Received: 21 September 2023 | Revised: 27 May 2024 | Accepted: 10 June 2024

DOI: 10.1002/env.2872

RESEARCH ARTICLE

WILEY

Marginal inference for hierarchical generalized linear mixed models with patterned covariance matrices using the Laplace approximation

Jay M. Ver Hoef[1] | **Eryn Blagg**[2] | **Michael Dumelle**[3] | **Philip M. Dixon**[2] | **Dale L. Zimmerman**[4] | **Paul B. Conn**[1]

[1]Marine Mammal Laboratory, NOAA-NMFS Alaska Fisheries Science Center, Seattle, Washington, USA

[2]Department of Statistics, Iowa State University, Ames, Iowa, USA

[3]United States Environmental Protection Agency, Corvallis, Oregon, USA

[4]Department of Statistics and Actuarial Science, The University of Iowa, Iowa City, Iowa, USA

Correspondence
Jay M. Ver Hoef, Marine Mammal Laboratory, NOAA-NMFS Alaska Fisheries Science Center, 7600 Sand Point Way NE, Seattle, WA 98115, USA.
Email: jay.verhoef@noaa.gov

Abstract

We develop hierarchical models and methods in a fully parametric approach to generalized linear mixed models for any patterned covariance matrix. The Laplace approximation is used to marginally estimate covariance parameters by integrating over all fixed and latent random effects. The Laplace approximation relies on Newton–Raphson updates, which also leads to predictions for the latent random effects. We develop methodology for complete marginal inference, from estimating covariance parameters and fixed effects to making predictions for unobserved data. The marginal likelihood is developed for six distributions that are often used for binary, count, and positive continuous data, and our framework is easily extended to other distributions. We compare our methods to fully Bayesian methods, automatic differentiation, and integrated nested Laplace approximations (INLA) for bias, mean-squared (prediction) error, and interval coverage, and all methods yield very similar results. However, our methods are much faster than Bayesian methods, and more general than INLA. Examples with binary and proportional data, count data, and positive-continuous data are used to illustrate all six distributions with a variety of patterned covariance structures that include spatial models (both geostatistical and areal models), time series models, and mixtures with typical random intercepts based on grouping.

KEYWORDS

generalized linear mixed models, Laplace approximation, marginalization, patterned covariance matrix, spatial statistics, time series

1 | INTRODUCTION

The classical linear model relies on a normal distribution that has continuous support on the real line, but many data are binary, counts, or positive continuous. Such data can be transformed to stabilize variances and create empirical distributions that are "near normal," allowing the use of classical linear models (e.g., Snedecor & Cochran, 1980, p. 288). For

example, a square root transformation can be used for count data. However, Nelder and Wedderburn (1972) introduced the generalized linear model (GLM, McCullagh & Nelder, 1989) as a natural extension to linear models, such as the Poisson distribution for counts, the Bernoulli distribution for binary data, etc., which have become very popular and generally preferred to data transformations (e.g., Warton & Hui, 2011). GLMs can be extended by introducing latent random effects via a linear mixed model to create a class of generalized linear mixed models (GLMMs, Breslow & Clayton, 1993). These latent random effects are usually assumed to be independent and identically distributed normal variables (Zeger & Karim, 1991), however it is also possible for the latent random effects to be temporally autocorrelated (e.g., Stiratelli et al., 1984; Zeger et al., 1988), spatially autocorrelated (e.g, Clayton & Kaldor, 1987; Diggle et al., 1998; Gotway & Stroup, 1997), or both (Cressie & Wikle, 2011, p. 380). A unifying framework for this literature is available through a hierarchical generalized linear mixed model (HGLMM, Lee & Nelder, 1996).

1.1 | Hierarchical generalized linear mixed models

Most GLMs are motivated by the exponential family of distributions (e.g., Fisher, 1934; Lehmann & Casella, 2006). However, most common software packages use iteratively-reweighted least-squares (Wedderburn, 1974) to fit GLMs through their first two moments (mean and variance functions). This procedure allows a flexible class of models that can be fit though a single inferential procedure (quasi-likelihood), which is desirable because some GLMs like the quasi-Poisson have no true likelihood.

By contrast, we will take a fully parametric approach to create covariance dependence for GLMMs through a hierarchical construction. We will use the notation $[\mathbf{y}|\boldsymbol{\mu}]$ to denote any joint probability density/mass function of the vector of random variables $\mathbf{y}$ conditional on a vector of parameters, or other fixed variables, $\boldsymbol{\mu}$. In some cases, we may want to model multiple vectors of responses (as with time series applications) or to condition on more than one set of parameters, in which case an expression might look more like $[\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_k|\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_\ell, \boldsymbol{\phi}]$. As a simple example, consider a joint probability mass function that consists of the product of independent negative binomial distributions, which can be parameterized with a mean vector and a common extra parameter that allows for overdispersion. We could write this as $[\mathbf{y}|\boldsymbol{\mu}, \phi]$, where the $\boldsymbol{\mu}$ represents the mean vector and ϕ is the overdispersion parameter.

For the hierarchical construction of the generalized linear mixed models that we consider in this article, we allow the mean to vary by other random variables $\mathbf{w}$, and we condition on these random variables, $[\mathbf{y}|\mathbf{g}^{-1}(\mathbf{w}), \boldsymbol{\phi}]$, through the multivariate mean function $\mathrm{E}(\mathbf{y}) = \mathbf{g}^{-1}(\mathbf{w})$. For the Poisson distribution, each element of $\mathbf{g}(\cdot)$ is often the log function, and in general $\mathbf{g}(\cdot)$ is called the (multivariate) link function (McCullagh & Nelder, 1989). Link functions are monotonic so that each element of $\mathbf{g}^{-1}(\cdot)$ is one-to-one with each element of $\mathbf{g}(\cdot)$. Recall that the mean of a Poisson distribution must be positive, and if $\mathbf{g}(\cdot)$ has each element as the log function, then $\mathbf{g}^{-1}(\cdot)$ has each element as the exponential function so $\boldsymbol{\mu} = \mathbf{g}^{-1}(\mathbf{w})$ always has positive elements, which allows each element of $\mathbf{w}$ to be unconstrained.

We will only consider models where $\mathbf{w}$ is $n \times 1$ and has a multivariate normal distribution that is constructed through the linear mixed model,

$$\mathbf{w} = \mathbf{X}\boldsymbol{\beta} + \sum_{k=1}^{q} \mathbf{Z}_k \mathbf{r}_k + \boldsymbol{\epsilon}, \tag{1}$$

where $\mathbf{X}$ is an $n \times p$ full rank design matrix of explanatory variables, $\boldsymbol{\beta}$ is a $p \times 1$ parameter vector of fixed effects, $\mathbf{Z}_k$ is a design matrix for the kth random effect vector $\mathbf{r}_k$, and $\boldsymbol{\epsilon}$ represents independent and identically-distributed Gaussian error. We assume that $\mathrm{E}(\mathbf{r}_k) = \mathbf{0}$ for all k, $\mathrm{E}(\boldsymbol{\epsilon}) = \mathbf{0}$, $\mathrm{var}(\mathbf{r}_k) = \mathbf{V}_k$, $\mathrm{cov}(\mathbf{r}_j, \mathbf{r}_k) = \mathbf{0}$ when $j \neq k$, and $\mathrm{var}(\boldsymbol{\epsilon}) = \sigma_0^2\mathbf{I}$. We use the notation $[\mathbf{w}|\mathbf{X}, \boldsymbol{\beta}, \{\mathbf{Z}_k\}, \{\mathbf{V}_k\}, \boldsymbol{\theta}]$, or, more simply, $[\mathbf{w}|\mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma}_\theta]$, to indicate the probability density function $\mathbf{w} \sim N(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}_\theta)$ where

$$\boldsymbol{\Sigma}_\theta = \sum_k \mathbf{Z}_k \mathbf{V}_k \mathbf{Z}_k' + \sigma_0^2\mathbf{I}.$$

The covariance matrices $\{\mathbf{V}_k\}$ for $k = 1, \ldots, q$ can have additional covariance parameters beyond σ_0^2, all of which are contained in the vector $\boldsymbol{\theta}$. We will give more specific details on $\boldsymbol{\Sigma}_\theta$ later.

For the fully parametric, hierarchical models, a very general model can be constructed hierarchically as,

$$[\mathbf{y}, \mathbf{w}|\boldsymbol{\phi}, \mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma}_\theta] = [\mathbf{y}|\mathbf{g}^{-1}(\mathbf{w}), \boldsymbol{\phi}][\mathbf{w}|\mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma}_\theta], \tag{2}$$

where Berliner (1996) called $[\mathbf{y}|\mathbf{g}^{-1}(\mathbf{w}), \boldsymbol{\phi}]$ the *data model* and $[\mathbf{w}|\mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma}_\theta]$ the *process model*. As a concrete example, suppose that $[\mathbf{y}|\exp(\mathbf{w})]$ is Poisson, and $[\mathbf{w}|\mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma}_\theta]$ is multivariate normal, then the joint likelihood is

$$[\mathbf{y}, \mathbf{w}|\mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma}_\theta] = \left(\prod_{i=1}^{n} \frac{\exp(w_i)^{y_i} \exp(-\exp(w_i))}{y_i!}\right) \frac{\exp\left(-0.5(\mathbf{w} - \mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Sigma}_\theta^{-1}(\mathbf{w} - \mathbf{X}\boldsymbol{\beta})\right)}{(2\pi)^{n/2}|\boldsymbol{\Sigma}_\theta|^{1/2}},$$

and note the use of $\exp(w_i)$ for $\mathrm{E}(y_i|g^{-1}(w_i))$.

1.2 | Patterned covariance matrices

To construct a likelihood for (2) we will need parametric models for $\boldsymbol{\Sigma}_\theta$ in (1). There are few constraints here, and any valid covariance model for $\boldsymbol{\Sigma}_\theta$ is possible. For example, $\boldsymbol{\Sigma}_\theta$ may be constructed from typical mixed models where $\mathbf{Z}_k$ contains indicator variables for random intercepts, or explanatory variables for random slopes, and where $\mathbf{V}_k = \sigma_k^2\mathbf{I}$, and then $\boldsymbol{\Sigma}_\theta = \sum_k \sigma_k^2\mathbf{Z}_k\mathbf{Z}_k' + \sigma_0^2\mathbf{I}$. We can also consider time series models (e.g., Hamilton, 1994). For example, for a first-order autoregressive (AR1) model with i and j being integers, let $q = 1$ and $\mathbf{Z}_1 = \mathbf{I}$, then $\mathbf{V}_1$ has as its i,jth entry $\sigma_1^2\rho^{|i-j|}/(1-\rho^2)$, where $0 < \sigma_1^2$ and $-1 \le \rho < 1$. Similarly, we can have geostatistical models (e.g., Chiles & Delfiner, 1999), such as the exponential autocovariance model, where $q = 1, \mathbf{Z}_1 = \mathbf{I}$, and the i,jth element of $\mathbf{V}_1$ is $\sigma_1^2\exp(-\delta_{i,j}/\rho)$ where $\delta_{i,j}$ is Euclidean distance between the ith and jth locations, $0 < \sigma_1^2$, and $\rho > 0$. Other spatial covariance types include the conditional autoregressive (CAR, Besag, 1974; Cressie, 1993) and simultaneous autoregressive models (SAR, Whittle, 1954; Ver Hoef et al., 2018), moving average models in time series (e.g., Hamilton, 1994) and spatial statistics (Haining, 1978), spatio-temporal models (Cressie & Wikle, 2011, p. 380), and models on non-Euclidean topologies such as a sphere (e.g., the earth, Huang et al., 2011; Gneiting, 2013), and networks such as roads (Ver Hoef, 2018) and streams (Ver Hoef & Peterson, 2010). Because a covariance matrix can be constructed by summing covariance matrices as variance components, mixtures of all models mentioned above can create a rich set of patterned covariance matrices for modeling dependent structures. In what follows, we develop inference based on any valid covariance matrix.

1.3 | Inference for HGLMMs

The combination of the data model, $[\mathbf{y}|\mathbf{g}^{-1}(\mathbf{w}), \boldsymbol{\phi}]$, where any distribution could be used that matches the type of data, and the process model, $[\mathbf{w}|\mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma}_\theta]$, that can allow for any patterned covariance matrix, provides a hierarchical construction (2) that is a very rich and flexible class of models. This class of models is not new.

There are two broad methods of analysis. The most obvious method is Bayesian and computes the posterior distribution of all latent variables and parameters. Due to intractable integrals, this is usually achieved with Markov chain Monte Carlo (MCMC) methods (Gelfand & Smith, 1990; Gilks et al., 1996), of which there are now many varieties. Bayesian hierarchical models in our context have been extremely popular, beginning with spatial statistics (e.g., Clayton & Kaldor, 1987), clustered data (e.g., Zeger & Karim, 1991), time series (e.g., Berliner, 1996), and longitudinal data (Kleinman & Ibrahim, 1998) among others, and have been aided by the introduction of the `WinBUGS` software (Lunn et al., 2000), and, for `R`, the `spBayes` package (Finley et al., 2007).

A second approach attempts to estimate covariance parameters and fixed effects marginally while integrating out over all latent random effects. This can also be done using MCMC methods as a numerical integrator (e.g., Christensen, 2004; Zhang, 2002), but it can be quite slow, so a more popular and deterministic method uses a Laplace approximation (Tierney & Kadane, 1986). In particular, Rue et al. (2009) proposed integrated nested Laplace approximation (INLA) as approximate Bayesian inference when using Gaussian Markov random fields. They exploit the marginal specification of conditional autoregressive models for computational gains and use several first-order Laplace approximations to estimate fixed effects. Wood (2020) provided further computational gains with dense covariance matrices. Automatic differentiation is used in software `glmmTMB` (Brooks et al., 2017) as a general computational approach to fitting these models. Our development builds primarily on Evangelou et al. (2011) and Bonat and Ribeiro Jr (2016), who use a second-order Laplace

approximation with geostatistical covariance structures for binary and count data. The more general formulation is given by Bonat and Ribeiro Jr (2016), and we will point out differences from our development and that of Bonat and Ribeiro Jr (2016) in our Methods section.

Our general goal is to provide complete marginal inference, from estimating covariance parameters and fixed effects, to making predictions for unobserved data, for any patterned covariance matrix in the HGLMM framework. In particular, our goals are to: (1) find marginal maximum likelihood and restricted maximum likelihood estimates for covariance parameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$, (2) predict the latent values of $\mathbf{w}$, (3) estimate fixed effects $\boldsymbol{\beta}$, and (4) make predictions of new values of the process that generated $\mathbf{w}$ at unsampled times or locations.

The rest of this paper is organized as follows. In Section 2, we use the Laplace approximation to develop marginal maximum likelihood estimates for $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ using Newton–Raphson updates, which also leads to predictions of $\mathbf{w}$. From the predictions of $\mathbf{w}$ we develop estimators of $\boldsymbol{\beta}$ with proper confidence intervals and prediction of new values of the process generating $\mathbf{w}$ with proper prediction intervals. In Section 3, we conduct simulations to illustrate all methods and validate the earlier development, and we compare our methods to Bayesian methods using MCMC (through `R` package `spBayes`), to a Laplace approximation using automatic differentiation (through `R` package `glmmTMB`), and to an integrated nested Laplace approximation (through `R` package `INLA`). Section 4 provides three separate examples using real data sets to further illustrate the methods. We conclude with some discussion in Section 5.

2 | METHODS

When considering the hierarchical model formulation of the HGLMMs, we would like to marginalize the distribution $[\mathbf{w}, \mathbf{y}|\boldsymbol{\phi}, \mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma}_\theta] = [\mathbf{y}|\mathbf{g}^{-1}(\mathbf{w}), \boldsymbol{\phi}][\mathbf{w}|\mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma}_\theta]$ over $\mathbf{w}$ and be free of $\boldsymbol{\beta}$ as well to obtain a distribution of only the data and variance/covariance parameters. The Laplace method helps achieve that.

2.1 | Laplace approximation to likelihoods of HGLMMs

First, consider integrating over $\boldsymbol{\beta}$ as well as $\mathbf{w}$,

$$[\mathbf{y}|\boldsymbol{\phi}, \mathbf{X}, \boldsymbol{\Sigma}_\theta] = \int_{\mathbf{w}\in\mathbb{R}^n} \int_{\boldsymbol{\beta}\in\mathbb{R}^p} [\mathbf{w}, \mathbf{y}|\boldsymbol{\phi}, \mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma}_\theta] d\boldsymbol{\beta} d\mathbf{w} = \int_{\mathbf{w}} [\mathbf{y}|\mathbf{g}^{-1}(\mathbf{w}), \boldsymbol{\phi}] \int_{\boldsymbol{\beta}} [\mathbf{w}|\mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma}_\theta] d\boldsymbol{\beta} d\mathbf{w}.$$

When $[\mathbf{w}|\mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma}_\theta]$ is Gaussian, $\int_{\boldsymbol{\beta}} [\mathbf{w}|\mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma}_\theta] d\boldsymbol{\beta}$ is the likelihood for restricted (also known as residual) maximum likelihood estimation (REML, see Appendix). Note that REML was originally derived as the likelihood of a set of $n - p$ independent linear combinations of the observations known as error contrasts (Patterson & Thompson, 1971, 1974), and there is little literature on its derivation from integration, except Harville (1974), who integrates a linear contrast of the data that results in an extra $\det(\mathbf{X}'\mathbf{X})$ in the likelihood. Alternatively, consider $[\mathbf{w}|\mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\Sigma}_\theta]$ where $\boldsymbol{\beta}$ has been replaced by its conditional (on $\mathbf{w}$) maximum likelihood (ML) estimator, $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\boldsymbol{\Sigma}_\theta^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}_\theta^{-1}\mathbf{w}$. Then, both cases are free of $\boldsymbol{\beta}$,

$$[\mathbf{w}|\mathbf{X}, \boldsymbol{\Sigma}_\theta] = \frac{1}{C_n} \exp[-(1/2)(\mathbf{w} - \mathbf{X}\hat{\boldsymbol{\beta}})'\boldsymbol{\Sigma}_\theta^{-1}(\mathbf{w} - \mathbf{X}\hat{\boldsymbol{\beta}})],$$

where for ML estimation $C_n = \sqrt{2\pi^{n/2}|\boldsymbol{\Sigma}_\theta|}$ and for REML estimation $C_n = \sqrt{2\pi^{(n-p)/2}|\boldsymbol{\Sigma}_\theta||\mathbf{X}'\boldsymbol{\Sigma}_\theta^{-1}\mathbf{X}|}$. Note that Bonat and Ribeiro Jr (2016) only considered the marginal likelihood integrated over $\mathbf{w}$, and did not consider the likelihood where $\boldsymbol{\beta}$ was also integrated out (as in REML estimation) or back-substituted (as in ML estimation).

Next, to obtain the marginal distribution of the data and covariance parameters, we need the integral,

$$[\mathbf{y}|\boldsymbol{\phi}, \mathbf{X}, \boldsymbol{\Sigma}_\theta] = \int_{\mathbf{w}} [\mathbf{y}|\mathbf{g}^{-1}(\mathbf{w}), \boldsymbol{\phi}][\mathbf{w}|\mathbf{X}, \boldsymbol{\Sigma}_\theta] d\mathbf{w}.$$

Let us denote $\ell(\mathbf{w}, \cdot) = \log([\mathbf{y}|\mathbf{g}^{-1}(\mathbf{w}), \boldsymbol{\phi}][\mathbf{w}|\mathbf{X}, \boldsymbol{\Sigma}_\theta])$, and consider $\int e^{\ell(\mathbf{w}, \cdot)} d\mathbf{w}$. Let $\mathbf{v}$ be the gradient vector with ith element

$$v_i(\boldsymbol{\phi}, \boldsymbol{\theta}) = \frac{\partial \ell(\mathbf{w}, \cdot)}{\partial w_i},$$

and let $\mathbf{H}$ be the Hessian matrix with i,jth element,

$$H_{i,j}(\boldsymbol{\phi},\theta) = \frac{\partial^2 \ell(\mathbf{w},\cdot)}{\partial w_i \partial w_j},$$

where in both $v_i(\boldsymbol{\phi},\theta)$ and $H_{i,j}(\boldsymbol{\phi},\theta)$ we show dependence on parameters $\boldsymbol{\phi}$ and θ. Henceforth, we assume that the log-likelihood is sufficiently well-behaved so that $\mathbf{H}$ is negative definite. Using the multivariate Taylor series expansion of $\ell(\mathbf{w},\cdot)$ around some point $\mathbf{a}$,

$$\int_{\mathbf{w}} e^{\ell(\mathbf{w},\cdot)} d\mathbf{w} \approx \int_{\mathbf{w}} e^{\ell(\mathbf{a},\cdot)+\mathbf{v}'(\mathbf{w}-\mathbf{a})+1/2(\mathbf{w}-\mathbf{a})'\mathbf{H}(\mathbf{w}-\mathbf{a})} d\mathbf{w}.$$

Now if $\mathbf{a}$ is a value at which $\mathbf{v}=\mathbf{0}$, then

$$\int_{\mathbf{w}} e^{\ell(\mathbf{w},\cdot)} d\mathbf{w} \approx e^{\ell(\mathbf{a},\cdot)} \int_{\mathbf{w}} e^{-1/2(\mathbf{w}-\mathbf{a})'(-\mathbf{H})(\mathbf{w}-\mathbf{a})} d\mathbf{w} = e^{\ell(\mathbf{a},\cdot)}(2\pi)^{n/2} |-\mathbf{H}_{\mathbf{a}}(\boldsymbol{\phi},\theta)|^{-1/2},$$

where $\mathbf{H}_{\mathbf{a}}(\boldsymbol{\phi},\theta)$ indicates $\mathbf{H}$ evaluated at $\mathbf{a}$ and we again show its dependence on $\boldsymbol{\phi}$ and θ. The result on the most right-hand side is familiar from the normalizing constant of a multivariate Gaussian distribution. Hence,

$$[\mathbf{y}|\boldsymbol{\phi},\mathbf{X},\boldsymbol{\Sigma}_\theta] = \int_{\mathbf{w}} e^{\ell(\mathbf{w},\cdot)} d\mathbf{w} \approx [\mathbf{y}|\mathbf{g}^{-1}(\mathbf{a}),\boldsymbol{\phi}][\mathbf{a}|\mathbf{X},\boldsymbol{\Sigma}_\theta](2\pi)^{n/2}|-\mathbf{H}_{\mathbf{a}}(\boldsymbol{\phi},\theta)|^{-1/2}. \tag{3}$$

2.2 | Marginal maximum likelihood for covariance parameters

From (3) an approximate marginal maximum likelihood estimator for $\boldsymbol{\phi}$ and θ depending on finding $\mathbf{a}$ is

$$\{\hat{\boldsymbol{\phi}},\hat{\theta}\} = \arg\max_{\boldsymbol{\phi},\theta}\Big(\log[\mathbf{y}|\mathbf{g}^{-1}(\mathbf{a}),\boldsymbol{\phi}] + \log[\mathbf{a}|\mathbf{X},\boldsymbol{\Sigma}_\theta] - \frac{1}{2}\log(|-\mathbf{H}_{\mathbf{a}}(\boldsymbol{\phi},\theta)|)\Big), \tag{4}$$

where we drop terms that do not contain $\boldsymbol{\phi}$ or θ. Note that $\log[\mathbf{a}|\mathbf{X},\boldsymbol{\Sigma}_\theta]$ has the same form of the log-likelihood for ML or REML as in standard Gaussian models, but here it is evaluated at $\mathbf{a}$, where for ML

$$\log[\mathbf{a}|\mathbf{X},\boldsymbol{\Sigma}_\theta] = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}_\theta| - \frac{1}{2}(\mathbf{a}-\mathbf{X}\hat{\boldsymbol{\beta}}_{\mathbf{a}})'\boldsymbol{\Sigma}_\theta^{-1}(\mathbf{a}-\mathbf{X}\hat{\boldsymbol{\beta}}_{\mathbf{a}}), \tag{5}$$

and for REML

$$\log[\mathbf{a}|\mathbf{X},\boldsymbol{\Sigma}_\theta] = -\frac{n-p}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}_\theta| - \frac{1}{2}\log|\mathbf{X}'\boldsymbol{\Sigma}_\theta^{-1}\mathbf{X}| - \frac{1}{2}(\mathbf{a}-\mathbf{X}\hat{\boldsymbol{\beta}}_{\mathbf{a}})'\boldsymbol{\Sigma}_\theta^{-1}(\mathbf{a}-\mathbf{X}\hat{\boldsymbol{\beta}}_{\mathbf{a}}), \tag{6}$$

where in both cases $\hat{\boldsymbol{\beta}}_{\mathbf{a}} = (\mathbf{X}'\boldsymbol{\Sigma}_\theta^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}_\theta^{-1}\mathbf{a}$. The result (4) depends on finding $\mathbf{a}$ such that $\mathbf{v}=\mathbf{0}$. To achieve this, we use Newton–Raphson, conditional on $\boldsymbol{\phi}$ and θ, which we describe next.

Assuming the conditional independence of the elements of $\mathbf{y}$ given $\mathbf{g}^{-1}(\mathbf{w})$, we have

$$\log([\mathbf{y}|\mathbf{g}^{-1}(\mathbf{w}),\boldsymbol{\phi}][\mathbf{w}|\mathbf{X},\boldsymbol{\Sigma}_\theta]) = \sum_{i=1}^{n}\log[y_i|g_i^{-1}(w_i),\boldsymbol{\phi}] - \frac{1}{2}(\mathbf{w}-\mathbf{X}\hat{\boldsymbol{\beta}})'\boldsymbol{\Sigma}_\theta^{-1}(\mathbf{w}-\mathbf{X}\hat{\boldsymbol{\beta}}) + C, \tag{7}$$

where C comprises terms that do not contain $\mathbf{w}$. Let $\mathbf{d}_{\boldsymbol{\phi}}$ be the vector with ith component,

$$d_i \equiv \frac{\partial \log[y_i|g_i^{-1}(w_i),\boldsymbol{\phi}]}{\partial w_i},$$

and note that

$$\frac{\partial[-\frac{1}{2}(\mathbf{w}-\mathbf{X}\hat{\boldsymbol{\beta}})'\boldsymbol{\Sigma}_\theta^{-1}(\mathbf{w}-\mathbf{X}\hat{\boldsymbol{\beta}})]}{\partial \mathbf{w}} = -\boldsymbol{\Sigma}_\theta^{-1}\mathbf{w} + \boldsymbol{\Sigma}_\theta^{-1}\mathbf{X}\hat{\boldsymbol{\beta}},$$

so the gradient of (7) is

$$\mathbf{v} = \mathbf{d}_{\boldsymbol{\phi}} - \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\mathbf{w} + \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{d}_{\boldsymbol{\phi}} - \mathbf{P}_{\boldsymbol{\theta}}\mathbf{w}, \tag{8}$$

where $\mathbf{P}_{\boldsymbol{\theta}} = \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} - \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\mathbf{X}(\mathbf{X}'\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}$. For the Hessian, let $\mathbf{D}_{\boldsymbol{\phi}}$ be a diagonal matrix with ith diagonal element,

$$D_{i,i} \equiv \frac{\partial^2 \log[y_i|g_i^{-1}(w_i), \boldsymbol{\phi}]}{\partial w_i^2}, \tag{9}$$

where all off-diagonal elements are zero because all second-order partial derivatives are 0 when $i \neq j$ due to conditional independence. Then the Hessian of (7) is

$$\mathbf{H} = \mathbf{D}_{\boldsymbol{\phi}} - \mathbf{P}_{\boldsymbol{\theta}}. \tag{10}$$

Note that (8) and (10) differ from the gradient and Hessian in Bonat and Ribeiro Jr (2016) because we used $\hat{\boldsymbol{\beta}}$ in (7), which contains $\mathbf{w}$, whereas they used $\boldsymbol{\beta}$. In fact, the gradient and Hessian in Bonat and Ribeiro Jr (2016) may be obtained from (8) and (10) by replacing $\mathbf{P}_{\boldsymbol{\theta}}$ with $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}$. This is an important difference, as it allows us to optimize for just the covariance parameters without having to do so for $\boldsymbol{\beta}$ simultaneously.

A table of d_i and $D_{i,i}$ for a few common distributions and link functions is given in Table 1. In Table 1, we use alternative parameterizations for the negative binomial, gamma, and beta distributions so that $\mathrm{E}(y) = \mu$. We also reparameterize the inverse Gaussian distribution. Details for all distributions are given in the Appendix.

Conditional on $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$, a Newton–Raphson update is,

$$\mathbf{w}^{[k+1]} = \mathbf{w}^{[k]} - \mathbf{H}^{-1}\mathbf{v},$$

and upon convergence we set $\mathbf{a} = \mathbf{w}$ in (4) for any evaluation of the likelihood for given $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$. Notice that this makes the marginal maximum likelihood doubly iterative, as we solve for $\mathbf{a}$ while optimizing for $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$. It is possible to use other maximization routines, such as the EM algorithm, but generally the Newton–Raphson algorithm converges rapidly (often around 10 iterations in our experience), and this was favored by Bonat and Ribeiro Jr (2016) also. However, on occasion, the stepsize needs to be adjusted so that $\mathbf{v}$ does not diverge. For example, it is easy and fast to check $\mathbf{v}^{[k+1]} = \mathbf{d}_{\boldsymbol{\phi}} - \mathbf{P}_{\boldsymbol{\theta}}\mathbf{w}^{[k+1]}$, and if $\mathbf{v}^{[k+1]}$ is "larger" than $\mathbf{v}$ by some criterion (e.g., largest or average element of $\mathbf{v}$), then take

$$\mathbf{w}^{[k+1]} = \mathbf{w}^{[k]} - \alpha\mathbf{H}^{-1}\mathbf{v},$$

where $0 < \alpha < 1$. In the simulations below, we check $\mathbf{v}^{[k+1]}$ in the manner described above, and set $\alpha = 0.1$ if the largest element of $\mathbf{v}^{[k+1]}$ is larger than the largest element of $\mathbf{v}$. The advantage of using Newton–Raphson is that it provides $\mathbf{H}$,

TABLE 1 Flexibility of the HGLMM, showing how different distributions can be matched with different patterned covariance matrices. We also show distributions, inverse link functions, and first and second partial derivatives with respect to w_i for various parts of the log-likelihood. Any of the covariance models in the last column can be matched to any of the distributions in the first column.

$\log[\mathbf{y}\|\mathbf{g}^{-1}(\mathbf{w}), \boldsymbol{\phi}]$		$-(1/2)\log\|-\mathbf{H}_{\mathbf{a}}(\boldsymbol{\phi}, \boldsymbol{\theta})\|$		$+\log[\mathbf{a}\|\mathbf{X}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}]$
Distribution	$\boldsymbol{\mu} = \mathbf{g}^{-1}(\mathbf{w})$	$\boldsymbol{d_i}$	$\boldsymbol{D_{i,i}}$	$\boldsymbol{\Sigma_\theta}$**-types**
Binomial	$\mu = \frac{\exp(\mathbf{w})}{1+\exp(\mathbf{w})}$	$y_i - \frac{n_i \exp(w_i)}{1+\exp(w_i)}$	$-\frac{n_i \exp(w_i)}{(1+\exp(w_i))^2}$	Random effects
Poisson	$\mu = \exp(\mathbf{w})$	$y_i - \exp(w_i)$	$-\exp(w_i)$	Geostatistical
Neg. Binomial	$\mu = \exp(\mathbf{w})$	$\frac{\phi(y_i - e^{w_i})}{\phi + e^{w_i}}$	$-\frac{\phi e^{w_i}(\phi + y_i)}{(\phi + e^{w_i})^2}$	Spatial areal
Gamma	$\mu = \exp(\mathbf{w})$	$-\phi + y_i \phi e^{-w_i}$	$-y_i \phi e^{-w_i}$	Time series
Inv. Gaussian	$\mu = \exp(\mathbf{w})$	$\phi\left(\frac{y}{2e^{w_i}} - \frac{e^{w_i}}{2y}\right) + \frac{1}{2}$	$-\frac{\phi(e^{2w_i} + y_i^2)}{2ye^{w_i}}$	Spatio-temporal
Beta	$\mu = \frac{\exp(\mathbf{w})}{1+\exp(\mathbf{w})}$	$\frac{-\phi e^{w_i} k_0(w_i\|\phi, y_i)}{(e^{w_i}+1)^2}$	$\frac{-\phi e^{2w_i} k_1(w_i\|\phi, y_i)}{(e^{w_i}+1)^4}$	

Note: $k_0(w_i|\phi, y_i) = \psi^{(0)}\left(\frac{\phi e^{w_i}}{1+e^{w_i}}\right) - \psi^{(0)}\left(\frac{\phi}{1+e^{w_i}}\right) + \log\left(\frac{1}{y_i} - 1\right)$. $k_1(w_i|\phi, y_i) = \phi\left(\psi^{(1)}\left(\frac{\phi e^{w_i}}{1+e^{w_i}}\right) + \psi^{(1)}\left(\frac{\phi}{1+e^{w_i}}\right)\right) - 2\sinh(w_i)\left(k_0(w_i|\phi, y_i) + 2\tanh^{-1}(1 - 2y_i)\right)$. $\psi^{(n)}(\cdot)$ is the nth derivative of the digamma function. sinh and tanh are the hyperbolic sine and tangent functions, respectively.

which is useful for making adjustments to variances when estimating fixed effects and making predictions, which we describe in the next section.

In summary, estimation of covariance parameters and $\mathbf{w}$ can be written in the following steps,

1. Get initial values for covariance parameters $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$. For example, for variance components, such as σ_0^2 and σ_1^2, apportion var($g(\mathbf{y})$) equally among each variance component so that $\sigma_0^2 + \sigma_1^2 = \text{var}(g(\mathbf{y}))$. If there are many explanatory variables, a linear model can be fit to $g(\mathbf{y})$) and residual variance could be used. For geostatistical models, the range parameter can be set such that the effective range (the value at which autocorrelation is 0.05) is 0.9 (or some other fraction) of the maximum distance encountered in the data set. For areal data with row-standardized neighborhood matrices, the dependence parameter is often bounded between 0 and 1, so 0.5 is a good starting value. In general, a coarse grid search is best, where that grid search is not too computationally demanding. For example, one could try all combinations of a small, medium, and large values for each parameter, where small, medium, and large are determined from the model and data.
2. Pick initial values for $\mathbf{w}$. For example, set $\mathbf{w} = g(\mathbf{y})$, or as the residuals from an ordinary least squares model fit to $g(\mathbf{y})$, or as deviance residuals from a generalized linear model fit (using iteratively reweighted least squares) to $\mathbf{y}$.
3. Use Newton–Raphson to estimate $\mathbf{w} = \mathbf{a}$ for given $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ in (4).
4. Evaluate the log-likelihood in (4) for $\mathbf{a}$, $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$.
5. Loop through steps 3 and 4 for different values of $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ while optimizing for the log-likelihood in step 4 until convergence.

2.3 | Inference for fixed effects

In order to estimate $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ it was necessary to optimize the likelihood for $\mathbf{w}$, which we called $\mathbf{a}$, using Newton–Raphson, for each evaluation of the likelihood. Upon convergence in estimating $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$, we also have optimized $\mathbf{w}$. Let us denote the optimizing value as $\hat{\mathbf{w}} = \mathbf{a}$.

Bonat and Ribeiro Jr (2016) proposed profile likelihood for estimating $\boldsymbol{\beta}$ and obtaining confidence intervals, but their proposal is computationally demanding and does not extend well to cases with many coefficients in $\boldsymbol{\beta}$. An alternative estimator of $\boldsymbol{\beta}$ may be obtained by replacing the unobserved $\mathbf{w}$ with its predicted value $\hat{\mathbf{w}} = \mathbf{a}$, obtained as described in the previous subsection, in the expression for what would be the generalized least squares estimator of $\boldsymbol{\beta}$ if in fact $\mathbf{w}$ *was* observed. This yields the estimator $\hat{\boldsymbol{\beta}} = \mathbf{B}\hat{\mathbf{w}}$, where $\mathbf{B} = (\mathbf{X}'\boldsymbol{\Sigma}_\theta^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}_\theta^{-1}$. In order to estimate the variance of $\hat{\boldsymbol{\beta}}$, it is convenient to condition on $\mathbf{w}$ as if it were observed, often called the law of total variance,

$$\text{var}(\hat{\mathbf{w}}) = \text{E}_{\mathbf{w}}[\text{var}(\hat{\mathbf{w}}|\mathbf{w})] + \text{var}_{\mathbf{w}}[\text{E}(\hat{\mathbf{w}}|\mathbf{w})]. \tag{11}$$

Due to the optimization of $\hat{\mathbf{w}}$ from the likelihood, we will assume that $\hat{\mathbf{w}}|\mathbf{w}$ is approximately distributed as $\text{N}(\mathbf{w}, \mathbf{F}_{\mathbf{w}}^{-1})$, where $\mathbf{F}_{\mathbf{w}}$ is the observed Fisher information, or, less strictly, that $\text{E}(\hat{\mathbf{w}}|\mathbf{w}) = \mathbf{w}$ and $\text{var}(\hat{\mathbf{w}}|\mathbf{w}) = \mathbf{F}_{\mathbf{w}}^{-1}$, approximately. Thus, for the second term in (11), $\text{var}_{\mathbf{w}}[\text{E}(\hat{\mathbf{w}}|\mathbf{w})] = \boldsymbol{\Sigma}_\theta$, which we approximate by $\boldsymbol{\Sigma}_{\hat{\theta}}$ after substituting estimated parameters $\hat{\boldsymbol{\theta}}$ for $\boldsymbol{\theta}$. For the first term in (11), the observed Fisher information is equivalent to $-\mathbf{H}_{\mathbf{w}}(\boldsymbol{\phi}, \boldsymbol{\theta})^{-1}$, where we show the dependence on parameters $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$, and on $\mathbf{w}$ that comes from $\mathbf{D}_{\boldsymbol{\phi}}$ in (10) (see the examples of $D_{i,i}$ in Table 1). To obtain the Fisher information would require taking the expectation, $\text{E}_{\mathbf{w}}[\text{var}(\hat{\mathbf{w}}|\mathbf{w})]$, but this is complicated, and Efron and Hinkley (1978) argue for using the observed Fisher information instead, so we simply replace $\mathbf{w}$ in $-\mathbf{H}_{\mathbf{w}}(\boldsymbol{\phi}, \boldsymbol{\theta})^{-1}$ with $\hat{\mathbf{w}} = \mathbf{a}$, and we also replace $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ by their estimates $\hat{\boldsymbol{\phi}}$ and $\hat{\theta}$, and denote this as $-\mathbf{H}_{\hat{\mathbf{w}}}(\hat{\boldsymbol{\phi}}, \hat{\theta})^{-1}$. Then an estimator of the covariance matrix of estimated fixed effects is

$$\widehat{\text{var}}(\hat{\boldsymbol{\beta}}) = \mathbf{B}[\text{var}(\hat{\mathbf{w}})]\mathbf{B}' = \mathbf{B}[-\mathbf{H}_{\hat{\mathbf{w}}}(\hat{\boldsymbol{\phi}}, \hat{\theta})^{-1}]\mathbf{B}' + \mathbf{C}_{\hat{\beta}}, \tag{12}$$

where $\mathbf{C}_{\hat{\beta}} = \mathbf{B}\boldsymbol{\Sigma}_{\hat{\theta}}\mathbf{B}'$, which simplifies to $\mathbf{C}_{\hat{\beta}} = (\mathbf{X}'\boldsymbol{\Sigma}_{\hat{\theta}}^{-1}\mathbf{X})^{-1}$, the usual estimated variance-covariance matrix of fixed effects when using generalized least squares if $\mathbf{w}$ were observed.

2.4 | Inference for prediction

So far, we have estimated $\boldsymbol{\theta}$, $\boldsymbol{\phi}$, and $\boldsymbol{\beta}$, predicted $\mathbf{w}$, and obtained estimated covariance matrices for $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{w}}$. Now let us consider the task of prediction for unsampled data, which may be in space, or time, or by design. We will denote the

unsampled $\{w_i\}$ by the vector $\mathbf{u}$. We can extend the linear model (1) as

$$\begin{pmatrix}\mathbf{w}\\ \mathbf{u}\end{pmatrix} \sim N\left(\begin{pmatrix}\mathbf{X}\boldsymbol{\beta}\\ \mathbf{X_u}\boldsymbol{\beta}\end{pmatrix}, \begin{pmatrix}\boldsymbol{\Sigma}_\theta & \boldsymbol{\Sigma}_{\mathbf{wu}}\\ \boldsymbol{\Sigma}'_{\mathbf{wu}} & \boldsymbol{\Sigma}_{\mathbf{uu}}\end{pmatrix}\right). \tag{13}$$

Our goal is the prediction of $\mathbf{u}$. If $\mathbf{w}$ was observed, the best linear unbiased predictor (BLUP) of $\mathbf{u}$ would be $\boldsymbol{\Lambda}\mathbf{w}$, where $\boldsymbol{\Lambda} = \mathbf{X_u}\mathbf{B} + \boldsymbol{\Sigma}'_{\mathbf{wu}}\boldsymbol{\Sigma}_\theta^{-1} - \boldsymbol{\Sigma}'_{\mathbf{wu}}\boldsymbol{\Sigma}_\theta^{-1}\mathbf{XB}$ (Goldberger, 1962). Since, however, $\mathbf{w}$ is unobserved, the empirical BLUP (eBLUP) of $\mathbf{u}$ may be obtained by substituting $\hat{\mathbf{w}}$ for $\mathbf{w}$ in this expression, yielding $\hat{\mathbf{u}} = \boldsymbol{\Lambda}\hat{\mathbf{w}}$.

To determine the sampling properties of $\hat{\mathbf{u}}$, we again we need to make some adjustments that account for the substitution of $\hat{\mathbf{w}}$ for $\mathbf{w}$ in the BLUP. Assuming again that $\hat{\mathbf{w}}$ is unbiased for $\mathbf{w}$, it is easily seen that this alternative predictor is unbiased, i.e., $\mathrm{E}(\boldsymbol{\Lambda}\hat{\mathbf{w}}) = \mathrm{E}(\mathbf{u})$. Now we want an estimator of the prediction error variance associated with this predictor, which is $\mathrm{var}(\hat{\mathbf{u}} - \mathbf{u}) = \mathrm{var}(\boldsymbol{\Lambda}\hat{\mathbf{w}} - \mathbf{u})$. Note that the prediction error variance of the BLUP is

$$\mathrm{var}(\boldsymbol{\Lambda}\mathbf{w} - \mathbf{u}) = \boldsymbol{\Sigma}_{\mathbf{uu}} - \boldsymbol{\Sigma}'_{\mathbf{wu}}\boldsymbol{\Sigma}_\theta^{-1}\boldsymbol{\Sigma}_{\mathbf{wu}} + \mathbf{K}\mathbf{C}_\beta\mathbf{K}', \tag{14}$$

where $\mathbf{K} = \mathbf{X_u} - \boldsymbol{\Sigma}'_{\mathbf{wu}}\boldsymbol{\Sigma}_\theta^{-1}\mathbf{X}$ (Goldberger, 1962). To obtain the prediction error variance of our alternative predictor, it is convenient to condition on $\mathbf{w}$ and $\mathbf{u}$ as we did in (11), that is,

$$\mathrm{var}(\boldsymbol{\Lambda}\hat{\mathbf{w}} - \mathbf{u}) = \mathrm{E}_{\mathbf{w},\mathbf{u}}[\mathrm{var}(\boldsymbol{\Lambda}\hat{\mathbf{w}} - \mathbf{u}|\mathbf{w},\mathbf{u})] + \mathrm{var}_{\mathbf{w},\mathbf{u}}[\mathrm{E}(\boldsymbol{\Lambda}\hat{\mathbf{w}} - \mathbf{u}|\mathbf{w},\mathbf{u})].$$

Owing to the assumed unbiasedness of $\hat{\mathbf{w}}$ for $\mathbf{w}$, we have $\mathrm{E}(\boldsymbol{\Lambda}\hat{\mathbf{w}} - \mathbf{u}|\mathbf{w},\mathbf{u}) = \boldsymbol{\Lambda}\mathbf{w} - \mathbf{u}$, and the variance of this is given by (14). Conditionally, $\mathrm{var}_{\hat{\mathbf{w}}}(\boldsymbol{\Lambda}\hat{\mathbf{w}} - \mathbf{u})$ does not depend on $\mathbf{u}$, so $\mathrm{E}_{\mathbf{w},\mathbf{u}}[\mathrm{var}(\boldsymbol{\Lambda}\hat{\mathbf{w}} - \mathbf{u}|\mathbf{w},\mathbf{u})] = \mathrm{E}_{\mathbf{w}}(\boldsymbol{\Lambda}[-\mathbf{H}_{\mathbf{w}}(\boldsymbol{\phi},\theta)^{-1}]\boldsymbol{\Lambda}')$, and, as we did in the previous section, rather than take expectation, we simply use the observed Fisher information by replacing $\mathbf{w}$ in $\mathbf{H}$ with its estimator $\hat{\mathbf{w}} = \mathbf{a}$ and replacing $\boldsymbol{\phi}$ and θ with $\hat{\boldsymbol{\phi}}$ and $\hat{\theta}$. Putting them together, we obtain

$$\widehat{\mathrm{var}}(\boldsymbol{\Lambda}\hat{\mathbf{w}} - \mathbf{u}) = \boldsymbol{\Lambda}[-\mathbf{H}_{\hat{\mathbf{w}}}(\hat{\boldsymbol{\phi}},\hat{\theta})^{-1}]\boldsymbol{\Lambda}' + \boldsymbol{\Sigma}_{\mathbf{uu}} - \boldsymbol{\Sigma}'_{\mathbf{wu}}\boldsymbol{\Sigma}_{\hat{\theta}}^{-1}\boldsymbol{\Sigma}_{\mathbf{wu}} + \mathbf{K}\mathbf{C}_{\hat{\beta}}\mathbf{K}'. \tag{15}$$

All covariance matrices depend on θ, although the notation makes this explicit only for the covariance matrix of $\mathbf{w}$. We replace θ in these matrices by its estimator $\hat{\theta}$, where the fitted covariance function that is used to estimate $\boldsymbol{\Sigma}_\theta$ is also used to estimate $\boldsymbol{\Sigma}_{\mathbf{wu}}$ and $\boldsymbol{\Sigma}_{\mathbf{uu}}$.

3 | SIMULATIONS

To test our methods, and compare to other methods, we perform some simulations so that we know the true values of all parameters and $\mathbf{w}$. First, we simulate a single data set to more carefully compare results with other methods, and then we conduct some simulation experiments to examine frequentist properties for mean-squared (prediction) error and confidence/prediction interval coverage. All code and results are available as an `R` package mentioned at the end of this article.

3.1 | A single simulated spatial poisson regression

We first illustrate our methods with a simulated spatial data set. We also compare our methods to a fully Bayesian approach by using the `R` package `spBayes` (Finley et al., 2007), and we compare our methods to an approach that uses the Laplace approximation, but with automatic differentiation rather than the analytical results that we developed throughout this article, by using the `R` package `glmmTMB` (Brooks et al., 2017). We created a square grid of 16×16 data locations on a $(0,1) \times (0,1)$ (unit square) domain. On this grid, we generated a single $\mathbf{w}$ from a multivariate normal distribution with zero mean and covariance matrix determined by evaluating an exponential autocovariance model, $\mathrm{cov}(w(\mathbf{s}_i), w(\mathbf{s}_j)) = \sigma_1^2 \exp(-\delta_{i,j}/\rho) + \sigma_0^2 \mathcal{I}(\delta_{i,j} = 0)$, where $\mathbf{s}_i$ $(i = 1, \ldots, 256)$ is the vector of spatial coordinates at location i, $\delta_{i,j}$ is Euclidean distance between the ith and jth locations, and $\mathcal{I}(\cdot)$ is the indicator function, equal to one if its argument is true and equal to 0 otherwise. We set $\sigma_1^2 = 1$, $\rho = 1$, and $\sigma_0^2 = 0.0001$. A single realization of this process is given in Figure 1a. To the

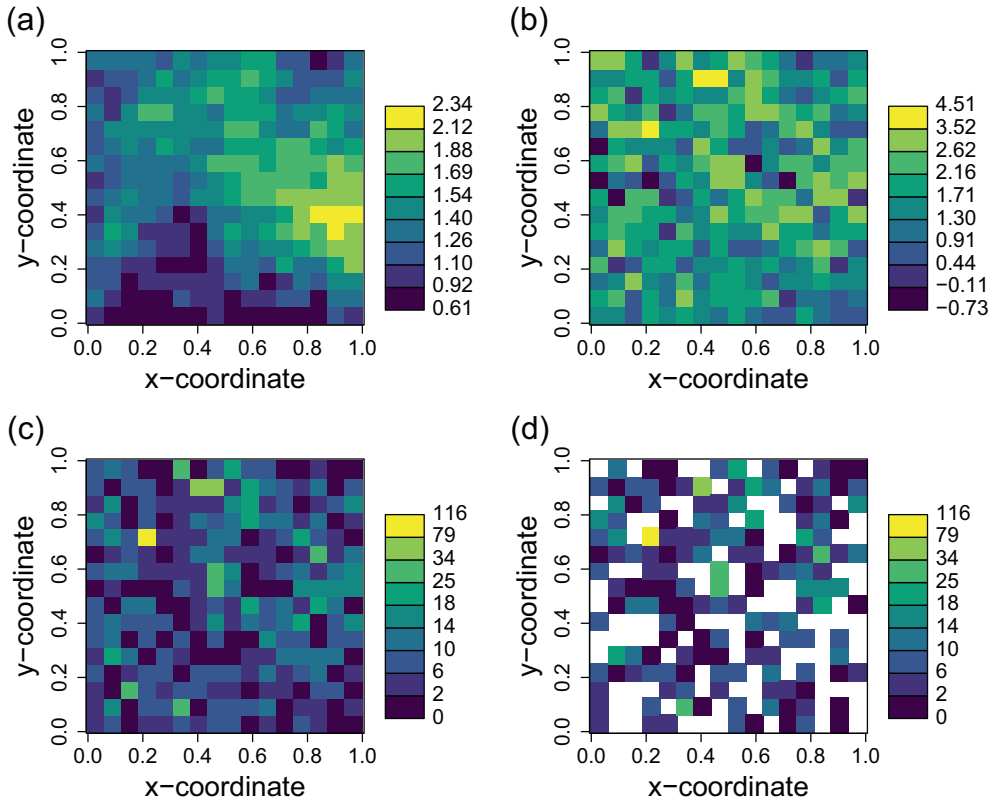

FIGURE 1 Simulated Poisson count data. (a) The true simulated $\mathbf{w}$ values. (b) The effects of explanatory variables added, $\mathbf{X}\boldsymbol{\beta} + \mathbf{w}$. (c) Simulated data from a Poisson distribution $Y \sim \text{Poi}(\exp(\mathbf{X}\boldsymbol{\beta} + \mathbf{w})$ (d) 100 response values removed at random, providing the final data set.

spatial random effects $\mathbf{w}$, we added fixed effects as a linear model

$$\text{E}(w_i) = \beta_0 + \beta_1 x_i + \beta_2 \tau_i + \beta_3 (x : \tau)_i,$$

where x_i was randomly and independently simulated from $\text{N}(0, 1)$, τ_i was randomly and independently simulated as a Bernoulli variable with probability $p = 0.5$, and $(x : \tau)_i$ was the interaction between the normally-distributed and Bernoulli-distributed explanatory variables (i.e., the product of x_i and τ_i). We set $\boldsymbol{\beta} = (0.5, 0.5, -0.5, 0.5)'$. Then a single realization of the fixed effects was added to the realized spatial random effects, $\mathbf{X}\boldsymbol{\beta} + \mathbf{w}$ (Figure 1b). Next, we created Poisson random variables under a HGLMM model by simulating $Y \sim \text{Poi}(\exp(\mathbf{X}\boldsymbol{\beta} + \mathbf{w}))$ (Figure 1c). Finally, we created 100 missing values at random from the grid (Figure 1d), where we set aside the true simulated values, leaving 156 observations for model fitting and 100 missing values for prediction.

The methods developed in this manuscript have been implemented in the `R` package `spmodel` (Dumelle et al., 2023). We fit models to the simulated data set using `spmodel`, `spBayes`, and `glmmTMB`, using a spatial exponential model for the covariance model for all three packages. All three packages assume the same HGLMM with a Poisson distribution. For `spBayes` we used a burn-in of 10,000 MCMC samples, and then used another 10,000 MCMC samples to summarize results. We used default priors as given in the `spBayes` examples. `glmmTMB` uses automatic differentiation rather than the analytical results that we develop in equations (4) through (10), and in Table 1. `glmmTMB` and `spmodel` require no extra input other than the response distribution and the form of the covariance matrix.

The estimated covariance parameters were somewhat different for all packages, and the partial sill, σ_1^2, and range ρ parameters for all three packages are shown on the HGLMM likelihood surface (for fixed nugget effect as estimated by `spmodel`) in Figure 2. Note that although the estimated covariance parameters among the three packages differ, they all lie along the ridge of high likelihood in Figure 2 (which actually shows -2 times the marginal loglikelihood), so that the ratio σ_1^2/ρ is very similar for all three packages, and it is this ratio that is important for estimation and prediction (Zhang, 2004).

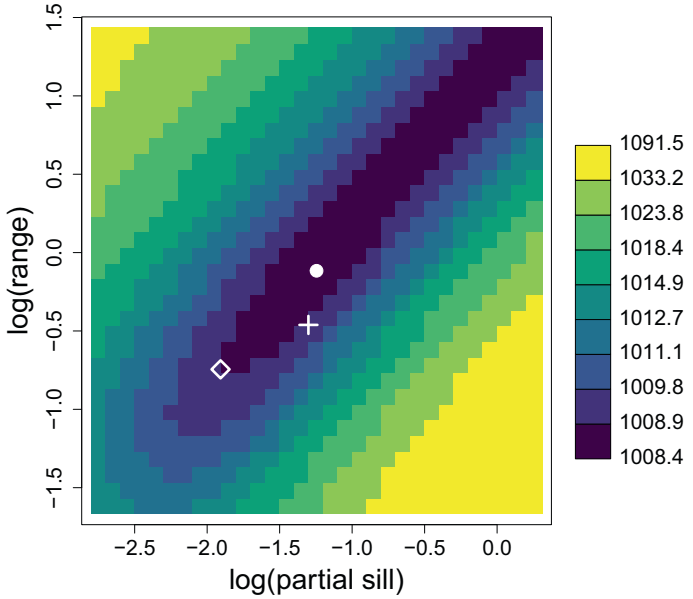

FIGURE 2 The marginal likelihood surface of two covariance parameters for the simulated data. The white circle shows the estimated values using the analytical method proposed here and implemented in `spmodel`, the plus symbol indicates the estimated values for `spBayes`, and the open diamond symbol indicates the estimated values using `glmmTMB`.

TABLE 2 Estimated fixed effects table for simulated data.

		Estimates			Standard error		
Effect	**True**	**spmodel**	**spBayes**	**glmmTMB**	**spmodel**	**spBayes**	**glmmTMB**
β_1	0.5	0.546	0.545	0.547	0.050	0.053	0.050
β_2	−0.5	−0.609	−0.615	−0.609	0.090	0.094	0.090
β_3	0.5	0.558	0.564	0.557	0.072	0.077	0.072

In fact, all three `R` packages gave very similar estimated fixed effects and standard errors (Table 2). Turning to prediction, Figure 3a shows the true simulated w-values (Figure 1a) and the predictions using `spmodel`, where the predictions are quite close to the true values. The predictions of `spmodel`, `spBayes`, and `glmmTMB` were virtually identical (Figure 3b,c). The prediction standard errors were slightly different, with those of `glmmTMB` about equal to `spmodel` on average, while those of `spBayes` were slightly higher (Figure 3d). The analysis of this simulated data set shows that all three methods yield very similar results.

3.2 | A simulation experiment for spatial poisson regression

The single data set provided insights on comparing estimates of three methods represented by `spmodel`, `spBayes`, and `glmmTMB`. In order to evaluate the bias of the estimators and predictors, mean-squared errors (MSE) of fixed effects, mean-squared prediction errors (MSPE), the coverage of confidence intervals and prediction intervals, and to compare computing time, we conducted a larger simulation experiment. We simulated the data exactly as in Figure 1 except the spatial locations were randomized within the unit square, rather than on a regular grid, and we used sample sizes of 125 and 525, in each case with 100 prediction locations. We simulated 2000 data sets to assess bias, MS(P)E, confidence/prediction interval coverage, and average computing times.

To estimate bias, we took the average of $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ (element-wise) over all 2000 simulated data sets, and to estimate mean-squared error (MSE) we took the average of $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^2$ (element-wise) over all 2000 simulated data sets. We also formed nominal 90% confidence intervals as $\hat{\beta}_j \pm 1.645\widehat{\text{se}}(\hat{\beta}_j)$, where, for `spmodel`, $\widehat{\text{se}}(\hat{\beta}_j)$ was the square root of the

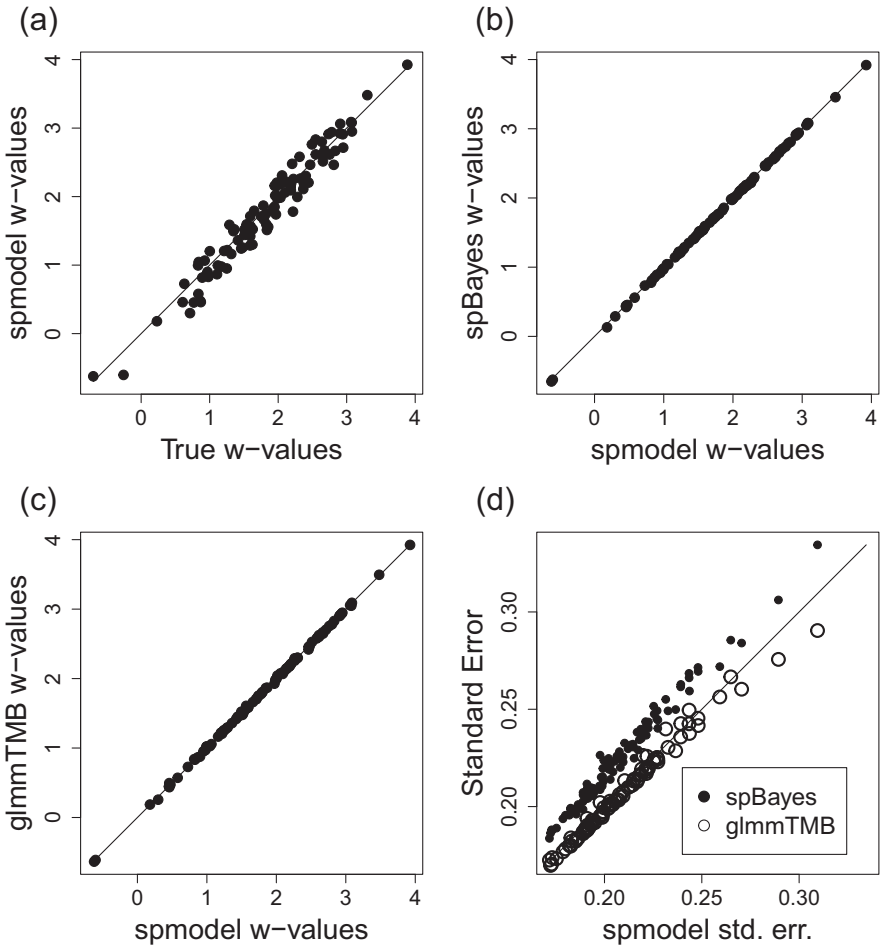

FIGURE 3 Comparisons among the three methods. (a) Predicted *w*-values using `spmodel` versus true *w*-values. (b) Predicted *w*-values using `spmodel` versus predicted *w*-values using `spBayes`. (c) Predicted *w*-values using `spmodel` versus predicted *w*-values using `glmmTMB`. (d) Prediction standard errors for `spBayes` and `glmmTMB` compared to `spmodel`.

$(j+1)$th diagonal element of (12) $(j = 0, 1, 2, 3)$. We computed the proportion of times, over the 2000 simulations, that the confidence intervals contained the true values. If we are estimating the variances of $\hat{\beta}_j$ well, the coverage should be close to 90%.

We also used the estimated covariance parameters in $\boldsymbol{\Sigma}_{\hat{\theta}}$ and the predicted $\hat{\mathbf{w}}$ to make predictions, using $\hat{\mathbf{u}} = \boldsymbol{\Lambda}\hat{\mathbf{w}}$ (recall that these are on the log scale), at all 100 prediction locations for each simulated data set. To estimate prediction bias, we used the average of the elements of $\hat{\mathbf{u}} - \mathbf{u}$ for each simulated data set, where $\mathbf{u}$ contains the simulated values at the 100 prediction locations, and then averaged those across the 2000 simulated data sets. We also formed 90% prediction intervals as $\hat{u}_k \pm 1.645\widehat{\text{se}}(\hat{u}_k)$ $(k = 1, \ldots, 100)$, where, for `spmodel`, $\widehat{\text{se}}(\hat{u}_k)$ was the square root of the kth diagonal element of (15). We computed the proportion of times, over the 100 predictions and 2000 simulations, that the prediction intervals contained the true values, which should be about 90%.

The results are shown in Table 3. There was very little estimation or prediction bias in any of the three methods, so those results are not shown. For small sample sizes, MSE, CI90, MSPE, and PI90 were very similar for all three methods, with a very small advantage to `spmodel`, with appropriate coverage for both confidence and prediction intervals. However, estimation and prediction times for `spBayes` were considerably higher. In fact, computation times for `spBayes` for sample size of 525 were prohibitive and were not included in the rest of the simulation study. For sample sizes of 525, again `spmodel` and `glmmTMB` provided similar results for MSE, CI90, MSPE, and PI90, however `spmodel` was much faster – several orders of magnitude faster for prediction. None of these results are surprising, as all three methods are fitting the same underlying model, but where `spBayes` uses MCMC to marginalize over all effects, which is quite slow, `glmmTMB` uses numerical differentiation for the Laplace approximation, which is slow for larger sample sizes, while we use analytical derivatives for the Laplace approximation, implemented in `spmodel`, which is the fastest method.

TABLE 3 MS(P)E and coverage for estimation of fixed effect β_1 and for prediction of $\mathbf{u}$ at unobserved locations.

	Geostatistical model, sample size = 125					
Package	**MSE**	**CI90**	**MSPE**	**PI90**	**time$_f$**	**time$_p$**
spmodel	0.0145	0.908	0.178	0.895	1.095	0.140
spBayes	0.0151	0.901	0.179	0.907	66.230	7.819
glmmTMB[a]	0.0149	0.906	0.184	0.891	0.926	2.335
	Geostatistical model, sample size = 525					
spmodel	0.00273	0.905	0.0921	0.909	6.458	0.322
glmmTMB	0.00276	0.900	0.0930	0.901	44.096	65.538
	CAR model, sample size = 125					
spmodel	0.0221	0.894	0.499	**0.879**	1.785	0.154
INLA	0.0245	**0.879**	0.522	**0.829**	1.058	
	CAR model, sample size = 525					
spmodel	0.00427	**0.873**	0.390	0.890	11.292	0.726
INLA[b]	0.00430	**0.882**	0.384	0.895	1.515	

Note: MSE is mean-squared error for β_1, CI90 is coverage for 90% confidence intervals for β_1, MSPE is mean-squared prediction error for 100 prediction locations, PI90 is coverage of 90% prediction intervals for 100 prediction locations. Values in bold are outside of the 99% confidence interval for 2000 independent Bernoulli trials with true value 0.9. The columns time$_f$ and time$_p$ are computing times (in seconds) for estimation and prediction, respectively.
[a]17/2000 failed to converge.
[b]13/2000 failed to converge.

The methods described in this manuscript are valid for any patterned covariance matrix, so we also simulated data from a conditional autoregressive (CAR) model (Besag, 1974; Ver Hoef et al., 2018). A popular R package for analyzing these data using a nested first-order Laplace approximation (Rue et al., 2009) is INLA (Rue et al., 2017). INLA uses models with sparse precision matrices such as CAR and first-order autoregressive (AR1) models for fast marginal computations such as those presented in this article. We simulated data on a 15×15 grid, and on a 25×25 grid, using a CAR model with first-order neighborhood weights (rook move), row-standardization, and a spatial dependence parameter of 0.999, resulting a covariance matrix $\mathbf{V} = (\text{diag}(\mathbf{W1}) - 0.999\mathbf{W})^{-1}$, where $\mathbf{W}$ is a binary matrix where a one indicates a neighbor and $\mathbf{1}$ is a vector of all ones. Note that while 0.999 may seem like very strong autocorrelation, it results in nearest neighbors having a correlation of approximately 0.85 (for the 15×15 grid), while those farthest apart in the grid have a correlation of approximately 0.40 (correlations are not stationary in CAR models). The same fixed effects were added as for the geostatistical simulations. For both sample sizes, the simulated values for 100 spatial locations were removed at random for prediction.

The results of spmodel and INLA are similar (Table 3). For smaller sample sizes, spmodel has lower MSE and MSPE, and better confidence and prediction interval coverage than INLA. However, for sample size 525, results are very similar, and INLA is very fast. Note that the methods in this article make no special use of the sparse precision matrix and certain fast computations used by INLA, so our methods can be adopted for simultaneous autoregressive models, spatial and temporal moving average models, and others not considered by INLA.

3.3 | A simulation experiment for other distributions

In addition to the Poisson distribution, we did similar simulations for all five of the other distributions in Table 1, and we make no further comparisons to other software. No ϕ was needed for the binomial (Bernoulli) distribution, but for the beta distribution we set $\phi = 10$, and for the negative binomial, gamma, and inverse Gaussian distributions we set $\phi = 1$ (parameterizations of the distributions are given in the Appendix). Estimation and prediction appeared to be unbiased in all five cases, so we only show the corrected 90% confidence interval coverage in Table 4. In all but one case, the intervals have close to 90% confidence and prediction interval coverage, the exception being the beta distribution, which undercovers slightly, especially for prediction. Note that we included a nugget effect when fitting all of these models. The

TABLE 4 Interval coverage for estimation of fixed effects $\boldsymbol{\beta}$ and for prediction of $\mathbf{u}$ at unobserved locations for the five distributions in Table 1 that were not included in Table 3; binomial (bino), beta, negative binomial (nbin), gamma (gamm) and inverse Gaussian (iGau).

Effect	bino	beta	nbin	gamm	iGau
$\hat{\beta}_0$	**0.718**	0.894	**0.750**	**0.726**	**0.645**
$\hat{\beta}_1$	0.906	**0.831**	0.911	0.897	0.895
$\hat{\beta}_2$	0.915	**0.843**	0.899	0.892	0.910
$\hat{\beta}_3$	0.912	**0.829**	0.907	0.891	0.897
$\hat{\mathbf{u}}$	0.891	**0.773**	0.905	0.909	**0.946**

Note: Coverage is for 90% confidence and prediction intervals, using the corrected versions in (12) and (15). Values in bold are outside of the 99% confidence interval for 2000 independent Bernoulli trials with true value 0.9.

prediction interval overcoverage (0.946) when using the inverse Gaussian distribution becomes 0.904 without a nugget effect. We address the nugget issue more in the Discussion and Conclusions section. More simulations will be necessary to fully characterize when the intervals corresponding to these distributions have shortcomings.

4 | EXAMPLES

We demonstrate the methods with three example data sets that use all of the distributions in Table 1, combined with covariance matrices developed through spatial autoregressive models, time series models, geostatistical models, and variance components models that include random effects. All data, code and results are available as an `R` package mentioned at the end of this article.

4.1 | 1980 presidential turnout in Texas

This data set contains the proportion of the population over age 19 that cast votes in the 1980 presidential election in the United States. The proportions are for each of the 254 counties in Texas. The data for the whole of the United States were collected and reported in Pace and Barry (1997), and are available in the `R` package `spData` (Bivand et al., 2023). We created a subset of the data for Texas only. The response variable is reported as a proportion, but we also created a binary variable by assigning a value of 1 to those proportions greater than 0.5 and assigning a value of zero otherwise. Here, we will fit the binomial distribution (actually a Bernoulli distribution because all sample sizes are one) to the binary response variable, and the beta distribution to the proportional response variable.

There are three explanatory variables in the data set: (1) proportion of population with college degrees, (2) proportion of home ownership, and (3) per capita income, where for all three variables the values are with respect to the total population over age 19 that were eligible to vote. A scatterplot of the logit of the proportional response variable for all three explanatory variables is given in Figure 4. In an attempt to linearize relationships, we cubed the explanatory variable for the proportion of home ownership and took the natural logarithm of per capita income. The linear model that we consider then is,

$$\mathbf{w} = \mathbf{X}\boldsymbol{\beta} + \mathbf{r}_1,$$

where $\mathbf{X}$ contains a column of ones for an overall mean and three columns for the (transformed) explanatory variables.

For the spatial random effects $\mathbf{r}_1$, we fit two spatial autoregressive models to the data for the 254 counties, a conditional autoregressive (CAR) model and a simultaneous autoregressive (SAR) model. These models rely on neighbor definitions, rather than distance directly. We defined a neighbor of a county as any other county whose centroid was within 150 km of the originating county's centroid. Using that definition, some counties had but a single neighbor, while others had many; the maximum was 38. Let $\mathbf{W} = (W_{i,j})$ denote a neighbor incidence matrix, where $W_{i,j} = 1$ if county j is a neighbor of county i, and $W_{i,j} = 0$ otherwise, where the diagonal is all zeros (a site is not a neighbor of itself). Let $\mathbf{W}_{rs}$ be a "row-standardized" version of $\mathbf{W}$, where the elements in any row of $\mathbf{W}$ are divided by their row sum, $W_{i,+} = \sum_j W_{i,j}$. Then the covariance

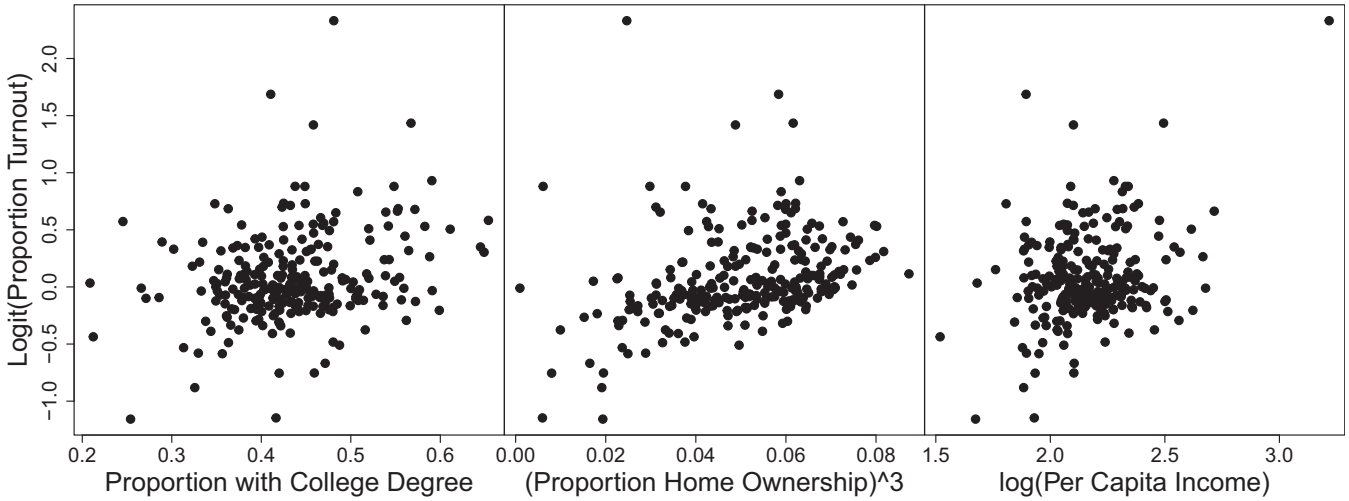

FIGURE 4 Scatterplot of the logit of voter-turnout response variable by the three explanatory variables. Note the transformations of some explanatory variables, where proportion of home ownership was cubed, and natural logs were taken of per capita income.

matrix for a CAR model is

$$\boldsymbol{\Sigma}_\theta = \sigma^2(\mathbf{I} - \rho\mathbf{W}_{rs})^{-1}\mathbf{M}_{rs},$$

where $\mathbf{M}_{rs}$ is a diagonal matrix with ith diagonal element $1/W_{i,+}$. The SAR covariance matrix is

$$\boldsymbol{\Sigma}_\theta = \sigma^2[(\mathbf{I} - \rho\mathbf{W}_{rs})(\mathbf{I} - \rho\mathbf{W}'_{rs})]^{-1}.$$

In both covariance matrices, ρ is a spatial dependence parameter whose maximum value is one (because of row-standardization), and we considered the parameter space $0 \leq \rho \leq 1$ and $\sigma^2 > 0$.

For the model with the binary response variable, the three explanatory variables, and the CAR covariance matrix, using (4) with REML we estimated $\hat{\sigma}^2 = 31.59$ and $\hat{\rho} = 0.9999$. Note that $\hat{\rho}$ is very close to the boundary. When $\rho = 1$ is fixed at the boundary, then the covariance matrix no longer exists, but it is possible to use an intrinsic conditional autoregressive model (ICAR, Besag et al., 1991). The results should be very similar. The ICAR model eliminates the need to estimate ρ. At the cost of a parameter, estimating ρ allows the models to remain in our HGLMM framework. Changing the model to a SAR covariance matrix, we estimated $\hat{\sigma}^2 = 0.681$ and $\hat{\rho} = 0.956$. The minimized value of the minus twice the log-likelihood in (4) was 747.51 for the CAR model, while it was 738.89 for the SAR model, indicating that the SAR model was a better choice. The number of parameters are the same for both models, and here we use the log-likelihood (4) to evaluate model performance in the same way as Rue et al. (2017). Model comparisons based on the loglikelihood, such as likelihood ratio tests, or model selection such as AIC (Akaike, 1973), are possible. Table 5 gives fixed effects estimates. Note the large difference in standard errors using the naive approach based on $\mathbf{C}_{\hat{\beta}}$ and the corrected version given by (12).

For the proportion turnout response variable, the beta distribution in Table 1 was used, and for the CAR covariance matrix we estimated $\hat{\sigma}^2 = 0.850$, $\hat{\rho} = 0.9992$, and $\hat{\phi} = 72.9$, while for the SAR covariance matrix we estimated $\hat{\sigma}^2 = 0.0125$, $\hat{\rho} = 0.942$, and $\hat{\phi} = 46.6$. The minimized value of minus twice the log-likelihood in (4) was -101.04 for the CAR model, while it was -98.32 for the SAR model, indicating the CAR model was a slightly better choice, in contrast to the binary data. Table 6 gives fixed effects estimates. As was the case for the binary data, there is a large difference in standard errors using the naive approach based on $\mathbf{C}_{\hat{\beta}}$ and the corrected version given by (12). The overall patterns of coefficient estimates and their standard errors are similar between SAR and CAR models in both Tables 5 and 6. In comparing Table 5 to Table 6, there appears to be more precision in the estimates in Table 6, especially regarding the significance of the per capita income variable. This is not surprising because the transformation of the proportional turnout data into binary data invariably results in a loss of information.

The predicted $\hat{\mathbf{w}}$ values are shown in Figure 5. A spatial visualization of the binary data shows some apparent clustering of 1's and 0's (Figure 5a). The predicted $\hat{\mathbf{w}}$ values for the binary data using a SAR model have the highest values in

TABLE 5 Estimated fixed effects table for Texas turnout data using binary response variable.

	SAR model					CAR model	
Effect	**Est.**	**s.e.$_u$**	**s.e.$_c$**	***z*-val.**	***p*-val.**	**Est.**	**s.e.$_c$**
Intercept	−4.747	0.775	2.302	−2.062	0.0392	−4.402	6.791
College	5.083	1.136	3.717	1.367	0.1715	4.681	3.817
Home-owner	70.243	3.572	12.911	5.440	<0.0001	69.960	13.365
Income	−0.688	0.421	1.377	−0.500	0.6173	−0.611	1.443

Note: The estimates are given by Est., while s.e.$_u$ is the naive standard error using only $\mathbf{C}_\beta$ from Section 2.3 and s.e.$_c$ is the corrected standard error using (12). The *z*-val. is the estimate divided by the corrected standard error, and the *p*-val. is the probability of exceeding the modulus of the *z*-value if the effect were truly zero assuming a standard normal distribution for the *z*-value (infinite sample size).

TABLE 6 Estimated fixed effects table for Texas turnout data using proportional response variable.

	CAR model					SAR model	
Effect	**Est.**	**s.e.$_u$**	**s.e.$_c$**	***z*-val.**	***p*-val.**	**Est.**	**s.e.$_c$**
Intercept	−1.384	0.457	0.496	−2.789	0.0053	−1.613	0.253
College	0.619	0.283	0.416	1.488	0.1368	0.407	0.407
Home-owner	8.609	0.801	1.272	6.770	<0.0001	8.711	1.301
Income	0.340	0.112	0.157	2.168	0.0302	0.470	0.142

Note: The headings are the same as for Table 5.

the northern Texas "pan-handle" and in central Texas (Figure 5b), and the pattern is similar for the predicted $\hat{\mathbf{w}}$ values for the binary data when using a CAR model (Figure 5c). A logit transformation of the raw proportional turnout data are shown in Figure 5d and the predicted $\hat{\mathbf{w}}$ values for the SAR model (Figure 5e) appear to smooth the raw data, with a similar spatial pattern to the binary data (Figure 5b) and to the predicted $\hat{\mathbf{w}}$ values using a CAR model (Figure 5f).

4.2 | Harbor seal counts in Alaska

For over 30 years, aerial surveys of harbor seals throughout Alaska have been flown by the Marine Mammal Laboratory of the Alaska Fisheries Science Center, part of the US government NOAA Fisheries. These surveys, which are performed primarily during the late summer months when harbor seals are molting, are the primary method for monitoring and estimating the abundance of harbor seals (Muto et al., 2022). Based on genetic sampling, all seals in Alaska have been divided into 12 different "stocks," or genetic populations. Abundance estimates are created for each stock, and here we will use the stock known as the Sitka/Chatham Strait population. This data set consists of 716 observations in the years 1998, 2003, 2008–2011, and 2015. To regulate body temperature, seals often haul out of the water, which is how they are counted more easily than when they are in the water. All known harbor seal haul-out sites for this population were collected into 74 sample polygons. Some sample polygons were counted multiple times per year, while others were skipped in some years. For each aerial count, explanatory variables included time-from-low-tide and time-of-day.

We use Poisson and negative binomial models in Table 1 to formulate hierarchical GLMMs for the seal count data. We consider a specific case of the model in (1):

$$\mathbf{w} = \mathbf{X}\boldsymbol{\beta} + \mathbf{r}_1 + \mathbf{Z}_2\mathbf{r}_2 + \boldsymbol{\epsilon},$$

where $\mathbf{X}$ contains a column for an overall mean; 73 columns with indicators of a mean effect for each sample polygon (as deviations from the mean for the first polygon, which was absorbed into the overall mean); a column for time-of-day, which is the elapsed fraction of the day from solar noon (the time when the sun is at the zenith); a column for time-from-low-tide, which is in hours from low tide (tide cycles last about 12 h in this area); and columns for squared time-of-day and squared time-from-low-tide. Thus, $\mathbf{X}$ has 78 columns. The $n \times 1$ random effect $\mathbf{r}_1$ is assumed to have the

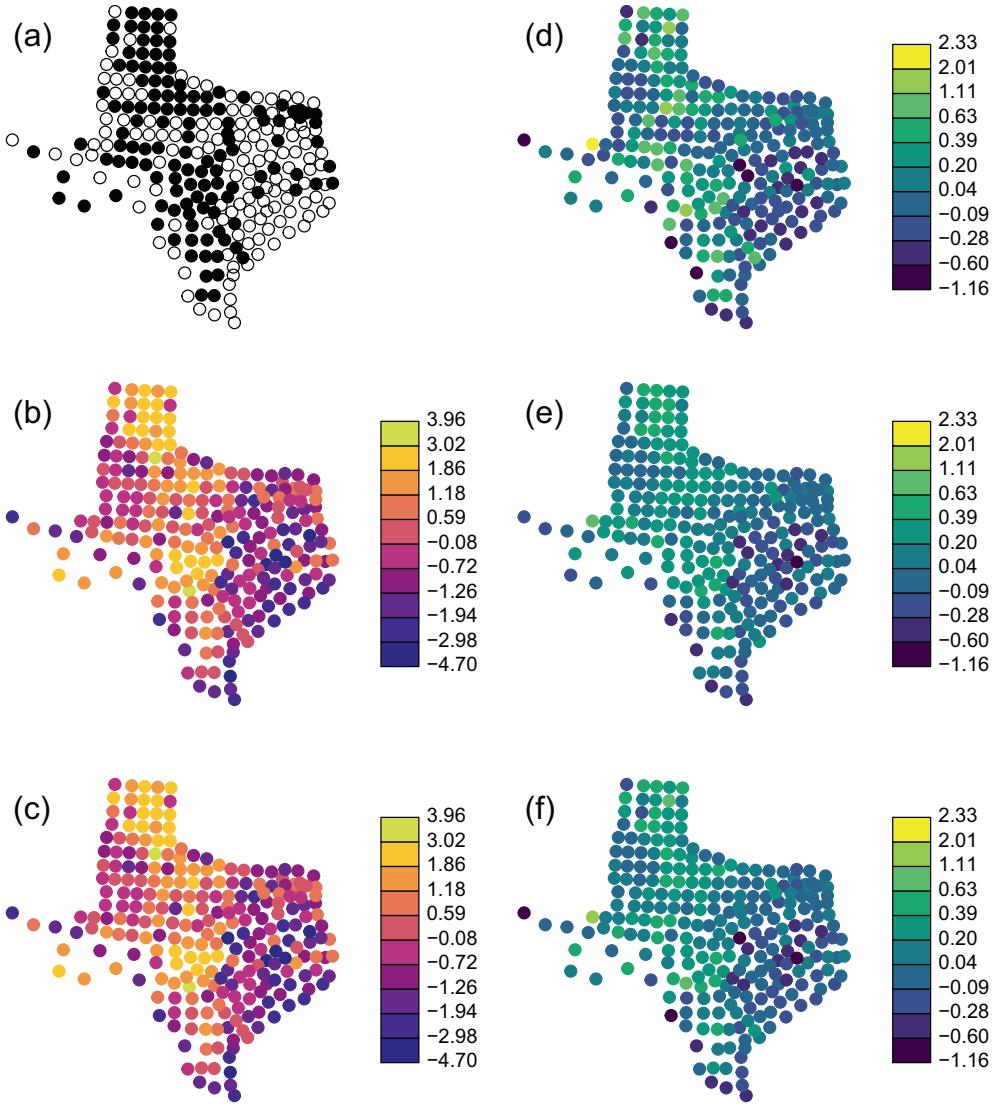

FIGURE 5 Raw data and predicted spatial random effects ($\mathbf{w}$) for the Texas turnout data. (a) Raw binary data, where open circles are zeros and solid circles are ones, (b) predicted $\hat{\mathbf{w}}$ using SAR model for binary data, (c) predicted $\hat{\mathbf{w}}$ using CAR model for binary data, (d) logit-transformed proportional turnout data, (e) predicted $\hat{\mathbf{w}}$ using SAR model for proportional turnout data, (f) predicted $\hat{\mathbf{w}}$ using CAR model for proportional turnout data.

covariance structure of a stationary first-order autoregressive (AR1) time series model, so $\text{cov}(r_{1,i}, r_{1,j}) = \sigma_1^2 \rho^{|t_i - t_j|}/(1-\rho^2)$, where t_i and t_j are the years of the counts for the ith and jth observations. Moreover, we assume that counts from distinct polygons are independent of each other, so the autocovariance only occurs among years for a given polygon, yielding a covariance matrix with a block diagonal structure. However, for some polygons there are repeated measures within year, so $\mathbf{Z}_2$ is a design matrix created as an interaction between polygon and year, and $\sigma_2^2 \mathbf{Z}_2 \mathbf{Z}_2'$ allows for additional correlation among repeated samples per year within a polygon. Not all years had repeated measures, and $\mathbf{Z}_2$ had 518 columns and we denote σ_2^2 as the variance of $\mathbf{r}_2$. When using a model with a Poisson distribution, we allow for further uncorrelated overdispersion by using $\boldsymbol{\epsilon}$ as given in (1), where $\text{var}(\boldsymbol{\epsilon}) = \sigma_0^2$, but for the negative binomial distribution, which directly allows for overdispersion, we set $\boldsymbol{\epsilon} = \mathbf{0}$.

For the Poisson model with AR1 covariance structure, using (4) with REML we estimated $\hat{\sigma}_0^2 = 0.858$, $\hat{\sigma}_1^2 = 0.648$, $\hat{\sigma}_2^2 = 1.48 \times 10^{-6}$ and $\hat{\rho} = 0.938$, while for the negative binomial distribution with the AR1 covariance model, we estimated $\hat{\sigma}_1^2 = 14.92$, $\hat{\sigma}_2^2 = 0.0057$, $\hat{\rho} = 0.9992$, and $\hat{\phi} = 1.529$. The minimized value of minus twice the loglikelihood in (4) was 9607.4 for the Poisson model, while it was 9437.6 for the negative binomial model, with an equal number of parameters, indicating that the negative binomial was a better choice. Table 7 gives fixed effects estimates for the negative binomial

model, except that we include the naive and corrected variances for the Poisson model as well. It is especially interesting that when ϵ was included in the $\mathbf{w}$ values for the Poisson model, the diagonal elements of $\mathbf{C}_\beta$ were large and increased only slightly when adjusted by $\mathbf{B}(-\mathbf{H}_\mathbf{a}^{-1})\mathbf{B}'$ as given in (12). This is in contrast to the negative binomial model, whose diagonal elements of $\mathbf{C}_\beta$, as reflected by s.e.$_u$ in Table 7, are very small. Yet, s.e.$_c$ is very similar for the negative binomial and Poisson models, despite their apparently different covariance structures.

The fitted explanatory variables allow insight into seal behavior, indicating that seals prefer to haul out of the water around midday and at low tides. The model confirms that counts are highest at these times, as the coefficients in Table 7 are plotted in Figure 6. This shows changes on the log scale, where all other explanatory variables are held fixed at zero, so Figure 6 may be interpreted as the log of the proportional change in expected counts with unit change in the explanatory variable. Note that we remain in the linear modeling framework on the link scale. It would be possible to extend our models to generalized additive models on the link scale, for example, Ruppert et al. (2003), but we make no attempt to do so here.

Our ultimate goal is to predict the abundance of seals at each site, using the results in Section 2.4. Predicted w-values, after exponentiating, are shown in Figure 7 for 4 of the 74 different sites, which are labeled AC10, AC11, BC01, and BC02. For each year, we predicted $\hat{\mathbf{u}}$ in (13) with 90% prediction intervals using (15), and then exponentiated both predictions and prediction intervals. The errors are large, but this is not unreasonable given the relatively few observations per site. For this model, we borrowed strength across sites for estimating the autocorrelation parameter ρ, assuming all sites had the same amount of autocorrelation among $\mathbf{w}$. In general, the predictions tend to shrink toward the overall mean for the site, but for site BC02 especially, the predictions are greater than the observed values. This can be explained because the

TABLE 7 Estimated fixed effects and associated quantities, under the negative binomial and Poisson models, for the harbor seal count data.

	Negative binomial					Poisson	
Effect	**Est.**	**s.e.$_u$**	**s.e.$_c$**	***z*-val.**	***p*-val.**	**s.e.$_u$**	**s.e.$_c$**
time-from-low-tide	−0.061	0.0019	0.036	−1.705	0.0882	0.039	0.041
(time-from-low-tide)^2	−0.057	0.0012	0.020	−2.806	0.0050	0.024	0.024
hour-of-day	−0.234	0.0049	0.084	−2.787	0.0053	0.097	0.100
(hour-of-day)^2	−0.581	0.0066	0.114	−5.104	0.0000	0.130	0.134

Note: The headings are the same as for Table 5.

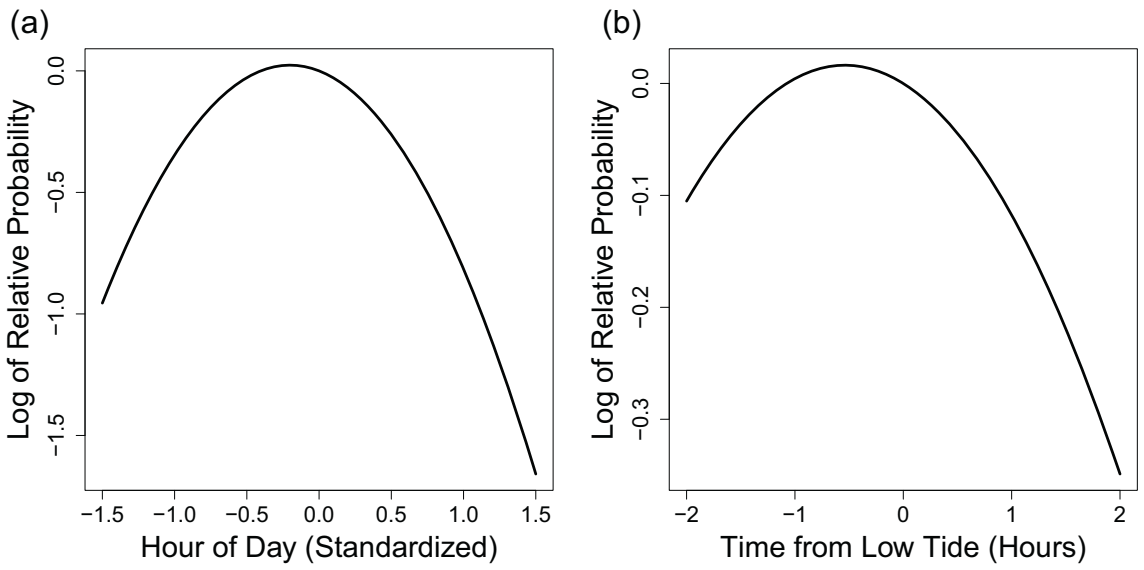

FIGURE 6 Fitted effects of (a) hour-of-day and (b) time-from-low-tide on harbor seal counts. The fitted effect shows the log of the expected proportional change when all other covariates are held at zero.

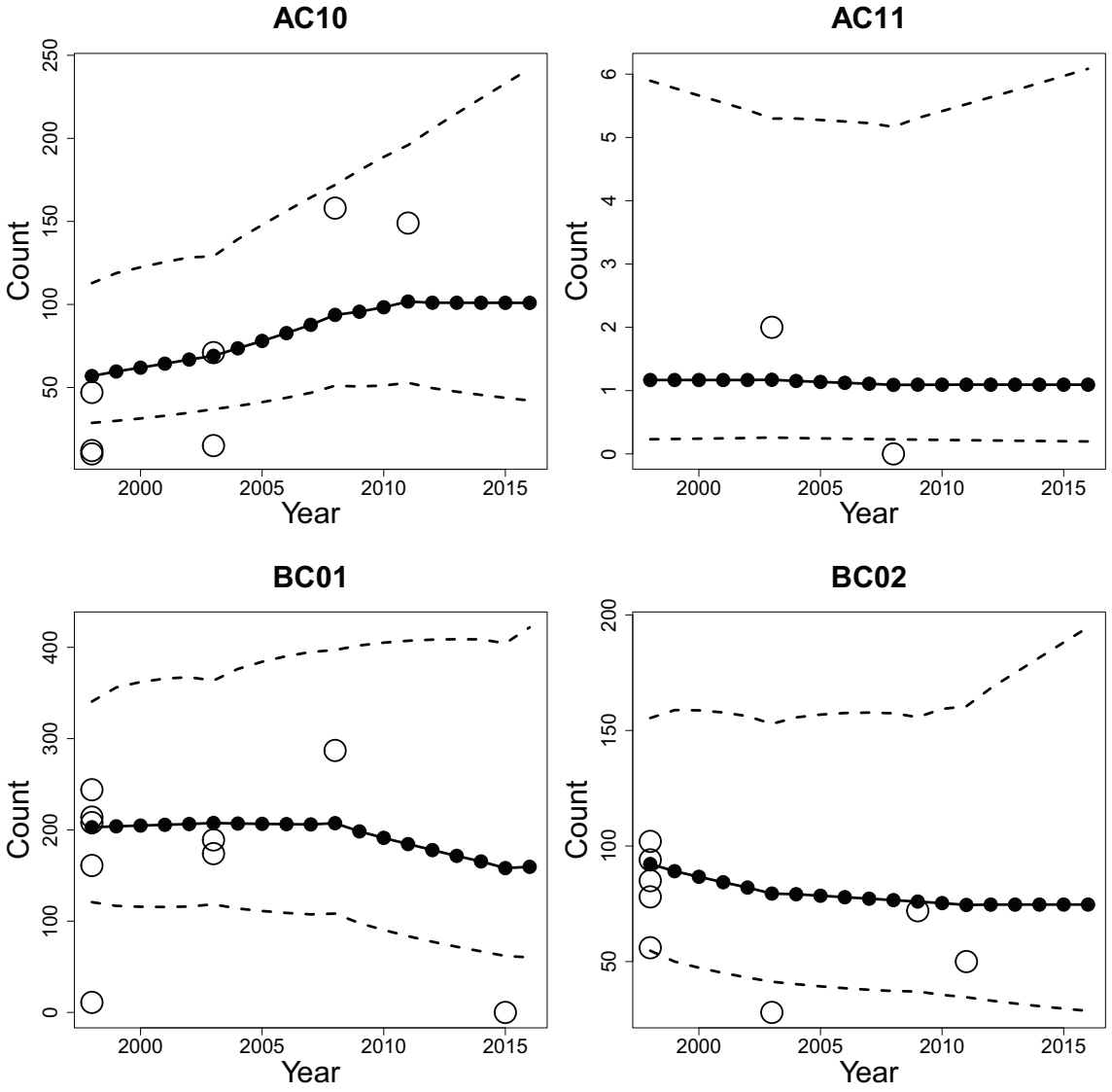

FIGURE 7 Predicted *w*-values for 4 of the 74 sites. Open circles are raw counts, and solid circles are predicted *w*-values connected by a solid line. Note that predictions are made for optimal sighting conditions so can be expected to be, on average, greater than the count values. The dashed line shows the prediction intervals.

predictions are standardized to optimal conditions for the explanatory variables (time-of-day and time-to-low-tide). Site BC02 was counted in suboptimal conditions on almost all occasions, which is entirely possible because a high tide may occur at solar noon. It is impossible to optimize explanatory variables through a sampling design without being willing to wait and sample only when a low tide occurs at near solar noon.

4.3 | Heavy metal concentrations in Moss

Cape Krusenstern National Park is in northwest Alaska, USA, and nearby is the Red Dog mine, where zinc, lead, cadmium and other heavy metals are mined. Trucks haul ore to the coast from the Red Dog Mine on a road that traverses Cape Krusenstern National Park. There is speculation that dust escapes into the environment from those trucks. Mosses obtain much of their nutrients from the air, so they are ideal biomonitors for heavy metals attached to airborne dust. In 2001 (Hasselbach et al., 2005) and again in 2006 (Neitlich et al., 2017), mosses were sampled for heavy metals, with the sampling being more dense near the road. Current annual growth of moss tissue was sampled, ground, homogenized,

and then sent for laboratory analysis. Here, we just consider lead concentrations, although many other elements were analyzed. Potentially important explanatory variables that we include are distance-from-haul-road, side-of-the-road (north or south), and year of sample. There are 365 records in the data set, with 244 from 2001 and 121 from 2006.

Lead concentrations are inherently positive and are often skewed, which led Hasselbach et al. (2005) and Neitlich et al. (2017) to transform the response to the log scale. Instead, here we use the gamma and inverse Gaussian models given in Table 1. For the covariance structure, we consider a special case of the linear model (1),

$$\mathbf{w} = \mathbf{X}\boldsymbol{\beta} + \mathbf{r}_1 + \mathbf{Z}_2\mathbf{r}_2 + \mathbf{Z}_3\mathbf{r}_3 + \boldsymbol{\epsilon},$$

where $\mathbf{X}$ contains a column for an overall mean, an indicator column for year 2006 (2001 is absorbed into the overall mean), log of distance to road (in meters), and an indicator for south of the road (north is absorbed into the overall mean). We also consider an interaction between distance to road and the side of the road.

The random effect $\mathbf{r}_1$ is assumed to have a geostatistical autocovariance structure given specifically by the exponential model, $\text{cov}(r_1(\mathbf{s}_i), r_1(\mathbf{s}_j)) = \sigma_1^2 \exp(-\delta_{i,j}/\rho)$, where $\mathbf{s}_i$ is a vector containing the spatial coordinates of the ith location and $\delta_{i,j}$ is the Euclidean distance between locations $\mathbf{s}_i$ and $\mathbf{s}_j$. The parameter σ_1^2 is often called the partial sill, and ρ is the range parameter, which controls the distance-decay rate of the autocovariance with distance. The variance of $\boldsymbol{\epsilon}$, σ_0^2, is often called the nugget effect. We assume that responses from different years are independent. Within year and location, at some sites, duplicate samples were obtained to account for microscale variation; that is, one handful of moss was grabbed and then another (the distance between grabs was assumed to be zero). Hence, $\mathbf{Z}_2$ is a design matrix with indicator variables for location; this causes increased autocorrelation for any samples from the same location. Some samples were ground into two replicate samples for laboratory analysis, as there can be some variation in the machines that measure concentration or the way it is homogenized. Therefore, $\mathbf{Z}_3$ is a design matrix that contains indicator variables for a duplicate nested within location; this causes repeated replicates to have higher autocorrelation due to coming from a common duplicate.

Using the gamma distribution with the exponential covariance model and maximum likelihood (rather than REML), minus twice the log-likelihood was equal to 2989.128, and, from the fixed effects table, it appeared that the main effect for side-of-road was not significant. We used ML rather than REML because REML does not provide nested models for likelihood comparisons when fixed effects are changing (Verbeke & Molenberghs, 2000, p. 75). We refit the model without that main effect, and minus twice the log-likelihood from (4) was 2990.334. Using either AIC or a likelihood ratio test, we have evidence to drop the main effect for side-of-road from the model. After doing so, the marginal estimates of the covariance parameters are $\hat{\sigma}_1^2 = 0.1703$, $\hat{\sigma}_2^2 = 0.0635$, $\hat{\sigma}_3^2 = 0.0267$, $\hat{\sigma}_0^2 = 0.000023$, $\hat{\rho} = 9.075$, and $\hat{\phi} = 367.9$. We fit an inverse Gaussian model with the same mean and covariance structure, for which minus twice the log-likelihood was 2989.99, which is almost identical to the value for the gamma model. The covariance parameters for the inverse Gaussian model were estimated to be $\hat{\sigma}_1^2 = 0.1704$, $\hat{\sigma}_2^2 = 0.0633$, $\hat{\sigma}_3^2 = 0.0267$, $\hat{\sigma}_0^2 = 0.000561$, $\hat{\rho} = 9.064$, and $\hat{\phi} = 446.4$, which are almost identical to their counterparts for the gamma model. Interestingly, if we fit a normal spatial model to the log concentrations we obtain estimates $\hat{\sigma}_1^2 = 0.1702$, $\hat{\sigma}_2^2 = 0.0634$, $\hat{\sigma}_3^2 = 0.0267$, $\hat{\sigma}_0^2 = 0.0028$, and $\hat{\rho} = 9.057$, which are almost identical to estimates for both the gamma and inverse Gaussian models.

Estimates of the fixed effects are almost identical for all three models, so only those for the gamma model are given (Table 8). Note that for this example, in contrast to the previous two examples, there is little difference in the standard errors of the estimated fixed effects based on s.e.$_u$ and s.e.$_c$. This can happen when essentially all of the variation is captured by $\mathbf{w}$ and the contribution of $\log[\mathbf{y}|\mathbf{g}^{-1}(\mathbf{w}), \boldsymbol{\phi}]$ is small in comparison. This contribution is controlled in part by ϕ for the gamma and inverse Gaussian distributions, whose estimated values are very large, and the variance of the distribution is inversely proportional to ϕ (see the Appendix).

TABLE 8 Estimated fixed effects for the moss lead data when using the gamma distribution.

Effect	Est.	s.e.$_u$	s.e.$_c$	z-val.	p-val.
Intercept	8.062	0.2031	0.2043	39.47	<0.0001
Year	−0.441	0.2233	0.2236	−1.97	0.0488
Distance to Road	−0.577	0.0184	0.0186	−31.07	<0.0001
Distance-to-road:Southside	−0.112	0.0117	0.0117	−9.51	<0.0001

Note: The headings are the same as for Table 5, where here, for the p-val, we used a t-distribution with $365 - 4 = 361$ degrees of freedom.

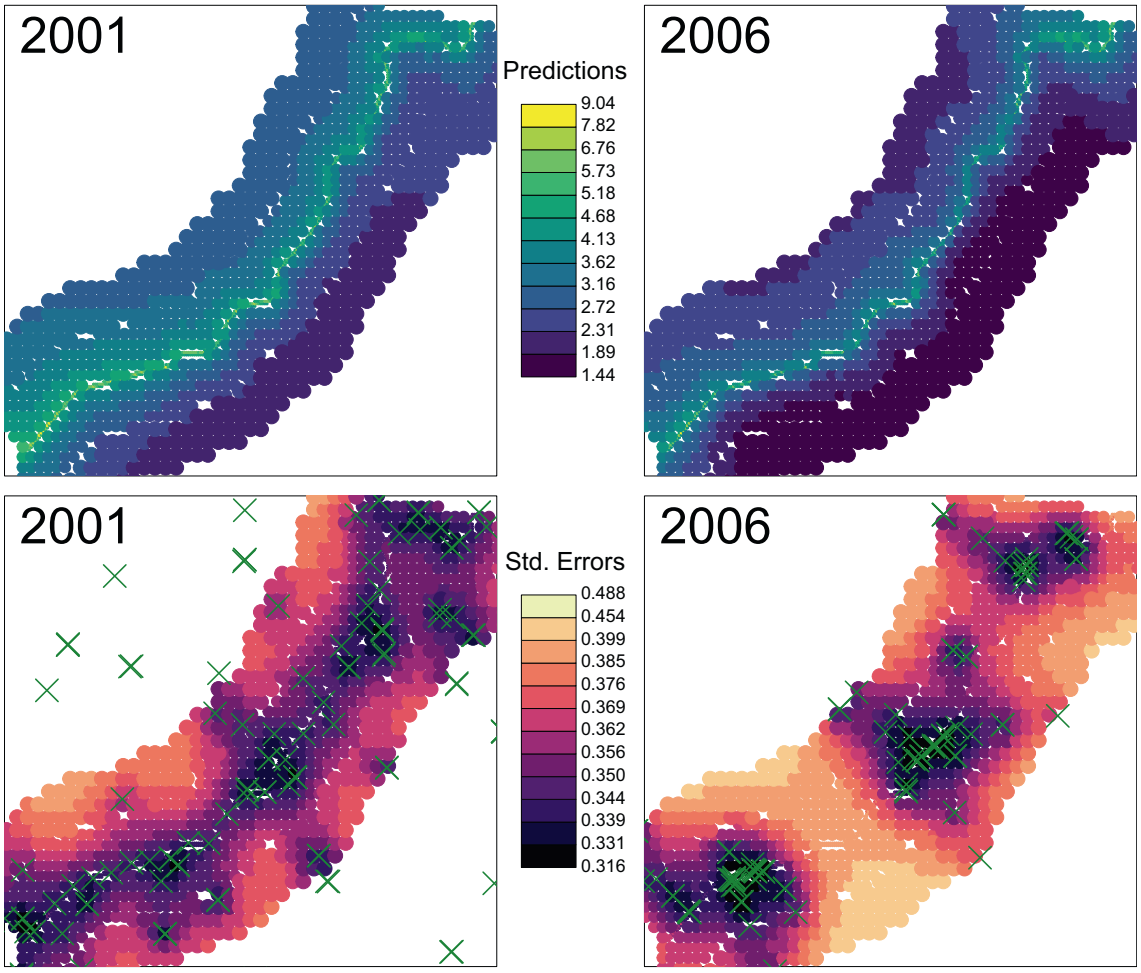

FIGURE 8 Prediction and their standard errors for 2001 and 2006 at locations near the haul road through Cape Krusenstern National Park, Alaska. The green × symbols show sample locations.

Predictions of the w-values are similar for both the gamma and inverse Gaussian distribution models, and they are both very similar to a normal model for the log concentrations. Predictions of $\mathbf{w}$ by years 2001 and 2006 under the gamma model are shown in Figure 8. The prediction locations were divided into three groups, one which was closely spaced near the haul road, with each successive group having more coarsely spaced locations as they get farther from the road. It is clear that predicted values are largest near the road (Figure 8), but also that the predicted values generally decreased from 2001 to 2006, likely due to a change from 2001 to 2006 where coverings were used on the trucks hauling ore on the road. The prevailing winds are from the south, and it is clear that predicted values are higher on the north side of the road. The prediction standard errors show the typical pattern in geostatistics, i.e., they are smaller near a sample or in dense concentrations of samples (Figure 8) than farther away.

5 | DISCUSSION AND CONCLUSIONS

We have developed a very flexible framework for modeling binary, count, positive continuous, and other data types in a hierarchical generalized linear mixed model framework. Virtually any data type can be accommodated by the many distributions that are known in statistics, and these distributions can be matched to virtually any patterned covariance matrix, where a short list is given in Table 1. Our examples illustrate all of the distributions in Table 1, and, for covariance matrices, we used CAR and SAR spatial autoregressive models, AR1 time series models, and exponential geostatistical models. We also showed how complex covariance matrices can be created by mixing random effects with other covariance structures. Any covariance matrix is possible in the HGLMM framework, including spatio-temporal, covariances for data

on a sphere, covariances derived for linear networks such as streams and roads, etc. Using the Laplace approximation, the resulting log-likelihood is composed of the log-likelihood of the data distribution, the ML or REML log-likelihood for normally-distributed data, and the determinant of a Hessian matrix (4).

We have developed marginal inference for three of the most common objectives in linear models. First, in order to estimate fixed effects and make predictions, we must estimate all covariance parameters, which is accomplished from (4). Then, it is necessary to adjust variances for the fact that $\mathbf{w}$ is latent in the model, and not observed, which is accomplished from (12) when estimating fixed effects and from (15) when predicting at unsampled locations.

The models can be computationally demanding, as they require computing the determinant of the Hessian matrix and, in our implementation, its inverse as well. Optimizing the likelihood is doubly iterative as Newton–Raphson updates are used during likelihood optimization for covariance parameters, requiring $\mathbf{H}^{-1}$ for each update. While this may limit the size of data sets for our HGLMM framework, we would like to point out some time-saving features. First, note from (10) that

$$\mathbf{H} = \left[\mathbf{D}_{\boldsymbol{\phi}} - \boldsymbol{\Sigma}_{\theta}^{-1}\right] + \left[\boldsymbol{\Sigma}_{\theta}^{-1}\mathbf{X}\right]\left(\mathbf{X}'\boldsymbol{\Sigma}_{\theta}^{-1}\mathbf{X}\right)^{-1}\left[\mathbf{X}'\boldsymbol{\Sigma}_{\theta}^{-1}\right],$$

where we add brackets to show that this has Sherman-Morrison-Woodbury form $\mathbf{A} + \mathbf{BCB}'$ (Sherman & Morrison, 1949; Woodbury, 1950). If $\mathbf{A}$ is $n \times n$ but has a fast inverse, and $\mathbf{C}$ has small dimension, then the inverse $(\mathbf{A} + \mathbf{BCB}')^{-1}$ can be made much faster than a full $n \times n$ inverse. For example, consider our second example on harbor seals with 716 records at 74 sample sites. We assumed a time series model within site, but independence between sites, giving the covariance matrix a block diagonal structure. Thus, $\boldsymbol{\Sigma}_{\theta}^{-1}$ has a block-wise inverse, and $\mathbf{D}_{\boldsymbol{\phi}}$ is diagonal, so $[\mathbf{D}_{\boldsymbol{\phi}} - \boldsymbol{\Sigma}_{\theta}^{-1}]^{-1}$ can be inverted block-wise, which is much faster than a single inverse for the whole $n \times n$ matrix. Similarly, $\mathbf{Z}_k\mathbf{Z}_k'$ is often block-diagonal, and multiple variance components can use the Sherman-Morrison-Woodbury theorem recursively (Dumelle et al., 2021).

There are many big data approaches to HGLMMs, but in general they require some modification to the full covariance matrix in order to gain computational speed. Hence, they consider only a subset of possible covariance matrices. For example, Zilber and Katzfuss (2021) only consider geostatistical models (Matérn only in their examples), and note that they use pseudo-data and require ordering spatial data, which changes the true spatial covariance matrix. Bradley et al. (2020) use conjugate multivariate priors that are matched to the response distribution, and are not fully general. Our models make no compromise on the covariance matrix, and hence we make no claim that these models will fit fast enough for data sets with many thousands of records.

The methods given in this paper compare favorably with existing methods. We emphasize that, in the space provided, we could not perform extensive simulations with all factors of sample sizes, different covariance structures, different parameter configurations and so forth, to compare against all other methods and software. We chose a couple of simulations and three popular `R` packages. The methods in this paper, as implemented in `spmodel`, were faster than `spBayes` and `glmmTMB` for geostatistical models, and better than `INLA` for CAR models with small sample sizes. Note, however, that we do not claim to be experts with `spBayes`, `glmmTMB` or `INLA`, and the developers of these packages, and users with more extensive experience, might be able to make these packages run faster and produce better results.

An important consideration for these models is the interplay of the independent component $\boldsymbol{\epsilon}$ in (1) and $\boldsymbol{\phi}$ in $[\mathbf{y}|\mathbf{g}^{-1}(\mathbf{w}), \boldsymbol{\phi}]$. The parameters in $\boldsymbol{\phi}$ often control variance, and can be confounded with $\boldsymbol{\epsilon}$. As an extreme example, suppose that $[\mathbf{y}|\mathbf{g}^{-1}(\mathbf{w}), \boldsymbol{\phi}]$ is a normal distribution where $\mathbf{g}^{-1}(\mathbf{w})$ has the identity function for each element and ϕ has but one element – the variance parameter. Then ϕ and σ_0^2 will not be identifiable. More often, $\boldsymbol{\phi}$ controls how variance is related to the mean, but we expect that there still can be some confounding. For any particular data set, this can be investigated through log-likelihood plots of σ_0^2 and ϕ, similar to Figure 2, or with more experience on how these parameters interact for particular models. For example, in our harbor seal example, we included an independent component (nugget effect) for the latent w-values to absorb overdispersion, but for the negative binomial we excluded the nugget effect because we assumed that the dispersion parameter ϕ absorbed overdispersion.

The HGLMM framework in this paper can be contrasted to the mixed model extension of GLMs. The GLM framework is inspired by the regular exponential family of distributions, and these lead to what are called the "canonical" link functions. For example, the canonical link function for the gamma distribution is $-1/\mu$, but it is often changed to $w = g(\mu) = 1/\mu$. However, that implies that $\mu = g^{-1}(w) = 1/w$, but because w can be negative, it is possible for μ to have negative values. In a moment-based modeling framework using pseudo-likelihood with iteratively reweighted least squares, this can be tolerated if the values stay fairly close to the parameter space, and it allows for a wide variety of link functions, which provides a great amount of flexibility. However, in the HGLMM framework, which is fully parametric,

the evaluation of the log-likelihood for $[\mathbf{y}|\mathbf{g}^{-1}(\mathbf{w}), \boldsymbol{\phi}]$ is not possible if $\mathbf{g}^{-1}(\mathbf{w})$ is outside of the parameter space for the mean. For HGLMMs, link and mean functions must be chosen to respect the parameter space.

We have given a broad outline of marginal inference under the HGLMM. There are many topics to explore that were not mentioned. For example, we may want to make inference on predictions where $\mathbf{w}$ is back-transformed as $\mathbf{g}^{-1}(\mathbf{w})$, and where the variability of $\mathbf{y}|\mathbf{w}$ is added. We may also desire inference for further functions of $\mathbf{g}^{-1}(\mathbf{w})$ such as block averages. Likewise, we may want inferences on random effects (best linear unbiased predictions) of $\mathbf{r}_i$ in (1). Like most linear models, we can consider linear combinations of $\boldsymbol{\beta}$, or contrasts of $\boldsymbol{\beta}$ parameters, in making inferences on fitted models, treatment effects, etc. We only covered the basic framework is this paper and there are many further research topics to develop.

ACKNOWLEDGMENTS

The project received financial support from the National Marine Fisheries Service, NOAA and the U.S. Environmental Protection Agency (EPA). The findings and conclusions in the paper are those of the author(s) and do not necessarily represent the views of the reviewers nor the EPA or the National Marine Fisheries Service, NOAA. Any use of trade, product, or firm names does not imply an endorsement by the US Government. NOAA nor the EPA endorse any commercial products, services, or enterprises.

DATA AVAILABILITY STATEMENT

All data, code, and results for the simulations and applications are available as an `R` package in the Github repository, https://github.com/USEPA/mhglmm.laplace.

ORCID

Jay M. Ver Hoef https://orcid.org/0000-0003-4302-6895
Michael Dumelle https://orcid.org/0000-0002-3393-5529
Philip M. Dixon https://orcid.org/0000-0002-1778-0686
Dale L. Zimmerman https://orcid.org/0000-0003-1212-4089
Paul B. Conn https://orcid.org/0000-0002-2801-299X

REFERENCES

Akaike, H. (1973). *Information theory and an extension of the maximum likelihood principle*. In B. Petrov & F. Csaki (Eds.), *Second International Symposium on Information Theory* (pp. 267–281). Akademiai Kiado.

Berliner, L. M. (1996). *Hierarchical Bayesian time series models*. In *Maximum Entropy and Bayesian Methods: Santa Fe, New Mexico, USA, 1995 Proceedings of the Fifteenth International Workshop on Maximum Entropy and Bayesian Methods* (pp. 15–22). Springer.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society, Series B*, *36*, 192–236.

Besag, J., York, J., & Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, *43*, 1–20.

Bivand, R., Nowosad, J., & Lovelace, R. (2023). spData: datasets for spatial analysis. R Package Version 2.2.2.

Bonat, W. H., & Ribeiro, P. J., Jr. (2016). Practical likelihood analysis for spatial generalized linear mixed models. *Environmetrics*, *27*, 83–89.

Bradley, J. R., Holan, S. H., & Wikle, C. K. (2020). Bayesian hierarchical models with conjugatefull-conditional distributions for dependent data From the natural exponential family. *Journal of the American Statistical Association*, *115*, 2037–2052.

Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, *88*, 9–25.

Brooks, M. E., Kristensen, K., Van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Mächler, M., & Bolker, B. M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, *9*, 378.

Chiles, J.-P., & Delfiner, P. (1999). *Geostatistics: Modeling spatial uncertainty*. John Wiley & Sons.

Christensen, O. F. (2004). Monte Carlo maximum likelihood in model-based geostatistics. *Journal of Computational and Graphical Statistics*, *13*, 702–718.

Clayton, D., & Kaldor, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, *43*, 671–681.

Cressie, N., & Wikle, C. K. (2011). *Statistics for Spatio-temporal data*. John Wiley & Sons.

Cressie, N. A. C. (1993). *Statistics for spatial data* (Revised ed.). John Wiley & Sons.

Diggle, P. J., Tawn, J. A., & Moyeed, R. A. (1998). Model-based geostatistics (with discussion). *Journal of the Royal Statistical Society, Series C: Applied Statistics*, *47*, 299–326.

Dumelle, M., Higham, M., & Ver Hoef, J. M. (2023). spmodel: Spatial statistical modeling and prediction in R. *PLOS One*, *18*, e0282524.

Dumelle, M., Ver Hoef, J. M., Fuentes, C., & Gitelman, A. (2021). A linear mixed model formulation for spatio-temporal random processes with computational advances for the product, sum, and product–sum covariance functions. *Spatial Statistics*, *43*, 100510.

Efron, B., & Hinkley, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. *Biometrika*, *65*, 457–483.

Evangelou, E., Zhu, Z., & Smith, R. L. (2011). Estimation and prediction for spatial generalized linear mixed models using high order Laplace approximation. *Journal of Statistical Planning and Inference*, *141*, 3564–3577.

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.

Fisher, R. A. (1934). Two new properties of mathematical likelihood. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, *144*, 285–307.

Gelfand, A. E., & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, *85*, 398–409.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). *Introducing Markov Chain Monte Carlo*. In *Markov Chain Monte Carlo in practice* (pp. 1–19). Chapman and Hall.

Gneiting, T. (2013). Strictly and non-strictly positive definite functions on spheres. *Bernoulli*, *19*, 1327–1349.

Goldberger, A. S. (1962). Best linear unbiased prediction in the generalized linear regression model. *Journal of the American Statistical Association*, *57*, 369–375.

Gotway, C. A., & Stroup, W. W. (1997). A generalized linear model approach to spatial data analysis and prediction. *Journal of Agricultural, Biological, and Environmental Statistics*, *2*, 157–178.

Haining, R. P. (1978). The moving average model for spatial interaction. *Transactions of the Institute of British Geographers*, *3*, 202–225.

Hamilton, J. D. (1994). *Time series analysis* (Vol. *2*). Princeton University Press.

Harville, D. A. (1974). Bayesian inference for variance components using only error contrasts. *Biometrika*, *61*, 383–385.

Hasselbach, L., Ver Hoef, J. M., Ford, J., Neitlich, P., Crecelius, E., Berryman, S., Wolk, B., & Bohle, T. (2005). Spatial patterns of cadmium and lead deposition on and adjacent to National Park Service lands in the vicinity of Red Dog Mine, Alaska. *Science of the Total Environment*, *348*, 211–230.

Huang, C., Zhang, H., & Robeson, S. M. (2011). On the validity of commonly used covariance and variogram functions on the sphere. *Mathematical Geosciences*, *43*, 721–733.

Kleinman, K. P., & Ibrahim, J. G. (1998). A semi-parametric Bayesian approach to generalized linear mixed models. *Statistics in Medicine*, *17*, 2579–2596.

Lee, Y., & Nelder, J. A. (1996). Hierarchical generalized linear models. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*, 619–656.

Lehmann, E. L., & Casella, G. (2006). *Theory of point estimation*. Springer Science & Business Media.

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, *10*, 325–337.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). Chapman & Hall Ltd..

Muto, M. M., Helker, V. T., Delean, B. J., Young, N. C., Freed, J. C., Angliss, R. P., Friday, N. A., Boveng, J. M., Breiwick, B. M., Brost, M. F., Cameron, P. J., Clapham, J. L., Crance, S. P., Dahle, M. E., Dahlheim, B. S., Fadely, M. C., Ferguson, L. W., Fritz, K. T., Goetz, R. C., … Zerbini, A. N. (2022). *Alaska marine mammal stock assessments, 2021*. U.S. Department of Commerce, NOAA Technical Memorandum NMFS-AFSC-441. 295 p.

Neitlich, P. N., Hoef, J. M. V., Berryman, S. D., Mines, A., Geiser, L. H., Hasselbach, L. M., & Shiel, A. E. (2017). Trends in spatial patterns of heavy metal deposition on national park service lands along the Red Dog Mine haul road, Alaska, 2001–2006. *PLOS One*, *12*, e0177936.

Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society, Series A: General*, *135*, 370–384.

Pace, R. K., & Barry, R. (1997). Quick computation of spatial autoregressive estimators. *Geographical analysis*, *29*, 232–247.

Patterson, H., & Thompson, R. (1974). *Maximum likelihood estimation of components of variance*. In *Proceedings of the 8th International Biometric Conference* (pp. 197–207). Biometric Society.

Patterson, H. D., & Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, *58*, 545–554.

Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *71*, 319–392.

Rue, H., Riebler, A., Sørbye, S. H., Illian, J. B., Simpson, D. P., & Lindgren, F. K. (2017). Bayesian Computing with INLA: A review. *Annual Review of Statistics and Its Application*, *4*, 395–421.

Ruppert, D., Wand, M. P., & Carroll, R. J. (2003). *Semiparametric regression*. Campbridge University Press.

Sherman, J., & Morrison, W. J. (1949). Adjustment of an inverse matrix corresponding to changes in the elements of a given column or a given row of the original matrix. *Annals of Mathematical Statistics*, *20*, 621.

Snedecor, G. W., & Cochran, W. G. (1980). *Statistical methods* (7th ed.). Iowa State University.

Stiratelli, R., Laird, N., & Ware, J. H. (1984). Random-effects models for serial observations with binary response. *Biometrics*, *40*, 961–971.

Tierney, L., & Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, *81*, 82–86.

Ver Hoef, J. M. (2018). Kriging models for linear networks and non-Euclidean distances: Cautions and solutions. *Methods in Ecology and Evolution*, *9*, 1600–1613.

Ver Hoef, J. M., & Peterson, E. (2010). A moving average approach for spatial statistical models of stream networks (with discussion). *Journal of the American Statistical Association*, *105*, 6–18.

Ver Hoef, J. M., Peterson, E. E., Hooten, M. B., Hanks, E. M., & Fortin, M.-J. (2018). Spatial autoregressive models for statistical inference from ecological data. *Ecological Monographs*, *88*, 36–59.
Verbeke, G., & Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. Springer-Verlag Inc.
Warton, D. I., & Hui, F. K. (2011). The arcsine is asinine: the analysis of proportions in ecology. *Ecology*, *92*, 3–10.
Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*, *61*, 439–447.
Whittle, P. (1954). On stationary processes in the plane. *Biometrika*, *41*, 434–449.
Wood, S. N. (2020). Simplified integrated nested Laplace approximation. *Biometrika*, *107*, 223–230.
Woodbury, M. A. (1950). Inverting modified matrices, Statistical Research Group, Princeton N.J., published. *Memorandum Report*, *42*, 1–4.
Zeger, S. L., & Karim, M. R. (1991). Generalized linear models With random effects; A Gibbs sampling approach. *Journal of the American Statistical Association*, *86*, 79–86.
Zeger, S. L., Liang, K.-Y., & Albert, P. S. (1988). Models for longitudinal data: A generalized estimating equation approach. *Biometrics*, *44*, 1049–1060.
Zhang, H. (2002). On estimation and prediction for spatial generalized linear mixed models. *Biometrics*, *58*, 129–136.
Zhang, H. (2004). Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association*, *99*, 250–261.
Zilber, D., & Katzfuss, M. (2021). Vecchia–Laplace approximations of generalized Gaussian processes for big non-Gaussian spatial data. *Computational Statistics & Data Analysis*, *153*, 107081.

How to cite this article: Ver Hoef, J. M., Blagg, E., Dumelle, M., Dixon, P. M., Zimmerman, D. L., & Conn, P. B. (2024). Marginal inference for hierarchical generalized linear mixed models with patterned covariance matrices using the Laplace approximation. *Environmetrics*, e2872. https://doi.org/10.1002/env.2872

APPENDIX

A.1 Derivation of REML from integration

Consider a multivariate normal distribution for a general linear model,

$$[\mathbf{y};\boldsymbol{\beta},\boldsymbol{\theta}] = \frac{\exp\left(-\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right)}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}}, \tag{A1}$$

where $\mathbf{y}$ is an $n \times 1$ vector for the response variable, $\mathbf{X}$ is a $n \times p$ design matrix of explanatory variables, $\boldsymbol{\beta}$ is a $p \times 1$ vector of fixed effects, $\boldsymbol{\theta}$ contains covariance parameters of the $n \times n$ covariance matrix $\boldsymbol{\Sigma}$. It is possible to obtain REML equations by integrating out the fixed effects $\boldsymbol{\beta}$,

$$\int_{\mathbb{R}^p} f(\mathbf{y};\boldsymbol{\beta},\boldsymbol{\theta})d\boldsymbol{\beta},$$

to obtain a likelihood that is a function of just the covariance parameters $\boldsymbol{\theta}$ and the data $\mathbf{y}$. In particular

$$-2\ln\left(\int_{\mathbb{R}^p} f(\mathbf{y};\boldsymbol{\beta},\boldsymbol{\theta})d\boldsymbol{\beta}\right) = (n-p)\ln(2\pi) + \ln|\boldsymbol{\Sigma}| + \ln|\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X}| + (\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}),$$

where $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{y}$.

Proof. Write (A1) as

$$\begin{aligned}[\mathbf{y};\boldsymbol{\beta},\boldsymbol{\theta}] &= \frac{\exp\left(-\frac{1}{2}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}+\mathbf{X}\hat{\boldsymbol{\beta}}-\mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}+\mathbf{X}\hat{\boldsymbol{\beta}}-\mathbf{X}\boldsymbol{\beta})\right)}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}}, \\ &= \frac{\exp\left(-\frac{1}{2}[(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}) + (\mathbf{X}\hat{\boldsymbol{\beta}}-\mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Sigma}^{-1}(\mathbf{X}\hat{\boldsymbol{\beta}}-\mathbf{X}\boldsymbol{\beta}) + C\right)}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}},\end{aligned}$$

where $C = 2(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\boldsymbol{\Sigma}^{-1}(\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}) = 0$. Factor out terms that do not contain $\boldsymbol{\beta}$,

$$\int_{\mathbb{R}^p} [\mathbf{y}; \boldsymbol{\beta}, \theta] d\boldsymbol{\beta} = M \int_{\mathbb{R}^p} \exp\Big(-\frac{1}{2}(\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Sigma}^{-1}(\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta})\Big) d\boldsymbol{\beta},$$

where $M = \exp[-\frac{1}{2}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})]/[(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}]$. Notice that

$$\begin{aligned}
&\int_{\mathbb{R}^p} \exp\Big(-\frac{1}{2}(\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Sigma}^{-1}(\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta})\Big) d\boldsymbol{\beta}, \\
=&\int_{\mathbb{R}^p} \exp\Big(-\frac{1}{2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\Big) d\boldsymbol{\beta}, \\
=&2\pi^{p/2}|(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})|^{-1/2},
\end{aligned}$$

by recalling that, for positive definite $\mathbf{A}_{m\times m}$ and any conformable $\mathbf{x} \neq \mathbf{0}$,

$$\int_{\mathbb{R}^m} \exp(-\mathbf{x}'\mathbf{A}\mathbf{x}/2) d\mathbf{x} = (2\pi)^{m/2}|\mathbf{A}|^{-1/2}.$$

Hence, we arrive at

$$[\mathbf{y}; \theta] = \int_{\mathbb{R}^p} [\mathbf{y}; \boldsymbol{\beta}, \theta] d\boldsymbol{\beta} = \frac{\exp\Big(-\frac{1}{2}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})\Big)}{(2\pi)^{(n-p)/2}|\boldsymbol{\Sigma}|^{1/2}|\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X}|^{1/2}},$$

and taking $-2\ln[\mathbf{y}; \theta]$ we obtain the desired result. ∎

A.2 Distribution parameterizations

A.2.1 Negative binomial distribution

For the negative binomial, y_i is a non-negative integer with probability density function (PDF)

$$[y|\mu, \phi] = \frac{\Gamma(y + \phi)}{\Gamma(\phi)y!}\left(\frac{\mu}{\mu + \phi}\right)^y\left(\frac{\phi}{\mu + \phi}\right)^\phi,$$

where $0 < \mu < 1$, $0 < \phi$, $\mathrm{E}(Y) = \mu$, $\mathrm{var}(Y) = \mu + \mu^2/\phi$, and $\Gamma(\cdot)$ is the gamma function.

A.2.2 Gamma distribution

For the gamma distribution, y_i is positive with PDF

$$[y|\mu, \phi] = \frac{1}{\Gamma(\phi)}\left(\frac{\phi}{\mu}\right)^\phi y^{\phi-1}\exp\left(\frac{-y\phi}{\mu}\right),$$

where $0 < \mu$, $0 < \phi$, $\mathrm{E}(Y) = \mu$, and $\mathrm{var}(Y) = \mu^2/\phi$.

A.2.3 Beta distribution

For the beta distribution, $0 < y_i < 1$ with PDF

$$[y|\mu, \phi] = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)}y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1},$$

where $0 < \mu < 1$, $0 < \phi$, $\mathrm{E}(Y) = \mu$, and $\mathrm{var}(Y) = \mu(1-\mu)/(1+\phi)$.

A.2.4 Inverse Gaussian distribution

The inverse Gaussian distribution is usually written as,

$$[y;\mu,\lambda] = \sqrt{\frac{\lambda}{2\pi y^3}}\exp\left(-\frac{\lambda(y-\mu)^2}{2\mu^2 y}\right),$$

where $y > 0$, $\mu > 0$, and $\lambda > 0$. In this parameterization λ is a shape parameter, and $\mathrm{E}(Y) = \mu$ and $\mathrm{var}(Y) = \mu^3/\lambda$. In order to keep μ positive and w unconstrained in (1), we let $\boldsymbol{\mu} = \exp(\mathbf{w})$. However, under this construction, from (9), we obtain

$$D_{i,i} = \frac{(e^{w_i} - 2y_i)}{\phi e^{2w_i}},$$

and some $D_{i,i}$ can be positive whenever $e^{w_i} > 2y_i$, which can lead to $\mathbf{H}$ in (10) being singular. We propose an alternative parameterization. For inverse Gaussian models, λ is often scaled, and here we do so by taking $\phi = \lambda/\mu = \lambda/\exp(w)$, yielding a μ-scaled-λ inverse Gaussian model,

$$[y;\mu,\lambda] = \sqrt{\frac{\phi\exp(w)}{2\pi y^3}}\exp\left(-\frac{\phi(y-\exp(w))^2}{2\exp(w)y}\right),$$

where $\phi > 0$ and now $\mathrm{var}(Y) = \mu^2/\phi$. Under this parameterization, we have

$$D_{i,i} = -\frac{\phi(e^{2w_i} + y_i^2)}{2ye^{w_i}},$$

which is always negative, and so (10) is always well-behaved. Under this construction, we also have

$$d_i = \phi\left(\frac{y}{2e^{w_i}} - \frac{e^{w_i}}{2y}\right) + \frac{1}{2}.$$