Semi-Implicit Approaches for Large-Scale Bayesian Spatial Interpolation

Sébastien Garneau*
Department of Epidemiology, Biostatistics and Occupational Health,
McGill University, sebastien.garneau@mail.mcgill.ca
and
Carlos T.P. Zanini
Department of Statistical Methods,
Federal University of Rio de Janeiro
and
Alexandra M. Schmidt
Department of Epidemiology, Biostatistics and Occupational Health,
McGill University

October 23, 2025

*Dr. Schmidt is grateful for financial support from the Natural Sciences and Engineering Research Council (NSERC) of Canada (Discovery Grant RGPIN-2024-04312) and the endowed University Chair, McGill University.

Abstract

Spatial statistics often rely on Gaussian processes (GPs) to capture dependencies across locations. However, their computational cost increases rapidly with the number of locations, potentially needing multiple hours even for moderate sample sizes. To address this, we propose using Semi-Implicit Variational Inference (SIVI), a highly flexible Bayesian approximation method, for scalable Bayesian spatial interpolation. We evaluated SIVI with a GP prior and a Nearest-Neighbour Gaussian Process (NNGP) prior compared to Automatic Differentiation Variational Inference (ADVI), Pathfinder, and Hamiltonian Monte Carlo (HMC), the reference method in spatial statistics. Methods were compared based on their predictive ability measured by the CRPS, the interval score, and the negative log-predictive density across 50 replicates for both Gaussian and Poisson outcomes. SIVI-based methods achieved similar results to HMC, while being drastically faster. On average, for the Poisson scenario with 500 training locations, SIVI reduced the computational time from roughly 6 hours for HMC to 130 seconds. Furthermore, SIVI-NNGP analyzed a simulated land surface temperature dataset of 150,000 locations while estimating all unknown model parameters in under two minutes. These results highlight the potential of SIVI as a flexible and scalable inference technique in spatial statistics.

Keywords: Dimension Reduction, Exponential Family, Gaussian Process, Variational Inference

1 Introduction

Easier access to satellite data and Geographic Information Systems (GIS) has led to an increase in geographically referenced data. Such data naturally occur in a variety of fields, such as epidemiology, public health, and environmental sciences (Yu et al. 2017, Blangiardo et al. 2016, Rodriguez-Idiazabal et al. 2025). The core principle of analyzing spatial data can be summarized by Tobler's first law of geography: "everything is related to everything else, but near things are more related than distant things". While this law is a simplification, it captures the fundamental notion of spatial dependence, which is at the heart of spatial statistics.

Gaussian processes (GPs) have been used predominantly to capture such dependence. A GP is a stochastic process for which any finite set of random variables follows a multivariate normal distribution, with mean given by the GP's mean function and covariance given by the GP's covariance function. The covariance function specifies how the correlation between locations decays with distance.

However, the use of a GP comes with an important drawback: evaluating the likelihood of an n-dimensional GP requires $\mathcal{O}(n^3)$ operations and $\mathcal{O}(n^2)$ memory. Hence, models using a GP become computationally intensive even for moderately large datasets. Proposed solutions to this problem include stochastic partial differential equations (Lindgren et al. 2011), low-rank approximations (Banerjee et al. 2008), and using conditional independence structures (Vecchia 1988), among others.

Variational inference (VI), a Bayesian technique that aims to approximate the posterior by a distribution in a pre-specified family of variational distributions, provides a computationally efficient alternative to Markov Chain Monte Carlo (MCMC) for large-scale models that include a GP component, such as those commonly used in spatial statistics. Among existing VI methods for spatial data, most are restricted to Gaussian outcomes (Ren et al. 2011,

Song & Datta 2025), or require specific approximations to be developed to handle specific outcome types (Lee & Lee 2025) . Therefore, there remains a need for VI methods that could theoretically be flexible enough to accommodate a wide range of outcome types.

In this paper, we propose the use of Semi-implicit variational inference (SIVI) (Yin & Zhou 2018) for Bayesian geostatistical models. We evaluate SIVI through simulation studies where we compare it to automatic differentiation variational inference (ADVI) (Kucukelbir et al. 2017) and Pathfinder (Zhang et al. 2022), two commonly used VI algorithms, and a modified Hamiltonian Monte Carlo (HMC) (Neal 2011) algorithm known as the No-U-Turn sampler (NUTS) by Hoffman et al. (2014). To further improve computational efficiency, we also consider the combination of SIVI with the nearest-neighbor Gaussian process (NNGP) (Datta et al. 2016) prior. The proposed SIVI-based approaches only require the log joint distribution of the observed data and model parameters to be differentiable, so they apply to a broad range of model specifications. In this sense, we explore Gaussian and Poisson outcomes in detail to exemplify the proposed methodology.

The remainder of this paper is organized as follows. In Section 2, we review the Bayesian framework for geostatistics and the NNGP approach. Section 3 briefly introduces variational inference methods, focusing on the variational methods used in comparison to the proposed methods. Section 4 presents the proposed variational inference procedures in the context of Bayesian geostatistical models. The simulation studies are described in Section 5. We then illustrate the proposed methods on a large-scale dataset in Section 6. Finally, Section 7 concludes the paper with a discussion of the findings and potential areas for future research.

2 Geostatistics

In geostatistics, also referred to as point-referenced data, observations are collected at a set of n fixed locations $\mathcal{S} = \{s_1, \ldots, s_n\}$ within a region of interest $\mathcal{D} \subset \mathbb{R}^d$, where typically $d = 2$ or $d = 3$. At each location $s \in \mathcal{S}$, let $y(s)$ denote the observed outcome and let $\mathbf{x}(s)$ be a $p \times 1$ vector of covariates with associated $p \times 1$ vector of coefficients $\boldsymbol{\beta}$. Following Diggle et al. (1998), a general model for point-referenced observations within the region $\mathcal{D}$ can be expressed as

$$h(\mathbb{E}[y(s) \mid w(s)]) = \mathbf{x}(s)^\top \boldsymbol{\beta} + w(s), \quad s \in \mathcal{D}, \tag{1}$$

where $h(\cdot)$ is a known link function and $w(s)$ is a spatial random effect that captures the spatial dependence of observations at distinct locations. Observed data $\mathbf{y} = (y(s_1), \ldots, y(s_n))^\top$ are assumed to be conditionally independent across locations $s_1, \ldots, s_n$ and to follow a distribution in the exponential family given $\mathbf{w} = (w(s_1), \ldots, w(s_n))^\top$. The process $\{w(s) : s \in \mathcal{D}\}$ is typically modeled as a Gaussian process with zero mean function and covariance function $C(s_i, s_j) = Cov(w(s_i), w(s_j))$, in which case we denote $w(s) \sim \text{GP}(0, C)$. Throughout this paper, we adopt the exponential covariance function given by $C(s_i, s_j) = \sigma^2 exp(-\phi\ d(s_i, s_j))$, where $\sigma^2 > 0$ is the marginal variance of the spatial process, also referred to as the partial sill, $d(s_i, s_j)$ is the Euclidean distance between locations s_i and s_j, and $\phi > 0$ is the range parameter controlling the strength of the spatial correlation as distance between locations increases.

In this paper, we focus on two specific examples in the exponential family of distributions for $\mathbf{y}$: the Gaussian (section 2.1) and the Poisson (section 2.2). In a Bayesian framework, a prior distribution is assigned to all the unknown parameters $\boldsymbol{\theta}$ to complete the model specification. For instance, in the Poisson case, $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \sigma^2, \phi, \mathbf{w}\}$. Common choices include

independent Gaussian priors for the regression coefficients $\boldsymbol{\beta}$, inverse-gamma priors for the variance parameters σ^2 and τ^2 (see Gelman (2006) for alternatives), and either an inverse-gamma or a uniform prior for ϕ.

2.1 Gaussian Spatial Model

First, we consider the Gaussian case, where the multivariate normal likelihood of observed data $\mathbf{y}$ can be expressed in two equivalent forms. The first one, referred to as the conditional model, keeps the random effects vector $\mathbf{w}$ in the model such that $\mathbf{y} \mid \boldsymbol{\beta}, \mathbf{w}, \tau^2 \sim \text{MVN}(\mathbf{X}\boldsymbol{\beta} + \mathbf{w}, \tau^2\mathbf{I}_n)$, where $\mathbf{X}$ is the $n \times p$ matrix with its i-th row representing the observed vector of covariates $\mathbf{x}(s_i)$, $\mathbf{I}_n$ is the identity matrix of dimension n, and $\tau^2 > 0$, referred to as the nugget, captures the measurement errors. The second formulation, referred to as the marginal model, integrates out the random effects vector $\mathbf{w}$ obtaining $\mathbf{y} \mid \boldsymbol{\beta}, \tau^2, \sigma^2, \phi \sim \text{MVN}(\mathbf{X}\boldsymbol{\beta}, \mathbf{C} + \tau^2\mathbf{I}_n)$, where $\mathbf{C}$ denotes the $n \times n$ covariance matrix with entries $C_{ij} = Cov(w(s_i), w(s_j)),\ i, j \in \{1, \ldots, n\}$.

For the Gaussian case, we assigned independent standard Gaussian priors to the components of $\boldsymbol{\beta}$ and inverse-gamma priors for σ^2, τ^2, and ϕ. Specifically, we centered the prior distribution of σ^2 and τ^2 on the empirical variance of the observed data to ensure that they are on the correct scale by setting the shape to 2 and the scale to the observed response variance $\widehat{\text{Var}}(\mathbf{y}) = \frac{1}{n-1}\sum_{i=1}^{n}(y_{s_i} - \bar{y})^2$ with $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_{s_i}$. For the range parameter ϕ, we set inverse-gamma prior with shape equal to a and scale equal to b such that $E[\phi] = \frac{6}{\max d_{ij}}$ and $Pr[\phi < 2E(\phi)] \approx 0.99$. We used an informative prior on ϕ to address well-known identifiability issues in geostatistics (Zhang 2004, Tang et al. 2021). This prior was designed to softly constrain the range of ϕ, playing a similar role to the uniform prior commonly used in the literature for the range parameter.

2.2 Poisson Spatial Model

Second, we consider the Poisson case with the canonical logarithmic link function (i.e., $h(u) = \log(u)$). In this case, the conditional model is given by $y(s) \mid \lambda(s) \sim \text{Poisson}(\lambda(s))$, where $\lambda(s)$ is the conditional mean of the distribution as defined by Equation (1). Unlike the Gaussian case, the marginal model does not have a closed form and will not be considered.

For the Poisson case, we assigned a Gaussian prior with mean one and unit variance to the intercept β_0, and standard Gaussian prior to the other regression coefficients β_k for $k \neq 0$, inverse-gamma priors with shape two and scale 0.1 for σ^2. An inverse-gamma prior with shape a and scale b for ϕ, where a and b are obtained the same way as in the Gaussian case. Those priors are centered on values ensuring that they would be in an appropriate scale.

2.3 Bayesian Inference in Geostatistics

With the model specification complete, the quantity of interest then becomes the posterior distribution $\pi(\boldsymbol{\theta} \mid \mathbf{y})$. The posterior distribution does not usually have an analytical form and is typically approximated using one of many possible computational methods for Bayesian inference, such as sampling-based Markov chain Monte Carlo (MCMC) methods, optimization-based approaches like variational inference (see Blei et al. (2017) for an introduction), and other methods, such as INLA (Rue et al. 2009).

While MCMC methods have desirable theoretical properties, they become computationally expensive for large sample sizes, as each iteration requires computing the inverse and determinant of an $n \times n$ covariance matrix, needing $\mathcal{O}(n^3)$ operations and $\mathcal{O}(n^2)$ memory. To alleviate this issue, Datta et al. (2016) proposed an alternative prior for the spatial random effects known as the nearest-neighbor Gaussian process (NNGP). The key idea

behind NNGP is to factorize the joint distribution of the random effects using conditional distributions: $p(\mathbf{w}) = p(w_{s_1}) \prod_{i=2}^{n} p(w_{s_i} \mid w_{s_1}, \ldots, w_{s_{i-1}})$.

However, these conditional distributions remain computationally expensive as n grows large since each w_{s_i} depends on all previous locations. Vecchia's approximation (Vecchia 1988) addresses this issue by restricting the conditioning set of location s_i to at most M of its nearest neighbors:$p(\mathbf{w}) \approx \prod_{i=1}^{n} p(w_{s_i}|\mathbf{w}_{N(i)})$, where $N(i)$ denotes the set of at most M nearest neighbors of location s_i within $s_1, \ldots, s_{i-1}$. Datta et al. (2016) showed that this formulation results in a valid Gaussian process with a sparse precision matrix. The main advantage of using an NNGP prior is that computing the likelihood only involves covariance matrices of dimension $M \times M$, where M is usually much smaller than the number of observed locations n, which greatly reduces the computational and memory costs. Predictions at new locations are performed sequentially: each new location is conditioned on its M nearest neighbors among the set of observed locations.

Next, in section 3, we introduce variational inference and describe two flexible algorithms for approximate Bayesian inference in the context of spatial models.

3 Review of Variational Inference Methods

Variational inference (VI) is an optimization-based approach to approximate complex posterior distributions. The idea is to select a family of approximating distributions Q (namely the variational family) and find the member $q^*(\boldsymbol{\theta}) \in Q$ that is the closest to the true posterior $\pi(\boldsymbol{\theta} \mid \mathbf{Y})$ in terms of the Kullback-Leibler (KL) divergence: $q^*(\boldsymbol{\theta}) = \text{argmin}_{q \in Q} KL(q(\boldsymbol{\theta}) \parallel \pi(\boldsymbol{\theta}|\mathbf{Y}))$, with the KL divergence between two densities $q(\boldsymbol{\theta})$ and $p(\boldsymbol{\theta})$ being defined as $KL(q(\boldsymbol{\theta}) \parallel p(\boldsymbol{\theta})) = \int [\log q(\boldsymbol{\theta}) - \log p(\boldsymbol{\theta})]\, q(\boldsymbol{\theta}) d\boldsymbol{\theta} = \mathbb{E}_q(\log q(\boldsymbol{\theta}) - \log p(\boldsymbol{\theta}))$. For example, if Q is restricted to the family of Gaussian distributions, then, $q^*(\boldsymbol{\theta})$ corre-

sponds to the Gaussian distribution with the smallest KL divergence to the true posterior distribution. The mean-field approximation, one of the most commonly used restrictions when defining the variational family Q, assumes that the variational distribution can be factorized in blocks of independent subsets of parameters: $q(\boldsymbol{\theta}) = \prod_{k=1}^{K} q(\boldsymbol{\theta}_k), \quad \boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K\}$.

Directly solving this problem is possible even in the usual case where $\pi(\boldsymbol{\theta} \mid \mathbf{y})$ is known up to a multiplicative normalizing constant. By defining the evidence lower bound (ELBO) as $\text{ELBO} = -\mathbb{E}_q[\log q(\boldsymbol{\theta})] + \mathbb{E}_q[\log p(\boldsymbol{\theta}, \mathbf{y})]$ we can rearrange the KL divergence as

$$\begin{aligned} KL(q(\boldsymbol{\theta}) \parallel \pi(\boldsymbol{\theta} \mid \mathbf{y})) &= \mathbb{E}_q[\log q(\boldsymbol{\theta})] - \mathbb{E}_q[\log \pi(\boldsymbol{\theta} \mid \mathbf{y})] \\ &= \mathbb{E}_q[\log q(\boldsymbol{\theta})] - \mathbb{E}_q[\log p(\boldsymbol{\theta}, \mathbf{y}) - \log(p(\mathbf{y}))] \\ &= \mathbb{E}_q[\log q(\boldsymbol{\theta})] - \mathbb{E}_q[\log p(\boldsymbol{\theta}, \mathbf{y})] + \log p(\mathbf{y}) \\ &= -\text{ELBO} + \log p(\mathbf{y}), \end{aligned}$$

where $\log p(\mathbf{y})$ is referred to as the log-evidence. Since $\log p(\mathbf{y})$ does not depend on q, minimizing the KL divergence is equivalent to maximizing the ELBO with respect to $q \in Q$. Notice that the ELBO is indeed a lower bound on the log evidence, i.e, $ELBO \leq \log p(\mathbf{y})$ since the KL divergence is non-negative. In addition, the ELBO can be rewritten as $\text{ELBO} = \mathbb{E}_q[\log p(\mathbf{y} \mid \boldsymbol{\theta})] - KL(q(\boldsymbol{\theta}) \parallel \pi(\boldsymbol{\theta}))$. The first term rewards variational distributions that explain the observed data well, while the second term penalizes divergence from the prior, acting as a regularization term.

Next, in sections 3.1 and 3.2, we review two VI procedures for GPs in a Bayesian geostatistical context.

3.1 Automatic Differentiation Variational Inference

Kucukelbir et al. (2017) introduced automatic differentiation variational inference (ADVI), a general framework that automates the derivation of variational inference algorithms. The key idea is to transform all model parameters into an unconstrained real-valued space. If $\boldsymbol{\theta}$ represents the original parameters, an invertible and differentiable mapping $T : \boldsymbol{\theta} \mapsto \boldsymbol{\xi} \in \mathbb{R}^K$ is applied so that the joint distribution $p(\mathbf{y}, \boldsymbol{\theta})$ can be expressed as

$$p(\mathbf{y}, \boldsymbol{\xi}) = p(\mathbf{y}, \boldsymbol{\theta})\Big|_{\boldsymbol{\theta}=T^{-1}(\boldsymbol{\xi})} \times |J_{T^{-1}(\boldsymbol{\xi})}|,$$

where $|J_{T^{-1}(\boldsymbol{\xi})}|$ denotes the determinant of the Jacobian of the inverse transformation T^{-1}. This allows a reparameterizable variational family, typically a multivariate Gaussian, to be defined in the unconstrained space of the joint model parameters.

Two common choices for the variational Gaussian are the mean-field Gaussian and the full-rank Gaussian. Mean-field assumes that the covariance matrix is diagonal: $q(\boldsymbol{\xi}) = \prod_{k=1}^{K} N(\xi_k|\mu_k, \sigma_k^2)$ for $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_K)$ with variational parameters $\boldsymbol{\nu} = (\mu_1, \ldots, \mu_K, \sigma_1^2, \ldots, \sigma_K^2)$. Full-rank assumes $q(\boldsymbol{\xi}) = MVN(\boldsymbol{\xi} \mid \mathbf{m}, \mathbf{S})$ with variational parameters $\boldsymbol{\nu} = (\mathbf{m}, \mathbf{L})$ where $\mathbf{S} = \mathbf{L}\mathbf{L}^\top$ with $\mathbf{L}$ denoting a lower triangular matrix. The simplification of the covariance structure under the mean-field assumption leads to smaller computation times, but ignores posterior correlations, whereas the full-rank Gaussian imposes no constraints on the covariance matrix, allowing correlation between the parameters a posteriori to be captured.

Finally, by reparameterizing $\boldsymbol{\xi} = \mathbf{m} + \mathbf{L}\boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim MVN(\mathbf{0}, \mathbf{I})$ in the case of the full rank variational family assumption, we can write

$$\text{ELBO} = \mathbb{E}_{\boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{I})} \left\{ \log p(\mathbf{y}, T^{-1}(\boldsymbol{\xi})) + \log \mid J_{T^{-1}(\boldsymbol{\xi})} \mid - \log N(\boldsymbol{\xi};\ \mathbf{m}, \mathbf{L}\mathbf{L}^\top) \right\}$$

which can be approximated via Monte Carlo and optimized with respect to the variational parameters $\boldsymbol{\nu} = (\mathbf{m}, \mathbf{L})$ by gradient-based algorithms using automatic differentiation soft-

ware. As an example, Kucukelbir et al. (2015) describes the implementation of ADVI under Stan software (Stan Development Team 2025).

Regarding geostatistical models, the full-rank Gaussian is often a logical choice when using a conditional model since we expect the spatial random effects to be correlated. Doing so, however, leads to a large and dense covariance matrix, making the algorithm computationally expensive.

3.2 Pathfinder

Zhang et al. (2022) proposed Pathfinder, a variational method for differentiable probability densities. Pathfinder begins from random initializations in the tails of the posterior distribution and uses quasi-Newton optimization trajectories to move toward regions of high posterior density. Along each optimization trajectory, Pathfinder constructs a sequence of local multivariate Gaussian approximations using gradient and curvature information. The approximation that achieves the highest ELBO is selected as the final variational distribution for that trajectory.

When using multiple optimization trajectories, also referred to as multi-path Pathfinder, each trajectory can be run independently of the others. Posterior samples are then obtained by applying importance resampling across the draws from all trajectories (see Zhang et al. (2022) for the specifics of the implementation).

Pathfinder has the potential to be highly parallelizable, as each trajectory can be run independently of the others. Algorithmically, Pathfinder is similar to ADVI as they both use the same variational family. The main difference is how the optimization is performed. While ADVI uses gradient-based optimization, Pathfinder uses quasi-Newton optimization. Since Pathfinder depends on the quality of the local Gaussian approximations, it has similar issues to ADVI when applied to conditional geostatistical models.

Next, in section 4, we introduce semi-implicit variational inference by Yin & Zhou (2018) and present our proposed implementation of it using both full GP and NNGP priors for the latent spatial effects.

4 A Semi-Implicit Variational Inference Approach for Spatial Models

Yin & Zhou (2018) proposed semi-implicit variational inference (SIVI), a method that uses neural networks to define a flexible variational mixture model to approximate complex posterior distributions. Given observations $\mathbf{y}$, latent variables $\boldsymbol{\theta}$, likelihood $p(\mathbf{y} \mid \boldsymbol{\theta})$, and prior $\pi(\boldsymbol{\theta})$, the posterior distribution $\pi(\boldsymbol{\theta} \mid \mathbf{y})$ is approximated by a variational distribution $h_{\boldsymbol{\nu}}(\boldsymbol{\theta}) = \int q(\boldsymbol{\theta} \mid \boldsymbol{\psi}) q_{\boldsymbol{\nu}}(\boldsymbol{\psi}) d\boldsymbol{\psi}$, where the augmented mixing variables $\boldsymbol{\psi}$ are defined as $\boldsymbol{\psi} = g(\boldsymbol{\epsilon}; \boldsymbol{\nu})$, where $g(\cdot)$ represents a multilayer perceptron with variational parameters $\boldsymbol{\nu}$ (weights and biases) applied to random noise $\boldsymbol{\epsilon} \sim q(\boldsymbol{\epsilon})$. Therefore, the neural network transforms random noise $\boldsymbol{\epsilon}$ into an augmented mixing random variable $\boldsymbol{\psi}$, which implicitly defines $\boldsymbol{\psi} \sim q_{\boldsymbol{\nu}}(\boldsymbol{\psi})$. The vector of neural network parameters $\boldsymbol{\nu}$ is optimized by maximizing the ELBO, defined as

$$\text{ELBO} = \mathbb{E}_{h_{\boldsymbol{\nu}}}[\log p(\mathbf{y}, \boldsymbol{\theta}) - \log h_{\boldsymbol{\nu}}(\boldsymbol{\theta})]. \tag{2}$$

In SIVI, the conditional distribution $q(\boldsymbol{\theta} \mid \boldsymbol{\psi})$ must be explicit, meaning that the distribution has a closed parametric form, but the distribution $q_{\boldsymbol{\nu}}(\boldsymbol{\psi})$ is allowed to be implicitly defined via the transformation $\boldsymbol{\psi} = g(\boldsymbol{\epsilon}; \boldsymbol{\nu})$ with $\boldsymbol{\epsilon} \sim q(\boldsymbol{\epsilon})$. Another key assumption is that $q(\boldsymbol{\theta} \mid \boldsymbol{\psi})$ is reparameterizable, meaning that a sample can be generated from it by transforming random noise.

However, the ELBO defined in equation (2) is often intractable since it requires computing $h_{\boldsymbol{\nu}}(\boldsymbol{\theta})$, which may be only implicitly available. To address this, and also to avoid degeneracy issues, Yin & Zhou (2018) proposed a computable lower bound on the ELBO, given by

$$\underline{\mathcal{L}}_K = \mathbb{E}_{\boldsymbol{\psi}\sim q_{\boldsymbol{\nu}}(\boldsymbol{\psi})}\mathbb{E}_{\boldsymbol{\theta}\sim q(\boldsymbol{\theta}|\boldsymbol{\psi})}\mathbb{E}_{\widetilde{\boldsymbol{\psi}}_1,\ldots,\widetilde{\boldsymbol{\psi}}_K \overset{\text{iid}}{\sim} q_{\boldsymbol{\nu}}(\boldsymbol{\psi})}\left[\log p(\mathbf{y},\boldsymbol{\theta}) - \log\left(\frac{1}{K+1}\left\{q(\boldsymbol{\theta}\mid\boldsymbol{\psi}) + \sum_{k=1}^{K} q(\boldsymbol{\theta}\mid\widetilde{\boldsymbol{\psi}}_k)\right\}\right)\right], \tag{3}$$

where the added sum over k prevents SIVI from degenerating into a point mass density. It is important to highlight the fundamental distinction between $\widetilde{\boldsymbol{\psi}}_1,\ldots,\widetilde{\boldsymbol{\psi}_K}$ and $\boldsymbol{\psi}$. In equation (3), $\boldsymbol{\theta}$ is generated in two steps: 1) simulating $\boldsymbol{\psi}\sim q(\boldsymbol{\psi})$ and 2) conditionally on $\boldsymbol{\psi}$, simulating $\boldsymbol{\theta}\sim q(\boldsymbol{\theta}\mid\boldsymbol{\psi})$. Additionally, notice that, although we compute $q(\boldsymbol{\theta}\mid\widetilde{\boldsymbol{\psi}}_k)$, we never sample $\boldsymbol{\theta}$ conditionally on $\widetilde{\boldsymbol{\psi}}_1,\ldots,\widetilde{\boldsymbol{\psi}_K}$.

In practice, the expectations in Equation (3) are approximated by Monte Carlo. Specifically, we used one sample for each $\widetilde{\boldsymbol{\psi}}_k$, J samples for $\boldsymbol{\psi}$ and one sample of $\boldsymbol{\theta}$ for each $\boldsymbol{\psi}$. Then the resulting computable Monte Carlo approximation of the lower bound $\underline{\mathcal{L}}_K$ is

$$\widehat{\underline{\mathcal{L}}}_K = \frac{1}{J}\sum_{j=1}^{J}\log\left\{\frac{p(\mathbf{y},\boldsymbol{\theta}_j)}{\frac{1}{K+1}\left[q(\boldsymbol{\theta}_j\mid\boldsymbol{\psi}_j) + \sum_{k=1}^{K} q(\boldsymbol{\theta}_j\mid\widetilde{\boldsymbol{\psi}}_k)\right]}\right\}, \tag{4}$$

where $\boldsymbol{\psi}_j = g(\boldsymbol{\epsilon}_j;\boldsymbol{\nu}),\ \boldsymbol{\epsilon}_j \overset{\text{iid}}{\sim} q(\boldsymbol{\epsilon}),\ j=1,\ldots,J$, $\widetilde{\boldsymbol{\psi}}_k = g(\widetilde{\boldsymbol{\epsilon}}_k;\boldsymbol{\nu}), \widetilde{\boldsymbol{\epsilon}}_k \overset{\text{iid}}{\sim} q(\boldsymbol{\epsilon}),\ k=1,\ldots,K$, and $\boldsymbol{\theta}_j \sim q(\boldsymbol{\theta}\mid\boldsymbol{\psi}_j),\ j=1,\ldots,J$. Recall that $\boldsymbol{\epsilon},\ \widetilde{\boldsymbol{\epsilon}_1},\ldots,\widetilde{\boldsymbol{\epsilon}_K} \overset{\text{iid}}{\sim} q(\boldsymbol{\epsilon})$ implies that the parameterizations $\boldsymbol{\psi} = g(\boldsymbol{\epsilon};\boldsymbol{\nu})$ and $\widetilde{\boldsymbol{\psi}}_k = g(\widetilde{\boldsymbol{\epsilon}}_k;\boldsymbol{\nu})$ are differentiable functions of the variational parameters $\boldsymbol{\nu}$, therefore, the gradients for optimization of (4) with respect to $\boldsymbol{\nu}$ can be calculated using automatic differentiation software. Typically, $q(\boldsymbol{\epsilon})$ is a standard Gaussian distribution.

For geostatistical models, the conditional distribution $q(\boldsymbol{\theta}\mid\boldsymbol{\psi})$ can be specified such that all elements of $\boldsymbol{\theta}$ are conditionally independent, that is, $q(\boldsymbol{\theta}\mid\boldsymbol{\psi}) = \prod_{k=1}^{K} q(\theta_k\mid\boldsymbol{\psi})$. It is im-

portant to note that this conditional distribution is not the overall variational distribution on $\boldsymbol{\theta}$. Correlations between parameters are captured implicitly through the neural network structure defining $q_{\boldsymbol{\nu}}(\boldsymbol{\psi})$ since $h_{\boldsymbol{\nu}}(\boldsymbol{\theta}) = \int q(\boldsymbol{\theta} \mid \boldsymbol{\psi}) q_{\boldsymbol{\nu}}(\boldsymbol{\psi}) d\boldsymbol{\nu}$ and the components of $q_{\boldsymbol{\nu}}(\boldsymbol{\psi})$ are not independent.

The specific choice of the component distributions $q(\theta_k \mid \boldsymbol{\psi})$ can vary. The key requirement is that the selected distributions produce valid parameter values (e.g., non-negative support for variance terms). All hyperparameters of these conditional distributions (i.e., $\boldsymbol{\psi}$) are generated by the neural network, with appropriate transformation applied when necessary. For example, exponentiating the network output to guarantee that variance terms are greater than zero.

In this context, several choices are natural due to conjugacy considerations in the Gaussian case. Specifically, we assigned independent Gaussian variational distributions to the fixed effects $\boldsymbol{\beta}$ (i.e., $q(\beta_k \mid \boldsymbol{\psi}) = N(\mu_{\beta_k}, \sigma^2_{\beta_k})$, inverse-gamma distributions to the variance parameters σ^2 and τ^2 (i.e., $q(\sigma^2 \mid \boldsymbol{\psi}) = \text{IG}(a_{\sigma^2}, b_{\sigma^2}), q(\tau^2 \mid \boldsymbol{\psi}) = \text{IG}(a_{\tau^2}, b_{\tau^2})$), and Gaussian distributions with common variance to the random effects (i.e., $q(\mathbf{w} \mid \boldsymbol{\psi}) = \text{MVN}(\mu_{\mathbf{w}}, \sigma^2_{\mathbf{w}} \mathbf{I}_n)$). There is no conjugate choice for the range parameter ϕ. We assigned a log-normal distribution, which provided a reasonable distribution alternative (i.e., $q(\phi \mid \boldsymbol{\psi}) = LN(\mu_\phi, \sigma^2_\phi)$. Hence, for the Gaussian case, $\boldsymbol{\psi} = (\mu_{\beta_0}, \ldots, \mu_{\beta_k}, \sigma^2_{\beta_0}, \ldots, \sigma^2_{\beta_k}, a_{\sigma^2}, b_{\sigma^2}, a_{\tau^2}, b_{\tau^2}, \mu_\phi, \sigma^2_\phi, \boldsymbol{\mu}_{\mathbf{w}}, \sigma^2_{\mathbf{w}})$.

For the Poisson case, as no conjugate choices were available, we used the same component distributions as in the Gaussian case. The procedure is summarized in Algorithm 1, which describes the implementation of SIVI for the conditional Gaussian case with an intercept and one covariate using pseudo-code.

Algorithm 1 Semi-Implicit Variational Inference for the Conditional Gaussian Case

Inputs: response data $\mathbf{y}$, observed covariates $\mathbf{X}$ $J \in \mathbb{N}$, $K \in \mathbb{N}$, initial variational parameters values $\boldsymbol{\nu}_0$, learning rate $\eta \in (0, 1)$, neural network architecture $g(\,\cdot\,;\,\boldsymbol{\nu})$

Initialize the neural network parameters $\boldsymbol{\nu} = \boldsymbol{\nu}_0$

for each training iteration **do**

for $j = 1, \ldots, J$ **do**

Generate a random noise vector: $\boldsymbol{\epsilon}_j \overset{\text{iid}}{\sim} q(\boldsymbol{\epsilon})$

Sample the hyperparameters using the neural network:

$\boldsymbol{\psi}_j = (\mu_{\beta_0,j}, \sigma^2_{\beta_0,j}, \mu_{\beta_1,j}, \sigma^2_{\beta_1,j}, a_{\sigma^2,j}, b_{\sigma^2,j}, a_{\tau^2,j}, b_{\tau^2,j}, \mu_{\phi,j}, \sigma^2_{\phi,j}, \boldsymbol{\mu}_{\mathbf{w},j}, \sigma^2_{\mathbf{w},j}) = g(\boldsymbol{\epsilon}_j; \boldsymbol{\nu})$

Draw parameters from the conditional distributions:

$\beta_{0,j} \sim N(\mu_{\beta_0,j}, \sigma^2_{\beta_0,j})$

$\beta_{1,j} \sim N(\mu_{\beta_1,j}, \sigma^2_{\beta_1,j})$

$\sigma^2_j \sim \text{IG}(a_{\sigma^2,j}, b_{\sigma^2,j})$

$\tau^2_j \sim \text{IG}(a_{\tau^2,j}, b_{\tau^2,j})$

$\phi_j \sim \text{LN}(\mu_{\phi,j}, \sigma^2_{\phi,j})$

$\mathbf{w}_j \sim \text{MVN}(\boldsymbol{\mu}_{\mathbf{w},j}, \sigma^2_{\mathbf{w},j}\mathbf{I}_n)$

end for

Generate K auxiliary samples of hyperparameters: $\widetilde{\boldsymbol{\psi}_k} = g(\widetilde{\boldsymbol{\epsilon}_k}; \boldsymbol{\nu}), \widetilde{\boldsymbol{\epsilon}_k} \overset{\text{iid}}{\sim} q(\boldsymbol{\epsilon}),\ k = 1, \ldots, K,$

Compute $\widehat{\underline{\mathcal{L}}}_K$ using Eq. (4):

$\widehat{\underline{\mathcal{L}}}_K = \frac{1}{J} \sum_{j=1}^{J} \log \left\{ \frac{p(\mathbf{y}, \boldsymbol{\theta}_j)}{\frac{1}{K+1}\left[q(\boldsymbol{\theta}_j|\boldsymbol{\psi}_j) + \sum_{k=1}^{K} q(\boldsymbol{\theta}_j|\widetilde{\boldsymbol{\psi}_k})\right]} \right\}$

Update the neural network parameters: $\boldsymbol{\nu} \leftarrow \boldsymbol{\nu} + \eta \nabla_{\boldsymbol{\nu}} \widehat{\underline{\mathcal{L}}}_K$

▷ The gradient is computed via automatic differentiation

end for

Draw approximate posterior samples from the variational distribution $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m \overset{\text{iid}}{\sim} q_{\boldsymbol{\nu}}(\boldsymbol{\theta})$

Output: Final neural network parameters $\boldsymbol{\nu}$ and samples $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m \overset{\text{iid}}{\sim} q_{\boldsymbol{\nu}}(\boldsymbol{\theta})$

4.1 SIVI-NNGP

Since SIVI only requires the log joint distribution of the observed data and model parameters, replacing the original GP prior by an NNGP prior is straightforward. To implement an NNGP prior, the spatial locations must be ordered, typically by one of the coordinates, and the maximum number of neighbors each location could have, M, is specified.

Let $N(i)$ be the set of neighbors for location s_i. $N(i)$ can contain at most M locations among $s_1, \ldots, s_{i-1}$, corresponding to the M locations with the smallest Euclidean distances to s_i.

Following Datta et al. (2016), let $\mathbf{c}_{i,N(i)}$ denote the $1 \times M$ vector of covariances between location s_i and its neighbors in $N(i)$ and let $C_{N(i)}$ denote the $M \times M$ covariance matrix of the locations in $N(i)$. Then, the factorized likelihood as described in section 2.3 becomes $p(\mathbf{w}) \approx \prod_{i=1}^{n} N(w_{s_i} | b_i^\top \mathbf{w}_{N(i)}, f_i)$, where $b_i = \mathbf{c}_{i,N(i)} C_{N(i)}^{-1}$ and $f_i = \mathrm{Cov}(s_i, s_i) - \mathbf{c}_{i,N(i)} C_{N(i)}^{-1} \mathbf{c}_{i,N(i)}^\top$.

In our implementation, we ordered by the first coordinate and set $M = 10$. All other parameters of the SIVI algorithm were kept identical to those used in the full GP implementation described in the previous section.

To implement SIVI-NNGP, the only change required in Algorithm 1 happens in the numerator of $\widehat{\underline{\mathcal{L}}}_K$ (Equation (4)). Specifically, the prior on $\mathbf{w}$ is replaced by the NNGP prior in the computation of $p(\mathbf{y}, \boldsymbol{\theta}_j)$.

Next, in section 5, we describe the simulation studies and their results.

5 Simulation Studies

In this section, we perform different simulation studies to evaluate the performance of SIVI and SIVI-NNGP in comparison with HMC and other VI methods. We first describe the

general settings that are shared between the Gaussian and Poisson models, including the sample sizes, software, hardware (CPU and GPU), and the criteria used for comparison. Then, in Sections 5.1 and 5.2, we present the implementation details, including the data generation mechanism, and the results for the Gaussian and Poisson outcomes, respectively.

We considered three sample sizes, $n \in (70, 170, 520)$. For each sample size, data were split into a training set and a validation set of size 20, resulting in training sets of size 50, 150, and 500, respectively. For each sample size, 50 independent replicates were generated. In each replicate, spatial locations were randomly sampled from a square of length 50.

We compared the proposed SIVI method and two commonly used variational inference methods (ADVI and Pathfinder) to the No-U-Turn Sampler (NUTS) by Hoffman et al. (2014), an adaptive Hamiltonian Monte Carlo (HMC) algorithm (Neal 2011) widely used for spatial modeling (see Aiello et al. (2023) for an applied example and chapter 10 of the Stan User's Guide (Stan Development Team 2025) for details on implementing GPs using HMC). Further, we evaluated the proposed SIVI-NNGP approach, which implements SIVI with an NNGP prior for the spatial random effects rather than the full GP prior, and compared it to the traditional NNGP implementation fitted using NUTS.

We implemented the SIVI-based inference algorithms in Python using TensorFlow (Abadi et al. 2015), whereas the other methods were implemented in R via CmdStanR (Gabry et al. 2025). All computations were performed on Rorqual, a high-performance computation cluster operated by Calcul Québec and part of the Digital Research Alliance of Canada.

The HMC algorithm was run with four chains simulated in parallel for 2,000 iterations each, including 1,000 warm-up iterations, resulting in 1,000 posterior samples per chain. For ADVI, we used the full-rank Gaussian implementation, as preliminary testing showed poor performance for the mean-field variation. For the two largest sample sizes, 100 Monte Carlo samples were used to approximate the gradient during stochastic gradient descent;

for the smallest sample size, the full set of observations was used. We also used 100 Monte Carlo samples to evaluate the ELBO. For Pathfinder, we used 30 independent paths, each with a maximum of 1,000 steps along the optimization trajectory. We set the history size to 6 for the inverse Hessian approximation, and 50 Monte Carlo samples were used to evaluate the ELBO.

For SIVI, each auxiliary variable $\boldsymbol{\psi}$ for the variational model was defined as an output from a deep neural network architecture $g(\,\cdot\,;\,\boldsymbol{\nu}))$ with weights and biases $\boldsymbol{\nu}$ with five intermediate hidden layers containing 2048, 1500, 1000, 800, and 600 neurons with a ReLU activation function, respectively, with a 100-dimensional standard Gaussian noise vector as input. The network was trained for 1,000 iterations. We used 50 Monte Carlo samples to approximate expectations ($J = 50$) and set $K = 1,000$ in Equation (4). To accelerate computations, we enabled TensorFlow graph mode by annotating functions with `@tf.function` and replacing Python loops with TensorFlow loops. Further, for SIVI-NNGP, we set the maximum neighborhood size to 10 (i.e., $M = 10$).

Table 1 summarizes the computational resources allocated to each method in terms of the number of CPU cores and GPU usage. HMC used four cores (one for each chain), ADVI was restricted to a single-core implementation since it does not currently support multiple cores, and although Pathfinder is highly parallelizable, we restricted the number of cores to six, representing a realistic setting. SIVI-based methods used two cores and $\frac{1}{8}$th of the computing power of an NVIDIA H100 GPU. The CPU-only methods were run on an AMD EPYC 9654 (Zen 4) processor, and the CPU-GPU methods were run on an Intel Xeon Gold 6448Y processor.

Table 1: Computational resources allocated to each method

Method	Number of cores	GPU
HMC / NNGP	4	-
ADVI	1	-
Pathfinder	6	-
SIVI / SIVI-NNGP	2	$\frac{1}{8}$ of a NVIDIA H100

Notes: HMC = Hamiltonian Monte Carlo; ADVI = Automatic Differentiation Variational Inference; SIVI = Semi-Implicit Variational Inference; NNGP = Nearest-Neighbor Gaussian Process.

The comparison among inference procedures was twofold. First, we assessed predictive performance on the validation set using three metrics: the continuous ranked probability score (CRPS), the interval score (IS), and the negative log-predictive density (NLPD) (see Gneiting & Raftery (2007) for more details). The CRPS measures the accuracy of the probabilistic predictions by comparing the predictive distribution with the observed outcomes. The IS measures the quality of uncertainty quantification, balancing the length of the credible interval with penalties when the true outcome falls outside the interval. The NLPD quantifies how likely the observed data are to be from the predictive density. Together, these metrics provide assessments of predictive accuracy, uncertainty quantification, and calibration. Second, we evaluated parameter estimation by examining posterior means across replicates, summarized using box plots for each parameter, and the runtime of the training process. Sections 5.1 and 5.2 describe the specifics of Gaussian and Poisson outcome simulations, respectively.

5.1 Gaussian Case

For the Gaussian model, observations were generated using the conditional model described in Section 2.1, with parameter values $\beta_0 = 0$, $\beta_1 = 0.5$, $\sigma^2 = 1$, $\tau^2 = 0.25$, and $\phi = 0.1$. Under this covariance specification, the correlation between two locations separated by a distance of approximately 30 (roughly 40% of the maximum possible distance) is about 0.05. The covariate $x(s)$ was defined as the standardized version of a sample from a Poisson distribution with mean 3: $x^*(s) \sim \text{Poisson}(3), x(s) = \frac{x(s)^* - \bar{x}(s)^*}{\text{SD}(x(s)^*)}$ associated with the coefficient β_1 and also an intercept associated to the coefficient β_0.

In the simulations, both the marginal and conditional formulations were evaluated for all methods. However, for the largest sample size ($n = 520$), some formulations were omitted due to computational constraints. Specifically, HMC was only applied to the marginal model, while ADVI was applied only to the conditional model. Pathfinder was excluded from the comparison at $n = 520$, as its conditional formulation performed poorly in the smaller sample sizes compared to the other methods. For the largest dataset, we additionally included the traditional NNGP implementation and the proposed SIVI-NNGP approach.

The predictive scores and the runtimes of the considered methods are presented in Figure 1. Panel *D* shows that the SIVI-based methods became faster than the alternatives once the training sample size reached 150, with the difference widening at the largest training size of 500. For $n_{train} = 500$, SIVI-NNGP required, on average, 110 seconds, compared to 710 seconds for NNGP fitted with HMC. The conditional and marginal SIVI implementations required 140 seconds and 106 seconds on average, respectively. On the other hand, ADVI was consistently slower than HMC for a given formulation: taking on average 2.8 hours when $n_{train} = 500$, compared to 1.2 hours for HMC. The methods performed similarly across the three prediction scores (CRPS, NLPD, and IS), with the conditional formulation

of Pathfinder performing slightly worse.

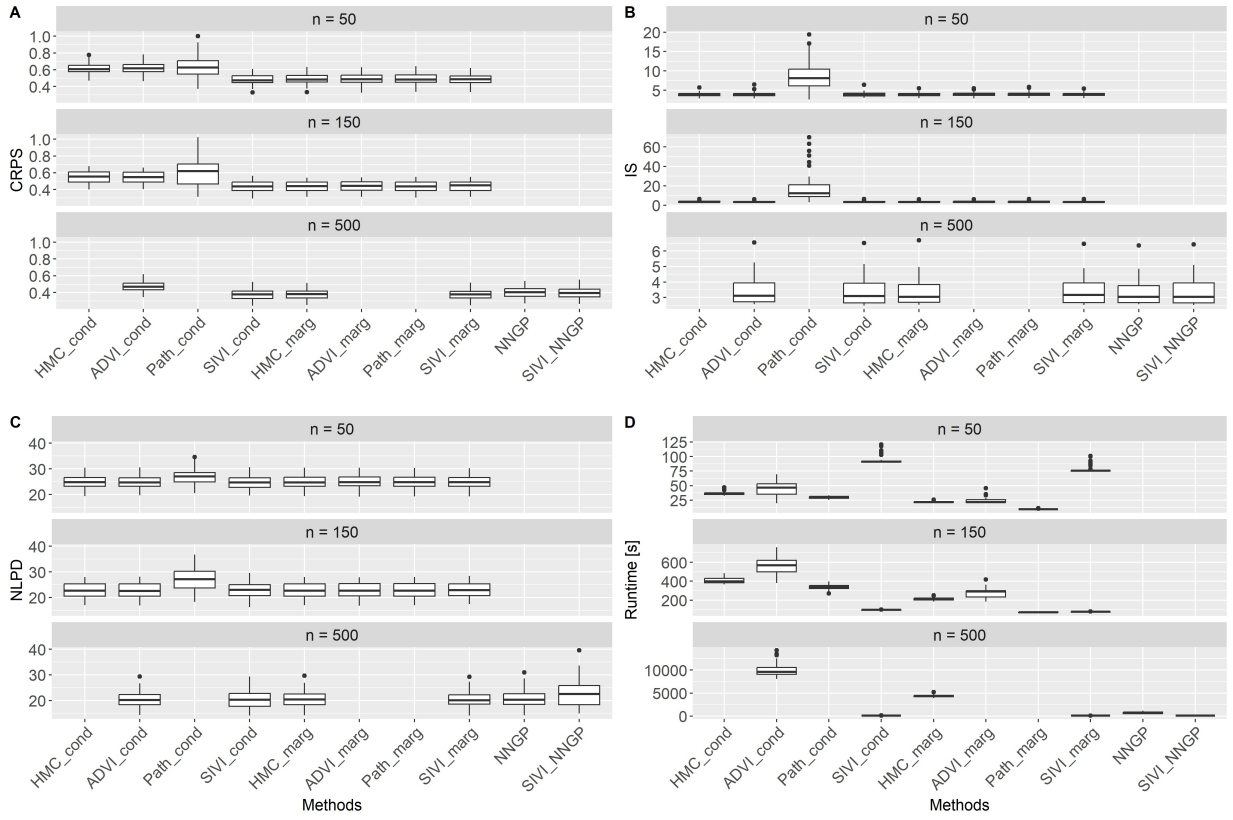

Figure 1: Prediction Scores and Runtimes of 50 Replicates for the Gaussian Models. An empty column indicate that the corresponding method was not evaluated for that scenario. All prediction scores were computed on 20 held-out locations. A: Continuous Ranked Probability Score (CRPS), B: Interval Score (IS), C: Negative log-predictive density (NLPD), D: Runtime. Methods: HMC = Hamiltonian Monte Carlo, ADVI = Automatic Differentiation Variational Inference, Path = Pathfinder, SIVI = Semi-Implicit Variational Inference, NNGP = Nearest-Neighbor Gaussian Process. Formulations: cond = Conditional, marg = Marginal

Posterior means of the model parameters are presented in Figure 2. Estimates were generally similar across methods. Conditional Pathfinder struggled with the variance terms (σ^2 and τ^2) and the range parameter ϕ. Conditional SIVI, SIVI-NNGP, and NNGP methods

tended to slightly overestimate the range parameter, implying a correlation structure that decayed quicker than expected with distance. In addition, conditional SIVI, SIVI-NNGP, and NNGP methods tended to underestimate the error variance (τ^2).

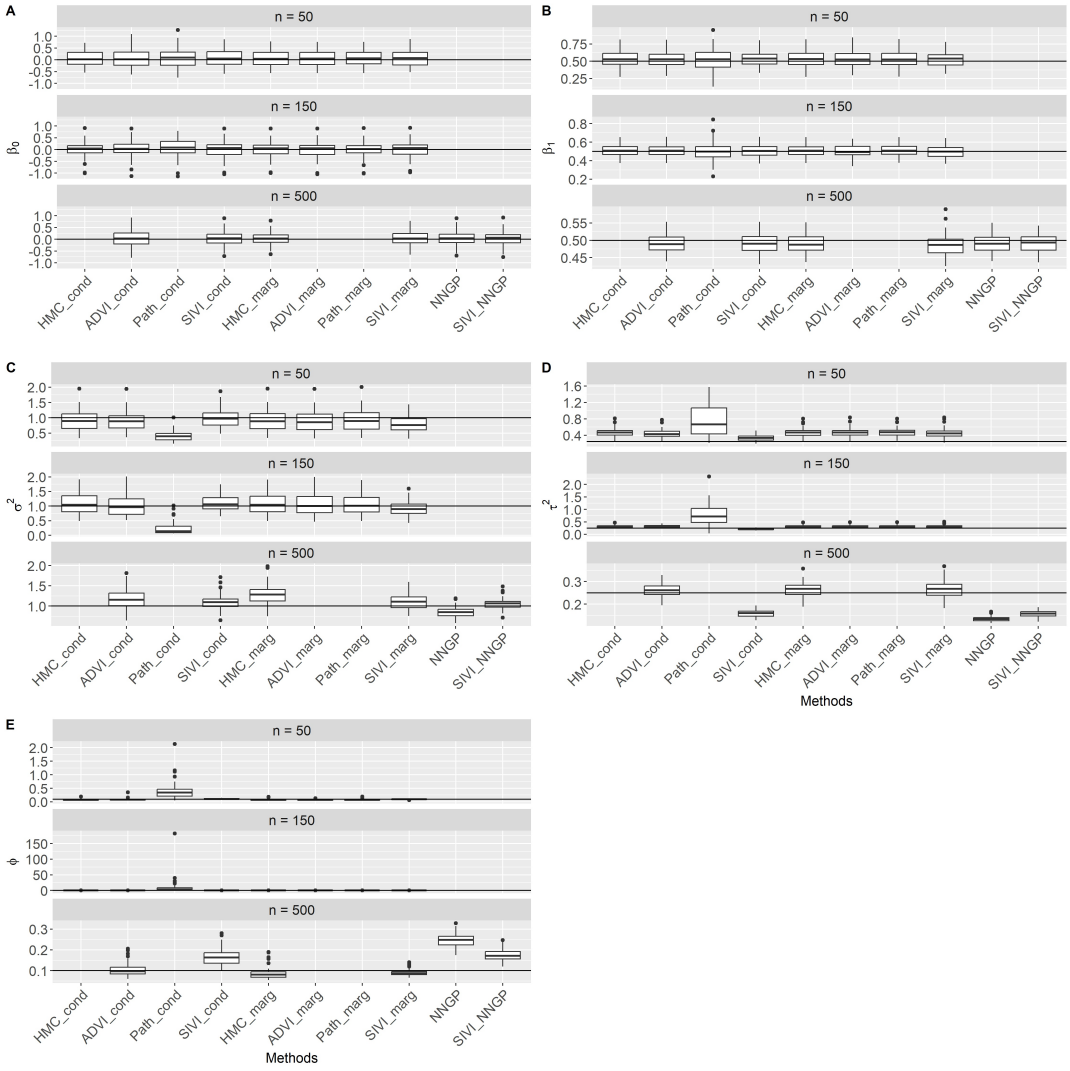

Figure 2: Posterior means of 50 Replicates for the Gaussian Models. An empty column indicate that the corresponding method was not evaluated for that scenario. The horizontal black line corresponds to the true value. A: β_0, B: β_1, C: σ^2, D:τ^2, E: ϕ. Methods: HMC = Hamiltonian Monte Carlo, ADVI = Automatic Differentiation Variational Inference, Path = Pathfinder, SIVI = Semi-Implicit Variational Inference, NNGP = Nearest-Neighbor Gaussian Process. Formulations: cond = Conditional, marg = Marginal.

Overall, these results highlight the advantages of SIVI and SIVI-NNGP. Both methods scale more efficiently than the alternatives, while achieving similar predictive scores to the reference HMC method with Marginal SIVI and Conditional SIVI showing slightly better results than SIVI-NNGP in terms of NLPD.

5.2 Poisson Case

For the Poisson model, observations were generated using the model described in Section 2.2, with parameter values $\beta_0 = 1.5$, $\beta_1 = 0.25$, $\sigma^2 = 0.1$, and $\phi = 0.1$. The covariate $x(s)$ was generated from a standard Gaussian distribution being associated with regression coefficient β_1. An intercept associated with β_0 was also considered.

Pathfinder was not considered for the Poisson case due to its worse performance in the conditional formulation of the Gaussian case. All other non-NNGP methods were included in the comparison.

The predictive scores and the runtimes for all methods applied to the Poisson model are presented in Figure 3. For the two smallest sample sizes, the methods performed similarly across all predictive scores. However, for the largest sample size, ADVI resulted in worse predictive performances than the other methods. In terms of computation times, SIVI required an average of 132 seconds for the largest sample size, compared to 9.7 hours for ADVI and 6.1 hours for HMC.

Posterior means of the model parameters are shown in Figure 4. ADVI tended to underestimate the intercept β_0, overestimate the partial sill (σ^2), and had difficulties estimating the range parameter ϕ. The SIVI methods overestimated the range parameter, implying a correlation structure that decayed faster with distance than expected.

These results reinforce the findings from the Gaussian case: SIVI scales much more effi-

ciently than the alternative ADVI and HMC methods while exhibiting predictive performance almost indistinguishable from the reference HMC approach. This also highlights that SIVI is flexible enough to perform well even for non-Gaussian outcomes.

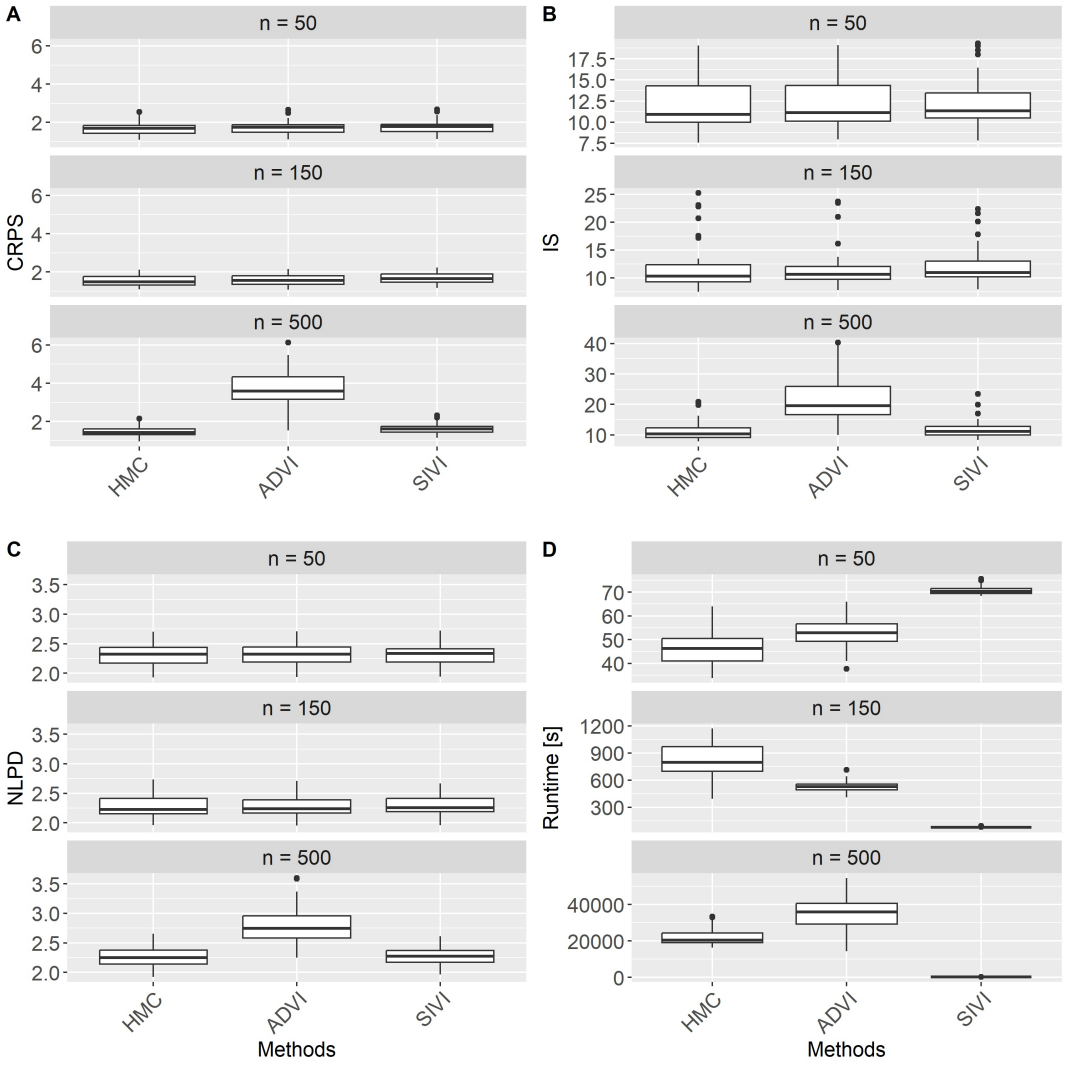

Figure 3: Prediction Scores and Runtime of 50 Replicates for the Poisson Model. All Prediction Scores Were Evaluated on 20 Held-Out Locations. A: Continuous Ranked Probability Score(CRPS), B: Interval Score (IS), C: Negative log-predictive density (NLPD), D: Runtime. Methods: HMC = Hamiltonian Monte Carlo, ADVI = Automatic Differentiation Variational Inference, SIVI = Semi-Implicit Variational Inference.

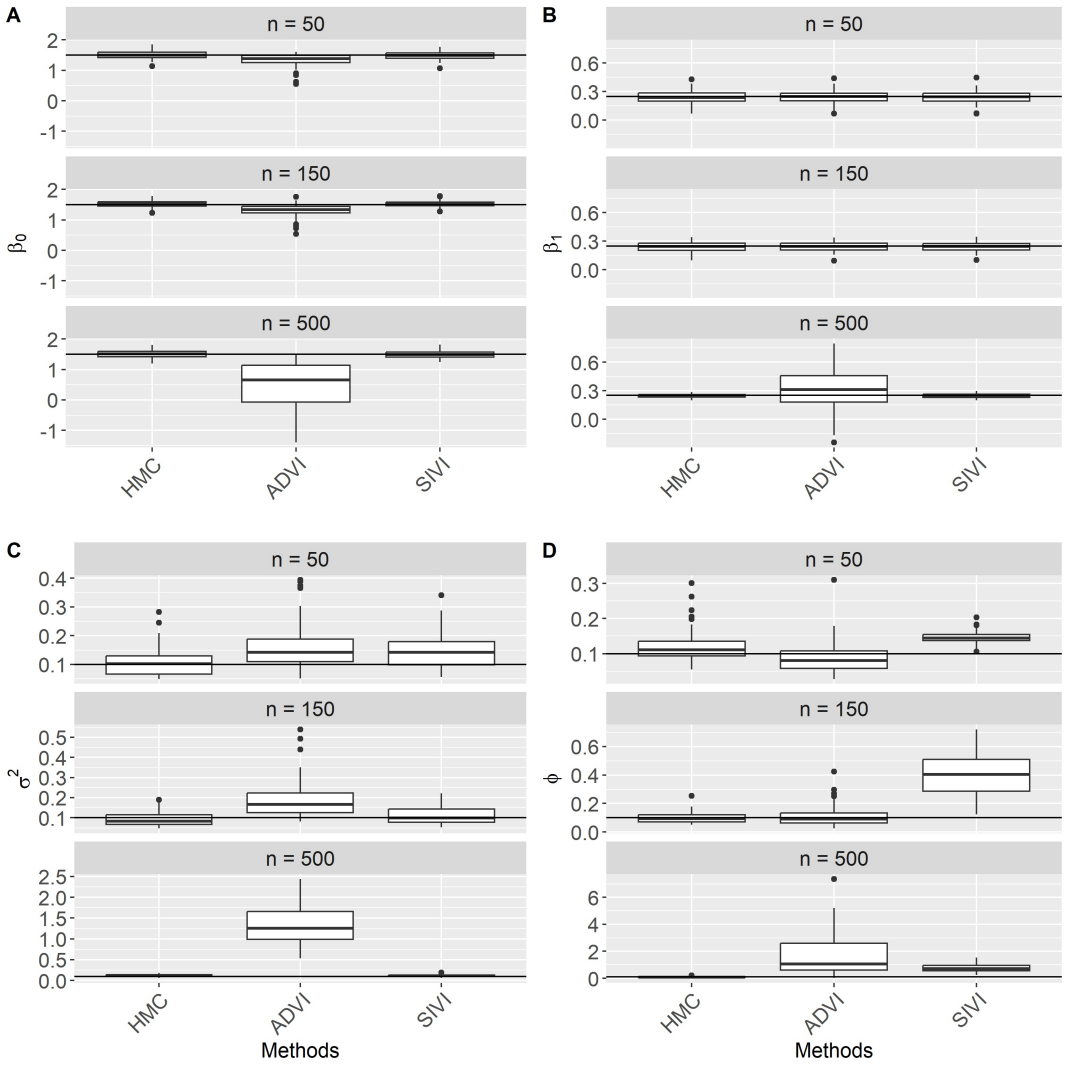

Figure 4: Posterior means of 50 Replicates for the Poisson model. The horizontal black line corresponds to the true value. A: β_0, B: β_1, C: σ^2, D: ϕ. Methods: HMC = Hamiltonian Monte Carlo, ADVI = Automatic Differentiation Variational Inference, SIVI = Semi-Implicit Variational Inference

Next, in section 6, we demonstrate the scalability of SIVI-NNGP by applying it to a large-scale spatial dataset.

6 Application to a Large Dataset

Heaton et al. (2019) conducted a comprehensive comparison of spatial prediction methods for Gaussian outcomes. The methods were evaluated on two large datasets of 150,000 locations: one real and one simulated. The real dataset consists of daytime land surface temperature measured by the Terra instrument onboard the MODIS satellite on August 4, 2016, covering longitudes of -95.91 to -91.28 and latitudes of 34.30 to 37.07. This dataset can be viewed as a misspecification case, since the true covariance structure is unlikely to follow an exponential model.

The simulated dataset, which imitates the real one, corresponds to a correctly specified case, as observations were simulated from a Gaussian process with a constant mean, an exponential correlation structure, and a nugget effect. In both datasets, locations were split into a training set of 105,569 locations and a validation set of 44,431 locations.

We applied SIVI-NNGP to the simulated dataset rather than the real dataset since our current implementation uses an exponential correlation structure and does not account for anisotropy. The neural network variational parameters were trained for 1,000 iterations with a 100-dimensional Gaussian noise vector, i.e. $\boldsymbol{\epsilon} \overset{q}{\sim} N(0, 1)$, as input . In Equation (4), we used $J = 10$ Monte Carlo samples to approximate expectations and set $K = 10$ to avoid numerical overflow during the ELBO computation steps using the complete set of training locations ($n_{train} = 105,539$). Figure 5 compares the predicted and true values of the 44,431 validation locations. It also shows the posterior predictive standard deviations of the predicted values. Table 2 presents a comparison of the predictive scores obtained

by SIVI-NNGP with the other methods reported in Heaton et al. (2019) on the validation locations.

The lowest CRPS score reported in Heaton et al. (2019) is 0.43. SIVI-NNGP's CRPS of 0.47 places it among the best-performing methods. Similarly, the lowest root mean squared error (RMSE) reported in Heaton et al. (2019) is 0.83. Given SIVI-NNGP's RMSE of 0.89, it places it among the best performing methods. Among the 13 methods considered in Heaton et al. (2019), only one method (Gapfill (Gerber et al. 2018)) completed the task in less time. However, it produced worse predictive scores (CRPS of 0.64 and RMSE of 1.00). Among all methods considered in Heaton et al. (2019), only one, reported as NNGP-conjugate, achieved similar predictive scores (CRPS of 0.46 and RMSE of 0.88) in a similar amount of time (approximately two minutes). However, this speed was achieved by *fixing parameters* in the covariance structure. The NNGP implementation that did not fix any parameters took approximately 45 minutes to complete, whereas SIVI-NNGP took less than two minutes without fixing any parameters.

The comparison of runtimes between SIVI-NNGP and the methods reported in Heaton et al. (2019) is not entirely direct. SIVI-NNGP was run on different hardware and under a different model specification (the NNGP approach in Heaton et al. (2019) used the marginal formulation, whereas SIVI-NNGP used the conditional one). Nevertheless, the results clearly show that SIVI-NNGP scales effectively to large datasets while achieving good predictive performance.

Table 2: Predictive Performance Computed on the 44,431 Validation Locations and Runtime of SIVI-NNGP compared to the methods reported in Heaton et al. (2019)

Score	SIVI-NNGP	Other Methods (best - worst)
CRPS	0.47	(0.43 - 0.74)
RMSE	0.89	(0.83 - 1.31)
Runtime [minutes]	1.8	(0.63 - 2888.89)

Notes: SIVI-NNGP = Semi-Implicit Variational Inference with Nearest-Neighbor Gaussian Process; CRPS = Continuous Ranked Probability Score; RMSE = Root Mean Squared Error

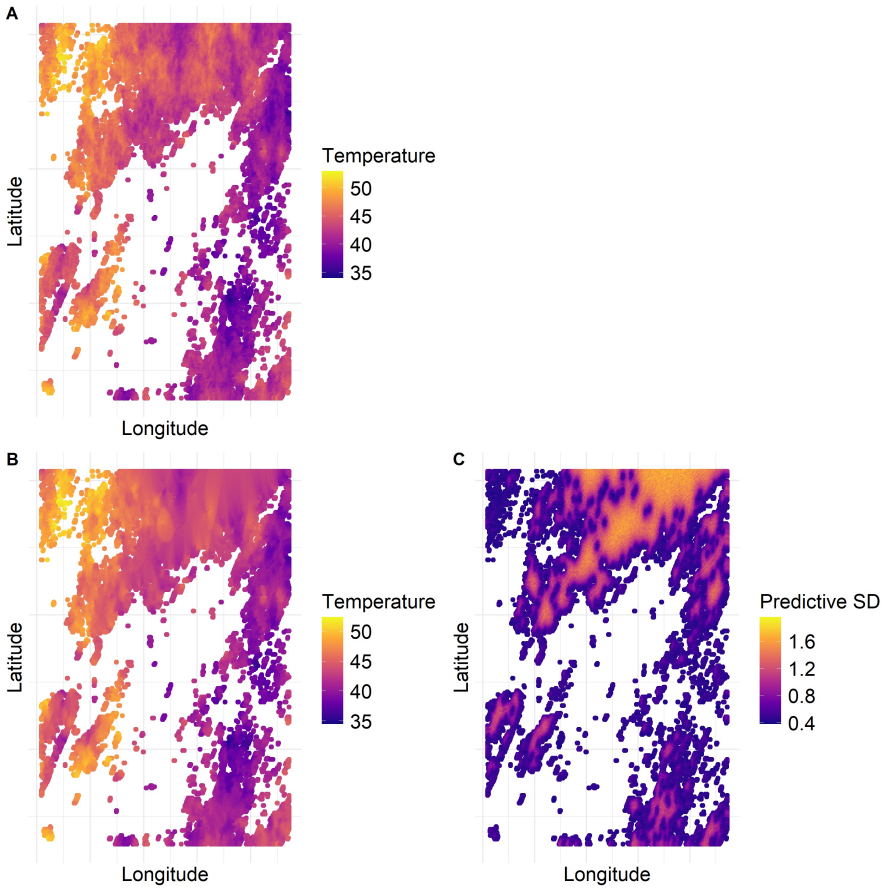

Figure 5: Posterior means and predictive standard deviations (SD) of the predicted values (panel B and C, respectively) compared to the simulated land surface temperature (panel A) of Heaton et al. (2019), covering longitudes -95.91 to -91.28 and latitudes 34.30 to 37.07.

7 Discussion

In this paper, we introduced SIVI alongside two other variational inference methods, ADVI and Pathfinder, and applied them to Bayesian geostatistical models. These methods were compared against HMC, used as the benchmark for performance evaluation. Both Gaussian and Poisson outcomes were considered, with sample sizes of 70, 170, and 520. We also implemented SIVI-NNGP, a combination of the SIVI algorithm with the NNGP prior, and compared it to an NNGP model fitted using HMC for Gaussian outcomes at a sample size of 520 observations.

Predictive performances were assessed using CRPS, IS, and NLPD on 20 held-out locations over 50 replicates. Overall, the methods offered comparable predictive performance across outcome types and sample sizes. The main exceptions were the conditional Pathfinder implementation for the Gaussian case and the ADVI implementation for the Poisson case at the largest sample size, both of which performed slightly worse.

Parameter estimation results were similarly consistent across outcome types and sample sizes. Adding to the previously mentioned exceptions, the conditional SIVI, SIVI-NNGP, and NNGP methods tended to overestimate the range parameter (ϕ) and underestimate the error variance (τ^2). In terms of computational efficiency, the SIVI-based methods achieved the fastest runtimes for the two largest sample sizes in both outcome types. On the other hand, ADVI was consistently slower than HMC.

To further evaluate the scalability of SIVI-NNGP, we applied it to a large dataset of 150,000 locations (105,569 training and 44,431 validation locations) previously analyzed in Heaton et al. (2019). SIVI-NNGP took less than 2 minutes to complete and achieved CRPS and RMSE scores that place it among the best-performing methods reported in Heaton et al. (2019).

Overall, these results highlight that SIVI-based methods can achieve predictive performance similar to HMC at a fraction of the computational cost, without needing to fix any parameters in the covariance structure. Furthermore, they demonstrate the potential of SIVI-based methods for large-scale geostatistical modeling. As a sensitivity analysis, we also explored a different variational distribution for the random effects in the conditional models. We used a multivariate Student-t distribution with four degrees of freedom. The results obtained (not shown) were almost identical to those reported here.

Note that memory usage was not evaluated in our comparison. While SIVI reduced computation times, it still required an important quantity of memory for large sample sizes, as it does not approximate the covariance structure in the likelihood as some other methods do. In practice, SIVI-NNGP helps with this aspect, but further optimization could improve memory efficiency. Also, our implementations of the SIVI-based methods relied on a standard neural network architecture and intuitive variational distributions. More advanced neural network designs or other variational families may improve the estimation of parameters.

Possible direction for future work include examination of other correlation functions, exploration of anisotropic covariance structures that allow correlation to vary with both distance and direction, and extending the framework to spatio-temporal contexts.

8 Data Availability Statement

The data used in Section 6 from Heaton et al. (2019) is freely available at: https://github.com/finnlindgren/heatoncomparison.

9 Acknowledgments

This research was enabled in part by support provided by Calcul Québec (www.calculquebec.ca) and the Digital Alliance of Canada (www.alliancecan.ca/en). We also acknowledge the use of ChatGPT-5 for assistance with debugging TensorFlow code and Grammarly for language improvement.

10 Disclosure statement

The authors report there are no conflicts of interest to declare.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y. & Zheng, X. (2015), 'TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems'. Software available from tensorflow.org. **URL:** *https://www.tensorflow.org/*

Aiello, L., Fontana, M. & Guglielmi, A. (2023), 'Bayesian Functional Emulation of CO_2 Emissions on Future Climate Change Scenarios', *Environmetrics* **34**(8), e2821.

Banerjee, S., Gelfand, A. E., Finley, A. O. & Sang, H. (2008), 'Gaussian Predictive Process Models for Large Spatial Data Sets', *Journal of the Royal Statistical Society Series B: Statistical Methodology* **70**(4), 825–848.

Blangiardo, M., Finazzi, F. & Cameletti, M. (2016), 'Two-Stage Bayesian Model to Evaluate the Effect of Air Pollution on Chronic Respiratory Diseases Using Drug Prescriptions', *Spatial and Spatio-temporal Epidemiology* **18**, 1–12. Environmental Exposure and Health.

Blei, D. M., Kucukelbir, A. & McAuliffe, J. D. (2017), 'Variational Inference: A Review for Statisticians', *Journal of the American Statistical Association* **112**(518), 859–877.

Datta, A., Banerjee, S., Finley, A. O. & Gelfand, A. E. (2016), 'Hierarchical Nearest-Neighbor Gaussian Process Models for Large Geostatistical Datasets', *Journal of the American Statistical Association* **111**(514), 800–812.

Diggle, P. J., Tawn, J. A. & Moyeed, R. A. (1998), 'Model-Based Geostatistics', *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **47**(3), 299–350.

Gabry, J., Češnovar, R., Johnson, A. & Bronder, S. (2025), *CmdStanR: R Interface to 'CmdStan'*. R package version 0.9.0, https://discourse.mc-stan.org.
URL: *https://mc-stan.org/cmdstanr/*

Gelman, A. (2006), 'Prior Distributions for Variance Parameters in Hierarchical Models', *Bayesian Analysis* **1**(3), 515–533.

Gerber, F., de Jong, R., Schaepman, M. E., Schaepman-Strub, G. & Furrer, R. (2018), 'Predicting Missing Values in Spatio-Temporal Remote Sensing Data', *IEEE Transactions on Geoscience and Remote Sensing* **56**(5), 2841–2853.

Gneiting, T. & Raftery, A. E. (2007), 'Strictly Proper Scoring Rules, Prediction, and Estimation', *Journal of the American Statistical Association* **102**(477), 359–378.

Heaton, M. J., Datta, A., Finley, A. O., Furrer, R., Guinness, J., Guhaniyogi, R., Gerber, F., Gramacy, R. B., Hammerling, D., Katzfuss, M., Lindgren, F., Nychka, D. W., Sun,

F. & Zammit-Mangion, A. (2019), 'A Case Study Competition Among Methods for Analyzing Large Spatial Data', *Journal of Agricultural, Biological and Environmental Statistics* **24**(3), 398–425.

Hoffman, M. D., Gelman, A. et al. (2014), 'The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo.', *Journal of Machine Learning Research* **15**(1), 1593–1623.

Kucukelbir, A., Ranganath, R., Gelman, A. & Blei, D. M. (2015), Automatic Variational Inference in Stan, *in* 'Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 1', NIPS'15, MIT Press, Cambridge, MA, USA, p. 568–576.

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A. & Blei, D. M. (2017), 'Automatic Differentiation Variational Inference', *Journal of Machine Learning Research* **18**(14), 1–45.

Lee, J. H. & Lee, B. S. (2025), 'A Scalable Variational Bayes Approach to Fit High-Dimensional Spatial Generalized Linear Mixed Models'.
URL: *https://arxiv.org/abs/2402.15705*

Lindgren, F., Rue, H. & Lindström, J. (2011), 'An Explicit Link Between Gaussian Fields and Gaussian Markov Random Fields: the Stochastic Partial Differential Equation Approach', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73**(4), 423–498.

Neal, R. M. (2011), MCMC Using Hamiltonian Dynamics, *in* S. Brooks, A. Gelman, G. L. Jones & X.-L. Meng, eds, 'Handbook of Markov Chain Monte Carlo', Chapman and Hall/CRC, pp. 116–162.

Ren, Q., Banerjee, S., Finley, A. O. & Hodges, J. S. (2011), 'Variational Bayesian Methods for Spatial Data Analysis', *Computational Statistics & Data Analysis* **55**(12), 3197–3217.

Rodriguez-Idiazabal, L., Martinez-Beneito, M. A., Quintana, J. M., Garcia-Asensio, J., Legarreta, M. J., Larrea, N. & Barrio, I. (2025), 'Understanding the COVID-19 Pandemic Through Bayesian Spatio-Temporal Modeling of Several Outcomes', *Spatial and Spatio-temporal Epidemiology* **54**, 100737.

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**(2), 319–392.

Song, J. & Datta, A. (2025), 'Fast Variational Bayes for Large Spatial Data'.
URL: *https://arxiv.org/abs/2507.12251*

Stan Development Team (2025), *Stan User's Guide*. Version 2.37.
URL: *https://mc-stan.org/users/documentation/*

Tang, W., Zhang, L. & Banerjee, S. (2021), 'On Identifiability and Consistency of the Nugget in Gaussian Spatial Process Models', *Journal of the Royal Statistical Society Series B: Statistical Methodology* **83**(5), 1044–1070.

Vecchia, A. V. (1988), 'Estimation and Model Identification for Continuous Spatial Processes', *Journal of the Royal Statistical Society: Series B (Methodological)* **50**(2), 297–312.

Yin, M. & Zhou, M. (2018), Semi-Implicit Variational Inference, *in* J. Dy & A. Krause, eds, 'Proceedings of the 35th International Conference on Machine Learning', Vol. 80 of *Proceedings of Machine Learning Research*, PMLR, pp. 5660–5669.

Yu, W., Liu, Y., Ma, Z. & Bi, J. (2017), 'Improving Satellite-Based PM2.5 Estimates in China Using Gaussian Processes Modeling in a Bayesian Hierarchical Setting', *Scientific Reports* **7**(1), 7048.

Zhang, H. (2004), 'Inconsistent Estimation and Asymptotically Equal Interpolations in Model-Based Geostatistics', *Journal of the American Statistical Association* **99**(465), 250–261.

Zhang, L., Carpenter, B., Gelman, A. & Vehtari, A. (2022), 'Pathfinder: Parallel Quasi-Newton Variational Inference', *Journal of Machine Learning Research* **23**(306), 1–49.