Received: 16 June 2021 | Accepted: 30 March 2022

DOI: 10.1111/rssc.12565

ORIGINAL ARTICLE

ROYAL STATISTICAL SOCIETY | Series C Applied Statistics

A nearest-neighbour Gaussian process spatial factor model for censored, multi-depth geochemical data

Tilman M. Davies[1] | **Sudipto Banerjee**[2] | **Adam P. Martin**[3] | **Rose E. Turnbull**[3]

[1]Department of Mathematics & Statistics, University of Otago, Dunedin, New Zealand

[2]Department of Biostatistics, University of California Los Angeles, Los Angeles, USA

[3]GNS Science, Dunedin, New Zealand

Correspondence
Tilman M. Davies, Department of Mathematics & Statistics, University of Otago, Dunedin, New Zealand.
Email: tilman.davies@otago.ac.nz

Abstract

We investigate the relationships between local environmental variables and the geochemical composition of the Earth in a region spanning over 26,000 km^2 in the lower South Island of New Zealand. Part of the Southland–South Otago geochemical baseline survey—a pilot study pre-empting roll-out across the country—the data comprise the measurements of 59 chemical trace elements, each at two depth prescriptions, at several hundred spatial sites. We demonstrate construction of a hierarchical spatial factor model that captures inter-depth dependency; handles imputation of left-censored readings in a statistically principled manner; and exploits sparse approximations to Gaussian processes to deliver inference. The voluminous results provide a novel impression of the underlying processes and are presented graphically via simple web-based applications. These both confirm existing knowledge and provide a basis from which new research hypotheses in geochemistry might be formed.

KEYWORDS

Bayesian model, geochemistry, geostatistics, multivariate outcome, spatial prediction

1 | INTRODUCTION

The distribution of elements in soil, rock, water and atmosphere is of fundamental importance to human health, animal well-being, plant growth and mineral formation and as such, affects quality of life and economic prosperity. The study of the distribution, concentration and circulation of chemical elements in the environment is known as *geochemistry* (Goldschmidt, 1954). Geochemistry studies are applied at all scales from atomic to planetary and have been used to study the near-surface distribution and concentration of elements in soil at a regional (<100,000 km^2), national and even global scale (Plant et al., 2001). Such studies are commonly referred to as *geochemical baseline soil studies* and were initially undertaken in support of mineral exploration (Johnson et al., 2005). Geochemical baseline soil studies have since evolved to address research questions relevant in areas of public health (Plant et al., 2001; Turnbull et al., 2019), environmental regulation (Darnley et al., 1995), forensic studies (Reimann & de Caritat, 2012), soil fertility (Clare, 1981), pollution (Deely et al., 1992; Martin et al., 2018), agriculture (Martin et al., 2017; Webber, 1981), forestry, water supply and irrigation (Purchase & Fergusson, 1986) and transport and urbanisation (Fergusson et al., 1980). They have now been undertaken on all continents and at a variety of scales (Herselman et al., 2005; Matschullat et al., 2012; National Soil Survey Office, 1998; Reimann & de Caritat, 2012; Rogers et al., 2017; Smith et al., 2013). In all these studies, a common aim is the provision of a spatial description of the geochemical variability within the study area.

1.1 | Role of statistics and common challenges

Both parametric and nonparametric techniques are frequently used in the analysis of geochemical data. Multivariate exploratory techniques such as ANOVA, independent component analysis, principal component analysis, linear discriminant analysis and multidimensional scaling are popular (Grunsky & de Caritat, 2019; Lado et al., 2008; Singer & Kouda, 2001). Also commonly used are inverse-distance weighted interpolation, kriging and regression-kriging (Bartier & Keller, 1996; Lado et al., 2008; Martin et al., 2016; Reimann, 2005a, 2005b). Establishing and attempting to explain 'class membership', the identification of subgroups of similarly dispersed elements, is often of interest. To that end, model-based clustering, logistic regression and machine learning methods such as support vector machines, random forests and neural networks have also been deployed (van den Berg et al., 2006; Grunsky & de Caritat, 2019; Rissmann et al., 2019).

Due to practical limitations, geochemical baseline soil survey data often contain samples with element concentrations below the lower method detection limit (Farnham et al., 2002), yielding left-censored observations. This is challenging for many statistical treatments which require a complete dataset. Many studies use crude substitution, for example replacement by half the lower limit value (Clarke, 1998) or simple imputation techniques such as mean substitution or marginal regression on a subset of predictors (see Graham, 2012; Sanford et al., 1993; Singer & Kouda, 2001, for examples in geochemistry). The potential bias inherent in such approaches is unappealing from both inferential and predictive perspectives.

Other challenges to the interpretation of geochemical data include missing values, merging, levelling different datasets (reconciling soil samples analysed at different laboratories at different times), adequate spatial coverage, sample design and closure (Grunsky & de Caritat, 2019). Perhaps more crucially, the previous statistical analyses have generally disregarded comparing geochemical samples taken at multiple depths, despite the routine collection of those

data. Depending on the specific research question of interest, this can be profoundly detrimental to making inference—information on geochemical variation between depths can shed light on anthropogenic contamination from, for example, agricultural, vehicular or industrial sources. Geochemistry studies typically assume a source of anthropogenic contamination when the topsoil is rich in certain heavy metals (i.e. lead, arsenic, cadmium) relative to the subsoil. Such an assumption may be based on expertly derived thresholds, for example, the topsoil–subsoil difference method (Kapička et al., 2001), or an arbitrary figure such as readings above 1.5 times the interquartile range of a regional dataset, designed to account for natural background variation (e.g. the geoaccumulation index method; Müller, 1979).

The high dimensionality of the response variable is also a challenge of note. Even those methods that have readily available multivariate analogues (e.g. kriging) often become unwieldy as the number of outcomes under examination grows. Many studies have therefore simply restricted attention to a relatively modest subset of elements (e.g. <10 elements) and many models tend to be spatially coarse (e.g. grids of square-kilometres in interpolated models) due to the increased computational expense associated with finer-scale, higher-dimensional modelling (although exceptions do exist, e.g. Lado et al., 2008).

In part, the result of the aforementioned challenges of geochemical survey data has left a need to consider models that can cope with the complexities therein in a more 'self-contained' fashion. While the breadth of research questions relevant to the analysis of geochemical data does not support a one-size-fits-all approach, there remains tremendous inferential value in constructing models that can simultaneously cope with high-dimensional response variables, multi-depth observations and censoring, while facilitating other desirable operations such as point-specific predictions and the associated uncertainty in a statistically principled manner.

1.2 | Current focus: The Southland–South Otago geochemical baseline survey

In New Zealand, a geochemical baseline soil survey—the Southland–South Otago geochemical baseline survey—was undertaken over the southern region of the country to map the distribution and concentration of dozens of chemical trace elements in soil for human health, pollution and mineral exploration purposes (Martin et al., 2016; Rattenbury et al., 2014). It covers approximately 26,000 km^2; encompassing an area marginally greater than Wales. Being a single survey with samples analysed consecutively on one instrument at one laboratory, some of the aforementioned limitations of studying regional geochemical baseline soil data, like merging or levelling multiple datasets, are not present. Furthermore, by comparison to the known bedrock geology, land use, mineralisation and anthropogenic pollution, the study is also assumed to be of adequate spatial coverage to detect regional-scale element patterns (Martin et al., 2016, 2018). However, other previously mentioned challenges—high-dimensional response, multi-depth observations, censoring—remain. It is of interest to accommodate, not ignore, these complexities.

Our overarching goal is to formulate a model capable of (a) quantifying to what degree the extraneous variables (e.g. climate, rock type, urbanisation) control the spatial distribution of trace elements in soil at two depths; and (b) confidently predicting element in soil patterns in data-poor areas. Doing so is important for informing local planning and making science-based policy decisions. For example, predictions of how element trends might evolve under changing rainfall

expected with climate change, or how varying land use (agriculture, urban) and land cover (native vegetation, pasture) will affect soil element concentrations at different depths, provides crucial scientific insight. The other important point is spatial scale. Making predictions of how elements in soil patterns vary at the regional scale has limited use with respect to making science-based policy decisions at a suburb scale, or even an individual property/paddock scale, which is a primary goal of policy makers and geochemists. In this light, constructing a model capable of fine-scale spatial predictions has obvious value.

Generally speaking, joint modelling of multiple spatially dependent outcomes requires developing multivariate spatial processes. If the number of outcomes is small or moderate, then the *linear model of coregionalisation* (LMC—see Gelfand et al., 2004; Wackernagel, 2003) is a popular method for building valid multivariate spatial processes. The LMC forms each element of a multivariate spatial process as a linear combination of a collection of latent spatial processes that, while dependent over space, are independent of each other. The association among the elements of the multivariate spatial process is determined by the coefficients in the linear transformation.

If the number of outcomes is very large, then dimension reduction is sought. This leads to *spatial factor models*. Spatial factor models have been explored in different contexts including by Wang and Wall (2003), Lopes et al. (2008), Ren and Banerjee (2013) and Taylor-Rodriguez et al. (2019). While the aforementioned approaches differ in the precise formulation of the multivariate process, they are all variants of extending the usual factor models to spatial models by modelling the factors as independent spatial processes. This is equivalent to an LMC in spatial modelling such that the number of latent spatial processes is much less than the number of outcomes.

A unique complication of our current application is presented by the multi-depth nature of geochemical survey data. The models we devise here will further extend the concept of the spatial factor model in a 'multistage' or 'multiresolution' sense, with factors specific to each depth. Of particular interest is precisely how one might design or impose any inter-depth relationship between factors. As we shall see, this can be guided by both prevailing scientific knowledge and the research objectives at hand.

1.3 | Article layout

The balance of the article is structured as follows. In Section 2 we provide some basic plots of the data and perform some simple exploratory analyses. In Section 3 our focus turns to design of a spatial factor model geared to explain the variation in the multiple elements across space and depth. The fitting algorithm and some practical decisions needed to produce the final fit are described in Section 4. Key results are reported in Sections 5 and 6 offers some concluding remarks and details of future research objectives.

2 | DATA AND EXPLORATORY ANALYSIS

The dataset motivating the current work is drawn from the Southland–South Otago geochemical baseline survey of New Zealand (Martin et al., 2016; Rattenbury et al., 2014), named after the regions of the country the sampling sites straddle. Figure 1 shows the study region and $n = 333$ sampling locations in the South Island.

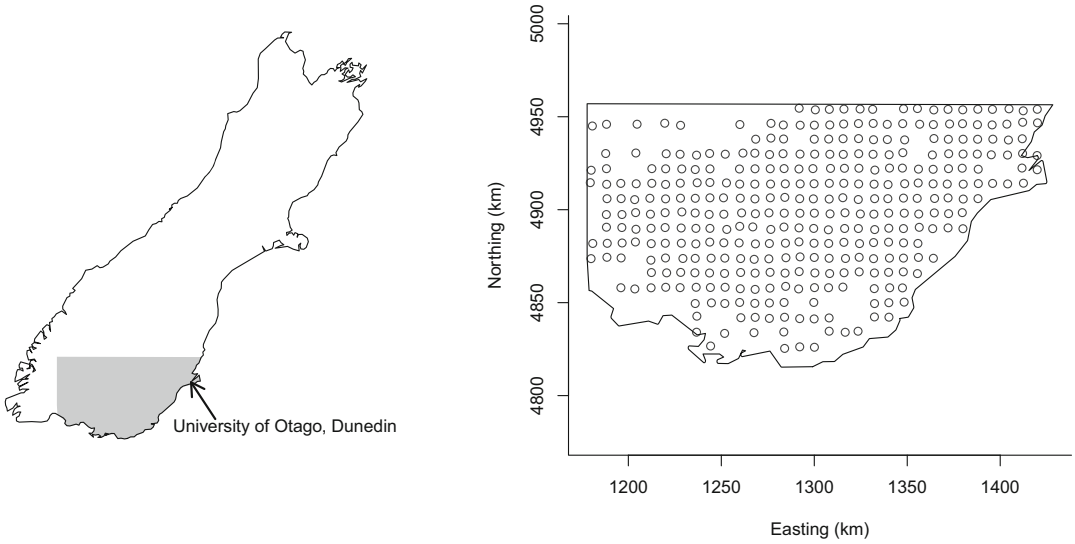

FIGURE 1 Study region limits in the mainland South Island of Aotearoa/New Zealand (left), and spatial locations of 333 sampling sites using the New Zealand Transverse Mercator coordinate system scaled to kilometres

2.1 | Response data

The survey grid design followed international protocols for regional geochemical baseline design (Darnley et al., 1995; Rawlins et al., 2012). Gaps and minor irregularities in the survey design were unavoidable due to geographic barriers (e.g. swamps, mountains) and land access permission challenges. Soil samples were hand-dug using a coring tool. At each site, extraction was performed at $m = 2$ different depths, shallow (Depth A, 0–20 cm) and deep (Depth B, 50–70 cm). Spatial locations were recorded using GPS, and subsequently projected to New Zealand Transverse Mercator (NZTM) coordinates and scaled to kilometres. All soil samples were analysed for a suite of major, minor and trace elements using the inductively coupled plasma mass spectrometry (ICP-MS) technique on an aqua regia partial digest. Analyses were undertaken by Bureau Veritas Minerals Laboratories (BVML) in Vancouver, Canada. Prior to analysis, each sample was sieved to extract the sub-180 μm fraction, and a split of 0.5 g from this fraction was taken for acid digestion. To ensure the quality and accuracy of the data, a comprehensive quality assurance/quality control (QAQC) programme was completed involving regular analysis of duplicates, replicates and survey soil standards, as well as international reference materials and an in-house laboratory pulp duplicate and blank. For a comprehensive outline of sample preparation and analytical methodology, and QAQC protocols, the reader is referred to Martin et al. (2016). A total of 65 elements were measured which, for reasons noted below, we express in log-parts-per-billion (log-ppb).

As is typical in these surveys, the sensitivity of the equipment and procedures used to obtain the measurements is not limitless, yielding a lower-bound limit-of-detection, or *method detection limit* (MDL), which is specific to each element. In many cases the MDL is surpassed, resulting in left-censored outcomes. For the purposes of our analysis, we cull from the dataset a small number of elements that are excessively (i.e. almost 100%) censored, and focus on readings of the remaining $q = 59$ distinct elements.

Thus, $N = mnq = 39{,}294$ individual depth-site-element outcomes are targeted for modelling. Of these $N_{\text{cens}} = 1884$ are left-censored due to the respective MDLs. There are only $N_{\text{miss}} = 3$

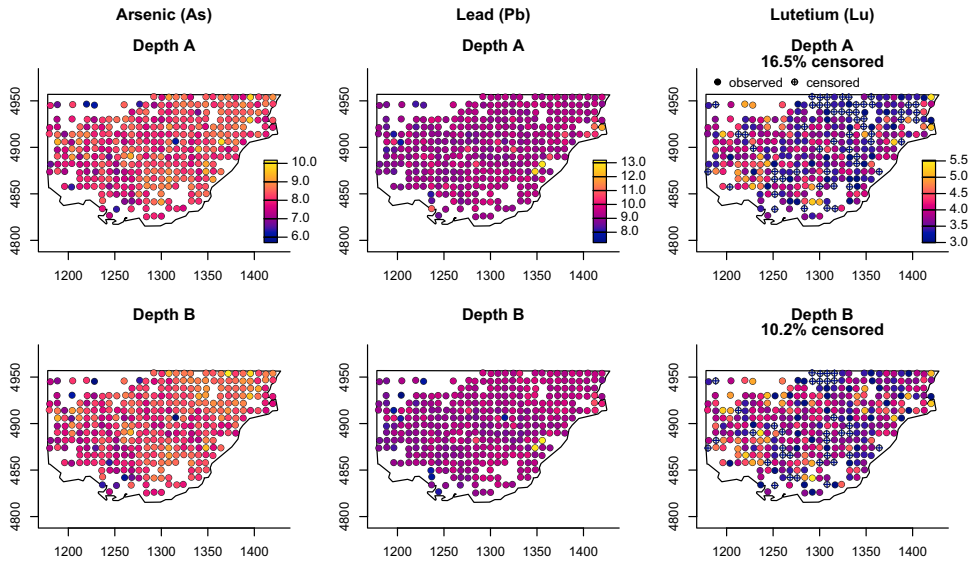

FIGURE 2 Responses for arsenic, lead and lutetium (left to right respectively). Common colour scale between depths; units in log-ppb. Filled dots (•) represent observed data and crossed circles (⊕) denote left-censored values [Colour figure can be viewed at wileyonlinelibrary.com]

missing values, all of which correspond to carbon (C). The 59 elements of interest are listed in greater detail in section 1 of Supplement A.

Figure 2 gives an example of the response data for arsenic (As); lead (Pb); and the rare-earth element lutetium (Lu). Censoring is apparent in the latter, and each of these shows strong spatial similarity between depths—a feature exhibited by all elements in the dataset.

To further investigate the scaling of the responses, Figure 3 provides comparative histograms of the distributions of the untransformed uncensored Depth A readings of As and Lu and their log-transformed versions. An additional histogram in Figure 3 shows the maximum-likelihood estimates of the Box–Cox power parameter $\lambda^\star$, computed separately for each element at both depths (after omitting censored/missing values). This distribution of estimates is centred close to zero with a mean of approximately 0.00593. As such, we settled for use of log-ppb in contrast to something more general like the Box–Cox transform with a non-zero $\lambda^\star$, as the log transform (a) serves to satisfactorily remove notable right-skewness in the distributions of the element readings; (b) estimation of the 'power' parameter $\lambda^\star$ for the Box–Cox function does not indicate a systematic or substantial favour for non-zero values of $\lambda^\star$ (cf. Figure 3); and (c) the log transform is commonly used and familiar in the geochemical sciences (see e.g. Aitchison et al., 2000; Gazley et al., 2020; McKinley et al., 2016).

2.2 | Environmental variables

Alongside the trace element readings, various environmental variables are obtained at each sampling site. Estimating the association between the response variables and these predictors forms an important part of the analysis, both to provide an overall fixed-effect adjustment and to quantify the relationships with the individual variables in their own right.

Figure 4 displays the seven variables at hand: stratigraphy; lithology; soil type; vegetation type; terrain slope; mean daily air temperature; and mean annual rainfall. The first four of these are

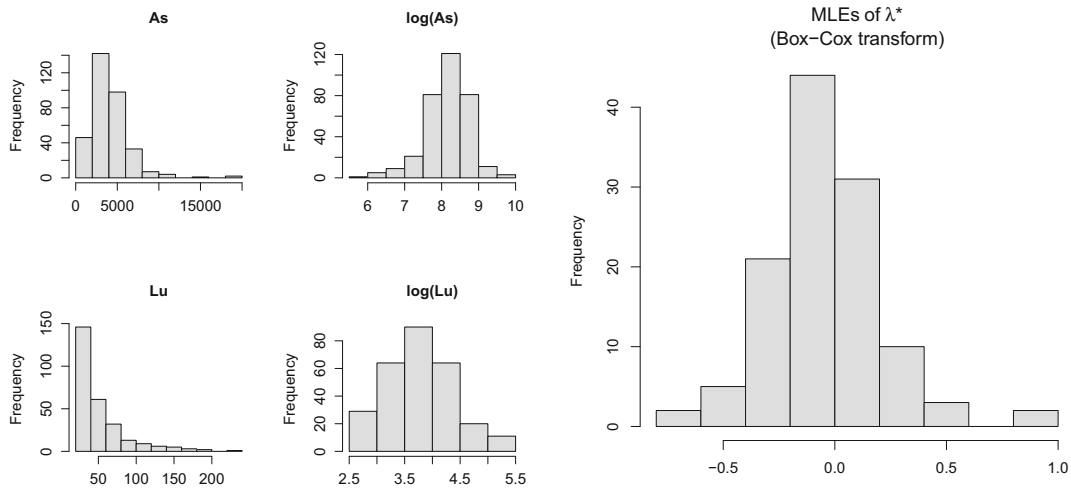

FIGURE 3 Distributions of the (uncensored) untransformed Depth A responses for As and Lu, contrasted with their log-transformed counterparts (left, panel of four histograms), and the collection of maximum likelihood estimates of $\lambda^\star$ for the Box–Cox transform applied to all elements at both depths (right, large histogram)

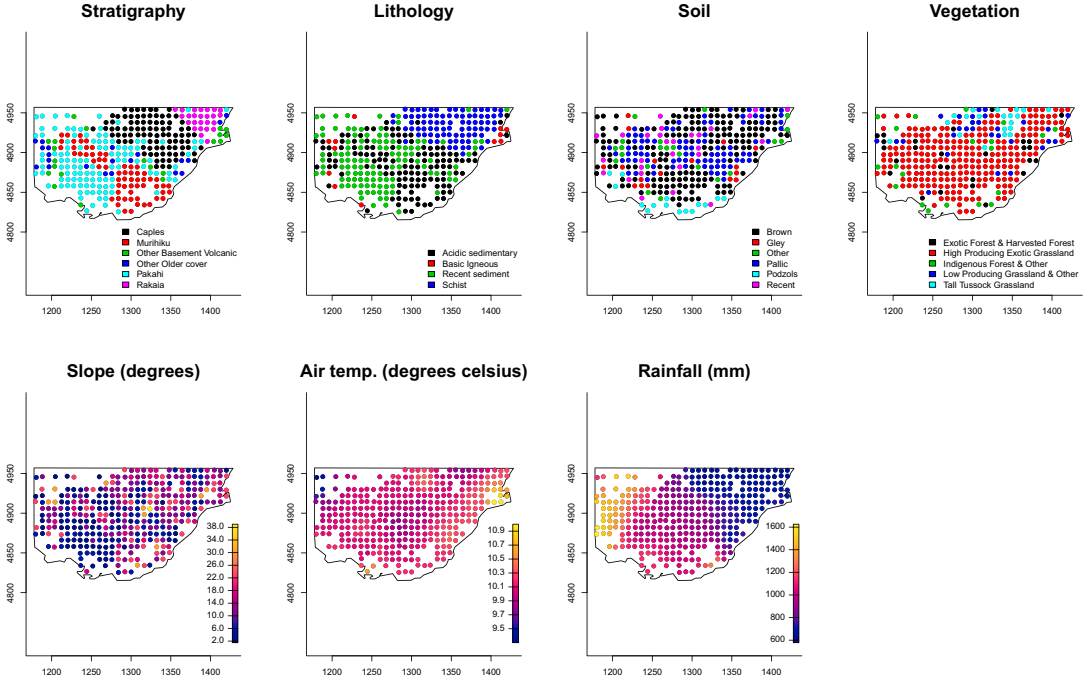

FIGURE 4 Seven environmental variables obtained at each sampling site; those in the bottom row are treated as continuous [Colour figure can be viewed at wileyonlinelibrary.com]

categorical variables and the remaining three are treated as continuous. In section 2 of Supplement A we offer some commentary on their meaning and interpretation in the current context.

2.3 | Principal component, hierarchical clustering and variogram analyses

Visualisation of the raw data as in Figure 2 reflects existing knowledge in geochemistry. Not only do we tend to observe very similar spatial variation between depths, but many different elements themselves share similar spatial distributions. This is due to a variety of reasons, both environmental and anthropogenic.

Elementary exploratory techniques reveal such structure. A dendrogram resulting from Euclidean-distance hierarchical clustering appears in Figure 5. This clearly illustrates the first clusters are formed, for the most part, by the observations at both depths for a specific element. The graph also shows that the $mq = 118$ elements at both depths collapse into relatively few groups, relatively quickly thereafter.

Scree plots of a PCA appear in Figure 6. The PCA is performed separately on the depth-specific residuals following a fixed-effect linear regression to remove the potential influence of the environmental variables detailed in Section 2.2. This analysis provides another two pieces of useful information. First, it supports the interpretation of structure evident in the dendrogram—relatively few components (up to around 10) describe a substantial proportion of the (estimated residual) variance in the 59 responses. Second, the striking similarity of the curves corresponding to the two depths—this hints at simplified or shared modelling strategies for the between-depth relationship; ideas we pursue in the model design itself.

Lastly, we interrogate several individual elements using variogram models. Unsurprisingly, even after accounting for the environmental predictors (using the same linear regression trend as in the PCA), all elements display lingering spatial dependence as judged by sustained breaches of 95% envelopes computed via Monte-Carlo permutation—it is acknowledged in the geochemical literature that the variables contributing to the spatial variation in these trace elements are numerous and complex (Martin et al., 2016; Sparks, 2003). A typical example of this is

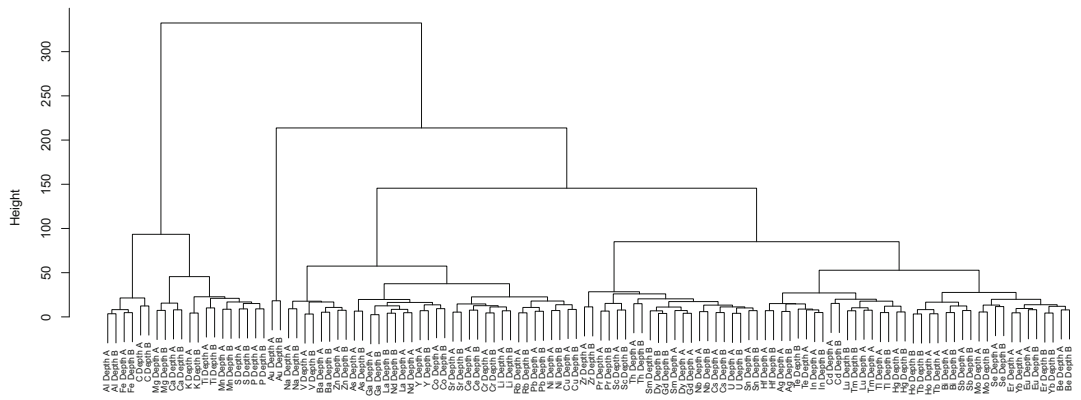

FIGURE 5 Cluster dendrogram of the response data

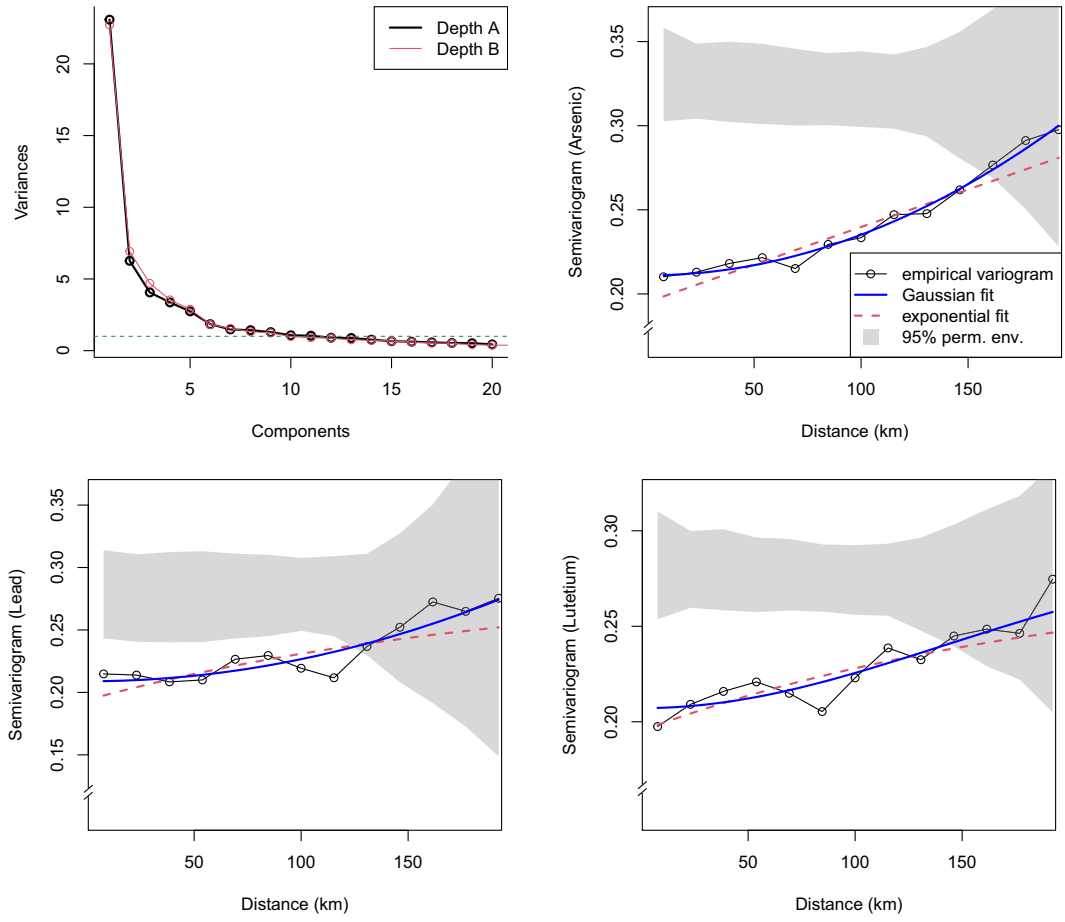

FIGURE 6 Top left: Scree plots of PCA for each depth, after adjusting for the environmental variables in a linear regression; unit variance is marked with a dotted horizontal line. Top right, bottom left, bottom right respectively: Variogram analysis of the Depth A responses for arsenic, lead and lutetium [Colour figure can be viewed at wileyonlinelibrary.com]

provided in the top-right, bottom-left and bottom-right panels side of Figure 6, which shows the estimated empirical variogram of the data for As, Pb and Lu at Depth A (corresponding to the data in the top row of Figure 2). Such analyses also permit simple explorations of the nature of the spatial autocorrelation. Superimposed atop the empirical variogram for each of the three elements are two variogram models fitted via least squares—one using an exponential covariance function, the other using a Gaussian (squared exponential) covariance. While both fitted curves succeed in picking up the overall dependence, the Gaussian fit is somewhat better able to adapt to stronger dependence at small distances, resulting in a better match to the empirical variogram—behaviour we note is mimicked by the other elements.

As a check for anisotropy, we additionally explore directional variograms. The estimated semivariances in the four cardinal directions are sufficiently similar such that the omnidirectional curves are taken as sensible summaries of the underlying processes. Using the three example elements discussed in this section, we provide their directional variograms in section 3 of Supplement A.

3 | MODEL DESIGN

We propose a hierarchical model for multivariate geochemical observations as outlined in Section 2.1, with q outcomes/elements each measured at m depths across n sites $\mathcal{S} = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$. While censoring is permitted, as is missingness, for ease of notation we shall for the moment ignore such entries—they are dealt with at the model fitting stage in a typical data augmentation step described in Section 4.2. Furthermore, although our motivating dataset has particular features of note (e.g. we need only deal with two depths, and environmental variables are the same at each depth and for each element), here we present the model in a slightly more general form for the sake of exposition.

We follow customary notations. Let $\mathrm{N}_d(\boldsymbol{a}, \mathbf{B})$ denote a normal distribution of dimension d with mean and variance/covariance structures $\boldsymbol{a}$ and $\mathbf{B}$ respectively; and let $\mathrm{N}_d(\mathbf{x}|\boldsymbol{a}, \mathbf{B})$ be the corresponding normal density evaluated at $\mathbf{x}$. The quantity $\mathbf{0}$ denotes a vector of zeros with size clear by context; $\mathbf{I}_d$ the $d \times d$ identity matrix; $\mathbf{1}[\cdot]$ the indicator function; $\oplus$ the matrix direct sum; and $\otimes$ the Kronecker product.

3.1 | A factor framework

Let $Y_i^{(j)}(\mathbf{s})$ represent the measurement corresponding to the ith element at the jth depth and generic spatial location $\mathbf{s}$. We assume the following observation equation

$$Y_i^{(j)}(\mathbf{s}) = \mathbf{x}_i^{(j)}(\mathbf{s})^\top \boldsymbol{\beta}_i^{(j)} + \sum_{\ell=1}^{r} \lambda_{i\ell}^{(j)} f_\ell^{(j)}(\mathbf{s}) + \epsilon_i^{(j)}(\mathbf{s}); \qquad i = 1, \ldots, q; \quad j = 1, \ldots, m. \tag{1}$$

Here, each $\mathbf{x}_i^{(j)}(\mathbf{s})$ is the $p_i \times 1$ vector of covariates (including an intercept) at site $\mathbf{s}$ with the hth element giving the value of $x_{ih}^{(j)}(\mathbf{s})$ for $h = 1, 2, \ldots, p_i$. Indexing by i and j reflects the fact that we could allow for the collection of explanatory variables to be different for each possible outcome and each possible depth if necessary. This is paired with a corresponding $p_i \times 1$ vector of regression coefficients $\boldsymbol{\beta}_i^{(j)}$. Given the large value of q, we achieve dimension reduction through latent factors and their corresponding loads. Thus, variation beyond those explained by the covariates are captured through a linear combination of $r \ll q$ factors $f_\ell^{(j)}(\mathbf{s})$. The values for the ℓth factor at depth j receive an outcome-specific load of $\lambda_{i\ell}^{(j)}$. These factors are considered in greater detail in Section 3.2. Residual measurement error is left to $\epsilon_i^{(j)}(\mathbf{s}) \overset{ind}{\sim} N(0, \delta_i^{2(j)})$, assumed to be independent across outcomes i and depths j.

Collecting (1) over the q elements yields the vector-valued model,

$$\mathbf{Y}^{(j)}(\mathbf{s}) = \mathbf{X}^{(j)}(\mathbf{s})\boldsymbol{\beta}^{(j)} + \boldsymbol{\Lambda}^{(j)}\mathbf{f}^{(j)}(\mathbf{s}) + \boldsymbol{\epsilon}^{(j)}(\mathbf{s}); \qquad j = 1, \ldots, m, \tag{2}$$

where $\mathbf{Y}^{(j)}(\mathbf{s}) = [Y_1^{(j)}(\mathbf{s}), \ldots, Y_q^{(j)}(\mathbf{s})]^\top$ is $q \times 1$, $\mathbf{X}^{(j)}(\mathbf{s}) = \bigoplus_{i=1}^{q} \mathbf{x}_i^{(j)}(\mathbf{s})^\top$ is the $q \times p$ (with $p = \sum_i p_i$) design matrix, $\boldsymbol{\beta}^{(j)} = [\boldsymbol{\beta}_1^{(j)\top}, \ldots, \boldsymbol{\beta}_q^{(j)\top}]^\top$ is the $p \times 1$ vector of coefficients, $\boldsymbol{\Lambda}^{(j)} = [\boldsymbol{\lambda}_1^{(j)}, \ldots, \boldsymbol{\lambda}_r^{(j)}]$ with $\boldsymbol{\lambda}_\ell^{(j)} = [\lambda_{1\ell}^{(j)}, \ldots, \lambda_{q\ell}^{(j)}]^\top$ is the $q \times r$ factor loading matrix, $\mathbf{f}^{(j)}(\mathbf{s}) = [f_1^{(j)}(\mathbf{s}), \ldots, f_r^{(j)}(\mathbf{s})]^\top$ is the $r \times 1$ collection of factor values and $\boldsymbol{\epsilon}^{(j)}(\mathbf{s}) = [\epsilon_1^{(j)}(\mathbf{s}), \ldots, \epsilon_q^{(j)}(\mathbf{s})]^\top$ is the $q \times 1$ vector of errors with $\boldsymbol{\epsilon}^{(j)}(\mathbf{s}) \sim \mathrm{N}_q(\mathbf{0}, \mathbf{D}^{(j)})$, where $\mathbf{D}^{(j)} = \bigoplus_{i=1}^{q} \delta_i^{2(j)} = \mathrm{diag}[\delta_1^{2(j)}, \ldots, \delta_q^{2(j)}]$. We also define $\boldsymbol{\delta}^{2(j)} = \{\delta_1^{2(j)}, \ldots, \delta_q^{2(j)}\}$.

Proceeding with further concatenation over the locations in $\mathcal{S}$, we obtain

$$\mathbf{Y}^{(j)} = \mathbf{X}^{(j)}\boldsymbol{\beta}^{(j)} + (\mathbf{I}_n \otimes \boldsymbol{\Lambda}^{(j)})\mathbf{f}^{(j)} + \boldsymbol{\epsilon}^{(j)}; \qquad j = 1, \ldots, m, \tag{3}$$

where $\mathbf{Y}^{(j)} = [\mathbf{Y}^{(j)}(\mathbf{s}_1)^\top, \ldots, \mathbf{Y}^{(j)}(\mathbf{s}_n)^\top]^\top$ is the $nq \times 1$ vector obtained by concatenating $\mathbf{Y}^{(j)}(\mathbf{s}_k)$ for $k = 1, 2, \ldots, n$. The $nq \times p$ design matrix is given by concatenating the $\mathbf{X}^{(j)}(\mathbf{s}_k)$s, namely $\mathbf{X}^{(j)} = [\mathbf{X}^{(j)}(\mathbf{s}_1)^\top, \ldots \mathbf{X}^{(j)}(\mathbf{s}_n)^\top]^\top$, giving an $nq \times p$ structure. The factors are arranged similarly to form the $nr \times 1$ vector $\mathbf{f}^{(j)} = [\mathbf{f}^{(j)}(\mathbf{s}_1)^\top, \ldots, \mathbf{f}^{(j)}(\mathbf{s}_n)^\top]^\top$ operated on by the $nq \times nr$ block diagonal loading matrix $\mathbf{I}_n \otimes \boldsymbol{\Lambda}^{(j)}$. Lastly we also concatenate the error terms to give $\boldsymbol{\epsilon}^{(j)} = [\boldsymbol{\epsilon}^{(j)}(\mathbf{s}_1)^\top, \ldots, \boldsymbol{\epsilon}^{(j)}(\mathbf{s}_n)^\top]^\top$ as the $nq \times 1$ vector of residuals, with $\boldsymbol{\epsilon}^{(j)} \sim \mathrm{N}_{nq}(\mathbf{0}, \mathbf{I}_n \otimes \mathbf{D}^{(j)})$. Finally, we collect the responses over depth to obtain

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\Lambda}\mathbf{f} + \boldsymbol{\epsilon}, \tag{4}$$

where $\mathbf{Y}(\boldsymbol{\beta})$ is $mnq \times 1$ ($mp \times 1$) obtained by concatenating the $\mathbf{Y}^{(j)}$s ($\boldsymbol{\beta}^{(j)}$s) in Equation (3), $\mathbf{X} = \bigoplus_{j=1}^{m}\mathbf{X}^{(j)}$ is $mnq \times mp$, $\boldsymbol{\Lambda} = \bigoplus_{j=1}^{m}\{\mathbf{I}_n \otimes \boldsymbol{\Lambda}^{(j)}\}$ is $mnq \times mnr$, $\mathbf{f} = [\mathbf{f}^{(1)\top}, \ldots, \mathbf{f}^{(m)\top}]^\top$ is $mnr \times 1$, and $\boldsymbol{\epsilon} \sim \mathrm{N}_{mnq}(\mathbf{0}, \mathbf{D})$, where $\boldsymbol{\epsilon}$ is $mnq \times 1$ formed by concatenating $\boldsymbol{\epsilon}^{(j)}$s in Equation (3), $\mathbf{D} = \bigoplus_{j=1}^{m}\left\{\mathbf{I}_n \otimes \mathbf{D}^{(j)}\right\}$ is $mnq \times mnq$ diagonal comprising elements in $\boldsymbol{\delta}^2 = \{\boldsymbol{\delta}^{2(1)}, \ldots, \boldsymbol{\delta}^{2(m)}\}$. For ease of interpretation, in section 4.1 of Supplement A we provide some more explicit matrix diagrams of the various structures defined here.

3.2 | Spatial factors and model identifiability

The preceding formulation presents a very flexible modelling framework. The dimension reduction, which manifests as the $(q \times r) \times (r \times 1)$ second term on the right-hand side of (2), offers computational savings when r is small relative to q. More importantly, in the current application where sub-groups of different elements are known to exhibit similar spatial trends, the fundamental statistical premise of a factor model is also intuitively sensible from a scientific standpoint.

We introduce spatially correlated factors in Equation (1) by specifying

$$f_\ell^{(j)}(\mathbf{s}) \sim \mathrm{GP}\left(0, C^{(j)}(\cdot, \cdot; \boldsymbol{\phi}_\ell)\right); \qquad \ell = 1, \ldots, r; \quad j = 1, \ldots, m, \tag{5}$$

where 'GP' denotes *Gaussian process* (a spatial process where, for any finite collection of spatial sites $\mathcal{T} \in \mathbb{R}^2$, the vector $\mathbf{f}_\ell^{(j)}(\mathbf{t})$ for $\mathbf{t} \in \mathcal{T}$ has a corresponding multivariate normal distribution), following the notation of Banerjee et al. (2014), Finley et al. (2007). These processes are taken to be mutually independent with respect to ℓ. Centred about zero, we have $\mathrm{Cov}[f_\ell^{(j)}(\mathbf{s}), f_\ell^{(j)}(\mathbf{s}')] = C^{(j)}(\mathbf{s}, \mathbf{s}'; \boldsymbol{\phi}_\ell)$, where $C^{(j)}(\cdot, \cdot; \boldsymbol{\phi}_\ell)$ is a positive definite spatial covariance function controlled by parameter $\boldsymbol{\phi}_\ell$. The r independent factors are therefore represented as a *multivariate* Gaussian process (MVGP)

$$\mathbf{f}^{(j)}(\mathbf{s}) \sim \mathrm{MVGP}(\mathbf{0}, C^{(j)}(\cdot, \cdot; \boldsymbol{\phi})) \tag{6}$$

with $\boldsymbol{\phi} = \{\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_r\}$. Here, we have a matrix-valued multivariate cross-covariance function in C, where $C^{(j)}(\mathbf{s}, \mathbf{s}'; \boldsymbol{\phi})$ yields the $r \times r$ dependence between the r factors at locations $\mathbf{s}$ and $\mathbf{s}'$. The aforementioned independence between different factors implies $C^{(j)}(\mathbf{s}, \mathbf{s}'; \boldsymbol{\phi})$ is an $r \times r$ diagonal

matrix with ℓth diagonal entry $C^{(j)}(\mathbf{s}, \mathbf{s}'; \boldsymbol{\phi}_\ell)$ for $\ell = 1, 2, \ldots, r$. From Equation (6), we obtain $\mathbf{f}^{(j)} \sim \mathrm{N}_{nr}(\mathbf{0}, \boldsymbol{\Gamma}_{\boldsymbol{\phi}}^{(j)})$ for $j = 1, 2, \ldots, m$, where $\boldsymbol{\Gamma}_{\boldsymbol{\phi}}^{(j)}$ is the $nr \times nr$ covariance matrix with the $r \times r$ matrix $C^{(j)}(\mathbf{s}_{k_1}, \mathbf{s}_{k_2}; \boldsymbol{\phi})$ as the (k_1, k_2)th block for $k_1, k_2 = 1, 2, \ldots, n$.

However, the flexibility we are afforded by such a model is not without pitfalls. To wit, the key issue of identifiability requires careful treatment; workarounds depend on the specifications for the various model components. Identifiability ensures substantive interpretation of latent factors. Violation of identifiability means that certain latent factors cannot be uniquely extracted from data casting questions on their substantive interpretation. It is well known (see e.g. Anderson, 2003) that a latent factor model is not identifiable due to rotational indeterminacy, that is, the distribution of the zero-centred outcomes is invariant to orthogonal transformations (rotations) of the latent factors and the loading matrix. With independent factors, a widely used approach is to fix certain elements of the loading matrix to constant values, usually to zeroes, such as restricting the loading matrix to be an upper or lower triangular matrix with strictly positive diagonal elements (Lopes & West, 1999). When factors are modelled as spatial processes, Ren and Banerjee (2013) argue that non-identifiability is restricted to sign-alternating reflectors and permutations and impose an ordering on the spatial decay parameters to ensure identifiability in theory. In practice, however, such methods need not be effective in uniquely identifying the individual components when numerical separation of some of the spatial decay parameters is violated. Predictive inference is still robust, but interpretation of the underlying associations among the outcomes can be unreliable.

Recent work by Zhang and Banerjee (2021) proposes classes of spatial factor models using a Bayesian matrix-normal formulation. While computationally attractive for predictive inference, such methods do not easily accommodate the richer structures in the factors that we desire in our application. In our case prediction is naturally still of interest. However, it is also desirable to gain a sense of the overall strength of between-depth association in these residual spatial effects, owing to the scientific interest in both environmental and anthropogenic inputs. Hence, in Section 3.3 we introduce additional structure in the spatial factors to account for association across depths while recognising the challenges of identifying all the factor loadings from such additional structure. Instead, we conduct a PCA analysis (cf. Figure 6) as a part of preliminary exploration and fix the factor loading matrix using the eigenvectors obtained from the PCA analysis.

3.3 | Multi-depth construction

We offer further structure to the specification of the factors as we seek to leverage and quantify the between-depth similarity. First, focus on the first (shallow) depth, $j = 1$, which from Section 3.2 we have as $\mathbf{f}^{(1)} \sim \mathrm{N}_{nq}(\mathbf{0}, \boldsymbol{\Gamma}_{\boldsymbol{\phi}}^{(1)})$. In

$$\boldsymbol{\Gamma}_{\boldsymbol{\phi}}^{(1)} = \left[C^{(1)}(\mathbf{s}_{k_1}, \mathbf{s}_{k_2}; \boldsymbol{\phi})\right]_{k_1, k_2 = 1, \ldots, n} = \left[\bigoplus_{\ell=1}^{r} C^{(1)}(\mathbf{s}_{k_1}, \mathbf{s}_{k_2}; \boldsymbol{\phi}_\ell)\right]_{k_1, k_2 = 1, \ldots, n}$$

we assign unit variance; that is, $C^{(1)}(\mathbf{s}, \mathbf{s}'; \boldsymbol{\phi}_\ell)$ is a spatial *correlation* function. In addition, we assume stationarity and isotropy thereof; the latter assumption supported by the exploratory analysis (see the directional variogram examples in Supplement A). The Matérn class of functions (see

e.g. Chilés & Delfiner, 2012) is popular in practice. Given the findings of our exploratory analysis, we opt for the Gaussian correlation function of the form

$$C^{(1)}(\mathbf{s},\mathbf{s}';\boldsymbol{\phi}_\ell) = \exp\left\{-(\phi_\ell\|\mathbf{s}-\mathbf{s}'\|)^2\right\}; \qquad \phi_\ell > 0, \tag{7}$$

which we found to be better suited to the sample data than the exponential correlation function. Note $\|\mathbf{s}-\mathbf{s}'\|$ denotes Euclidean distance between $\mathbf{s}$ and $\mathbf{s}'$ and, in such a case, the parameter $\boldsymbol{\phi}_\ell = \phi_\ell$ is simply a single scalar component. A slightly more relaxed approach would be to implement the general Matérn covariance function itself, although this would require an additional smoothness parameter be sampled as part of the fitting algorithm. We prefer to avoid this route in the current application given the computational cost of involving additional parameters in the global dependence structure, coupled with the fact that this particular parameter is often poorly identified in practice (see e.g. Stein, 1999).

We now propose to model the factors at subsequent depths $j > 1$ as a function of the values at depth $j = 1$. Specifically,

$$\begin{aligned} f_\ell^{(j)}(\mathbf{s}) &= \alpha_j f_\ell^{(1)}(\mathbf{s}) + \eta_\ell^{(j)}(\mathbf{s}); \qquad \ell = 1,\ \ldots\ ,r; \quad j = 2,\ \ldots\ ,m; \\ \Rightarrow \mathbf{f}^{(j)}(\mathbf{s}) &= \alpha_j \mathbf{f}^{(1)}(\mathbf{s}) + \boldsymbol{\eta}^{(j)}(\mathbf{s}), \end{aligned} \tag{8}$$

where additional independent error is given by $\eta_\ell^{(j)}(\mathbf{s}) \sim \mathrm{N}_1(0, \sigma_\ell^{2(j)})$. Hence,

$$\boldsymbol{\eta}^{(j)}(\mathbf{s}) \sim \mathrm{N}_r(\mathbf{0}, \mathbf{S}^{(j)}) \tag{9}$$

where $\boldsymbol{\eta}^{(j)}(\mathbf{s}) = [\eta_1^{(j)}(\mathbf{s}),\ \ldots\ ,\eta_r^{(j)}(\mathbf{s})]^\top$ and $\mathbf{S}^{(j)} = \bigoplus_{\ell=1}^r \sigma_\ell^{2(j)}$ is $r \times r$ diagonal; note also that $\boldsymbol{\eta}^{(\cdot)}(\mathbf{s})$ is independent of $\mathbf{f}^{(j)}(\mathbf{s})$ for all j and we write $\boldsymbol{\sigma}^2 = \{\boldsymbol{\sigma}^{2(2)},\ \ldots\ ,\boldsymbol{\sigma}^{2(m)}\}$ where $\boldsymbol{\sigma}^{2(j)} = \{\sigma_1^{2(j)},\ \ldots\ ,\sigma_r^{2(j)}\}$.

Turn to the jth-depth factor values across all sites; that is, the $nr \times 1$ vector $\mathbf{f}^{(j)}$ used in the model (3). Coupling (6) with (8) we have

$$\begin{aligned} \mathrm{Cov}[\mathbf{f}^{(j)}(\mathbf{s}), \mathbf{f}^{(j)}(\mathbf{s}')] &= \mathrm{Cov}\left[\alpha_j \mathbf{f}^{(1)}(\mathbf{s}) + \boldsymbol{\eta}^{(j)}(\mathbf{s}), \alpha_j \mathbf{f}^{(1)}(\mathbf{s}') + \boldsymbol{\eta}^{(j)}(\mathbf{s}')\right] \\ &= \alpha_j^2 C^{(1)}(\mathbf{s},\mathbf{s}';\boldsymbol{\phi}) + \mathbf{S}^{(j)}\mathbf{1}[\mathbf{s}=\mathbf{s}'], \end{aligned}$$

and therefore by considering $\mathbf{s}, \mathbf{s}' \in \mathcal{S}$,

$$\mathbf{f}^{(j)} \sim \mathrm{N}_{nr}\left(\mathbf{0}, \alpha_j^2 \boldsymbol{\Gamma}_{\boldsymbol{\phi}}^{(1)} + \mathbf{I}_n \otimes \mathbf{S}^{(j)}\right); \qquad j = 2,\ \ldots\ ,m. \tag{10}$$

Thus, to complete the full covariance matrix for the entire collection of factor values at all depths, $\mathbf{f}$ as appears in Equation (4), we must also obtain the between-depth factor dependencies by applying $\mathrm{Cov}[\mathbf{f}^{(j_1)}(\mathbf{s}), \mathbf{f}^{(j_2)}(\mathbf{s}')]$ for $j_1, j_2 \in 1,\ \ldots\ ,r; j_1 \neq j_2$. Doing so it is easy to show that

$$\mathbf{f} \sim \mathrm{N}_{mnr}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\alpha},\boldsymbol{\phi},\boldsymbol{\sigma}^2}), \tag{11}$$

where

$$\boldsymbol{\Sigma}_{\boldsymbol{\alpha},\boldsymbol{\phi},\boldsymbol{\sigma}^2} = \boldsymbol{\alpha}\boldsymbol{\alpha}^\top \otimes \boldsymbol{\Gamma}_{\boldsymbol{\phi}}^{(1)} + \left[\mathbf{O}_{nr} \oplus \left\{\bigoplus_{j=2}^m \left(\mathbf{I}_n \otimes \mathbf{S}^{(j)}\right)\right\}\right] \tag{12}$$

for $\boldsymbol{\alpha} = [1, \alpha_2, \ldots, \alpha_m]^\top$ and $\mathbf{O}_{nr}$ denoting a $nr \times nr$ matrix of zeros. As we have done earlier, Section 4.2 of Supplement A elucidates on the matrix structures and other quantities defined in this section. We remark that our construction in Equation (8) bears similarities with dynamic spatial–temporal models in which there is a substantial literature (see, e.g. Stroud et al., 2001). While identifiability of the α_js for Bayesian inference will not require additional constraints as long as proper priors are assigned, we will use proper priors with positive support to reflect the plausibility of positive associations among the depth-specific spatial processes (see Equation (16) in Section 3.5).

The culmination of the preceding design is a collection of spatial factors in $\mathbf{f}$ that are permitted to co-vary in a particular way. The above definitions make it clear that for a given factor ℓ, the spatial effect at any depth $j > 1$ is linearly related (with slope controlled by α_j) to the ℓth spatial effect at the first depth $j = 1$. This relationship, however, does not extend to informing changes to other factors (in other words, the ℓ_1th factor at depth $j = 1$ does not inform the ℓ_2th factor; $\ell_1 \neq \ell_2$, at any depth).

As noted in Section 3.2, achieving practically meaningful identifiability with such structure in the factors $\mathbf{f}$ requires consideration of the loading matrix. In a purely non-spatial unit-variance setting, that is, simply defining $f_\ell^{(j)}(\mathbf{s}) \overset{\text{ind}}{\sim} \mathrm{N}_1(0, 1)$ for any given depth j, the within-site cross-covariance would be wholly determined by $\boldsymbol{\Lambda}^{(j)}\boldsymbol{\Lambda}^{(j)\top}$. However, under the between-depth prescription described above, the covariance structure evident in Equation (10) for the factors $\mathbf{f}^{(j)}$ at depths $j > 1$ confounds this relationship and ensuring identifiability while retaining meaningful interpretation becomes challenging. Hence, we assume that the spatial effects for any given element at depths subsequent to $j = 1$ are strongly associated, and fix $\boldsymbol{\Lambda}^{(j)}$ for all j using the PCA analysis shown in Figure 6 at depth $j = 1$. Let $\mathbf{E}_1$ be the $n \times q$ matrix of zero-centred residuals following a standard fixed-effect regression of the $j = 1$ depth observations, with each column scaled to have unit variance. Then, with $\mathbf{V}$ representing the $q \times q$ matrix whose columns are the eigenvectors of $\mathbf{E}_1^\top\mathbf{E}_1$, for each $j = 1, \ldots, m$ we fix each $q \times r$ matrix $\boldsymbol{\Lambda}^{(j)}$ to take on the first r columns of $\mathbf{V}$; each column multiplied by the square root of its corresponding eigenvalue. This corresponds to an optimal choice for the loading matrix within the context of probabilistic PCA (Tipping & Bishop, 1999); also see section 12.2 in Bishop (2006). This allows us to distil the overall between-depth relationship for all elements using the parameters in $\boldsymbol{\alpha}$, where α_j models the association for the factors at depth j with that at depth 1 (taken as a baseline to construct the joint distribution of the factors).

3.4 | Sparsity-inducing NNGP

Motivated by the computational intractability of geostatistical modelling problems with a large number of observations, the work by Datta et al. (2016a) introduced the *nearest-neighbour Gaussian process* (NNGP), which is designed as a direct approximation to a corresponding GP. The NNGP offers massive computational savings and scalability while providing an accurate and precise representation of the full GP. The basic idea behind the technique is to approximate the dependence structure of the full GP using only the measurements in the immediate spatial neighbourhood of a given location $\mathbf{s}$. This 'immediate neighbourhood' is defined in terms of a directed acyclic graph (DAG)—locations are ordered, then the spatially nearest ν observations with indexes less than that corresponding to $\mathbf{s}$ are identified. It transpires that the result is a sparsity-induced 'reconstruction' of the Cholesky decomposition of the target covariance matrix. See Datta et al. (2016a, 2016b) for a detailed technical treatment.

While the number of sites n in the present application is certainly modest, the high dimensionality of the response q coupled with the fact we seek to model at m depths compounds the computational task ahead. The key bottleneck lies in Equation (11), with repeated decomposition of the $mnr \times mnr$ matrix $\mathbf{\Sigma}_{\boldsymbol{\alpha},\boldsymbol{\phi},\boldsymbol{\sigma}^2}$ required for in-turn access to all the parameters in $\{\boldsymbol{\alpha}, \boldsymbol{\phi}, \boldsymbol{\sigma}^2\}$. Even for small r, we found this to be computationally prohibitive using the full specification. To this end, we turn to the NNGP to assist in the model fitting.

Given the r-dimensional multivariate outcome for the factors, we seek to replace the MVGP that leads to Equation (11) with a matching multivariate NNGP (MVNNGP). Ultimately, this permits us to rewrite (11) as a product of independent r-dimensional normals—this is the backbone of the computational savings.

We make use of the walkthrough provided in Banerjee (2017) for the univariate NNGP, as well as the model definitions and justifications in Taylor-Rodriguez et al. (2019) for the MVNNGP, to appropriately form the approximation for the current application. Specifically, the reader is directed to Sections 3.1 and 3.2 of Banerjee (2017) and Sections 3.2 and 3.3 of Taylor-Rodriguez et al. (2019). Additionally, Sections S2 and S3 of the online supplementary material of the latter paper offer key guidance for the blueprint that follows here.

The multi-depth nature of our modelling problem requires a modified approach to defining the local neighbourhood of a given observation. First, define the new index $k^\star \in \mathcal{K}^\star$ with

$$\mathcal{K}^\star = \{k + (j-1)n : k = 1, \ldots, n; j = 1, \ldots, m\}, \tag{13}$$

which references spatial sites in blocks by depth. This augmented index is able to point uniquely to both spatial site *and* depth. The reverse map to depth, $j \leftarrow k^\star$ is given by the function

$$g(k^\star) = \begin{cases} 1 & \text{for } 1 \le k^\star \le n \\ 2 & \text{for } n < k^\star \le 2n \\ \vdots & \\ m & \text{for } (m-1)n < k^\star \le mn \end{cases},$$

and the reverse map to spatial site, $k \leftarrow k^\star$ is found with

$$h(k^\star) = k^\star - (g(k^\star) - 1)n.$$

Let

$$\mathbf{f}^\star(\mathbf{s}_{k^\star}) = \mathbf{f}^{(g(k^\star))}(\mathbf{s}_{h(k^\star)}) = \left[f_1^{(g(k^\star))}(\mathbf{s}_{h(k^\star)}), \ldots, f_r^{(g(k^\star))}(\mathbf{s}_{h(k^\star)})\right]^\top$$

refer to the $r \times 1$ vector of factors for site-depth $k^\star$, and for any set $\mathcal{U} = \{u_1, \ldots, u_T\}$ satisfying $\emptyset \subset \mathcal{U} \subseteq \mathcal{K}^\star$, let

$$\mathbf{f}^\star_U = \left[\mathbf{f}^{(g(u_1))}(\mathbf{s}_{h(u_1)})^\top, \ldots, \mathbf{f}^{(g(u_T))}(\mathbf{s}_{h(u_T)})^\top\right]$$

be the $rT \times 1$ vector of factors at the T site-depth locations in $\mathcal{U}$.

Next, let $N(k^\star) = \{\overline{k}^\star : \overline{k}^\star < k^\star\}$ denote the full set of all *directed* neighbours of site-depth $k^\star$, with cardinality $|N(k^\star)|$. From this, take $\overline{k}^\star_1, \ldots, \overline{k}^\star_{|N(k^\star)|}$ to be the entries of $N(k^\star)$ arranged in increasing order of Euclidean distance to the target site $k^\star$, satisfying

$$\|\mathbf{s}_{h(\bar{k}_1^\star)} - \mathbf{s}_{h(k^\star)}\| \leq \|\mathbf{s}_{h(\bar{k}_2^\star)} - \mathbf{s}_{h(k^\star)}\| \leq \ \ldots \ \leq \|\mathbf{s}_{h(\bar{k}^\star_{|N(k^\star)|})} - \mathbf{s}_{h(k^\star)}\|$$

Note that when equalities occur in the above sequence, we treat the offending positions as interchangeable. Indeed, due to the multi-depth nature of the data, it is quite possible to encounter a $h(\bar{k}^\star_{(\cdot)}) = h(k^\star)$ (in which case the spatial distance is taken as zero, and the between-depth correlation as per the definitions in Section 3.3 ultimately comes into play).

Now, suppose we restrict attention to, at most, the nearest ν neighbours of site-depth $k^\star$ from its set $N(k^\star)$. Define the *reduced* neighbour set as

$$\mathcal{N}_\nu(k^\star) = \left\{ \bar{k}_c^\star \ : \ c = 1, \ \ldots \ , \min\left[\nu, |N(k^\star)|\right] \right\}, \tag{14}$$

and let $\nu_{k^\star} = |\mathcal{N}_\nu(k^\star)| = \min[\nu, |N(k^\star)|]$ be the specific number of restricted neighbours of $k^\star$. Due to Equation (14), note that $\nu_{k^\star}$ is capped at ν for all site-depths, that is, $\max(\nu_1, \ \ldots \ , \nu_{mn}) = \nu$.

The final quantities required for the MVNNGP approximation are the so-called *kriging weights* and *variances*, which we find based on the reduced neighbour set. Let

$$R(k^\star) = \{rk^\star - (r-1) + r' \ : \ r' = 0, \ \ldots \ , r-1\}$$

identify the positions of the vector $\mathbf{f}$ that correspond to the $k^\star$th site-depth; similarly apply $R(\mathcal{N}_\nu(k^\star))$ to identify the positions in turn for each member of the neighbour set. Then a standard appeal to the properties of the multivariate Gaussian distribution yields the kriging weight for $k^\star$ as

$$\mathbf{B}_{k^\star} = \mathbf{\Sigma}_{\boldsymbol{\alpha},\boldsymbol{\phi},\sigma^2}\left[R(k^\star), R\left(\mathcal{N}_\nu(k^\star)\right)\right] \mathbf{\Sigma}^{-1}_{\boldsymbol{\alpha},\boldsymbol{\phi},\sigma^2}\left[R\left(\mathcal{N}_\nu(k^\star)\right), R\left(\mathcal{N}_\nu(k^\star)\right)\right]$$

where for any matrix $\boldsymbol{M}$, $\boldsymbol{M}[A_1, A_2]$ provides the $|A_1| \times |A_2|$ sub-matrix indexed by rows and columns in the sets A_1 and A_2 respectively. The corresponding kriging variance is

$$\boldsymbol{\zeta}_{k^\star} = \mathbf{\Sigma}_{\boldsymbol{\alpha},\boldsymbol{\phi},\sigma^2}\left[R(k^\star), R(k^\star)\right] - \mathbf{B}_{k^\star}\mathbf{\Sigma}_{\boldsymbol{\alpha},\boldsymbol{\phi},\sigma^2}\left[R\left(\mathcal{N}_\nu(k^\star)\right), R(k^\star)\right].$$

Computational accessibility now becomes apparent when we consider the following, based on the density implied by Equation (11):

$$\mathrm{N}_{mnr}(\mathbf{f}|\mathbf{0}, \mathbf{\Sigma}_{\boldsymbol{\alpha},\boldsymbol{\phi},\sigma^2}) \approx \prod_{k^\star=1}^{mn} \mathrm{N}_r(\mathbf{f}^\star(\mathbf{s}_{k^\star})|\mathbf{B}_{k^\star}\mathbf{f}^\star_{\mathcal{N}_\nu(k^\star)}, \boldsymbol{\zeta}_{k^\star}). \tag{15}$$

Due to the enforced restriction on the number of directed neighbours of each site, the quantities on the right-hand side of (15) can be kept to a manageable size for fast and efficient evaluation: $\mathbf{B}_{k^\star}$ is $r \times \nu_{k^\star} r$; $\mathbf{f}^\star_{\mathcal{N}_\nu(k^\star)}$ is $\nu_{k^\star} r \times 1$; and $\boldsymbol{\zeta}_{k^\star}$ is $r \times r$. These can be calculated on-the-fly with no need to construct and invert the full $\mathbf{\Sigma}_{\boldsymbol{\alpha},\boldsymbol{\phi},\sigma^2}$ or requisition large chunks of computer memory, and the independence reflected by the product over the $k^\star$ further allows parallel computations if desired. We consider the choice of ν in Section 4.4.

3.5 | Posterior

We complete our model specification by placing priors on all unknown parameters. The full-data joint posterior is written as

$$
\begin{aligned}
p\left(\boldsymbol{\beta}, \mathbf{f}, \delta^2, \boldsymbol{\alpha}, \boldsymbol{\phi}, \sigma^2 | \mathbf{Y}, \mathbf{X}, \boldsymbol{\Lambda}\right) \propto & \, \mathrm{N}_{mnq}(\mathbf{Y}|\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\Lambda}\mathbf{f}, \mathbf{D}) \times \mathrm{N}_{mp}\left(\boldsymbol{\beta}|\mathbf{0}, c_\beta \mathbf{I}_{mp}\right) \\
& \times \prod_{k^\star=1}^{mn} \mathrm{N}_r\left(\mathbf{f}^\star(\mathbf{s}_{k^\star})|\mathbf{B}_{k^\star}\mathbf{f}^\star_{\mathcal{N}_v(k^\star)}, \boldsymbol{\zeta}_{k^\star}\right) \times \prod_{j=1}^{m}\prod_{i=1}^{q} \mathrm{IG}\left(\delta_i^{2(j)}\middle| a_{\delta^2}, b_{\delta^2}\right) \\
& \times \prod_{j=2}^{m} \mathrm{UNIF}\left(\alpha_j | a_{\boldsymbol{\alpha}}, b_{\boldsymbol{\alpha}}\right) \times \prod_{j=2}^{m}\prod_{\ell=1}^{r} \mathrm{UNIF}\left(\sigma_\ell^{2(j)}\middle| a_{\sigma^2}, b_{\sigma^2}\right) \\
& \times \prod_{\ell=1}^{r} \mathrm{UNIF}(\phi_\ell | a_{\boldsymbol{\phi}}, b_{\boldsymbol{\phi}}),
\end{aligned} \tag{16}
$$

where $\mathrm{IG}(\cdot|a, b)$ is the inverse gamma density with shape and rate parameters a and b respectively; and $\mathrm{UNIF}(\cdot|a, b)$ the uniform density with limits $a < b$.

Choices of normal and inverse gamma priors for $\boldsymbol{\beta}$ and the $\delta_i^{2(j)}$ respectively were made for reasons of conjugacy. We set a vague $c_\beta = 100$ along with $a_{\delta^2} = 2$ and $b_{\delta^2} = 0.18$; with a shape of 2, the inverse gamma has an infinite variance with mean equal to the rate. The value of b_{δ^2} was gleaned from exploratory semivariograms and pilot runs of the algorithm. Uniform priors are suggested for all remaining parameters. A wide $a_{\sigma^2} = 0$ and $b_{\sigma^2} = 100$ interval is chosen for all components of $\boldsymbol{\sigma}^2$ in our analysis. The variable components of $\boldsymbol{\alpha}$ are constrained to be positive, lying between $a_{\boldsymbol{\alpha}} = 0$ and $b_{\boldsymbol{\alpha}} = 2$; our decision on these bounds is to encompass values evenly around 1 (which would indicate factors at depths $j > 1$ are merely a copy of the factors at depth $j = 1$ with some random noise) while forbidding excessive values. Lastly, we choose the range of the $\boldsymbol{\phi}$s to be relevant to the spatial scale of the study region, following the same strategy as in Wang and Wall (2003) and Ren and Banerjee (2013). Given the parametrisation of (7), we set $a_{\boldsymbol{\phi}} = \sqrt{-\log(0.05)/d_{0.9}} = 0.0114$ and $b_{\boldsymbol{\phi}} = \sqrt{-\log(0.01)/d_0} = 0.3480$, where d_0 and $d_{0.9}$ are, respectively, the minimum and 0.9th quantile of the collection of pairwise Euclidean distances between any two sampling sites. These limits provide the 'strongest' and 'weakest' values for any given ϕ_ℓ in terms of spatial dependency of the r factors, found as those which would result in a correlation of 0.05 at $d_{0.9}$ ($a_{\boldsymbol{\phi}}$), and effective independence with a correlation of 0.01 at d_0 ($b_{\boldsymbol{\phi}}$).

4 | IMPLEMENTATION

Here we discuss the necessary ingredients for fitting and predicting from the model described in Section 3, and further discussion on other decisions made in our implementation.

4.1 | Sampling algorithm

We derive a typical Metropolis-within-Gibbs algorithm for model fitting. To start with, the regression coefficients $\boldsymbol{\beta}$ and the residual error variances in $\mathbf{D}$ are updated via Gibbs steps. Their full conditionals are

$$
\boldsymbol{\beta}|\cdots \sim \mathrm{N}_{mp}\left(\mathbf{V}_\beta\left\{\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{D}^{-1}(\mathbf{Y} - \boldsymbol{\Lambda}\mathbf{f})\right\}, \mathbf{V}_\beta\right),
$$

where $\mathbf{V}_\beta^{-1} = \mathbf{X}^\top\mathbf{D}^{-1}\mathbf{X} + c_\beta\mathbf{I}_{mp}$, and

$$\delta_i^{2(j)} | \cdots \sim \text{IG}\left(a_{\delta^2} + \frac{n}{2}, b_{\delta^2} + \frac{1}{2}\sum_{k=1}^{n}\left\{ Y_i^{(j)}(\mathbf{s}_k) - \mathbf{x}_i^{(j)}(\mathbf{s}_k)^\top \boldsymbol{\beta}_i^{(j)} - \sum_{\ell=1}^{r} \lambda_{i\ell}^{(j)} f_\ell^{(j)}(\mathbf{s}_k) \right\}^2 \right)$$

for $i = 1, \ldots, q; j = 1, \ldots, m$.

The factor values in $\mathbf{f}$ are also updated using Gibbs steps; exploiting the MVNNGP representation permits each collection of r values to be updated in turn according to site-depth $k^\star \in \mathcal{K}^\star$. To find the appropriate full conditional distribution, first define

$$\mathbf{B}_{k^\star}^{[c]} = \mathbf{B}_{k^\star}[\{1, \ldots, r\}, \{(c-1)r+1, \ldots, (c-1)r+r\}]$$

as the $r \times r$ sub-matrix of the kriging weight $\mathbf{B}_{k^\star}$ that relates the cth neighbour of $k^\star$, $\bar{k}_c^\star$, to $k^\star$ itself; $c = 1, \ldots, \nu_{k^\star}$. Further define $\mathcal{P}(k^\star) = \{\check{k}^\star : k^\star \in \mathcal{N}_v(\check{k}^\star)\}$ as the set of all site-depths that have $k^\star$ as one of its $\nu_{\check{k}^\star}$ nearest neighbours, and specifically identify $z(k^\star|\check{k}^\star)$ as the index of the set $\mathcal{N}_v(\check{k}^\star)$ that yields the point of interest $k^\star$ (i.e. using the notation of the neighbours from Equation (13), we have $\bar{\check{k}}^\star_{z(k^\star|\check{k}^\star)} = k^\star$). We can then re-write the mean of the r-dimensional normal in Equation (14) as a sum involving the $\mathbf{B}_{k^\star}^{[c]}$, and recognise the expression

$$\begin{aligned}
&\text{N}_q\left(\mathbf{Y}^{(g(k^\star))}(\mathbf{s}_{h(k^\star)}) \middle| \mathbf{X}^{(g(k^\star))}(\mathbf{s}_{h(k^\star)})\boldsymbol{\beta}^{(g(k^\star))} + \boldsymbol{\Lambda}^{(g(k^\star))}\mathbf{f}^\star(\mathbf{s}_{k^\star}), \mathbf{D}^{(g(k^\star))}\right) \\
&\quad \times \text{N}_r\left(\mathbf{f}^\star(\mathbf{s}_{k^\star}) \middle| \sum_{c=1}^{\nu_{k^\star}} \mathbf{B}_{k^\star}^{[c]} \mathbf{f}^\star(\mathbf{s}_{\bar{k}_c^\star}), \boldsymbol{\zeta}_{k^\star}\right) \\
&\quad \times \prod_{\check{k}^\star \in \mathcal{P}(k^\star)} \text{N}_r\left(\mathbf{f}^\star(\mathbf{s}_{\check{k}^\star}) \middle| \mathbf{B}_{\check{k}^\star}^{[z(k^\star|\check{k}^\star)]} \mathbf{f}^\star(\mathbf{s}_{k^\star}) + \sum_{\substack{u=1 \\ u \neq z(\check{k}^\star)}}^{\nu_{\check{k}^\star}} \mathbf{B}_{\check{k}^\star}^{[u]} \mathbf{f}^\star(\mathbf{s}_{\bar{\check{k}}_u^\star}), \boldsymbol{\zeta}_{\check{k}^\star}\right)
\end{aligned}$$

as the portion of the posterior distribution (15) that contains all instances of $\mathbf{f}^\star(\mathbf{s}_{k^\star})$. It can thereafter be shown that the full conditional for $\mathbf{f}^\star(\mathbf{s}_{k^\star})$ is

$$\begin{aligned}
\mathbf{f}^\star(\mathbf{s}_{k^\star}) | \cdots \sim \text{N}_r\Bigg(\mathbf{V}_{k^\star} \Bigg\{ &\boldsymbol{\zeta}_{k^\star}^{-1} \sum_{c=1}^{\nu_{k^\star}} \mathbf{B}_{k^\star}^{[c]} \mathbf{f}^\star(\mathbf{s}_{\bar{k}_c^\star}) + \sum_{\check{k}^\star \in \mathcal{P}(k^\star)} \left(\mathbf{B}_{\check{k}^\star}^{[z(k^\star|\check{k}^\star)]}\right)^\top \boldsymbol{\zeta}_{\check{k}^\star}^{-1} \boldsymbol{\chi}_{\check{k}^\star} \\
&+ \left(\boldsymbol{\Lambda}^{(g(k^\star))}\right)^\top \left(\mathbf{D}^{(g(k^\star))}\right)^{-1} \left(\mathbf{Y}^{(g(k^\star))}(\mathbf{s}_{h(k^\star)}) - \mathbf{X}^{(g(k^\star))}(\mathbf{s}_{h(k^\star)})\boldsymbol{\beta}^{(g(k^\star))}\right) \Bigg\}, \mathbf{V}_{k^\star} \Bigg),
\end{aligned}$$

where

$$\boldsymbol{\chi}_{\check{k}^\star} = \mathbf{f}^\star(\mathbf{s}_{\check{k}^\star}) - \sum_{\substack{u=1 \\ u \neq z(k^\star|\check{k}^\star)}}^{\nu_{\check{k}^\star}} \mathbf{B}_{\check{k}^\star}^{[u]} \mathbf{f}^\star(\mathbf{s}_{\bar{\check{k}}_u^\star})$$

and

$$\mathbf{V}_{k^\star}^{-1} = \boldsymbol{\zeta}_{\check{k}^\star}^{-1} + \sum_{\check{k}^\star \in \mathcal{P}(k^\star)} \left(\mathbf{B}_{\check{k}^\star}^{[z(k^\star|\check{k}^\star)]}\right)^\top \boldsymbol{\zeta}_{\check{k}^\star}^{-1} \mathbf{B}_{\check{k}^\star}^{[z(k^\star|\check{k}^\star)]} + \left(\boldsymbol{\Lambda}^{(g(k^\star))}\right)^\top \left(\mathbf{D}^{(g(k^\star))}\right)^{-1} \boldsymbol{\Lambda}^{(g(k^\star))}.$$

The remaining parameters are updated using Metropolis steps. Per iteration, we require $m(r+1)-1$ separate evaluations of the dependence structure of $\mathbf{f}$ to successfully update all components in $\{\boldsymbol{\alpha}, \boldsymbol{\phi}, \sigma^2\}$. Our MVNNGP approximation greatly eases the computational burden involved.

4.2 | Data augmentation

The model permits handling of censored and missing values in a straightforward manner via data augmentation. Essentially this involves treating any unknown values as additional parameters in the model: we sample their values during fitting using the Gaussian distribution, and 'fill-in' the relevant positions of $\mathbf{Y}$ with these imputed values, yielding the augmented-complete dataset $\mathbf{Y}^*$.

Fridley and Dixon (2007) offer a good description of this procedure for left-censored and missing values in spatial regression which we adapt to the current case. First, let $\{s_1, \ldots, s_{N_{\text{miss}}}\} \subset \{1, \ldots, mnq\}$ and $\{t_1, \ldots, t_{N_{\text{cens}}}\} \subset \{1, \ldots, mnq\}$ identify the N_{miss} and N_{cens} missing and censored index positions of the observed data, respectively, as arranged in $\mathbf{Y}$. Furthermore, let the notation $\mathbf{W}[u]$ extract the uth element from any given vector $\mathbf{W}$. Now consider the oth iteration of the fitting algorithm. Once all the model parameters have been updated, we sample as

$$\mathbf{Y}_{(o+1)}^*[s] \sim \mathrm{N}_1\left(\left\{\mathbf{X}\boldsymbol{\beta}_{(o)}\right\}[s] + \left\{\boldsymbol{\Lambda}\mathbf{f}_{(o)}\right\}[s], \mathbf{D}_{(o)}[s,s]\right); \qquad s \in \{s_1, \ldots, s_{N_{\text{miss}}}\},$$

where the subscript (o) denotes the samples values at the oth iteration. The data augmentation step is almost identical for the left-censored values, except we sample from the truncated normal density

$$\mathbf{Y}_{(o+1)}^*[t] \sim \mathrm{N}_1^{(-\infty, \tau(t)]}\left(\left\{\mathbf{X}\boldsymbol{\beta}_{(o)}\right\}[t] + \left\{\boldsymbol{\Lambda}\mathbf{f}_{(o)}\right\}[t], \mathbf{D}_{(o)}[t,t]\right); \qquad t \in \{t_1, \ldots, t_{N_{\text{cens}}}\},$$

with limits $(-\infty, \tau(t)]$, where $\tau(t)$ provides the MDL of the observation indexed by t.

When all the required entries in $\mathbf{Y}_{(o+1)}^*$ have been sampled, updating the model parameters for the $(o+1)$th iteration then proceeds after treating this vector of responses as the full dataset $\mathbf{Y} \leftarrow \mathbf{Y}^*$ in the algorithm described above. Further information on general theoretical and practical aspects of data augmentation techniques can be found in Gelman et al. (2013). Fridley and Dixon (2007) also provide simulations assessing its performance in a spatial setting, finding it to be far preferable than heavily biased 'quick-fix' options as noted in Section 1.1 (such as replacing all censored entries by half their MDL).

4.3 | Prediction

Prediction of the response at a previously unmeasured location $\hat{\mathbf{s}} \notin \mathcal{S}$ is a common goal for such analyses. In our case, this is a two-step process, performed post-fit, which involves first generating a sample for $\mathbf{f}^{(j)}(\hat{\mathbf{s}})$ at a given depth j, followed by generating the desired $\mathbf{Y}^{(j)}(\hat{\mathbf{s}})$.

Let $\mathcal{W}(\hat{\mathbf{s}}) = \{\hat{k}_1^\star, \ldots, \hat{k}_\omega^\star\} \subset \mathcal{K}^\star$ represent the indexes of the ω nearest site-depth neighbours of location $\hat{\mathbf{s}}$ from the original set of site-depths as identified by Equation (13), that is, $\|\mathbf{s}_{h(\hat{k}_1^\star)} - \hat{\mathbf{s}}\| \leq \ldots \leq \|\mathbf{s}_{h(\hat{k}_\omega^\star)} - \hat{\mathbf{s}}\|$. Under this prescription, the sequence of neighbours provides sub-sequences of the identical *spatial* neighbours across the m depths. For each retained posterior sample of $\mathbf{f}$, $\boldsymbol{\alpha}$, $\boldsymbol{\phi}$ and σ^2, we nominate a desired depth j and draw

$$\mathbf{f}^{(j)}(\hat{\mathbf{s}}) \sim \mathrm{N}_r(\hat{\mathbf{B}}_{\hat{\mathbf{s}}^{(j)}} \mathbf{f}^\star_{\mathcal{W}(\hat{\mathbf{s}})}, \hat{\boldsymbol{\zeta}}_{\hat{\mathbf{s}}^{(j)}}); \qquad j = 1, \ldots, m,$$

where, using the definitions in Section 3.3, $\hat{\mathbf{B}}_{\hat{\mathbf{s}}^{(j)}} = \hat{\boldsymbol{\Sigma}}^{(j)}_{\boldsymbol{\alpha},\boldsymbol{\phi}}(\hat{\mathbf{s}}) \boldsymbol{\Sigma}^{-1}_{\boldsymbol{\alpha},\boldsymbol{\phi},\sigma^2}[R(\mathcal{W}(\hat{\mathbf{s}})), R(\mathcal{W}(\hat{\mathbf{s}}))]$ and $\hat{\boldsymbol{\zeta}}_{\hat{\mathbf{s}}^{(j)}} = \alpha_j^2 \mathbf{I}_r + \mathbf{S}^{(j)} - \hat{\mathbf{B}}_{\hat{\mathbf{s}}^{(j)}} \left\{ \hat{\boldsymbol{\Sigma}}^{(j)}_{\boldsymbol{\alpha},\boldsymbol{\phi}}(\hat{\mathbf{s}}) \right\}^\top$, with

$$\hat{\boldsymbol{\Sigma}}^{(j)}_{\boldsymbol{\alpha},\boldsymbol{\phi}}(\hat{\mathbf{s}}) = \alpha_j \left[\alpha_{g(\hat{k}_1^\star)} C^{(1)} \left(\hat{\mathbf{s}}, \mathbf{s}_{h(\hat{k}_1^\star)}; \boldsymbol{\phi} \right), \ldots, \alpha_{g(\hat{k}_\omega^\star)} C^{(1)} \left(\hat{\mathbf{s}}, \mathbf{s}_{h(\hat{k}_\omega^\star)}; \boldsymbol{\phi} \right) \right]$$

providing the $r \times \omega r$ matrix of the cross-covariances for site $\hat{\mathbf{s}}$, at depth j, with each of the site-depths in $\mathcal{W}(\hat{\mathbf{s}})$. Note that $\alpha_1 = 1$ and $\mathbf{S}^{(1)} = \mathbf{O}_r$ are constant. Using the corresponding samples of $\boldsymbol{\beta}$ and δ^2, a realisation $\mathbf{Y}^{(j)}(\hat{\mathbf{s}})$ may then be readily generated as

$$\mathbf{Y}^{(j)}(\hat{\mathbf{s}}) \sim \mathrm{N}_q(\mathbf{X}^{(j)}(\hat{\mathbf{s}}) \boldsymbol{\beta}^{(j)} + \boldsymbol{\Lambda}^{(j)} \mathbf{f}^{(j)}(\hat{\mathbf{s}}), \mathbf{D}^{(j)}), \tag{17}$$

which requires knowledge of the values of the covariates at $\hat{\mathbf{s}}$. Summary statistics may subsequently be swept out of the collection of generated predictions produced by Equation (17) across all retained posterior samples of the model parameters. The mean and standard deviation serve as suitable expected-value point predictions and prediction standard errors respectively.

4.4 | Number of factors, number of neighbours

The age-old issue of choosing the number of factors, r, remains crucial in our spatial factor framework. Technically, model selection scores such as the deviance information criterion (DIC) and the widely applicable information criterion (WAIC) might be used, but we have found these do not tend to perform well for the sole purpose of finding r. Goodness of fit naturally improves as r is increased, and this often overwhelms data-driven estimates of model complexity, leading to steady declines in the scores even as r grows beyond computationally feasible values. Alternatively, sophisticated dynamic factor selection might be considered as part of the model definition. Ren and Banerjee (2013) proposed such a method, but the numeric stability thereof can be fickle in practice, with no guarantees the dynamic selection will settle on a given r. There is also a substantial additional computational cost to the fitting algorithm for such steps, further burdening an already expensive exercise.

A more heuristic approach to choosing r is therefore favoured for the current application. We opt to inspect the root mean squared predictive error (RMSPE) for an increasing sequence of factors, obtained using pilot runs of the fitting algorithm. Based on prevailing knowledge of the geochemical variation, which suggests the number of factors ought to be modest relative to q, 'candidate' models based on $r = 2, \ldots, 15$ factors are considered. The scree plot in Figure 6 suggests 15 factors ought to be more than capable of explaining the bulk of the residual variation in the data. Of the $N - N_{\text{miss}} - N_{\text{cens}}$ observations of the raw response data that are neither missing nor censored, we randomly choose 1000 more (indexed, say, by the set $\mathcal{U}^\star$) and omit them. Pilot fits

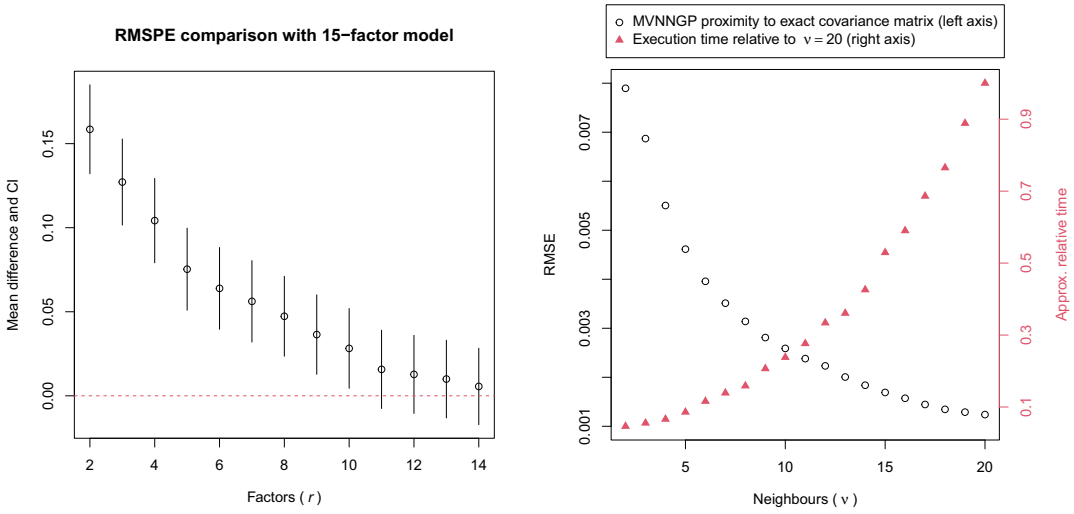

FIGURE 7 Left: Mean difference and associated 95% confidence intervals in RMSPE estimates relative to the $r = 15$ model based on pilot runs with 1000 artificial missing values. Right: Root mean square error between exact covariance matrix and MVNNGP approximation thereto for a varying number of neighbours ν, along with approximate computation time relative to $\nu = 20$ [Colour figure can be viewed at wileyonlinelibrary.com]

of the candidate models are obtained using the fitting algorithm, and the 1000 artificially missing values are imputed as described in Section 4.2. For $u \in \mathcal{U}$ we find $\widehat{\text{RMSPE}}_u = \{H^{-1}\sum_{o=1}^{H}(\mathbf{Y}^*_{(o)}[u] - \mathbf{Y}[u])^2\}^{0.5}$, where the sum is evaluated over the H retained posterior samples of the model fit.

The left panel of Figure 7 summarises the results presented as the difference in the mean RMSPE across all $u \in \mathcal{U}$, and associated 95% confidence interval, comparing all candidate models with $2 \leq r < 15$ with that of the most complex model at $r = 15$. The first interval to overlap zero is that corresponding to the 11-factor model. Thus, we proceed with $r = 11$ for the final model fit presented in the following section.

The other choice to be made concerns the MVNNGP approximation itself. Recall from Section 3.4 that we must set the maximum number of neighbours ν, which governs the extent of the sparse approximation to the 'exact' dependence structure of the full MVGP. A greater ν leads to improved approximation accuracy but with an increased computational cost. Datta et al. (2016a, 2016b) and subsequent works (e.g. Banerjee, 2017) have noted the approximation (15) to be excellent in many different cases with surprisingly few neighbours, such as 5, 10 or 20. It is instructive to investigate this choice for the current application, which we do by way of comparing the overall proximity of the exact covariance matrix for the full MVGP, $\boldsymbol{\Sigma}_{\boldsymbol{\alpha},\boldsymbol{\phi},\sigma^2}$ as in Equation (11), to that recoverable from the MVNNGP approximation in Equation (15).

To do this, we first extract sensible parameter values for all of $\boldsymbol{\sigma}^2$, $\boldsymbol{\phi}$ and $\boldsymbol{\alpha}$ from the $r = 11$ pilot run used above. We then construct the full $mnr \times mnr$ matrix $\boldsymbol{\Sigma}_{\boldsymbol{\alpha},\boldsymbol{\phi},\sigma^2}$ using (12), and construct the MVNNGP approximation thereto, $\tilde{\boldsymbol{\Sigma}}^{(\nu)}_{\boldsymbol{\alpha},\boldsymbol{\phi},\sigma^2}$, by inverting $\tilde{\boldsymbol{\Sigma}}^{(\nu)-1}_{\boldsymbol{\alpha},\boldsymbol{\phi},\sigma^2} = (\mathbf{I}_{mnr} - \mathbf{B}^\top)\boldsymbol{\zeta}^{-1}(\mathbf{I}_{mnr} - \mathbf{B})$, where $\mathbf{B}[R(k^\star), R(\mathcal{N}_\nu(k^\star))] = \mathbf{B}_{k^\star}$ and $\boldsymbol{\zeta} = \bigoplus_{k^\star} \boldsymbol{\zeta}_{k^\star}$ for all $k^\star \in \mathcal{K}^\star$, as per the notation in Section 3.4. We repeat this for $\nu = 2, \ldots, 20$, each time finding the element-wise root mean square error (RMSE) between the two matrices as

$$\text{RMSE}_\nu = \sqrt{(mnr)^{-2} \sum_{k_1^\star, k_2^\star \in \mathcal{K}^\star} \left(\tilde{\boldsymbol{\Sigma}}^{(\nu)}_{\boldsymbol{\alpha},\boldsymbol{\phi},\sigma^2}\left[k_1^\star, k_2^\star\right] - \boldsymbol{\Sigma}_{\boldsymbol{\alpha},\boldsymbol{\phi},\sigma^2}\left[k_1^\star, k_2^\star\right] \right)^2}.$$

Results are given in the right-hand panel of Figure 7, showing the proximity to the full matrix improving with increasing ν. Overlaid on this plot is an impression of timing, with the approximate computational cost per iteration relative to the most complex approximation at $\nu = 20$ also shown—increasing steadily with ν. In terms of predictive performance, we found little difference for different values of ν. Based on these findings and the established literature on choosing ν, we set $\nu = 10$. As can be seen in Figure 7, this appears to strike a good balance between approximation accuracy and computational requirements.

5 | RESULTS

Three parallel Markov chains with varying starting values are run. Following a sizeable burn-in period to ensure convergence, each of these proceeds for a long run of 100,000 iterations, thinning by a factor of $\frac{1}{10}$ to temper storage requirements. Results are based on the collated set of 30,000 retained iterations from these three independent runs. The R language (R Core Team, 2021) was used for implementation, with computational bottlenecks farmed out to C++ via the 'Rcpp' and 'RcppArmadillo' packages (Eddelbuettel, 2013; Eddelbuettel & Sanderson, 2014). The running time under these conditions on a local desktop machine (4.2 GHz Quad-Core Intel i7 CPU, 32Gb RAM) was approximately 150 h, although we note from inspecting the output that meaningful inference may be readily achieved with far shorter runs, that is, within 48–96 h. The dataset and code are supplied as Supplement B.

To assist exploring the vast array of results, including the raw data (with imputed values for censored and missing observations); predictive surfaces; and estimated posterior distributions for the various model parameters, we have packaged up the output so that it may be easily presented using Shiny (Chang et al., 2021). The main application is accessible at https://www.stats.otago.ac.nz/nzgeochem/ and provides drop-down menus where the user can select from the list of modelled elements (c.f. Supplement A) as well as the desired output: the raw data (optionally displaying imputed missing and censored values where applicable); pixel images of the model prediction over the study region for both depths; estimated posteriors for all components of the regression parameters $\boldsymbol{\beta}$; the residual variances $\boldsymbol{\delta}^2$; the spatial correlation scale parameters $\boldsymbol{\phi}$; the inter-depth variances $\boldsymbol{\sigma}^2$; and the inter-depth dependency parameter α; and finally, site-specific posterior means for the 11 factors $\mathbf{f}$.

To provide further inferential value we have also expanded upon the 2D predictive images available in the above app. A secondary Shiny application located at https://www.stats.otago.ac.nz/nzgeochem_3dpred/ provides mouse-rotatable 3D graphics of the predictive surfaces in various forms, facilitated by the 'rgl' package (Murdoch & Adler, 2021).

We now make use of these visual applications to explore selected results. All estimates and plots that follow may be obtained by accessing the above links.

5.1 | Fixed effects

An exhaustive geochemical examination of the full set of regression parameters will be pursued elsewhere. To provide a glimpse of how these estimates may prove contextually useful, however, we will briefly consider some estimated components of $\boldsymbol{\beta}$ for scandium (Sc), iron (Fe) and sodium (Na). Figures 8–10 provide plots of selected regression parameters for these three elements respectively. In each case, a solid vertical line marks the posterior mean and dashed lines delineate a 95%

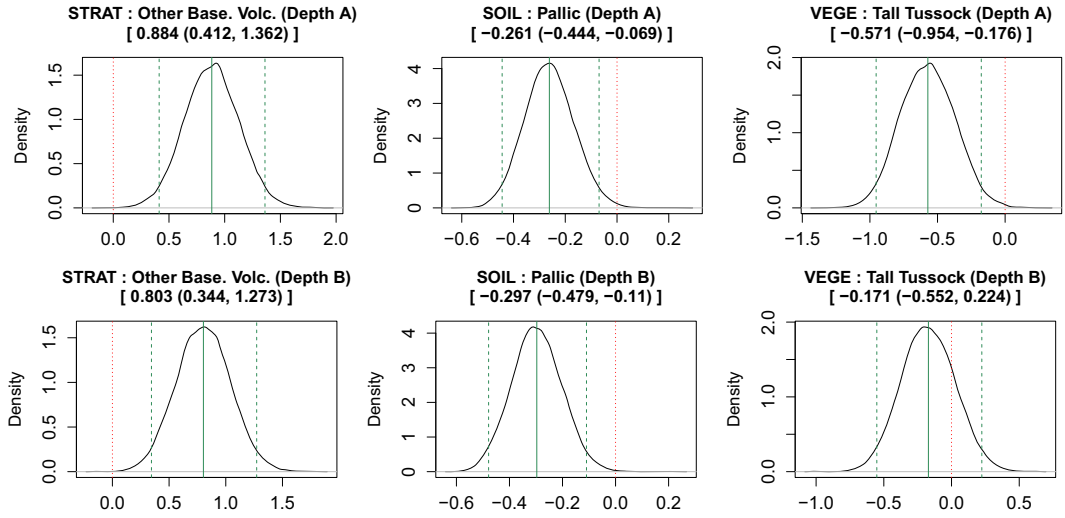

FIGURE 8 Noteworthy estimated posterior densities for regression parameters pertaining to Sc [Colour figure can be viewed at wileyonlinelibrary.com]

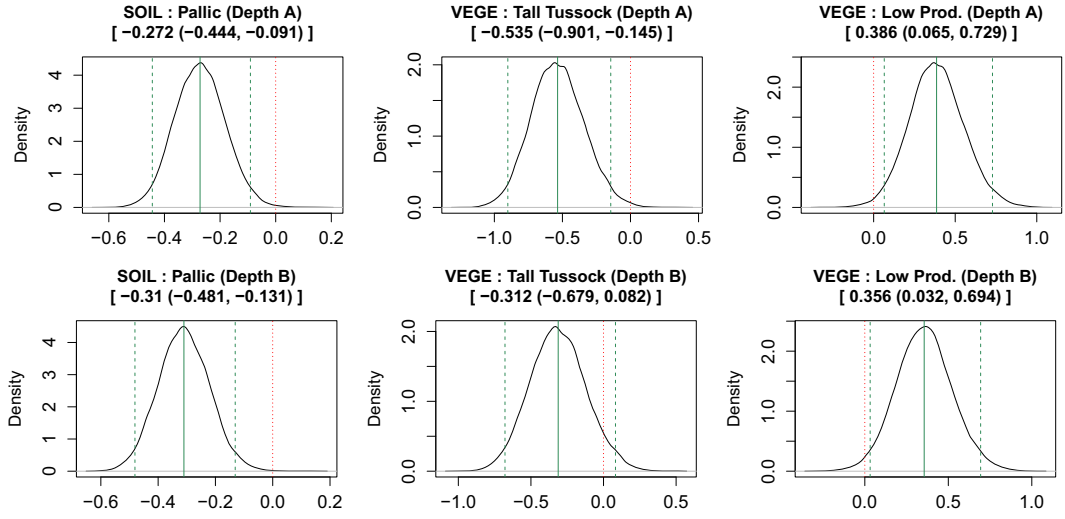

FIGURE 9 Noteworthy estimated posterior densities for regression parameters pertaining to Fe [Colour figure can be viewed at wileyonlinelibrary.com]

credible interval (with their numeric values listed in the titles of each panel). Refer to Supplement A for further explanation of these predictors and the abbreviations used.

Beginning with Sc, we note significant positive effects at both depths for STRAT : Other Base. Volc. with respect to the reference level of STRAT. This aligns with existing geological knowledge that scandium is typically more enriched in mafic igneous rocks (captured by this category) in comparison to other rock types in the survey area. A negative effect at both depths is also observed for SOIL : Pallic. Pallic soils are typically derived from schist and sandstone rock sources (Hewitt, 2010), which have relatively low scandium in comparison to igneous rock types. Similarly, note that a significant negative effect is observable at Depth A for VEGE : Tall Tussock but not at the deeper Depth B; these grasslands are most abundant in areas with underlying schist rock.

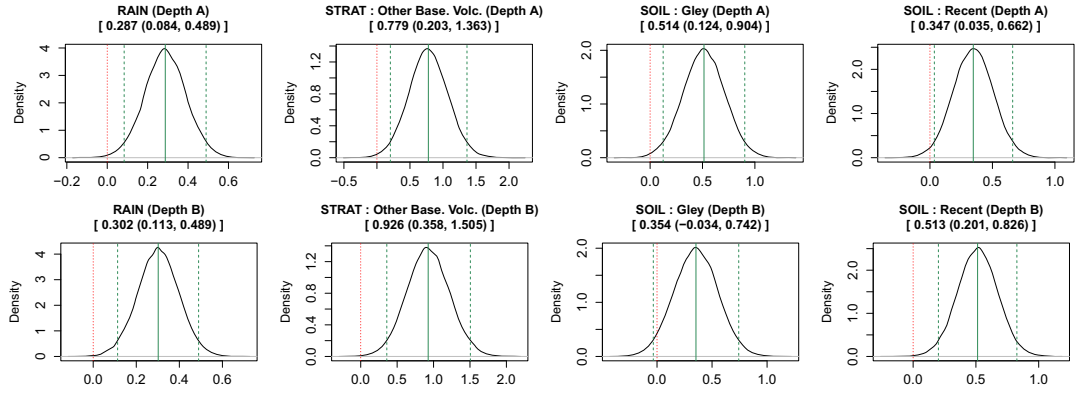

FIGURE 10 Noteworthy estimated posterior densities for regression parameters pertaining to Na [Colour figure can be viewed at wileyonlinelibrary.com]

Examining Fe reveals a similar story. Pallic soils have a low content of iron oxide minerals relative to other soil types (Hewitt, 2010). This is a result of being derived mainly from Fe-poor schist rocks in the study area and is reflected as the negative effects at both depths for SOIL : Pallic as well as the Depth A negative effect for VEGE : Tall Tussock. Interestingly, we also observe significantly positive effects for VEGE : Low. Prod. This type of vegetation traditionally has superphosphate fertiliser added to it, which is rich in Fe (Marshall & Hill, 1952), reflecting an anthropogenic effect. Native vegetation does not receive direct fertiliser application, and other types of vegetation such as high producing grassland (VEGE : High Prod.) receives a mix of Fe-bearing superphosphate and Fe-poor or Fe-free fertilisers.

Lastly, Na varies across the study. Rocks with Na-bearing minerals (e.g. Na-feldspar; pyroxene, nepheline) are most prevalent in igneous rocks found in STRAT : Murihiku and STRAT : Other Base. Volc., as can be seen by significant positive relationships at both depths, relative to other STRAT types. The only exception is Depth B STRAT : Pakahi, where a weak, positive relationship is observed. Rocks forming the Pakahi Supergroup are geologically young and often sourced from the erosion and redeposition of other rock types along flood plains. The Depth B, weak positive relationship with Na thus reflects the relationships in the source rocks from STRAT : Murihiku and STRAT : Other Base. Volc. Recent soils form on surfaces that are morphologically young relative to others around them. In the study area, this typically occurs on Pakahi Supergroup rocks. Thus, the positive relationship between Na and SOIL : Recent reflects the geological input from igneous rocks represented by STRAT : Murihiku and STRAT : Other Base. Volc. It is interesting to note that the location of STRAT : Murihiku and STRAT : Other Base. Volc. coincides with elevated rainfall and lower air temperatures, and the correlation between Na, RAIN and ATEMP is not causative. One exception is the positive relationship between Na and SOIL : Gley—gley soils tend to be water-logged and may concentrate mobile forms of Na.

5.2 | Random effects

The factors are each found to exhibit moderate-to-strong spatial dependency through the components of $\boldsymbol{\phi}$. The minimum lower and maximum upper values across all eleven 95% credible intervals corresponding to each component of $\boldsymbol{\phi}$ are 0.086 and 0.135 respectively. Recall from Section 3.5 that the uniform prior limits imposed on the correlation scale parameters are (0.0114, 0.3480), strong to weak. The results therefore suggest there remains detectable

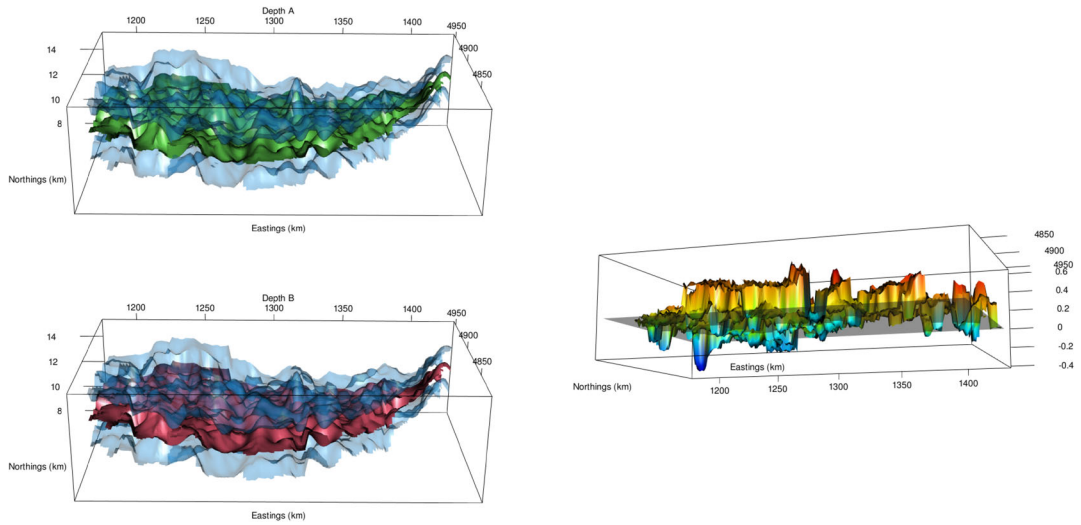

FIGURE 11 Screenshots of interactive 3D graphics for predictive surfaces of Na. Left: model predictions (top: Depth A/green; bottom: Depth B/red) with ±2 prediction errors (blue surfaces). Right: Difference between the Depth A and Depth B predictive surfaces; a grey plane marks zero [Colour figure can be viewed at wileyonlinelibrary.com]

residual spatial variation in the data. We note a posterior expected value of 0.895 along with a narrow credible interval of (0.874, 0.916) for α. As per Equation (8), this is indicative of notable between-depth association among the residual spatial factors across all elements. This is useful from a geochemistry perspective—a similarity of residual effects implies any between-depth distinctions in the responses have been adequately described by the depth-specific predictor variables.

5.3 | Predictions

The predictive surfaces, calculated as per Section 4.3 on a fine grid of spatial coordinates laid over the study region, can also offer valuable insight. These are best viewed in the secondary (3D prediction) Shiny application where the user can interact with the graphics for differing perspective. For illustration we show screenshots for sodium (Na) in Figure 11.

The predictions at both depths follow similar overall features as visible on the left of Figure 11. There is a noticeable upturned peak in sodium readings predicted around the populated centre of Dunedin on the eastern border; a partial reflection of the positive effect of geology seen in STRAT : Oth. Bas. Vol. Looking to the difference surface, we observe that concentrations of Na are generally predicted to be higher in Depth A soils when compared to Depth B soils, although there are clearly some sub-regions where the reverse is true.

6 | CONCLUSION AND FUTURE RESEARCH

Geochemical survey data form the basis of many and varied research questions, often revolving around an understanding of the spatial variation in element concentrations. Among many inherent complexities are a high-dimensional outcome variable, multi-depth observations that

reflect different environmental and anthropogenic imperatives, and missing or censored records. Standard tools of multivariate statistics and geostatistics are usefully deployed in simplified scenarios, such as in analysing a subset of the survey data, although this naturally limits inferential and predictive ability.

In this work we have demonstrated how one might construct a model to cope with these complexities. We considered a spatial factor model to handle the multivariate response, whereby predictors are permitted to have different effects on the observations of a given element for each of the two depths, while simultaneously imposing a direct between-depth relationship on the individual factors. The fitting algorithm imputes censored and missing values in a data augmentation step, and we leverage methodological advancements in the approximation of Gaussian spatial processes to further increase computational accessibility.

The model output is comprehensive, permitting inference and prediction at an individual-depth/element/location level. As illustrated in the previous section, our results both align with existing knowledge and lend support to other conjectures on the nature of the geochemical composition of the region. For ease of interpretation we chose to present the results in a mainly graphical way using a pair of web-based applications. These plots and predictions offer chemists and policy makers a uniquely useful view of the chemical landscape. Using them, researchers can objectively quantify which factors influence soil chemistry. Furthermore, they facilitate the formation of new ideas regarding probable soil chemistry, and the factors influencing that soil chemistry, in data-poor areas of the survey region. For policy makers, the interactive plots are a visual aid to understanding the variation in soil chemistry and factors influencing it that are more effective than static plots alone.

Our design represents merely one possible way in which we might seek to model such data, and we shall pursue alternatives in future work. In situations where the emphasis lies firmly on prediction, for example, there is value in exploring a design similar to that of Groth et al. (2018), in which each component of the multivariate response is regressed on the others. This in turn opens up new options for capturing and describing between-depth relationships. Another option would be to consider a model in terms of explicit 3D space (some forays into using 3D kriging and interpolation for e.g. mineral deposits exist; see Wang & Huang, 2012), which in theory could permit prediction in a continuous fashion below the soil surface. Such an approach would at least require the exact depths of the individual soil samples be known (as opposed to being assigned to categories); the curse of dimensionality would likely dictate a requirement of sizeable datasets for meaningful inference; and careful consideration of the form of the directional dependence functions would be needed. A third avenue of pursuit could constitute leveraging recent advancements, namely *stitching*, in specifying richer forms of spatial cross-covariance structures to aid the subsequent parameter-heavy consequences of highly multivariate responses (Dey et al., 2021).

As it stands, our 'layered' approach to the issue of depth retains the relative simplicity of 2D dependence structures and aligns with the common focus of 'inter-depth' behaviours in geochemical research studies; see for example, Lawrence et al. (2015) and Turnbull et al. (2019). As is also particularly relevant to the current analysis, the model presented here could be readily extended to incorporate samples from different surveys across the country, in which a data-driven levelling of the various samples would make an appealing feature.

ACKNOWLEDGEMENTS

Tilman M. Davies was supported in part by the Royal Society of New Zealand, Marsden Fund grants 15-UOO-092 and 19-UOO-191. Sudipto Banerjee was supported in part by federal grants

NSF/DMS 1916349 and 2113778; and NIH/NIEHS 1R01ES027027. GNS Science contribution funded by the Ministry of Business Innovation and Employment core funding. The authors thank Greg Trounson (Department of Mathematics & Statistics, University of Otago) for assistance in deploying and debugging the Shiny applications. The anonymous referees and an associate editor are thanked for their constructive critiques which led to notable improvements to the presentation.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available in the supplementary material of this article.

ORCID

Tilman M. Davies https://orcid.org/0000-0003-0565-1825
Sudipto Banerjee https://orcid.org/0000-0002-2239-208X

REFERENCES

Aitchison, J., Barceló-Vidal, C., Martín-Fernández, J. & Pawlowsky-Glahn, V. (2000) Logratio analysis and compositional distance. *Mathematical Geosciences*, 32, 271–275.

Anderson, T. (2003) *An introduction to multivariate statistical analysis*, 3rd edition, Hoboken, NJ: Wiley.

Banerjee, S. (2017) High-dimensional Bayesian geostatistics. *Bayesian Analysis*, 12(2), 583–614.

Banerjee, S., Carlin, B.P. & Gelfand, A.E. (2014) *Hierarchical modeling and analysis for spatial data*, 2nd edition, New York: Chapman & Hall/CRC.

Bartier, P.M. & Keller, C.P. (1996) Multivariate interpolation to incorporate thematic surface data using inverse distance weighting (IDW). *Computers & Geosciences*, 22(7), 795–799.

van den Berg, R.A., Hoefsloot, H.C., Westerhuis, J.A., Smilde, A.K. & van der Werf, M.J. (2006) Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics*, 7(1), 142.

Bishop, C. (2006) *Pattern recognition and machine learning*. New York: Springer-Verlag.

Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y. et al. (2021) `shiny`: Web application framework for R. R package version 1.6.0.

Chilés, J. & Delfiner, P. (2012) *Geostatistics: modeling spatial uncertainty*, 2nd edition, Hoboken, NJ: Wiley.

Clare, N.T. (1981) Chemistry in animal disease and production. In: Williams, P.P. (ed), *Chemistry in a young country*, Christchurch, New Zealand: New Zealand Institute of Chemistry Inc., pp. 65–77.

Clarke, J.U. (1998) Evaluation of censored data methods to allow statistical comparisons among very small samples with below detection limit observations. *Environmental Science & Technology*, 32(1), 177–183.

Darnley, A.G., Björklund, A., Bølviken, B., Gustavsson, N., Koval, P.V., Plant, J.A. et al. (1995) *A global geochemical database for environmental and resource management*. Paris, France: UNESCO.

Datta, A., Banerjee, S., Finley, A.O. & Gelfand, A.E. (2016a) Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association*, 111, 800–812.

Datta, A., Banerjee, S., Finley, A.O. & Gelfand, A.E. (2016b) On nearest-neighbor Gaussian process models for massive spatial data. *Wiley Interdisciplinary Reviews: Computational Statistics*, 8, 162–171.

Deely, J.M., Tunicliff, J.C., Orange, C.J. & Edgerley, W.H.L. (1992) Heavy metals in surface sediments of Waiwhetu Stream, Lower Hutt, New Zealand. *New Zealand Journal of Marine and Freshwater Research*, 26, 417–427.

Dey, D., Datta, A. & Banerjee, S. (2021) Graphical Gaussian process models for highly multivariate spatial data. *Biometrika*, asab061. Available from: https://doi.org/10.1093/biomet/asab061

Eddelbuettel, D. (2013) *Seamless R and C++ integration with Rcpp*. New York: Springer.

Eddelbuettel, D. & Sanderson, C. (2014) `RcppArmadillo`: accelerating R with high-performance C++ linear algebra. *Computational Statistics and Data Analysis*, 71, 1054–1063.

Farnham, I.M., Singh, A.K., Stetzenbach, K.J. & Johannesson, K.H. (2002) Treatment of nondetects in multivariate analysis of groundwater geochemistry data. *Chemometrics and Intelligent Laboratory Systems*, 60(1), 265–281.

Fergusson, J.E., Hayes, R.W., Tan, S.Y. & Sim, H.T. (1980) Heavy metal pollution by traffic in Christchurch, New Zealand: lead and cadmium content of dust, soil and plant samples. *New Zealand Journal of Science*, 23, 293–310.

Finley, A.O., Banerjee, S. & Carlin, B.P. (2007) `spBayes`: an R package for univariate and multivariate hierarchical point-referenced spatial models. *Journal of Statistical Software*, 19, 1–24.

Fridley, B.L. & Dixon, P. (2007) Data augmentation for a Bayesian spatial model involving censored observations. *Environmetrics*, 18, 107–123.

Gazley, M.F., Martin, A.P., Turnbull, R.E., Frontin-Rollet, G. & Strong, D.T. (2020) Regional patterns in standardised and transformed pathfinder elements in soil related to orogenic-style mineralisation in southern New Zealand. *Journal of Geochemical Exploration*, 217, 106593.

Gelfand, A.E., Schmidt, A.M., Banerjee, S. & Sirmans, C.F. (2004) Nonstationary multivariate process modeling through spatially varying coregionalization. *Test*, 13, 263–312.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A. & Rubin, D.B. (2013) *Bayesian data analysis*, 3rd edition, Boca Raton, FL: CRC Press.

Goldschmidt, V.M. (1954) *Geochemistry*. Oxford, UK: Clarendon Press.

Graham, J.W. (2012) *Missing data: analysis and design*. New York: Springer.

Groth, C.P., Banerjee, S., Ramachandran, G., Stenzel, M.R. & Stewart, P.A. (2018) Multivariate left-censored Bayesian modeling for predicting exposure using multiple chemical predictors. *Environmetrics*, 29, e2505.

Grunsky, E. & de Caritat, P. (2019) State-of-the-art analysis of geochemical data for mineral exploration. *Geochemistry: Exploration, Environment, Analysis*, 20(2), 217–232.

Herselman, J., Steyn, C. & Fey, M. (2005) Baseline concentration of Cd, Co, Cr, Cu, Pb, Ni and Zn in surface soils of South Africa: research in action. *South African Journal of Science*, 101(11), 509–512.

Hewitt, A.E. (2010) *New Zealand soil classification*, 3rd edition, Canterbury, New Zealand: Manaaki Whenua Press.

Johnson, C.C., Breward, N., Ander, E.L. & Ault, L. (2005) G-BASE: Baseline geochemical mapping of Great Britain and Northern Ireland. *Geochemistry: Exploration, Environment, Analysis*, 5(4), 347–357.

Kapička, A., Petrovský, E., Jordanova, N. & Podrázský, V. (2001) Magnetic parameters of forest top soils in Krkonoše mountains, Czech Republic. *Physics and Chemistry of the Earth, Part A (Solid Earth and Geodesy)*, 26(11), 917–922.

Lado, L.R., Hengl, T. & Reuter, H.I. (2008) Heavy metals in European soils: a geostatistical analysis of the FOREGS Geochemical database. *Geoderma*, 148(2), 189–199.

Lawrence, C.R., Harden, J.W., Xu, X., Schulz, M.S. & Trumbore, S.E. (2015) Long-term controls on soil organic carbon with depth and time: a case study from the Cowlitz River Chronosequence, WA USA. *Geoderma*, 247–248, 73–87.

Lopes, H.F. & West, M. (1999) Model uncertainty in factor analysis, Technical report, Institute of Statistics and Decision Sciences, North Carolina, USA: Duke University.

Lopes, H.F., Salazar, E. & Gamerman, D. (2008) Spatial dynamic factor analysis. *Bayesian Analysis*, 3(4), 759–792.

Marshall, H. & Hill, W. (1952) Composition and properties of superphosphate. Effect of aluminum and iron content on curing behavior. *Industrial & Engineering Chemistry*, 44(7), 1537–1540.

Martin, A.P., Turnbull, R.E., Rattenbury, M.S., Cohen, D.R., Hoogewerff, J., Rogers, K.M. et al. (2016) The regional geochemical baseline soil survey of southern New Zealand: design and initial interpretation. *Journal of Geochemical Exploration*, 167, 70–82.

Martin, A.P., Turnbull, R.E., Rissman, C.W.F. & Rieger, P. (2017) Heavy metal and metalloid concentrations in soils under pasture of southern New Zealand. *Geoderma Regional*, 11(Supplement C), 18–27.

Martin, A.P., Ohneiser, C., Turnbull, R.E., Strong, D.T. & Demler, S. (2018) Soil magnetic susceptibility mapping as a pollution and provenance tool: an example from southern New Zealand. *Geophysical Journal International*, 212(2), 1225–1236.

Matschullat, J., Höfle, S., da Silva, J., Mello, J., Melo Jr., G., Pleßow, A. et al. (2012) A soil geochemical background for northeastern Brazil. *Geochemistry: Exploration, Environment, Analysis*, 12, 197–209.

McKinley, J.M., Hron, K., Grunsky, E.C., Reimann, C., de Caritat, P., Filzmoser, P. et al. (2016) The single component geochemical map: fact or fiction? *Journal of Geochemical Exploration*, 162, 16–28.

Müller, G. (1979) Schwermetalle in den Sedimenten des Rheins: Veränderungen seit 1971. *Umschau in Wissenschaft und Technik*, 79, 778–783.

Murdoch, D. & Adler, D. (2021) rgl: 3D Visualization Using OpenGL. R package version 0.108.3. Available from: https://CRAN.R-project.org/package=rgl [Accessed 3rd March 2022].

National Soil Survey Office. (1998) *Soils of China*. Beijing, People's Republic of China: China Agricultural Press.

Plant, J., Smith, D., Smith, B. & Williams, L. (2001) Environmental geochemistry at the global scale. *Applied Geochemistry*, 16, 1291–1308.

Purchase, N.G. & Fergusson, J.E. (1986) The distribution and geochemistry of lead in river sediments, Christchurch, New Zealand. *Environmental Pollution Series B (Chemical and Physical)*, 12(3), 203–216.

R Core Team. (2021) *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rattenbury, M.S., Martin, A.P., Turnbull, R.E. & Christie, A.B. (2014) *Sampling methodology for a regional multi-element geochemical baseline survey*. Lower Hutt, New Zealand: GNS Science. GNS Science Report Number: 2014/62.

Rawlins, B.G., McGrath, S.P., Scheib, A.J., Breward, N., Cave, M., Lister, T.R. et al. (2012) *The advanced soil geochemical atlas of England and Wales*. Keyworth, Nottingham, UK: British Geological Survey.

Reimann, C. (2005a) Geochemical mapping: technique or art? *Geochemistry: Exploration, Environment, Analysis*, 5(4), 359–370.

Reimann, C. (2005b) Sub-continental-scale geochemical mapping: sampling, quality control and data analysis issues. *Geochemistry: Exploration, Environment, Analysis*, 5(4), 311–323.

Reimann, C. & de Caritat, P. (2012) New soil composition data for Europe and Australia: demonstrating comparability, identifying continental-scale processes and learning lessons for global geochemical mapping. *Science of the Total Environment*, 416, 239–252.

Ren, Q. & Banerjee, S. (2013) Hierarchical factor models for large spatially misaligned data: a low-rank predictive process approach. *Biometrics*, 69, 19–30.

Rissmann, C.W.F., Pearson, L.K., Beyer, M., Couldrey, M.A., Lindsay, J.L., Martin, A.P. et al. (2019) A hydrochemically guided landscape classification system for modelling spatial variation in multiple water quality indices: Process-attribute mapping. *Science of the Total Environment*, 672, 815–833.

Rogers, K.M., Turnbull, R.E., Martin, A.P., Baisden, W.T. & Rattenbury, M.S. (2017) Stable isotopes reveal human influences on southern New Zealand soils. *Applied Geochemistry*, 82, 15–24.

Sanford, R.F., Pierson, C.T. & Crovelli, R.A. (1993) An objective replacement method for censored geochemical data. *Mathematical Geology*, 25, 59–80.

Singer, D.A. & Kouda, R. (2001) Some simple guides to finding useful information in exploration geochemical data. *Natural Resources Research*, 10(2), 137–147.

Smith, D.B., Cannon, W.F., Woodruff, L.G., Solano, F., Kilburn, J.E. & Fey, D.L. (2013) *Geochemical and mineralogical data for soils of the conterminous United States*. United States Geological Survey (USGS), Data Series 801. Available from: https://pubs.usgs.gov/ds/801/ [Accessed 16th June 2021].

Sparks, D.L. (2003) *Environmental soil chemistry*, 2nd edition, San Diego, CA: Academic Press.

Stein, M.L. (1999) *Interpolation of spatial data: some theory for kriging*. New York: Springer-Verlag.

Stroud, J.R., Müller, P. & Sansó, B. (2001) Dynamic models for spatiotemporal data. *Journal of the Royal Statistical Society, Series B*, 63, 673–689.

Taylor-Rodriguez, D., Finley, A.O., Datta, A., Babcock, C., Andersen, H., Cook, B.D. et al. (2019) Spatial factor models for high-dimensional and large spatial data: an application in forest variable mapping. *Statistica Sinica*, 29(3), 1155–1180.

Tipping, M.E. & Bishop, C.M. (1999) Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3), 611–622.

Turnbull, R.E., Rogers, K., Martin, A.P., Rattenbury, M.S. & Morgan, R. (2019) Human impacts recorded in chemical and isotopic fingerprints of soils from Dunedin City, New Zealand. *Science of the Total Environment*, 673, 455–469.

Wackernagel, H. (2003) *Multivariate geostatistics: an introduction with applications*. Berlin, Germany: Springer-Verlag.
Wang, G. & Huang, L. (2012) 3D geological modeling for mineral resource assessment of the Tongshan Cu deposit, Heilongjiang Province, China. *Geoscience Frontiers*, 3(4), 483–491.
Wang, F. & Wall, M.M. (2003) Generalized common spatial factor model. *Biostatistics*, 4, 569–582.
Webber, J. (1981) Trace metals in agriculture. In Lepp, N.W. (Ed.) *Effect of heavy metal pollution on plants*. Pollution Monitoring Series, Netherlands: Springer, pp. 159–184.
Zhang, L. & Banerjee, S. (2021) Spatial factor modeling: a Bayesian matrix-normal approach for misaligned data. *Biometrics*. Available from: https://doi.org/10.1111/biom.13452

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Davies, T.M., Banerjee, S., Martin, A.P. & Turnbull, R.E. (2022) A nearest-neighbour Gaussian process spatial factor model for censored, multi-depth geochemical data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 71(4), 1014–1043. Available from: https://doi.org/10.1111/rssc.12565