} + \log(t))\mathrm{d}t, \tag{4.1}$$

where $\mathrm{E}[e^{\max(\mathbf{U})}] = \int_0^\infty \Pr(\max(\mathbf{U}) > \log(t))\mathrm{d}t$. The probabilities of the marginals producing an exceedance are given as

$$\Pr(X_{0,j} > 0) = \frac{\mathrm{E}[e^{U_j}]}{\mathrm{E}[e^{\max(\mathbf{x})}]}.$$

Kiriliouk *et al.* (2019) proposed a definition for f_U that allows for a simpler construction. Let $\mathbf{V} \in \mathbb{R}^d$ be a random vector of independent components such that its joint density is the product of the independent marginal densities, $f_v(\boldsymbol{\nu}) = \prod_{j=1}^d f_j(\nu_j)$. No restrictions are placed on f_v, allowing almost any density to be used. Some distributions, however, result in simpler integrals of closed-form and thus are preferred to others. The reverse exponential distribution is one of such distributions, having the form

$$f_j(\nu_j) = \alpha_j e^{\alpha_j(\nu_j + \beta_j)},$$

where the support for ν_j is $(-\infty, -\beta_j)$, the scale parameter is $\alpha_j > 0$, and the location parameter is $\beta_j \in \mathbb{R}$. This distribution is equivalent to the exponential distribution with the small difference that the negative term in the exponential is introduced because $\nu_j + \beta_j < 0$. Introducing only the bivariate case, one can substitute the product of these densities in (4.1) such that $f_U \equiv f_V$, yields the h_U density in closed form

$$h_U(\mathbf{x};\mathbf{1},\mathbf{0}) = \frac{(e^{-\max(\mathbf{x}+\boldsymbol{\beta})})^{\sum_{j=1}^2 \alpha_j + 1}}{E[e^{\max(\mathbf{U})}]} \frac{1}{1 + \sum_{j=1}^2 \alpha_j} \prod_{j=1}^2 \alpha_j (e^{x_j + \beta_j})^{\alpha_j}. \tag{4.2}$$

Because (4.2) is fitted to standardised components at the unit scale, analogously to a copula model, the model captures the dependence between components, with $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)$ and $\boldsymbol{\beta} = (\beta_1, \beta_2)$ as dependence parameters. Kiriliouk *et al.* (2019) explores the possibility of common parameters, $\alpha = \alpha_1 = \alpha_2$ and/or $\beta = \beta_1 = \beta_2$ on standardised data, but uses unique α_j and unique β_j to fit the model at real scale. To ensure the identifiability of the location parameter, the parameter for the last component, β_2, was permanently fixed to 0 (Kiriliouk *et al.*, 2019).

The standard approach to fitting the MGPD is using a censored likelihood (Smith, 1997; Ledford and Tawn, 1997), where the density in (4.2) is the contribution of an trans-

formed observation $(\boldsymbol{Y}-\boldsymbol{u})$ to the likelihood only when it is considered "sufficiently large", exceeding a threshold $\boldsymbol{m}$, for $\boldsymbol{m} \leq \boldsymbol{0}$, making the contribution of each observation to the likelihood relative. Observations with only one component exceeding its threshold are partially censored; and observations where no exceedances are fully censored. This is done for two reasons. First, the marginals of the MGPD place density on a lower endpoint when it is not an exceedance and can incorrectly influence the likelihood. The second reason is because the parameters that control dependence have been shown to be larger than when using censored estimation (Huser *et al.*, 2016). Let $C \subset D = \{1, ...d\}$ contain the indices for the elements of $\boldsymbol{Y}-\boldsymbol{u}$ that fall below $\boldsymbol{m}$, i.e., $Y_j - u_j \leq m_j$ for $j \in C$, and $Y_j - u_j > m_j$ for $j \in D/C$, with at least one such $Y_j > u_j$. Each realization of $\boldsymbol{Y}$ then has the likelihood distribution

$$h^C\left(y_{D/C} - u_{D/C}, m_C; \boldsymbol{\theta}\right) = \int_{\substack{\times(-\infty, u_j+m_j] \\ j \in C}} h(y-u; \boldsymbol{\theta}) \mathrm{d}y_C, \tag{4.3}$$

Multiplying all contributions to the likelihood, the censored likelihood function for this model has the form

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} h^{C_i}(y_{i,D/C_i} - u_{D/C_i}, m_{C_i}; \boldsymbol{\theta}), \tag{4.4}$$

where $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{1}, \boldsymbol{0})$ is a vector of covariates, and h^C is the likelihood contribution defined in (4.3).

As an advantage, censored likelihood produces a more stable likelihood estimation and avoids the known bias in parameter estimation that has been known to affect dependence (Huser *et al.*, 2016).

It is possible to assess the fit of the MGPD on the dependence using χ - the coefficient of tail dependence introduced in Section 2.2.2 which we will denote as χ_{MGPD}. For the U representation, dependence is solely defined by the dependence parameter $\boldsymbol{\alpha}$. In the bivariate case where there are two distinct α_j, χ_{MGPD} is defined as

$$\chi_{MGPD} = 1 - \left(\frac{1+\alpha_{(1)}^{-1}}{1+\alpha_{(2)}^{-1}}\right)^{1+\alpha_{(2)}} \frac{\alpha_{(1)}}{\alpha_{(2)}} \frac{1}{1+\alpha_1+\alpha_2},$$

where $\alpha_{(1)} = \max\{\alpha_1, \alpha_2\}$ and $\alpha_{(2)} = \min\{\alpha_1, \alpha_2\}$ (Kiriliouk *et al.*, 2019). In the case where $\alpha_1 = \alpha_2 > 0$, χ_{MGPD} is defined as

$$\chi_{MGPD} = 1 - \frac{1}{1+2\alpha}.$$

4.3 Exponential Factor Copula Model (EFC)

Sub-asymptotic models were originally developed to improve poor convergence of extreme data to limiting distributions (Lugrin *et al.*, 2021). A side effect of the improved conver-

gence is a flexible dependence structure, rendering them a suitable alternative for situations in which the constant dependence models (max-stable and threshold-stable models) prove too rigid. Although many sub-asymptotic models exist which are used for decaying dependence cases - cases where dependence decays at increasing values but does not reach asymptotic independence, (Wadsworth and Tawn, 2012; Huser *et al.*, 2017; Huser and Wadsworth, 2019; Castro-Camilo and Huser, 2020), we focus on the Exponential Factor Copula model (EFCM), used by Castro-Camilo and Huser (2020), which is a special case of the factor copula models proposed by Krupskii *et al.* (2018).

A copula is a multivariate function with uniform margins. In this framework, the univariate marginal distributions and the dependence structure are handled separately, with the copula modelling only the dependence structure.

For a D-dimensional multivariate random variable $X = \{X_1, ..., X_D\}$, with distribution $F(x_1, ..., x_D) = P(X_1 \leq x_1, ..., X_D \leq x_D)$ and marginals $F_j(x_j) = P(X_j \leq x_j)$, the copula C has distribution

$$C(u_1, ..., u_D) = P(U_1 \leq u_1, ..., U_D \leq u_D),$$

where $0 \leq u_j \leq 1$, and U_j is a unit scale random function defined as $U_j = F_j(X_j), j = 1, ..., D$. Sklar's Theorem (Sklar, 1959) proved that for any multivariate distribution F with continuous marginal distributions $F_1, .., F_D$, a unique copula exists such that

$$F(x_1, ..., x_D) = C\{F_1(x_1), ..., F_D(x_D)\},$$

enabling any multivariate distribution to be expressed in terms of a copula and its marginal distributions. Furthermore, the copula can be directly expressed in terms of F and its marginal distributions as

$$C(u_1, ..., u_D) = F\{F_1^{-1}(u_1), ..., F_D^{-1}\}.$$

The density of the copula can be obtained through differentiation as

$$c(u_1, ..., u_D) = \frac{f\{F_1^{-1}(u_1), ..., F_D^{-1}(u_D)\}}{\prod_{i=1}^{D} f_i\{F_i^{-1}(u_i)\}}.$$

Factor copula models are descendants of the spatial factor models. These models are based on the assumption of a common latent random factor for all spatial locations (Wang and Wall, 2003; Hogan and Tchernis, 2004; Irincheeva *et al.*, 2012). They can be interpreted as the case where an unobserved variable explains the dependence structure. Krupskii and Joe (2015) extended the theoretical framework of a common factor to explain the dependence structure to copulas and later spatial copulas (Krupskii *et al.*, 2018). The

spatial approach by Krupskii *et al.* (2018) is based on the process

$$W(\boldsymbol{s}) = Z(\boldsymbol{s}) + V, \tag{4.5}$$

where $\boldsymbol{s} \in \mathbb{R}^n$, Z is a Gaussian Process (GP), and V is a common factor with an arbitrary distribution that is independent of location. Castro-Camilo and Huser (2020) proposed using an exponential factor copula (EFC) where V is an exponentially-distributed random variable with parameter λ. W results in a Gaussian location mixture, i.e., a standard Gaussian process with an exponentially distributed mean. They used this model to capture the spatial dependence of extreme precipitation in the contiguous USA. The covariance structure of $Z(\boldsymbol{s})$ was obtained using a stationary Matern correlation function $\rho(h)$.

In the bivariate setting, Z has a stationary 2×2 correlation matrix Σ_Z:

$$\Sigma_Z = \begin{pmatrix} 1 & \rho(h) \\ \rho(h) & 1 \end{pmatrix},$$

where $\rho(h)$ is the stationary correlation between components which is solely a function of the distance between them.

V and Z are independent, allowing the joint distribution of W to be expressed as

$$F_2^W(w_1, w_2) = \lambda \int_0^\infty \Phi_2(w_1 - \nu, w_2 - \nu; \Sigma_Z) \times \exp(-\lambda\nu)\mathrm{d}\nu,$$

with density

$$f_2^W(w_1, w_2) = \lambda \int_0^\infty \phi_D(w_1 - \nu, w_2 - \nu; \Sigma_Z) \times \exp(-\lambda\nu)\mathrm{d}\nu,$$

where $\Phi_2(\cdot; \Sigma)$ and $\phi_2(\cdot, \Sigma)$ are the multivariate standard normal distribution and density (respectively) with correlation matrix Σ. The marginal distribution F can then be expressed as

$$F_1^W(w; \lambda) = \Phi(w) - \exp(\lambda^2/2 - \lambda w)\Phi(w - \lambda).$$

As with peaks-over threshold models, inference for the EFC model can be carried out using censored likelihood (Smith, 1997; Ledford and Tawn, 1997), similarly to (4.4) for the MGPD. The full log-likelihood is obtained by summing the log-likelihood contributions of locations where all components exceed their respective thresholds (non-censored; NC); locations where only one component exceeds its respective threshold (partially-censored; PC); and locations with no exceedances (fully-censored; FC). The full log-likelihood is

given as

$$\ell(\theta) = \sum_{i \in NC} \log f_2^W(w_{i1}, w_{i2}; \theta) - \sum_{i \in NC} \sum_{j=1}^{2} \log f_1^W(w_{ij}; \lambda) + N_{FC} \times \log F_2^W(w_1^*, w_2^*; \theta) + \sum_{i \in PC} \log \partial_{J_i} F_2^W(\max(w_{i1}, w_1^*), \max(w_{i2}, w_2^*)) - \sum_{i \in PC} \sum_{j=J_i} \log f_1^W(w_{ij}; \lambda), \tag{4.6}$$

where $\theta = (\lambda, \rho)$, u_j^* for $j = \{1, 2\}$ are the marginal thresholds in the uniform scale, u_{ij} are the scores of each observation at uniform scale, $w_j^* = (F_1^W)^{-1}(u_j^*; \lambda)$, $w_{ij} = (F_1^W)^{-1}(u_{ij}; \lambda)$, J_i indicates the component which exceeds its respective threshold at that location, NC is the index of the non-censored locations, FC indicates the fully-censored locations and PC indicates the partially-censored locations.

Maximising the censored likelihood (4.6), allows for the estimation of model parameters λ and ρ.

Estimation of the dependence in the EFC model can be summarised using the coefficient of tail dependence, $\chi_{EFC}(u)$ (Castro-Camilo and Huser, 2020), obtained using

$$\chi_{EFC}(u) = 2 - f(u) - g(u)h(u),$$

where

$$f(u) = \frac{1 - \Phi\{z(u), z(u)\, \Sigma_Z\}}{1 - u}, \quad g(u) = \frac{\exp\{\lambda^2/2 - \lambda z(u)\}}{1 - u}, \quad h(u) = 2\Phi\{\lambda\sqrt{(1-\rho)/2}; \Omega\},$$

and where Φ is a bivariate Gaussian distribution with correlation $\rho(h)$, and u represents a data percentile above the smallest appropriate threshold u_0. The covariance matrix Ω can be expressed as

$$\Omega = \begin{pmatrix} 1 & -\sqrt{(1-\rho(h))/2} \\ -\sqrt{(1-\rho(h))/2} & 1 \end{pmatrix}.$$

A simulation study was performed to gain intuition and assess the EFC model's capacity to capture different scenarios of decaying dependence. The model was fitted in the frequentist framework by maximising the censored likelihood in `R` using the `optim` function and the L-BFGS optimisation method in a similar manner to the code provided in the supplementary material of Castro-Camilo and Huser (2020).

4.3.1 Simulation Study: Investigating the EFC's Performance with Decaying Dependence

The work presented in this chapter focuses on comparing two extreme value models with distinct dependence structures, the MGPD and the EFC model. Understanding the MGPD dependence paradigm is straightforward, as it is constant throughout the range of the data, even at extreme percentiles, and can be easily estimated from model parameters. The EFC model, on the other hand, assumes a flexible dependence and thus can capture decaying dependence. To gain an intuition for the flexible dependence structure of the EFC model as well as assess its performance, we performed a simulation study where we tested the ability of the EFC model to recover the true dependence structure (through the coefficient of tail dependence χ) as well as the true model parameters.

Bivariate data are simulated directly from the model in (4.5), which requires the simulation of a bivariate normal distribution (Z), with 0 mean and Σ_Z correlation matrix, and the addition of an exponentially-distributed random variable of rate λ. Nine simulation scenarios denoted as simulations A through I (Table 4.1) were set up to investigate a wide range of parameter sets that can mimic the real data, ranging from weak to strong correlation between components ($\rho = 0.4$ to $\rho = 0.9$, respectively); and weak to strong decay in the tail dependence structure ($\lambda = 5$ to $\lambda = 20$, respectively). For each simulation scenario, a total of $n = 3000$ observations were simulated. The process was repeated 1000 times for each scenario.

Table 4.1: Parameters ρ and λ of the data-generating process, inducing different decaying dependence scenarios A to I.

Simulation	λ	ρ
A	5	0.4
B	5	0.7
C	5	0.9
D	10	0.4
E	10	0.7
F	10	0.9
G	20	0.4
H	20	0.7
I	20	0.9

Once the data were simulated using the parameters in Table 4.1, they were transformed to a uniform scale in order to satisfy requirements for the use of a copula. A uniform scale was obtained for both simulated components using the non-parametric rank transformation

$$u_i = \frac{\text{rank}(y_i)}{N+1}, \quad i = 1, ..., N,$$

where y_i represents the i-th observation of an individual component. The EFC model was fitted to all 1000 simulations of each scenario. A numerical summary of the results in terms of root mean square error (RMSE) for the model parameters and coefficient of tail dependence χ is given in Table 4.2.

Table 4.2 shows that the estimation of the dependence parameters, $\hat{\lambda}$ and $\hat{\rho}$, improves with increasing values of λ. Standard errors are high in simulations A to C, where tail decay is weak ($\lambda = 5$). For $\hat{\lambda}$, standard errors range from 8.83 to 11.91, while they range from 0.02 to 0.07 for $\hat{\rho}$. For simulations where tail decay is strong, meaning $\lambda = 20$ as in scenarios G to I, uncertainty is reduced to the range 1.8 to 2.33 for $\hat{\lambda}$ and 0.01 to 0.03 for $\hat{\rho}$. Differences between the χ and $\hat{\chi}$, as shown in the last column, are generally lower for simulations G to I than A to F.

Table 4.2: Median of estimated model parameters using censored likelihood for the nine simulated scenarios with standard deviation given in parenthesis.

Simulation	λ	$\hat{\lambda}$	$\hat{\lambda}$**-RMSE**	ρ	$\hat{\rho}$	$\hat{\rho}$**-RMSE**	$\chi-$**RMSE**
A	5	4.24(8.83)	10.29	0.4	0.39(0.07)	0.07	0.02(0.01)
B	5	3.6(11.91)	13.53	0.7	0.69(0.04)	0.05	0.02(0.01)
C	5	3.39(10.66)	11.95	0.9	0.89(0.02)	0.02	0.01(0.01)
D	10	9.6(5.86)	5.85	0.4	0.39(0.06)	0.06	0.01(0.01)
E	10	8.93(6.92)	7.04	0.7	0.69(0.04)	0.05	0.02(0.01)
F	10	7.4(6.75)	7.14	0.9	0.89(0.02)	0.02	0.01(0.01)
G	20	19.65(2.03)	2.33	0.4	0.4(0.03)	0.03	0.01(0.01)
H	20	19.58(1.8)	2.17	0.7	0.7(0.02)	0.02	0.01(0.01)
I	20	21.52(1.84)	2.65	0.9	0.9(0.01)	0.01	0.01(0.01)

The estimates for the coefficients of tail dependence for the 1000 simulations of each scenario are shown in Figure 4.1. From the true $\chi(u)$ values shown as the black lines, it is possible to see that the dependence decays at increasing values of u across all simulations, as intended, but simulations with $\rho = 0.4$ and $\rho = 0.7$ (A, B, D, E, G, and H) show stronger decay than those with $\rho = 0.9$ (C, F, and G). We can also see the behaviour noted in Table 4.2; that is, the accuracy of estimation improves as λ increases. Indeed, the most extreme percentiles of the data for simulations G to I (bottom row in Figure 4.1) are estimated more closely than A to F (first and second row in Figure 4.1).

While the estimation of $\chi(u)$ improves with a stronger decay in tail dependence, the results of the simulation show that the model accurately estimates model parameters for all simulation scenarios. Moreover, dependence is captured appropriately, even at high percentiles for all scenarios and does so with consistency by showing small RMSE values in the estimation of $\hat{\chi}$ against the true value.

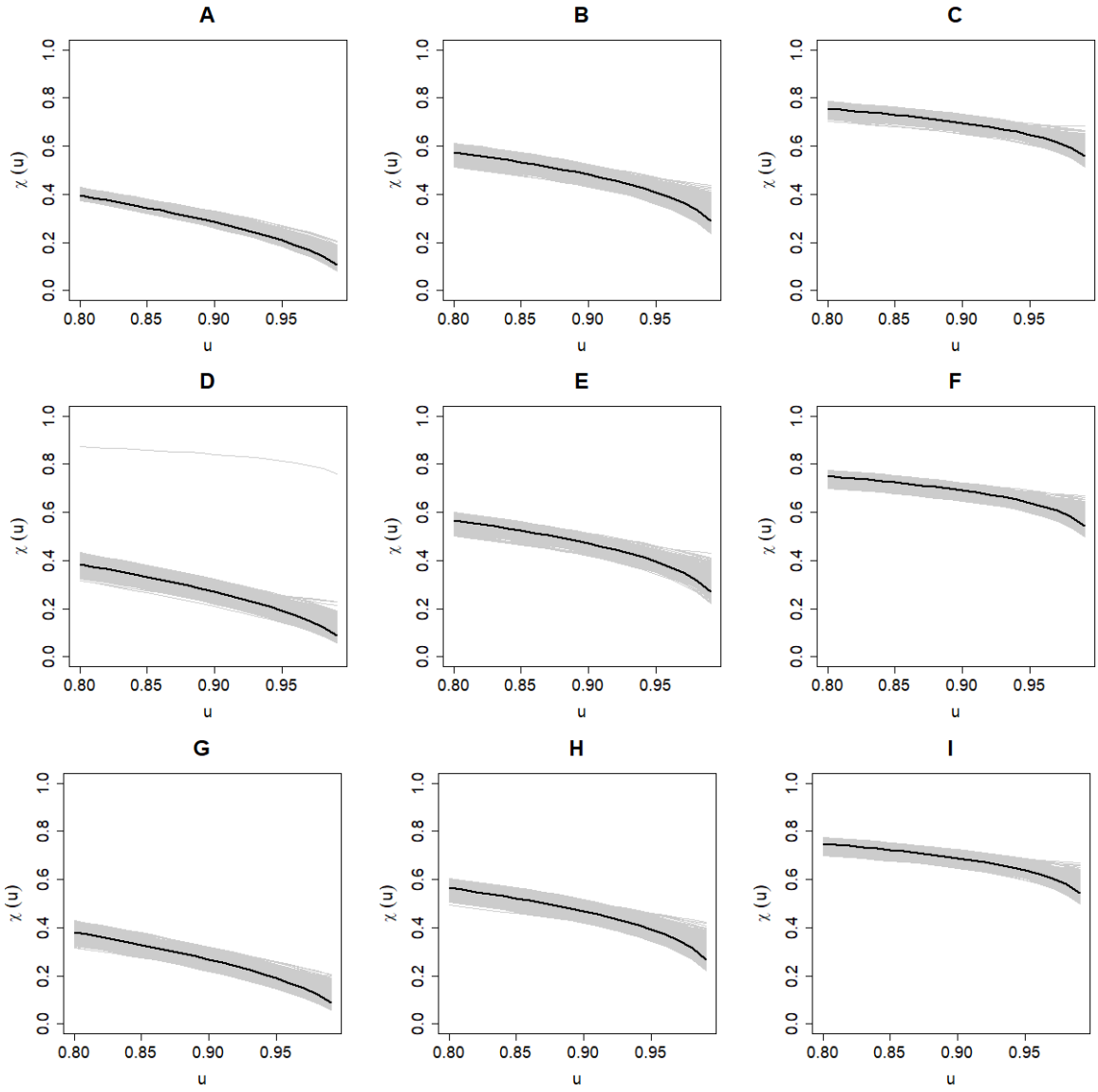

Figure 4.1: The grey lines correspond to tail dependence coefficients (χ) for percentiles between 0.8 and 0.95 for the 1000 simulations performed for the simulation scenarios A-F. The line in black indicates the true tail dependence coefficient for the simulated data.

4.4 Extremal Dependence of Heavy Metal Contaminants

The case study presented in this chapter uses data from the G-BASE dataset from the BGS for the observations in the Glasgow Conurbation. A description and exploratory analysis of the data is provided in 3.2. Here, we explore the dependence structure by the MGPD, a constant dependence model, and the EFC, a subasymptotic dependence model, to assess the dependence behaviour between the important pairs of HM contaminants: As, Cr, Cu, Ni, Pb, and Zn.

The first step in the modelling process is to find an appropriate threshold for the data. Mean residual life plots (2.14) were evaluated for each element individually to ascertain a sensible common threshold for all contaminants. Given that threshold values (u) that exceed the smallest appropriate threshold u_0, i.e., $u > u_0$, are considered suitable for extreme value analysis, the largest u_0 from the individual contaminants was chosen as a common threshold. In this case, the common smallest appropriate threshold across all contaminants was found to be the 82nd quantile. In the second step, we fit a flexible, parametric quantile regression (QR) to each contaminant individually to obtain a surface of the estimated 82nd quantiles in order to identify exceedances of this threshold. The QR model for each included covariates that have been shown to affect HM concentrations, such as elevation, slope, aspect, multiresolution index of valley bottom flatness (MMRVBF), the complementary multiresolution index of the ridge top flatness (MRRTF) and topographic wetness index (TWI) obtained from the Ordinance Survey (OS) digital terrain model (`https://www.ordnancesurvey.co.uk/products/os-terrain-50`) and proximity to nearest road using the Open Roads data set of the OS (`https://www.ordnancesurvey.co.uk/products/os-open-roads`), as advised by the modelling selection processed carried out by Johnson *et al.* (2017) for the purpose of mapping heavy metal contaminants in soil. The approach resembles that in Youngman (2019) and was fitted using the `evgam` package in R (Youngman, 2022). Observations of each contaminant exceeding their respective, modelled 82nd quantile were considered exceedances. A GPD is then fitted to individual contaminant exceedances using maximum likelihood, and the data are transformed into an approximated uniform scale using the inverse probability transform of the fitted cdf.

The MGPD and the EFC were fitted to the bivariate threshold exceedances in a uniform scale using a censored likelihood; the estimated parameters are given in Table 4.3. Figures 4.2, 4.3 and 4.4 show plots of the estimated and empirical χ values for both models along with 95% confidence intervals obtained using 500 point-wise bootstrap samples.

The χ plots show the diversity of possible dependence structures in pairs of HM elements with concentrations exceeding the 82nd quantile. From Figures 4.2, 4.3 and 4.4 and Table 4.3, we can see that certain elements have a higher dependence than others. Pairs Pb-Cu, Pb-Zn, Zn-Cu, Cu-Ni, Zn-Ni, Pb-Ni and Pb-Ni have χ values above 0.5 at

$u = 0.82$ quantile, but Zn-Ni and Pb-Ni remain above 0.4 at higher quantiles ($u = 0.99$). Low-dependence pairs are Cr-Cu, Pb-Cr, and Cr-As, as they have χ values below 0.4 through all quantiles. There is also a wide range of diversity in extremal dependence. While some pairs display the expected non-constant dependence structures, i.e., substantial decay at high quantiles, the most notable cases are pairings with Cr, such as Cr-Ni, Cr-Zn and Cr-Cu.

The simulation study performed in 4.3.1 showed that the EFC model is capable of capturing both constant and non-constant dependence structures; however, the case study shows it is less capable of capturing a nearly constant dependence in real data. Pairs As-Cu, As-Ni, Pb-As, Zn-Cu, Zn-Ni generally display stable dependence at lower percentiles, better captured by the MGPD model of constant dependence, but the decaying dependence at extreme percentiles is captured more appropriately by the EFC model. Strongly non-stationary pairs such as Cr-Cu, Cr-Zn, Cu-Ni, Pb-Cr, Pb-Cu, Pb-Ni profit from the flexibility awarded by the EFC model with an estimated χ within the data's bootstrapped confidence intervals. Pairs for which both models behave poorly, such as As-Zn, Cr-As, Cr-Ni, Pb-As, and Pb-Zn tend to show a preference for the EFC model at high quantiles. A further discussion of these results is given in Chpater 8.

Table 4.3: Summary of model outputs for the Exponential Factor Copula and the Multivariate Generalized Pareto. RMSE values for the discrepancy between modelled and empirical dependence between pairs, as measured by χ.

			EFC			**MGPD**		
Pair	$\hat{\chi}_{0.82}$	$\hat{\chi}_{0.99}$	λ	ρ	RMSE	$\{\alpha_1, \alpha_2\}$	$\{\beta_1, \beta_2\}$	RMSE
As-Cu	0.47	0.4	28.30	0.70	0.09	1.9, 0.4	0.34, 0	0.10
Cr-As	0.34	0.14	24.90	0.60	0.12	3.6, 0.8	0.36, 0	0.13
As-Ni	0.46	0.18	23.60	0.70	0.09	2.2, 0.5	0.33, 0	0.09
Pb-As	0.48	0.29	20.60	0.70	0.08	2, 0.6	0.51, 0	0.09
As-Zn	0.46	0.22	11.80	0.70	0.10	2.1, 0.5	0.35, 0	0.10
Cr-Cu	0.39	0.07	25.90	0.50	0.04	3.7, 0.8	0.26, 0	0.11
Cu-Ni	0.63	0.43	11.70	0.80	0.03	1.1, 0.5	0.61, 0	0.09
Pb-Cu	0.71	0.50	13.70	0.90	0.05	0.8, 0.3	0.5, 0	0.13
Zn-Cu	0.68	0.43	11.70	0.80	0.04	1.1, 0.4	0.46, 0	0.07
Cr-Ni	0.54	0.18	26.30	0.60	0.07	3.5, 0.7	0.4, 0	0.13
Pb-Cr	0.38	0.22	26.90	0.50	0.05	3.7, 0.7	0.39, 0	0.10
Cr-Zn	0.41	0.04	25.20	0.50	0.04	3.8, 0.8	0.4, 0	0.11
Pb-Ni	0.56	0.36	11.80	0.80	0.06	1.4, 0.5	0.51, 0	0.12
Pb-Zn	0.68	0.36	18.50	0.80	0.04	1.1, 0.4	0.5, 0	0.08
Zn-Ni	0.59	0.32	23.10	0.80	0.03	1.4, 0.5	0.53, 0	0.09

4.5 Discussion and Conclusion

Fields without natural replications have been historically neglected as potential applications of EVT. Extremes in soil are usually modelled using conventional geostatistical techniques, which typically model the mean under Gaussian assumptions. These Gaussian assumptions have been proven unsuitable to model extreme observations (Coles, 2001) and predict extreme behaviour. The detrimental effect of exposure to HM contamination puts pressure on practitioners to produce accurate and useful predictions of extreme values.

The application of EVT to soil data is challenging. The lack of replication, as mentioned before, has proved to be the most significant obstacle. In the univariate setting, a natural solution to the lack of replicates treats the data as a replicated set from a single site. In this chapter, however, the focus is on the extremal dependence of contamination by multiple HM elements. The extension of the univariate solution to our scenario is to neglect spatial dependence and treat each component as coming from a non-spatial process. This allows us to characterise the extremal dependence between components but not how that extremal dependence changes with location. Two models for extremal dependence were fitted to observed data under this framework - the EFC model of (Castro-Camilo and Huser, 2020) and the MGPD (Kiriliouk *et al.*, 2019). The EFC model is a flexible, sub-asymptotic model. For HM pairs in the Clyde River Basin, this model produced better fits for non-constant dependence structures and was particularly good at capturing decaying dependence at very high percentiles. Pairs that displayed a more constant dependence proved harder to estimate at lower percentiles but were often still appropriately captured at the high percentiles, indicating the need to specify application-specific percentile ranges of interest, as the dependence in this range could help guide dependence modelling.

The second model is the MGPD, an asymptotic threshold-stable model with a constant dependence through the entire support of the data values. Despite its lack of flexibility, the model proved useful for pairs that displayed constant dependence. However, the model's rigidity prevented it from accurately capturing the decaying dependence at higher percentiles. The relative success of the MGPD was unexpected, mainly because nearly constant dependence structures such as the ones found between some pairs are not common in environmental applications (Huser and Wadsworth, 2022).

Extremal models are necessary to address applications concerning HM contamination in the soil. For the bivariate case, both sub-asymptotic (Huser and Wadsworth, 2022) and asymptotic models (Coles, 2001) may play a useful role depending on the dependence structure. Moreover, they are potentially helpful in identifying sources of contamination. Pairs that exhibit high and constant extremal dependence can be perceived as coming from the same source, as they have high concentrations at the same time, while pairs where asymptotic dependence is weak can be treated as not coming from the same contaminating source. However, the models have limitations. Neither one is particularly good

at capturing the whole range of the data. While the MGPD model seems to be better at capturing the lower quantiles, the EFC model usually experiences improvement at higher quantiles. Careful consideration must be given to the choice of threshold and the range for extremal inference when performing model selection.

Additionally, another notable limitation is that divorcing the data from their spatial dimension produces overall results lacking spatial variability - a characteristic especially useful in applications to HM contamination. The most organic extension of this work is to incorporate a spatial structure in the data. Modest work has been done on the subject, which showed the usefulness of the λ-madogram (Naveau *et al.*, 2009) and other spatial metrics of spatial dependence without Gaussian assumptions (Demangeot, 2020).

Further extensions of this work include increasing dimensionality. Once the spatial correlation is accounted for, it may be useful for the application to extend the model to include more than two elements; however, this must be done carefully. In soil, observed values are typically known as compositional and are measured in parts per million (ppm) - the relative number of "parts" of that element from a total of one million. Although this is not a perfect measurement, it introduces spurious correlation as (theoretically) the sum of parts should add up to a million. However, the use of compositional analysis (CoDa; Aitchison 1985) in spatial analysis has been controversial, as some consider that compositional values must be modelled together to account for the inherent correlation, i.e., if one component increases, the rest must decrease to maintain the sum to a million constraint (in ppm), known in spatial modelling as spurious spatial covariance (Pawlowsky-Glahn and Egozcue, 2016), although a few models do exist (Martins *et al.*, 2016). Others consider that a transformation, such as log-ratio transformation, is enough to bypass the sum constraint hurdle and use geostatistical models freely given the change of support (Clarotto *et al.*, 2022). Additionally, the definition of exceedances is more complex at $d > 2$ dimensions, as exceedance occur in one or various components simultaneously. Extensions of the model presented here to $d > 2$ or the spatial dimension would therefore have to address the concerns above to present a workable model of multivariate or spatial dependence for the application of HM soil contamination.

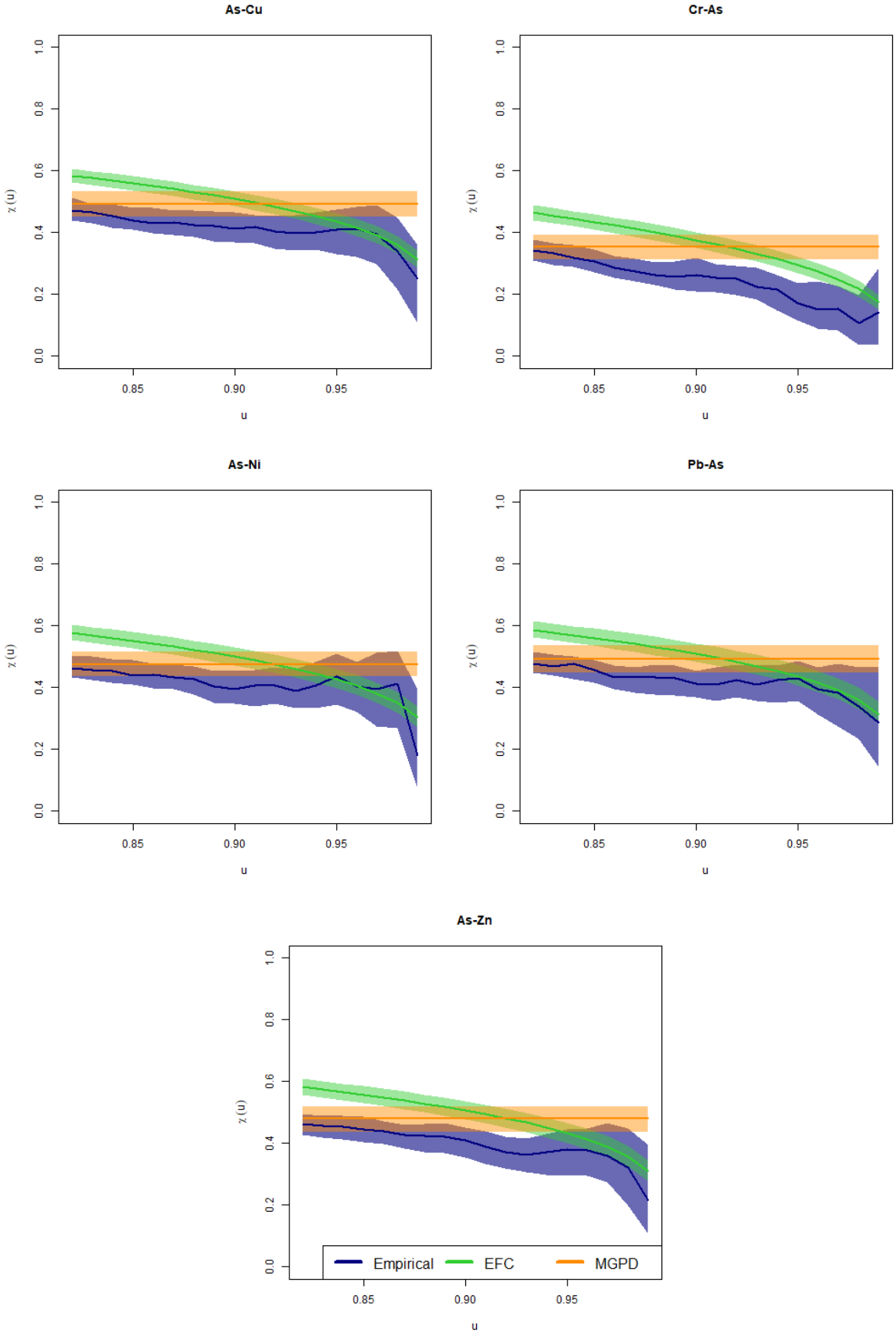

Figure 4.2: MGPD and EFC models for all 5 possible pairs of As and the other elements: Cr, Cu, Ni, Pb, and Zn. 95% confidence intervals are presented using point-wise bootstrap for 500 samples.

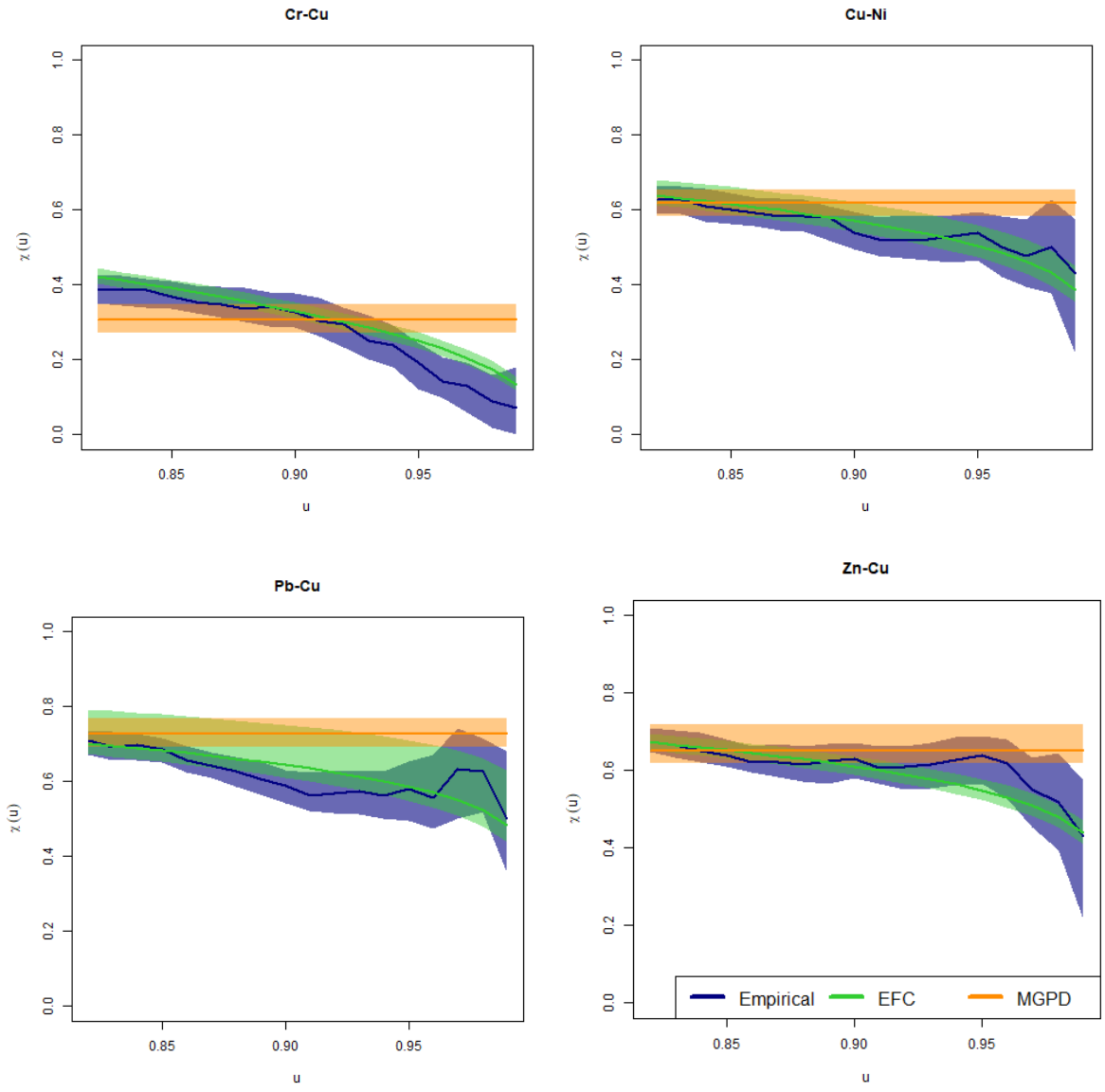

Figure 4.3: MGPD and EFC models for all 4 possible pairs of Cu and the elements, Cr, Ni, Pb, and Zn. 95% confidence intervals are presented using point-wise bootstrap for 500 samples.

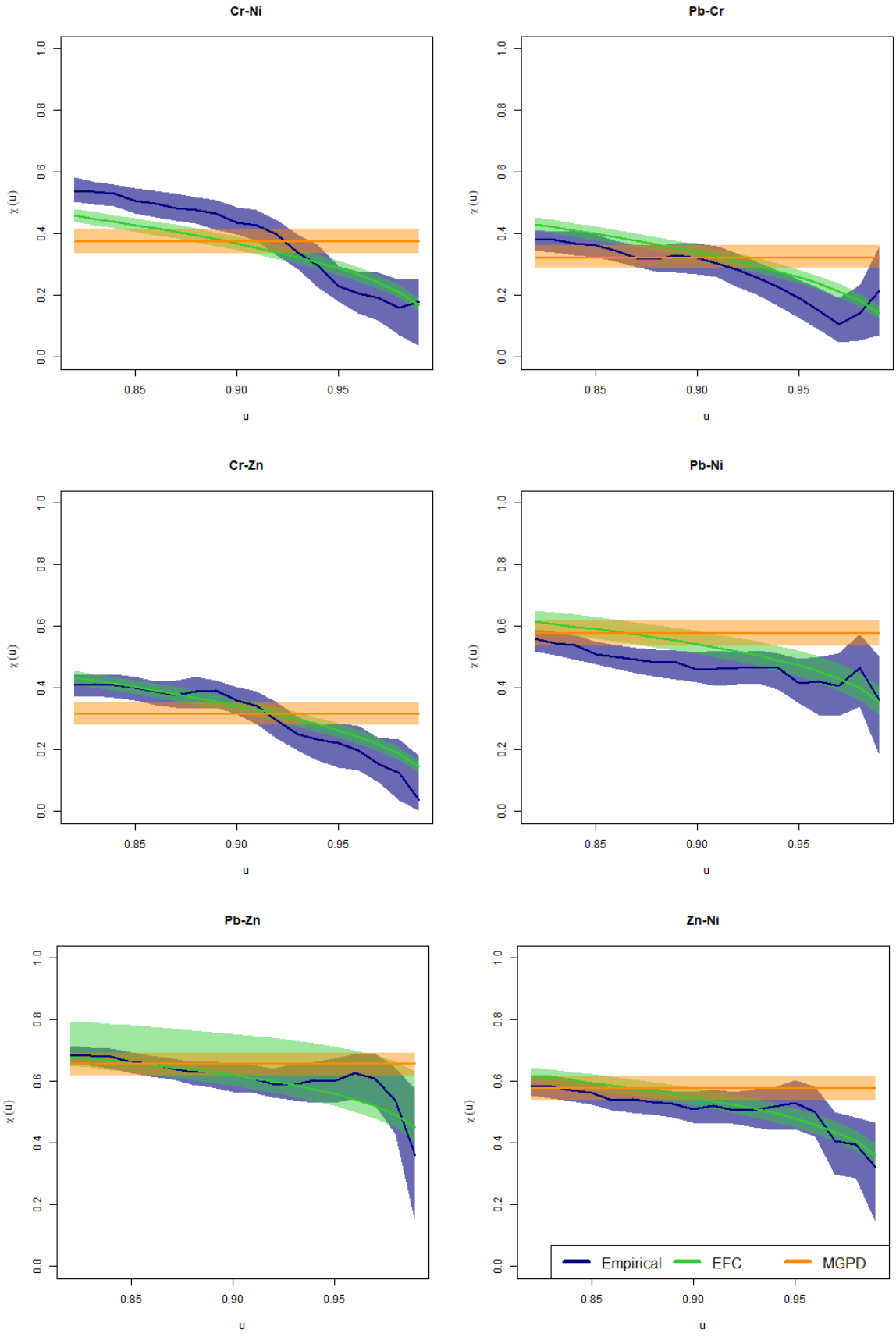

Figure 4.4: MGPD and EFC models for all 6 possible pairs of between Cr, Ni, Pb, and Zn. 95% confidence intervals are presented using point-wise bootstrap for 500 samples.

Chapter 5

Spatial Modelling of Heavy Metal Contamination

5.1 Heavy Metal Soil Contamination

Heavy metal soil contamination is the excessive accumulation of heavy metals (HMs) such as As, Cu, Cr, Ni, Pb, and Zn in the soil (Su *et al.*, 2014; Mishra *et al.*, 2019; Tang *et al.*, 2019). While trace amounts of HMs are indispensable in the environment, acute and chronic exposure to high concentrations can pose a significant risk to public health (Morais *et al.*, 2012). Therefore, modelling the spatial distribution of contaminant concentrations is of interest to policy makers and public health practitioners.

The main sources of HMs in urban and rural environments are often anthropogenic (Gómez-Sagasti *et al.*, 2012). In urban areas, primary sources include industrial waste and residue, chemical manufacturing, sewage, atmospheric deposition, and combustion of fossil fuels (Tang *et al.*, 2019; Kupka *et al.*, 2021). HM soil contamination is generally characterised by its wide spatial distribution, strong latency, irreversibility, and complex multivariate nature (Su *et al.*, 2014). The remediation of contaminated soils is relatively slow compared to the remediation of contaminated water or air (Kupka *et al.*, 2021). For this reason, understanding the extent and intensity of the contamination, particularly in urban and densely-populated areas, is essential to establish preventive public health measures and mitigate the impacts of soil HM contamination (Gómez-Sagasti *et al.*, 2012).

Spatial samples of HM concentrations come from soil surveys (Tóth *et al.*, 2016), which use spatial sampling schemes that prioritise the spatial distribution as dictated by each contaminant's properties and other soil properties (Khlifi and Hamza-Chaffai, 2010). The distribution of HM contaminants is known to be heavy-tailed (Marchant *et al.*, 2011), which is commonly addressed by performing a Box-Cox transformation, and modelling the transformed values using a geostatistical approach, also known as transgaussian kriging when using a kriging model (Diggle and Ribeiro, 2007; Lado *et al.*, 2008; Lv *et al.*, 2015).

However, the heavy-tail of the distribution persists even after transformation and is not captured using the Gaussian framework (Marchant *et al.*, 2011, 2010). If the contaminating concentrations (above-baseline concentrations of a contaminant) are in the tail of the distribution, then Gaussian models are not appropriate to assess risk in this context, as they are known to underestimate extreme values at the tails. Risk or exposure models that use these approaches can, therefore, provide misleading information to the public.

Extreme Value Theory (EVT), as described in detail in Section 2.2, is the natural statistical framework for the analysis of extreme values. Classical EVT distributions define extreme values as the maximum value inside a block (block-maxima) or values exceeding a given threshold (threshold exceedances). For the definition of the block-maxima approach, let $\{X_1, X_2, \cdots, X_n\}$ be a set of n iid continuous random variables and $M_n = \max\{X_1, \cdots, X_n\}$ represent the maxima. The extremal types theorem (Coles, 2001) states that if there exist sequences $\{a_n > 0\}$ and $\{b_n\}$ such that the normalisation of M_n as $M_n^* = \frac{M_n - b_n}{a_n}$ converges to a non-degenerate function G as $n \to \infty$, then G belongs to the family of generalised extreme value distributions (GEVDs) (Fisher and Tippett, 1928; Gnedenko, 1943; Mises, 1954). Threshold exceedances, also known as the peak-over-threshold approach (POT), represent an alternative to the block-maxima approach and are defined as observations of X that exceed a given threshold u, i.e., $X - u | X > u$. If the conditions for the limiting characterisation of M_n^* given above hold, threshold exceedances converge to the generalised Pareto distribution (GPD) when $u \to \infty$, with distribution function characterised by a scale $\sigma > 0$ and a shape $\xi \in \mathbb{R}$ parameter, and given by

$$H(y; \sigma, \xi) = 1 - \left(1 + \frac{\xi y}{\sigma}\right)^{-1/\xi}, \tag{5.1}$$

defined on $\{y : y > 0 \text{ and } 1 + \frac{\xi y}{\sigma} > 0\}$.

However, an essential requirement for the application of EVT is that the data contain replications in time at each location. In environmental applications, temporal replicates are generally available for phenomena with a temporal dimension. However, soil surveys are generally unreplicated, meaning replications in time are rarely available and preventing the use of EVT for applications such as mapping HM contamination. Additionally, the multivariate nature of HM contamination shows the presence of concomitant extremes, which may be of special interest for public health planning due to the complexity posed by the risks associated with multiple contaminants. It is clear then that a suitable spatial statistical modelling of HM contamination requires the use of multivariate spatial models that do not underestimate the marginal extreme behaviours, are able to capture extremal dependence between contaminants across space, and can handle unreplicated data.

The model proposed in this chapter addresses the gap between classical EVT based on replications and unreplicated multivariate spatial settings by using a continuous mixture of

non-extreme and extreme distributions, representing a novel statistical undertaking. The model is extended to the bivariate case using a coregionalisation framework, accounting for the extremal dependence between contaminants. We assume the distribution of each contaminant can be decomposed into two components, representing the body and the tail of the distribution. The body of the distribution is composed of a combination of natural pedogenic processes and diffuse background contamination and represents the majority of the observations (Ander *et al.*, 2013). The tail contains the extreme concentrations from contaminating anthropogenic or natural processes (Marchant *et al.*, 2010). Each component is modelled using suitable distributions for extreme and non-extreme concentrations and woven together using a continuous mixture model representation. In principle, the body and tail processes of two contaminants can be differently affected by the same spatial factors, so we allow the mixture components to share spatial effects across variables inside a coregionalisation framework. Inference on the model is performed using the integrated nested Laplace approximation (INLA; Rue *et al.* 2009), which allows Bayesian inference for the class of latent Gaussian models (LGMs) and can be fitted in `R` using the package `R-INLA`. Coregionalisation models are straightforward to fit with `R-INLA` (Krainski *et al.*, 2018); however, mixture models lack the LGM representation needed for INLA, so we adapt a conditional approach following Gomez-Rubio (2017). The model is assessed in its bivariate representation using a simulation study and implemented in a case study using data from the Geochemical Baseline Survey of the Environment (G-BASE) in the Glasgow Conurbation. Important byproducts of our modelling approach are risk maps showing probabilities of two contaminants jointly exceeding their respective safety values as defined by the soil guideline values (SGV; Cole and Jeffries 2009).

The remainder of this chapter proceeds as follows. The marginal mixture model is first introduced for the univariate case in Section 5.2.1. The bivariate extension, known as the coregionalised mixture model, is presented in Section 5.2.2. In Section 5.2.5, we detail the simulation study carried out to assess the performance of the bivariate model. Section 5.3 shows the results of the model applied to the case study of the G-BASE data in the Glasgow Conurbation focusing on the chromium-lead (Cr-Pb) pair. Finally, Section 5.4 provides a discussion of the methods presented.

5.2 Development of the Coregionalised Mixture Model

5.2.1 Univariate Body-Tail Mixture Model

At its most basic setting, a mixture model is a convex combination of K distributions to represent the different underlying groups in the data Y (McLachlan *et al.*, 2019), repre-

sented as

$$Y \sim \sum_{k=1}^{K} p_k f_k(y, \theta_k),$$

where $f_k(\cdot|\theta_k)_{k=1}^{K}$ is a set of parametric distributions, one for each latent group in the data, and $\mathbf{p} = (p_1, ..., p_K)$ are their associated weights defined with $\sum_{k=1}^{K} p_k = 1$.

Modelling marginal HM concentration distributions under this framework is suitable, as HM contaminant distributions can be assumed to come from two distinct processes - one for the baseline observations, also known as the body of the distribution, and another for the extreme concentrations (Lv *et al.*, 2015) - resulting in the body and tail portions of the distribution - if Y is a random variable defined in space and observed at locations $y(\boldsymbol{s})$ for $\boldsymbol{s} \in \mathcal{S} \subset \mathbb{R}^2$, a reasonable mixture model for HM contaminants can be defined as

$$y(\boldsymbol{s}) \sim p f_B(y(\boldsymbol{s})|\boldsymbol{\theta}_B) + (1-p) f_T(y(\boldsymbol{s})|\boldsymbol{\theta}_T), \quad \boldsymbol{s} \in \mathcal{S}, \tag{5.2}$$

where f_B is the distribution applicable to the body, f_T is the distribution of the tail, $\boldsymbol{\theta}_B$ and $\boldsymbol{\theta}_T$ are model parameters for f_B and f_T, respectively, and p is the mixing proportion, representing the probability that an observation will be in the body.

Although different distributions can be assigned to f_B and f_T, common approaches in the literature have used finite mixtures of Gaussian distributions where $2 < K < \infty$ distributions are used (Lin *et al.*, 2010; Zhu *et al.*, 2021). While this might be an appropriate approach for the observations in the body of the distribution, f_B, correctly characterising HM extremes motivates the use of a generalised Pareto distribution (GPD) for the tail f_T. However, the use of the GPD for f_T has been implausible given the GPD's reliance on temporal replications at each location. To circumvent this need for replications, we assume spatial stationarity of threshold exceedances and model transform the tail of each HM contaminant to a Gaussian distribution using a stationary GPD as in (5.1). To accomplish this, we first select an appropriate threshold u and define the threshold exceedances y_T as $y_T(\boldsymbol{s}) = y(\boldsymbol{s}) - u | y(\boldsymbol{s}) > u$. Then, we transform the tail values to a common Gaussian scale using the probability integral transform, giving rise to new tail values

$$y'_T(\boldsymbol{s}) = \Phi^{-1}\left(\hat{H}_T(y_T(\boldsymbol{s}; \hat{\sigma}, \hat{\xi})\right), \quad y_T(\boldsymbol{s}) > 0, \tag{5.3}$$

where $\hat{H}_T$ is the fitted GPD with scale $\hat{\sigma}$ and shape $\hat{\xi}$, Φ is the standard normal distribution, and $y_T(\boldsymbol{s})$ represents the threshold exceedances of y.

The final mixture model for marginal contaminants that incorporates (5.2) and (5.3) is

$$y(\boldsymbol{s}) \sim p f_B(y(\boldsymbol{s})|\boldsymbol{\theta}_B) + (1-p) f_T(y'(\boldsymbol{s})|\boldsymbol{\theta}_T), \quad \boldsymbol{s} \in \mathcal{S}. \tag{5.4}$$

where $f_B(\cdot)$ and $f'_T(\cdot)$ are Gaussian densities with parameter vectors $\boldsymbol{\theta}_B \equiv \boldsymbol{\theta}_B(s) =$

$(\eta_B(s), \tau_B)$ and $\boldsymbol{\theta}_T \equiv \boldsymbol{\theta}_T(s) = (\eta_T(s), \tau_T)$, respectively, $y'(\boldsymbol{s})$ the same as (5.3), and p is the mixture coefficient but can also be considered as $p = P(y(\boldsymbol{s}) > u)$. If the GPD fit is good, we should have $\eta_T(\boldsymbol{s}) = 0$ and $\tau(\boldsymbol{s}) = 1$. Still, in (5.4) we estimate $\boldsymbol{\theta}_T$ to accommodate for possible deviations of $y_T(\boldsymbol{s})$ from the GPD fit. Further flexibility is provided by allowing the Gaussian means, $\eta_B(\boldsymbol{s})$ and $\eta_T(\boldsymbol{s})$, to change with locations and other covariates, being referred to in the future as linear predictors.

5.2.2 Bivariate Extension: Coregionalised Mixture Model

The extension of the model in (5.4) to the bivariate setting was made to enable the assessment of the risk posed by two contaminants at the same location. Extending extreme value models to the bivariate setting, however, requires the consideration of extremal dependence between variables (Coles 2001, Ch. 8). Extremal dependence, along with other properties of the data such as unreplicated observations and the spatial dimension, motivated the use of the coregionalisation framework of Krainski *et al.* (2018) and was discussed in Chapter 4. This approach allows for multivariate response models with latent Gaussian characterisations to be modelled jointly using Bayesian inference by enabling different likelihood functions for each variable and modelling the dependence structure between variables by introducing shared components in the linear predictor. In the bivariate case, let $y_1(\boldsymbol{s})$ and $y_2(\boldsymbol{s})$ be two spatial variables for locations $\boldsymbol{s} \in \mathcal{S}$. Simple linear predictors for a in general coregionalisation model can be defined as

$$\begin{aligned} y_1(\boldsymbol{s}) &\sim N(\eta_1(\boldsymbol{s}), \tau_1), \\ y_2(\boldsymbol{s}) &\sim N(\eta_2(\boldsymbol{s}), \tau_2), \\ \eta_1(\boldsymbol{s}) &= \alpha_1 + z_1(\boldsymbol{s}) \\ \eta_2(\boldsymbol{s}) &= \alpha_2 + \lambda z_1(\boldsymbol{s}) + z_2(\boldsymbol{s}), \qquad \text{for } \boldsymbol{s} \in \mathcal{S}, \end{aligned} \tag{5.5}$$

where $\eta_1(\boldsymbol{s})$ and $\eta_2(\boldsymbol{s})$ are the linear predictors of y_1 and y_2 at $\boldsymbol{s}$, respectively, α_1 and α_2 are intercepts, $z_1(\boldsymbol{s})$ and $z_2(\boldsymbol{s})$ are spatial random effects (Lindgren *et al.*, 2011), and λ is a scaling coefficient for the shared spatial effect (Krainski *et al.*, 2018). Inside this framework, dependence is induced in the linear predictors of y_1 and y_2 through the shared component, $z_1(\boldsymbol{s})$. The construction of the linear predictors is otherwise flexible. Here, while y_1 has a single spatial effects term, y_2 can have a second spatial effects term z_2, to capture residual spatial dependence unique to y_2. This framework can be extended beyond the bivariate case, but the implementation is restricted due to the tradeoff between accuracy and computational costs. Additionally, the linear predictors can also contain linear and non-linear effects, as well as more shared components of different forms. Inference on the model can be readily performed using integrated nested Laplace approximations (INLA; Rue *et al.* 2009; Van Niekerk *et al.* 2023). For full details on Bayesian inference

and INLA, please see Section 2.3.2.

Combining (5.4) and (5.5) results in a bivariate spatial extreme mixture model for unreplicated data that combines bivariate mixture models for an accurate representation of the body and tail of the distribution of each contaminant using a Gaussian-GPD composition while a coregionalisation structure incorporates the spatial dependencies within a latent Gaussian model framework. Specifically, we construct two spatial mixture models with $K = 2$ mixing components for variables $y_1(\boldsymbol{s})$ and $y_2(\boldsymbol{s})$ at locations $\boldsymbol{s} \in \mathcal{S}$. The components account for the body and tail observations of each variable. The spatial mixture models are defined as

$$\begin{aligned} y_1(\boldsymbol{s}) &\sim p_1 f_{B_1}(y_1(\boldsymbol{s})|\boldsymbol{\theta}_{B_1}) + (1-p_1) f_{T_1}(y_1(\boldsymbol{s})|\boldsymbol{\theta}_{T_1}), \\ y_2(\boldsymbol{s}) &\sim p_2 f_{B_2}(y_2(\boldsymbol{s})|\boldsymbol{\theta}_{B_2}) + (1-p_2) f_{T_2}(y_2(\boldsymbol{s})|\boldsymbol{\theta}_{T_2}), \end{aligned} \tag{5.6}$$

where, for $i \in \{1, 2\}$, f_{B_i} is the density of the non-extreme observations in the body, f_{T_i} is the density of the extremes in the tail, $\boldsymbol{\theta}_{B_i} = (\eta_{B_i}, \tau_{B_i})$ and $\boldsymbol{\theta}_{T_i} = (\eta_{T_i}, \tau_{T_i})$ are the body and tail parameters, respectively, and p_i is the mixing proportion or the probability that an observation belongs to the body of the distribution.

For the mixture models in (5.6), we explored the inclusion of shared components on the body and tails, as in (5.5), to account for dependence at non-extreme and extreme values, respectively. Even though the shared components are flexible and can be tailored for each application, components should be linked only when necessary since increasing the number of shared components considerably increases computational costs. Our coregionalised mixture model shares spatial components only in the tails to account for extremal dependence, while the non-extreme components have shared dependence only through common covariates. Therefore, the body and tail linear predictors can be expressed as

$$\begin{aligned} \eta_{B_1}(\boldsymbol{s}) &= \alpha_{B_1} + z_{B_1}(\boldsymbol{s}) + \sum_{j \in \mathcal{J}} \beta_{B_{1j}} x_j(\boldsymbol{s}), & \eta_{T_1}(\boldsymbol{s}) &= \alpha_{T_1} + z_{T_1}(\boldsymbol{s}) + \sum_{j \in \mathcal{J}} \beta_{T_{1j}} x_j(\boldsymbol{s}), \\ \eta_{B_2}(\boldsymbol{s}) &= \alpha_{B_2} + z_{B_2}(\boldsymbol{s}) + \sum_{j \in \mathcal{J}} \beta_{B_{2j}} x_j(\boldsymbol{s}), & \eta_{T_2}(\boldsymbol{s}) &= \alpha_{T_2} + \lambda_T z_{T_1}(\boldsymbol{s}) + z_{T_2}(\boldsymbol{s}) + \sum_{j \in \mathcal{J}} \beta_{T_{2j}} x_j(\boldsymbol{s}), \end{aligned} \tag{5.7}$$

where for $i = (1, 2)$, z_{B_i} and z_{T_i} are random spatial effects, x_j are covariates, α_{B_i} and α_{T_i} are intercepts, $\beta_{B_{ij}}$ and $\beta_{T_{ij}}$ are coefficients corresponding to the covariates x_j for the body and tail respectively, and λ_T is a scaling coefficient for the shared random spatial effect z_{T_1}.

5.2.3 Inference for the Coregionalised Mixture Model

Inference for the model is based on the conditional latent Gaussian field framework proposed by Gomez-Rubio (2017), where we replace the Markov chain Monte Carlo (MCMC) inference with a simple conditional approach based on conditioning parameters *a priori*, similar to the importance sampling approach to conditional mixture models by Berild *et al.* (2022). In this model, only p_1 and p_2 are defined *a priori*, through a grid-search of possible values. The definition of p_1 and p_2 inform the classification of y as y_B or y_T, belonging to the body or tail of y, which in turn enables the identifiability of GPD in (5.3) in the unreplicated setting.

The default implementation of the GPD likelihood in INLA links the linear predictor to a fixed α-quantile of the distribution, similar to the relationship between the linear predictor and the mean parameter in the Gaussian case. This parametrisation implicitly assumes replicates at each $\boldsymbol{s} \in \mathcal{S}$. However, the transformation in (5.3) enables the model to be fitted following the usual geostatistical design of single replicates over space. The spatial random effects, z, are fitted using a stochastic partial differential equation approach (SPDE; Lindgren *et al.* 2011). Under these specifications, we used all non-informative standard Gaussian priors (mean 0 and unit standard deviation) for the intercept parameters $\boldsymbol{\alpha} = (\alpha_{B_i}, \alpha_{T_i})$ and regression coefficients $\boldsymbol{\beta} = (\beta_{B_{ij}}, \beta_{T_{ij}})$. The precision parameters, τ_{B_i} and τ_{T_i} where given penalised complexity (PC) priors (Simpson *et al.*, 2017). PC priors are defined as

$$P(\tau^{-1/2} < \nu_0^{-1/2}) = \alpha_\nu,$$

where τ is the precision, and ν_0 and α_ν are prior hyperparameters. To encourage identifiability, we chose to encourage small values of the standard deviation. Given the transformation in (5.3), we chose $\nu_0 = 2$ and $\alpha_\nu = 0.05$ for τ_{T_i}. For the precision parameters of the body distributions, τ_{B_i}, we chose hyperparameter values using empirical knowledge, so that $\nu_0 = \hat{\tau}_{B_i,y_i}/2$ and $\alpha_\nu = 0.05$, where $\hat{\tau}_{B_i,y_i}$ is the empirical precision; the PC prior therefore penalises precision smaller than half the empirical precision. In terms of the standard deviation, the PC prior penalises standard deviation values that are larger than twice the empirical estimated standard deviation of y_i. The choice of these hyperparameters is robust, and no discernible change is obtained with different hyperparameter values. A simulation study is performed in Section 5.2.5 to assess the model's performance.

The model was fitted in `R` using the `R-INLA` package where the priors and the coregionalisation framework has been previously implemented. The model is computationally expensive, fitting a dataset of approximately 3000 bivariate observations in 3 hours in a machine with 4 CPU cores at 2.21 GHz and 16GB RAM. The code is freely accessible in github (`https://github.com/danicuba-stats/BivariateExtremeMix`).

5.2.4 Bivariate Risk Assessment

Probability maps visualising the probability of a joint exceedance in both contaminants are calculated as a byproduct of the model and constitute a useful tool for risk assessment. Specifically, the maps show $\Pr(y_1(s_i) > u_1 | y_2(s_i) > u_2)$, where u_1 and u_2 are possible safety guidance values for y_1 and y_2, respectively, provided by regional safety regulations. The probability of joint exceedance at any location is obtained by sampling from the posterior predictive distribution, which can be done using a multi-step Monte Carlo method summarised in Figure 5.1. However, it is first necessary to sample from the posteriors of the linear predictor and the hyperparameters (denoted as in Figure 5.1). These samples are then used to obtain samples of the posterior predictive distribution of each component after a back-transformation of the tail distribution and mixing of the distributions according to the mixture proportions p_1 and p_2 as in (5.6). The joint exceedance probabilities are then computed empirically by counting the number of times the samples exceed u_1 and u_2 at the same time and dividing it by the total number of samples collected. This process is repeated 1000 to obtain measures of uncertainty of the Monte Carlo procedure.

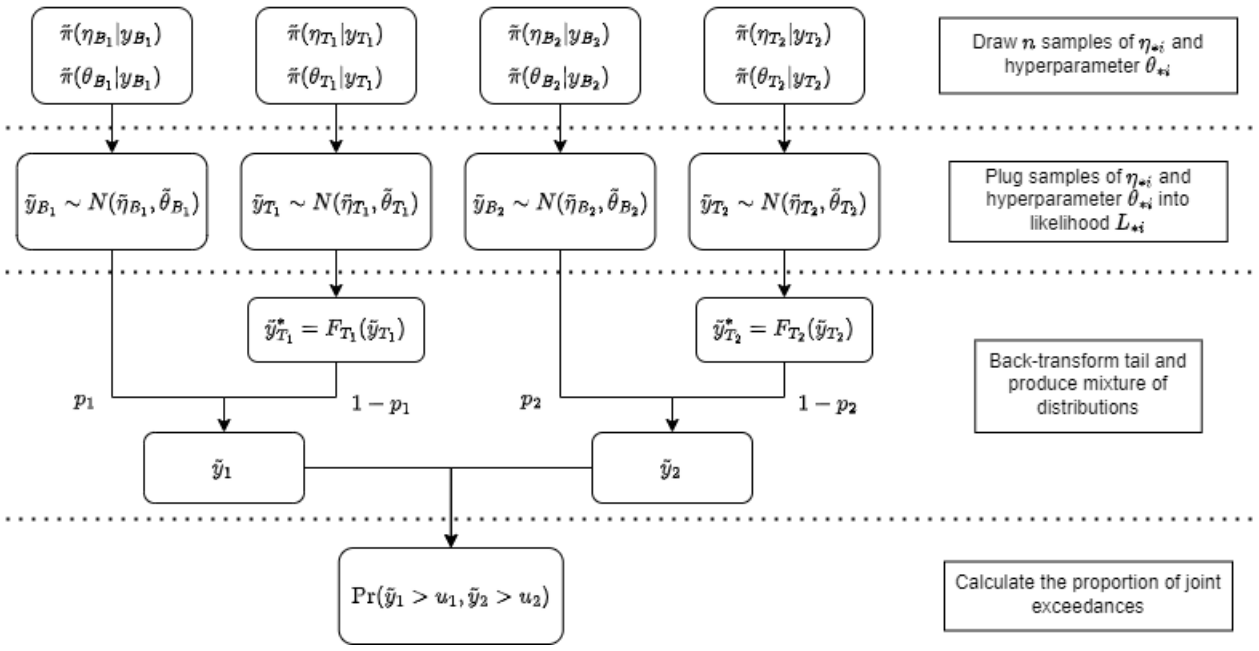

Figure 5.1: Flow chart of the Monte Carlo method used to sample from posterior predictive distributions of y_1 and y_2 and obtain a map of the probability of joint exceedance of threshold u_1 and u_2 respectively. $\tilde{\pi}(\cdot|\cdot)$ denote posterior distributions of the corresponding parameters and hyperparameters. The process is repeated 1000 times.

5.2.5 Simulation Study: Investigating the Performance of the Coregionalised Mixture Model

An extensive simulation study was performed to assess the performance of the bivariate coregionalised mixture model. We show the construction and results of four different simulation scenarios that mimic real-life HM soil concentrations.

Specifications of the Data-Generating Process

The data were simulated directly from the model in (5.6) and (5.7), taking into account the adjustment proposed in (5.3). A total of $N = 1000$ simulations of $n = 1000$ observations were generated over the region $\mathcal{S} = [0, 100]^2$ for two response variables, y_1 and y_2. The random spatial effects are simulated as Gaussian Processes (GP) with Matern covariance function (Matérn, 1960) defined as

$$C_\nu(d) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu}\frac{d}{\rho}\right)^\nu \mathrm{K}_\nu \left(\sqrt{2\nu}\frac{d}{\rho}\right),$$

where d is the Euclidean distance between two observations, Γ is the gamma function, K_ν is the modified Bessel function of the second kind, ρ is the range of spatial dependence, and $\nu = 1$ is the smoothness parameter. Parameter values for the proposed simulation scenarios are given in Table 5.1.

Simulation scenarios A and B consider heavy-tail distributions with different levels of extremal spatial dependence through the weight of the shared spatial effect, λ, as shown in (5.7). Scenario A has $\lambda = 0.25$ and Scenario B as $\lambda = 0.9$, corresponding to weak and strong extremal dependence between components, respectively. Each scenario is further subdivided into two variations representing different mixture proportions: Variation 1 at $p = 0.75$ and variation 2 at $p = 0.9$. This gives rise to four scenarios, A1, A2, B1 and B2. The data are simulated using two covariates, x_j for $j = \{1, 2\}$, with standard uniform distributions. The values for the remaining parameters are given in Table 5.1.

Table 5.1: Parameter values for the bivariate scenarios A and B, representing small and large spatial extremal dependence through the weight of the shared spatial effect λ. Each scenario is further subdivided into variations 1 and 2, resulting in four scenarios: A1, A2, B1, and B2. Variations 1 and 2 represent two mixture proportions, $p = 0.75$ and $p = 0.9$, respectively.

Parameter	**A**	**B**
Variation 1 : (p_1, p_2)	(0.75, 0.75)	(0.75, 0.75)
Variation 2 : (p_1, p_2)	(0.9, 0.9)	(0.9, 0.9)
$(\alpha_{B_1}, \alpha_{T_1})$	(1,0)	(1,0)
$(\alpha_{B_2}, \alpha_{T_2})$	(1,0)	(1,0)
$(\beta_{B_{1j}}, \beta_{T_{1j}})$	(0.1,0.25)	(0.1,0.25)
$(\beta_{B_{2j}}, \beta_{T_{2j}})$	(0.1,0.25)	(0.1,0.25)
λ	0.25	0.9
(ρ_1, ρ_2)	(10, 15)	(10, 15)
$(\sigma_{T_1}, \sigma_{T_2})$	(1,1)	(1,1)
(ξ_1, ξ_2)	(0.05, 0.25)	(0.5, 0.25)

Classification of Body and Tail

The model requires *a priori* classification of observations as belonging to the body or tail of the distribution, as mentioned in Section 5.2.3, which is considered a preprocessing step to enable model fitting. While the choice of mixture proportion p_i also defines threshold u_i, given that the proportion of observations exceeding threshold u_i is the same as p_i, setting all exceedances from u_i as belonging to the tail results in an upper truncation for the body, where $i = \{1, 2\}$. As a result, we propose a classification based on the Metropolis-Hastings algorithm, which results in a soft boundary between body and tail. The classification is as follows.

1. First, fit stationary proposal distributions for the body and the tail. For the body, f_B as a Gaussian distribution fitted using $f_B(y(\boldsymbol{s})) \equiv \mathrm{N}(\eta, \tau^{-1/2})$, where η is the mean parameter and τ is the precision. For the tail, fit a stationary GPD distribution using $f_T(y(\boldsymbol{s})) \equiv \mathrm{GPD}(u_i, \sigma, \xi)$, where u_i is the threshold value chosen *a priori*, σ is the scale parameter, and ξ is the shape parameter. These distributions are fitted using maximum likelihood estimation to speed up the process.

2. For each observation $s_m = s_1, ..., s_n$, compute the density of $y(s_m)$ under f_B and under f_T, denoted as $p_B(y(s_j))$ and $p_T(y(s_j))$, respectively.

3. Obtain the classification ratio for the m-th observation to evaluate its membership to the body or tail using $p_\alpha = \min\left\{1, \frac{p_T(y(s_j))}{p_B(y(s_j))}\right\}$.

4. Draw a random sample from a uniform distribution $u_\alpha \sim \mathrm{U}(0, 1)$.

5. Assign the observation $y(s_j)$ as belonging to the tail if $p_\alpha \geq u_\alpha$.

6. Repeat the process $n_c = 100$ times for each observation and assign the membership that appears the most number of times in the n_c samples.

Results of Simulation Scenario A1

Results of the simulations are assessed using Q-Q plots, root mean square error (RMSE), and true-parameter 95% coverage probability. Although only results for simulation A1 are shown in this section, results for the remaining scenarios, A2, B1, and B2 are shown in Appendix A. A discussion for all scenarios is provided in this section.

Figure 5.2 shows the Q-Q plots of the results for simulation A1. The figure shows that the model performs better for y_1 (variable 1), displaying smaller variability at higher values than y_2 (variable 2). However, the mean and median of the data are well captured. Table 5.2 shows the coverage probability for the parameters of the linear predictor, $\boldsymbol{\alpha} = (\alpha_{B_i}, \alpha_{T_i})$ and $\boldsymbol{\beta} = (\beta_{B_{ij}}, \beta_{T_{ij}})$. We can see that they are well captured and have coverage probabilities close to 0.95, which is optimal, with the exception of α_{T_1}, which might account for the abrupt behaviour around the transition between body and tail. Of the remaining parameters, only λ has lower coverage probabilities, indicating the fit is not as good in y_2 as in y_1, as is expected given the asymmetric construction of the model.

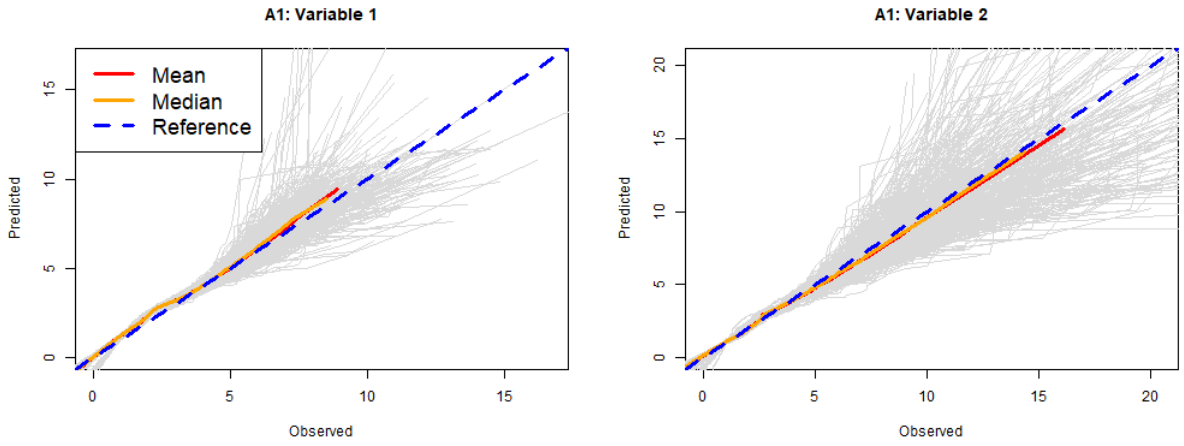

Figure 5.2: Q-Q plots of all simulations (grey) of bivariate scenario A1, for variables 1 and 2. The mean and median of the simulations are shown in red and orange, respectively, while the reference line is given in blue.

Table 5.2: Summary of results of A1. The table shows the parameter's true value; estimated parameter mean, median and standard deviation, 95% coverage probability, and the mean RMSE.

Parameter	True Val	Mean	Median	Sd	Coverage Pr	RMSE	MAE
α_{B_1}	1.00	1.07	1.06	0.07	0.95	0.10	0.08
α_{T_1}	0.00	0.05	0.05	0.09	0.74	0.10	0.08
α_{B_2}	1.00	1.16	1.16	0.07	0.99	0.18	0.16
α_{T_2}	0.00	-0.00	0.01	0.09	0.99	0.09	0.07
$\beta_{B_{11}}$	0.10	0.05	0.05	0.04	0.95	0.07	0.05
$\beta_{B_{12}}$	0.25	0.15	0.14	0.02	0.99	0.10	0.10
$\beta_{T_{11}}$	0.10	0.02	0.04	0.05	0.95	0.10	0.08
$\beta_{T_{12}}$	0.25	0.16	0.15	0.04	0.86	0.09	0.09
$\beta_{A_{21}}$	0.10	0.05	0.05	0.03	0.99	0.06	0.05
$\beta_{A_{22}}$	0.25	0.15	0.15	0.03	0.99	0.10	0.10
$\beta_{T_{21}}$	0.10	0.03	0.04	0.04	0.99	0.08	0.07
$\beta_{T_{22}}$	0.25	0.18	0.18	0.03	0.87	0.17	0.17
ρ_{B_1}	5.00	4.85	3.43	2.39	0.99	2.33	2.08
ρ_{T_1}	10.00	10.23	10.20	0.16	0.99	0.28	0.23
ρ_{B_2}	5.00	5.89	4.97	4.10	0.95	4.09	2.67
ρ_{T_2}	15.00	21.77	19.36	6.97	0.79	9.58	6.77
λ	0.75	1.05	1.01	0.37	0.95	0.46	0.36
ξ_1	0.05	0.10	0.10	0.09	0.95	0.10	0.08
ξ_2	0.25	0.24	0.25	0.14	0.89	0.16	0.13

The results for A2 show a pattern similar to that of A1. The Q-Q plots in Figure A.1 in Appendix A show a larger variability displayed in the tail of y_2 than in y_1. Additionally, y_2 is slightly underestimated at extreme values. Table A.1 summarises the parameter estimates. Once again, the linear coefficients are well captured by the model while λ is consistently overestimated, similar to A1.

The performance assessments of the classification of observations as body or tail for A1 and A2 are given in Table A.2 in the Appendix. Carried out *a priori* with the method in Section 5.2.5, the classification of both is approximately $\sim 90\%$ of the observations begin classified correctly.

Scenario B had a larger value for the weight of the shared spatial component, $\lambda = 0.9$, indicating stronger extremal dependence between variables. Figure A.2 in the Appendix shows that for variable 1, the model performs as expected, similarly to results in A1 and A2. Variable 2 follows the pattern seen in A1 and A2, where variability is increased in the tail. However, in this scenario, the mean and median show a slight overestimation at the extremes. Table A.3 in the Appendix shows the model has good coverage probabilities for most parameters, except α_{T_1}. The large standard deviation of the range parameters, ρ_1 and ρ_2, show the model struggles to calculate the range of the spatial components, a possible explanation for the increased variability in the tails of the distribution.

The results for B2 are shown in Figure A.3 in the Appendix. The model performs similarly to B1, with the mean showing sensitivity to large values and a well captured median for both variables. The summary of the estimated model parameters in Table A.4 show that the model correctly estimates most parameters, with the exception of ρ_1 and ρ_2, which experience an even bigger variability in estimation than previous scenarios. The coverage probability of the ranges, ρ_1 and ρ_2 seems to be lower for Variation 2, meaning the model struggles to accurately estimate the range when there are fewer extreme observations.

The classification of the observations for B1 and B2 (Table A.5) is similar to scenario A. The relatively lower specificity values indicate a poorer categorisation of the tail observations than the body. However, the results indicate the model did not suffer a loss of power with a larger imbalance in classes. Overall, the simulations show that the model proposed in this chapter performs well under various scenarios that mimic real soil data.

5.3 Application to Cr-Pb Soil Contamination in the Glasgow Conurbation

5.3.1 Summary of Data

In this case study, we apply the bivariate coregionalised mixture model to jointly model concentrations of chromium and lead. Data for this application are from the G-BASE survey (Johnson *et al.*, 2005) performed by the British Geological Survey, and consist of 2750 topsoil observations in the Glasgow Conurbation. The original scale of the data is parts per million, but the data is preprocessed using a log transformation for improved properties. Even though these data are discussed at length in Section 3.2, a brief summary is provided in this section. Figure 5.3 shows histograms of Cr and Pb after the log transformation, showing the heavy-tailed nature of contaminant distributions, with Pb enjoying better symmetry and Cr displaying a heavier tail.

The maps in Figure 5.4 provide an assessment of the spatial patterns of these two contaminants. In the top row of the figure, the full range of concentrations of Cr and Pb after a log transformation, respectively, is given. The maps display a continuous scale of observations up to the 95th quantile. Observations beyond this quantile are censored and shown in orange. In this scale, the 95th quantile corresponds to 5.198 log(ppm) for Cr and 6.095 log(ppm) for Pb. The maxima, shown in red, are 8.582 log(ppm) and 9.204 log(ppm) for Cr and Pb respectively. The map shows an agglomeration of high values just south of the Clyde in central Glasgow, while other high values can also be found in Coatbridge, East Kilbride, and Wishaw to the south and southeast. To the southwest, high values are seen around Paisley and further south towards Clyde Muirshiel Regional

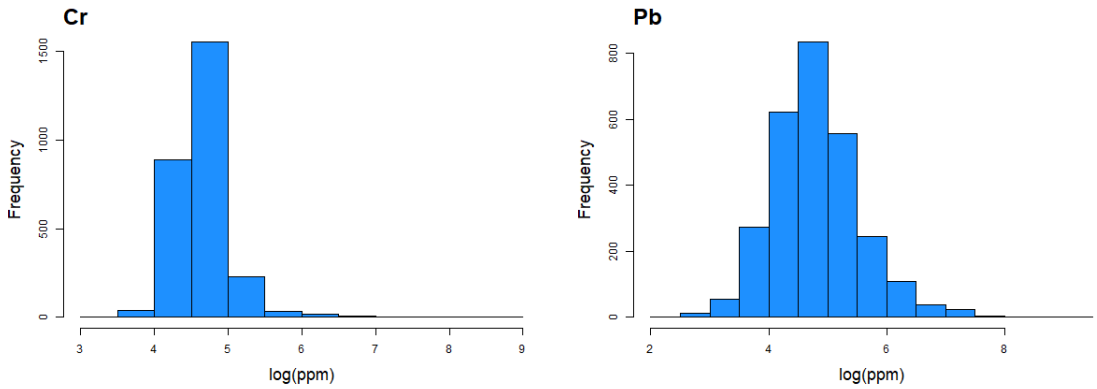

Figure 5.3: Histogram of Cr (left) and Pb (right) in log(ppm) scale.

Park. High values for Pb are found near Greenock Port and Dumbarton, especially along the major motorways (M74 and A82) towards The Trossachs National Park and the port of Greenock.

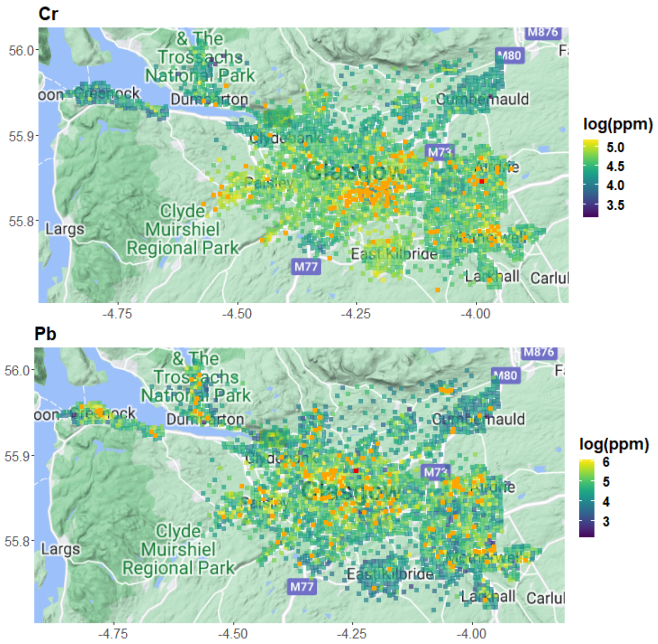

Figure 5.4: Censored map of log concentrations of Cr (top) and Pb (bottom) where observations above the 95th percentile are orange (5.198 for Cr and 6.095 for Pb), and the maximum observation is marked in red (8.582 for Cr and 9.204 for Pb).

The specification of the model in (5.7) allows for the inclusion of covariate information $x_j(\boldsymbol{s})$ as linear fixed effects to inform the linear predictors. Following Johnson *et al.* (2017), we obtained terrain and topography variables for model covariates, including elevation, slope, aspect, multiresolution index of valley bottom flatness (MRVBF), the complementary multiresolution index of the ridge top flatness (MRRTF), and topographic wetness index (TWI) from a digital elevation model (DEM) by the Ordinance Survey

(OS; https://www.ordnancesurvey.co.uk/products/os-terrain-50) at a resolution of 1:50000, roughly equivalent to a 1km × 1km grid. Processing of the data was done in R using RSAGA (Brenning, 2008). Variables providing proximal traffic information, such as the type of nearest road, primary (A) or secondary (B), and distance to the nearest primary and secondary roads, were obtained Open Roads data set of the OS (https://www.ordnancesurvey.co.uk/products/os-open-roads).

5.3.2 Assessing the Body-Tail Classification

The *a priori* classification of observations as body or tail depends on the parameters p_1 and p_2, corresponding to the mixture proportions of y_1 and y_2 respectively, which indicate the probability of an observation belonging to the body of the distribution. Unlike the simulation study, p_1 and p_2 are not known in real-world applications and are computationally prohibitive to estimate using the methods available when this analysis were carried out (such as INLA within MCMC); therefore, an exhaustive search for appropriate values is performed using a grid of values from 0.75 to 0.99 in increments of 0.01, and the values for p_1 and p_2 that yield the best DIC and predictive RMSE values are selected. In the case of the Cr-Pb pair, the best model performance was given by $p_1 = 0.98$ and $p_2 = 0.95$. A prediction using a spatial binomial model is made over a rectangular region representing the spatial extent of the observations to obtain the classification of prediction locations (Figure 5.5). The binomial model captures the pattern of extreme observations in Figure 5.4, where extreme concentrations of Cr are clustered south of the Clyde in Glasgow and Wishaw to the south east. The predictions for Pb also capture the true pattern, with extreme or tail observations found along the M74, the A82, and Coatbridge.

5.3.3 Results of Model Validation

Our bivariate spatial extreme mixture model in (5.6) and (5.7) is fitted to the rectangular area in Figure 5.5 representing the spatial extent of the observations to obtain a continuous prediction of the concentrations of Cr and Pb in the river basin. Model validation is performed using k-fold cross-validation, with $k = 20$ where every fold removes 5% of the observations and performs predictions on the removed locations. Figure 5.6 compares the cross-validation predictions of the coregionalised mixture model to a non-mixture Gaussian model fitted in INLA where both components are fitted independently of each other and provides smoothed 95% credible intervals. The larger width of the credible intervals of the coregionalised mixture model is expected due to the heavier tail of the GPD distribution compared to the Gaussian model and the limited number of observations at higher threhsolds. For Cr, we see the deviation is greater, which is understandable given the heavier tail of the Cr distribution as corroborated by its estimated kurtosis. On the

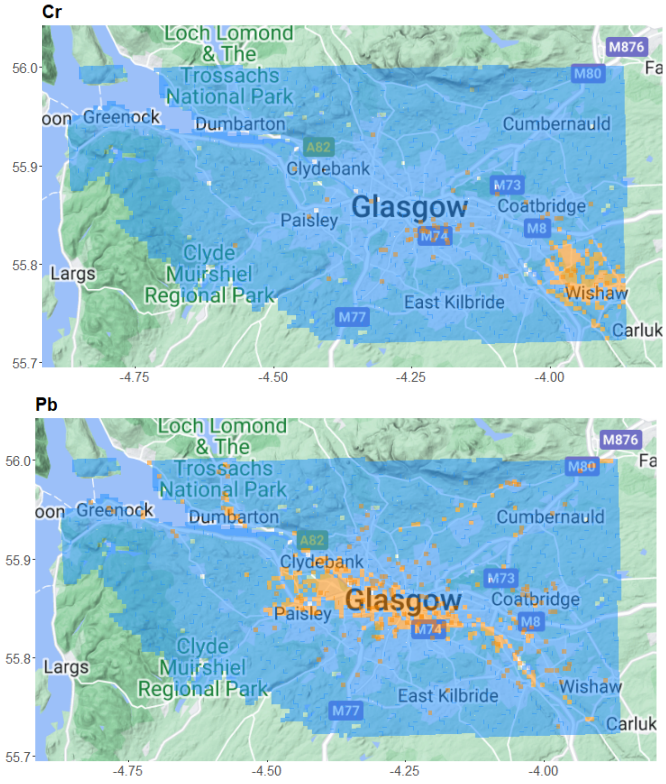

Figure 5.5: Classification map of observations for Cr (top) and Pb (bottom) as body (blue) or tail (orange) using the mixture proportions $p_1 = 0.98$ and $p_2 = 0.95$ respectively.

other hand, the difference between models is less distinct for Pb and can be explained due to lower kurtosis. Overall, the coregionalised mixture model shows an improvement over the Gaussian model in capturing extreme values and modelling heavy-tailed distributions.

Figure 5.7 provides maps of the estimated marginal concentrations. The figures show that Cr has new predicted areas of contamination in the Wishaw area to the southeast. The area between the city centre and East Kilbride to the south experiences higher concentrations too, which matches the observed data. Other areas of raised predicted Cr concentrations are west of Paisley (to the west of the City of Glasgow), and Coatbrige to the east. Pb shows similar trends to those anticipated. The M74 road around the city centre and to the south through Wishaw are singled out as having especially high concentrations, as does the Port of Greenock and the A82 towards the Trossachs National Park. Additional predicted areas of contamination include the M80 near Falkirk to the southeast. Overall, higher concentrations can be found along more densely populated areas from Paisley to the west to Coatbridge in the east and around major A and M roads with heavy traffic.

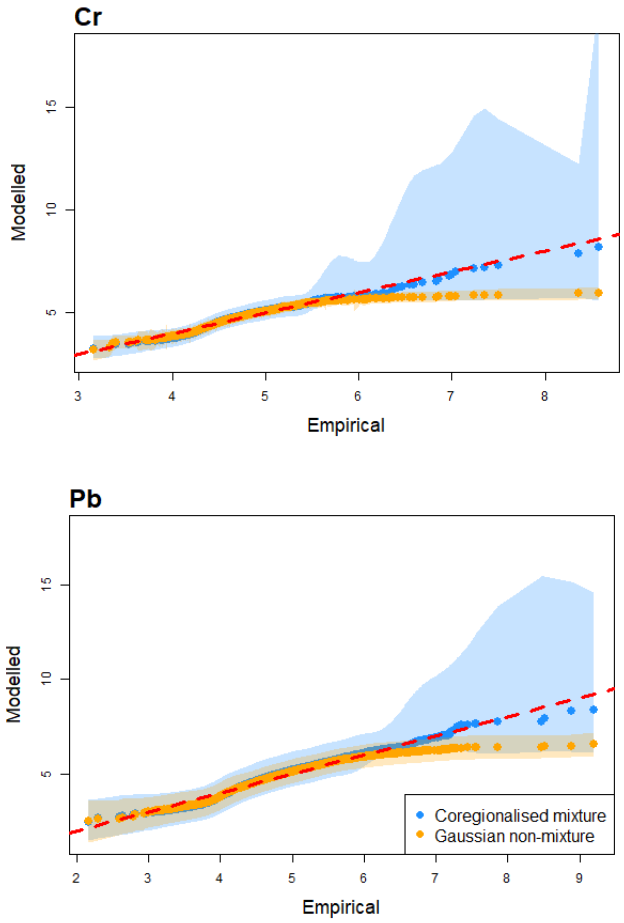

Figure 5.6: Q-Q plots for the coregionalised mixture model (blue) with smoothed 95% credible intervals and the non-mixture independent Gaussian models (orange) predictions of the log concentrations of Cr (top) and Pb (bottom). Mixture proportions are $p_1 = 0.98$ and $p_2 = 0.95$.

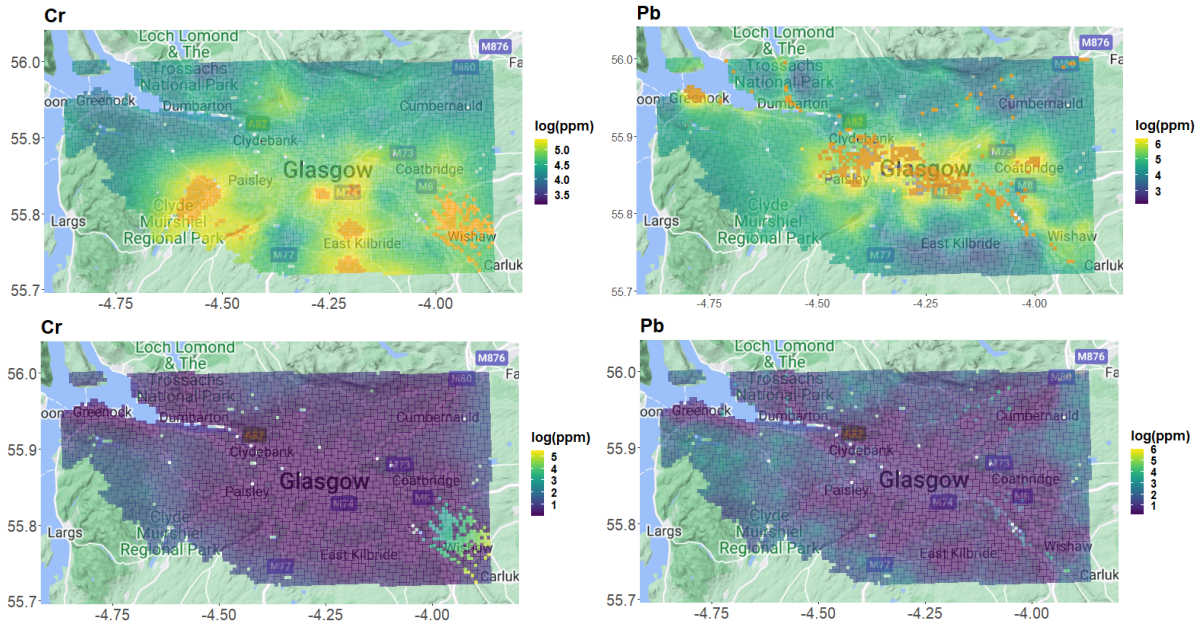

Figure 5.7: *Top*: maps of the results of the coregionalised mixture model with observations exceeding the 95th quantile in colour orange (same values as in Figure 5.4) for Cr (left) and Pb (right). *Bottom:* width of the 95% credible intervals for the coregionalised mixture model for both contaminants.

5.3.4 Risk Assessment of Cr-Pb Joint Exceedance

Assessing the risk posed by both contaminants simultaneously is possible using joint exceedance probability maps. The Environment Agency in the UK published soil guidance values (SGVs) for most HM elements (Cole and Jeffries, 2009), indicating what concentration thresholds are considered safe. For a full list of SGVs and more information on the policy behind HM contamination, please see Section 3.1.5. In residential areas with plants or food production (SGV1), Cr is recommended to stay under 130ppm or 4.87 in log(ppm), and Pb under 200ppm or 5.30 log(ppm). In residential areas without plant or food production (SGV2), the SGV is 200ppm for Cr and 310ppm for Pb, or 5.29 log(ppm) and 5.74 log(ppm) for Cr and Pb, respectively. Using the process described in Figure 5.1, we computed the probabilities of joint exceedance of SGV1 and SGV2 using samples from the posterior predictive distribution. The pointwise estimates in Figure 5.8 show that Cr and Pb have high probabilities of exceeding SGV1 in the south, southeast, and east of the city of Glasgow. These areas are well-known for legacy contamination due to historical chromium ore processing and other chromium-producing industries (CL:AIRE, 2007). The map of the width of the 95% credible interval shows that uncertainties given by the confidence intervals are small and do not affect interpretation. The coefficient of variation of these estimated probabilities, defined as the ratio of the standard deviation to the mean, of the first column displayed at the bottom of Figure 5.8 shows that areas of high probability exhibit small relative variation, whereas the areas of low probability,

mainly to the north of the city, show larger variation in relation to the mean.

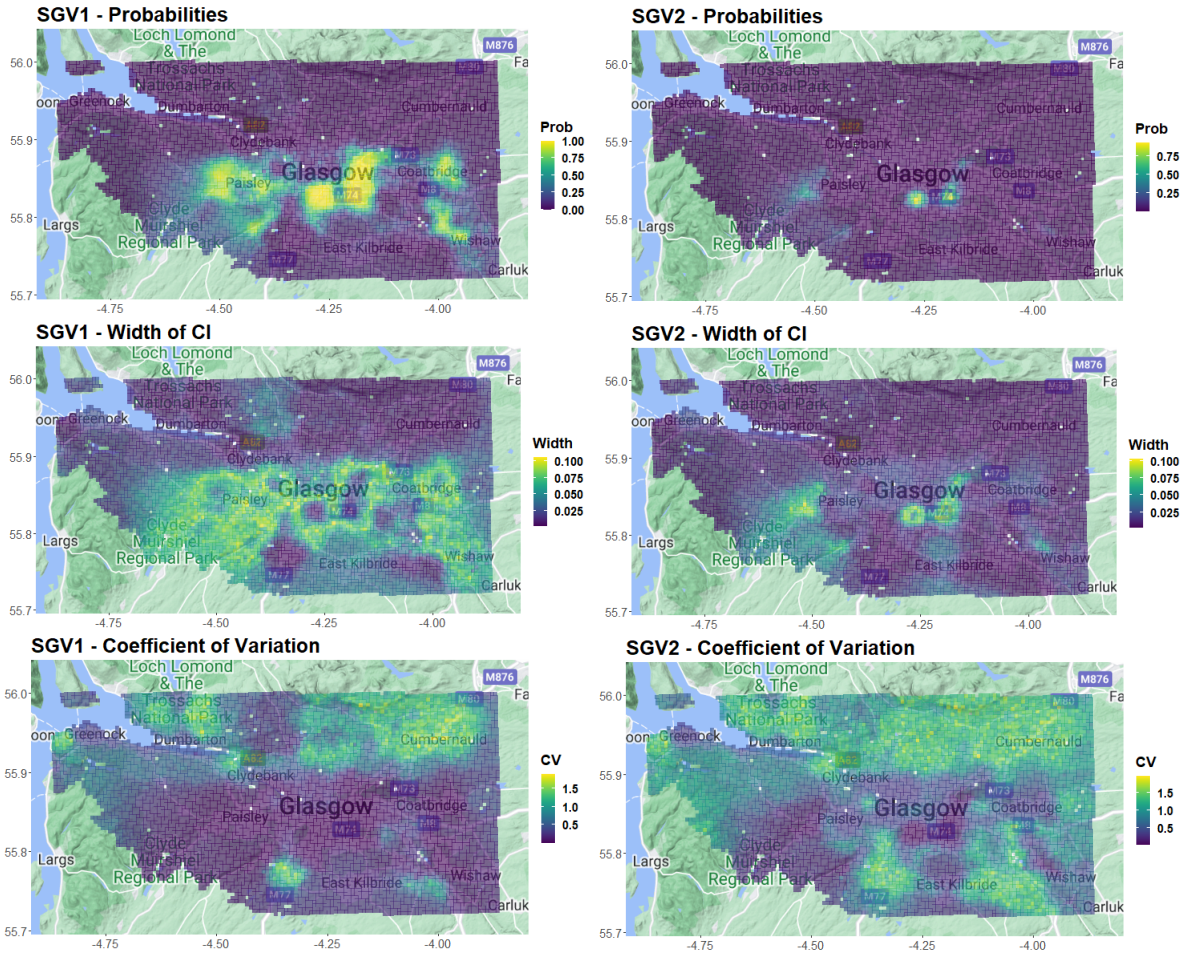

Figure 5.8: *Top:* Probability maps for joint exceedances SGV1 (left) and SGV2 (right). *Middle:* Width of 95% confidence intervals of the exceedance probabilities as described in Section 5.2.2 for SGV1 and SGV2, respectively. *Bottom:* Maps of the coefficient of variation for the estimated exceedance probabilities.

The maps for SGV2 (second column in Figure 5.8) show a similar result where there is only a high probability of both contaminants exceeding the threshold in two contamination hotspots in well-known areas to the south and southeast of the city centre. Additionally, there are higher probabilities to the southwest, near Paisley. Uncertainty maps show the two hotspots of contamination with high certainty, while contamination detected to the southwest is more uncertain. Areas of low probability, such as the north, show greater variability in relation to the mean, similar to SGV1.

5.4 Discussion and Future Work

Maps of geochemical concentrations are generally produced using geostatistical models under a Gaussian framework. Such models do not account for the various processes re-

sponsible for the spatial distribution of geochemical concentrations but rather model all observations as realisations of a single Gaussian process, resulting in an underestimation of extreme values and measures of risk . When more than one contaminant is present, classical geostatistical models not only under-estimate extreme concentrations, but also do not account for the extremal dependence between contaminants. We propose to partition the distribution of each contaminant into body and tail, representing non-extreme and extreme concentrations, and weaving them together inside a coregionalised mixture model framework. The body component, representing non-extreme observations linked to a combination of natural pedogenic processes and diffuse background contamination, is modelled using a Gaussian model common to geochemical applications. The tail, containing extreme observations related to contaminating anthropogenic or natural processes, is modelled using an adapted extreme value distribution. While EVT is an attractive framework to capture extremal behaviour, it requires replications at each location, which are not commonly available in geochemical datasets. For this reason, a transformation is applied to the tail under a stationary GPD, and later modelled under a Gaussian likelihood following the usual geostatistical setting of a single replication. Our proposed framework combines ideas from coregionalisation models, mixture models, spatial latent Gaussian models, and EVT to capture non-extremal concentrations as well as the extremal behaviour and dependence between two contaminants at high concentrations. It produces continuous maps of estimated marginal concentrations as well as risk maps in the form of joint exceedance probabilities.

We fit our model to Cr and Pb concentrations in the Glasgow Conurbation in the west of Scotland. The data was first transformed using a log transformation, and kurtosis provided evidence of the deviation of the tails from a Gaussian distribution. The results of the model show that there are areas of Cr contamination to the south and southwest of Glasgow, southwest of Paisley, to the south around East Kilbride, and to the east near Wishaw. Areas of high Pb concentrations are found around the Glasgow city centre and seem to be contained to areas around the Clyde River and the Port of Greenock. Marginal credible intervals are the widest for observations belonging to the tail, which is expected for extreme observations due to the limited sample size. The model shows that joint contamination has a high probability of exceeding SGV1 and SGV2 in the south and southeast of the city of Glasgow, areas known to be affected by legacy industrial HM contamination. A comparison with a classical Gaussian model shows that the joint modelling of two contaminants using the coregionalised mixture model is an improvement, particularly at the tail. Furthermore, the shared spatial random component z_{T_1} models the factors affecting the extreme values of both contaminants while the weight of the shared spatial component, λ, models the dependence between both tail distributions. High values of λ can account for strong extremal dependence between components while $\lambda = 0$ indicates

extremal independence. For this application, we constrained the dependence structure to only account for extremal dependence through a shared spatial random effect. However, the framework is flexible and can be easily extended to capture dependence through other terms in both the body and tail or extended beyond the bivariate case. Future work can be developed to jointly model p_1 and p_2 in space and integrate this modelling with the coregionalised mixture model. A natural way to do this is through a hierarchical Bayesian model where inference is carried out using simulation-based approaches, such as MCMC.

Chapter 6

Data Fusion for Extremes

6.1 Data Fusion for Extremes in Air Quality Monitoring

While invisible to the naked eye, airborne solid and liquid particles are ubiquitous and responsible for human and environmental health. In good air quality conditions, airborne particles play an essential role in the hydrological cycle, atmospheric circulation, and the necessary existence of greenhouse and trace gases (Pöschl, 2005). Air quality degradation due to air pollution can have significant adverse effects on the environment and cause chronic and acute damage to human health, effectively decreasing quality of life and increasing mortality rates (H R Anderson, 2009).

Air pollution refers to hazardous gaseous chemicals or airborne particles in the environment and can have both natural and anthropogenic sources. In nature, some physical occurrences, such as volcanic activity or forest fires, can release large amounts of hazardous pollutants into the air. Many more anthropogenic sources of air pollution exist, however, including industrial facilities, fossil fuel combustion for energy and transportation, and excessive use of fertilisers (Kampa and Castanas, 2008) among others covered in detail in Section 3.3.

The risk posed by air pollution can be determined using physical primary parameters of the polluting particles: concentration, size, structure, and chemical composition (Pöschl, 2005). The most common categories are gaseous pollutants, persistent organic pollutants, heavy metals, and particulate matter. While all air pollutants lower air quality, the scope of this work is concerned with particulate matter (PM). PM is a general term for a mixture of particles suspended in breathing air varying in size and composition, often categorised by size as smaller than 2.5 µm ($PM_{2.5}$) or 10 µm (PM_{10}).

$PM_{2.5}$ and PM_{10} have harmful effects on the human body due to their abundance in urban settings and the wide variety of sizes and compositions. This type of pollution

poses a significant risk to public health and has been linked to increased levels of mortality and premature death (Kyung and Jeong, 2020). It increases the occurrence of and exacerbates cardiovascular and cerebrovascular diseases through mechanisms of systemic inflammation, direct and indirect coagulation activation, and translocation of systemic circulation (Anderson *et al.*, 2012). It can also have harmful effects on the respiratory system. While coarser particles, such as PM_{10}, are deposited in the upper respiratory tracts, the finer particles, such as $PM_{2.5}$, can reach the lung alveoli and result in chronic obstructive pulmonary disease (COPD) (Don D. Sin *et al.*, 2023), bronchial asthma and lung cancer, among other chronic respiratory conditions (Kampa and Castanas, 2008; Don D. Sin *et al.*, 2023; Kyung and Jeong, 2020).

Another significant negative effect of PM pollution on human health is that it can include all other common air pollutants, including heavy metals, organic compounds, biological matter, particle carbon core, and reactive gases known to be harmful (see Section 3.3). Chronic and acute exposure to PM pollution occurs through inhalation and ingestion through deposition on food and water. However, the most significant impacts on public health are caused by episodes of heavy and extremely heavy levels of pollution in the air. Zhang *et al.* (2021) showed that heavy and extremely heavy $PM_{2.5}$ pollution events substantially increased hospital admissions for cardiovascular disease. Similarly, extreme events of PM pollution are also linked to the exacerbation of asthmatic symptoms in children and adults (Anderson *et al.*, 2012). These events of extreme PM pollution are therefore targeted in policy at global and regional scales. In 2021, the World Health Organization (WHO) published new air quality guidelines (AQG) for air pollution recommending short-term exposure to not exceed a 24-hour average of $15\,\mu m^3$ and $45\,\mu m^3$ for more than three days a year for $PM_{2.5}$ and PM_{10}, respectively (WHO, 2021). The UK also has guidelines stipulating that $PM_{2.5}$ concentrations should not exceed $20\,\mu m^3$ as a yearly average (more in Section 3.3.4).

Meeting the suggested AQGs, however, is not a trivial task from the policy or technical points of view. Air quality management requires efficient coordination in the government apparatus, affecting economic, legal, public health, and political spheres. Compliance with suggested standards, such as WHO (2006), was commonly achieved through wide catch-all policies that focused on a single pollutant at the time and targeted well-known local and national sources of pollution (Martenies *et al.*, 2015). As air quality has improved, simple policies for comprehensively reducing pollutant emissions became more difficult to design and implement. Given the adverse health effects on populations living and working in areas of low air quality, an inevitable requirement for effective policy is its ability to target more localised sources of pollution in both space and time. Therefore, identifying "hotspots" of air pollution, or areas that more commonly experience extreme episodes of PM pollution, is a priority for decision-makers.

The accurate spatial and temporal identification of these extreme episodes is limited by data availability (Martenies *et al.*, 2015). For example, PM concentration data in the UK are available through the Automatic Urban and Rural Network (AURN), run by the Department for Environment, Food and Rural Affairs (DEFRA). While the network is extensive, comprising 171 working stations collecting data for various periods since 1972, the spatial coverage is insufficient to inform targeted local policies required for the effective improvement of air quality. The temporal coverage of the stations is not uniform and can contain missing observations. The interpolation of measurements taken by in-situ equipment (such as AURN observation stations), in both space and time, is necessary on a fine scale to accurately capture extreme local pollution events, but frequently exhibit gaps in coverage. Alternative sources can provide data with improved properties, such as comprehensive spatial coverage and complete historical long-term records, but are typically unable to capture local nuances that are seen in data from in-situ observation stations. The Copernicus Atmosphere Monitoring Service (CAMS) is one such source of alternative data and is run by the European Centre for Medium-Range Weather Forecasts (ECMWF), representing a global reanalysis of atmospheric composition with global coverage at daily and sub-daily temporal scales (for details see Section 3.4). However, remote-sensing sources are known to underestimate extreme values (Palharini *et al.*, 2020; Ståhl *et al.*, 2024), requiring a pre-processing step for improved representation of extreme events.

Integrating different data streams with often different spatial and temporal characteristics is described as data fusion. It results in an improved representation of the phenomenon that combines the desired properties from each source. Kriging methods are commonly used for data fusion in the context of air quality monitoring (Ferreira *et al.*, 2000; Künzli *et al.*, 2005; Beauchamp *et al.*, 2017, 2018; Xie *et al.*, 2017). They usually present a reliable interpolation of the mean values but tend to smooth extreme values due to their intrinsic Gaussian assumption (Gressent *et al.*, 2020), which in many cases leads to an underestimation of extreme events. Many alternatives to kriging have been proposed recently for data fusion; see, e.g. Fuentes and Raftery (2005); Bogaert and Fasbender (2007); Banerjee *et al.* (2015); Gengler and Bogaert (2016); Wilkie *et al.* (2019); Villejo *et al.* (2023). However, all of these techniques are developed under a Gaussian framework and, therefore, share the same limitations with kriging in capturing extreme events.

Alternatives for the specific purpose of fusing extreme values exist under a myriad of frameworks, such as quantile regression neural networks or automated regression-based statistical downscaling (Bürger *et al.*, 2012), mixture of Gaussian distributions (Ebtehaj and Foufoula-Georgiou, 2010), or conditioning model parameters on remote-sensing observations (Hundecha and Bárdossy, 2008). However, as flexible as these approaches may be, they lack the theoretical justification behind extreme values provided by extreme value

theory (EVT). Furthermore, they are not capable of retaining temporal information on the occurrence of an extreme value, thus providing less specific information about threshold exceedances.

The research presented in this chapter proposes a model that extends the hierarchical spatiotemporal data fusion model of Wilkie *et al.* (2019) in the context of extremes by using a generalised Pareto likelihood in a Bayesian framework to target threshold exceedances and maintain their temporal structure by the introduction of a zero-inflated modelling adjustment to the generalised Pareto distribution known as the Dirac-delta generalised Pareto distribution. The model performs data fusion by linking data from observation networks and remote-sensing sources through the scale parameter of nonstationary GPDs. Similar ideas have been explored in Healy *et al.* (2023), but contrary to their approach, our method uses flexible spatiotemporal Bayesian hierarchical structures to capture extremal dependence. In this way, we use both datasets to provide an improved representation of the threshold exceedances over space and time. The result is a dataset, complete in space and time, that appropriately calibrates modelled observations to accurately capture local threshold exceedances as informed by in-situ measurements over a limited number of locations. The work in this chapter is structured as follows. Section 6.2 provides a detailed description of the method proposed for this application. Section 6.3 applies the proposed method to the case study of air quality monitoring in the Greater London region and provides a description of the results and a comparison to alternative modelling approaches. Finally, Section 6.4 provides a conclusion and discussion on the results.

6.2 Development of Data Fusion for Extremes Model

6.2.1 The Non-Parametric Data Fusion Model of Wilkie *et al.* (2019)

Wilkie *et al.* (2015) proposed a non-parametric downscaling model to fuse data with different spatial supports inside a Bayesian hierarchical framework. Specifically, they fused remote-sensed data of log(chlorophyll-a) concentrations from the European Space Agency's ENVISAT satellite at Lake Balaton, Hungary, to data taken *in-situ* at limited locations inside the lake. The approach was motivated by the model originally proposed by Gelfand *et al.* (2003) and later developed by Berrocal *et al.* (2010), where an underlying (true) process is not assumed, but rather, assumed to be sampled using in-situ measuring equipment; these in-situ measurements are then linked to the remote-sensing data via a linear regression model with spatially-varying coefficients. To describe Berrocal *et al.* (2010), let $Y(\boldsymbol{s})$ for locations $\boldsymbol{s} = \{\boldsymbol{s}_1, ..., \boldsymbol{s}_n\}$ where $\boldsymbol{s}_i \in \mathbb{R}^2$ represent the data taken in-situ, and $x(\boldsymbol{B})$ be observations from remote sensing sources for the grid cells $B_i = \{B_1, ..., B_n\}$ that

correspond to each of the observations as the nearest grid centroid by geodesic distance. Defining a location $\boldsymbol{s}_i$ as existing inside the grid cell B_i allows the model to be defined as

$$Y(\boldsymbol{s}_i) = \alpha(\boldsymbol{s}_i) + \beta x(B_i) + \epsilon(\boldsymbol{s}_i), \quad \epsilon(\boldsymbol{s}_i) \sim N(0, \sigma^2), \tag{6.1}$$

with

$$\begin{aligned}\alpha(\boldsymbol{s}_i) &= \alpha + \tilde{\alpha}(\boldsymbol{s}_i),\\ \beta(\boldsymbol{s}_i) &= \beta + \tilde{\beta}(\boldsymbol{s}_i),\end{aligned}$$

where ϵ represents a Gaussian error term with variance σ^2. In this model, α and β are the additive and multiplicative bias of the remote-sensing data, while $\tilde{\alpha}(\boldsymbol{s}_i)$ and $\tilde{\beta}(\boldsymbol{s}_i)$ represent local adjustments. They are spatially varying, and defined as a bivariate zero-mean Gaussian distribution, as in Schmidt and Gelfand (2003). Working in the Bayesian framework, $\tilde{\boldsymbol{\alpha}} = (\tilde{\alpha}(\boldsymbol{s}_1), ..., \tilde{\alpha}(\boldsymbol{s}_n))$ and $\tilde{\boldsymbol{\beta}} = (\tilde{\beta}(\boldsymbol{s}_1), ..., \tilde{\beta}(\boldsymbol{s}_n))$ are given the priors

$$\tilde{\boldsymbol{\alpha}} \sim N(\mathbf{0}, \sigma_\alpha^2 \exp(-\phi_\alpha \Sigma_{\text{data}})), \quad \text{and} \quad \tilde{\boldsymbol{\beta}} \sim N(\mathbf{1}, \sigma_\beta^2 \exp(-\phi_\beta \Sigma_{\text{data}})), \tag{6.2}$$

where Σ_{data} is an $n \times n$ matrix of Euclidean distances between in-situ data locations defined as $\Sigma_{\text{data}} = |y_j - y_k|$ for $j = 1, ..., n$ and $k = 1, ..., n$, σ_α^2 and σ_β^2 are spatial variances, and ϕ_α and ϕ_β are spatial decay parameters, equivalent to those in an exponential covariance function.

Wilkie *et al.* (2019) used only the spatially-varying coefficients idea of (6.1), and extended it to the spatiotemporal case using smooth curves fitted using basis functions of some dimension d (Ramsay and Silverman, 2006), further enabling the temporal interpolation at each location. The model is defined as

$$\begin{aligned}
\mathbf{y}_i \mid \mathbf{c}_i, \sigma_y^2 &\sim \mathrm{N}_{q_i}\left(\boldsymbol{\Phi}_i \mathbf{c}_i, \sigma_y^2 \mathbf{I}_{q_i}\right),\\
\left(\sigma_y^2\right)^{-1} &\sim \mathrm{Ga}\left(a_y, b_y\right),\\
c_{ij} \mid \alpha_{ij}, \beta_{ij}, d_{ij}, \sigma_c^2 &\sim \mathrm{N}\left(\alpha_{ij} + \beta_{ij} d_{ij}, \sigma_c^2\right),\\
\boldsymbol{\alpha}_j \mid \sigma_\alpha^2 &\sim \mathrm{N}_n\left(\mathbf{0}, \sigma_\alpha^2 \exp\left(-\phi_\alpha \boldsymbol{\Sigma}_{\text{data}}\right)\right),\\
\boldsymbol{\beta}_j \mid \sigma_\beta^2 &\sim \mathrm{N}_n\left(\mathbf{1}, \sigma_\beta^2 \exp\left(-\phi_\beta \boldsymbol{\Sigma}_{\text{data}}\right)\right),\\
\left(\sigma_\alpha^2\right)^{-1} &\sim \mathrm{Ga}\left(a_\alpha, b_\alpha\right),\\
\left(\sigma_\beta^2\right)^{-1} &\sim \mathrm{Ga}\left(a_\beta, b_\beta\right),\\
\left(\sigma_c^2\right)^{-1} &\sim \mathrm{Ga}\left(a_c, b_c\right),\\
\mathbf{x}_i \mid \mathbf{d}_i, \sigma_x^2 &\sim \mathrm{N}_{l_i}\left(\Psi_i \mathbf{d}_i, \sigma_x^2 \mathbf{I}_{p_i}\right),\\
\left(\sigma_x^2\right)^{-1} &\sim \mathrm{Ga}\left(a_x, b_x\right),\\
\mathbf{d}_i &\sim \mathrm{N}_m\left(\boldsymbol{\mu}_d, \Sigma_d\right),
\end{aligned} \tag{6.3}$$

where

- $i = 1, ..., n$, represents point locations of in-situ data.
- $\mathbf{y}_i = (y_{i1}, ..., y_{iq_i})^T$ represents in-situ measurements at point locations i for times 1 to q_i.
- $\mathbf{x}_i = (x_{i1}, ..., x_{il_i})^T$ represents the data taken from the centroid for grid-cell that covers the point location i for times 1 to l_i.
- $\mathbf{\Phi}$ is a $q_i \times m$ matrix of basis functions that fits a smooth through the time series to $\mathbf{y}_i$ at location i.
- $\mathbf{\Psi}$ is a $l_i \times m$ matrix of basis functions that fits a smooth through the time series to $\mathbf{x}_i$ at location i.
- Σ_{data} is an $n \times n$ matrix of distances between the n in-situ locations.
- ϕ_α and ϕ_β are the spatial decay parameters.
- $\boldsymbol{\alpha}_j$ and $\boldsymbol{\beta}_j$ are the additive and multiplicative bias between the mean parameter of x and y across time, as defined by the j-th basis function of Φ and Ψ, respectively. The two parameters have a similar interpretation to those in (6.2).
- $\boldsymbol{c}_i = (c_{i1}, ..., c_{iq_i})^T$ and $\boldsymbol{d}_i = (d_{i1}, ..., d_{il_i})^T$ are the coefficients of matrix of basis functions Φ and Ψ for the i-th location.
- $a_y, b_y, a_\alpha, b_\alpha, a_\beta, b_\beta, a_c, b_c, a_x, b_x, \mu_d$ and Σ_d are hyperparameters chosen a priori. Wilkie *et al.* (2019) sets $a_y, a_\alpha, a_\beta, a_c, a_x$ to 2 and $b_y, b_\alpha, b_\beta, b_c, b_x$ equal to 1, meaning precision priors for all parameters were set to $\text{Ga}(2, 1)$. The hyperparameters of $\boldsymbol{d}_i$ are set as $\boldsymbol{\mu}_d = \mathbf{0}$ and $\Sigma_d = \{1, 2, ..., m\} \times \boldsymbol{I}_m$, which reflects the lack of prior knowledge of the behaviour of $\boldsymbol{d}_i$.

This model defines a linear relationship between the mean of remote sensing and in situ observations, $\Psi_j \mathbf{d}_i$ and $\Phi_i \mathbf{c}_i$, respectively. Consequently, the model fuses the mean of the in-situ and the remote-sensing data at each location, smoothing extreme high and low observations. While this suits Wilkie *et al.* (2019)'s application on water quality monitoring in Lake Balaton, it may not be suitable for applications where extremes are the target.

6.2.2 Data Fusion for Extremes (ExDF)

The model we propose is inspired by (6.3) and targets extreme values defined as exceedances of a threshold. We here define extremes as exceedances over a large threshold,

for which the generalised Pareto distribution (GPD) is a suitable approximation; see Section 2.2 for more details. Let

$$\begin{aligned} \mathbf{y}_i^* &= \{\mathbf{y}_i - y_{iu} \mid \mathbf{y}_i > y_{iu}\} \\ \mathbf{x}_i^* &= \{\mathbf{x}_i - x_{iu} \mid \mathbf{x}_i > x_{iu}\}, \end{aligned} \tag{6.4}$$

where y_{iu} and x_{iu} represent the threshold at the u-th percentile of $\mathbf{y}_i$ and $\mathbf{x}_i$, respectively, and u is close enough to 1 so that the GPD is a suitable approximation of the behaviour of y_i^* and x_i^*. When non-threshold exceedances are removed, a naive adaptation of (6.3) for threshold exceedances would be to replace the Gaussian likelihood of $\mathbf{y}_i$ and $\mathbf{x}_i$ by the GPD likelihood, as

$$\begin{aligned} \mathbf{y}_i^* \mid \boldsymbol{c}_i &\sim \text{GPD}\left(\exp(\Phi_i \boldsymbol{c}_i), \xi\right), \\ c_{ij} \mid \alpha_{ij}, \beta_{ij}, d_{ij}, \sigma_c^2 &\sim \text{N}\left(\alpha_{ij} + \beta_{ij} d_{ij}, \sigma_c^2\right), \\ \boldsymbol{\alpha}_j \mid \sigma_\alpha^2 &\sim \text{N}_n\left(\mathbf{0}, \sigma_\alpha^2 \exp\left(-\phi_\alpha \Sigma_{\text{data}}\right)\right), \\ \boldsymbol{\beta}_j \mid \sigma_\beta^2 &\sim \text{N}_n\left(\mathbf{1}, \sigma_\beta^2 \exp\left(-\phi_\beta \Sigma_{\text{data}}\right)\right), \\ \left(\sigma_\alpha^2\right)^{-1} &\sim \text{Ga}\left(a_\alpha, b_\alpha\right), \\ \left(\sigma_\beta^2\right)^{-1} &\sim \text{Ga}\left(a_\beta, b_\beta\right), \\ \left(\sigma_c^2\right)^{-1} &\sim \text{Ga}\left(a_c, b_c\right), \\ \mathbf{x}_i^* \mid \boldsymbol{d}_i &\sim \text{GPD}\left(\exp(\Psi_i \boldsymbol{d}_i), \xi\right), \\ \boldsymbol{d}_i &\sim \text{N}_m\left(\boldsymbol{\mu}_d, \Sigma_d\right), \end{aligned} \tag{6.5}$$

where $\exp(\boldsymbol{\Phi}_i \boldsymbol{c}_i)$ represent the scale parameters for the q_i time points available for $\mathbf{y}_i^*$, $\exp(\boldsymbol{\Psi}_i \boldsymbol{d}_i)$ are the scale parameters for the p_i available time points at $\mathbf{x}_i^*$, and ξ is a fixed shape parameter shared between $\mathbf{y}_i^*$ and $\mathbf{x}_i^*$. Keeping ξ constant is not a direct conversion from the previous model, but it is often done in practice to ensure identifiability (Youngman, 2019), reduce the computation burden of inference, and reduce uncertainty. This approached was considered restrictive and one of the drawbacks of this version of the model.

This model has multiple shortcomings. First, it only permits the fusion of threshold exceedances under the unstated assumption that non-threshold exceedances are not observed. In this way, non-threshold exceedances do not contribute to the likelihood but are missing altogether. Furthermore, the missing observations (non-threshold exceedances) are therefore interpolated using the GPD likelihood, meaning the model assumes all observations in the time period are, or should be, threshold exceedances. This is inaccurate since non-threshold exceedances were removed artificially during the data processing step in (6.4). Second, the apriori fixing of the shape parameter, ξ, introduces uncertainty and bias into the model, as it would have to be estimated in a previous step using a separate mechanism and the model would have to be appropriately defined to propagate the

uncertainty of this estimation.

To address the first concern, we proposed a censored approach where we censor all non-exceeding observations at 0; so we define $\mathbf{y}_i^*$ and $\mathbf{x}_i^*$ as

$$\begin{aligned} \mathbf{y}_i^* &= \begin{cases} 0, & \text{for } y_i \leq y_{iu} \\ y_i - y_{iu} & \text{for } y_i > y_{iu} \end{cases} \\ \mathbf{x}_i^* &= \begin{cases} 0, & \text{for } x_i \leq x_{iu} \\ x_i - x_{iu} & \text{for } x_i > x_{iu}. \end{cases} \end{aligned} \tag{6.6}$$

If $u > 0.5$, the resulting $\mathbf{y}_i^*$ and $\mathbf{x}_i^*$ will have a majority of zeroes. We propose accommodating these values similarly to Weglarczyk *et al.* (2005) and Couturier and Victoria-Feser (2010), who used a zero-inflated GPD mixture model called the Dirac-delta generalised Pareto distribution (δ-GPD). Weglarczyk *et al.* (2005) used this model and a few similar variants to perform frequency analysis of hydrologic data in arid and semi-arid regions, where rain is infrequent, resulting in a truncation at zero. Couturier and Victoria-Feser (2010) used the same δ-GPD to model radio audience data, which is full of true zeros but also values under the limit of detection that are censored at zero. The δ-GPD is defined as

$$f(y|p,\sigma,\xi) = (1-p)\delta(y) + \frac{p}{\sigma}\left(1 + \frac{\xi y}{\sigma}\right)^{-1/\xi - 1} \Delta_0(y), \tag{6.7}$$

where σ and ξ are the scale and shape parameters of the GPD, respectively, $p \in [0,1]$ is the probability of a threshold exceedance, $\delta(y)$ is the Dirac delta function with density only at $y = 0$, and $\Delta_0(y)$ is the unit step function, equalling 1 when $y > 0$. The δ-GPD is a good fit for the problem for two reasons. First, under the δ-GPD values of 0 contribute $1 - p$ to the likelihood and, therefore, are accounted for in the model. Second, simulating from this model is straightforward and can generate zeroes, which simulation from the GPD cannot do, allowing us to model the spatiotemporal locations of non-exceedances and sizes of exceedances.

Couturier and Victoria-Feser (2010) suggest extending the model in (6.7) to include covariates in the parameter p through a GLM framework. Specifically, for time t at location y_i, they propose

$$p_{y_i t} = \nu^{-1}(\mathbf{Z}_i[t,\cdot]\boldsymbol{\lambda}_i) = \frac{\exp(\mathbf{Z_i}[\mathbf{t},\cdot]\boldsymbol{\lambda}_i)}{1 + \exp(\mathbf{Z}_i[t,\cdot]\boldsymbol{\lambda}_i)}, \tag{6.8}$$

where $\mathbf{Z}_i$ is a matrix of covariates and $[t,\cdot]$ denotes only the t-th row, $\boldsymbol{\lambda}_i$ is a vector of coefficients, and ν^{-1} is the inverse of the *logit* link function. We investigated the use of (6.8) using different configurations for $\mathbf{Z}_i$ and found that, for every time point $t = 1, ..., q_i$,

in $\mathbf{y}_i^*$, the best performance is obtained incorporating the previous, present, and future information in $\mathbf{x}_i^*$. Specifically, at every location i, we define the corresponding covariate matrix $\mathbf{Z}_i$ as

$$\mathbf{Z}_i = \begin{bmatrix} 1 & 0 & \mathbb{1}(x_{i1}^* > 0) & \mathbb{1}(x_{i2}^* > 0) \\ 1 & \mathbb{1}(x_{i1}^* > 0) & \mathbb{1}(x_{i2}^* > 0) & \mathbb{1}(x_{i3}^* > 0) \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ 1 & \mathbb{1}(x_{i(l_i-1)}^* > 0) & \mathbb{1}(x_{i(l_i)}^* > 0) & 0 \end{bmatrix},$$

which uses the indicator for the occurrence of a threshold exceedance in $\mathbf{x}_i^*$ at times $t-1$, t, and $t+1$ as covariate information to predict a threshold exceedance in $\mathbf{y}_i^*$ at time t. This choice of $\mathbf{Z}_i$ has a physical interpretation, as it is not uncommon for remote-sensing or modelled data to have a lagged reaction to physical occurrences.

We incorporate additional flexibility into the model in (6.5) by estimating ξ inside the model. To ensure identifiability and reduce the computational burden, we assume that shape parameters do not vary with time and space and estimate ξ_y for all $\mathbf{y}_i^*$ and ξ_x for all $\mathbf{x}_i^*$. In the spirit of Castro-Camilo *et al.* (2022), we impose restrictions to ξ_y and ξ_x through prior distributions. These restrictions aim to preserve the usual MLE properties and the existence of first and second moments for both spatiotemporal exceedance processes. As mentioned in Section 2.2.1, these properties are achieved by restricting the shape parameter to the interval $[-0.5, 0.5]$. Additionally, the priors have a shrinking effect, whereby simpler or more parsimonious models (i.e., with shape parameter equal to 0) are chosen. After some exploration, we found that the above can be achieved using a scaled Laplace prior with density function

$$\text{Laplace}_\xi(x; \mu, b) = \begin{cases} \frac{1}{b}\left(\frac{\exp\left(\frac{-|x-\mu|}{b}\right)}{2-\exp\left(\frac{-0.5-\mu}{b}\right)-\exp\left(\frac{\mu-0.5}{b}\right)}\right) & \text{for } -0.5 \leq x \leq 0.5, \\ 0 & \text{otherwise,} \end{cases} \tag{6.9}$$

where $\mu \in \mathbb{R}$ is the location parameter and $b > 0$ is the scale, or diversity, parameter chosen to encourage parsimony; see Section 6.3.2. Note that the scaling factor in (6.9) is obtained by integrating the usual Laplace density in $[-0.5, 0.5]$.

By incorporating (6.8) and (6.9) to our modelling strategy, the new data fusion for

extremes model (ExDF) is defined as

$$
\begin{aligned}
\mathbf{y}_i^* \mid \mathbf{c}_i &\sim \delta\text{-GPD}\left(\exp(\Phi_i \mathbf{c}_i), \xi_y, \mathbf{p}_{yi}\right), \\
p_{y_i t} &= \text{logit}(\lambda_{i0} + \lambda_{i1}\mathbb{1}_{x^*_{i(t-1)}} + \lambda_{i2}\mathbb{1}_{x^*_{it}} + \lambda_{i3}\mathbb{1}_{x^*_{i(t+1)}})^{-1}, \\
c_{ij} \mid \alpha_{ij}, \beta_{ij}, d_{ij}, \sigma_c^2 &\sim \text{N}\left(\alpha_{ij} + \beta_{ij} d_{ij}, \sigma_c^2\right), \\
\xi_y &\sim \text{Laplace}_\xi(\mu_y, b_y) \quad \text{where } -0.5 \leq \xi_y \leq 0.5, \\
\boldsymbol{\alpha}_j \mid \sigma_\alpha^2 &\sim \text{N}_n\left(\mathbf{0}, \sigma_\alpha^2 \exp\left(-\phi_\alpha \Sigma_{\text{data}}\right)\right), \\
\boldsymbol{\beta}_j \mid \sigma_\beta^2 &\sim \text{N}_n\left(\mathbf{1}, \sigma_\beta^2 \exp\left(-\phi_\beta \Sigma_{\text{data}}\right)\right), \\
\left(\sigma_\alpha^2\right)^{-1} &\sim \text{Ga}\left(a_\alpha, b_\alpha\right), \\
\left(\sigma_\beta^2\right)^{-1} &\sim \text{Ga}\left(a_\beta, b_\beta\right), \\
\left(\sigma_c^2\right)^{-1} &\sim \text{Ga}\left(a_c, b_c\right), \\
\mathbf{x}_i^* \mid \mathbf{d}_i &\sim \delta\text{-GPD}\left(\exp(\Psi_i \mathbf{d}_i), \xi_x, \mathbf{p}_{xi}\right), \\
p_{x_i t} &= \mathbb{1}(x_{it}^* > 0), \\
\mathbf{d}_i &\sim \text{N}_m\left(\boldsymbol{\mu}_d, \Sigma_d\right), \\
\xi_x &\sim \text{Laplace}_\xi(\mu_x, b_x), \quad \text{where } -0.5 \leq \xi_x \leq 0.5, \\
\lambda_{i0} &\sim N(\mu_{\lambda_0}, \sigma_{\lambda_0}^2), \\
\lambda_{i1} &\sim N(\mu_{\lambda_1}, \sigma_{\lambda_1}^2), \\
\lambda_{i2} &\sim N(\mu_{\lambda_2}, \sigma_{\lambda_2}^2), \\
\lambda_{i3} &\sim N(\mu_{\lambda_3}, \sigma_{\lambda_3}^2),
\end{aligned}
\tag{6.10}
$$

where

- $i, \mathbf{y}_i, \mathbf{x}_i, \boldsymbol{\Phi}, \boldsymbol{\Psi}, \Sigma_{\text{data}}, \phi_\alpha$, ϕ_β, $\boldsymbol{\alpha}_j$, $\boldsymbol{\beta}_j$, $\boldsymbol{c}_i$, and $\boldsymbol{d}_i$ are defined as in (6.3).
- $\lambda_{i0}, \lambda_{i1}, \lambda_{i2}$ and λ_{i3} are the linear coefficients for the prediction of the probability of a threshold exceedance.
- $\mathbb{1}_{x_{i(t-1)}}$ is the indicator function with value 1 when $x_{i(t-1)} > 0$ and 0 otherwise. A similar definition is given for $\mathbb{1}_{x_{it}}$ and $\mathbb{1}_{x_{i(t+1)}}$.
- $\mathbf{p}_{y_i} = (p_{y_i 1}, ..., p_{y_i q_i})$ and $\mathbf{p}_{x_i} = (p_{x_i 1}, ..., p_{x_i l_i})$ represent the probability of a threshold occurrence at t for $\mathbf{y}_i^*$ and $\mathbf{x}_i^*$, respectively. Because x_{it}^* is always known, $\mathbf{p}_{x_i}$ is not estimated, but rather, it is a binary indicator with 1 indicating a threshold exceedance and 0 indicating a value under the threshold.
- $a_\alpha, b_\alpha, a_\beta, b_\beta, a_c, b_c, \mu_d$ and Σ_d are the same hyperparameters as in (6.3).
- $\mu_y, b_y, \mu_x, b_x, \mu_{\lambda_0}, \mu_{\lambda_1}, \mu_{\lambda_2}, \mu_{\lambda_3}, \sigma_{\lambda_0}^2, \sigma_{\lambda_1}^2, \sigma_{\lambda_2}^2$ and $\sigma_{\lambda_3}^2$ are hyperparameters for the estimation of the shape and probability parameters.

6.2.3 Performing Inference: Metropolis-Hastings MCMC

The model in (6.3) was fitted by (Wilkie *et al.*, 2019) using MCMC (see Section 2.4.1). Since all the priors and likelihoods were conjugate, the resulting posteriors had a closed-form solution and could be easily sampled using Gibb's sampler. However, this is not the case in our model, given that the δ-GPD has no conjugate priors. For this reason, all parameters are fitted using Metropolis-Hastings sampling, whereby the posterior is sampled from an acceptance/rejection process where the probability of acceptance/rejection of a parameter is the product of its density and the density (or likelihood) of the parameters that depend on it. The process order is more easily understood from the chart in Figure 6.1.

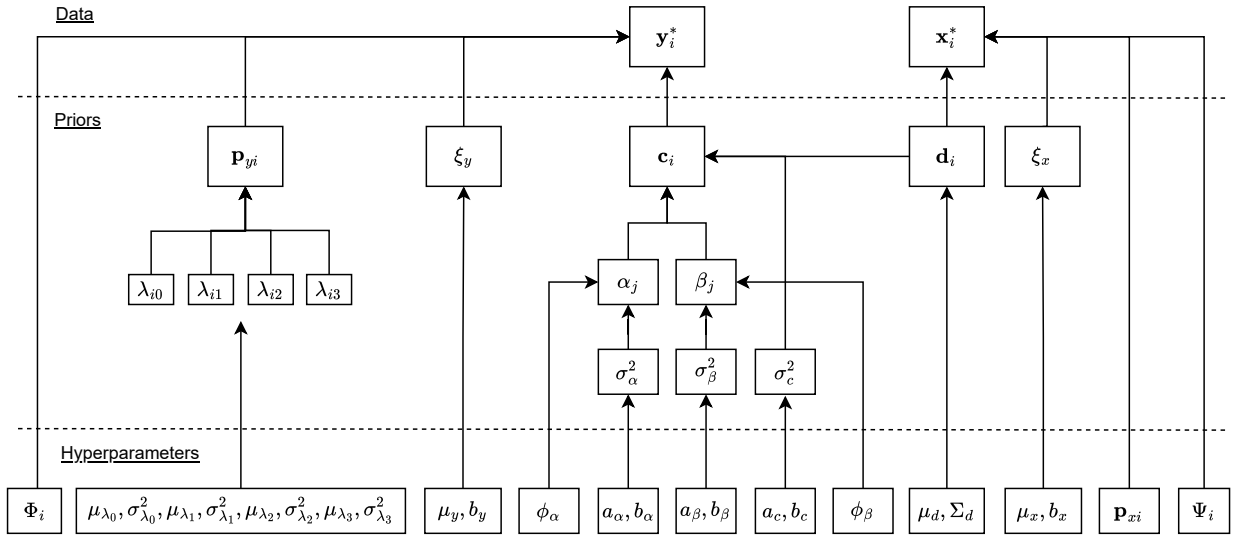

Figure 6.1: Hierarchy of the data fusion model for threshold exceedances defined in (6.10) and referred to as the ExDF model.

The model was originally fitted in `R` using MCMC; however, this approach was too slow to achieve converge of model parameters. Consequently, the code was into `C++` for computational efficiency with the help of `Rcpp`. The code is now freely available on Github (`https://github.com/danicuba-stats/DataFusion_for_Extremes`).

6.3 Application to $PM_{2.5}$ Air Pollution in the Greater London Area

The case study chosen to showcase the model in (6.10) is $PM_{2.5}$ air pollution in the Greater London area using the AURN and the EAC4 as in-situ measurements and modelled data, respectively. Our aim is to combine the highly localised, high-quality information of the

threshold exceedances at each location given by in-situ observation stations (AURN) with the complete spatial coverage and temporal coverage of the remote-sensing data (EAC4) to obtain a fused dataset that provides reliable estimates of threshold exceedances of $PM_{2.5}$ at locations where no in-situ observation station is present. An added byproduct of our approach is the retention of the temporal information of extremes and non-extreme occurrences, meaning our fusion approach could be used in combination with classical data fusion approaches for the bulk of the distribution to provide a full-range fused dataset. See Section 6.4 for a discussion of such approaches. Given that guidelines of $PM_{2.5}$ are given in 24-hour averages, the temporal scale in this application is daily means for the year 2022, the most complete year in the AURN dataset for the Greater London region.

Measurements of the AURN observation stations were in sub-daily frequencies with some missing observations but were aggregated to the 24-hour mean. Spatial coverage is poor and biased, with only 12 sites in the region with sufficient observations for the year 2022 mostly centred around more densely populated areas. For the full description of the data and an exploratory analysis, please see Section 3.4.2. The pre-processing of the data discussed in (6.6) is done for the AURN and the EAC4 data using $u = 0.8$, a threshold suitable for extreme value analysis for both datasets according to stability in the mean residual life plots and because it is low enough to obtain $n \approx 75$ observations for the year 2022.

The EAC4 dataset is a reanalysis model and provides modelled observations of $PM_{2.5}$ concentrations in a $0.1° \times 0.1°$ scale grid with global coverage. The data are smooth in time and space, with complete spatial and temporal coverages. For more details and an exploratory analysis, please see Section 3.4.1.

6.3.1 Differences between AURN and EAC4 Data

The EAC4 is a modelled representation of $PM_{2.5}$ measurements and the AURN data are in-situ measurements of $PM_{2.5}$. The EAC4 data are associated to their grid centroids, and each AURN station is also then matched to the nearest EAC4 centroid, as shown in Figure 6.2.

Each pair of AURN-EAC4 observations were plotted in a Q-Q plot shown in Figure 6.3. Data from the two sources are highly correlated as expected, but two major discrepancies between them are discernible. First, the bottom left corner of the plot shows most sites are above the 1-1 reference line, meaning the EAC4 overestimates very small values. The second is the expected behaviour around large values - the EAC4 data underestimates large values. This is consistently the case for sites further away from the city centre, such as A, B, G, and I to L. As we can see from Figure 6.3, site C shows more alignment along the 1-1 line between the EAC4 data set and AURN measurements.

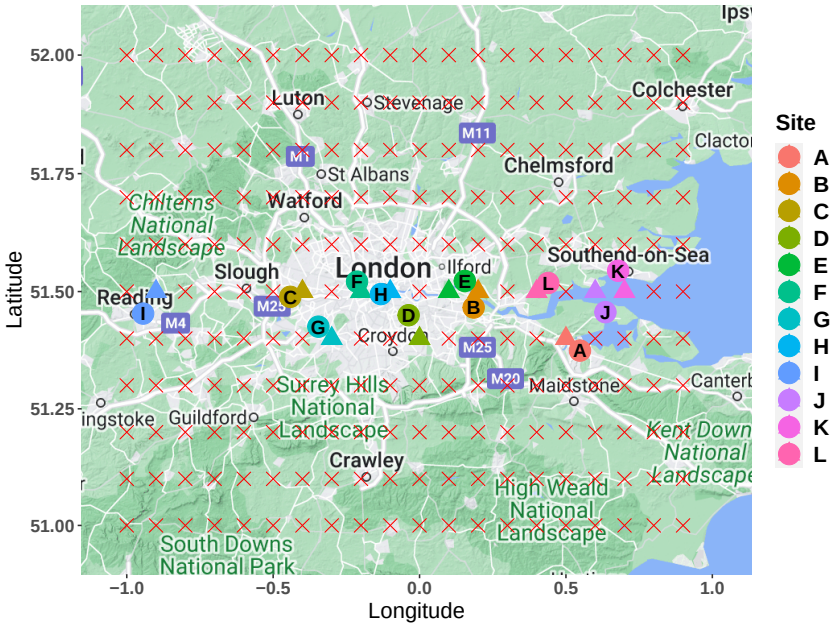

Figure 6.2: Map of the Greater London region with coloured circles denoting locations where an AURN observation station is located and coloured triangles indicating the nearest centroid in the EAC4 grid. The red crosses indicate the remaining EAC4 grid centroids.

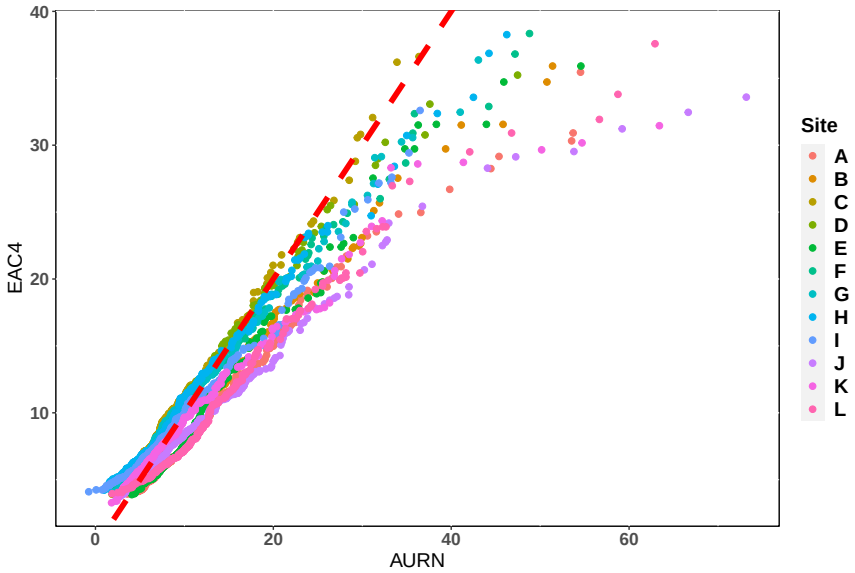

Figure 6.3: Q-Q plot for data from each AURN site (in different colours) and the nearest cell-centroid from the EAC4 grid.

6.3.2 Choice of Hyperparameter Values

In this section, we summarise the choices of hyperparameter values for our model in (6.10), which is the first step in fitting a Bayesian hierarchical framework. Recall that all hyperparameters are listed in the bottom line of the diagram in Figure 6.1. A sensitivity analysis using a leave-one-site-out approach was performed to ascertain the optimal number of

dimensions d for the basis functions Φ_i and Ψ_i, which model the temporal trend of y_i and x_i, respectively. The plots in Figure 6.4 show the changes in RMSE, MAE, mean coverage of the 95% predictive intervals and mean width of the 95% predictive intervals (mean PI width) with increasing d from $d = 5$ to $d = 150$ averaged across all locations. Only PI coverage is consistently capturing most observations which is due to the wide confidence bands produced by the GPD distribution. We can see in the figure that increasing d will always improve model fit. However, we can also see elbow plots for RMSE, MAE, and mean PI showing that improvements are marginal beyond a certain dimension. Additionally, the computational costs of increasing d are high, given the number of parameters is to $8nd$ where n is the number of locations, therefore, it is of interest to use the minimum suitable value of d. For these reasons, we fix the basis dimensions to $d = 60$ in the following, as it is the smallest dimension beyond which improvements in RMSE, MAE, and mean PI width are marginal.

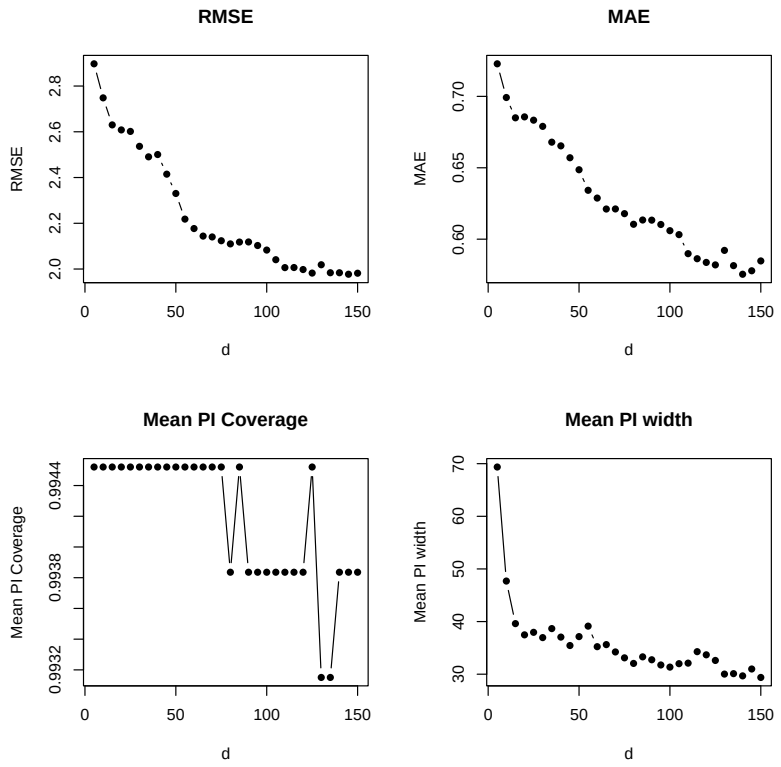

Figure 6.4: *Top:* RMSE and MAE estimates with increases in dimension d of the basis functions Φ_i and Ψ_i in (6.9). *Bottom:* Changes in the mean coverage of the 95% predictive intervals and the mean width of 95% predictive intervals.

The hyperparameters, μ_y, b_y, μ_x and b_x associated with the priors over the shape parameters ξ_y and ξ_x, were chosen to provide parsimony in the GPD likelihood by penalising deviations from $\xi_y = \xi_x = 0$. For this reason, $\mu_x, \mu_y = 0$ and $b_x, b_y = 0.05$. The posterior distribution of ξ_x and ξ_y is robust to these decisions, as shown in the trace plots in

Appendix B.

For the vector of hyperparameters $\boldsymbol{\lambda}_i = (\lambda_{i0}, \lambda_{i1}, \lambda_{i2}, \lambda_{i3})$, those related to the logistic regression component of the model (see (6.8)), a study was conducted to assess the sensitivity of the model to the prior specifications of these parameters. Figure 6.5 shows changes in the posterior of λ_j given different prior specifications, ranging from highly informative ($\mu_{\lambda_j} = \text{logit}^{-1}(0)$) to very uninformative ($\mu_{\lambda_j} = \text{logit}^{-1}(0.5)$) for $j = 0, 1, 2, 3$. As seen in the figure, all components of $\boldsymbol{\lambda}_i$ are robust to these choices and always converge to very similar values. Therefore, for simplicity, all $\boldsymbol{\lambda}_i$ are set to $\mu_{\lambda_j} = 0$ and $\sigma_{\lambda_j} = 1, j = 0, 1, 2, 3$.

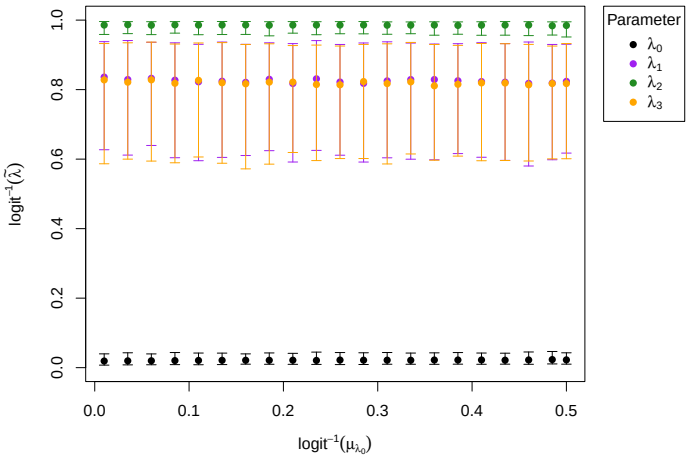

Figure 6.5: Mean estimates and 95% credible intervals for the posterior samples of $\boldsymbol{\lambda}_i$ against various values for μ_{λ_j}, for $j = \{0, 1, 2, 3\}$

Finally, the exponential decay parameters ϕ_α and ϕ_β were chosen by estimating proxy parameters. Regression models, $y_i^* = a_i + b_i x_i^*$, were fitted at each location i. Variograms of the parameters a and b were estimated, and exponential models were fitted. The resulting exponential decay parameters were $\phi_a, \phi_b = 1.6$, and can be considered a proxy for the exponential decay of ϕ_α and ϕ_β.

Computation and Convergence

The complexity of the model in (6.7) is mostly due to the number of parameters as defined by d. CPU memory limitations turn prohibitive when fitting the model at $d = 60$ - equivalent to 2932 parameters when using 12 sites - in base R. For this reason, the model was coded in C++, which cut computation time for half a million samples of the posterior predictive distribution from close to 100 hours to 4 hours.

To assess the convergence of the model, two MCMC chains were run at randomised initial parameter values. Each chain consisted of 3 million sampling iterations, with 1

million samples discarded as burn-in. The convergence of the parameters is strong, and trace and density plots for some parameters are given in Appendix B. The Gelman-Rubin convergence metric was also estimated, resulting in point estimates of 1, demonstrating strong convergence across all parameters. The process also demonstrated that convergence is achieved with as little as half a million samples with 100,000 sample burn-in.

Goodness-of-fit

Overall goodness of fit of our model at specific locations can be visualised in Figure 6.6 and 6.7. At the fitted sites in Figure 6.6, the mean of the predictive posterior distribution follows the data appropriately, identifying threshold exceedances $y > 0$ as well as non-threshold exceedances (censored as $y = 0$). The 95% credible intervals have high coverage of the AURN observations, with the model failing to cover only 2 observations on average, equivalent to a 0.978 coverage probability. The width of the credible intervals, however, is large. At $t = 83$, the yearly maximum, the upper bound of the credible intervals is 120 for site D and 85 for site I. Density plots for the shape parameters, ξ_x and ξ_y, are given in the Appendix B. The estimated values posterior means are $\xi_x = -0.22$ with $(-0.34, 0.02)$ as the 95% credible interval, and $\xi_y = -0.10$ with $(-0.216, 0.05)$ as the 95% credible interval.

Figure 6.7 provides Q-Q plots for the fitted sites D and I, shown in Figure 6.6, as well as 95% point-wise confidence intervals for the ExDF fitted values obtained via bootstrap, where 1000 samples of the posterior predictive distribution where sampled with replacement, and the quantiles of interest were estimated. This process was repeated 1000 times to obtain robust results. The plots show a similar story as that of Figure 6.6 where the ExDF fitted values are a better representation of the AURN threshold exceedances than the EAC4 data, especially at high values. Similar figures for the remaining sites can be seen in Appendix B.

6.3.3 Investigating Exceedance Probability

A summary of the posterior means and 95% credible intervals for $\boldsymbol{\lambda}_i$ averaged across all fitted sites A to K when predicting over site L is given in Table 6.1. The table shows that, on average, $\text{logit}^{-1}(\hat{\lambda}_0) = 0.05$, meaning that the model assigns a 0.05 probability of exceeding the threshold to all times points, irrespective of the presence of a threshold exceedance in the EAC4 dataset. $\hat{\lambda}_2$, the regression coefficient linked to time t, has an average value of 0.9 in the probability scale, meaning there is a 0.9 probability of observing an exceedance in y (in-situ measurements) given an exceedance in x (modelled data) at the same time. Both $\hat{\lambda}_1$ and $\hat{\lambda}_3$ have confidence intervals inclusive of 0 in the logit scale and 0.5 in the probability scale, showing that the lagged presence of exceedance in x, whether $t-1$ or $t+1$, is not a strong linear predictor of observing an exceedance at t in y.

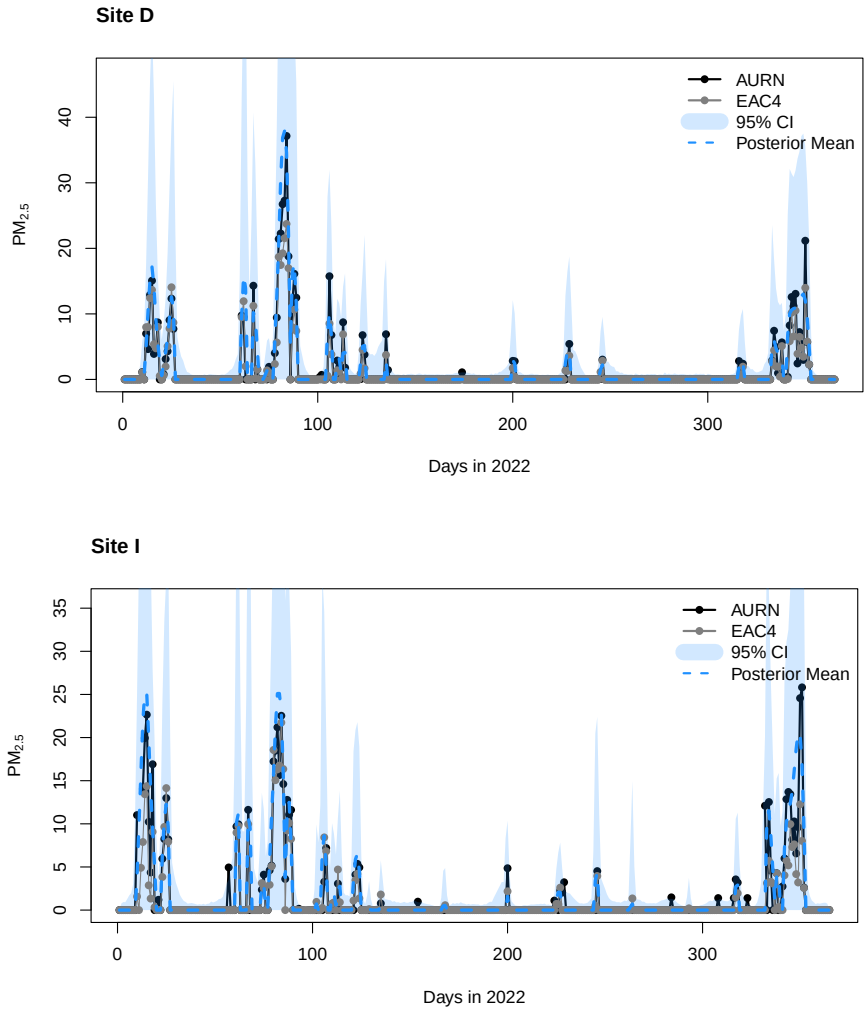

Figure 6.6: $PM_{2.5}$ measurements from the site D (top) and I (bottom) from the AURN are shown in black, EAC4 values are shown in grey, and posterior mean is shown in blue along with the 95% credible band.

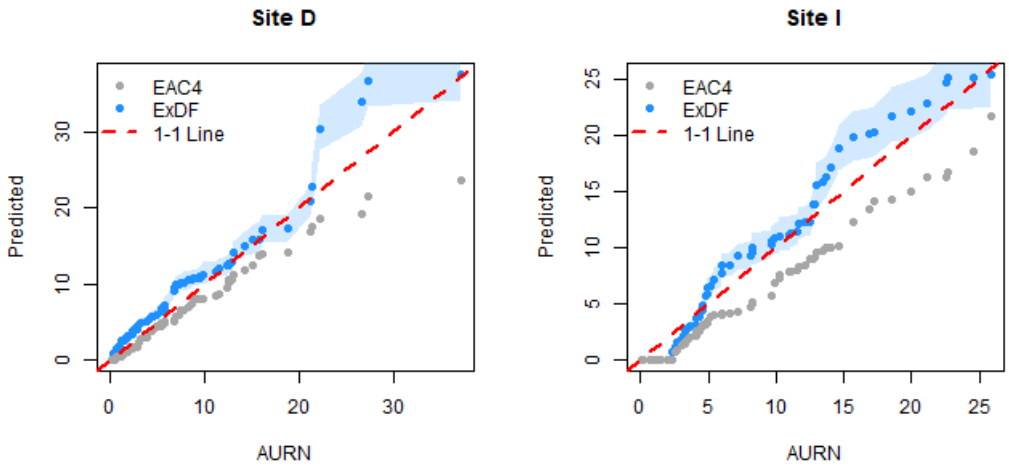

Figure 6.7: Q-Q plot of $PM_{2.5}$ measurements at site D(left) and I (right) for the ExDF and EAC4 models in blue and grey, respectively, against the true observations from the AURN observation stations at those locations. Point-wise 95% confidence intervals are given for the ExDF data.

However, we kept them since they still contribute to the predictive ability of our model.

Table 6.1: Posterior means and 95% credible interval for $\hat{\lambda}_j$ and $logit(\hat{\lambda}_j)$ for $j = \{0, 1, 2, 3\}$, averaged over the 11 fitted sites when predicting over site L.

Parameter	$\hat{\lambda}_j$	$\textbf{logit}^{-1}(\hat{\lambda}_j)$
λ_0	-2.95(-3.82,-2.07)	0.05(0.02,0.11)
λ_1	0.85(-0.53,2.23)	0.7(0.37,0.90)
λ_2	2.19(0.81,3.58)	0.9(0.69,0.97)
λ_3	0.85(-0.55,2.23)	0.7(0.37,0.90)

The classification performed by the logistic regression was investigated using classic classification metrics, such as accuracy, precision, recall, specificity, and F1 Score. Table 6.2 shows the summary for the sites used in model fitting, A to K, along with the mean classification metrics of the EAC4 data for those sites. As seen in the table, the correct classification of observations into threshold exceedances and non-exceedances is high across all metrics for all fitted sites. Only sites E and I have lower metrics, with F1 Scores < 0.8. While this is still a good classification, it is notable that different sites differ in classification performance and are likely to reflect the classification of the EAC4 data due to the importance of λ_2, the regression coefficient linked to t.

Table 6.2: Classification metrics for the logistic regression component of our model for fitted sites A to K. Table includes a comparison with the EAC4 data.

Site	Accuracy	Precision	Recall	Specificity	F1 Score
A	0.94	0.84	0.89	0.96	0.86
B	0.96	0.87	0.95	0.97	0.91
C	0.96	0.87	0.92	0.97	0.89
D	0.97	0.90	0.95	0.97	0.92
E	0.91	0.79	0.76	0.95	0.77
F	0.94	0.83	0.89	0.95	0.86
G	0.95	0.85	0.92	0.96	0.88
H	0.92	0.77	0.85	0.93	0.81
I	0.90	0.75	0.79	0.93	0.77
J	0.95	0.84	0.92	0.96	0.88
K	0.95	0.86	0.92	0.96	0.89
EAC4	0.95	0.87	0.87	0.97	0.87

6.3.4 Model Validation through Leave-one-site-out

A leave-one-site-out cross-validation (LOSO-CV) procedure was performed on the twelve sites available in the Greater London region to assess the model's predictive performance where "true" observations are available, meaning measurements from an AURN station are available. For each iteration of this cross-validation, a site is removed from the training set and used as a test location, the model is fitted to the available locations in the training set, and a prediction is made for the test location. Given that the focus of the proposed model is on threshold exceedances and the prediction of an exceeding occurrence has already been evaluated, only the prediction of the size of the threshold exceedance is evaluated here. A summary of the results is given in Table 6.3 alongside a comparison between the data fusion for extremes proposed in this chapter (ExDF), the Gaussian model of Wilkie *et al.* (2019) (GausDF) and the prediction available using the EAC4 modelled data. The comparison is made using RMSE and MAE. Additionally, an evaluation of the probabilistic prediction of the GausDF and the ExDF models is done through the continuous rank probability score (CRPS) defined as

$$\text{CRPS}(F, y) = \int_{\mathbb{R}} [F(x) - \mathbb{1}(x \geq y)]^2 \, \mathrm{d}x,$$

where F is the cumulative distribution function of the predictive distribution, and $y \in \mathbb{R}$ is the true value. The table shows two important results. The first is that both data fusion approaches improved the representation of threshold exceedances in the EAC4 data. Of the twelve sites, the RMSE and MAE show that the GausDF and the ExDF models outperform the EAC4 data for 7 and 10 sites, respectively. However, this result is not homogeneous across locations. Indeed, the EAC4 shows better performance at sites C and

G. Although site C is best described by the EAC4 data, which was already suspected from Figure 6.3, it is followed closely by the ExDF model. A bigger difference is seen in site G.

The second main result from table 6.3 is that the ExDF model outperforms the GausDF model at every site. Evidence for this result is the lower RMSE, MAE and CRPS values at every site, with lower CRPS values indicating a better probabilistic prediction. This result was expected, as the ExDF approach is tailored for threshold exceedances, unlike the GausDF and the EAC4 models.

Table 6.3: Results of the LOSO-CV comparisons between GausDF, ExDF and the EAC4 data. Table provides RMSE, MAE, and CPRS values for sites A to L for each of the three data sources. Values in bold indicate the minimum value using that metric at that site.

	RMSE			**MAE**			**CRPS**	
Site	GausDF	ExDF	EAC4	GausDF	ExDF	EAC4	GausDF	ExDF
A	3.12	**1.68**	7.57	2.79	**0.94**	6.44	0.52	**0.35**
B	2.25	**1.20**	5.01	2.12	**0.83**	4.03	0.52	**0.36**
C	2.21	1.10	**0.62**	1.66	0.68	**0.50**	0.52	**0.36**
D	1.71	**1.21**	1.94	1.22	**0.77**	0.99	0.52	**0.38**
E	5.30	**2.20**	5.89	5.12	**1.64**	5.23	0.47	**0.29**
F	3.89	**1.66**	3.12	3.75	**1.01**	2.00	0.48	**0.34**
G	3.53	3.09	**2.29**	3.35	2.38	**1.57**	0.49	**0.19**
H	2.92	**2.38**	2.60	2.56	**1.66**	1.75	0.50	**0.18**
I	3.80	**2.66**	2.83	3.66	**2.45**	2.57	0.50	**0.15**
J	6.67	**5.50**	10.29	5.90	**3.42**	8.09	0.48	**0.18**
K	3.33	**3.11**	6.78	2.30	**2.12**	4.59	0.51	**0.18**
L	5.30	**2.10**	8.49	5.03	**0.98**	7.31	0.49	**0.34**

The Q-Q plots in Figures 6.8 to 6.10 help visualise the validation results shown in Table 6.3. Sites A, B, and D in Figure 6.8 show that the ExDF produce a better representation of the true values at that location. The EAC4 data seems to be better only at site C. For this site, the similarity between the ExDF model and the EAC4 data is evident, while the GausDF model shows larger discrepancies with the true values for smaller threshold exceedances. The figure also shows that, for these sites, the GausDF model diverges from the true values most at smaller threshold exceedances.

In Figure 6.9 (sites E to H), a similar pattern is visible. Threshold exceedances at sites E, F, and H are better captured by the ExDF model, while the GausDF and EAC4 experience a varied performance. Only at site G is the EAC4 data a better fit (as noted in Table 6.3). The figure shows that none of the three models are good at capturing the largest exceedances at these four locations, especially for values $PM_{2.5}> 30$. However, the EAC4 is a closer fit for smaller threshold exceedances, followed by the ExDF and the GausDF models, meaning the EAC4's representation of the extremes deteriorates the further away from the mean, as expected in non-extreme models.

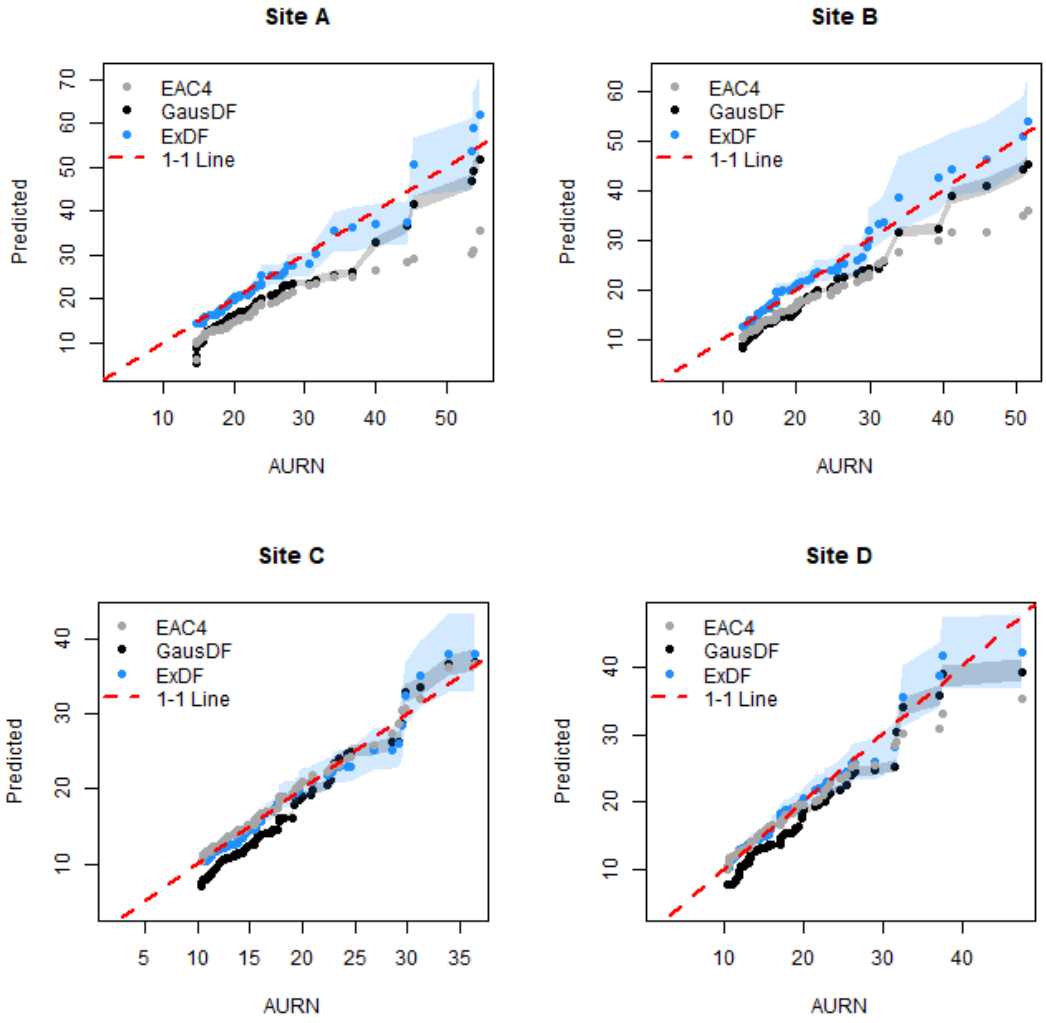

Figure 6.8: Q-Q plots of LOSO-CV results for sites A to D. The figures compare the GausDF model (Wilkie *et al.*, 2019) in black, the EAC4 data in grey, and the data fusion for extremes model (ExDF) in blue. Point-wise 95% confidence intervals are given for the ExDF and GausDF models.

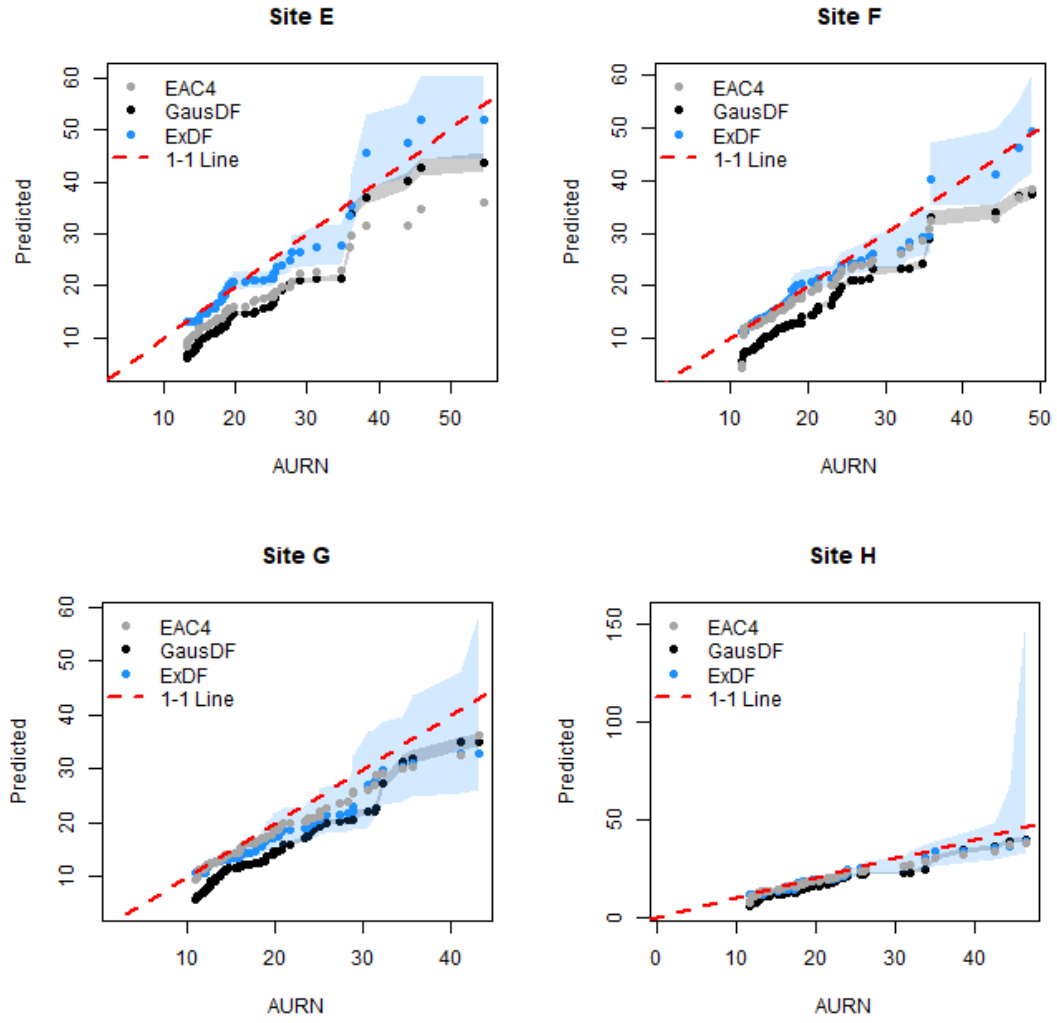

Figure 6.9: Q-Q plots of LOSO-CV results for sites E to H. The figures compare the GausDF model (Wilkie *et al.*, 2019) in black, the EAC4 data in grey, and the data fusion for extremes model (ExDF) in blue. Point-wise 95% confidence intervals are given for the ExDF and GausDF models.

Finally, Figure 6.10 shows model fits for sites I to L. Threshold exceedances at sites I and L are well captured by the ExDF model. For sites J and K, a pattern is visible across all data sources: underestimation of the largest threshold exceedances. While the ExDF confidence intervals still capture this behaviour, it is clear to see that these sites behave differently from the others. Further investigation into the results showed that, although these observation stations are available at locations near the coast, the nearest EAC4 cell centroids, from where the data for the EAC4 source was extracted, are directly over the North Sea. No other sites experience this disparity between conditions at the AURN site and the nearest EAC4 centroid, which could explain the poor performance of all models at this location. A possible solution to this problem is choosing a different grid-cell in the EAC4 grid, requiring a possible reevaluation of the nearest-centroid rule for data extraction.

Overall, the figures show that the ExDF model produces a better representation of the threshold exceedances of the in-situ measurements than the readily available EAC4 data and the Gaussian approach. Larger confidence intervals are seen for the point-wise estimates of the threshold exceedances for the ExDF compared to those of the GausDF. The figures also shine a light on the limitations of data fusion models. At locations where the remote-sensing or modelled data do not reflect the true conditions, improvements using data fusion models over the EAC4 data are small and objectively negligible.

6.3.5 Maps of $PM_{2.5}$ Expected Shortfall from EAC4 and ExDF data

Predictions over the Greater London area were made to highlight the improvements in the EAC4 $PM_{2.5}$ measurements when AURN measurements are fused using the ExDF model proposed in this chapter. Figure 6.11 shows the expected shortfall, defined as $\mathrm{E}(Y - u|Y > u)$, and the range of the threshold exceedances, defined as $\max - \min$, for the EAC4 and the ExDF data. The shortfall estimates in the top row show that the EAC4 are smooth in space and experience little variability. The largest shortfall estimate is $6.2\,\mu g/m^3$ observed in the London city centre, followed by the region in the southwest. In the ExDF data, a different spatial pattern emerges. The largest values in the ExDF data show high values on the east coast, with the largest shortfall value as $9.07\,\mu g/m^3$ near Southend-on-Sea. Differences are also noticeable around the London city centre, where the EAC4 experiences higher shortfall values and the ExDF experiences lower. Other patterns, however, seem to be in agreement. For example, the northwest corner for both data sources has lower values, while a pattern of high-low-high values from the city centre to the southeast corner is visible.

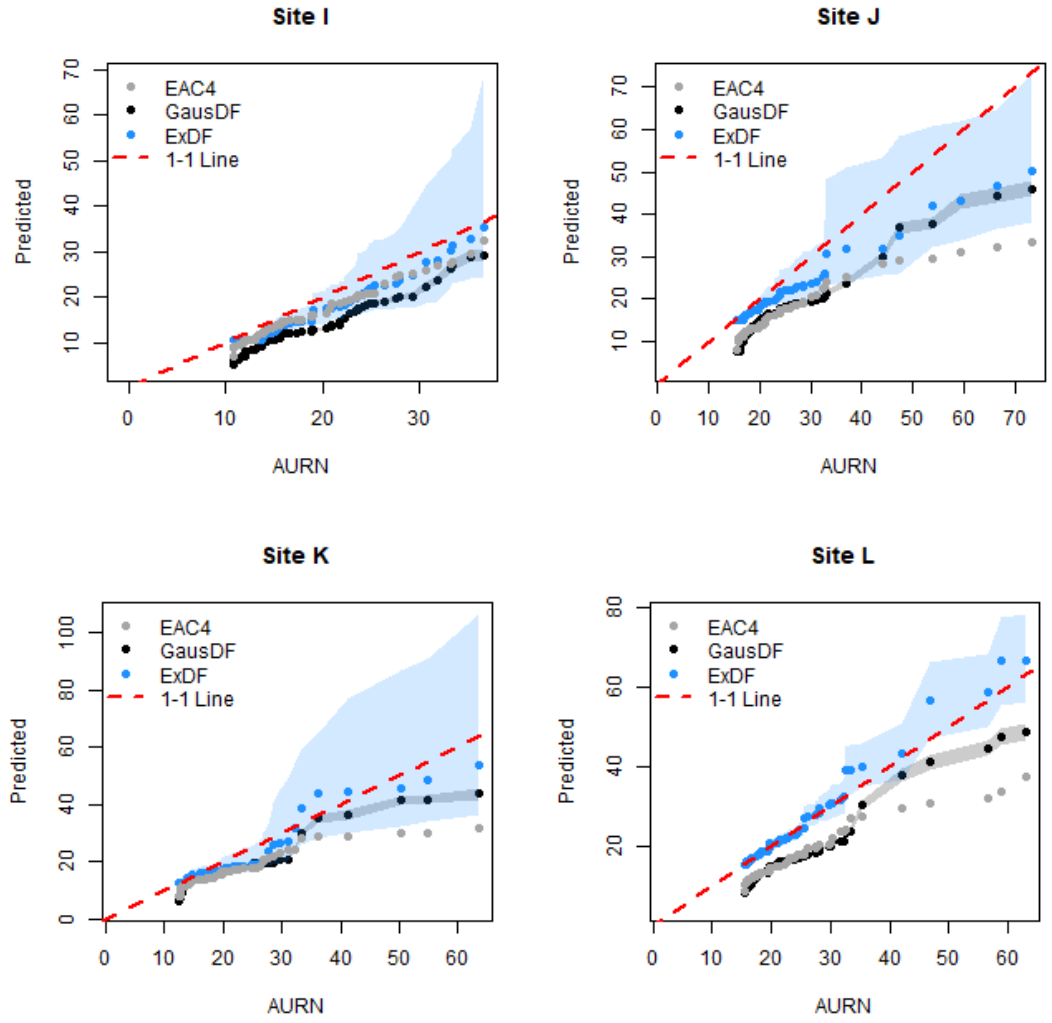

Figure 6.10: Q-Q plots of LOSO-CV results for sites I to L. The figures compare the GausDF model (Wilkie *et al.*, 2019) in black, the EAC4 data in grey, and the data fusion for extremes model (ExDF) in blue. Point-wise 95% confidence intervals are given for the ExDF and GausDF models.

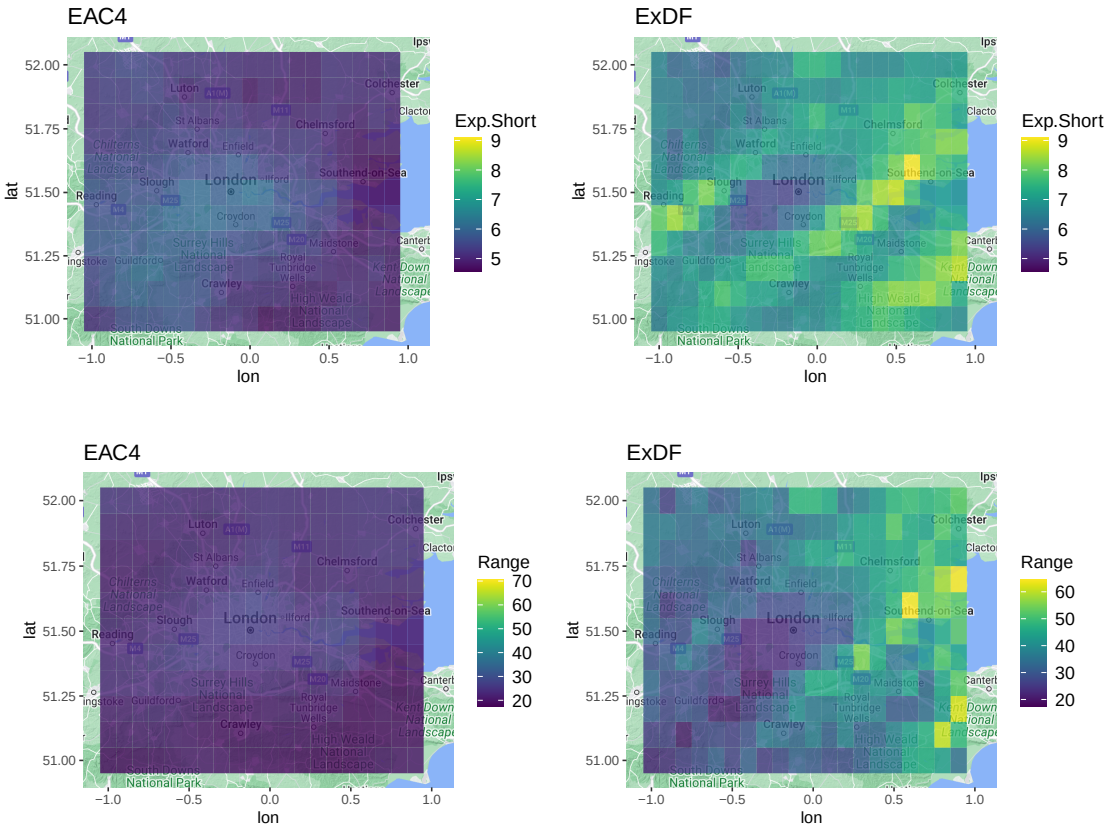

Figure 6.11: *Top*: Map of $PM_{2.5}$ expected shortfall from the EAC4 data and the ExDF model for the year 2022. *Bottom:* Map of the range of exceedances for the EAC4 data and the ExDF model for the year 2022.

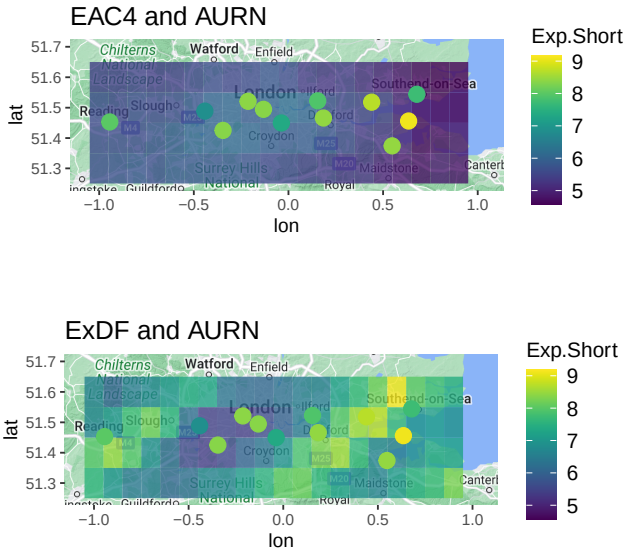

Figure 6.12: Cropped map of the Greater London area showing $PM_{2.5}$ expected shortfall from the EAC4 (top) and ExDF (bottom) data along with the empirical shortfall computed at AURN observation stations for the year 2022.

The range figures on the bottom row of Figure 6.11 show a similar pattern as the shortfall, with the EAC4 threshold exceedances exhibiting overall narrower ranges with raised values around the London city centre, and the ExDF data showing wide ranges on the east coast and narrower ranges further west. Overall, the figures provide similar information. In the EAC4 data, higher $PM_{2.5}$ values are consistently found near the London city centre. The ExDF data, on the other hand, shows higher exceedances on the east coast, especially at locations directly on the coast and around the mouth of the river Thames. A pattern of higher values of $PM_{2.5}$ on the coast is also seen in the literature. For perspective, Figure 6.12 provides a cropped map with overlayed expected shorfall values from the AURN observation stations. The figures show that the variability of the ExDF map is much closer to the in-situ measurements than the EAC4, even capturing the high values observed near the coast. Yang *et al.* (2023) showed that values are generally higher around the coast due to emissions from maritime transport and various related sources, as well as salt content from the sea.

Overall, the maps highlight the difference between threshold exceedances in modelled and fused data - a proxy for the difference between modelled and in-situ measurements - which is significant, both in terms of spatial patterns and measurement values.

6.4 Discussion and Future Work

The ExDF model proposed in (6.7) is an extension of the data fusion model proposed by Wilkie *et al.* (2019) for fusing extreme values from in-situ and remote-sensing sources to exploit the spatial coverage of remote-sensing observations and retain the accuracy of in-situ data. It links the two datasets through the scale parameter of the GPD using a flexible regression whose parameters are subject to change in space and time. Unlike previous models that use EVT for data fusion, the model presented here retains information about the time of a threshold exceedance by exploiting the information provided by the remote-sensing source. As such, the approach is the symbiotic combination of two mechanisms. The first models the probability of a threshold exceedance in the in-situ data using threshold exceedances in the remote-sensing observations and classifies observations as threshold and non-threshold exceedances, while the second models the magnitude of the exceedance.

A generalised linear model (GLM) approach predicts the probability of observing a threshold exceedance in the in-situ data, given that one has been observed in the remote-sensing data. The covariates in the GLM are the lagged indicators of remote-sensing data exceedances at times $t-1$, t and $t+1$. The model shows that only x_{it} highly affects the probability of exceedance while $x_{i(t-1)}$ and $x_{i(t+1)}$ have little influence on the probability of exceedance. Results from the predictions show that they are dependent on the remote-sensing observations - meaning the ExDF model only predicts an exceedance if one is observed in the EAC4 data. Nevertheless, this mechanism is helpful for accurately estimating extremes under a GPD likelihood by mitigating the smoothing effect of the spline functions and increasing the temporal accuracy of threshold exceedance predictions. Changes to improve this classification can be made by re-assessing the GLM structure or by integrating covariates that are known to affect $PM_{2.5}$concentrations. For example, Jin *et al.* (2022) show that appropriate covariates are meteorological variables such as dew point temperature, temperature, humidity, relative humidity, precipitation, potential evapotranspiration rate, and windspeed. Other auxiliary data include normalised difference vegetation index (NDVI), enhanced vegetation index (EVI), population density, and Keetch-Byram drought index (KBDI). Including these auxiliary variables in the model can improve the estimated probability of exceedance and is a natural extension of the model. The predicted probabilities are incorporated into the ExDF model through a Dirac-delta generalised Pareto distribution, which allows the preservation of the temporal location of threshold and non-threshold exceedances in the time series by censoring non-threshold exceedances. As mentioned in Section 6.2, this model feature is highly desirable as it allows our model to be used with classical data fusion approaches for the bulk of the distribution to provide a full-range fused dataset. Although this is beyond the scope of this chapter, we envision that such a strategy should incorporate careful assessment of the impact of

using different models for the fusion of bulk and exceedances values, especially around the threshold. Alternatively, a more seamless full-range data fusion approach could be conceived using continuous models that are known to adequately capture extreme and non-extreme observations, such as the extended generalised Pareto distribution (Naveau *et al.*, 2016).

The magnitude of threshold exceedances is predicted using a GPD likelihood inside the Dirac-delta framework proposed by Weglarczyk *et al.* (2005) and Couturier and Victoria-Feser (2010). While modelling threshold exceedances with a GPD is intuitive, the model is still subject to assumptions and impositions in the name of parsimony or practicality. For example, the assumption of independence between ξ_x and ξ_y was made to simplify the number of estimated parameters. An increase in complexity (and the number of model parameters) could allow the inclusion of a linear or non-linear relationship between the two. However, this was discarded as increasing the number of the parameters can result in challenges in terms of estimation and inference, potentially jeopardising the identifiability of the linear relationship between scale parameters (namely parameters c_i and d_i), which already accounts for the relationship between the two datasets. Another assumption is the spatial exponential decay imposed in the model through parameters ϕ_α and ϕ_β. In this model, we propose to estimate the exponential decay parameters *a priori* by modelling the variogram of proxy parameters and utilising the estimated exponential decay coefficient. Although this method has worked well, the spatial structure might not be appropriate for all applications and is restricted to second-order stationarity.

The results in Section 6.3 show the model performs well, and the posterior predictive mean of the prediction seems an appropriate candidate for point predictions, yielding non-zero observations for threshold exceedances and zero values for censored non-exceedances. However, it is limited by the quality of the prediction made in the remote-sensing data and is also subject to its bias as highlighted by Maraun and Widmann (2018).

Finally, the maps of expected shortfall using the EAC4 and ExDF models in Section 6.3.5 highlight the difference between the two models. The EAC4 data are spatially smooth and exhibit low variability, while the ExDF model displays a wider variability. The spatial pattern in the EAC4 data shows a slight increase in $PM_{2.5}$ concentrations in the London city centre and the southwest, while the ExDF show higher, much higher concentrations at coastal locations. The maps show that a data fusion model for extreme values can help reveal different spatial patterns and provide better information as to the location and the intensity of $PM_{2.5}$ pollution, which is not otherwise seen in the EAC4 dataset. Further conclusions and discussion about the practical implications of this work are given in Section 8.2.

Chapter 7

EVA 2023 Data Challenge: The Wee Extremes Team

7.1 Motivation and Context of Data Challenge

The latest Extreme Value Analysis (EVA) Conference was held in the summer of 2023 in Milan, Italy. As is customary, the organising committee set up a Data Challenge event to engage conference attendees or groups interested in EVA. PhD students at the University of Glasgow participated in the challenge under the team name "The Wee Extremes". Although the team submitted entries for all four challenges posed by the organisers, this thesis covers only challenges 2 and 4, denoted C2 and C4, respectively, as those represent the author's contribution to the team entry. Challenges 1 and 3, denoted C1 and C3, respectively, were developed by other members of the team and will consequently not feature in this chapter.

The event consisted of four challenges. C1 focused on modelling univariate nonstationary extremes with missing observations using covariate information. C2 also addressed univariate extreme modelling but focused on estimating return values for extrapolatory probabilities. As an added complication, the challenge required the optimisation of an application-specific loss function, a more realistic scenario when extreme models are required for decision-making. C3 and C4, focused on multivariate challenges in extreme value analysis. In the first part of C3, participants were asked to find the probability that an extreme value will be exceeded at three locations simultaneously; in the second part, the participants were asked for the probability that only two of the three locations would observe an exceedance. In C4, the problem of high dimensionality was approached, and participants were asked to consider two different regions with 25 observation stations in each region. The first part of the challenge required participants to estimate the probability that all observations in a region will exceed a region-specific high value simultaneously, while the second part required the estimation of the probability that all observation sta-

tions exceed the same high value (Rohrbeck *et al.*, 2023).

In the remainder of this chapter, we will provide the work for which the author was responsible - C2 and C4. Section 7.3 provides further details for the problem posed in C2, describes the data, details the methodology developed for the problem, and provides the results. Section 7.4 provides similar information for C4. A summary of the conclusions and a discussion in light of the truth being revealed in Rohrbeck *et al.* (2023) is provided in Section 7.5.

7.2 Some Utopian Context

The challenges the organisers pose are common problems encountered by extreme value analysis in environmental applications. However, the problems were set on an alternate planet called Utopia, where the environment resembles that of planet Earth but retains significant differences.

Various features of Utopia are given to facilitate the modelling process. For example, the Utopian year consists of 300 days. It is both seasonal and cyclical, enjoying only 2 seasons in a year and with a cyclical period of 70 years. The data are generally presented as $Y_{i,t}$, where $i \in \mathcal{I}$ denotes different locations. Despite the spatial aspect of these data, no location information is provided, and $Y_{i,t}$ are considered to be independent and identically distributed given the set of covariates. In the multivariate challenges, C3 and C4, the marginal distributions of Y are identical over space and time, and follow a standard Gumbel distribution.

Finally, in C4, we consider the fact that the planet's government is subdivided into two independent regional governments, U_1 and U_2, each responsible for 25 towns.

7.2.1 C2 - Extremes of a Univariate Random Variable

The data provided for the challenge consists of samples of a single vector $Y \in [0, \infty)$ of 70,000 observations. The minimum value in the data is 0.01, while the maximum is 210. The data display a heavy tail, as shown in the histogram in Figure 7.1, with 50% of the observations existing in the interval $[25.3, 42.4]$.

When the focus is on threshold exceedances, the mean residual life plot (Coles, 2001) in the right-hand side of Figure 7.1 shows a linear behaviour starting at $u_0 = 77$ (corresponding to the 95% quantile of the data), as the plot is roughly linear before and after roughly corresponding to the threshold-stability property of the GPD (see 2.2.1). Therefore, we consider $u_0 = 77$ to be an appropriate threshold for the GPD distribution.

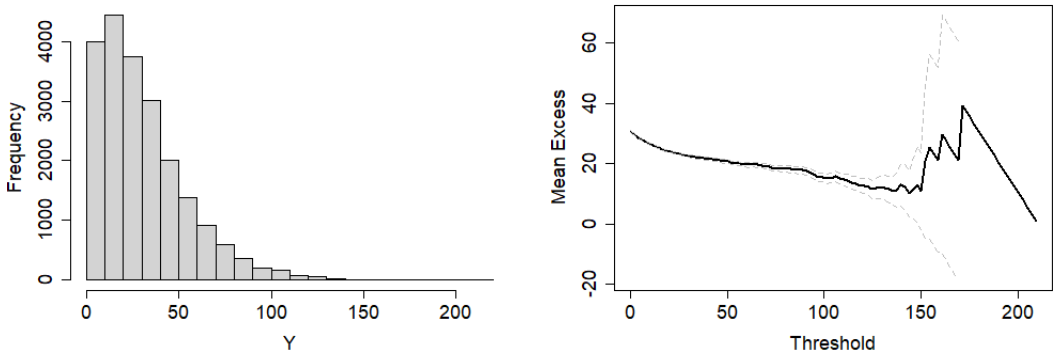

Figure 7.1: *Left:* Histogram of response variable Y in C2 displaying support in $[0, \infty)$ and a heavy-tail. *Right:* Mean residual life plot of the data displaying linearity at approximately $u = 77$

7.2.2 C4 - Multivariate Dataset for U_1 and U_2

The data were given for two governmental regions, U_j for $j = 1, 2$. Each region had $i = 1, ..., 25$ locations, where each location consisted of a time series spanning $10,000$ days. We denote the time series as $\mathbf{Y}_j = (Y_{1j}, ..., Y_{25j})$ for $j = (1, 2)$. No spatial information was given to enable a spatial modelling approach.

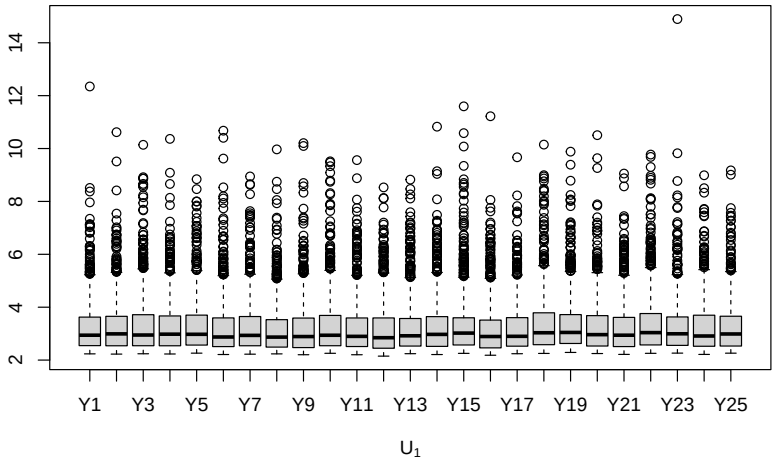

Figure 7.2: Boxplot of observations exceeding the 90th-quantile for all locations $i = 1, ..., 25$ in U_1.

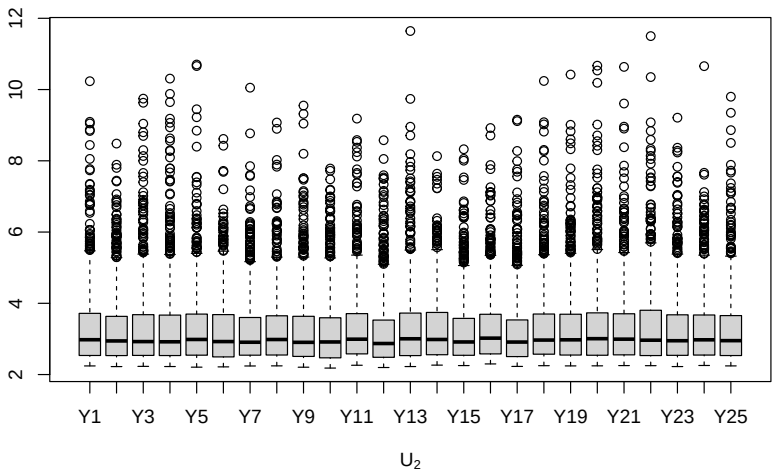

Figure 7.3: Boxplot of observations exceeding the 90th-quantile for all locations $i = 1, ..., 25$ in U_2.

Figures 7.2 and 7.3 show boxplots for observations exceeding the 90th-quantile at each location in regions U_1 and U_2. No clearly discernible pattern or outliers arise in U_1 or U_2. However, the extremal dependence between sites, measured using $\chi(u) = \Pr(Y_i > u | Y_{k \neq i} > u)$, shown in Figure 7.4 shows a diverse set of dependence structures across locations in each region. In U_1, the figure shows three asymptotically independent clusters starting at $\chi(0.9) = \{0.1, 0.3, 0.4\}$ and one asymptotically dependent cluster at $\chi(0.9) = 0.5$. In U_2, there are two asymptotically independent clusters at $\chi(0.9) = \{0.1, 0.3\}$, one asymptotically dependent cluster with decaying dependence at $\chi(0.9) = \{0.4\}$ and an asymptotically dependent cluster with almost constant dependence at $\chi(0.9) = 0.5$.

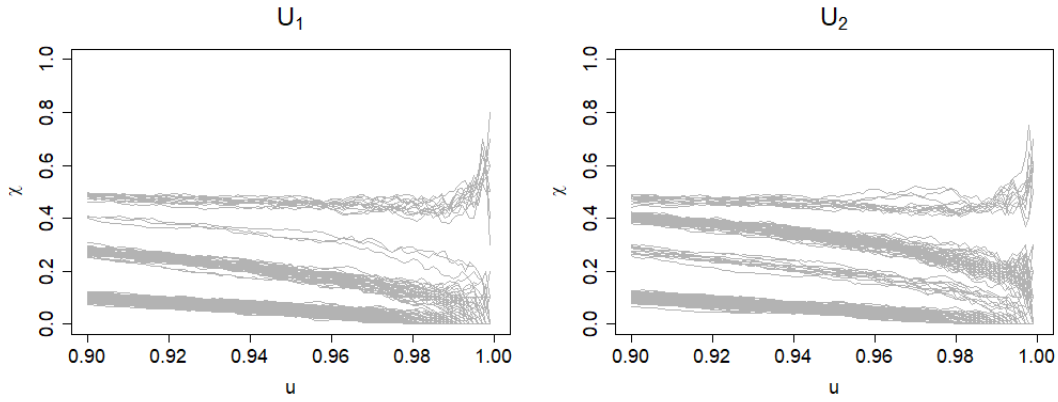

Figure 7.4: *Left:* Plots of the coefficient of tail dependence χ for all possible pairs in region U_1 for thresholds $u > 0.9$. *Right:* Plots of the coefficient of tail dependence χ for all possible pairs in region U_2 for thresholds $u > 0.9$.

7.3 C2 - Univariate Extrapolation with Arbitrary Loss Function

In C2, the challenge is extrapolating a quantile value for an event of interest Y given a loss function subjective to the practical limitations of the application. The probability of the quantile of interest, q_{C2} is defined as

$$\Pr(Y > q_{\text{C2}}) = \frac{1}{300T}, \tag{7.1}$$

where $T = 200$. Given that the annual cycle in Utopia is 300 days, the problem is concerned with finding the return value exceeded once in a return period of 200 years.

Practical concerns regarding the preparedness for the extreme occurrence Y dictate that underestimating the magnitude of the extreme event would be more costly than the expense incurred in the infrastructure necessary to prepare for an event that represents an overestimation. The loss function reflecting this asymmetric decision is described in Rohrbeck *et al.* (2023) as

$$L(q, \hat{q}) = \begin{cases} 0.9(0.99q - \hat{q}) & \text{if } 0.99q > \hat{q}, \\ 0 & \text{if } |q - \hat{q}| \leq 0.01q, \\ 0.1(\hat{q} - 1.01q) & \text{if } 1.01q < \hat{q}, \end{cases} \tag{7.2}$$

where q is the true quantile and $\hat{q}$ is an estimate of q. An illustration of the loss function is given in Figure 7.5.

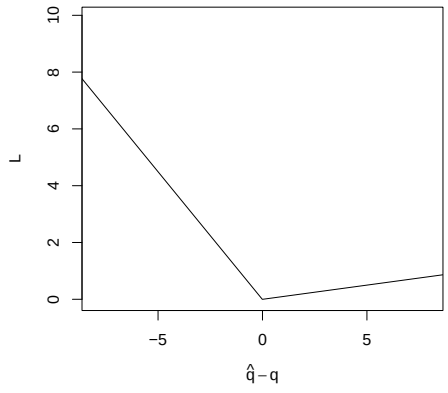

Figure 7.5: Visualisation of loss function in (7.2) for a true value of $q = 1$

7.3.1 Extreme Weighted Bootstrap for Extrapolation

The problem is well-posed for extreme value analysis, as extrapolating into the tail of the distribution and estimating a return value has been an attractive output of the field, particularly for heavy-tailed data as seen in the histogram in Figure 7.1. Return values can be estimated for the classical extreme distributions (GEVD and GPD) in the univariate case. For the GPD, the return value y_r is estimated for a return period r using

$$y_r = u + \frac{\sigma}{\xi}[(rp_u)^{\xi} - 1],$$

where $p_u = \Pr(Y > u)$ is the probability of exceeding the threshold u, and σ and ξ are the scale and shape parameters of the GPD, respectively. The accuracy of the estimation of any return value is dependent on various factors: (1) the quality of historical data, (2) the length of the record, (3) the characteristics of the true data-generating distributions, e.g., stationary vs nonstationary, (4) the appropriateness (or misspecification) of the statistical model, and (5) the model-fitting mechanism (Mackay and Jonathan, 2020). To address these issues, an approach inspired by Jonathan *et al.* (2021), who provided a systematic review of return value estimation and concluded that the best estimator is the mean of different quantile estimates for the annual maximum event, was proposed using a weighted data bootstrap model (Hall and Maesono, 2000) with a focus on extreme values. Bootstrapping data for the purpose of extreme value theory has been used before by de Haan and Zhou (2024) for improved extreme value estimators and Varga *et al.* (2016) who used it for improved uncertainty estimates of return values.

Hall and Maesono (2000) propose an extension of the classical bootstrap formulation where sampling weights, $w_i = \{w_1, ..., w_n\}$, are assigned to each observation in the original data Y_i, defining the sampling priority for that observation, resulting in a bootstrapped sample

$$\mathbf{Y}_n^* = (\underbrace{Y_1, \dots Y_1}_{w_1 * n \text{ times}}, \dots, \underbrace{Y_n, \dots Y_n}_{w_n * n \text{ times}}).$$

While various weighted strategies exist, only Varga *et al.* (2016) applied it in the extreme value context to improve uncertainty estimates of the estimation of return values of precipitation in Hungary. They proposed the multinomial or exponential distributions as the generating distributions for the weights, which were used to weight the contribution of each observation to the likelihood. The strategy was considered insufficient, as it failed to improve extrapolation and did not incorporate a loss function. Our focus on extremes and extrapolation requires a weighting strategy that assigns higher weights to extreme values for better representation, and that could be used in conjunction with the loss function in (7.2). For this reason, we chose to define the bootstrap sampling probabilities, $\mathbf{w}$, as a scaled version of $\arctan\big(y_{(i)}\big)$, where y_i is the i-th ordered observation, that results in an

"S" shape as seen in Figure 7.6. Specifically, the sampling weight w_i for sorted observation $y_{(i)}$, $i = 1, \ldots, n$, is

$$w_i = \frac{w_i^*}{\sum_{i=1}^{n} w_i^*}, \quad \text{where} \quad w_i^* = \frac{y_i^* - \min\{y_i^*\}}{\max\{y_i^* - \min y_i^*\}} \quad \text{and} \quad y_i^* = \arctan\left\{\Phi(\widehat{F}_Y^{-1}(y_i))\right\}, \tag{7.3}$$

where $\widehat{F}_Y(\cdot)$ is the empirical cdf of Y and $\Phi(\cdot)$ is the cdf of a standard normal distribution. Both the arctan and Φ, are required to obtain the desired "S" shape to assign small values a low probability of sampling and large values a high probability of sampling, as the representation of extreme values in each sample is important to capture the underlying extremal behaviour and improve extrapolation into the tail.

The bootstrapping procedure using the weights in (7.3) is used to produce B extreme-rich samples at every iteration. A GPD is fitted to each sample and used to predict the desired quantile value. The algorithm can be summarised as follows

1. For each iteration $b = \{1, ..., B\}$ sample a set of n observations with replacement using w_i in (7.3) as the probability of drawing the i-th quantile from the original data, $i = 1, \ldots, n$.

2. Fit a stationary GPD model, $\hat{F}_b$, to exceedances over the 99.5% empirical quantile ($p_u = 0.995$).

3. Predict the high quantiles $q_{(j)} < q_{\text{C2}}$, where $q_{(j)} > \hat{F}_Y^{-1}(p_u)$, $j = 1, ..., n_u$, corresponding to the j-th ordered observation over the threshold u_0. Using the fact that

$$\hat{F}_b^{-1}\left(\frac{\hat{F}_Y(q_{(j)}) - (1 - p_u)}{p_u}\right) = \hat{q}_{(j)},$$

   calculate $L(q_{(j)}, \hat{q}_{(j)})$ using (7.2) and compute the total loss for that sample $L_b = \sum_{j=1}^{n_u} L(q_{(j)}, \hat{q}_{(j)})$.

4. Predict the desired quantile q_{C2} using $\hat{F}_b$ to obtain $\hat{q}_{\text{C2}}$.

7.3.2 Application to C2

The procedure detailed above was repeated for $B = 1000$ iterations, resulting in a range of bootstrap predictions assessed by their respective total loss using the arbitrary loss function defined in (7.2). The right-hand side of Figure 7.6 shows these bootstrap predictions for quantiles exceeding the threshold $u = 119.33$, corresponding to the 99.5th empirical quantile. As seen from the figure, the predicted values have a large range, which increases at larger quantiles, meaning uncertainty around any estimate increases with higher predictions. The point estimate for the iteration yielding the smallest total loss L_b was proposed

as the final answer. In this application, it was $\hat{q}_{C2} = 239$ with 95% confidence interval $(169.2, 363.2)$, reflecting wide uncertainty around the estimation of $\hat{q}_{C2}$ under the arbitrary loss function.

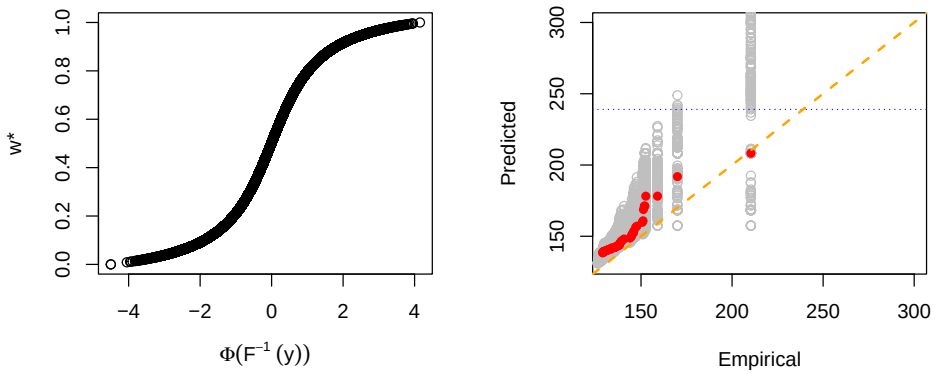

Figure 7.6: *Left:* Bootstrap sampling weights before rescaling for transformed observations y_i^* where extreme observations are more likely to be sampled at every bootstrap iteration. *Right:* Bootstrapped quantile predictions based on a GPD fitted on exceedances of the 99.5% empirical quantile. The red points represent the best prediction, i.e., the sample that minimises the loss function, and the blue line is our final prediction for q_{C2}.

7.4 C4 - Probability of High-Dimensional Simultaneous Extreme Event

In C4, the Utopian government required the probability that the variables $Y_{i,j}$ would exceed a threshold simultaneously, i.e.,

$$\Pr(Y_{i,j} > s_i : i_j = 1, ..., 50; j = 1, 2),$$

where i represents one of the 25 locations under the responsibility of the regional government U_j.

The challenge was further subdivided into two problems. In the first, a different design is allowed for each regional government, U_1 and U_2, where each region is assigned a different threshold, namely s_1 and s_2, respectively. Here, we consider s_1 to be the marginal level exceeded once in a year on average, equivalent to a probability of $\phi_1 = 1/300$ given that there are 300 days in the Utopian year, while s_2 is the level exceeded once a month, equivalent to a probability of $\phi_2 = 12\phi_1$. In this challenge, the value exceeded once a year is given as $s_1 = 5.702$, and the one exceeded once a month is $s_2 = 3.199$. The probability associated with the simultaneous exceedance of s_1 and s_2 at U_1 and U_2 is denoted as p_1,

and its estimate is $\hat{p}_1$. In the second problem, only the probability of a yearly exceedance is considered, p_2, so that $\phi_1 = \phi_2$ and $s_1 = s_2 = 5.702$.

7.4.1 PPCA and Application to C4

While there are methods capable of dealing with high-dimensional data for multivariate and spatial extremes (see Huser and Wadsworth, 2022), the choice of dependence structure can be restrictive. Figure 7.4 shows the estimated χ values for thresholds $u > 0.9$. As seen in the figure, the data display diverse dependence structures, with pairs of locations displaying strong dependence, weak dependence, asymptotic dependence, and asymptotic independence. The diverse dependence classes, coupled with practical limitations, were central to choosing a dimension-reduction approach, specifically, probabilistic principal component analysis (PPCA, Tipping and Bishop, 1999), which maps the existing extremal dependence structure onto an asymptotically independent setting through a latent Gaussian structure.

PPCA is an extension of principal component analysis (PCA) that assumes a probabilistic model. It is considered a dimensionality reduction method that assumes a lower-dimensional latent Gaussian model framework. Let $Y = \{\mathbf{y}_1, ..., \mathbf{y}_n\}$ where every $\mathbf{y}_i \in \mathbb{R}^d$ are d-dimensional vectors. The underlying structure is assumed to be linear, depending on some latent variable $\mathbf{z}_i \in \mathbb{R}^k$ where $k \leq d$, i.e.,

$$\mathbf{y}_i = \mathbf{W} z_i + \epsilon,$$

where $\mathbf{W}$ is $d \times k$ dimensional matrix of the principal component axes, and ϵ is a Gaussian error term with zero mean and σ^2 variance. By integrating the latent variable $\mathbf{z}_i$, we have that $\mathbf{y}_i \sim N(0, \mathbf{W}\mathbf{W}^T + \sigma^2 I)$, which in turn enables the estimation of $\mathbf{W}$ and σ^2 via maximum likelihood. PCA arises as a special case when $\sigma^2 = 0$. Predictions can then be made using a multivariate Gaussian distribution of dimension d. In PCA methods, there are as many principal components as there are variables. The first principal components are said to account for the most variance in the data and often carry information about the majority of the observations, while the smallest principal components contain information about outliers and extremes. For this reason, we chose to use $d = 25$ to estimate $\hat{p}_1$ and $d = 50$ to estimate $\hat{p}_2$. The model provided the point estimates $\widehat{p}_1 = 2.9 \times 10^{-9}$ for the first problem, and $\widehat{p}_2 = 5.4 \times 10^{-10}$ for the second.

7.5 Discussion and Future Work

A discussion of the results was facilitated by the reveal of the truth provided by the organisers in Rohrbeck *et al.* (2023). C2 posed a challenge focused on extrapolating extremes

in a univariate setting using an arbitrary loss function that reflected the limitations of the application. To tackle the problem, we proposed a weighted bootstrap approach with a tailored weighting strategy to prioritise the sampling of extreme values and mitigate the bias presented by non-extreme values. Applying the model resulted in a point estimate of $\hat{q_{C2}} = 239$ with 95% confidence interval $(169.2, 363.2)$. Rohrbeck *et al.* (2023) provided a true value of $q = 196.6$, which was overestimated by our modelling approach. We believe there are two major reasons for this behaviour: one conceptual reason and one technical.

The data provided for C2 spanned 70 years, while the question required estimating a value likely to be exceeded only once in 200 years (q_{C2}). Working with this information, we formed the assumption that an estimate of this quantile, $\hat{q}_{\text{C2}}$, would, therefore, be larger than the values observed in the data. Using this assumption as a guiding principle, we preferred models and techniques that met this assumption. This assumption proved erroneous, as the true value was revealed to be smaller than the maximum observed in the data.

The second reason is the weighting strategy. The assumption above encouraged a weighting strategy that resulted in the over-sampling of extremes. This over-sampling introduced bias and resulted in overestimating q_{C2}. Although the confidence intervals captured the true value, these were wide and thus included unreasonable predictions under the loss function. Future work on this approach involves a thorough search and comparison for weighting strategies that result in accurate point estimates and narrower confidence intervals.

C4 centred around estimating the probability of joint exceedance in a high-dimensional setting. As such, there was a strong focus on the extremal dependence between pairs in both regions. Our PPCA approach, which assumes a latent Gaussian structure, did not account for the dependence between locations. The results of the model yielded $\widehat{p}_1 = 2.9 \times 10^{-9}$ for the first problem, and $\widehat{p}_2 = 5.4 \times 10^{-10}$ for the second, both of which overestimated the true values $p_1 = 8.4 \times 10^{-23}$ and $p_2 = 5.4 \times 10^{-25}$, respectively. Figure 7.4 showed that the data had clusters of locations with different dependence structures. Unfortunately, time constraints did not allow us to pursue a modelling approach based on clusters of dependence structures, which was later revealed as the correct approach in (Rohrbeck *et al.*, 2023).

Chapter 8

Conclusions and Future Work

The research presented in this thesis develops novel methodologies to tackle the statistical modelling of spatial extremes with applications in heavy metal (HM) soil contamination and $PM_{2.5}$ air pollution. The new methodologies combine existing spatial approaches and extreme value theory to tackle non-replicated extremes for HM contamination and data fusion for extremes of $PM_{2.5}$ air pollution. Chapters 2 and 3 are provided as necessary background. Chapter 2 provides the statistical background used as the foundation of the proposed models, covering geostatistical models, extreme value theory, methods for Bayesian inference, and common data fusion approaches. In contrast, Chapter 3 provides the environmental context for the two pollution case studies.

The motivation behind modelling bivariate heavy metal contamination in the Glasgow Conurbation (Chapters 4 and 5) was to improve the spatial estimates of the extreme values of contaminant concentrations while accounting for the extremal dependence between them. The project was undertaken in collaboration with the British Geological Survey, which kindly provided the G-BASE data set for the application. First, Chapter 4 provides an in-depth comparison of the extremal dependence between contaminant pairs under two models with different extremal dependence structures, ignoring spatial variability. Chapter 5 develops a methodology to model these contaminants in space by proposing a bivariate coregionalised mixture model fitted using INLA under the Bayesian inference framework.

In Chapter 6, a data fusion model tailored for extreme values was developed in Chapter 6, building on the Gaussian approach of Wilkie *et al.* (2019). The project was motivated by the need to improve the representation of extreme values in remote-sensing or modelled data by fusing them with spatially sparse in-situ measurements using a Bayesian hierarchical model.

Finally, Chapter 7 provides the work presented at the EVA 2023 Data Challenge. The chapter covers two challenges, C2 and C4, where C2 focuses on extrapolating a high quantile for a univariate random variable to minimise an arbitrary loss function. C4, on the other hand, is about estimating the probability of simultaneous exceedance in a high-

dimensional setting. Both problems represent common challenges in modelling extreme values.

8.1 On Modelling Bivariate Heavy Metal Soil Contamination

8.1.1 Extremal Dependence between Contaminants

In Chapter 4, we present an investigation of the extremal dependence structures in heavy metal contaminants in the Glasgow Conurbation by using two existing models with different extremal dependence structures - the multivariate generalised Pareto distribution (MGPD) using the approach proposed by Kiriliouk *et al.* (2019) and the exponential factor copula model (EFC) proposed by Castro-Camilo and Huser (2020). These models represent two distinct classes of extremal dependence. The MGPD produces a constant dependence, meaning the probability of simultaneously experiencing high values in both contaminants is always unchanged, even for very extreme values. Conversely, the EFC is a subasymptotic model capable of capturing decaying dependence - meaning the dependence between components decreases at increasing values but does not reach asymptotic independence.

Both models were fitted to the fifteen possible pairs between As, Cr, Cu, Ni, Pb, and Zn. The results shine a light on the diversity in extremal dependence structures between these contaminants in the Glasgow Conurbation and provide insight into their possible sources. The empirical dependence between pairs with As shown in Figure 4.2, i.e., As-Cu, As-Ni, As-Pb, and As-Zn, show strong, near constant dependence ($\chi \approx 0.5$), only displaying a decline at high quantiles ($u > 0.95$). As such, these pairs are better represented by the rigid dependence of the MGPD. The EFC is a better fit only at very high quantiles, $u > 0.95$, when the decay in χ is evident. The results imply that As is a common joint byproduct of processes that release Cu, Ni, Zn, and Pb into the soil. Only the As-Cr pair exhibits a different extremal dependence structure altogether - it is weaker ($\chi < 0.4$) at lower quantiles and decays consistently at increasing quantiles. For this pair, the MGPD captures dependence only at lower quantiles and the EFC only at very high quantiles. However, both models overestimate the dependence at most quantiles and fail to capture it appropriately, showing that alternative dependence models could be needed.

The dependence between pairs with Cu shown in Figure 4.3, i.e., Cu-Ni, Cu-Pb, and Cu-Zn, is strong ($\chi > 0.6$) and displays minor decay until larger quantiles. Although the dependence between Cu-Ni and Zn-Cu is nearly constant, Zn-Cu is better captured by the MGPD, and Cu-Ni is better captured by the EFC. The pair Pb-Cu displays weak decay and is better captured by the EFC. The strong dependence between most of these

pairs indicates a common source of contamination and possibly similar mobility pathways, as their dependence remains high for the full range of observations. Only the pair Cr-Cu exhibits strong decay and approaches asymptotic independence. The EFC model accomplishes a close fit to the dependence throughout the range of observations. As such, it provides evidence that Cr and Cu may not come from a common source or might have different mobility pathways in the region.

For the pairs in Figure 4.4, that is, the remaining pairs of Cr, Ni, and Zn not covered previously, various dependence structures are visible. Pairs with Cr, i.e., Cr-Ni, Pb-Cr, and Cr-Zn, experience decaying dependence structures. Cr-Ni displays the strongest decay from $\chi = 0.58$ to $\chi = 0.05$, which is only captured by the EFC at $u > 0.925$, while the MGPD model performs poorly. The other two pairs, Pb-Cr and Cr-Zn, have decaying structures better captured by the EFC model. The remaining three pairs in the figure, i.e., Pb-Ni, Pb-Zn, and Zn-Ni, all display very weak decay or near-constant dependence. The Pb-Ni pair, however, is mainly overestimated by both models, with only EFC capturing dependence at $u > 0.92$. The Pb-Zn pair shows strong dependence $\chi = 0.7$, with decay only at high quantiles and is similarly captured by both models. Finally, the EFC better captures the Zn-Ni pair, although near-constant dependence can be seen up to $u = 0.95$. From the results, it can be seen that Cr-Ni have a strong relationship that decays rapidly, providing evidence of different sources of extreme values or mobility capacity at elevated values in the region.

The results from these comparisons provide three main conclusions. First, dependence structures in heavy metal contaminants in the soil are diverse. While it is commonly known that decaying dependence is a common feature in environmental applications (Castro-Camilo and Huser, 2020), each pair should be examined individually, as models with rigid dependence structures, such as the MGPD, can prove useful. Contrastingly, despite the EFC's more flexible dependence structure, it is still insufficient to capture some of the decaying dependence found between these contaminants, but it is generally better at capturing the decay at high quantiles ($u > 0.95$). Finally, the results show that Cr has a different extremal behaviour from most other contaminants. Chen *et al.* (2024) further supports this conclusion by showing that Cr has a different leaching behaviour than other heavy metal contaminants, being retained in the soil for longer periods and unaffected by normal or extreme precipitation events. These results show the importance of investigating extremal dependence in an exploratory phase, as a spatial model for these applications must consider the different joint tail behaviour of specific contaminant pairs.

8.1.2 Bivariate Coregionalised Mixture Model

In Chapter 5, a spatial model for bivariate heavy metal contaminants was developed. It combines two different modelling frameworks - mixture models and coregionalised models.

The mixture model is a natural approach to the heavy-tailed distributions of heavy metal concentration in the soil because it exploits the difference between baseline and trace concentrations, which comprise most of the data, and the extreme values, i.e., the tail of the distribution. The coregionalisation framework, on the other hand, enables joint modelling of the contaminants. The chapter proposes constructing the coregionalised structure that models the contaminant bodies independently using covariates in a linear predictor. The marginal tails are first modelled using a generalised Pareto distribution (GPD) and then transformed into a Gaussian distribution. After they are transformed, they are modelled using linear predictors that share a scaled random spatial effect. The scaling coefficient regulates the dependence between tails, thus accommodating the extremal dependence between components.

A comparison between the model results and the conventional kriging approach shows that the proposed coregionalised mixture model is better at capturing the extreme values and accounts for extremal dependence. While both models provide smooth concentration maps for the marginal contaminants, the coregionalised mixture model can also provide maps for the probabilities of joint exceedance that account for the extremal dependence between components. A limitation of the model is that the membership of each observation as belonging to the body or the tail must be assigned a priori, potentially introducing bias in the modelling process.

Maps of the probability of exceeding soil guideline values for residential use with growing produce (SGV1) and residential use without produce (SGV2) are shown. The probability of observing joint exceedances of the SGV1 is highest directly south of the river Clyde in the city centre, followed by areas immediately to the east of the city centre, including two villages, and a small region to the west. The contaminated areas near the city centre are historical industrial zones well known for industrial metal-smelting and ship-building. The villages to the east of Glasgow have long histories of mining work and ore processing, which are known to have produced high concentrations of heavy-metal byproducts. The probabilities of exceedance north of the river Clyde are generally low, which, unlike the south of the Clyde, did not have such a vigorous industrial activity, providing further evidence of historical industrial activity as the source of the contamination. The probabilities of exceeding SGV2 are generally lower and only present south of the Clyde near the city centre of Glasgow, areas known for high historical industrial activity. As expected, the maps of uncertainties show that the model performs better in areas with higher sample densities.

The maps of the probability of the joint exceedance of soil guideline values are useful outputs of this model. Unlike their Gaussian counterparts, these maps account for the heavy-tailed distributions of contaminants and are suitable for extrapolation using extreme value theory. As a result, they are better for risk assessment than the non-extreme

alternatives. This is particularly useful for policymakers and urban planners, who require this information to plan appropriate land use and mitigate the effects of heavy metal soil contamination in the Glasgow Conurbation.

8.1.3 Future Work

The research undertaken for this application consists of investigating the dependence structure of heavy metal soil contaminants and a novel bivariate spatial approach for unreplicated observations that combines extreme value theory, mixture models, and the coregionalisation framework. While the coregionalised mixture model was motivated by this application, it can be used in other applications where extreme values are of outmost importance and no temporal replicates are available. Methodologically, we suggest some avenues for possible future work.

- A flexible and joint estimation of p_1 and p_2, the mixture parameters for the two contaminants. The parameters p_1 and p_2 refer to the proportion of observations that belong to the body of the distributions. By contrast, $1 - p_1$ and $1 - p_2$ define the proportion of observations in the tail, defining the number of extreme observations. In the current approach, p_1 and p_2 are fixed a priori by performing a grid search over possible values and selecting the best-performing parameter values under model selection metrics. While this is sufficient for the initial development of the model presented here, it can be unrealistic in some applications, as the probability of observing an extreme value depends on the location. Developing spatially-variant p_1 and p_2 could provide more accurate results, but the lack of temporal replication could produce identifiability problems. Additionally, incorporating the estimation of p_1 and p_2 into the modelling framework would help with uncertainty quantification and provide more information about the classification of the observations as body or tail.

- Exploring the integration of different extremal dependence structures into the model. While the extremal dependence between contaminants was explored separately from the spatial modelling, and the spatial modelling incorporates extremal dependence between components through latent structures, it would be useful to explore alternative extremal dependence structures within the class of asymptotic dependence and independence models.

- Exploring different coregionalisation structures. The model presented here only accounted for dependence through one shared component between contaminant tails. While this is an elegant solution to account for extremal dependence, coregionalisation in INLA is flexible, and additional constructions can be explored accounting

for dependence in other ways. The flexibility of these constructions, however, will be constrained by the lack of temporal replications.

8.2 On Data Fusion for $PM_{2.5}$ Pollution Extremes

Chapter 6 extended the Gaussian model of Wilkie *et al.* (2019) to propose a data fusion approach tailored to extreme values defined as exceedances of the 80th-quantile marginal threshold. The aim was to enhance threshold exceedances of $PM_{2.5}$ in the EAC4 in the Greater London area to better represent in-situ observations, as those provided by the spatially-sparse AURN network. The process consisted of two main steps. First, at each location, the data from EAC4 and AURN, x and y, respectively, were preprocessed to censor non-threshold exceedances at 0. In the second step, the model is fitted to the data. The likelihood for both data sources is the Dirac-delta generalised Pareto distribution (δ-GPD), a variant of the GPD that includes one additional parameter, p, which denotes the probability of observing an exceedance. Under this likelihood, a non-threshold exceedance has a density of $1 - p$, while a threshold exceedance has a density under the GPD density scaled by p. In this way, the δ-GPD incorporates information about the non-threshold exceedances in a way that the GPD cannot. This mimics the so-called point process representation of extremes, where the time of occurrence and sites of exceedances are approximated by a bivariate point process (Coles, 2001, Ch. 7). Other than the incorporation of the additional parameter p, the δ-GPD is similar to the GPD, having a shape (ξ), scale (σ), and location $\mu = 0$ parameter. In this model, p is estimated using a logistic regression approach inside a hierarchical model structure. It estimates $p_{y_i t}$, the probability of observing an exceedance at location i and time t, using the occurrence of an extreme at $x^*_{i(t-1)}$, x^*_{it}, and $x^*_{i(t+1)}$ so that lagged EAC4 information can be considered. The results from this part of the model showed that only x^*_{it} has a significant influence on $p_{y_i t}$, with the regression coefficients for $x^*_{i(t-1)}$ and $x^*_{i(t+1)}$ being statistically insignificant. These results indicate that the lagged presence of an exceedance carries little information, and an exceedance in the AURN data is mostly informed by an exceedance in the EAC4 data, limiting our model to the accuracy of the EAC4 data at predicting an exceedance.

The shape parameters, ξ_x and ξ_y, are constant across time and space but modelled separately for each data source. The mode of the posterior distributions gives negative values, where $\xi_x \approx -0.25$ and $\xi_y \approx -0.15$. The data fusion occurs in the estimation of the scale parameters. A basis function of d dimensions provides the temporal structure to estimate a different scale parameter at each time point for the EAC4 data, σ_{xt}. This estimated scale parameter at time t is then used in a regression-type model to estimate the GPD scale of the AURN data, σ_{yt}, using the same basis function. Therefore, the model assigns a GPD distribution at each time point of y where ξ_y is kept constant, and the σ_{yt}

is modelled using a flexible basis function and σ_{xt} as a covariate.

The model is fitted using MCMC via Metropolis-Hastings, which enables the exploration of the whole parameter space. Predictions can then be used to interpolate missing observations in the AURN data or enhance EAC4 data to approximate AURN observations for locations where no AURN observation station exists. A leave-one-site-out cross-validation procedure (LOSO-CV) compares the predictive ability of the data fusion for extremes model (ExDF), the Gaussian approach (GausDF) of Wilkie *et al.* (2019), and the EAC4 data to mimic the threshold exceedances measured by the 12 AURN observation stations available in the area. They show that the ExDF is better at capturing the threshold exceedances of the in-situ measurements at 10 of the 12 locations, outperforming the GausDF and EAC4 models.

Finally, maps of the expected shortfall for the ExDF and the EAC4 models are given. The maps show that the ExDF data have greater variability, values closer to those observed in the nearby AURN stations, and that a different spatial pattern is visible with higher $PM_{2.5}$ concentrations near the coast.

8.2.1 Future Work

While the model presents an improvement over its Gaussian counterpart, there are various directions for possible future work.

- Improved estimation of p. Currently, the model relies on exceedances in the EAC4 data to indicate a threshold exceedance in the predicted in-situ measurements. Work in Section 6.3.3 shows that the occurrence of an exceedance in the EAC4 is not a perfect estimator of exceedances in the AURN data. The problem can be addressed in various ways. First, it is possible to consider non-threshold exceedance information of the EAC4 data. As the two data sources are highly correlated, high but not extreme values in the EAC4 data generally indicate high values in the AURN data. This information is lost during the censoring process, so exploring alternative ways to utilise it could improve the estimation of p. The second is the incorporation of covariates, known to have an influence on $PM_{2.5}$, to improve the estimates of p. Meteorological, geographical, or ecological variables, such as relative humidity, elevation, and NDVI are known to affect $PM_{2.5}$ and easily-accessible from remote sensing sources.

- A body-tail approach to fuse the complete distribution. While including non-threshold exceedance information could prove beneficial in estimating p, a valuable extension of this work is the fusion of the entire distribution, not just the tail. If a GPD likelihood is retained for threshold exceedances, the definition of a suitable threshold

and the transition between body and tail must be considered. Body and tail approach incorporating extreme value distributions are not new in the literature and are sometimes called extreme mixture models (see, e.g., the review by Scarrott and MacDonald, 2012). An alternative approach is the extended GPD model proposed by Naveau *et al.* (2016).

- Increasing the temporal dimension. In the case study demonstrated for this chapter, only the year 2022 was considered. Extending the period of interest might prove beneficial for the model fit, as more threshold exceedances are observed. Additionally, long-term trends of exceedances could be captured providing insights into the extremal behaviour of $PM_{2.5}$ across years. This, however, is dependent on the efficiency of the model fitting process, which is limited in the current set up.

- Improving computational efficiency. The code for the model was written in C++ to improve running time. While this was a significant step in the feasibility of the model fitting, work can be done to make this model more efficient. This can be done through several mechanisms. First, exploring localised inference approaches could allow parallelisation of the model, effectively reducing computation time. Second, moving from CPU usage to GPU could prove invaluable and reduce the need for CPU resources.

8.3 On the EVA 2023 Data Challenge

The EVA 2023 data challenge was part of the EVA 2023 conference in Milan, Italy. The organisers posed 4 challenges, C1 to C4 (Rohrbeck *et al.*, 2023). This thesis covered only C2 and C4, as they are the author's contribution to a team entry.

In C2, the challenge comprised estimating an extrapolatory high quantile while minimising an arbitrary loss function that incorporated application-specific information. To tackle this challenge, we proposed using an extreme-weighted bootstrap approach. The weights in the weighted-bootstrap defined the sampling probability in the bootstrap sampling, and were assigned so that extreme values had a higher probability of being sampled, and low values had a small probability of being sampled. A GPD was fitted for each bootstrap sample, and the arbitrary loss function was used to calculate the observed quantiles. The total loss was obtained by summing the estimated loss for the observed quantiles, and the sample that minimised the loss was then used to estimate the desired extrapolatory quantile. The method had mixed success. The confidence intervals captured the true value, but the point estimate overestimated it.

For C4, the organisers asked participants to estimate the probability of observing joint exceedances at 50 locations. The first part of the challenge was calculating the

probability that twenty-five of these locations exceeded some safety threshold s_1 while the rest exceeded a different threshold s_2 at the same time. The second part was estimating the probability of all fifty sites exceeding s simultaneously. Due to time constraints, we used probabilistic principal component analysis (PPCA), a fast dimension reduction method that does not rely on EVT. Due to the dependent structures in the data and the independence assumption of PPCA, the estimated probabilities overestimated the true probabilities.

8.3.1 Reflections

- Exploring different weighting strategies. The extreme-weighted bootstrap approach for C2 had mixed success by providing a point-estimate that was much larger than the true value but capturing the true value in the 95% confidence intervals. An exploration of different weighting strategies could prove beneficial with the aim of improved estimation of extrapolatory quantiles and uncertainty around the estimation, which would also produce narrower confidence intervals.

- Exploring PCA approaches for extreme value analysis for C4. Drees and Sabourin (2021) proposes an asymptotically independent framework for PCA of extreme values. Under a multi-stage framework, the stage following marginal modelling could identify clusters as the sets of locations that are asymptotically dependent. The Drees and Sabourin (2021) model could then be applied to model inter-cluster dependence under the assumption of asymptotic independence.

Appendix A

Appendix for Chapter 5: Results of Simulation Study

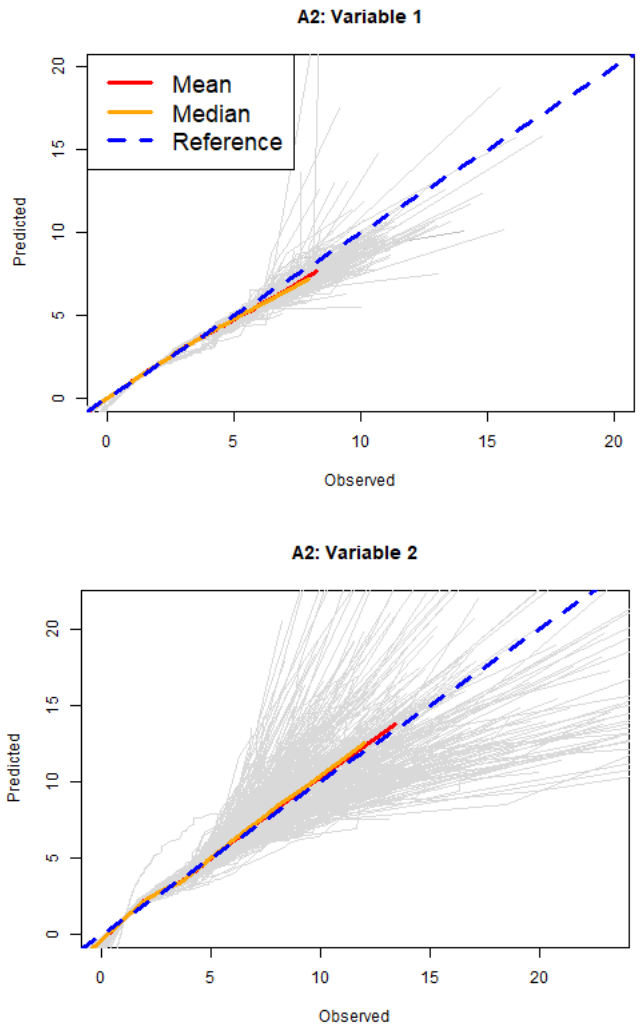

Figure A.1: Q-Q plots of all simulations (grey) for simulation study A2, for variables 1 and 2. The mean and median of the simulations are shown in red and orange, respectively, while the reference line is in blue.

Table A.1: Summary of results for simulation study A2. The table shows the parameter's true value; estimated parameter mean, median and standard deviation; the 95% coverage probability; and the mean RMSE and MAE.

Parameter	True Val	Median	Mean	Sd	Coverage.pr	RMSE	MAE
α_{A1}	1.00	1.01	1.01	0.07	0.95	0.07	0.06
α_{T1}	0.00	-0.06	-0.05	0.06	0.99	0.09	0.07
α_{A2}	1.00	1.01	1.01	0.07	0.94	0.07	0.06
α_{T2}	0.00	-0.06	-0.06	0.07	0.99	0.09	0.07
β_{A1_1}	0.10	0.08	0.08	0.03	0.99	0.03	0.03
β_{A1_2}	0.25	0.20	0.20	0.02	0.99	0.05	0.05
β_{T1_1}	0.10	0.03	0.03	0.10	0.98	0.12	0.10
β_{T1_2}	0.25	0.08	0.08	0.06	0.82	0.18	0.17
β_{A2_1}	0.10	0.08	0.08	0.03	0.99	0.03	0.02
β_{A2_2}	0.25	0.20	0.20	0.02	0.99	0.05	0.05
β_{T2_1}	0.10	0.03	0.03	0.09	0.99	0.12	0.10
β_{T2_2}	0.25	0.08	0.08	0.07	0.81	0.18	0.17
τ_1	1.00	2.52	0.96	4.80	0.79	5.03	3.36
τ_2	1.00	2.77	1.66	4.50	0.79	4.83	3.33
ρ_{z1}	5.00	5.08	5.20	1.22	0.98	1.22	0.80
ρ_{z3}	10.00	14.99	15.18	1.26	0.97	5.14	4.99
ρ_{z3}	5.00	3.49	3.03	2.10	0.99	2.59	2.16
ρ_{z4}	15.00	14.42	13.91	2.34	0.99	2.41	2.07
λ	0.90	0.91	0.88	0.22	0.92	0.27	0.20
ξ_1	0.05	0.12	0.13	0.13	0.94	0.15	0.12
ξ_2	0.25	0.32	0.32	0.13	0.96	0.15	0.12

Table A.2: Classification metrics including accuracy, precision, sensitivity, and specificity for simulation scenarios A1 and A2.

Scenario	Variable	Accuracy	Precision	Sensitivity	Specificity
A1	1	0.89	0.95	0.58	0.99
	2	0.89	0.96	0.57	0.99
A2	1	0.91	0.77	0.62	0.97
	2	0.91	0.78	0.62	0.98

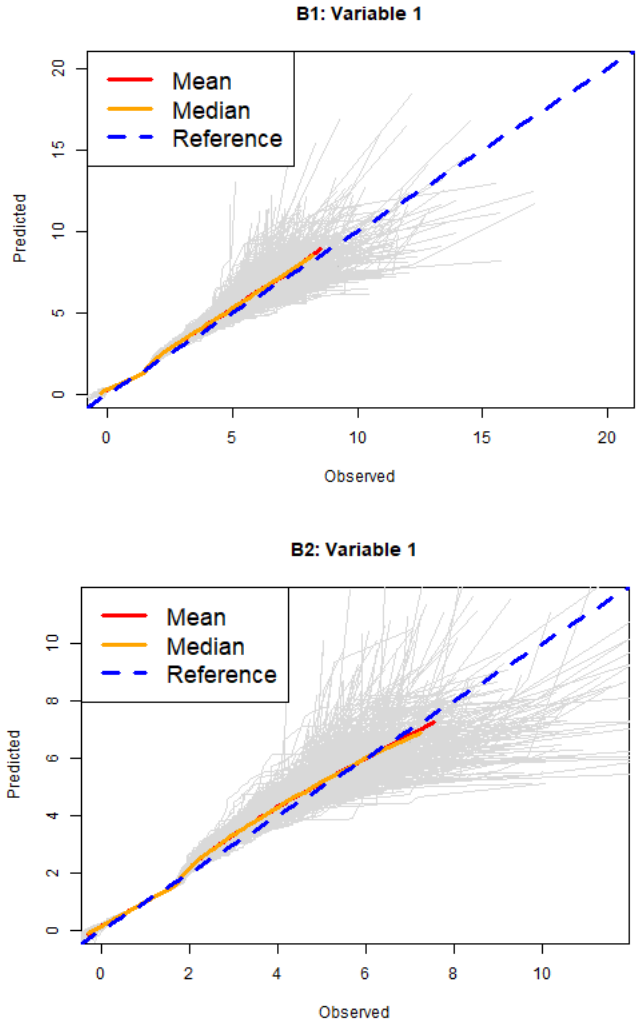

Figure A.2: Q-Q plots of simulations (grey) for simulation scenario B1, for variables 1 and 2. The mean and median of the simulations are shown in red and orange, respectively, while the reference line is given in blue.

Table A.3: Summary of results for simulation scenario B1. The table shows the parameter's true value; estimated parameter mean, median and standard deviation; the 95% coverage probability; and the mean RMSE.

Parameter	True Val	Median	Mean	Sd	Coverage.pr	RMSE	MAE
α_{B1}	1.00	0.93	0.93	0.07	0.91	0.10	0.08
α_{T1}	0.00	-0.11	-0.11	0.05	0.93	0.12	0.11
α_{B2}	1.00	0.93	0.93	0.07	0.88	0.10	0.08
α_{T2}	0.00	-0.10	-0.10	0.05	0.99	0.11	0.10
β_{B1_1}	0.10	0.05	0.05	0.03	0.99	0.06	0.05
β_{B1_2}	0.25	0.14	0.14	0.02	0.99	0.12	0.11
β_{T1_1}	0.10	0.05	0.05	0.06	0.99	0.08	0.06
β_{T1_2}	0.25	0.13	0.13	0.04	0.92	0.13	0.12
β_{B2_1}	0.10	0.06	0.06	0.03	0.99	0.05	0.05
β_{B2_2}	0.25	0.14	0.14	0.02	0.94	0.11	0.11
β_{T2_1}	0.10	0.04	0.04	0.06	0.99	0.09	0.07
β_{T2_2}	0.25	0.09	0.10	0.04	0.71	0.16	0.16
τ_1	1.00	2.71	1.31	3.91	0.72	4.27	1.71
τ_2	1.00	1.69	1.32	0.92	0.85	1.14	0.69
ρ_{z1}	5.00	4.43	2.77	3.75	0.99	4.67	2.73
ρ_{z3}	10.00	11.32	10.32	2.08	0.99	2.82	1.37
ρ_{z3}	5.00	5.63	5.28	3.96	0.99	4.01	3.03
ρ_{z4}	15.00	19.16	17.82	6.19	0.99	10.41	6.08
λ	0.90	0.85	0.86	0.26	0.83	0.26	0.22
ξ_1	0.05	0.09	0.09	0.07	0.92	0.08	0.07
ξ_2	0.25	0.29	0.29	0.09	0.92	0.10	0.08

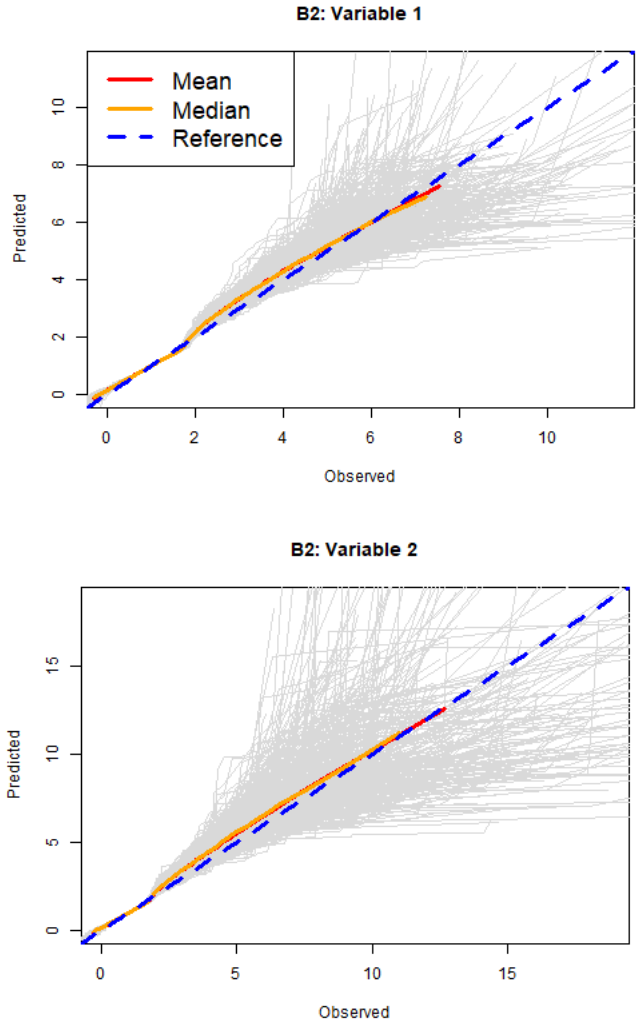

Figure A.3: Q-Q plots of all simulations of scenario B2, for variables 1 and 2. The mean and median of the simulations are shown in red and orange, respectively, while the reference line is given in blue.

Table A.4: Summary of results for simulation scenario B2. The table shows the parameter's true value; estimated parameter mean, median and standard deviation; the 95% coverage probability; and the mean RMSE.

Parameter	True Val	Median	Mean	Sd	Coverage.pr	RMSE	MAE
α_{B1}	1.00	1.00	1.00	0.07	0.99	0.07	0.05
α_{T1}	0.00	0.36	0.36	0.14	0.93	0.38	0.36
α_{B2}	1.00	1.01	1.01	0.07	0.99	0.07	0.06
α_{T2}	0.00	0.31	0.30	0.15	0.84	0.34	0.31
β_{B1_1}	0.10	0.08	0.09	0.02	0.99	0.03	0.02
β_{B1_2}	0.25	0.20	0.20	0.02	0.99	0.05	0.05
β_{T1_1}	0.10	0.03	0.03	0.10	0.99	0.12	0.10
β_{T1_2}	0.25	0.08	0.08	0.07	0.82	0.18	0.17
β_{B2_1}	0.10	0.08	0.09	0.03	0.99	0.03	0.02
β_{B2_2}	0.25	0.21	0.21	0.02	0.99	0.05	0.04
β_{T2_1}	0.10	0.02	0.01	0.09	0.98	0.13	0.10
β_{T2_2}	0.25	0.03	0.03	0.06	0.55	0.23	0.22
τ_1	1.00	1.19	0.33	2.57	0.99	2.57	1.26
τ_2	1.00	1.44	0.40	3.14	0.97	3.17	1.40
ρ_{z1}	5.00	4.83	5.02	1.26	0.96	1.27	0.82
ρ_{z3}	10.00	14.80	14.92	1.26	0.96	4.96	4.80
ρ_{z3}	5.00	3.75	3.34	2.24	0.98	2.56	2.08
ρ_{z4}	15.00	17.99	18.04	1.15	0.99	3.20	2.99
λ	0.90	1.09	1.05	0.22	0.88	0.29	0.22
ξ_1	0.05	0.10	0.11	0.20	0.96	0.20	0.16
ξ_2	0.25	0.31	0.33	0.22	0.95	0.23	0.18

Table A.5: Classification metrics including accuracy, precision, sensitivity, and specificity for simulation scenarios B1 and B2.

Scenario	Variable	Accuracy	Precision	Sensitivity	Specificity
B1	1	0.88	0.92	0.55	0.98
	2	0.88	0.93	0.56	0.99
B2	1	0.93	0.61	0.91	0.93
	2	0.93	0.61	0.91	0.93

Appendix B

Appendix for Chapter 6: Diagnostic trace and density plots

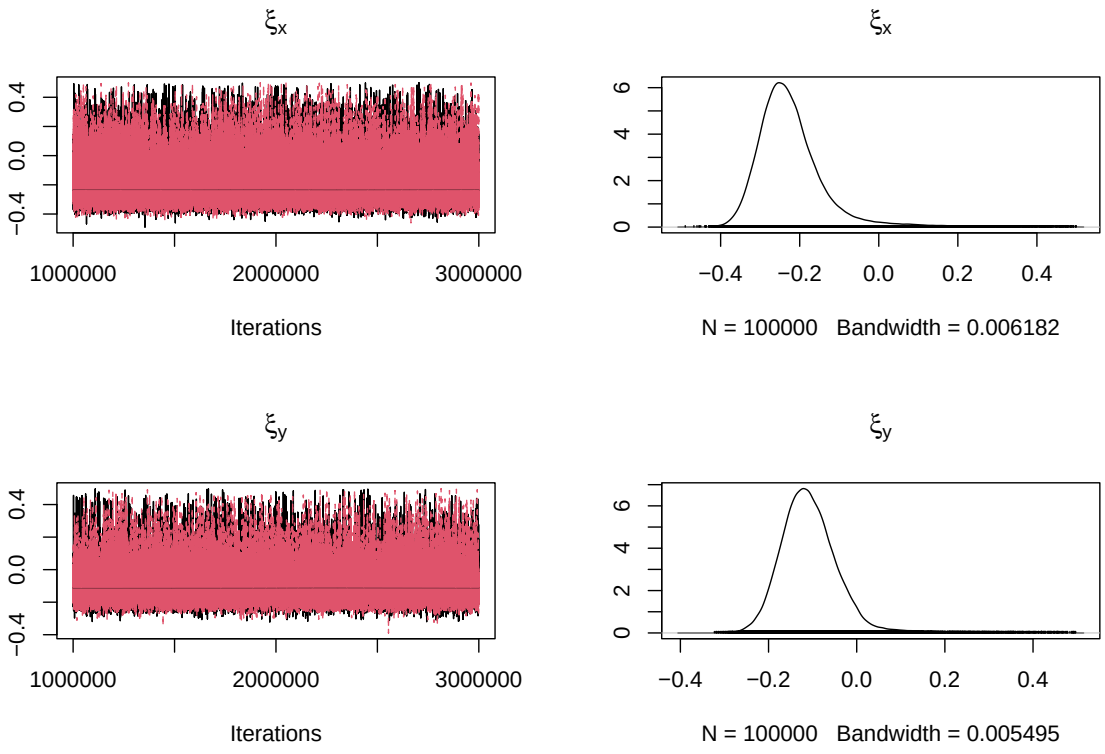

Figure B.1: Trace and density plots for the shape parameters ξ_x and ξ_y showing two different chains in black and red.

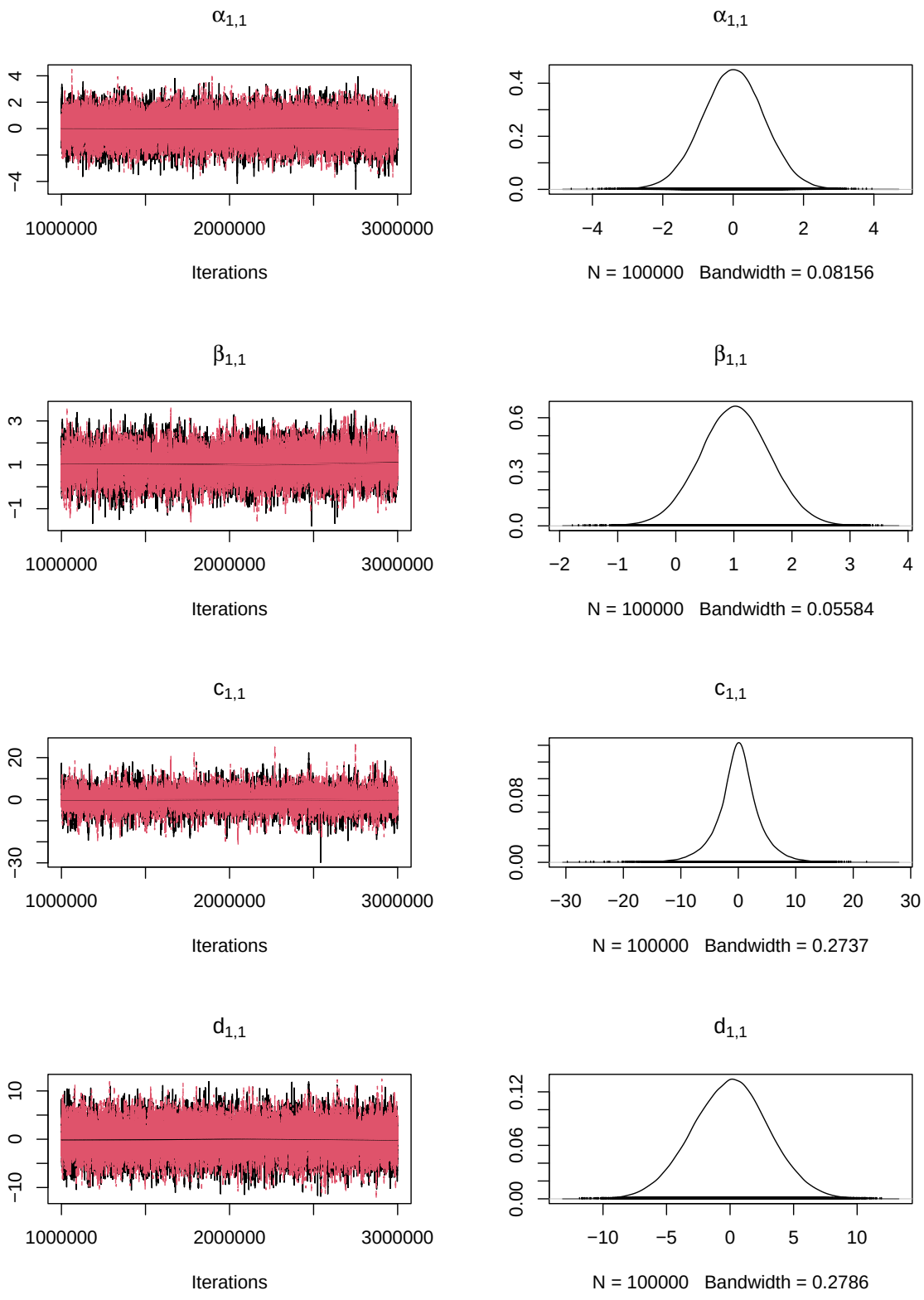

Figure B.2: Trace and density plots for parameters $\alpha_{1,1}, \beta_{1,1}, c_{1,1}$ and $d_{1,1}$ showing two different chains in black and red.

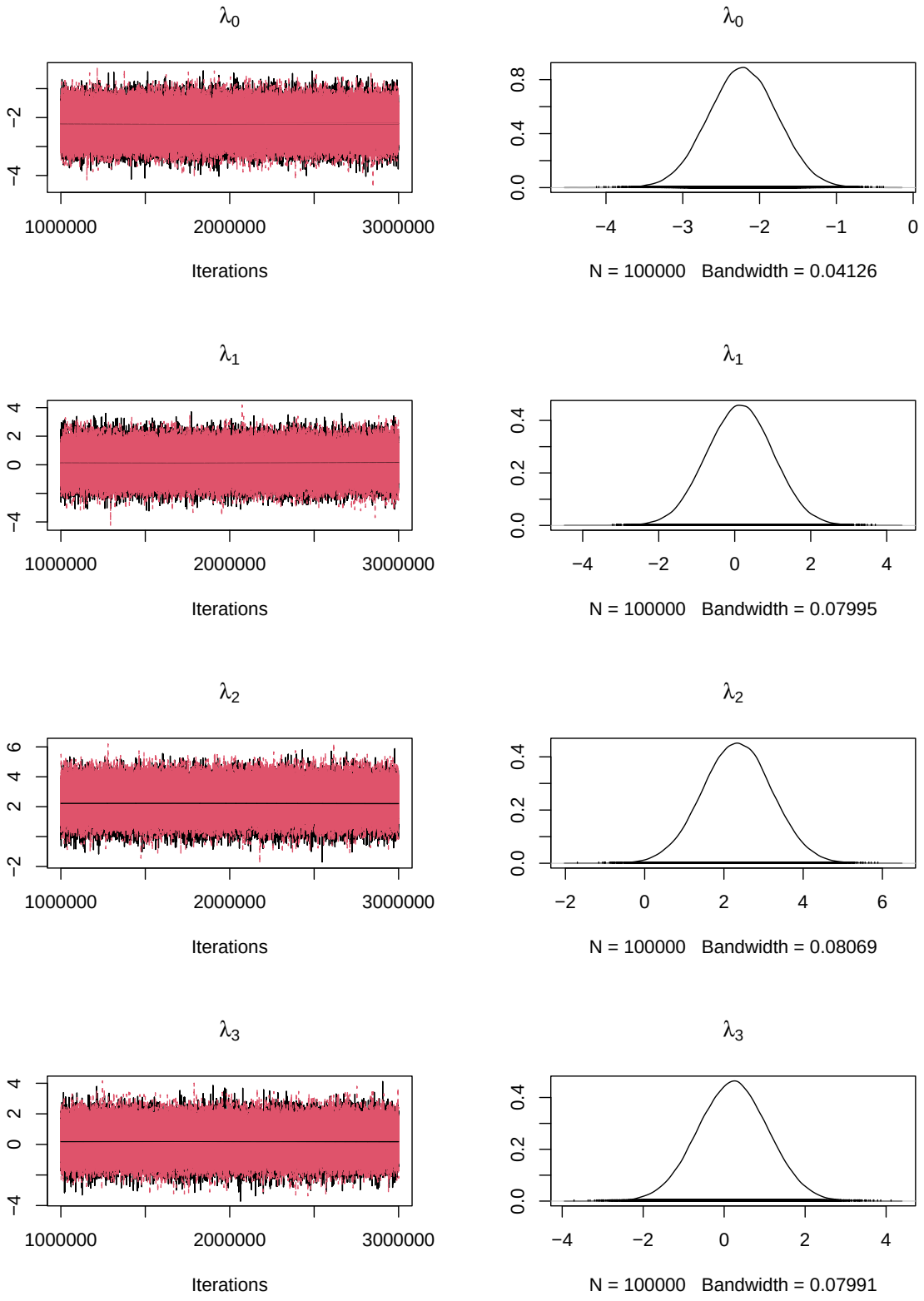

Figure B.3: Trace and density plots for parameters λ_0, λ_1, λ_2, λ_3 for Site A showing two different chains in black and red.

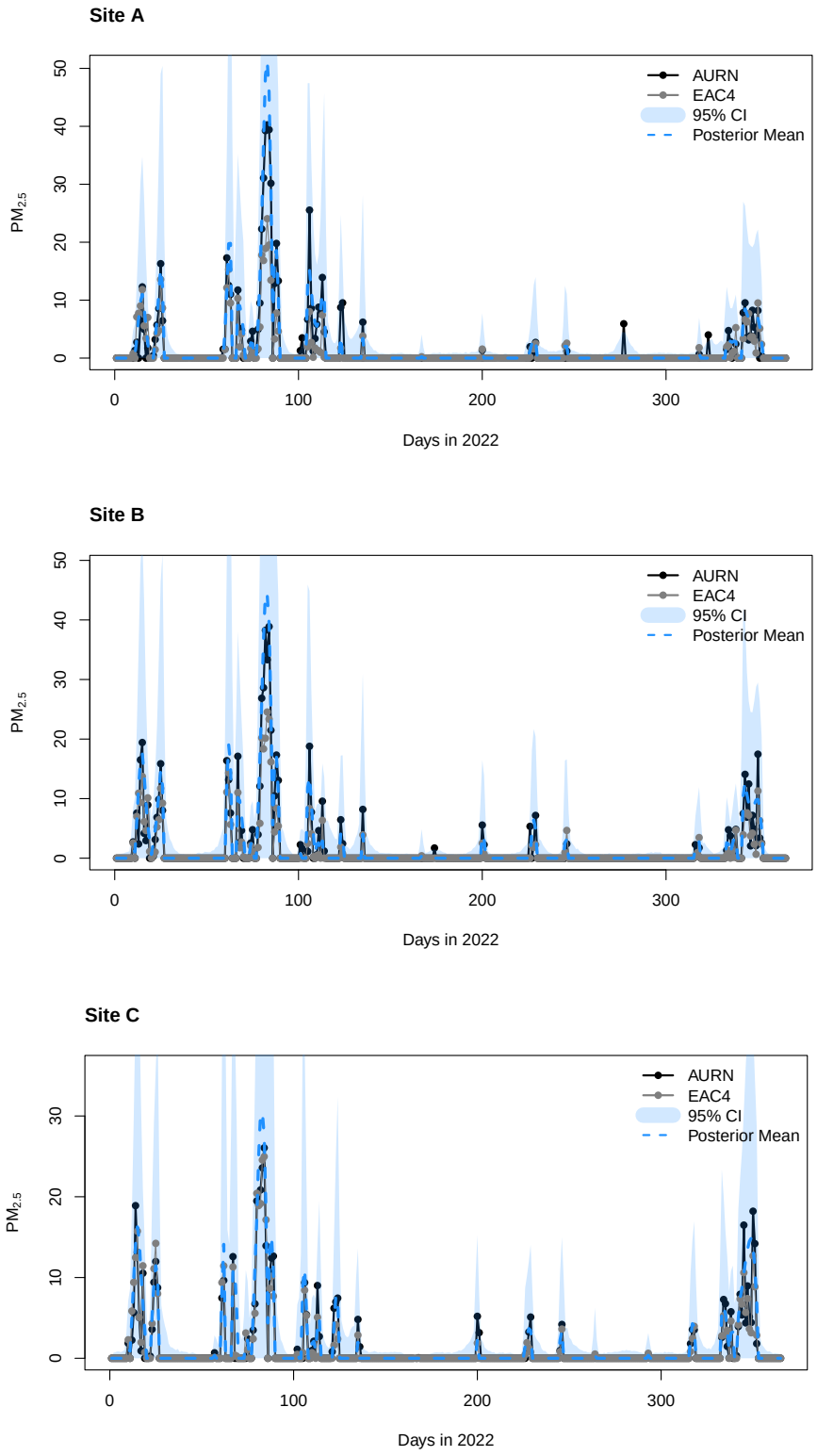

Figure B.4: Times series of fitted values and confidence bands of the ExDF model, along with the true AURN measurements and EAC4 observations for sites A, B,and C.

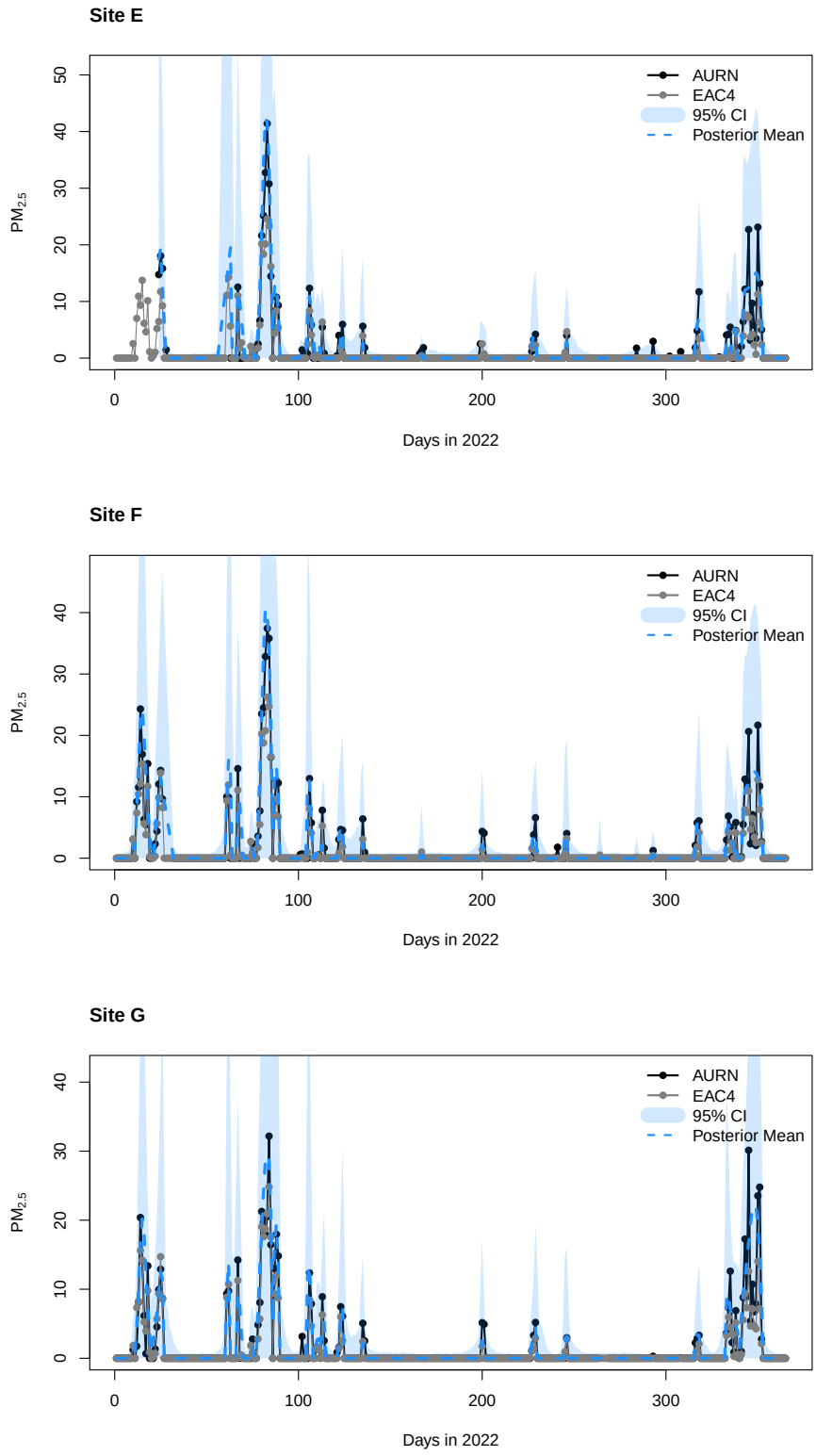

Figure B.5: Times series of fitted values and confidence bands of the ExDF model, along with the true AURN measurements and EAC4 observations for sites E, F, and G.

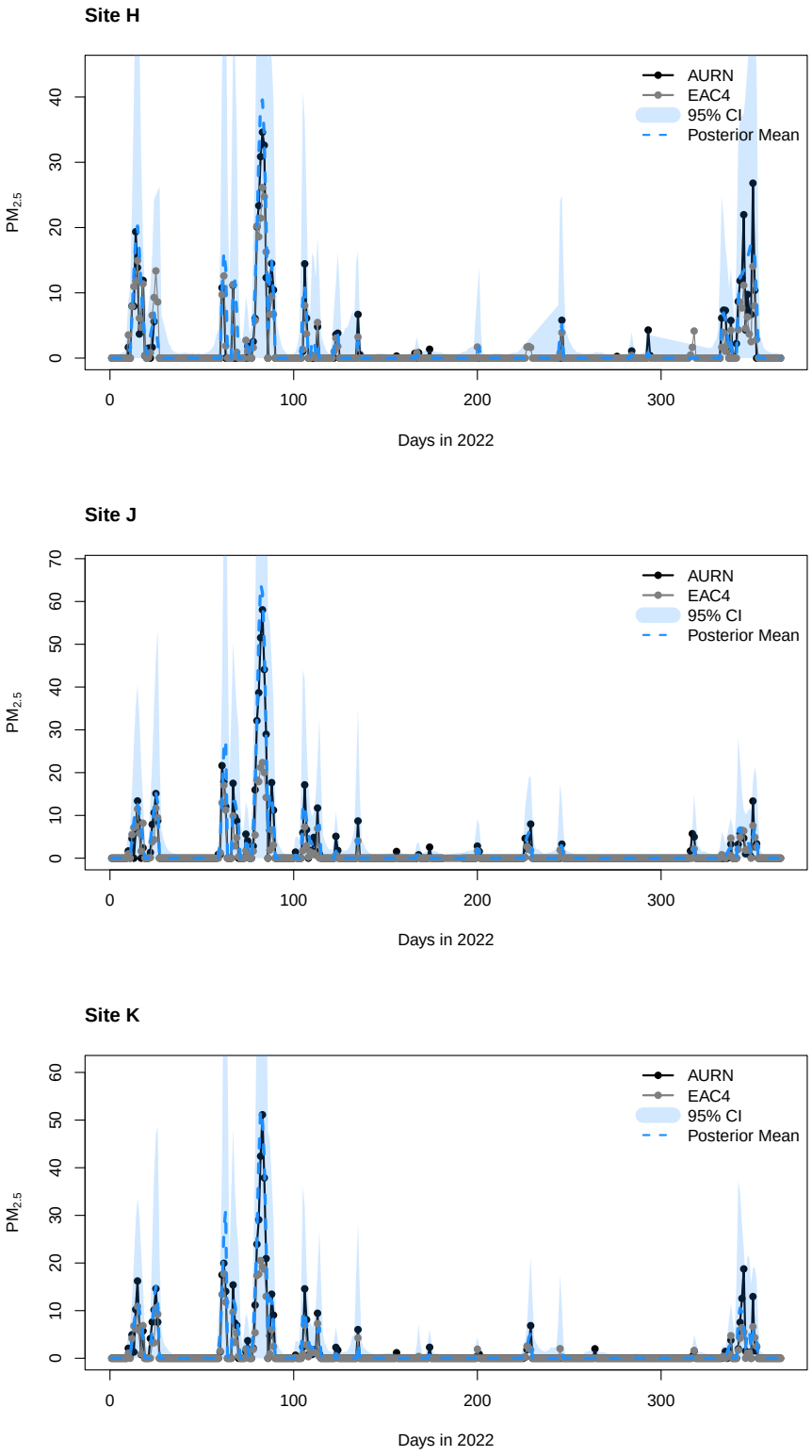

Figure B.6: Times series of fitted values and confidence bands of the ExDF model, along with the true AURN measurements and EAC4 observations for sites H, J, and K.

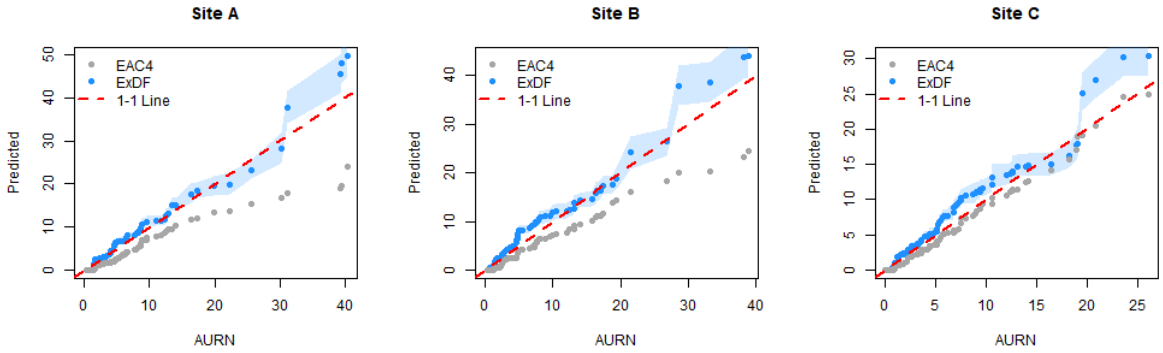

Figure B.7: Q-Q plot of $PM_{2.5}$ measurements at site A (left), B (middle), and C (right) for the ExDF and EAC4 models in blue and grey, respectively, against the true observations from the AURN observation stations at those locations. Point-wise 95% confidence intervals are given for the ExDF data.

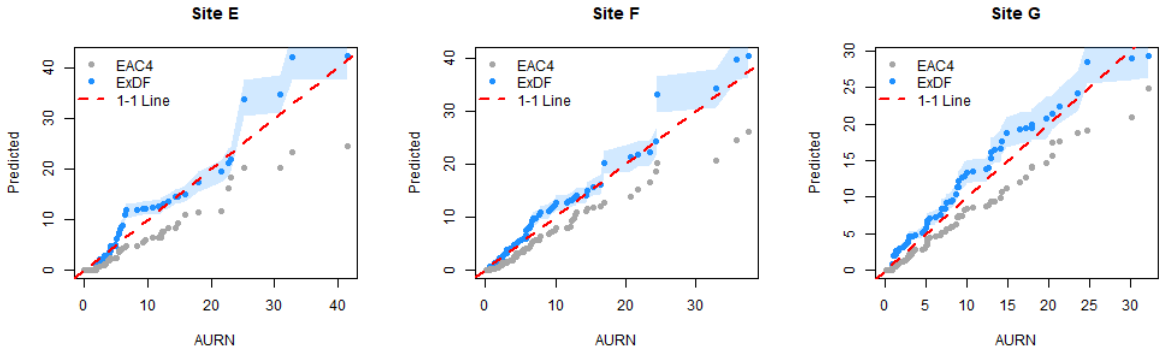

Figure B.8: Q-Q plot of $PM_{2.5}$ measurements at site E (left), F (middle), and G (right) for the ExDF and EAC4 models in blue and grey, respectively, against the true observations from the AURN observation stations at those locations. Point-wise 95% confidence intervals are given for the ExDF data.

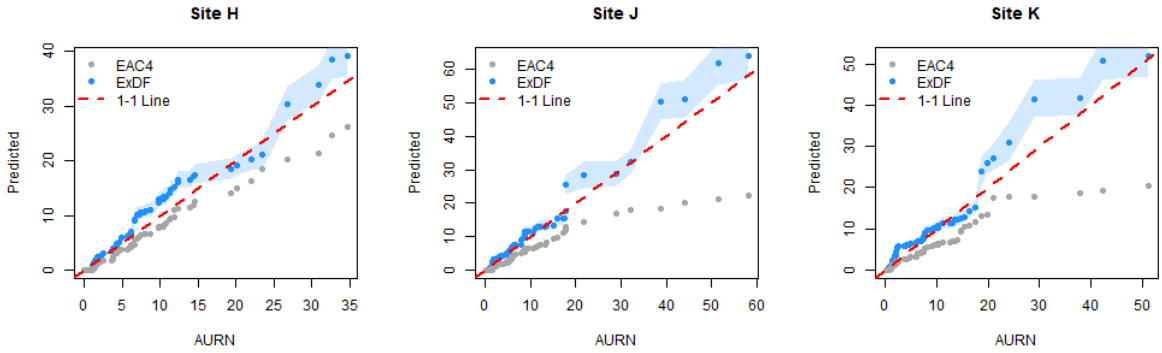

Figure B.9: Q-Q plot of $PM_{2.5}$ measurements at site H (left), J (middle), and K (right) for the ExDF and EAC4 models in blue and grey, respectively, against the true observations from the AURN observation stations at those locations. Point-wise 95% confidence intervals are given for the ExDF data.

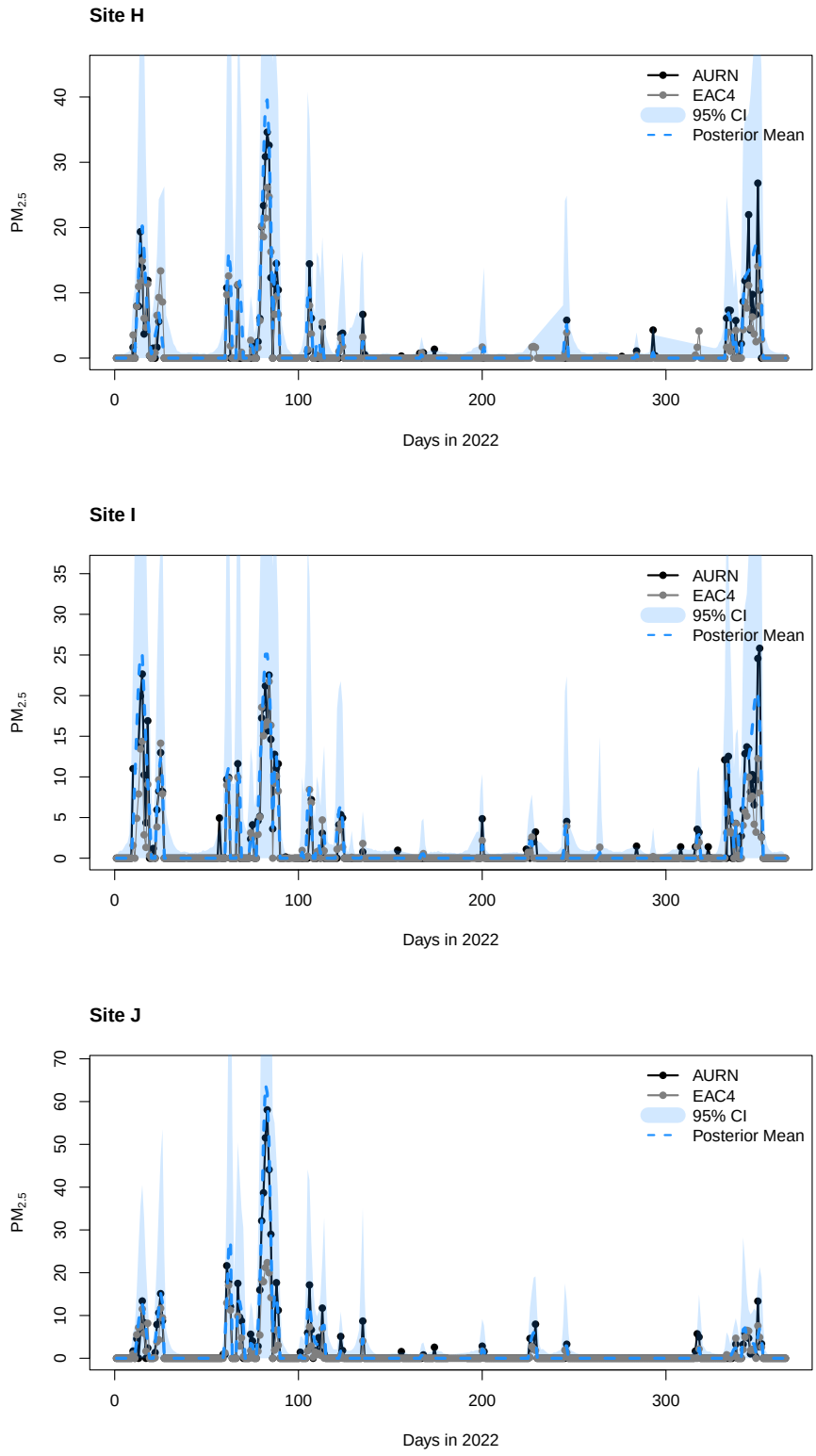

Figure B.10: Times series of fitted values and confidence bands of the ExDF model, along with the true AURN measurements and EAC4 observations for sites H, I,and J.

Bibliography

Adamo, P., Iavazzo, P., Albanese, S., Agrelli, D., De Vivo, B. and Lima, A. (2014) Bioavailability and soil-to-plant transfer factors as indicators of potentially toxic element contamination in agricultural soils. *Science of The Total Environment* **500-501**, 11–22.

Agyeman, P. C., Kingsley, J., Kebonye, N. M., Khosravi, V., Borůvka, L. and Vašát, R. (2023) Prediction of the concentration of antimony in agricultural soil using data fusion, terrain attributes combined with regression kriging. *Environmental Pollution* **316**, 120697.

Aitchison, J. (1985) A General Class of Distributions on the Simplex. *Journal of the Royal Statistical Society. Series B (Methodological)* **47**(1), 136–146.

Alkema, L., Raftery, A. E. and Clark, S. J. (2007) Probabilistic projections of HIV prevalence using Bayesian melding. *The Annals of Applied Statistics* **1**(1), 229–248.

Alloway, B. J. (2013) Sources of Heavy Metals and Metalloids in Soils. In *Heavy Metals in Soils: Trace Metals and Metalloids in Soils and their Bioavailability*, ed. B. J. Alloway, pp. 11–50. Dordrecht: Springer Netherlands.

Amaral Turkman, M. A., Feridun Turkman, K., de Zea Bermudez, P., Pereira, S., Pereira, P. and de Carvalho, M. (2021) Calibration of the Bulk and Extremes of Spatial Data. *REVSTAT-Statistical Journal* **19**(3), 309–325.

Anand, A., Ramadurai, G. and Vanajakshi, L. (2014) Data Fusion-Based Traffic Density Estimation and Prediction. *Journal of Intelligent Transportation Systems* **18**(4), 367–378.

Ander, E. L., Johnson, C. C., Cave, M. R., Palumbo-Roe, B., Nathanail, C. P. and Lark, R. M. (2013) Methodology for the determination of normal background concentrations of contaminants in English soil. *Science of The Total Environment* **454-455**, 604–618.

Anderson, J. O., Thundiyil, J. G. and Stolbach, A. (2012) Clearing the Air: A Review of the Effects of Particulate Matter Air Pollution on Human Health. *Journal of Medical Toxicology* **8**(2), 166–175.

Angon, P. B., Islam, M. S., Kc, S., Das, A., Anjum, N., Poudel, A. and Suchi, S. A. (2024) Sources, effects and present perspectives of heavy metals contamination: Soil, plants and human food chain. *Heliyon* **10**(7), e28357.

AQSR (2010) The Air Quality Standards Regulations 2010.

Banerjee, S., Carlin, B. P., Gelfand, A. E. and Banerjee, S. (2015) *Hierarchical Modeling and Analysis for Spatial Data*. 0th edition. Chapman and Hall/CRC.

Basith, S., Manavalan, B., Shin, T. H., Park, C. B., Lee, W.-S., Kim, J. and Lee, G. (2022) The Impact of Fine Particulate Matter 2.5 on the Cardiovascular System: A Review of the Invisible Killer. *Nanomaterials* **12**(15), 2656.

Beauchamp, M., de Fouquet, C. and Malherbe, L. (2017) Dealing with non-stationarity through explanatory variables in kriging-based air quality maps. *Spatial Statistics* **22**, 18–46.

Beauchamp, M., Malherbe, L., de Fouquet, C. and Létinois, L. (2018) A necessary distinction between spatial representativeness of an air quality monitoring station and the delimitation of exceedance areas. *Environmental Monitoring and Assessment* **190**(7), 441.

Becker, S., Sapkota, R. P., Pokharel, B., Adhikari, L., Pokhrel, R. P., Khanal, S. and Giri, B. (2021) Particulate matter variability in Kathmandu based on in-situ measurements, remote sensing, and reanalysis data. *Atmospheric Research* **258**, 105623.

Beirlant, J. (ed.) (2004) *Statistics of extremes: theory and applications*. Wiley series in probability and statistics. Hoboken, NJ: Wiley.

Bellows, B. C. (2005) Arsenic in Poultry Litter: Organic Regulations. *The National Sustaninable Agriculture Information Service* .

Beloconi, A., Kamarianakis, Y. and Chrysoulakis, N. (2016) Estimating urban $PM10$ and $PM2.5$ concentrations, based on synergistic MERIS/AATSR aerosol observations, land cover and morphology data. *Remote Sensing of Environment* **172**, 148–164.

Benaskeur, A. R. and Rhéaume, F. (2007) Adaptive data fusion and sensor management for military applications. *Aerospace Science and Technology* **11**(4), 327–338.

Berild, M. O., Martino, S., Gómez-Rubio, V. and Rue, H. (2022) Importance Sampling with the Integrated Nested Laplace Approximation. *Journal of Computational and Graphical Statistics* **31**(4), 1225–1237.

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), 176–197.

Boccaccio, M. (2023) Principles and Methods to Control Environmental Pollution. *Journal of Pollution Effects & Control* **11**(2), 1–2.

Bogaert, P. and Fasbender, D. (2007) Bayesian data fusion in a spatial prediction context: a general formulation. *Stochastic Environmental Research and Risk Assessment* **21**(6), 695–709.

Branca, J. J. V., Morucci, G. and Pacini, A. (2018) Cadmium-induced neurotoxicity: still much ado. *Neural Regeneration Research* **13**(11), 1879–1882.

Braverman, A. J. (2014) Data Fusion. In *Wiley StatsRef: Statistics Reference Online*. John Wiley & Sons, Ltd.

Brenning, A. (2008) Statistical geocomputing combining R and SAGA: The example of landslide susceptibility analysis with generalized additive models. In *SAGA – (Hamburger Beitraege zur Physischen Geographie und Landschaftsoekologie, vol. 19)*, pp. 23–32. J. Boehner, T. Blaschke, L. Montanarella.

Briffa, J., Sinagra, E. and Blundell, R. (2020) Heavy metal pollution in the environment and their toxicological effects on humans. *Heliyon* **6**(9), e04691.

Brooks, S., Gelman, A., Jones, G. and Meng, X.-L. (eds) (2011) *Handbook of Markov Chain Monte Carlo*. New York: Chapman and Hall/CRC.

Bürger, G., Murdock, T. Q., Werner, A. T., Sobie, S. R. and Cannon, A. J. (2012) Downscaling Extremes—An Intercomparison of Multiple Statistical Methods for Present Climate. *Journal of Climate* **25**(12), 4366–4388.

Carmona, A., Roudeau, S. and Ortega, R. (2021) Molecular Mechanisms of Environmental Metal Neurotoxicity: A Focus on the Interactions of Metals with Synapse Structure and Function. *Toxics* **9**(9), 198.

Carnevale, C., Angelis, E. D., Finzi, G., Turrini, E. and Volta, M. (2020) Application of Data Fusion Techniques to Improve Air Quality Forecast: A Case Study in the Northern Italy. *Atmosphere* **11**(3), 244.

Castanedo, F. (2013) A Review of Data Fusion Techniques. *The Scientific World Journal* **2013**, 1–19.

Castrignanò, A., A. Castrignanò, Castrignanò, A., R. Quarto, Quarto, R., A. Venezia, Venezia, A., Gabriele Buttafuoco and Buttafuoco, G. (2017) A geostatistical approach for modelling and combining spatial data with different support. *Advances in Animal Biosciences* **8**(2), 594–599.

Castro-Camilo, D. and Huser, R. (2020) Local Likelihood Estimation of Complex Tail Dependence Structures, Applied to U.S. Precipitation Extremes. *Journal of the American Statistical Association* **115**(531), 1037–1054.

Castro-Camilo, D., Huser, R. and Rue, H. (2022) Practical strategies for generalized extreme value-based regression models for extremes. *Environmetrics* **33**(6), e2742.

Chakraborty, S. C., Qamruzzaman, M., Zaman, M. W. U., Alam, M. M., Hossain, M. D., Pramanik, B. K., Nguyen, L. N., Nghiem, L. D., Ahmed, M. F., Zhou, J. L., Mondal, M. I. H., Hossain, M. A., Johir, M. A. H., Ahmed, M. B., Sithi, J. A., Zargar, M. and Moni, M. A. (2022) Metals in e-waste: Occurrence, fate, impacts and remediation technologies. *Process Safety and Environmental Protection* **162**, 230–252.

Chang, H. H., Hu, X. and Liu, Y. (2014) Calibrating MODIS aerosol optical depth for predicting daily PM2.5 concentrations via statistical downscaling. *Journal of Exposure Science & Environmental Epidemiology* **24**(4), 398–404.

Charkiewicz, A. E., Omeljaniuk, W. J., Nowak, K., Garley, M. and Nikliński, J. (2023) Cadmium Toxicity and Health Effects—A Brief Summary. *Molecules* **28**(18), 6620.

Chatterjee, A., Michalak, A. M., Kahn, R. A., Paradise, S. R., Braverman, A. J. and Miller, C. E. (2010) A geostatistical data fusion technique for merging remote sensing and ground-based observations of aerosol optical thickness. *Journal of Geophysical Research: Atmospheres* **115**(D20).

Chen, Z., Chen, Y., Liang, J., Sun, Z., Zhao, H. and Huang, Y. (2024) The Release and Migration of Cr in the Soil under Alternating Wet–Dry Conditions. *Toxics* **12**(2), 140.

Cheng, B., Alapaty, K. and Arunachalam, S. (2024) Spatiotemporal trends in PM2.5 chemical composition in the conterminous U.S. during 2006–2020. *Atmospheric Environment* **316**, 120188.

Chiles, J.-P. and Delfiner, P. (1999) *Geostatistics: modeling spatial uncertainty*. Number Book, Whole. Chichester;New York, N.Y;: Wiley.

Chilès, J.-P. and Delfiner, P. (2012) *Geostatistics: modeling spatial uncertainty*. Second edition. Wiley series in probability and statistics. Hoboken, NJ: Wiley.

Chmielewski, M., Kukiełka, M., Pieczonka, P. and Gutowski, T. (2020) Methods and analytical tools for assessing tactical situation in military operations using potential approach and sensor data fusion. *Procedia Manufacturing* **44**, 559–566.

Chopra, A. K., Pathak, C. and Prasad, G. (2009) Scenario of heavy metal contamination in agricultural soil and its management. *Journal of Applied and Natural Science* **1**(1), 99–108.

CL:AIRE (2007) Treatment of Chromium contamination and Chromium Ore Processing Residue. Technical report, CL:AIRE.

Clarotto, L., Allard, D. and Menafoglio, A. (2022) A new class of alpha-transformations for the spatial analysis of Compositional Data. *Spatial Statistics* **47**, 100570.

Cole, S. and Jeffries, J. (2009) *Using soil guideline values*. Bristol: Environmet Agency.

Coles, S. (2001) *An introduction to statistical modeling of extreme values*. Number Book, Whole. London;New York;: Springer.

Coles, S. and Casson, E. (1998) Extreme value modelling of hurricane wind speeds. *Structural Safety* **20**(3), 283–296.

Coles, S., Heffernan, J. and Tawn, J. (1999) Dependence Measures for Extreme Value Analyses. *Extremes* **2**(4), 339–365.

Couturier, D.-L. and Victoria-Feser, M.-P. (2010) Zero-inflated truncated generalized Pareto distribution for the analysis of radio audience data. *The Annals of Applied Statistics* **4**(4).

Cressie, N. and Johannesson, G. (2008) Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**(1), 209–226.

Cressie, N. A. and Helterbrand, J. D. (1994) Multivariate spatial statistical models pp. 179–188.

Cressie, N. A. C. (1993) *Statistics for spatial data*. Rev. ed edition. Wiley series in probability and mathematical statistics. New York: Wiley.

Dasarathy, B. (1997) Sensor fusion potential exploitation-innovative architectures and illustrative applications. *Proceedings of the IEEE* **85**(1), 24–38.

Dautov, R., Distefano, S. and Buyya, R. (2019) Hierarchical data fusion for Smart Healthcare. *Journal of Big Data* **6**(1), 19.

Davel, A. P., Lemos, M., Pastro, L. M., Pedro, S. C., de André, P. A., Hebeda, C., Farsky, S. H., Saldiva, P. H. and Rossoni, L. V. (2012) Endothelial dysfunction in the pulmonary artery induced by concentrated fine particulate matter exposure is associated with local but not systemic inflammation. *Toxicology* **295**(1-3), 39–46.

Davison, A. C. and Gholamrezaee, M. M. (2012) Geostatistics of extremes. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **468**(2138), 581–608.

Davison, A. C., Padoan, S. A. and Ribatet, M. (2012) Statistical Modeling of Spatial Extremes. *Statistical Science* **27**(2).

DEFRA (1995) The Environmental Protection Act 1990: Part 2A.

Demangeot, M. (2020) L'analyse spatiale des extrêmes à partir d'une unique réalisation: un point de vue géostatistique p. 181.

Diggle, P. and Ribeiro, P. J. (2007) *Model-based geostatistics*. Springer series in statistics. New York, NY: Springer.

Don D. Sin, Dany Doiron, Àlvar Agustí, Antonio Anzueto, Peter J. Barnes, Bartolomé R. Celli, Gerard J. Criner, David Halpin, MeiLan K. Han, Fernando J. Martínez, María Montes de, Alberto Papi, Ian Pavord, N. Roche, Dave Singh, Robert A. Stockley, M. Victorina Lopez Varlera, Jadwiga A. Wedzicha, Claus Volgelmeier and Jean Bourbeau (2023) Air pollution and COPD: GOLD 2023 committee report **61**(5), 2202469–2202469.

Drees, H. and Sabourin, A. (2021) Principal component analysis for multivariate extremes. *Electronic Journal of Statistics* **15**(1), 908–943.

Duan, Z., Du, F.-y., Yuan, Y.-d., Zhang, Y.-p., Yang, H.-s. and Pan, W.-s. (2013) [Effects of PM2.5 exposure on Klebsiella pneumoniae clearance in the lungs of rats]. *Zhonghua Jie He He Hu Xi Za Zhi = Zhonghua Jiehe He Huxi Zazhi = Chinese Journal of Tuberculosis and Respiratory Diseases* **36**(11), 836–840.

Dudek, A. and Baranowski, J. (2023) Spatial Modeling of Air Pollution Using Data Fusion. *Electronics* **12**(15), 3353.

Durrant-Whyte, H. F. (1988) Sensor models and multisensor integration. *The International Journal of Robotics Research* **7**(6), 97–113.

Ebtehaj, M. and Foufoula-Georgiou, E. (2010) Preservation of extremes in multi-sensor merging of precipitation. In *AGU Fall Meeting Abstracts*, volume 2010, pp. H21E–1086.

Eckel, H. E., Roth, U., Döhler, H., Nicholson, F. and Unvin, R. (2005) Assessment and reduction of heavy metal input into agro-ecosystems - Final report of the EU-Concerted Action AROMIS. KTBL-Schrift 432.

Engelke, S., de Fondeville, R. and Oesting, M. (2019) Extremal behaviour of aggregated data with an application to downscaling. *Biometrika* **106**(1), 127–144.

Engelke, S. and Ivanovs, J. (2021) Sparse Structures for Multivariate Extremes. *Annual Review of Statistics and Its Application* **8**(1), 241–270.

Environment Agency, E. (2009) CLEA 2009 Contaminated Land Exposure Assessment.

Erisman, J. W., Galloway, J. N., Seitzinger, S., Bleeker, A., Dise, N. B., Petrescu, A. M. R., Leach, A. M. and de Vries, W. (2013) Consequences of human modification of the global nitrogen cycle. *Philosophical Transactions of the Royal Society B: Biological Sciences* **368**(1621), 20130116.

Ermolin, M. S., Fedotov, P. S., Malik, N. A. and Karandashev, V. K. (2018) Nanoparticles of volcanic ash as a carrier for toxic elements on the global scale. *Chemosphere* **200**, 16–22.

Essou, G. R. C., Sabarly, F., Lucas-Picher, P., Brissette, F. and Poulin, A. (2016) Can Precipitation and Temperature from Meteorological Reanalyses Be Used for Hydrological Modeling? *Journal of Hydrometeorology* **17**(7), 1929 – 1950.

ETR (2023) The Environmental Targets (Biodiversity) (England) Regulations 2023.

Everett, C. J. and Frithsen, I. L. (2008) Association of urinary cadmium and myocardial infarction. *Environmental Research* **106**(2), 284–286.

Farina, A., Ortenzi, L., Ristic, B. and Skvortsov, A. (2014) Chapter 22 - Integrated Sensor Systems and Data Fusion for Homeland Protection. In *Academic Press Library in Signal Processing*, eds N. D. Sidiropoulos, F. Gini, R. Chellappa and S. Theodoridis, volume 2 of *Academic Press Library in Signal Processing: Volume 2*, pp. 1245–1320. Elsevier.

Fasbender, D., Obsomer, V., Radoux, J., Bogaert, P. and Defourny, P. (2007) Bayesian Data Fusion: Spatial and Temporal Applications. In *2007 International Workshop on the Analysis of Multi-temporal Remote Sensing Images*, pp. 1–6. Leuven, Belgium: IEEE.

Feigin, V. L., Roth, G. A., Naghavi, M., Parmar, P., Krishnamurthi, R., Chugh, S., Mensah, G. A., Norrving, B., Shiue, I., Ng, M., Estep, K., Cercy, K., Murray, C. J. L. and Forouzanfar, M. H. (2016) Global burden of stroke and risk factors in 188 countries,

during 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. *The Lancet Neurology* **15**(9), 913–924.

Feng, S., Gao, D., Liao, F., Zhou, F. and Wang, X. (2016) The health effects of ambient PM2.5 and potential mechanisms. *Ecotoxicology and Environmental Safety* **128**, 67–74.

Ferreira, F., Tente, H., Torres, P., Cardoso, S. and Palma-Oliveira, J. M. (2000) Air Quality Monitoring and Management in Lisbon. In *Urban Air Quality: Measurement, Modelling and Management*, eds R. S. Sokhi, R. San José, N. Moussiopoulos and R. Berkowicz, pp. 443–450. Dordrecht: Springer Netherlands.

Figueroa, A., Cameselle, C., Gouveia, S. and Hansen, H. K. (2016) Electrokinetic treatment of an agricultural soil contaminated with heavy metals. *Journal of Environmental Science and Health. Part A, Toxic/Hazardous Substances & Environmental Engineering* **51**(9), 691–700.

Fisher, R. A. and Tippett, L. H. C. (1928) Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Mathematical Proceedings of the Cambridge Philosophical Society* **24**(2), 180–190.

Forlani, C., Bhatt, S., Cameletti, M., Elias Krainski, Krainski, E. T., Elias Krainski and Blangiardo, M. (2020) A joint bayesian space-time model to integrate spatially misaligned air pollution data in R-INLA .

Foufoula-Georgiou, E., Ebtehaj, A. M., Zhang, S. Q. and Hou, A. Y. (2014) Downscaling Satellite Precipitation with Emphasis on Extremes: A Variational ‘1-Norm Regularization in the Derivative Domain. *Surv Geophys* .

Frechét, M. (1927) Sur la loi de probabilité de l’écart maximum. *Annales de la Société Polonaise de Mathématique T. 6 (1927)* .

Friberg, M. D., Zhai, X., Holmes, H. A., Chang, H. H., Strickland, M. J., Sarnat, S. E., Tolbert, P. E., Russell, A. G. and Mulholland, J. A. (2016) Method for Fusing Observational Data and Chemical Transport Model Simulations To Estimate Spatiotemporally Resolved Ambient Air Pollution. *Environmental Science & Technology* **50**(7), 3695–3705.

Friederichs, P. (2010) Statistical downscaling of extreme precipitation events using extreme value theory. *Extremes* **13**(2), 109–132.

Friederichs, P. and Thorarinsdottir, T. L. (2012) Forecast verification for extreme value distributions with an application to probabilistic peak wind prediction. *Environmetrics* **23**(7), 579–594.

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.

Fuller, R., Landrigan, P. J., Balakrishnan, K., Bathan, G., Bose-O'Reilly, S., Brauer, M., Caravanos, J., Chiles, T., Cohen, A., Corra, L., Cropper, M., Ferraro, G., Hanna, J., Hanrahan, D., Hu, H., Hunter, D., Janata, G., Kupka, R., Lanphear, B., Lichtveld, M., Martin, K., Mustapha, A., Sanchez-Triana, E., Sandilya, K., Schaefli, L., Shaw, J., Seddon, J., Suk, W., Téllez-Rojo, M. M. and Yan, C. (2022) Pollution and health: a progress update. *The Lancet Planetary Health* **6**(6), e535–e547.

Gandin, L. S. (1966) Objective analysis of meteorological fields. By L. S. Gandin. Translated from the Russian. Jerusalem (Israel Program for Scientific Translations), 1965. Pp. vi, 242: 53 Figures; 28 Tables. £4 1s. 0d. *Quarterly Journal of the Royal Meteorological Society* **92**(393), 447–447.

Garza-Lombó, C., Pappa, A., Panayiotidis, M. I., Gonsebatt, M. E. and Franco, R. (2019) Arsenic-induced neurotoxicity: A mechanistic appraisal. *Journal of biological inorganic chemistry : JBIC : a publication of the Society of Biological Inorganic Chemistry* **24**(8), 1305.

GBD (2016) Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015. *The Lancet* **388**(10053), 1659–1724.

Gelfand, A. E., Kim, H.-J., Sirmans, C. F. and Banerjee, S. (2003) Spatial Modeling With Spatially Varying Coefficient Processes. *Journal of the American Statistical Association* **98**(462), 387–396.

Gelfand, A. E. and Smith, A. F. M. (1990) Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association* **85**(410), 398–409.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. and Rubin, D. B. (2015) *Bayesian Data Analysis*. Third edition. New York: Chapman and Hall/CRC.

Gengler, S. and Bogaert, P. (2014) Bayesian data fusion for spatial prediction of categorical variables in environmental sciences. *AIP Conference Proceedings* **1636**(1), 88–93.

Gengler, S. and Bogaert, P. (2016) Integrating Crowdsourced Data with a Land Cover Product: A Bayesian Data Fusion Approach. *Remote Sensing* **8**(7), 545.

Ghigo, S., Bande, S., Ciancarella, L., Mircea, M., Piersanti, A., Righini, G., Baldasano, J., Basagaña, X., Cadum, E. and on, b. o. t. M. H. S. g. (2018) Mapping air pollutants at municipality level in Italy and Spain in support to health impact evaluations. *Air Quality, Atmosphere and Health* **11**(1), 69–82.

Gnedenko, B. (1943) Sur La Distribution Limite Du Terme Maximum D'Une Serie Aleatoire. *Annals of Mathematics* **44**(3), 423–453.

Goldhaber, S. B. (2003) Trace element risk assessment: essentiality vs. toxicity. *Regulatory Toxicology and Pharmacology: RTP* **38**(2), 232–242.

Gomez-Rubio, V. (2017) Mixture model fitting using conditional models and modal Gibbs sampling. arXiv:1712.09566 [stat].

Gressent, A., Malherbe, L., Colette, A., Rollin, H. and Scimia, R. (2020) Data fusion for air quality mapping using low-cost sensor observations: Feasibility and added-value. *Environment International* **143**, 105965.

Gualtieri, M., Ovrevik, J., Mollerup, S., Asare, N., Longhin, E., Dahlman, H.-J., Camatini, M. and Holme, J. A. (2011) Airborne urban particles (Milan winter-PM2.5) cause mitotic arrest and cell death: Effects on DNA, mitochondria, AhR binding and spindle organization. *Mutation Research* **713**(1-2), 18–31.

Gudendorf, G. and Segers, J. (2010) Extreme-Value Copulas. In *Copula Theory and Its Applications*, eds P. Jaworski, F. Durante, W. K. Härdle and T. Rychlik, pp. 127–145. Berlin, Heidelberg: Springer.

Gumbel, E. J. (1935) Les valeurs extrêmes des distributions statistiques. *Annales de l'institut Henri Poincaré* **5**(2), 115–158.

Guo, Z., Wang, H., Liu, Q. and Yang, J. (2014) A Feature Fusion Based Forecasting Model for Financial Time Series. *PLOS ONE* **9**(6), e101113.

Gómez-Sagasti, M. T., Alkorta, I., Becerril, J. M., Epelde, L., Anza, M. and Garbisu, C. (2012) Microbial Monitoring of the Recovery of Soil Quality During Heavy Metal Phytoremediation. *Water, Air, & Soil Pollution* **223**(6), 3249–3262.

H R Anderson (2009) Air pollution and mortality: A history. *Atmospheric Environment* **43**(1), 142–152.

de Haan, L. and Zhou, C. (2024) Bootstrapping Extreme Value Estimators. *Journal of the American Statistical Association* **119**(545), 382–393.

Haan, L. D. (1984) A Spectral Representation for Max-stable Processes. *The Annals of Probability* **12**(4).

Haan, L. d. and Ferreira, A. (2006) *Extreme Value Theory: An Introduction*. Springer series in operations research. New York ; London: Springer.

Habre, R., Moshier, E., Castro, W., Nath, A., Grunin, A., Rohr, A., Godbold, J., Schachter, N., Kattan, M., Coull, B. and Koutrakis, P. (2014) The effects of PM2.5 and its components from indoor and outdoor sources on cough and wheeze symptoms in asthmatic children. *Journal of Exposure Science & Environmental Epidemiology* **24**(4), 380–387.

Hall, D. and Llinas, J. (1997) An introduction to multisensor data fusion. *Proceedings of the IEEE* **85**(1), 6–23.

Hall, P. and Maesono, Y. (2000) A Weighted Bootstrap Approach to Bootstrap Iteration. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **62**(1), 137–144.

Hasanein, P. and Emamjomeh, A. (2019) Chapter 28 - Beneficial Effects of Natural Compounds on Heavy Metal–Induced Hepatotoxicity. In *Dietary Interventions in Liver Disease*, eds R. R. Watson and V. R. Preedy, pp. 345–355. Academic Press.

Hassani, S., Dackermann, U., Mousavi, M. and Li, J. (2024) A systematic review of data fusion techniques for optimized structural health monitoring. *Information Fusion* **103**, 102136.

He, D., Wu, S., Zhao, H., Qiu, H., Fu, Y., Li, X. and He, Y. (2017) Association between particulate matter 2.5 and diabetes mellitus: A meta-analysis of cohort studies. *Journal of Diabetes Investigation* **8**(5), 687–696.

He, Z., Shentu, J., Yang, X., Baligar, V., Zhang, T. and Stoffella, P. (2015) Heavy Metal Contamination of Soils: Sources, Indicators, and Assessment. *Journal of Environmental Indicators* **9**, 17–18.

Healy, D., Tawn, J., Thorne, P. and Parnell, A. (2023) Inference for extreme spatial temperature events in a changing climate with application to Ireland. *(to appear).* .

Hegazy, A. M. S. and Fouad, U. A. (2014) Evaluation of Lead Hepatotoxicity; Histological, Histochemical and Ultrastructural Study. *Forensic Medicine and Anatomy Research* **02**(03), 70.

Hogan, J. W. and Tchernis, R. (2004) Bayesian Factor Analysis for Spatially Correlated Data, with Application to Summarizing Area-Level Material Deprivation from Census Data. *Journal of the American Statistical Association* **99**(466), 314–324.

Hosseini Beinabaj, S. M., Heydariyan, H., Mohammad Aleii, H. and Hosseinzadeh, A. (2023) Concentration of heavy metals in leachate, soil, and plants in Tehran's landfill: Investigation of the effect of landfill age on the intensity of pollution. *Heliyon* **9**(1), e13017.

Hsiao, C.-L., Wu, K.-H. and Wan, K.-S. (2011) Effects of environmental lead exposure on T-helper cell-specific cytokines in children. *Journal of Immunotoxicology* **8**(4), 284–287.

Huang, H.-W., Lee, C.-H. and Yu, H.-S. (2019) Arsenic-Induced Carcinogenesis and Immune Dysregulation. *International Journal of Environmental Research and Public Health* **16**(15), 2746.

Huang, W., Li, T., Liu, J., Xie, P., Du, S. and Teng, F. (2021) An overview of air quality analysis by big data techniques: Monitoring, forecasting, and traceability. *Information Fusion* **75**, 28–40.

Hundecha, Y. and Bárdossy, A. (2008) Statistical downscaling of extremes of daily precipitation and temperature and construction of their future scenarios. *International Journal of Climatology* **28**(5), 589–610.

Hurtado-Díaz, M., Riojas-Rodríguez, H., Rothenberg, S. J., Schnaas-Arrieta, L., Kloog, I., Just, A., Hernández-Bonilla, D., Wright, R. O. and Téllez-Rojo, M. M. (2021) Prenatal PM2.5 exposure and neurodevelopment at 2 years of age in a birth cohort from Mexico city. *International Journal of Hygiene and Environmental Health* **233**, 113695.

Huser, R., Davison, A. C. and Genton, M. G. (2016) Likelihood estimators for multivariate extremes. *Extremes* **19**(1), 79–103.

Huser, R., Opitz, T. and Thibaud, E. (2017) Bridging asymptotic independence and dependence in spatial extremes using Gaussian scale mixtures. *Spatial Statistics* **21**, 166–186.

Huser, R., Opitz, T. and Wadsworth, J. (2024) Modeling of spatial extremes in environmental data science: Time to move away from max-stable processes. arXiv:2401.17430 [stat].

Huser, R. and Wadsworth, J. L. (2019) Modeling Spatial Processes with Unknown Extremal Dependence Class. *Journal of the American Statistical Association* **114**(525), 434–444.

Huser, R. and Wadsworth, J. L. (2022) Advances in statistical modeling of spatial extremes. *WIREs Computational Statistics* **14**(1), e1537.

Hyder, O., Chung, M., Cosgrove, D., Herman, J. M., Li, Z., Firoozmand, A., Gurakar, A., Koteish, A. and Pawlik, T. M. (2013) Cadmium Exposure and Liver Disease among US Adults. *Journal of Gastrointestinal Surgery : Official Journal of the Society for Surgery of the Alimentary Tract* **17**(7), 1265–1273.

Inness, A., Ades, M., Agustí-Panareda, A., Barré, J., Benedictow, A., Blechschmidt, A.-M., Dominguez, J. J., Engelen, R., Eskes, H., Flemming, J., Huijnen, V., Jones, L., Kipling, Z., Massart, S., Parrington, M., Peuch, V.-H., Razinger, M., Remy, S., Schulz, M. and Suttie, M. (2019) The CAMS reanalysis of atmospheric composition. *Atmospheric Chemistry and Physics* **19**(6), 3515–3556.

Irincheeva, I., Cantoni, E. and Genton, M. G. (2012) A Non-Gaussian Spatial Generalized Linear Latent Variable Model. *Journal of Agricultural, Biological, and Environmental Statistics* **17**(3), 332–353.

Issa, M. E., Helmi, A. M., Al-Qaness, M. A. A., Dahou, A., Abd Elaziz, M. and Damaševičius, R. (2022) Human Activity Recognition Based on Embedded Sensor Data Fusion for the Internet of Healthcare Things. *Healthcare* **10**(6), 1084.

Jedrychowski, W. A., Perera, F. P., Spengler, J. D., Mroz, E., Stigter, L., Flak, E., Majewska, R., Klimaszewska-Rembiasz, M. and Jacek, R. (2013) Intrauterine exposure to fine particulate matter as a risk factor for increased susceptibility to acute broncho-pulmonary infections in early childhood. *International Journal of Hygiene and Environmental Health* **216**(4), 395–401.

Jin, X., Ding, J., Ge, X., Liu, J., Xie, B., Zhao, S. and Zhao, Q. (2022) Machine learning driven by environmental covariates to estimate high-resolution PM2.5 in data-poor regions. *PeerJ* **10**, e13203.

Jinnagara Puttaswamy, S., Nguyen, H. M., Braverman, A., Hu, X. and Liu, Y. (2014) Statistical data fusion of multi-sensor AOD over the Continental United States. *Geocarto International* **29**(1), 48–64.

Johnson, C. C., Breward, N., Ander, E. L. and Ault, L. (2005) G-BASE: baseline geochemical mapping of Great Britain and Northern Ireland. *Geochemistry: Exploration, Environment, Analysis* **5**(4), 347–357.

Johnson, L., Bishop, T. and Birch, G. (2017) Modelling drivers and distribution of lead and zinc concentrations in soils of an urban catchment (Sydney estuary, Australia). *Science of The Total Environment* **598**, 168–178.

Jonathan, P., Randell, D., Wadsworth, J. and Tawn, J. (2021) Uncertainties in return values from extreme value analysis of peaks over threshold using the generalised Pareto distribution. *Ocean Engineering* **220**, 107725.

Journel, A. and Huijbregts, C. (1978) Mining Geostatistics. In *Mining Geostatistics*. London: Academic Press.

Kaiho, K. (2023) An animal crisis caused by pollution, deforestation, and warming in the late 21st century and exacerbation by nuclear war. *Heliyon* **9**(4), e15221.

Kallache, M., Vrac, M., Naveau, P. and Michelangeli, P.-A. (2011) Nonstationary probabilistic downscaling of extreme precipitation. *Journal of Geophysical Research: Atmospheres* **116**(D5).

Kampa, M. and Castanas, E. (2008) Human health effects of air pollution. *Environmental Pollution* **151**(2), 362–367.

Khalid, S., Shahid, M., Niazi, N. K., Murtaza, B., Bibi, I. and Dumat, C. (2017) A comparison of technologies for remediation of heavy metal contaminated soils. *Journal of Geochemical Exploration* **182**, 247–268.

Khanna, I., Khare, M., Gargava, P. and Khan, A. A. (2018) Effect of PM2.5 chemical constituents on atmospheric visibility impairment. *Journal of the Air & Waste Management Association* **68**(5), 430–437.

Khlifi, R. and Hamza-Chaffai, A. (2010) Head and neck cancer due to heavy metal exposure via tobacco smoking and professional exposure: a review. *Toxicology and Applied Pharmacology* **248**(2), 71–88.

Kim, Y.-J. and Kim, J.-M. (2015) Arsenic Toxicity in Male Reproduction and Development. *Development & Reproduction* **19**(4), 167–180.

Kiriliouk, A., Rootzén, H., Segers, J. and Wadsworth, J. L. (2019) Peaks Over Thresholds Modeling With Multivariate Generalized Pareto Distributions. *Technometrics* **61**(1), 123–135.

Kolmogorov, A. (1941) Interpolirovanie i ekstrapolirovanie statsionarnykh sluchainyk posledovatel'nostei. *Izv. Akad. Nauk SSSR* **5**, 3–14.

Krainski, E., Gómez-Rubio, V., Bakka, H., Lenzi, A., Castro-Camilo, D., Simpson, D., Lindgren, F. and Rue, H. (2018) *Advanced Spatial Modeling with Stochastic Partial Differential Equations Using R and INLA*. Chapman and Hall/CRC.

Kravchenko, A. (2003) Influence of spatial structure on accuracy of interpolation methods. *Soil Science Society of America Journal* **67**(5), 1564–1571.

Krige, D. G. (1951) A statistical approach to some basic mine valuation problems on the Witwatersrand. *Journal of the Southern African Institute of Mining and Metallurgy* **52**(6), 119–139.

Krupskii, P., Huser, R. and Genton, M. G. (2018) Factor Copula Models for Replicated Spatial Data. *Journal of the American Statistical Association* **113**(521), 467–479.

Krupskii, P. and Joe, H. (2015) Structured factor copula models: Theory, inference and computation. *Journal of Multivariate Analysis* **138**, 53–73.

Kulka, M. (2016) A review of paraoxonase 1 properties and diagnostic applications. *Polish Journal of Veterinary Sciences* **19**(1), 225–232.

Kupka, D., Kania, M., Pietrzykowski, M., Łukasik, A. and Gruba, P. (2021) Multiple Factors Influence the Accumulation of Heavy Metals (Cu, Pb, Ni, Zn) in Forest Soils in the Vicinity of Roadways. *Water, Air, & Soil Pollution* **232**(5), 194.

Kyung, S. Y. and Jeong, S. H. (2020) Particulate-Matter Related Respiratory Diseases. *Tuberculosis and Respiratory Diseases* **83**(2), 116.

Künzli, N., Jerrett, M., Mack, W. J., Beckerman, B., LaBree, L., Gilliland, F., Thomas, D., Peters, J. and Hodis, H. N. (2005) Ambient air pollution and atherosclerosis in Los Angeles. *Environmental Health Perspectives* **113**(2), 201–206.

Lado, L. R., Hengl, T. and Reuter, H. I. (2008) Heavy metals in European soils: A geostatistical analysis of the FOREGS Geochemical database. *Geoderma* **148**(2), 189–199.

Lamorie-Foote, K., Ge, B., Shkirkova, K., Liu, Q. and Mack, W. (2023) Effect of Air Pollution Particulate Matter on Ischemic and Hemorrhagic Stroke: A Scoping Review. *Cureus* **15**(10), e46694.

Larios, D. F., Barbancho, J., Rodríguez, G., Sevillano, J. L., Molina, F. J. and León, C. (2012) Energy efficient wireless sensor network communications based on computational intelligent data fusion for environmental monitoring. *IET Communications* **6**(14), 2189–2197.

Lark, R. (2002) Optimized spatial sampling of soil for estimation of the variogram by maximum likelihood. *Geoderma* **105**(1-2), 49–80.

Lark, R. M. (2000) A comparison of some robust estimators of the variogram for use in soil survey. *European Journal of Soil Science* **51**(1), 137–157.

Laskin, D. (2006) The Great London Smog. *Weatherwise* **59**(6), 42–45.

Lawrence, P. G., Roper, W., Morris, T. F. and Guillard, K. (2020) Guiding soil sampling strategies using classical and spatial statistics: A review. *Agronomy Journal* **112**(1), 493–510.

Ledford, A. W. and Tawn, J. A. (1997) Modelling Dependence Within Joint Tail Regions. *Journal of the Royal Statistical Society. Series B (Methodological)* **59**(2), 475–499.

Lentini, P., Zanoli, L., Granata, A., Signorelli, S. S., Castellino, P. and Dell'Aquila, R. (2017) Kidney and heavy metals - The role of environmental exposure (Review). *Molecular Medicine Reports* **15**(5), 3413–3419.

Lertxundi, A., Andiarena, A., Martínez, M. D., Ayerdi, M., Murcia, M., Estarlich, M., Guxens, M., Sunyer, J., Julvez, J. and Ibarluzea, J. (2019) Prenatal exposure to PM2.5 and NO2 and sex-dependent infant cognitive and motor development. *Environmental Research* **174**, 114–121.

Levasseur, P., Erdlenbruch, K. and Gramaglia, C. (2022) The health and socioeconomic costs of exposure to soil pollution: evidence from three polluted mining and industrial sites in Europe. *Journal of Public Health* **30**(10), 2533–2546.

Li, L., Shi, R., Zhang, L., Zhang, J. and Gao, W. (2014) The data fusion of aerosol optical thickness using universal kriging and stepwise regression in East China. Volume 9221.

Li, M., Wang, F., Jia, X., Li, W., Li, T. and Rui, G. (2021) Multi-source data fusion for economic data analysis. *Neural Computing and Applications* **33**(10), 4729–4739.

Liang, F., Gao, M., Xiao, Q., Carmichael, G. R., Pan, X. and Liu, Y. (2017) Evaluation of a data fusion approach to estimate daily PM2.5 levels in North China. *Environmental Research* **158**, 54–60.

Lin, Y.-C., Chi, W.-J. and Lin, Y.-Q. (2020) The improvement of spatial-temporal resolution of PM2.5 estimation based on micro-air quality sensors by using data fusion technique. *Environment International* **134**.

Lin, Y.-P., Cheng, B.-Y., Shyu, G.-S. and Chang, T.-K. (2010) Combining a finite mixture distribution model with indicator kriging to delineate and map the spatial patterns of soil heavy metal pollution in Chunghua County, central Taiwan. *Environmental Pollution* **158**(1), 235–244.

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: Link between Gaussian Fields and Gaussian Markov Random Fields. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73**(4), 423–498.

Liu, L., Li, W., Song, W. and Guo, M. (2018) Remediation techniques for heavy metal-contaminated soils: Principles and applicability. *Science of The Total Environment* **633**, 206–219.

Liu, Y., Zidek, J. V., Trites, A. W. and Battaile, B. C. (2016) Bayesian data fusion approaches to predicting spatial tracks: Application to marine mammals. *The Annals of Applied Statistics* **10**(3).

Lugrin, T., Tawn, J. A. and Davison, A. C. (2021) Sub-asymptotic motivation for new conditional multivariate extreme models. *Stat* **10**(1).

Luo, R. and Kay, M. (1989) Multisensor integration and fusion in intelligent systems. *IEEE Transactions on Systems, Man, and Cybernetics* **19**(5), 901–931.

Lv, J., Liu, Y., Zhang, Z., Dai, J., Dai, B. and Zhu, Y. (2015) Identifying the origins and spatial distributions of heavy metals in soils of Ju country (Eastern China) using multivariate and geostatistical approach. *Journal of Soils and Sediments* **15**(1), 163–178.

Maalouf, A. and Mavropoulos, A. (2023) Re-assessing global municipal solid waste generation. *Waste Management & Research* **41**(4), 936–947.

Mackay, E. and Jonathan, P. (2020) Assessment of return value estimates from stationary and non-stationary extreme value models. *Ocean Engineering* **207**, 107406.

Mahmood, Q., Mirza, N. and Shaheen, S. (2015) Phytoremediation Using Algae and Macrophytes: I. In *Phytoremediation: Management of Environmental Contaminants, Volume 2*, eds A. A. Ansari, S. S. Gill, R. Gill, G. R. Lanza and L. Newman, pp. 265–289. Cham: Springer International Publishing.

Manzione, R. L. and Castrignanò, A. (2019) A geostatistical approach for multi-source data fusion to predict water table depth. *Science of The Total Environment* **696**, 133763.

Maraun, D. and Widmann, M. (2018) *Statistical Downscaling and Bias Correction for Climate Research*. Cambridge University Press.

Marchant, B., Saby, N., Jolivet, C., Arrouays, D. and Lark, R. (2011) Spatial prediction of soil properties with copulas. *Geoderma* **162**(3-4), 327–334.

Marchant, B. P., Saby, N. P. A., Lark, R. M., Bellamy, P. H., Jolivet, C. C. and Arrouays, D. (2010) Robust analysis of soil properties at the national scale: cadmium content of French soils. *European Journal of Soil Science* **61**(1), 144–152.

Martenies, S. E., Wilkins, D. and Batterman, S. A. (2015) Health impact metrics for air pollution management strategies. *Environment International* **85**, 84–95.

Martinez, V. D., Vucic, E. A., Becker-Santos, D. D., Gil, L. and Lam, W. L. (2011) Arsenic Exposure and the Induction of Human Cancers. *Journal of Toxicology* **2011**, 431287.

Martins, A. B. T., Bonat, W. H. and Ribeiro, P. J. (2016) Likelihood analysis for a class of spatial geostatistical compositional models. *Spatial Statistics* **17**, 121–130.

Matheron, G. (1963) Principles of geostatistics. *Economic Geology* **58**(8), 1246–1266.

Matérn, B. (1960) *Spatial Variation: Stochastic Models and Their Application to Some Problems in Forst Survey and Other Sampling Investigations*. Esselte.

McLachlan, G. J., Lee, S. X. and Rathnayake, S. I. (2019) Finite Mixture Models. *Annual Review of Statistics and Its Application* **6**(1), 355–378.

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* **21**(1), 48–65.

Milton, A. H., Hussain, S., Akter, S., Rahman, M., Mouly, T. A. and Mitchell, K. (2017) A Review of the Effects of Chronic Arsenic Exposure on Adverse Pregnancy Outcomes. *International Journal of Environmental Research and Public Health* **14**(6), 556.

Mises, R. v. (1954) La distribution de la plus grande de n valeurs. *American Mathematical Society, Providence, RI* **II**, 271–294.

Mishra, S., Bharagava, R. N., More, N., Yadav, A., Zainith, S., Mani, S. and Chowdhary, P. (2019) Heavy Metal Contamination: An Alarming Threat to Environment and Human Health. In *Environmental Biotechnology: For Sustainable Future*, eds R. C. Sobti, N. K. Arora and R. Kothari, pp. 103–125. Singapore: Springer Singapore.

Mitra, S., Chakraborty, A. J., Tareq, A. M., Emran, T. B., Nainu, F., Khusro, A., Idris, A. M., Khandaker, M. U., Osman, H., Alhumaydhi, F. A. and Simal-Gandara, J. (2022) Impact of heavy metals on the environment and human health: Novel therapeutic insights to counter the toxicity. *Journal of King Saud University - Science* **34**(3), 101865.

Moon, K. J., Han, J. S., Ghim, Y. S. and Kim, Y. J. (2008) Source apportionment of fine carbonaceous particles by positive matrix factorization at Gosan background site in East Asia. *Environment International* **34**(5), 654–664.

Morabito, F. C., Simone, G. and Cacciola, M. (2008) 15 - Image fusion techniques for non-destructive testing and remote sensing applications. In *Image Fusion*, ed. T. Stathaki, pp. 367–392. Oxford: Academic Press.

Morais, S., E Costa, F. G. and Lourdes Pereir, M. D. (2012) Heavy Metals and Human Health. In *Environmental Health - Emerging Issues and Practice*, ed. J. Oosthuizen. InTech.

Munir, S., Mayfield, M. and Coca, D. (2021) Understanding spatial variability of no2 in urban areas using spatial modelling and data fusion approaches. *Atmosphere* **12**(2), 1–20.

Naveau, P., Guillou, A., Cooley, D. and Diebolt, J. (2009) Modelling pairwise dependence of maxima in space. *Biometrika* **96**(1), 1–17.

Naveau, P., Huser, R., Ribereau, P. and Hannart, A. (2016) Modeling jointly low, moderate, and heavy rainfall intensities without a threshold selection. *Water Resources Research* **52**(4), 2753–2769.

Neumann, T., Ebendt, R. and Kuhns, G. (2016) From finance to ITS: traffic data fusion based on Markowitz' portfolio theory. *Journal of Advanced Transportation* **50**(2), 145–164.

Nguyen, H., Cressie, N. and Braverman, A. (2012) Spatial Statistical Data Fusion for Remote Sensing Applications. *Journal of the American Statistical Association* **107**(499), 1004–1018.

Niekerk, J. v. and Rue, H. (2024) Low-rank Variational Bayes correction to the Laplace method. *Journal of Machine Learning Research* **25**(62), 1–25.

Noh, S. (2020) Intelligent Data Fusion and Multi-Agent Coordination for Target Allocation. *Electronics* **9**(10), 1563.

Okafor, N. U., Alghorani, Y. and Delaney, D. T. (2020) Improving Data Quality of Low-cost IoT Sensors in Environmental Monitoring Networks Using Data Fusion and Machine Learning Approach. *ICT Express* **6**(3), 220–228.

Padoan, S. A., Ribatet, M. and Sisson, S. A. (2010) Likelihood-Based Inference for Max-Stable Processes. *Journal of the American Statistical Association* **105**(489), 263–277.

Palharini, R. S. A., Vila, D. A., Rodrigues, D. T., Quispe, D. P., Palharini, R. C., de Siqueira, R. A. and de Sousa Afonso, J. M. (2020) Assessment of the Extreme Precipitation by Satellite Estimates over South America. *Remote Sensing* **12**(13), 2085.

Papastathopoulos, I. and Tawn, J. A. (2013) Extended generalised Pareto models for tail estimation. *Journal of Statistical Planning and Inference* **143**(1), 131–143.

Pawlowsky-Glahn, V. and Egozcue, J. J. (2016) Spatial analysis of compositional data: A historical review. *Journal of Geochemical Exploration* **164**, 28–32.

Pendergrass, D., Jacob, D., Zhai, S., Kim, J., Koo, J.-H., Bae, M. and Kim, S. (2021) Continuous Mapping of Fine Particulate Matter (PM2.5) Air Quality in East Asia by Application of a Random Forest Algorithm to GOCI Geostationary Satellite Data **2021**, A13A–06.

Pereira, S., Pereira, P., de Carvalho, M. and de Zea Bermudez, P. (2019) Calibration of extreme values of simulated and real data. In *Proceedings of International Workshop of Statistical Modelling*, volume I, pp. 147–150. INE, Guimaraes: (L. Meira-Machado and G. Soutinho, Eds.).

Pizzimenti, S., Toaldo, C., Pettazzoni, P., Dianzani, M. U. and Barrera, G. (2010) The "Two-Faced" Effects of Reactive Oxygen Species and the Lipid Peroxidation Product 4-Hydroxynonenal in the Hallmarks of Cancer. *Cancers* **2**(2), 338–363.

Pöschl, U. (2005) Atmospheric Aerosols: Composition, Transformation, Climate and Health Effects. *Angewandte Chemie International Edition* **44**(46), 7520–7540.

Ramanathan, V. and Carmichael, G. (2008) Global and regional climate changes due to black carbon. *Nature Geoscience* **1**(4), 221–227.

Ramsay, J. O. and Silverman, B. W. (2006) *Functional data analysis*. Second edition. Springer series in statistics. New York, NY: Springer.

Ran, M., Bai, X., Xin, F. and Xiang, Y. (2018) Research on Probability Statistics Method for Multi-sensor Data Fusion. In *2018 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)*, pp. 406–4065. Zhengzhou, China: IEEE.

Reczek, C. R. and Chandel, N. S. (2017) The Two Faces of Reactive Oxygen Species in Cancer. *Annual Review of Cancer Biology* **1**(Volume 1, 2017), 79–98.

Reeves, R. D., Baker, A. J. M., Jaffré, T., Erskine, P. D., Echevarria, G. and van der Ent, A. (2018) A global database for plants that hyperaccumulate metal and metalloid trace elements. *New Phytologist* **218**(2), 407–411.

Resnick, S. I. (1987) *Extreme Values, Regular Variation and Point Processes*. Springer Series in Operations Research and Financial Engineering. New York, NY: Springer.

Reyes, J. L., Molina-Jijón, E., Rodríguez-Muñoz, R., Bautista-García, P., Debray-García, Y. and Namorado, M. d. C. (2013) Tight Junction Proteins and Oxidative Stress in Heavy Metals-Induced Nephrotoxicity. *BioMed Research International* **2013**, 730789.

Ribeiro Sales, M. H., Souza, C. M. and Kyriakidis, P. C. (2013) Fusion of MODIS Images Using Kriging With External Drift. *IEEE Transactions on Geoscience and Remote Sensing* **51**(4), 2250–2259.

Richards, J. and Wadsworth, J. L. (2021) Spatial deformation for nonstationary extremal dependence. *Environmetrics* **32**(5), e2671.

Rohrbeck, C., Tawn, J. A. and Simpson, E. S. (2023) Editorial: EVA 2023 Data Challenge. *Extremes* .

Rootzén, H. and Nader, T. (2006) Multivariate Generalized Pareto Distributions. *Bernoulli* **12**(5), 917–930.

Rootzén, H., Segers, J. and L. Wadsworth, J. (2018a) Multivariate peaks over thresholds models. *Extremes* **21**(1), 115–145.

Rootzén, H., Segers, J. and Wadsworth, J. L. (2018b) Multivariate generalized Pareto distributions: Parametrizations, representations, and properties. *Journal of Multivariate Analysis* **165**, 117–131.

Rousseau, M.-C., Parent, M.-E., Nadon, L., Latreille, B. and Siemiatycki, J. (2007) Occupational exposure to lead compounds and risk of cancer among men: a population-based case-control study. *American Journal of Epidemiology* **166**(9), 1005–1014.

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.

Rue, H., Riebler, A., Sørbye, S. H., Illian, J. B., Simpson, D. P. and Lindgren, F. K. (2017) Bayesian Computing with INLA: A Review. *Annual Review of Statistics and Its Application* **4**(1), 395–421.

Ruggieri, F., Saavedra, J., Fernandez-Turiel, J. L., Gimeno, D. and Garcia-Valles, M. (2010) Environmental geochemistry of ancient volcanic ashes. *Journal of Hazardous Materials* **183**(1), 353–365.

Ryou, H. g., Heo, J. and Kim, S.-Y. (2018) Source apportionment of PM10 and PM2.5 air pollution, and possible impacts of study characteristics in South Korea. *Environmental Pollution* **240**, 963–972.

Scarrott, C. and MacDonald, A. (2012) A Review of Extreme Value Threshold Estimation and Uncertainty Quantification. *REVSTAT-Statistical Journal* **10**(1), 33–60.

Schlather, M. (2003) A dependence measure for multivariate and spatial extreme values: Properties and inference. *Biometrika* **90**(1), 139–156.

Schmidt, A. M. and Gelfand, A. E. (2003) A Bayesian coregionalization approach for multivariate pollutant data. *Journal of Geophysical Research: Atmospheres* **108**(D24).

Schneider, P., Castell, N., Vogt, M., Dauge, F. R., Lahoz, W. A. and Bartonova, A. (2017) Mapping urban air quality in near real-time using observations from low-cost sensors and model information. *Environment International* **106**, 234–247.

Schulze, F., Gao, X., Virzonis, D., Damiati, S., Schneider, M. R. and Kodzius, R. (2017) Air Quality Effects on Human Health and Approaches for Its Assessment through Microfluidic Chips. *Genes* **8**(10), 244.

Sharma, A., Kumar, V., Shahzad, B., Tanveer, M., Sidhu, G. P. S., Handa, N., Kohli, S. K., Yadav, P., Bali, A. S., Parihar, R. D., Dar, O. I., Singh, K., Jasrotia, S., Bakshi, P., Ramakrishnan, M., Kumar, S., Bhardwaj, R. and Thukral, A. K. (2019) Worldwide pesticide usage and its impacts on ecosystem. *SN Applied Sciences* **1**(11), 1446.

Sharma, S. and Mujumdar, P. P. (2022) Modeling concurrent hydroclimatic extremes with parametric multivariate extreme value models. *Water Resources Research* **58**(2).

Sheridan, S. C., Lee, C. C. and Smith, E. T. (2020) A Comparison Between Station Observations and Reanalysis Data in the Identification of Extreme Temperature Events. *Geophysical Research Letters* **47**(15), e2020GL088120.

Shoaib, M., Bosch, S., Incel, O. D., Scholten, H. and Havinga, P. J. M. (2014) Fusion of Smartphone Motion Sensors for Physical Activity Recognition. *Sensors* **14**(6), 10146–10176.

Simpson, D., Rue, H., Riebler, A., Martins, T. G. and Sørbye, S. H. (2017) Penalising Model Component Complexity: A Principled, Practical Approach to Constructing Priors. *Statistical Science* **32**(1).

Sklar, M. (1959) Fonctions de repartition an dimensions et leurs marges. *Publ. inst. statist. univ. Paris* **8**, 229–231.

Smith, R. (1997) Markov chain models for threshold exceedances. *Biometrika* **84**(2), 249–268.

Snider, G., Weagle, C. L., Murdymootoo, K. K., Ring, A., Ritchie, Y., Stone, E., Walsh, A., Akoshile, C., Anh, N. X., Balasubramanian, R., Brook, J., Qonitan, F. D., Dong, J., Griffith, D., He, K., Holben, B. N., Kahn, R., Lagrosas, N., Lestari, P., Ma, Z., Misra, A., Norford, L. K., Quel, E. J., Salam, A., Schichtel, B., Segev, L., Tripathi, S., Wang, C., Yu, C., Zhang, Q., Zhang, Y., Brauer, M., Cohen, A., Gibson, M. D., Liu, Y., Martins, J. V., Rudich, Y. and Martin, R. V. (2016) Variation in global chemical composition of $PM_{2.5}$: emerging results from SPARTAN. *Atmospheric Chemistry and Physics* **16**(15), 9629–9653.

Sterckeman, T., Douay, F., Proix, N., Fourrier, H. and Perdrix, E. (2002) Assessment of the Contamination of Cultivated Soils by Eighteen Trace Elements Around Smelters in the North of France. *Water, Air, and Soil Pollution* **135**(1), 173–194.

Ståhl, G., Gobakken, T., Saarela, S., Persson, H. J., Ekström, M., Healey, S. P., Yang, Z., Holmgren, J., Lindberg, E., Nyström, K., Papucci, E., Ulvdal, P., Ørka, H. O., Næsset, E., Hou, Z., Olsson, H. and McRoberts, R. E. (2024) Why ecosystem characteristics predicted from remotely sensed data are unbiased and biased at the same time – and how this affects applications. *Forest Ecosystems* **11**, 100164.

Su, C., Jiang, L. and Zhang, W. (2014) A review on heavy metal contamination in the soil worldwide: Situation, impact and remediation techniques. *Environmnetal Skeptics and Critics* **3**(2), 24–38.

Tang, J., Zhang, J., Ren, L., Zhou, Y., Gao, J., Luo, L., Yang, Y., Peng, Q., Huang, H. and Chen, A. (2019) Diagnosis of soil contamination using microbiological indices: A review on heavy metal pollution. *Journal of Environmental Management* **242**, 121–130.

Tellez-Plaza, M., Guallar, E., Howard, B. V., Umans, J. G., Francesconi, K. A., Goessler, W., Silbergeld, E. K., Devereux, R. B. and Navas-Acien, A. (2013) Cadmium Exposure and Incident Cardiovascular Disease. *Epidemiology (Cambridge, Mass.)* **24**(3), 421–429.

Tipping, M. E. and Bishop, C. M. (1999) Probabilistic Principal Component Analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **61**(3), 611–622.

Toulemonde, G., Ribereau, P. and Naveau, P. (2015) Applications of Extreme Value Theory to Environmental Data Analysis. In *Extreme Events*, pp. 7–21. American Geophysical Union (AGU).

Tóth, G., Hermann, T., Szatmári, G. and Pásztor, L. (2016) Maps of heavy metals in the soils of the European Union and proposed priority areas for detailed assessment. *Science of The Total Environment* **565**, 1054–1062.

Vallikannu, R., Kanpur Rani, V., Kavitha, B. and Sankar, P. (2023) An Analysis of Situational Intelligence for First Responders in Military. In *2023 International Conference on Artificial Intelligence and Applications (ICAIA) Alliance Technology Conference (ATCON-1)*, pp. 1–4.

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.

Varga, L., Rakonczai, P. and Zempléni, A. (2016) Applications of threshold models and the weighted bootstrap for Hungarian precipitation data. *Theoretical and Applied Climatology* **124**(3), 641–652.

Vaziri, N. D. (2008) Mechanisms of lead-induced hypertension and cardiovascular disease. *American Journal of Physiology. Heart and Circulatory Physiology* **295**(2), H454–465.

Viana, M., Pandolfi, M., Minguillón, M. C., Querol, X., Alastuey, A., Monfort, E. and Celades, I. (2008) Inter-comparison of receptor models for PM source apportionment: Case study in an industrial area. *Atmospheric Environment* **42**(16), 3820–3832.

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.

Wadsworth, J. L. and Tawn, J. A. (2012) Dependence modelling for spatial extremes. *Biometrika* **99**(2), 253–272.

Wan, X., Lei, M. and Chen, T. (2016) Cost–benefit calculation of phytoremediation technology for heavy-metal-contaminated soil. *Science of The Total Environment* **563-564**, 796–802.

Wang, F. and Wall, M. M. (2003) Generalized common spatial factor model. *Biostatistics (Oxford, England)* **4**(4), 569–582.

Wang, Y., Hu, X., Chang, H. H., Waller, L. A., Belle, J. H. and Liu, Y. (2018) A Bayesian Downscaler Model to Estimate Daily PM2.5 Levels in the Conterminous US. *International Journal of Environmental Research and Public Health* **15**(9), 1999.

Webster, R. and Oliver, M. A. (1992) Sample adequately to estimate variograms of soil properties. *Journal of Soil Science* **43**(1), 177–192.

Webster, R. and Oliver, M. A. (2007) *Geostatistics for environmental scientists*. Second edition. Statistics in practice. Chichester: Wiley.

Weglarczyk, S., Strupczewski, W. G. and Singh, V. P. (2005) Three-parameter discontinuous distributions for hydrological samples with zero values. *Hydrological Processes* **19**(15), 2899–2914.

WHO (2021) WHO global air quality guidelines. Particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide. Technical report, World Health Organization, Geneva: World Health Organization.

Wilkie, C. J., Miller, C. A., Scott, E. M., O'Donnell, R. A., Hunter, P. D., Spyrakos, E. and Tyler, A. N. (2019) Nonparametric statistical downscaling for the fusion of data of different spatiotemporal support. *Environmetrics* **30**(3), e2549.

Wilkie, C. J., Scott, E. M., Miller, C., Tyler, A. N., Hunter, P. D. and Spyrakos, E. (2015) Data Fusion of Remote-sensing and In-lake chlorophylla Data Using Statistical Downscaling. *Procedia Environmental Sciences* **26**, 123–126.

Wu, W. and Zhang, Y. (2018) Effects of particulate matter (PM2.5) and associated acidity on ecosystem functioning: response of leaf litter breakdown. *Environmental Science and Pollution Research International* **25**(30), 30720–30727.

Xie, X., Semanjski, I., Gautama, S., Tsiligianni, E., Deligiannis, N., Rajan, R. T., Pasveer, F. and Philips, W. (2017) A Review of Urban Air Pollution Monitoring and Exposure Assessment Methods. *ISPRS International Journal of Geo-Information* **6**(12), 389.

Xin, X., Shentu, J., Zhang, T., Yang, X., Baligar, V. C. and He, Z. (2022) Sources, Indicators, and Assessment of Soil Contamination by Potentially Toxic Metals. *Sustainability* **14**(23), 15878.

Xu, X., Tao, S., Huang, L., Du, J., Liu, C., Jiang, Y., Jiang, T., Lv, H., Lu, Q., Meng, Q., Wang, X., Qin, R., Liu, C., Ma, H., Jin, G., Xia, Y., Kan, H., Lin, Y., Shen, R. and Hu, Z. (2022) Maternal PM2.5 exposure during gestation and offspring neurodevelopment: Findings from a prospective birth cohort study. *Science of The Total Environment* **842**, 156778.

Xu, Z., dos Muchangos, L. S., Ito, L. and Tokai, A. (2023) Cost and health benefit analysis of remediation alternatives for the heavy-metal-contaminated agricultural land in a Pb–Zn mining town in China. *Journal of Cleaner Production* **397**, 136503.

Xue, J., Leung, Y. and Fung, T. (2017a) A Bayesian Data Fusion Approach to Spatio-Temporal Fusion of Remotely Sensed Images. *Remote Sensing* **9**(12), 1310.

Xue, T., Zheng, Y., Geng, G., Zheng, B., Jiang, X., Zhang, Q. and He, K. (2017b) Fusing Observational, Satellite Remote Sensing and Air Quality Model Simulated Data to Estimate Spatiotemporal Variations of PM2.5 Exposure in China. *Remote Sensing* **9**(3), 221.

Yaglom, A. M. (1987) *Correlation Theory of Stationary and Related Random Functions.* Volume I: Basic Results. New York: Springer-Verlag.

Yang, P., Drohan, P. J. and Yang, M. (2020) Patterns in soil contamination across an abandoned steel and iron plant: Proximity to source and seasonal wind direction as drivers. *CATENA* **190**, 104537.

Yang, X., Gao, Y., Li, Q., He, J., Liu, Y., Duan, K., Xu, X. and Ji, D. (2023) Maritime and coastal observations of ambient PM2.5 and its elemental compositions in the Bohai Bay of China during spring and summer: Levels, spatial distribution and source apportionment. *Atmospheric Research* **293**, 106897.

Yi, S.-M. and Hwang, I. (2014) Source Identification and Estimation of Source Apportionment for Ambient PM10 in Seoul, Korea. *Asian Journal of Atmospheric Environment* **8**(3), 115–125.

Youngman, B. D. (2019) Generalized Additive Models for Exceedances of High Thresholds With an Application to Return Level Estimation for U.S. Wind Gusts. *Journal of the American Statistical Association* **114**(528), 1865–1879.

Youngman, B. D. (2022) evgam: An R Package for Generalized Additive Extreme Value Models. *Journal of Statistical Software* **103**(3), 1–26.

Yu, L., Liou, I. W., Biggins, S. W., Yeh, M., Jalikis, F., Chan, L.-N. and Burkhead, J. (2019) Copper Deficiency in Liver Diseases: A Case Series and Pathophysiological Considerations. *Hepatology Communications* **3**(8), 1159–1165.

Yuan, F. and Zhan, H. (2022) Stock Market Investment Behavior Based on Behavioral Finance Based on Data Fusion Algorithm. *IETE Journal of Research* **0**(0), 1–7.

Zaiss, D. M. W., Gause, W. C., Osborne, L. C. and Artis, D. (2015) Emerging functions of amphiregulin in orchestrating immunity, inflammation, and tissue repair. *Immunity* **42**(2), 216–226.

Zambelli, B., Uversky, V. N. and Ciurli, S. (2016) Nickel impact on human health: An intrinsic disorder perspective. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* **1864**(12), 1714–1731.

Zefferino, R., Piccoli, C., Ricciardi, N., Scrima, R. and Capitanio, N. (2017) Possible Mechanisms of Mercury Toxicity and Cancer Promotion: Involvement of Gap Junction Intercellular Communications and Inflammatory Cytokines. *Oxidative Medicine and Cellular Longevity* **2017**, 7028583.

Zhang, L., Zhang, L., Teng, W. and Chen, Y. (2013) Based on Information Fusion Technique with Data Mining in the Application of Finance Early-Warning. *Procedia Computer Science* **17**, 695–703.

Zhang, Y., Ma, R., Ban, J., Lu, F., Guo, M., Zhong, Y., Jiang, N., Chen, C., Li, T. and Shi, X. (2021) Risk of Cardiovascular Hospital Admission After Exposure to Fine Particulate Pollution. *Journal of the American College of Cardiology* **78**(10), 1015–1024.

Zhao, J., Jing, X., Yan, Z. and Pedrycz, W. (2021) Network traffic classification for data fusion: A survey. *Information Fusion* **72**, 22–47.

Zhu, L., Zhang, L., Wang, J. and Lv, J. (2021) Combining finite mixture distribution, receptor model, and geostatistical simulation to evaluate heavy metals pollution in soils: Source and spatial pattern. *Land Degradation & Development* **32**(6), 2105–2115.

Zißner, P., Rettore, P. H. L., Santos, B. P., Loevenich, J. F. and Lopes, R. R. F. (2023) DataFITS: A Heterogeneous Data Fusion Framework for Traffic and Incident Prediction. *IEEE Transactions on Intelligent Transportation Systems* **24**(10), 11466–11478.