Environmetrics
September 2024, Volume 35, Issue 6, Pages e2871 (21p.)
https://doi.org/10.1002/env.2871
https://archimer.ifremer.fr/doc/00897/100863/

EM algorithm for generalized Ridge regression with spatial covariates

Obakrim Said [1, 2, *], Ailliot Pierre [3], Monbet Valerie [1], Raillard Nicolas [2]

[1] Univ Rennes CNRS IRMAR-UMR 6625 Rennes, France
[2] Unité Recherche et Développements Technologiques IFREMER Plouzané,France
[3] Laboratoire de Mathématiques de Bretagne Atlantique UMR CNRS 6205, Univ. Brest Brest,France

* Corresponding author : Said Obakrim, email address : saidobak@gmail.com

Abstract :

The generalized Ridge penalty is a powerful tool for dealing with multicollinearity and high-dimensionality in regression problems. The generalized Ridge regression can be derived as the mean of a posterior distribution with a Normal prior and a given covariance matrix. The covariance matrix controls the structure of the coefficients, which depends on the particular application. For example, it is appropriate to assume that the coefficients have a spatial structure when the covariates are spatially correlated. This study proposes an Expectation-Maximization algorithm for estimating generalized Ridge parameters whose covariance structure depends on specific parameters. We focus on three cases: diagonal (when the covariance matrix is diagonal with constant elements), Matérn, and conditional autoregressive covariances. A simulation study is conducted to evaluate the performance of the proposed method, and then the method is applied to predict ocean wave heights using wind conditions.

Keywords : conditional autoregressive, EM algorithm, generalized ridge, Matérn, spatial covariates

1 | INTRODUCTION

Recently, there has been a growing interest in the climate community in using statistical and data-driven methods as alternatives to computationally expensive, physics-based deterministic models (Reichstein et al. 2019). For instance, statistical methods are increasingly utilized for wave climate characterization, a crucial aspect for sectors such as marine engineering and coastal management (Camus et al. 2017; Charles, Idier, Delecluse, Déqué, & Le Cozannet 2012; Michel, Obakrim, Raillard, Ailliot, & Monbet 2022; Obakrim, Ailliot, Monbet, & Raillard 2023; Otto, Piter, & Gijsman 2021). Despite their promise, these methods face significant practical challenges. The high dimensionality of the data, due to a limited number of observations combined with a large number of covariates, increases the risk of overfitting. Additionally, strong spatial dependencies among covariates create issues with multicollinearity. To enhance predictive accuracy and physical interpretability, statistical and data-driven models must account for these complexities (Stevens et al. 2021). This study focuses on regression models with one dependent variable and covariates exhibiting strong spatial dependencies. An example of such a scenario is the prediction of ocean wave heights at a specific East Atlantic coastal location (Charles et al. 2012; Obakrim et al. 2023), where the regression model uses North Atlantic wind conditions as covariates represented by gridded data with significant spatial dependencies.

Classically, in regression models for spatial data, it is generally assumed that both the response variable and the covariates are available at each spatial location (see e.g. Heaton et al. (2019) for a review). In this context, methods that combine penalized likelihood approaches with geostatistical models to manage high dimensionality have been discussed in the literature (Chu, Zhu, & Wang 2011; Maranzano, Otto, & Fassò 2023). The most common way of regularizing a regression problem is to reduce the dimension of the covariate space. This reduction can be achieved by approximating the covariance function of the spatial covariate with a low-rank matrix, similar to the principles of PCA (see, for instance, Cressie and Johannesson (2008)), or through variable selection. Conventional variable selection methods, such as LASSO or SCAD penalties, are commonly employed in spatial statistical regression models (Liang, Cheng, Su, Xiao, & Song 2022), and spatial variable selection through cross-validation has been explored in previous work (Meyer, Reudenbach, Wöllauer, & Nauss 2019). However, we argue that regularization via variable selection may not be the most convenient approach when spatial covariates are highly correlated. When covariates at all spatial locations impact the response variable similarly, selecting specific variables could result in arbitrary choices. Instead, we propose implementing a Ridge regularization approach van Wieringen (2015) in the estimation procedure, which considers the spatial structure of the covariates and helps controlling the estimation variance.

Consider an experiment where we have the data $\{y, X\}$, of n observations of a continuous variable Y and $n \times d$ matrix of covariates X. Suppose that Y is related to X via a linear model

$$Y = X\beta + \epsilon, \tag{1}$$

where β are model coefficients and $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is the model error. We suppose that the intercept is either included in β (so that the first column of X is a vector of 1) or that Y and X are centered. In this study, we focus on the case where the response variable and covariates are climate variables. Climate data are known to be multicollinear and high-dimentional (Hessami, Gachon, Ouarda, & St-Hilaire 2008; Permatasari, Djuraidah, & Soleh 2017; Sungkawa, Rahayu, et al. 2019). In the case of multicollinearity or high-dimensionality, penalized linear regression methods, like Ridge regression, are needed to control the variance of the estimate. For example, Hessami et al. (2008) used Ridge regression to downscale precipitation and temperature in eastern Canada and pointed out that Ridge estimates are more robust than ordinary least squares estimates. Ridge estimator of the problem (1) is

$$\hat{\beta}_{\lambda}^{Ridge} = \arg\min_{\beta} -\ell(\beta, \sigma^2) + \lambda\|\beta\|^2 \tag{2}$$

where λ is the regularization parameter and $\ell(\beta, \sigma^2)$ is the log-likelihood of the model (1). The hyperparameter λ needs to be selected in order to get a trade-off between variance and bias (Hastie, Tibshirani, Friedman, & Friedman 2009), since for example high values of λ permit to reduce the variance but increase the bias of the model.

Boonstra, Mukherjee, and Taylor (2015) classified methods for selecting λ into goodness-of-fit-based and likelihood-based methods. Goodness-of-fit-based methods define a goodness of fit criterion (such as the mean squared error) and minimize it in terms of λ. The most common goodness-of-fit-based method is the k-fold cross-validation which consists of partitioning observations into k groups and estimating β k times for each λ leaving out one group. For each λ, a goodness of fit score is calculated, and λ with the minimum score value is chosen. The typical choices of k are 5 and 10, while setting $k = n$ leads to leave-one-out cross-validation (LOOCV). LOOCV leads to a better estimation of λ; however, it is computationally expensive given that it requires fitting the model n times (Patil, Wei, Rinaldo, & Tibshirani 2021). Generalized cross-validation (GCV) (Golub, Heath, & Wahba 1979) is an approximation of LOOCV that does not require fitting n models. GCV uses a weighted version of the predicted residual error sum of squares (PRESS) statistic (Allen 1974) as a goodness of fit criterion. One of the problems with goodness-of-fit-based methods is the selection of the grid for the search of the optimal λ, which influences the estimation.

Assuming that $Y|\beta \sim \mathcal{N}(X\beta, \sigma^2 I_n)$, Ridge regression can be derived as the mean of a posterior distribution with the prior $\beta \sim \mathcal{N}(0_d, \sigma^2\lambda^{-1} I_d)$ (van Wieringen 2015) and as in hierarchical linear regression, likelihood-based methods maximize the likelihood with respect to σ^2 and λ using for instance an iterative method (Boonstra et al. 2015). Unlike goodness-of-fit-based methods, the advantage of likelihood-based approaches is, on the one hand, that they do not require grid selection for the regularization parameters. On the other hand, likelihood-based methods can be generalized to consider any form of prior for the coefficients β, such as spatial dependence. For instance, Tew, Schmidt, and Makalic (2022) propose a diagonal covariance for β with varying variances. This covariance shape allows a local shrinkage.

When the covariates have a spatial structure, it is reasonable to suppose that coefficients have a spatial structure and a joint penalization of the coefficients is required (Tibshirani, Saunders, Rosset, Zhu, & Knight 2005). To do that, the generalized Ridge (van Wieringen 2015) can be used. Generalized Ridge extends the equation (2) by replacing the term $\lambda\|\beta\|^2$ to $\beta^T\Delta\beta$, where Δ is called the penalty matrix. In general, Δ depends on some regularization parameters (see, e.g., Goeman (2008) and Hemmerle (1975)); however, when the number of the regularization parameters is greater than 1, it is difficult to tune these parameters due to the combinatorial explosion. Generalized Ridge in the hierarchical linear model framework, is equivalent to suppose that $\beta \sim \mathcal{N}(0_d, \Sigma_\theta)$ where Σ_θ is a covariance matrix that depends on some parameters θ. Note that Σ_θ corresponds to the inverse of the penalty matrix Δ. The classical Ridge is a special case of this model when the covariance matrix Σ_θ is proportional to the identity matrix, and θ is the usual regularization parameter λ.

β is a latent field and this suggests using an Expectation Maximization algorithm to estimate the parameters. Indeed the EM algorithm (Dempster, Laird, & Rubin 1977) is probably the most usual method for estimating the parameters of a model with latent variables. The EM algorithm alternates between two steps: the expectation (E-step) and maximization (M-step) steps. The E-step calculates the conditional expectation of the log-likelihood given the observations and current parameters. In the M-step, the parameters are estimated by maximizing the conditional expectation of the log-likelihood calculated in the E-step. Note that EM algorithm for maximizing the likelihood in spatial models has already been used for instance for multivariate coregionalization models (Zhang (2007) Fassò and Finazzi (2011)).

In this study, we extend the algorithm in Bishop and Nasrabadi (2006), Tew et al. (2022) and Tew, Boley, and Schmidt (2024) and propose an EM algorithm to estimate the parameters of hierarchical linear regression when $\beta \sim \mathcal{N}(0, \Sigma_\theta)$. At first, we study the case where Σ_θ is diagonal with constant elements, which corresponds to the classical Ridge in equation (2) and the problem studied by Bishop and Nasrabadi (2006) and Tew et al. (2022). Then, our main contribution is to consider the case where the coefficients β have a spatial structure, especially when Σ_θ is the Matérn or the conditional autoregressive (CAR) covariance. To the best of our knowledge, this is the first time that it is proposed to regularize a spatial covariate linear regression by considering a latent parameter with a structure and calibrating the model using an EM algorithm in order to prevent the selection of the regularisation constant.

This paper is organized as follows. The proposed method and its special cases are presented in Section 2. Then, a simulation study is conducted in Section 3 to assess the performance of the proposed method. In section 4, we apply the methodology to oceanography data where the proposed methodology is used to predict significant wave height at a location in the Bay of Biscay using North Atlantic wind conditions as covariates. Finally, this study is concluded in Section 5.

2 | PROPOSED METHOD

As stated in the introduction, Ridge regression can be viewed as a hierarchical linear model where $\beta \sim \mathcal{N}(0_d, \sigma^2 \lambda^{-1} I_d)$. When there is a structure on the coefficients, it is unreasonable to consider all possible covariance functions as possible candidates for β. Therefore, we suppose that the covariance of β depends on some parameters θ, so that $\beta \sim \mathcal{N}(0_d, \Sigma_\theta)$. This motivates using the EM algorithm to find the maximum likelihood estimation of the parameters, where the model parameters are then $\Theta = (\sigma^2, \theta)$. The proposed method is described in this section, and three special cases of the covariance Σ_θ (the diagonal, Matérn, and CAR) are studied.

2.1 | EM algorithm for generalized Ridge

Consider the linear model (1) and assume that β is a latent variable that follows a normal distribution. We define the regression model hierarchically as

$$\begin{aligned} &\beta \sim \mathcal{N}(0_d, \Sigma_\theta) \\ &Y \mid \beta, \Theta \sim \mathcal{N}(X\beta, \sigma^2 I_n) \end{aligned} \tag{3}$$

where $\Theta = (\sigma^2, \theta)$. Note that for simplicity, we assume that the mean of β is zero. The EM algorithm for the case where β has a non-zero mean will be presented in the Appendix.

Given a sample $y = (y_1, ..., y_n)$, the complete log-likelihood is expressed as

$$\begin{aligned} \ln p(y, \beta; \Theta) &= \ln p(y \mid \beta; \sigma^2) + \ln p(\beta; \theta) \\ &= -\frac{1}{2}\left(n \ln(2\pi) + n \ln(\sigma^2) + \frac{1}{\sigma^2}\|y - X\beta\|^2 + d \ln(2\pi) + \ln(|\Sigma_\theta|) + \beta^T \Sigma_\theta^{-1} \beta\right) \end{aligned} \tag{4}$$

Maximum likelihood estimation consists of maximizing (4) with respect to the parameters Θ. This is usually done with the Expectation-Maximization algorithm in the latent variable context. The EM algorithm alternates between the E-step and M-step. In the E-step, the expectation $Q(\Theta|\Theta^{(i)})$ of the complete likelihood with respect to the posterior distribution of the latent variable β and the parameters $\Theta^{(i)}$ from the previous iteration i is calculated. In the M-step, the quantity $Q(\Theta|\Theta^{(i)})$ is maximized with respect to the parameters Θ.

The E-step and M-step are defined as follows

- E-step:

$$Q(\Theta|\Theta^{(i)}) = \mathbb{E}(\ln p(y, \beta; \Theta) \mid y, \Theta^{(i)}). \tag{5}$$

The posterior distribution of the latent variable β is a normal distribution with mean $\mu_{\beta|y}$ and covariance matrix $\Sigma_{\beta|y}$ such that

$$\begin{cases} \Sigma_{\beta|y} = (\Sigma_\theta^{-1} + \frac{1}{\sigma^2} X^T X)^{-1} \\ \mu_{\beta|y} = (X^T X + \sigma^2 \Sigma_\theta^{-1})^{-1} X^T y. \end{cases} \tag{6}$$

Note that $\mu_{\beta|y}$ defined in (6) is a generalized Ridge estimator (see e.g. van Wieringen (2015)) solution of the optimization problem

$$\mu_{\beta|y} = \arg\min_\beta \frac{\|y - X\beta\|^2}{\sigma^2} + \beta^T \Sigma_\theta^{-1} \beta \tag{7}$$

Therefore,

$$Q(\Theta|\Theta^{(i)}) = -\frac{1}{2}\left(\ln(|\Sigma_\theta|) + \mathrm{Tr}(\Sigma_\theta^{-1}\mathbb{E}(\beta\beta^T \mid y, \Theta^{(i)})) + n\ln(\sigma^2) + \frac{1}{\sigma^2}\mathbb{E}(\|y - X\beta\|^2 \mid y, \Theta^{(i)})\right) + C \tag{8}$$

where $C = (n + d)\ln(2\pi)$ is a constant and

$$\begin{cases} \mathbb{E}(\beta\beta^T|y; \Theta^{(i)}) = \Sigma_{\beta|y} + \mu_{\beta|y}\mu_{\beta|y}^T \\ \mathbb{E}(\|y - X\beta\|^2|y; \Theta^{(i)}) = \|y\|^2 - 2y^T X\mu_{\beta|y} + \mathrm{Tr}(X^T X\mathbb{E}(\beta\beta^T|y; \Theta^{(i)})), \end{cases} \tag{9}$$

and Tr denotes the trace operation.

- M-step:

In the M-step, the goal is to compute the updates of the parameters, denoted by $\Theta^{(i+1)}$, that maximize the objective function $Q(\Theta|\Theta^{(i)})$ given

$$\Theta^{(i+1)} = \arg\max_\Theta Q(\Theta|\Theta^{(i)}) \tag{10}$$

The closed-form update for the variance, denoted by $\sigma^{2,(i+1)}$, is computed as follows:

$$\sigma^{2,(i+1)} = \frac{1}{n}(\|y\|^2 - 2y^T X\mu_{\beta|y} + \mathrm{Tr}(X^T X\mathbb{E}(\beta\beta^T|y; \Theta^{(i)}))). \tag{11}$$

The update for the parameter θ, denoted by $\theta^{(i+1)}$ is found by maximizing the following optimization problem:

$$\theta^{(i+1)} = \arg\max_\theta \ln(|\Sigma_\theta^{-1}|) - \mathrm{Tr}(\Sigma_\theta^{-1}\mathbb{E}(\beta\beta^T \mid y, \Theta^{(i)})). \tag{12}$$

This optimization requires the parameterization of the covariance matrix Σ_θ, the details of which will be discussed in the subsequent section, highlighting certain special cases.

2.2 | Special cases

The M-step in equation (12) requires the maximization of $Q(\Theta|\Theta^{(t)})$ over the parameters of the covariance Σ_θ. In this study, we will explore three cases. First, we consider the case where Σ_θ is diagonal. Then, the case where β has a spatial structure, especially when the parametric covariance is the Matérn covariance function. Finally, we consider the conditional autoregressive model (CAR).

2.2.1 | Diagonal case

In the classical Ridge, the covariance matrix of the coefficients β is supposed to be diagonal such that

$$\Sigma_\theta = \sigma_\beta^2 \mathbf{I_d}. \tag{13}$$

Consequently, the determinant of the inverse covariance matrix can be expressed as:

$$|\Sigma_\theta^{-1}| = \prod_{i=1}^{d} \sigma_\beta^{-2} = \sigma_\beta^{-2d}. \tag{14}$$

Substituting this into (11), we obtain:

$$\sigma_\beta^{2,(i+1)} = \arg\max_{\sigma_\beta^2} \; -d\ln(\sigma_\beta^2) - \frac{1}{\sigma_\beta^2}\mathrm{Tr}(\mathbb{E}(\beta\beta^T \mid y, \Theta^{(i)})). \tag{15}$$

Setting the derivatives with respect to σ_β^2 to zero, we obtain the M-step

$$\sigma_\beta^{2,(i+1)} = \frac{\mathrm{Tr}(\mathbb{E}(\beta\beta^T \mid y, \Theta^{(i)}))}{d}. \tag{16}$$

Note that $\frac{1}{\sigma_\beta^2}$ corresponds to the regularization parameter λ in (1). In the classical Ridge regression framework, selecting the regularization parameter λ is essential and is typically achieved through cross-validation methods. Cross-validation methods have demonstrated efficacy in mitigating overfitting across a variety of statistical and machine learning applications. However, in our study, the hierarchical regression model defined in (3) inherently incorporates the regularization of the coefficients through the covariance Σ_θ. Through a simulation study, we show in Appendix A that the EM algorithm effectively estimates the regularization parameters without the need for cross-validation. As mentioned in the introduction, most Ridge regression cross-validation methods necessitate selecting a grid for the regularization parameters, a task that is not straightforward. Furthermore, when dealing with multiple regularization parameters, as elaborated in subsequent sections, applying cross-validation to select parameters becomes challenging due to combinatorial explosion. In contrast, the EM algorithm allows for the estimation of regularization parameters directly from the data. We provide a comparison of the two methods (cross-validation and EM algorithm) in the Appendix A.

It is important to note that some studies argue that the EM algorithm may be prone to overfitting (Andrews 2018; Tian, Xia, Zhang, & Feng 2011), leading to the consideration of hybrid approaches that combine the EM algorithm and cross-validation to mitigate this issue (Shinozaki & Ostendorf 2008; Takenouchi & Ikeda 2010). However, this aspect falls beyond the scope of our paper, which primarily focuses on demonstrating the efficiency of the EM algorithm in estimating regularization parameters.

2.2.2 | Spatial covariance functions

In spatial statistics applications, one may assume that β has a spatial structure. One way to do that is to assume that β has a parametric covariance function. There are many choices of covariance functions that are widely used for Gaussian processes and kriging (Schulz, Speekenbrink, & Krause 2018). In this study, we focus on the stationary Matérn covariance, which has the form

$$K(h; \phi, \kappa) = \frac{\sigma_\beta^2}{2^{\kappa-1}\Gamma(\kappa)} \left(\frac{h}{\phi}\right)^\kappa K_\kappa\left(\frac{h}{\phi}\right) \tag{17}$$

where h is the distance between two points, Γ is the Gamma function, and K_κ is the modified Bessel function (Abramowitz, Stegun, & Romer 1988). The Matérn function is parameterized by the variance parameter σ_β^2, the range parameter ϕ, and the smoothness parameter κ. The range parameter ϕ controls the decay rate with distance, with larger values of ϕ corresponding to more strongly correlated variables, and the smoothness parameter κ controls the mean-square differentiability of the spatial process.

By expressing $\Sigma_\theta = \sigma_\beta^2 R_\theta$, the maximization step for the covariance of β in (13) can be reformulated as follows:

$$(\sigma_\beta^{2,(i+1)}, \theta^{(i+1)}) = \arg\max_{\sigma_\beta^2, \theta} \; \ln(|R_\theta^{-1}|) - d\ln(\sigma_\beta^2) - \frac{1}{\sigma_\beta^2}\mathrm{Tr}(R_\theta^{-1}\mathbb{E}(\beta\beta^T \mid y, \Theta^{(i)})) \tag{18}$$

where R_θ is the Matérn correlation and $\theta = (\phi, \kappa)$. Since the variance parameter is constant and following Bachoc (2013), the optimization of the variance parameter σ_β^2 can be carried out separately with the correlation parameters ϕ and κ. Therefore,

$$\begin{aligned} \sigma_\beta^{2,(i+1)} &= \frac{\mathrm{Tr}(R_\theta^{-1}\mathbb{E}(\beta\beta^T \mid y, \Theta^{(i)}))}{d} \\ \theta^{(i+1)} &= \arg\max_\theta \; \ln(|R_\theta^{-1}|) - d\ln(\mathrm{Tr}(R_\theta^{-1}\mathbb{E}(\beta\beta^T \mid y, \Theta^{(i)}))). \end{aligned} \tag{19}$$

The solution to the optimization problem in equation (19) cannot be done analytically; therefore, numerical optimization algorithms are used. This study uses the quasi-Newton method L-BFGS-B to optimize the parameters. Given the difficulties in estimating Matérn parameters (Kaufman & Shaby 2013), we a priori fix the smoothness parameter as $\frac{3}{2}$, which gives the classical $\frac{3}{2}$-Matérn covariance function.

2.2.3 | Conditional autoregressive model

The M-step in equation (10) requires the inversion of the covariance matrix, which can be challenging for large matrices. This problem is wildly discussed in Gaussian processes literature (Ambikasaran, Foreman-Mackey, Greengard, Hogg, & O'Neil 2015;

Storkey 1999). Therefore, it can be numerically advantageous to parameterize the precision matrix (inverse of the covariance matrix) instead of the covariance matrix. This is motivated by the fact that the precision matrix $P_\theta = \Sigma_\theta^{-1}$ can be approximated by a sparse matrix (Tajbakhsh, Aybat, & Del Castillo 2020). In fact, the off-diagonal elements of the precision matrix correspond to the conditional covariance between two variables given the remaining variables. Therefore, conditionally independent variables have zero values in the precision matrix.

Gaussian Markov random fields (GMFs) are wildly used in spatial statistics (Cressie & Wikle 2015). GMFs models have a Markov property making them computationally and theoretically suitable (Rue 2001). Furthermore, (Rue & Tjelmeland 2002) demonstrated that a GMF model can approximate a Gaussian field with a Matérn correlation function and other families of correlation functions. Conditional autoregressive (CAR) models are classes of GMFs with well-defined joint Gaussian distribution (Cressie & Kapat 2008). This subsection will study cases where the coefficients β have the CAR model property. The joint distribution of a CAR is expressed as

$$\beta \sim \mathcal{N}(0, \tau^2 (I_d - \alpha H)^{-1} \Phi). \tag{20}$$

The distribution of β depends on unknown parameters α and τ^2, and many types of CAR models depend on the choice of the matrices H and Φ. Following Besag, York, and Mollié (1991), in this study, we consider the Weighted CAR (WCAR) model where

$$\Phi = \operatorname{diag}(|N_1|^{-1}, ..., |N_d|^{-1}) \tag{21}$$

where $|N_i|$ is the number of neighbors of location i and $H = \left(\frac{a_{ij}}{|N_i|}\right)_{d \times d}$; $i, j = 1, ..., d$, where a_{ij} is the (i, j) element of the adjacency matrix $A = (a_{ij})_{d\times d}$, where $a_{ij} = a_{ji} = 1$ if and only if location i and j are neighbors and otherwise $a_{ij} = 0$. Putting $P_\theta = \tau^{-2} \Phi^{-1}(I_d - \alpha H)$, the second part of the M-step in the equation (11) becomes

$$\theta^{(i+1)} = \arg\max_{\theta} \; \ln(|P_\theta|) - \operatorname{Tr}(P_\theta \mathbb{E}(\beta\beta^T | y, \theta^{(i)})) \tag{22}$$

where $\theta = (\tau^2, \alpha)$.

As for the Matérn covariance, the solution to the optimization problem (22) cannot be done analytically, and the numerical optimization algorithm L-BFGS-B is used. Note that the optimization of the variance parameter τ^2 can also be carried out separately with the parameter α.

Remark that this leads to a spatial extension of the fused Ridge method proposed in Goeman (2008). Indeed, when $\alpha = 1$

$$\beta P_\theta \beta^T = \frac{1}{\tau^2} \beta^T \Phi^{-1} (I_d - \alpha H) \beta = \frac{1}{2\tau^2} \sum_{(i,j)|a_{ij}=1} (\beta_i - \beta_j)^2. \tag{23}$$

and thus equation (7) becomes

$$\mu_{\beta|y} = \arg\min_{\beta} \frac{\|y - X\beta\|^2}{\sigma^2} + \frac{1}{2\tau^2} \sum_{(i,j)|a_{ij}=1} (\beta_i - \beta_j)^2 \tag{24}$$

This shows that any spatial coefficient variations will be penalized when solving (24). In this case, replacing the L2 norm with the L1 norm leads to the fused LASSO method proposed in Tibshirani et al. (2005). However, the matrix $(I_p - \alpha H)$ is not semi-positive definite when $\alpha = 1$ and thus Σ_θ is degenerate. Hereafter we impose the constraints $|\alpha| < 1$ to ensure that the precision matrix is positive definite. Another strategy would consist of adding a regular Ridge penalty (e.g., the discussion in van Wieringen (2015)).

3 | SIMULATION STUDY

In this section, a simulation study is conducted to assess the performance of the proposed method for estimating model parameters for the three cases: diagonal, Matérn, and CAR.

3.1 | Setup

This study focuses on using the proposed method for spatial applications. Therefore, we consider a 15×15 regular spatial grid in a square domain $[1, 15]^2$ where each location j has a covariate x_j. We generate $X = (x_{ij})_{n\times d}$ of n independent and identically distributed observations from a multivariate normal distribution with zero mean and a Matérn covariance with some arbitrary parameters $(\sigma_x^2, \phi_x, \kappa_x) = (6, 2, 3/2)$. Then, the coefficients β, kept the same for all observations, are simulated using either the diagonal, Matérn, or CAR case. Finally, for a given σ^2, Y is simulated from the normal distribution according to equation (3).

The parameters chosen for each case are:

- Diagonal: $\sigma^2 = 36$ and $\sigma^2_\beta = 7$
- Matérn: $\sigma^2 = 36$, $\sigma^2_\beta = 0.1$ and $\phi = 4$
- CAR: $\sigma^2 = 36$, $\tau^2 = 1$ and $\alpha = 0.9$

The parameters are chosen so that the results of the three methods are comparable. For the CAR model, we consider four neighbors to construct the adjacency matrix, and we chose $\alpha = 0.9$ to sufficiently smooth the resulting coefficients.

The EM algorithm is initialized with an arbitrary set of parameters, and the E-step and M-step are repeated until no further improvement can be made to the likelihood value or to limit the computational cost until a maximum number of iterations is reached. The computation time for one iteration on an i5-7500 CPU and 16Go computer is 0.16, 3, and 1.8 seconds for diagonal, Matérn, and CAR, respectively.

3.2 | Results

In this section, we present the results of the simulation study, outlined into five distinct segments. First, we employ a single simulation procedure to assess the accuracy of the estimated regression coefficients (Figure 1). Following this, we conduct three additional simulations using a Monte Carlo approach. These simulations assess the method's performance under varying conditions such as sample size, number of covariates, and residual variance (Figures 2 to 5). Lastly, we conduct a final Monte Carlo simulation to evaluate the method's robustness in estimating the regression coefficients in the case where the coefficients are estimated using a different covariance model that was used for simulation. In instances where the sample size (n) and the number of covariates (d) remain unchanged, we use a standard configuration where the sample size is set at 800, and the number of covariates is set at 250.

At first, one simulation is done for each case (diagonal, Matérn and CAR) with $n = 800$. The parameters are estimated using the EM algorithm presented in the previous section. Figure 1 shows the first simulation results. Left panels correspond to the true β, and right panels correspond to the estimated β using the EM algorithm. For all the cases, the EM algorithm does well in estimating the parameters, especially the variance σ^2.

To assess the influence of the sample size on the estimations, for each case, we perform 100 independent random simulations for each sample size varying from 50 to 850. For each simulation, the EM algorithm is used to estimate the parameters. Figure 2 shows the normalized root mean square error $NRMSE_\beta$ and $NRMSE_y$ for the three cases where

$$\begin{aligned} NRMSE_\beta &= \frac{\sqrt{\frac{1}{d}\sum_j^d(\beta_j - \hat{\beta}_j)^2}}{\hat{\sigma_\beta}} \\ NRMSE_y &= \frac{\sqrt{\frac{1}{n\prime}\sum_i^{n\prime}(y_i - \hat{y}_i)^2}}{\hat{\sigma}_y} \end{aligned} \tag{25}$$

where $\hat{\beta}_j$ and $\hat{y}_i$ are the estimated β_j and y_i and $\hat{\sigma_\beta}$ and $\hat{\sigma}_y$ are the sample standard deviation of β and y, respectively. $NRMSE_y$ is calculated in a test set (which is not used in the estimation) of size $n' = \frac{n}{2}$. For the three cases, $NRMSE_\beta$ and $NRMSE_y$ decrease as the sample size increases. However, the decay rate is quicker for the diagonal case, both for the prediction of the coefficients β and the response variable y. This disparity is expected as the parameter space for β is less restricted compared to the other cases where a spatial structure is constrained.

To evaluate the parameter estimates, we compare the EM estimates with the maximum likelihood estimates of the parameters, hereafter referred to as MLE, knowing the true β. More precisely, the MLE estimates are defined as

$$\Theta_{\text{MLE}} = \arg\max_{\Theta} -\frac{1}{2}\left(\ln(|\Sigma_\theta|) + \beta^T_{\text{true}}\Sigma^{-1}_\theta\beta_{\text{true}} + n\ln(\sigma^2) + \frac{1}{\sigma^2}\|y - X\beta_{\text{true}}\|^2\right) + C \tag{26}$$

where β_{true} is the true β simulated for each case with the parameters given in section 3.1. Along with the sample size, we are also interested in how the estimates behave when varying the number of covariates, d, and the variance parameter σ^2. Note that in practice, Θ_{MLE} cannot be found in practical applications, given that the true β is not observed (latent variable). Therefore, we expect the EM algorithm to provide less accurate estimates than MLE. However, we expect that by varying the sample size, the dimension, and the variance σ^2, the EM estimates asymptotically will be close to MLE estimates.

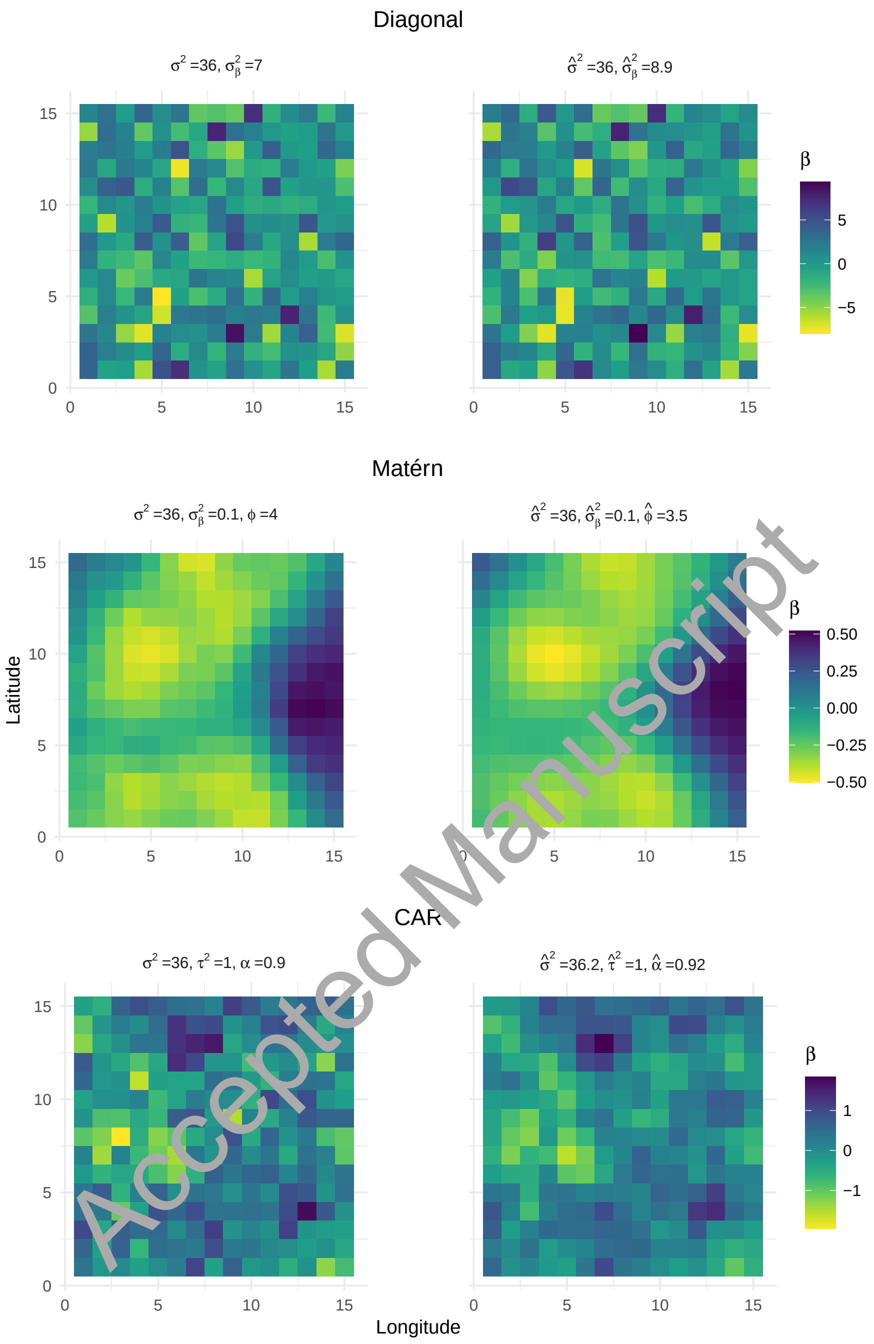

FIGURE 1 Simulation results for the three cases (diagonal, CAR, and Matérn). The left panels correspond to the true β coefficients with the true parameters given in section 3.1, and the right panels correspond to the β estimated when the sample size $n = 800$ and number of covriates $d = 225$

.

Figures 3 , 4 and 5 show boxplots of EM (red) and MLE (blue) estimates for the diagonal, Matérn and CAR cases as a function of sample size, number of covariates d, and variance σ^2. For the diagonal case, the estimate of σ^2 seems to converge to the true value of the parameter (blue line) when the sample size n increases as it does in the usual linear regression model. It is noteworthy that the variability in estimates of the spatial variance σ_β^2 does not significantly change as the sample size increases, but when n is large enough, EM and MLE seem to provide similar results. This is not unexpected since both methods are based on a single sample of the d-dimensional field β. As expected, the number of covariates d also affects the estimate of the parameter

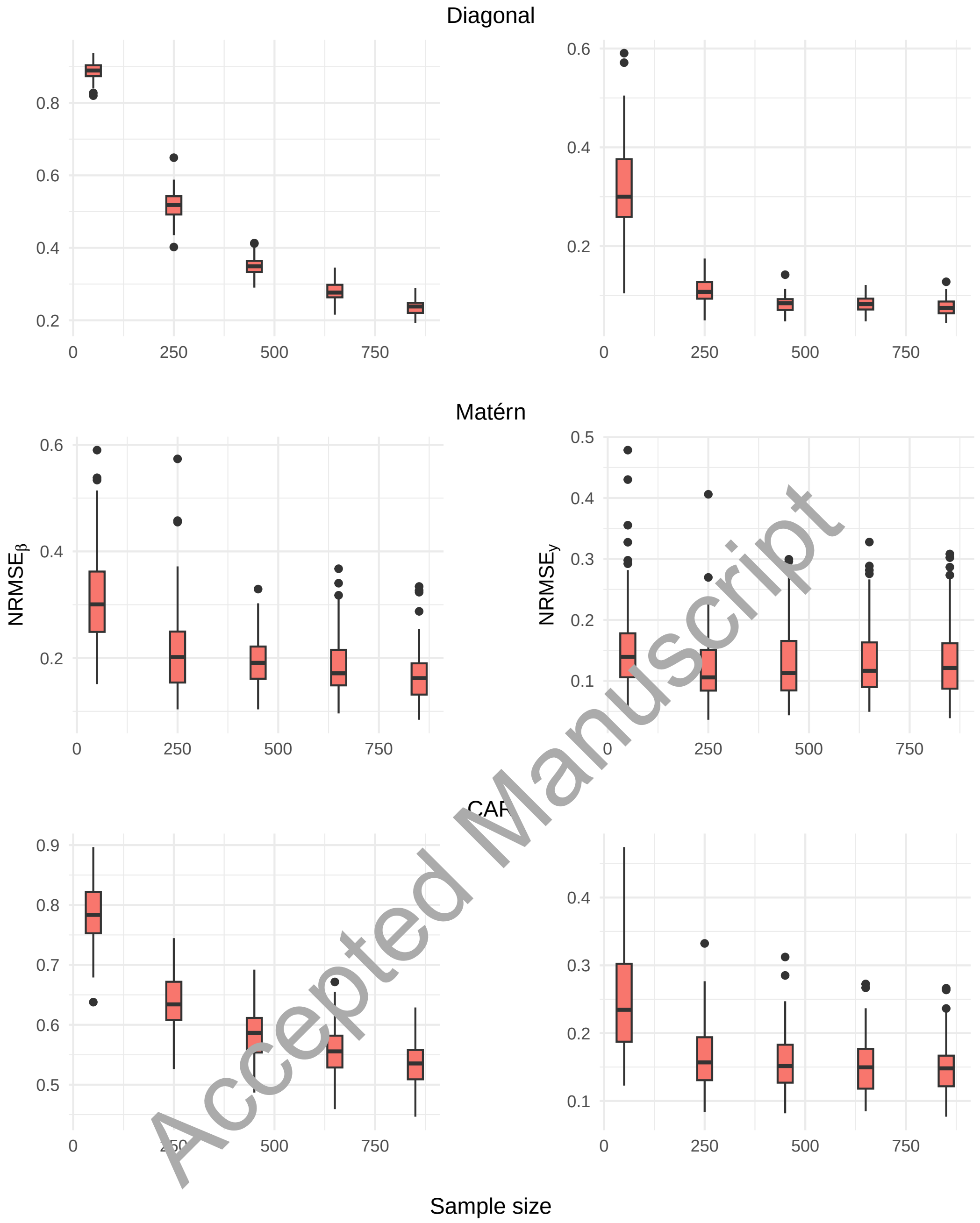

FIGURE 2 Results of $NRMSE_\beta$ (left panels) and $NRMSE_y$ (right panels) for the diagonal, CAR, and Matérn case as a function of the sample size n varying from 50 to 850 with a fixed number of covriates $d = 225$

.

σ^2_β, which converges towards the true value as d increases; however, no significant change is observed for σ^2 when d increases. The effect of the variance σ^2 on the estimation of σ^2_β is small, and we observe that for σ^2 larger than 100, the EM and MLE tend to underestimate σ^2_β. Similar behavior can be observed for the Matérn case: the variance parameter σ^2 seems to converge towards the actual value with increasing sample size. However, there is no significant change in the other parameters (the variance σ_β and the range ϕ). The number of covariates d mainly influences the parameters σ_β and ϕ, which describe the spatial structure of the d-dimensional field β, and as d increases, the estimates converge to the actual values. As for the diagonal case, the EM algorithm underestimates the parameters σ_β and ϕ when the variance σ^2 increases. Finally, for the CAR case, the sample size

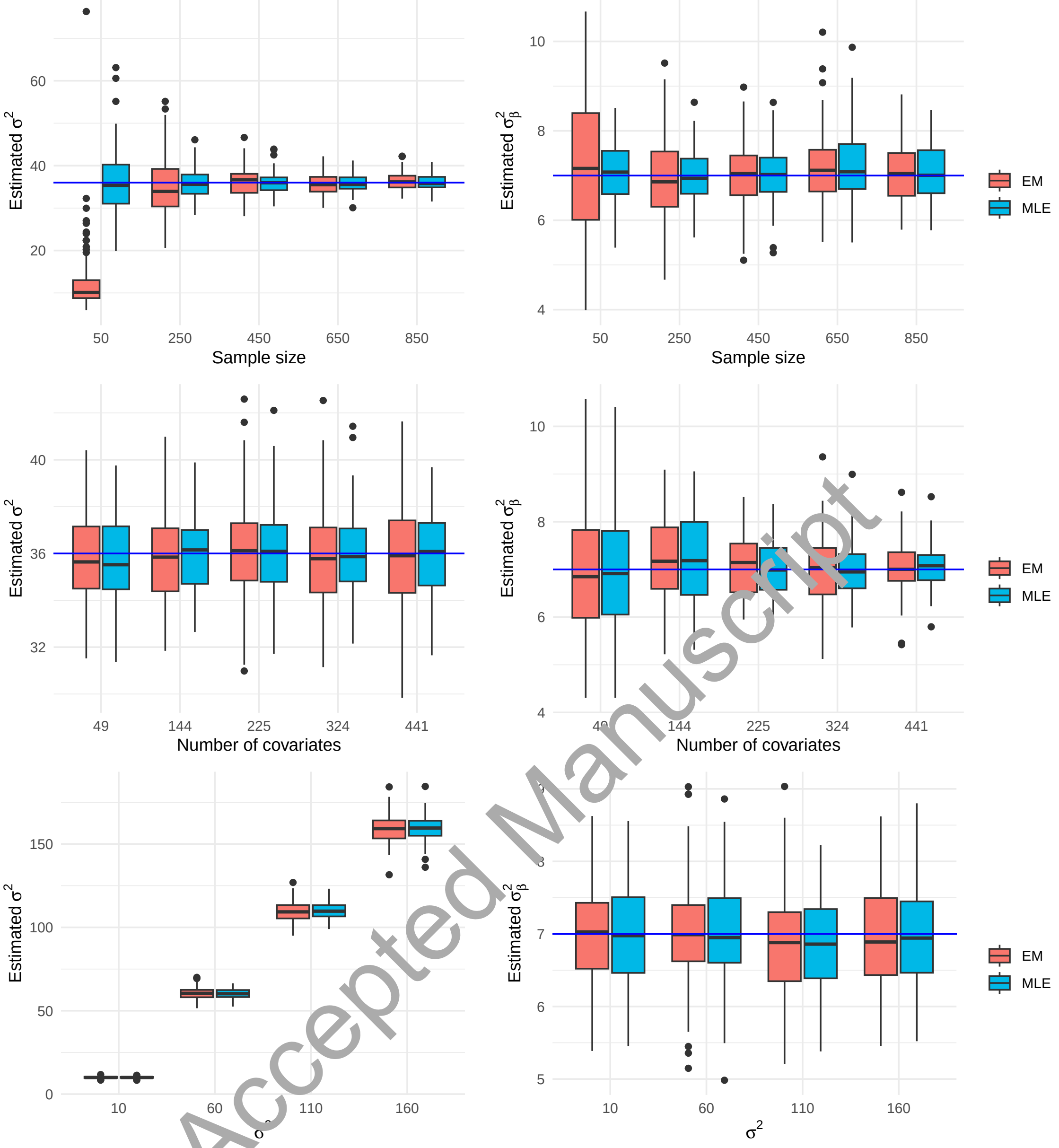

FIGURE 3 Estimated parameters in the case where the covariance of β is diagonal as a function of the sample size (with a fixed number of covariates $d = 225$), the number of covariates d (with a fixed sample size $n = 800$), and the variance σ^2 (with a fixed sample size $n = 800$ and number of covariates $d = 225$). Red boxes correspond to EM estimates and the blue ones to MLE estimates. The blue line corresponds to the true value of the parameter σ^2 and σ^2_β, which are equal to 36 and 7, respectively.

influences the parameters σ^2 and τ^2, but only slightly the correlation parameter α, which is mainly influenced by the number of covariates d. The variance σ^2 has a significant influence on τ^2, but only a small one on α.

Notice that in the high-dimensional scenario, the diagonal covariance underestimates the variance of residuals, which is not the case for the Matérn and CAR covariance cases. This underestimation and uncertainty in the diagonal case is a clear indication of overfitting, as the parameter space in this scenario is notably less constrained. However, it is important to note that this uncertainty in the high-dimensional case is generally more pronounced when no regularization is applied. If no regularization is applied in the high-dimensional case, there is not a unique solution to the regression problem given that the matrix $X^T X$ in equation (6) is not invertible. By introducing proper regularization through the covariance Σ_θ, the matrix $(X^T X + \sigma^2 \Sigma_\theta^{-1})$ in equation (6) becomes invertible, enabling a closed solution for Ridge regression.

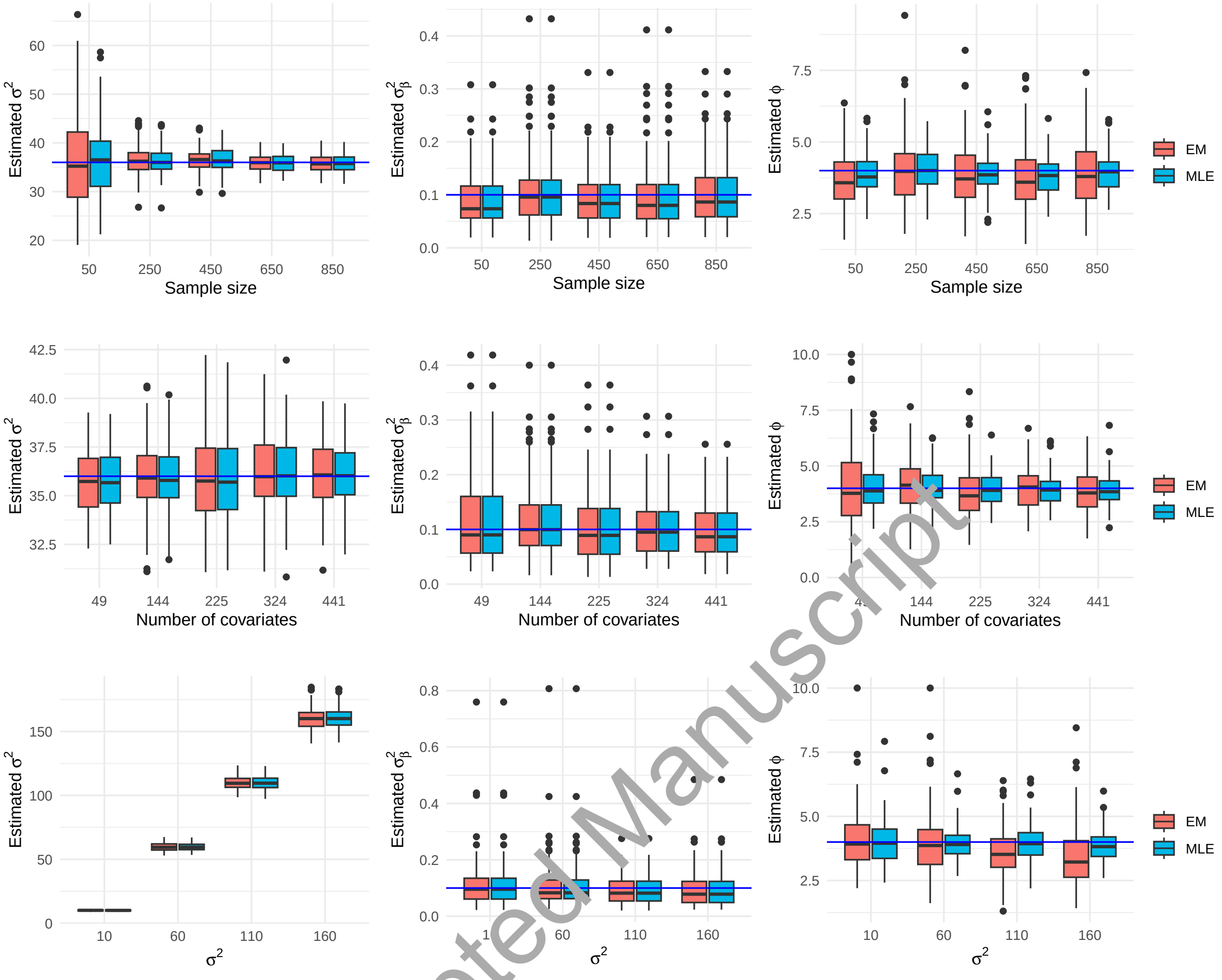

FIGURE 4 Estimated parameters in the case where the covariance of β is the Matérn as a function of the sample size (with a fixed number of covariates $d = 225$), the number of covariates d (with a fixed sample size $n = 800$), and the variance σ^2 (with a fixed sample size $n = 800$ and number of covariates $d = 225$)
. Red boxes correspond to EM estimates and the blue ones to MLE estimates. The blue line corresponds to the true value of the parameter σ^2, σ^2_β, and ϕ, which are equal to 36, 0.1, and 4, respectively.

To summarize:

- The sample size n mainly influences the estimation of the variance of the residuals σ^2.
- The parameters which describe the spatial structure of β are mainly influenced by the number of covariates d.
- As the variance σ^2 increases, EM underestimates the parameter σ^2_β of the diagonal and Matérn case, and the range parameter ϕ.
- EM estimates are close to MLE estimates in most cases when the sample size and the number of covariates d are large enough and the variance σ^2 is small.

Another interesting aspect worth investigating involves simulating coefficients β using one covariance model and estimating them using a different covariance model. The objective is to observe the sensitivity of estimation when the true covariance of the coefficients differs from the one used for model estimation. In this regard, we conducted 100 independent simulations of β utilizing the Matérn covariance function, and subsequently estimated the parameters utilizing three distinct cases: diagonal,

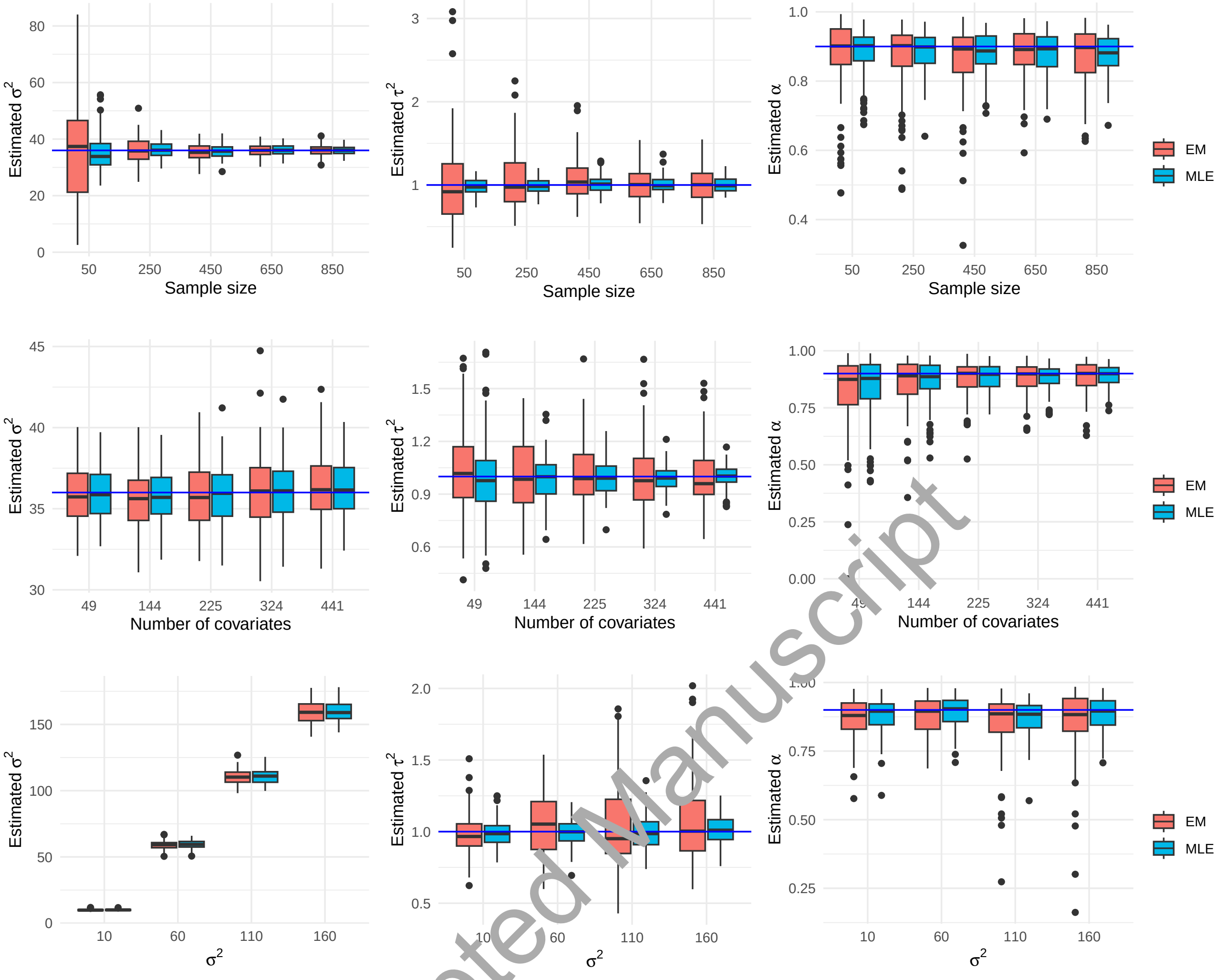

FIGURE 5 Estimated parameters in the case where the covariance of β is the CAR as a function of the sample size (with a fixed number of covariates $d = 225$), the number of covariates d (with a fixed sample size $n = 800$), and the variance σ^2 (with a fixed sample size $n = 800$ and number of covariates $d = 225$). Red boxes correspond to EM estimates and the blue ones to MLE estimates. The blue line corresponds to the true value of the parameter σ^2, σ_β^2, and α, which are equal to 36, 1, and 0.9, respectively.

Conditional Autoregressive (CAR), and Matérn covariances. The outcomes of this experiment are presented in Figure 6 . It is clear that using the Matérn covariance for the estimation gives better results in terms of $NRMSE_\beta$. Not surprisingly, the diagonal case is the worst model for estimating the coefficients. However, in terms of $NRMSE_y$, there is a small difference between the three methods.

4 | APPLICATION

The proposed method is applied to the problem of predicting the significant wave height (H_s) at a location in the Bay of Biscay using wind conditions over the North Atlantic (Figure 7), where the significant wave height is the average height of the highest third of the waves, a key measure of wave height that provides information about wave energy. In the Bay of Biscay, wave behavior is influenced by both local and global wind conditions, with some swells originating as far away as Cape Hatteras (Ardhuin & Orfila 2018). To accurately predict H_s,it is crucial to consider a large area covered by the swell generation (Obakrim et al. 2023).

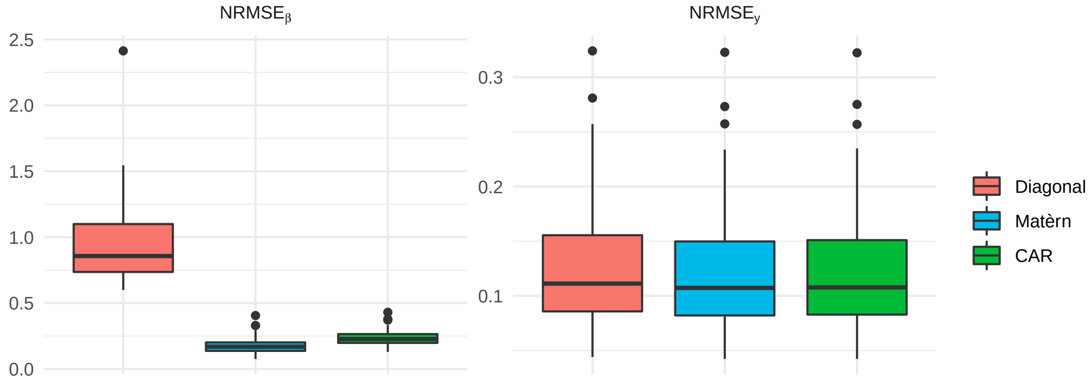

FIGURE 6 Results of the estimations when the true beta is simulated from Matérn with the parameters $\sigma^2 = 36$, $\sigma_\beta^2 = 0.1$ and $\phi = 4$ and sample size $n = 800$ and number of covariates $d = 225$. The left panel correspond to $NRMSE_\beta$ and the right one to $NRMSE_y$.

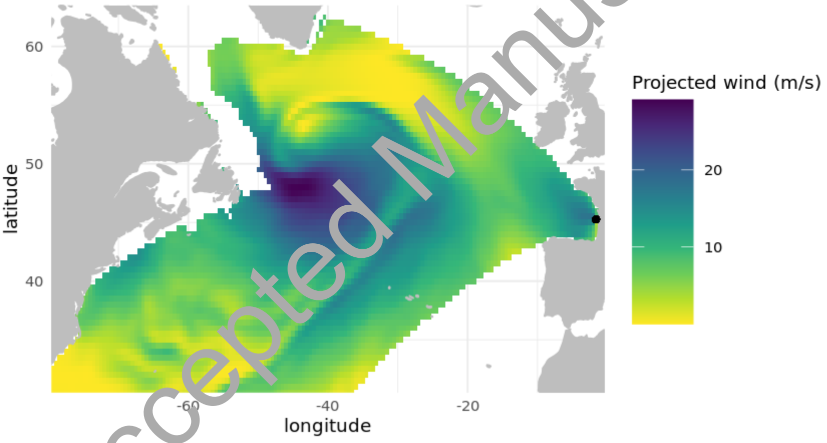

FIGURE 7 CFSR projected wind in the North Atlantic in 1994-01-01 00h:00. The black point represents the target point.

The classical method used to define a relevant spatial domain for the predictors, is to fit different models at different radii around the target point and select the optimal model, as shown in Michel et al. (2022); however, it could be computationally intensive to test such a number of models. One significant advantage of regularization methods, such as generalized Ridge, is their capacity to incorporate numerous covariates into the model while penalizing those that contribute minimally to the prediction (Hastie et al. 2009). For this study, we will therefore utilize a broad spatial domain, following previous studies Michel et al. (2022); Obakrim et al. (2023), for predicting wave height at the target location in the Bay of Biscay and use generalized ridge regression, allowing the regularization technique to automatically prioritize the most influential locations in terms of wave energy at the target point.

The data used for H_s comes from the Homere hindcast database (Boudière et al. 2013), and the wind data comes from Climate Forecast System Reanalysis (CFSR) (Saha et al. 2010). The wind data are pre-processed before being used as a predictor (see Obakrim et al. (2023) for the pre-processing procedure). We consider 23 years of H_s and wind data from 1994 to 2016 with a temporal resolution of 3 hours.

Method	r	RMSE(m)	bias(m)	Computation time (min)
Diagonal	0.941	0.414	-0.0004	16.77
Matérn	0.956	0.354	-0.04	40.9
CAR	0.957	0.352	-0.06	30.06

TABLE 1 Quantitative comparison of the diagonal, Matérn, and CAR methods in the validation set using the correlation (r), root mean square error (RMSE), bias, and computation time.

The regression problem is of the form

$$H_s(t) = \sum_{j=1}^{d} X_j(t)\beta_j + \epsilon(t) \quad t = 1, ..., n \tag{27}$$

where $X_j(t)$ is the predictor at time t and location j defined as

$$X_j(t; t_j, \alpha_j) = \tfrac{1}{2\alpha_j+1} \textstyle\sum_{i=t-t_j-\alpha_j}^{t-t_j+\alpha_j} W_j^2(i), \tag{28}$$
$$t_j + \alpha_j + 1 \leq t \leq t_j - \alpha_j + n$$

where W_j is the projected wind (figure 7) defined as

$$W_j = U_j\, cos\left(\frac{1}{2}(b_j - \theta_j)\right) \tag{29}$$

U_j is the wind speed, b_j is the great circle bearing, and θ_j is the wind direction at location j. α_j controls the length of the time window, and t_j is the mean travel time of waves which are estimated using the maximum correlation between H_s and the predictor

$$(\hat{t}_j, \hat{\alpha}_j) = arg \max_{t_j, \alpha_j} \left(corr(H_s, X_j^g(t_j, \alpha_j))\right) \tag{30}$$

where $\hat{t}_j$ and $\hat{\alpha}_j$ are the estimated t_j and α_j, respectively. Let $X = \{X_1, ..., X_d\}$ denote the predictor variables, with d representing the number of spatial locations, specifically $d = 5651$ in this case. The dataset is comprised of a sample size of $n = 67088$. Since the predictor has a spatial structure, it is reasonable to assume that the coefficients β also have a spatial structure so that nearby locations have close contributions to the waves at the target point. This assumption is equivalent to suppose that $\beta \sim \mathcal{N}(0, \Sigma_\theta)$. For the covariance Σ_θ, we will consider the cases of Matérn and CAR. For comparison, we also consider the diagonal case even though it does not consider any structure between coefficients.

The model's parameters (equation 27) are estimated using data from 1994 to 2013, and the model is evaluated in terms of correlation, RMSE, and bias, using a validation set from 2014 to 2016. Figure 8 shows the estimated posterior mean $\hat{\mu}_{\beta|y}$ of the coefficients β, defined in equation (6) and its corresponding lower and upper bound prediction interval Lower $\hat{\mu}_{\beta|y}$ and Upper $\hat{\mu}_{\beta|y}$, respectively, for the diagonal, Matérn, and CAR cases. The lower and upper bound prediction interval are defined as follows:

$$\text{Lower } \hat{\mu}_{\beta|y} = \hat{\mu}_{\beta|y} - 1.96 * \sqrt{\text{diag}(\hat{\Sigma}_{\beta|y})} \tag{31}$$
$$\text{Upper } \hat{\mu}_{\beta|y} = \hat{\mu}_{\beta|y} + 1.96 * \sqrt{\text{diag}(\hat{\Sigma}_{\beta|y})}$$

where $\text{diag}(\hat{\Sigma}_{\beta|y})$ is the diagonal of the estimated posterior covariance matrix of β defined in equation (6). Not surprisingly, the coefficients estimated with the diagonal covariance display no discernible physical spatial structure. Therefore, the assumption that close locations have close coefficients cannot be taken into account using the diagonal case. This motivates using the Matérn and CAR covariances. The Matérn and CAR covariances yield smoother coefficients with evident spatial structure. Moreover, coefficients for locations in proximity to the target point are notably larger, aligning with our prior assumption regarding the covariance. In addition, the prediction interval for β is relatively narrow, implying that the uncertainty surrounding the estimated coefficients is relatively low. It is worth noting that the CAR method incurs lower numerical computational costs compared to Matérn, particularly in terms of inverting the covariance matrix during each iteration of the optimization algorithm employed in the M-step.

Table 1 shows the results of the quantitative comparison between the three methods for predicting significant wave height in the validation set using correlation (r), root mean square error (RMSE), bias, and computation time in minutes. The computation

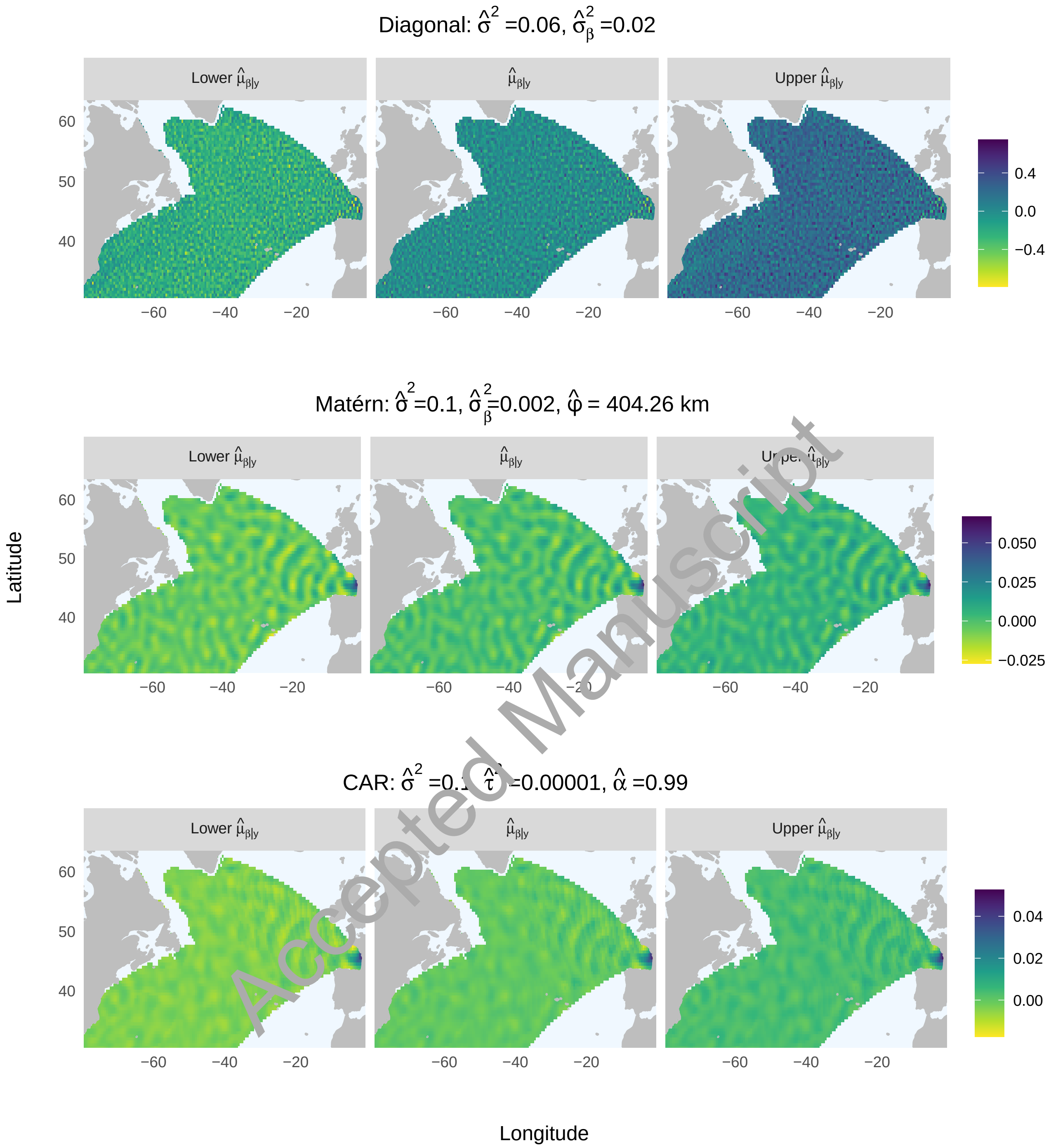
Diagonal: $\hat{\sigma}^2$ =0.06, $\hat{\sigma}^2_\beta$ =0.02
Lower $\hat{\mu}_{\beta|y}$
$\hat{\mu}_{\beta|y}$
Upper $\hat{\mu}_{\beta|y}$
0.4
0.0
−0.4
Matérn: $\hat{\sigma}^2$=0.1, $\hat{\sigma}^2_\beta$=0.002, $\hat{\phi}$ = 404.26 km
0.050
0.025
0.000
−0.025
CAR: $\hat{\sigma}^2$ =0.1, $\hat{\tau}^2$ =0.00001, $\hat{\alpha}$ =0.99
0.04
0.02
0.00
60
50
40
−60
−40
−20
Latitude
Longitude

FIGURE 8 The posterior mean estimate for β and its corresponding lower and upper bound prediction interval, denoted as Lower $\hat{\mu}_{\beta|y}$ and Upper $\hat{\mu}_{\beta|y}$, respectively, for the diagonal, Matérn, and CAR cases.

times were measured on a computer equipped with 28 cores and 115GB of RAM. In terms of correlation and RMSE, the diagonal method is the less accurate method. Therefore, adding the spatial structure in the covariance is advantageous in predicting the significant wave height. The CAR and Matérn methods lead to close results regarding r, RMSE, and bias. In Figure 9 , the time series of observed and predicted significant wave height (H_s) for February 2014 are showed for three distinct cases, along with the corresponding 95% prediction interval. The coverage probability is found to be lowest for the diagonal case. This is particularly evident during the event around 22 February, where the predicted values underestimate the observed H_s.

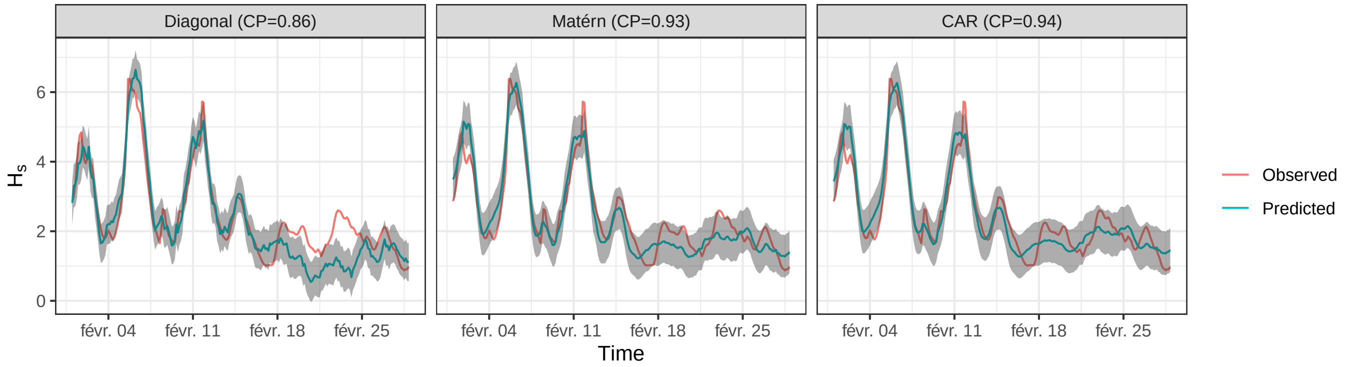

FIGURE 9 Time series of observed versus predicted H_s at the target location in February 2014 for the three cases. The gray shadow corresponds to the 95% prediction interval and CP is the coverage probability.

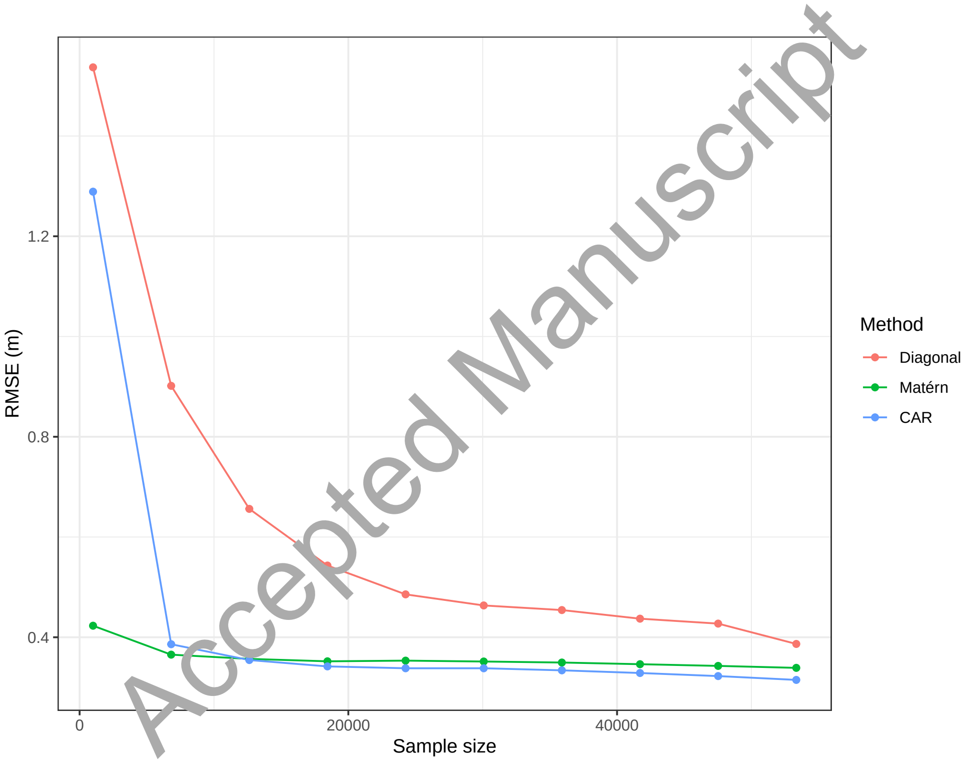

FIGURE 10 Root Mean Squared Error (RMSE) of the model, considering the three cases (diagonal, Matérn, and CAR), in predicting H_s within the validation set, plotted against the sample size utilized for training the model.

The diminished coverage probability in the diagonal case can be attributed to its limited capability to accurately predict swells originating from remote regions, primarily due to its neglect of spatial dependencies between neighboring locations.

In the simulation study, we studied the effect of the sample size on the estimation and we will use the same procedure in this application. We vary the sample size used for training, ranging from 1000 to 54000, , and analyzed its influence on predicting H_s in the validation period. The results of the experiments are showed in figure 10 . The diagonal covariance struggles with the challenges posed by high dimensionality, with the RMSE showing a consistent decline as the sample size grows. In the case of the CAR model, performance is also suboptimal, albeit better than the diagonal case in high dimensions. However, the RMSE for the CAR model decreases rapidly and stabilizes as the sample size reaches 6815. On the contrary, the Matérn model is less sensitive to the sample size, maintaining a relatively stable RMSE which aligns with the findings from the simulation experiments presented in section 3.2.

5 | CONCLUSIONS

This study proposes an EM algorithm for estimating generalized Ridge regression with spatial covariates. We study three cases: the diagonal, Matérn, and the CAR case. A simulation study is carried out to evaluate the performance of the algorithms, and the EM algorithm successfully estimates the parameters in all cases. We study the influence of the sample size, number of covariates, and the variance σ^2 on the estimation. The sample size mainly influences the variance parameter σ^2. The range parameter of the Matérn and the correlation parameter of the CAR are mainly influenced by the number of covariates d.

The proposed method is applied to the problem of downscaling the significant wave height in the Bay of Biscay using wind conditions over the North Atlantic. The Matérn method gives smooth coefficients with a clear spatial structure; however, CAR and Matérn methods lead to close results regarding r, RMSE, and bias. The Matérn covariance is clearly a better choice for spatial applications. However, estimating the parameters requires the inversion of the covariance matrix at each iteration of the optimization method in the M-step, which may be a computational bottleneck in many applications. To address this issue, instead of parameterizing the covariance matrix, one can parameterize the precision matrix directly as we do with the CAR method.

□

APPENDIX

A COMPARISON BETWEEN CROSS-VALIDATION AND EM

As stated in section 2, the EM algorithm can be used to estimate the regularization parameter in the classical Ridge regression framework. In this section, we compare the EM algorithm results with those obtained using least squares Ridge regression method with cross-validation (CV). The latter employs k-fold cross-validation to estimate the regularization parameter using *cv.glmnet* function in R (Friedman et al. 2023). We examine two distinct scenarios for the covariates: one where the covariates exhibit strong correlations and another where the correlations are weak. In the first case the covariates are simulated using a Matérn covariance characterized by parameters $(\sigma_x^2, \phi_x, \kappa_x) = (6, 2, 3/2)$, while the second scenario employs parameters $(\sigma_x^2, \phi_x, \kappa_x) = (6, 0.02, 3/2)$. Furthermore, we investigate two scenarios concerning the residuals of the regression model: one where the residuals are Gaussian distributed, similar to previous simulations, and another where the residuals do not follow a Gaussian distribution. In the latter case, we simulate the response variable Y according to the model:

$$Y = X\beta + \epsilon, \quad \text{where } \epsilon \sim U(2, 30). \tag{A.1}$$

Thus, we account for a total of four distinct scenarios. For each scenario, we conduct 100 independent random samples of coefficients β using the diagonal method, with parameters $\sigma^2 = 36$ and $\sigma_\beta^2 = 7$. In each simulation, we estimate the coefficients utilizing both the EM algorithm and the *cv.glmnet* method.

Figures A1 and A2 show the box plot of $NRMSE_\beta$ and $NRMSE_y$ in the Gaussian and non Gaussian cases, respectively, for the two scenarios of the covariates. Significant difference between the two methods is noticed, demonstrating better estimations by the EM algorithm for both β and the response variable, when the covariates are highly correlated. Conversely, when the covariate are weakly correlated, the distinction in performance is marginal or absent, particularly in the non Gaussian scenario. This means that the EM algorithm is less sensible by the multicolinearity problem than k-fold cross-validation, which may be explained by variability of cross-validation determining the best regularization parameter due to fold assignment (Algamal 2020).

B THE CASE WHERE β HAS A NON-ZERO MEAN

In this section, we consider the case where β has a non-zero mean as defined by the hierarchically model

$$\begin{aligned} \beta &\sim \mathcal{N}(\mu_\xi, \Sigma_\theta) \\ Y \mid \beta, \Theta &\sim \mathcal{N}(X\beta, \sigma^2 I_n) \end{aligned} \tag{B.1}$$

where $\Theta = (\sigma^2, \mu_\xi, \theta)$.

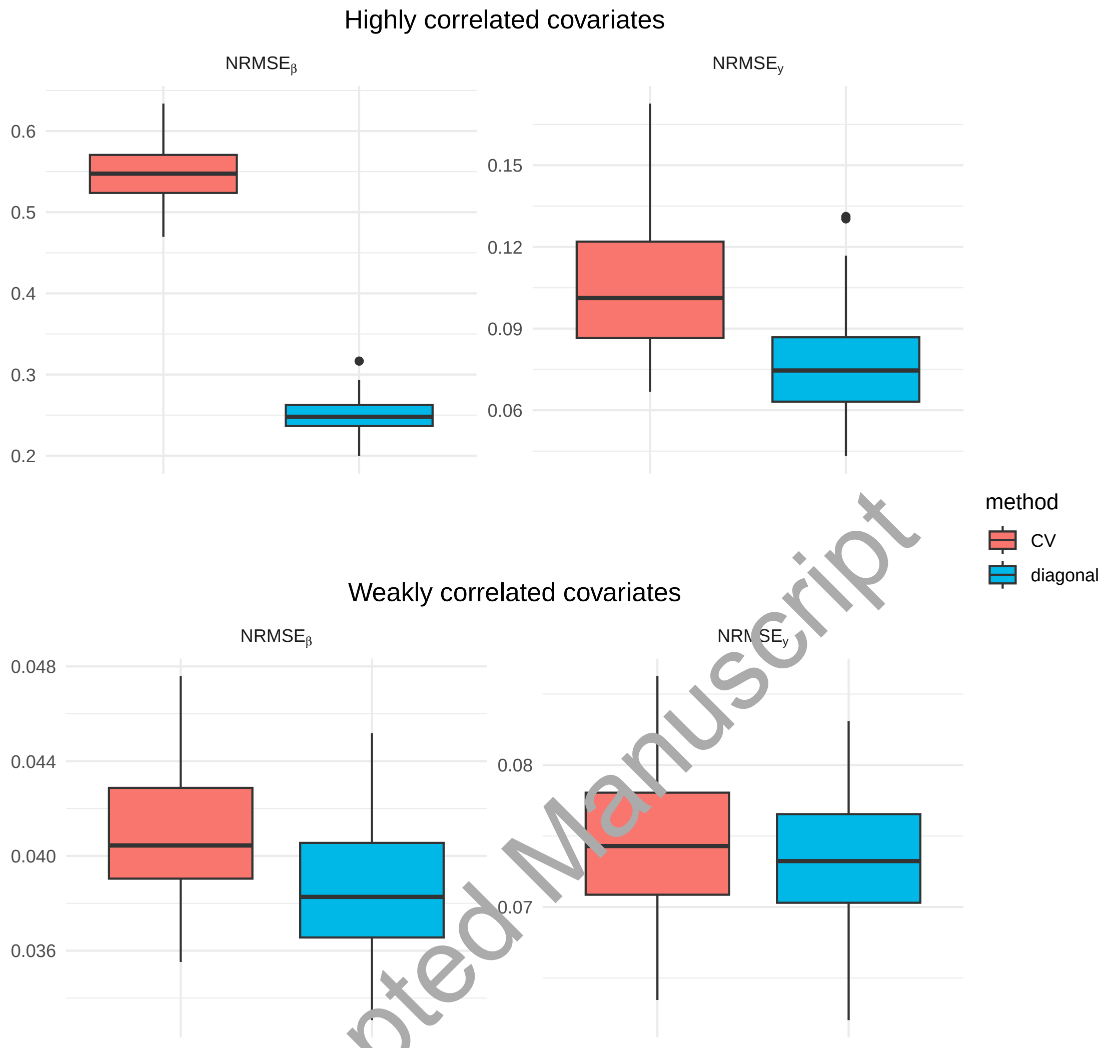

FIGURE A1 Results of estimating Ridge regression with the EM algorithm and 10-fold cross-validation in the Gaussian case when the covariates are highly correlated (top panel) and weakly correlated (bottom panel).

The complete log-likelihood is expressed as

$$\begin{aligned}\ln p(y,\beta;\Theta) &= \ln p(y \mid \beta;\sigma^2) + \ln p(\beta;\theta)\\ &= -\frac{1}{2}\left(n\ln(\sigma^2) + \frac{1}{\sigma^2}\|y + X\beta\|^2 + \ln(|\Sigma_\theta|) + \beta^T\Sigma_\theta^{-1}\beta - 2\beta^T\Sigma_\theta^{-1}\mu_\xi + \mu_\xi^T\Sigma_\theta^{-1}\mu_\xi\right) + C\end{aligned} \tag{B.2}$$

Where C is a constant. In the M-step, the quantity $Q(\Theta|\Theta^{(t)})$ is maximized with respect to the parameters Θ.

- E-step:

$$Q(\Theta|\Theta^{(t)}) = \mathbb{E}(\ln p(y,\beta;\Theta) \mid y,\Theta^{(t)}). \tag{B.3}$$

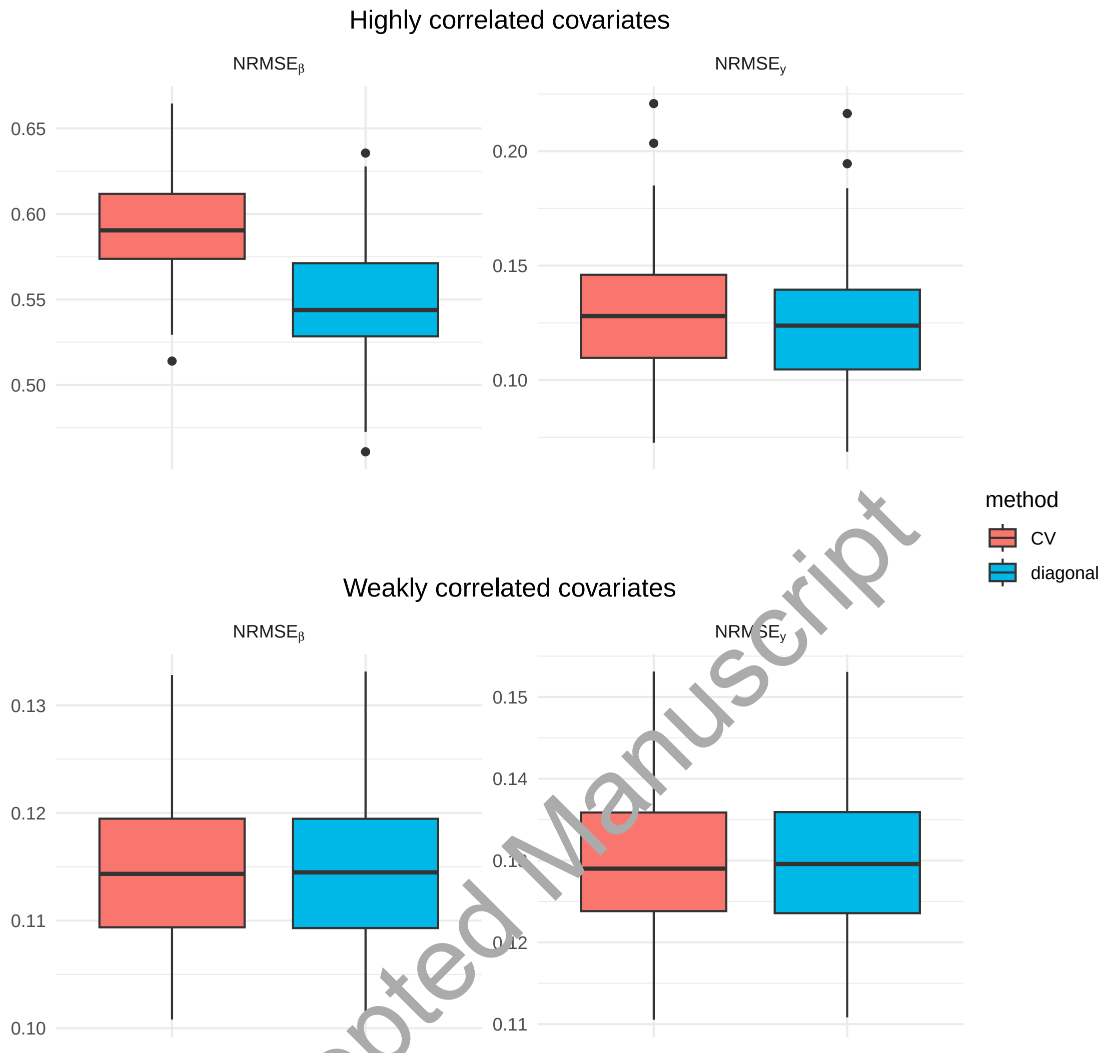

FIGURE A2 Results of estimating Ridge regression with the EM algorithm (diagonal) and 10-fold cross-validation (CV) in the non Gaussian case when the covariates are highly correlated (top panel) and weakly correlated (bottom panel).

The posterior distribution of the latent variable β is a normal distribution with mean $\mu_{\beta|y}$ and covariance matrix $\Sigma_{\beta|y}$ such that

$$\begin{cases} \Sigma_{\beta|y} = (\Sigma_\theta^{-1} + \frac{1}{\sigma^2}X^TX)^{-1} \\ \mu_{\beta|y} = \Sigma_{\beta|y}(\Sigma_\theta^{-1}\mu_\xi + \frac{1}{\sigma^2}X^Ty). \end{cases} \tag{B.4}$$

Therefore,

$$Q(\Theta|\Theta^{(t)}) = -\frac{1}{2}\left(\ln(|\Sigma_\theta|) + \mathrm{Tr}(\Sigma_\theta^{-1}\mathbb{E}(\beta\beta^T \mid y,\Theta^{(t)})) - 2\mu_{\beta|y}^T\Sigma_\theta^{-1}\mu_\xi + \mu_\xi^T\Sigma_\theta^{-1}\mu_\xi + n\ln(\sigma^2) + \frac{1}{\sigma^2}\mathbb{E}(\|y - X\beta\|^2 \mid y,\Theta^{(t)})\right) + C \tag{B.5}$$

where

$$\begin{cases} \mathbb{E}(\beta\beta^T|y;\Theta^{(t)}) = \Sigma_{\beta|y} + \mu_{\beta|y}\mu_{\beta|y}^T \\ \mathbb{E}(\|y - X\beta\|^2|y;\Theta^{(t)}) = \|y\|^2 - 2y^TX\mu_{\beta|y} + \mathrm{Tr}(X^TX\mathbb{E}(\beta\beta^T|y;\Theta^{(t)})) \end{cases} \tag{B.6}$$

- M-step:

The maximization step computes

$$\Theta^{(t+1)} = \arg\max_{\Theta} Q(\Theta|\Theta^{(t)}) \tag{B.7}$$

which leads to the following updates of the parameters

$$\begin{aligned}\sigma^{2,(t+1)} &= \frac{1}{n}(\|y\|^2 - 2y^T X \mu_{\beta|y} + \mathrm{Tr}(X^T X \mathbb{E}(\beta\beta^T | y; \Theta^{(t)}))) \\ (\xi^{(t+1)}, \theta^{(t+1)}) &= \arg\max_{\xi,\theta} \ \ln(|\Sigma_\theta^{-1}|) - \mathrm{Tr}(\Sigma_\theta^{-1} \mathbb{E}(\beta\beta^T \mid y, \Theta^{(t)})) + 2\mu_{\beta|y}^T \Sigma_\theta^{-1} \mu_\xi^{(t)} - \mu_{\xi^{(t)}}^T \Sigma_\theta^{-1} \mu_\xi^{(t)}\end{aligned} \tag{B.8}$$

References

Abramowitz, M., Stegun, I. A., & Romer, R. H. (1988). *Handbook of mathematical functions with formulas, graphs, and mathematical tables.* American Association of Physics Teachers.

Algamal, Z. Y. (2020). Shrinkage parameter selection via modified cross-validation approach for ridge regression model. *Communications in Statistics-Simulation and Computation*, *49*(7), 1922–1930.

Allen, D. M. (1974). The relationship between variable selection and data agumentation and a method for prediction. *technometrics*, *16*(1), 125–127.

Ambikasaran, S., Foreman-Mackey, D., Greengard, L., Hogg, D. W., & O'Neil, M. (2015). Fast direct methods for gaussian processes. *IEEE transactions on pattern analysis and machine intelligence*, *38*(2), 252–265.

Andrews, J. L. (2018). Addressing overfitting and underfitting in gaussian model-based clustering. *Computational Statistics & Data Analysis*, *127*, 160–171.

Ardhuin, F., & Orfila, A. (2018). Wind waves. *New Frontiers in Operational Oceanography*, 393–422.

Bachoc, F. (2013). *Parametric estimation of covariance function in gaussian process based kriging models. application to uncertainty quantification for computer experiments* (Unpublished doctoral dissertation). Université Paris-Diderot-Paris VII.

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), 1–20.

Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning* (Vol. 4) (No. 4). Springer.

Boonstra, P. S., Mukherjee, B., & Taylor, J. M. (2015). A small-sample choice of the tuning parameter in ridge regression. *Statistica Sinica*, *25*(3), 1185.

Boudière, E., Maisondieu, C., Ardhuin, F., Accensi, M., Pineau-Guillou, L., & Lepesqueur, J. (2013). A suitable metocean hindcast database for the design of marine energy converters. *International Journal of Marine Energy*, *3*, e40–e52.

Camus, P., Losada, I., Izaguirre, C., Espejo, A., Menendez, M., & Pérez, J. (2017). Statistical wave climate projections for coastal impact assessments. *Earth's Future*, *5*(9), 918–933.

Charles, E., Idier, D., Delecluse, P., Déqué, M., & Le Cozannet, G. (2012). Climate change impact on waves in the bay of biscay, france. *Ocean Dynamics*, *62*(6), 831–848.

Chu, T., Zhu, J., & Wang, H. (2011). Penalized maximum likelihood estimation and variable selection in geostatistics. *The Annals of Statistics*, *39*(5), 2607–2625.

Cressie, N., & Johannesson, G. (2008). Fixed rank kriging for very large spatial data sets. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *70*(1), 209–226.

Cressie, N., & Kapat, P. (2008). Some diagnostics for markov random fields. *Journal of computational and graphical statistics*, *17*(3), 726–749.

Cressie, N., & Wikle, C. K. (2015). *Statistics for spatio-temporal data.* John Wiley & Sons.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1–22.

Fassò, A., & Finazzi, F. (2011). Maximum likelihood estimation of the dynamic coregionalization model with heterotopic data. *Environmetrics*, *22*(6), 735–748.

Friedman, J., Hastie, T., Tibshirani, R., Narasimhan, B., Tay, K., Simon, N., … Yang, J. (2023). glmnet: Lasso and elastic-net regularized generalized linear models. *Astrophysics Source Code Library*, ascl–2308.

Goeman, J. J. (2008). Autocorrelated logistic ridge regression for prediction based on proteomics spectra. *Statistical Applications in Genetics and Molecular Biology*, *7*(2).

Golub, G. H., Heath, M., & Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, *21*(2), 215–223.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference,* *and prediction* (Vol. 2). Springer.

Heaton, M. J., Datta, A., Finley, A. O., Furrer, R., Guinness, J., Guhaniyogi, R., … others (2019). A case study competition among methods for analyzing large spatial data. *Journal of Agricultural, Biological and Environmental Statistics*, *24*, 398–425.

Hemmerle, W. J. (1975). An explicit solution for generalized ridge regression. *Technometrics*, *17*(3), 309–314.

Hessami, M., Gachon, P., Ouarda, T. B., & St-Hilaire, A. (2008). Automated regression-based statistical downscaling tool. *Environmental modelling & software*, *23*(6), 813–834.

Kaufman, C., & Shaby, B. A. (2013). The role of the range parameter for estimation and prediction in geostatistics. *Biometrika*, *100*(2), 473–484.

Liang, J., Cheng, Y., Su, Y., Xiao, S., & Song, Y. (2022). Variable selection for spatial logistic autoregressive models. *Mathematics*, *10*(17), 3095.

Maranzano, P., Otto, P., & Fassò, A. (2023). Adaptive lasso estimation for functional hidden dynamic geostatistical models. *Stochastic Environmental Research and Risk Assessment*, 1–23.

Meyer, H., Reudenbach, C., Wöllauer, S., & Nauss, T. (2019). Importance of spatial predictor variable selection in machine learning applications–moving from data reproduction to spatial prediction. *Ecological Modelling*, *411*, 108815.

Michel, M., Obakrim, S., Raillard, N., Ailliot, P., & Monbet, V. (2022). Deep learning for statistical downscaling of sea states. *Advances in Statistical Climatology, Meteorology and Oceanography*, *8*(1), 83–95.

Obakrim, S., Ailliot, P., Monbet, V., & Raillard, N. (2023). Statistical modeling of the space–time relation between wind and significant wave height. *Advances in Statistical Climatology, Meteorology and Oceanography*, *9*(1), 67–81.

Otto, P., Piter, A., & Gijsman, R. (2021). Statistical analysis of beach profiles–a spatiotemporal functional approach. *Coastal engineering*, *170*, 103999.

Patil, P., Wei, Y., Rinaldo, A., & Tibshirani, R. (2021). Uniform consistency of cross-validation estimators for high-dimensional ridge regression. In *International conference on artificial intelligence and statistics* (pp. 3178–3186).

Permatasari, S. M., Djuraidah, A., & Soleh, A. M. (2017). Statistical downscaling with gamma distribution and elastic net regularization: Case study: Monthly rainfall 1981-2013 at indramayu. In *The 2nd international conference on applied statistics (icas 2016)* (pp. 121–129).

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., et al. (2019). Deep learning and process understanding for data-driven earth system science. *Nature*, *566*(7743), 195–204.

Rue, H. (2001). Fast sampling of gaussian markov random fields. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *63*(2), 325–338.

Rue, H., & Tjelmeland, H. (2002). Fitting gaussian markov random fields to gaussian fields. *Scandinavian journal of Statistics*, *29*(1), 31–49.

Saha, S., Moorthi, S., Pan, H.-L., Wu, X., Wang, J., Nadiga, S., … others (2010). The ncep climate forecast system reanalysis. *Bulletin of the American Meteorological Society*, *91*(8), 1015–1058.

Schulz, E., Speekenbrink, M., & Krause, A. (2018). A tutorial on gaussian process regression: Modelling, exploring, and exploiting functions. *Journal of Mathematical Psychology*, *85*, 1–16.

Shinozaki, T., & Ostendorf, M. (2008). Cross-validation and aggregated em training for robust parameter estimation. *Computer Speech & Language*, *22*(2), 185–195.

Stevens, A., Willett, R., Mamalakis, A., Foufoula-Georgiou, E., Tejedor, A., Randerson, J. T., … Wright, S. (2021). Graph-guided regularized regression of pacific ocean climate variables to increase predictive skill of southwestern us winter precipitation. *Journal of climate*, *34*(2), 737–754.

Storkey, A. J. (1999). Truncated covariance matrices and toeplitz methods in gaussian processes. In *1999 ninth international conference on artificial neural networks icann 99.(conf. publ. no. 470)* (Vol. 1, pp. 55–60).

Sungkawa, I., Rahayu, A., et al. (2019). Extreme rainfall prediction using bayesian quantile regression in statistical downscaling modeling. *Procedia Computer Science*, *157*, 406–413.

Tajbakhsh, S. D., Aybat, N. S., & Del Castillo, E. (2020). On the theoretical guarantees for parameter estimation of gaussian random field models: A sparse precision matrix approach. *Journal of Machine Learning Research*, *21*(217), 1–41.

Takenouchi, T., & Ikeda, K. (2010). Theoretical analysis of cross-validation (cv)-em algorithm. In *International conference on artificial neural networks* (pp. 321–326).

Tew, S. Y., Boley, M., & Schmidt, D. (2024). Bayes beats cross validation: Efficient and accurate ridge regression via expectation maximization. *Advances in Neural Information Processing Systems*, *36*.

Tew, S. Y., Schmidt, D. F., & Makalic, E. (2022). *Sparse horseshoe estimation via expectation-maximisation.*

Tian, G., Xia, Y., Zhang, Y., & Feng, D. (2011). Hybrid genetic and variational expectation-maximization algorithm for gaussian-mixture-model-based brain mr image segmentation. *IEEE transactions on information technology in biomedicine*, *15*(3), 373–380.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., & Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(1), 91–108.

van Wieringen, W. N. (2015). Lecture notes on ridge regression. *arXiv preprint arXiv:1509.09169*.

Zhang, H. (2007). Maximum-likelihood estimation for multivariate spatial linear coregionalization models. *Environmetrics: The official journal of the International Environmetrics Society*, *18*(2), 125–139.