Received: 10 September 2022 | Revised: 8 March 2023 | Accepted: 11 April 2023

DOI: 10.1002/env.2805

RESEARCH ARTICLE

WILEY

Bayesian geostatistical modeling for discrete-valued processes

Xiaotian Zheng[1] | **Athanasios Kottas**[2] | **Bruno Sansó**[2]

[1]School of Mathematics and Applied Statistics, University of Wollongong, Wollongong, New South Wales, Australia

[2]Department of Statistics, University of California, Santa Cruz, California, USA

Correspondence
Xiaotian Zheng, School of Mathematics and Applied Statistics, University of Wollongong, Northfields Ave, Wollongong, NSW, 2522, Australia.
Email: xzheng@uow.edu.au

Funding information
National Science Foundation, Grant/Award Numbers: MMS 2050012, DMS 2153277, SES 1950902

Abstract

We introduce a flexible and scalable class of Bayesian geostatistical models for discrete data, based on nearest-neighbor mixture processes (NNMP), referred to as discrete NNMP. To define the joint probability mass function (pmf) over a set of spatial locations, we build from local mixtures of conditional pmfs using a directed graphical model, with a directed acyclic graph that summarizes the nearest neighbor structure. The approach supports direct, flexible modeling for multivariate dependence through specification of general bivariate discrete distributions that define the conditional pmfs. In particular, we develop a modeling and inferential framework for copula-based NNMPs that can attain flexible dependence structures, motivating the use of bivariate copula families for spatial processes. Moreover, the framework allows for construction of models given a pre-specified family of marginal distributions that can vary in space, facilitating covariate inclusion. Compared to the traditional class of spatial generalized linear mixed models, where spatial dependence is introduced through a transformation of response means, our process-based modeling approach provides both computational and inferential advantages. We illustrate the methodology with synthetic data examples and an analysis of North American Breeding Bird Survey data.

KEYWORDS

Bayesian hierarchical models, count data, nearest neighbors, spatial copula models, spatial generalized linear mixed models

1 | INTRODUCTION

Discrete geostatistical data arise in many areas, such as biology, ecology, and forestry. Such data sets consist of observations $\{y(\boldsymbol{s}_i)\}_{i=1}^n$ of random variables $\{Y(\boldsymbol{s}_i)\}_{i=1}^n$, where $\boldsymbol{s}_i = (s_{i1}, s_{i2})$ denotes a location in a continuous spatial domain $\mathcal{D} \subset \mathbb{R}^2$. Throughout the article, we assume that $Y(\boldsymbol{s}_i)$ is discrete, and takes values in $\mathbb{N} = \{0, 1, 2, \ldots\}$. As an example, consider observations for counts of a species of interest, commonly used to estimate the species distribution over a geographical domain.

To facilitate discussion in this section, we take $y_i \equiv y(\boldsymbol{s}_i)$ for brevity. Discrete geostatistical modeling concerns the joint probability mass function (pmf) $p(y_1, \ldots, y_n)$. The most common approach is the spatial generalized linear mixed model (SGLMM; Diggle et al., 1998), under which the joint pmf can be expressed as

$$p(y_1, \dots, y_n) = \int \cdots \int \prod_{i=1}^{n} p_i(y_i \mid z_i)\, p(z_1, \dots, z_n)\, dz_1 \cdots dz_n,$$

where $p_i(y_i \mid z_i)$ corresponds to an exponential family distribution, with z_i the spatial random effect in a linear mixed model for the transformed mean of y_i. The distribution for the random vector $(z_1, \dots, z_n)$ is induced by a spatial process over $\boldsymbol{D}$, typically a Gaussian process. The SGLMM provides a general modeling tool for geostatistical discrete data applications; see, for instance, Wikle (2002), Recta et al. (2012), Berrett and Calder (2016), and Zhang and Cressie (2020).

However, SGLMMs have certain properties that may be undesirable. First, since the spatial random effects are incorporated into the transformed mean, SGLMMs model spatial structure on a function of the response means, not on the observations directly. This is precisely the inferential focus in some cases, for example, when modeling the field of probabilities underlying spatial binary data. Nevertheless, due to the non-linear transformation induced by the link function, the model may impose a strong correlation between means over locations that are close, even though the corresponding observations may not be strongly correlated. More generally, the SGLMM may be limited when direct spatial modeling is of interest. In addition, the SGLMM specification poses computational challenges. Unlike Gaussian geostatistical models, the spatial random effects can not be marginalized out. Under simulation-based inference, estimating the spatial random effects requires sampling a large number of highly correlated parameters within a Markov chain Monte Carlo (MCMC) algorithm, which is likely to produce slow convergence, and a large memory footprint. Although efficient computational strategies have been explored in the literature (e.g., Christensen and Waagepetersen, 2002; Christensen et al., 2006; Guan and Haran, 2018; Sengupta and Cressie, 2013; Sengupta et al., 2016; Zhang, 2002), the computational challenge is unavoidable, especially for large spatial datasets.

An alternative to SGLMMs involves copula models which construct random fields given a pre-specified family of marginal distributions, with spatial dependence introduced through the multivariate distribution underlying a copula C. The joint cumulative distribution function (cdf) of the vector $(y_1, \dots, y_n)$ is given by $F(y_1, \dots, y_n) = C(F_1(y_1), \dots, F_n(y_n))$, where F_i is the cdf of y_i. The joint pmf is obtained by taking 2^n finite differences of C:

$$p(y_1, \dots, y_n) = \sum_{i_1=0,1} \cdots \sum_{i_n=0,1} (-1)^{i_1+\cdots+i_n}\, C(F_1(y_1 - i_1), \dots, F_n(y_n - i_n)). \tag{1}$$

Numerical computation of (1) requires 2^n n-dimensional copula evaluations. Unless n is very small, the computation is infeasible, necessitating approximate inference, for instance, through composite likelihoods (e.g., Bai et al., 2014; Kazianka, 2013; Kazianka and Pilz, 2010). In practice, spatial copula modeling typically turns to a Gaussian copula that provides simplicity in specifying spatial dependence (e.g., Han and De Oliveira, 2016; Madsen, 2009). The spatial Gaussian copula model (SGCM) has joint pmf

$$p(y_1, \dots, y_n) = \int_{\Phi^{-1}(F_1(y_1-1))}^{\Phi^{-1}(F_1(y_1))} \cdots \int_{\Phi^{-1}(F_n(y_n-1))}^{\Phi^{-1}(F_n(y_n))} N(z_1, \dots, z_n \mid \boldsymbol{0}, \boldsymbol{\Sigma})\, dz_1 \cdots dz_n, \tag{2}$$

where Φ is the cdf of the standard Gaussian distribution, and the covariance matrix $\boldsymbol{\Sigma}$ corresponds to a Gaussian process for the latent vector $(z_1, \dots, z_n)$, with $z_i = \Phi^{-1}(F_i(y_i))$. The joint pmf in (2) is more amenable to computation, although in general it still requires efficient approximations of a high-dimensional multivariate Gaussian integral (Han & De Oliveira, 2019), limiting the applicability of this class of models.

This article introduces a new class of spatial models for discrete geostatistical processes, along with a Bayesian framework for inference and prediction. This class is a discrete analogue of the nearest-neighbor mixture process (NNMP; Zheng et al., 2023), and is referred to as discrete NNMP. In general, multivariate discrete distributions are not as tractable as certain families of multivariate continuous distributions, in particular, the Gaussian family. This is the fundamental difficulty of process-based modeling for discrete geostatistical data. Both the SGLMM and SGCM resort to specifying spatial dependence through the latent vector $(z_1, \dots, z_n)$. Our methodology overcomes this difficulty by representing the joint pmf with respect to a fixed directed acyclic graph (DAG), that is,

$$p(y_1, \dots, y_n) = p(y_1) \prod_{i=2}^{n} p(y_i \mid \boldsymbol{y}_{\mathrm{Ne}(i)}),$$

where y_i is associated with vertex i in the DAG, and $\boldsymbol{y}_{\mathrm{Ne}(i)}$ corresponds to vertices j that have directed edges to i in the DAG, interpreted as the spatial nearest-neighbors of y_i. Factorization of the joint pmf according to a DAG implies conditional

independence, taking advantage of the local relationship between spatial locations. Spatial dependence for $(y_1, \dots, y_n)$ is introduced with appropriate models for the conditional pmfs $p(y_i \mid \boldsymbol{y}_{\mathrm{Ne}(i)})$. The discrete NNMP defines $p(y_i \mid \boldsymbol{y}_{\mathrm{Ne}(i)})$ as a mixture of conditional pmfs with spatially-dependent weights, such that each local conditional pmf depends on one element of $\boldsymbol{y}_{\mathrm{Ne}(i)}$. This reduces specification of a multivariate pmf $p(y_1, \dots, y_n)$ to that of bivariate pmfs that define the local conditional pmfs.

The primary goal of this article is the development of a copula-based modeling framework that provides generality to construct NNMPs given a discrete family for the marginal pmfs. We note that the copula approach considered in Zheng et al. (2023) does not apply to the discrete case. In fact, when the data are discrete, modeling and inference with general copulas present challenges (Genest & Nešlehová, 2007). Extra development is typically required for flexible modeling and efficient inference; see, for example, Panagiotelis et al. (2012), Smith and Khaled (2012), and Yang et al. (2020). Here, we carefully investigate the approach of using copulas to specify the bivariate pmfs of the NNMP. The resulting class of discrete copula NNMPs offers a new modeling tool for geostatistical discrete data. We show that the joint pmf $p(y_1, \dots, y_n)$ can be expressed using a collection of bivariate copulas and marginals, providing interpretability for model construction using different families of copulas. Our approach allows for the use of general bivariate copula families, which enhances model flexibility and enables the description of complex spatial dependencies. This feature compares favorably with the SGLMM that relies on a Gaussian process for the latent vector $(z_1, \dots, z_n)$, and with spatial copula models that are generally restricted to elliptical copulas, such as the SGCM in (2).

Regression modeling for discrete data differs from that for continuous data; common approaches such as additive or multiplicative models to include covariates are in general not available for discrete data. A further goal of this article is an extension of the first-order strict stationarity result in Zheng et al. (2023). The extension is key for discrete NNMPs, providing a constructive approach to develop models with spatially varying marginal pmfs $p_i(y_i)$. Hence, discrete NNMPs offer direct, process-based modeling of spatial dependence, while at the same time, they allow for incorporation of continuous or discrete covariates through the marginal distributions.

The discrete NNMP also offers advantages regarding computation. Direct modeling of the spatial dependence bypasses the need to work with a high-dimensional, correlated latent vector, which is unavoidable with SGLMMs. The structured mixture formulation of the NNMP requires only $4nL$ bivariate copula function evaluations for the joint pmf, where L is the number of mixture components. Since $L \ll n$, the numerical evaluation grows linearly in n. This results in a substantially lower computational cost, compared with the 2^n n-dimensional copula evaluations in (1), and with the $2n(n-1)$ bivariate copula evaluations using the composite likelihood approach. For efficient computation of discrete NNMPs for large data sets, we use uniform random variables to transform discrete variables into continuous ones. The proposed approach leverages the properties of wide families of copulas for continuous random vectors, and facilitates efficient estimation and prediction. We show through a simulation study that, compared with the popular SGLMM method, this approach yields reliable posterior inference at a much lower computational cost.

The article is organized as follows. In Section 2, we introduce NNMPs for discrete data, with copula-based discrete NNMPs developed in Section 3. Section 4 presents the Bayesian model formulation for inference, validation, and prediction, followed by illustration with synthetic and real datasets in Section 5. Finally, Section 6 concludes with a summary and discussion.

2 | NNMP MODELS FOR DISCRETE DATA

2.1 | Modeling framework

Consider a univariate, discrete-valued spatial process $Y(\boldsymbol{v})$ indexed by $\boldsymbol{v} \in \mathcal{D} \subset \mathbb{R}^p$, for $p \geq 1$. Let $\mathcal{S} = (\boldsymbol{s}_1, \dots, \boldsymbol{s}_n)$ be a reference set, where $\boldsymbol{s}_i \in \mathcal{D}$, for $i = 1, \dots, n$. We treat the set of locations in $\mathcal{S}$ as vertices of a fixed DAG, where each vertex $\boldsymbol{s}_i$ is associated with random variable $Y(\boldsymbol{s}_i)$. Equipped with the DAG, we can express the joint pmf $p(\boldsymbol{y}_{\mathcal{S}})$ as

$$p(\boldsymbol{y}_{\mathcal{S}}) = p(y(\boldsymbol{s}_1)) \prod_{i=2}^{n} p(y(\boldsymbol{s}_i) \mid \boldsymbol{y}_{\mathrm{Ne}(\boldsymbol{s}_i)}), \tag{3}$$

where $\boldsymbol{y}_{\mathcal{S}} = (y(\boldsymbol{s}_1), \dots, y(\boldsymbol{s}_n))^\top$, and $\mathrm{Ne}(\boldsymbol{s}_i) \subset \{\boldsymbol{s}_1, \dots, \boldsymbol{s}_{i-1}\}$ is the set of vertices that have directed edges to $\boldsymbol{s}_i$. The vector $\boldsymbol{y}_{\mathrm{Ne}(\boldsymbol{s}_i)}$ consists of random variables associated with $\mathrm{Ne}(\boldsymbol{s}_i)$. Typically, the elements of $\mathrm{Ne}(\boldsymbol{s}_i)$ are selected according to a specified distance function. In this article, the set $\mathrm{Ne}(\boldsymbol{s}_i)$ contains the first i_L elements in $\{\boldsymbol{s}_1, \dots, \boldsymbol{s}_{i-1}\}$ that are closest to $\boldsymbol{s}_i$

with respect to Euclidean distance, where $i_L = (i-1) \wedge L$. The elements in $\text{Ne}(\boldsymbol{s}_i)$ are ordered in ascending order, denoted as $\text{Ne}(\boldsymbol{s}_i) = (\boldsymbol{s}_{(i1)}, \ldots, \boldsymbol{s}_{(i,i_L)})$.

For continuous-valued spatial processes, expression (3) has been explored for fast likelihood computation (e.g., Katzfuss and Guinness, 2021; Vecchia, 1988), by viewing (3) as an approximation to the joint density of a Gaussian process realization. Datta et al. (2016) posit a parent Gaussian process, extending (3) to a nearest-neighbor Gaussian process by using the parent process to derive the conditionals in (3). Since the joint density under these models can be regarded as a factorization according to a DAG, they are often referred to as DAG-based models that provide scalability; see, for example, Peruzzi et al. (2022) and Jin et al. (2023) for some recent developments. Instead of treating the right-hand-side of (3) as an approximation, Zheng et al. (2023) use it as a direct model for $p(\boldsymbol{y}_S)$. Focusing on continuous data, the NNMP of Zheng et al. (2023) defines the conditionals in (3) with a structured mixture, which provides a general strategy to construct non-Gaussian spatial processes.

In this section, we extend the NNMP approach for discrete data. Before we proceed with details, we note that the factorization in (3) requires an ordering on the locations as they are not naturally ordered. Hereafter, we adopt a random ordering. For the data examples considered in Section 5, we found no discernible differences in model performance with two different random orderings. In fact, when it comes to prediction at a location outside $\mathcal{S}$, the effect of the ordering disappears; nearest neighbors of such a location are a subset of $\boldsymbol{y}_S$, based on the model formulation given below.

Constructing a discrete NNMP involves two steps. The first step consists of defining the joint pmf $p(\boldsymbol{y}_S)$ by modeling the conditional pmf in (3) as

$$p(y(\boldsymbol{s}_i) \mid \boldsymbol{y}_{\text{Ne}(\boldsymbol{s}_i)}) = \sum_{l=1}^{i_L} w_l(\boldsymbol{s}_i)\, f_{\boldsymbol{s}_i,l}(y(\boldsymbol{s}_i) \mid y(\boldsymbol{s}_{(il)})), \tag{4}$$

where $w_l(\boldsymbol{s}_i) \geq 0$ for every $\boldsymbol{s}_i \in \mathcal{S}$ and for all l, and $\sum_{l=1}^{i_L} w_l(\boldsymbol{s}_i) = 1$.

There are two model elements in (4) that describe spatial variability: the mixture component pmfs $f_{\boldsymbol{s}_i,l}$, and the weights $w_l(\boldsymbol{s}_i)$. We defer the specification of the pmfs $f_{\boldsymbol{s}_i,l}$ to the next section where the use of bivariate copulas is introduced as a general strategy. The weights are defined as increments of a logit Gaussian cdf $G_{\boldsymbol{s}_i}$, that is, $w_l(\boldsymbol{s}_i) = G_{\boldsymbol{s}_i}(r_{\boldsymbol{s}_i,l}) - G_{\boldsymbol{s}_i}(r_{\boldsymbol{s}_i,l-1})$, for $l = 1, \ldots, i_L$. Here, $0 = r_{\boldsymbol{s}_i,0} < r_{\boldsymbol{s}_i,1} < \ldots < r_{\boldsymbol{s}_i,i_L-1} < r_{\boldsymbol{s}_i,i_L} = 1$ are random cutoff points such that $r_{\boldsymbol{s}_i,l} - r_{\boldsymbol{s}_i,l-1} = k'(\boldsymbol{s}_i, \boldsymbol{s}_{(il)}) / \sum_{l=1}^{i_L} k'(\boldsymbol{s}_i, \boldsymbol{s}_{(il)})$, for some bounded kernel $k' : \mathcal{D} \times \mathcal{D} \to [0,1]$. Convenient choices for k' are kernels that compute the correlation between two points. The underlying Gaussian distribution for $G_{\boldsymbol{s}_i}$ has mean $\mu(\boldsymbol{s}_i) = \gamma_0 + \gamma_1 s_{i1} + \gamma_2 s_{i2}$, and variance κ^2. This formulation allows for spatial dependence among the weights through $\mu(\boldsymbol{s}_i)$. Also, the random cutoff points can flexibly reflect the neighbor structure of $\boldsymbol{s}_i$. We refer to Zheng et al. (2023) for additional details of the model for the weights.

The second step completes the construction of a valid stochastic process over $\mathcal{D}$ by extending (4) to an arbitrary finite set of locations outside $\mathcal{S}$, denoted as $\mathcal{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r)$, where $\mathcal{U} \subset \mathcal{D} \setminus \mathcal{S}$. In particular, we define the pmf of $\boldsymbol{y}_\mathcal{U}$ conditional on $\boldsymbol{y}_S$ as

$$p(\boldsymbol{y}_\mathcal{U} \mid \boldsymbol{y}_S) = \prod_{i=1}^{r} p(y(\boldsymbol{u}_i) \mid \boldsymbol{y}_{\text{Ne}(\boldsymbol{u}_i)}) = \prod_{i=1}^{r} \sum_{l=1}^{L} w_l(\boldsymbol{u}_i)\, f_{\boldsymbol{u}_i,l}(y(\boldsymbol{u}_i) \mid y(\boldsymbol{u}_{(il)})), \tag{5}$$

where the weights and conditional pmfs are defined analogously to Equation (4), and the points $(\boldsymbol{u}_{(i1)}, \ldots, \boldsymbol{u}_{(iL)})$ in $\text{Ne}(\boldsymbol{u}_i)$ are the first L locations in $\mathcal{S}$ that are closest to $\boldsymbol{u}_i$.

Given (4) and (5), a discrete-valued spatial process over $\mathcal{D}$ is well defined. For any finite set $\mathcal{V} \subset \mathcal{D}$ that is not a subset of $\mathcal{S}$, the joint pmf over $\mathcal{V}$ is obtained by marginalizing $p(\boldsymbol{y}_\mathcal{U} \mid \boldsymbol{y}_S) p(\boldsymbol{y}_S)$ over $\boldsymbol{y}_{S \setminus \mathcal{V}}$, where $\mathcal{U} = \mathcal{V} \setminus \mathcal{S}$. Practically, Equations (4) and (5) serve different purposes. The reference set $\mathcal{S}$ is often reserved for observed data, so model estimation is based on (4), while spatial prediction at new locations outside the reference set relies on (5). Henceforth, we use

$$p(y(\boldsymbol{v}) \mid \boldsymbol{y}_{\text{Ne}(\boldsymbol{v})}) = \sum_{l=1}^{L} w_l(\boldsymbol{v})\, f_{\boldsymbol{v},l}(y(\boldsymbol{v}) \mid y(\boldsymbol{v}_{(l)})), \tag{6}$$

to characterize discrete NNMPs, where $\boldsymbol{v}$ is a generic location in $\mathcal{D}$. The neighbor set $\text{Ne}(\boldsymbol{v})$ contains the first L locations in $\mathcal{S}$ that are closest to $\boldsymbol{v}$. We place these locations in ascending order according to distance, denoted as $\text{Ne}(\boldsymbol{v}) = (\boldsymbol{v}_{(1)}, \ldots, \boldsymbol{v}_{(L)})$.

Finally, we note that the discrete NNMP involves selecting the neighborhood size L. Our prior model for the spatially varying weights supports the strategy of using an over-specified L that gives a large neighbor set, with important neighbors

assigned large weights a posteriori. For specific data examples, a sensitivity analysis for L can be further carried out to find an optimal L according to standard model comparison metrics or scoring rules. This is illustrated with the real data application; see Section 5.3 and the Supporting Information.

2.2 | Model construction with spatially varying marginals

The key ingredient in constructing discrete NNMPs lies in the specification of the mixture component conditional pmfs $f_{\boldsymbol{v},l}$. There are many avenues to specify $f_{\boldsymbol{v},l}$. As each conditional pmf corresponds to a bivariate random vector, say $(U_{\boldsymbol{v},l}, V_{\boldsymbol{v},l})$, our strategy is to model $f_{\boldsymbol{v},l}$ through its bivariate pmf, denoted as $f_{U_{\boldsymbol{v},l},V_{\boldsymbol{v},l}}$. Let $f_{U_{\boldsymbol{v},l}}$ and $f_{V_{\boldsymbol{v},l}}$ be the marginal pmfs of $(U_{\boldsymbol{v},l}, V_{\boldsymbol{v},l})$, such that $f_{\boldsymbol{v},l} \equiv f_{U_{\boldsymbol{v},l}|V_{\boldsymbol{v},l}} = f_{U_{\boldsymbol{v},l},V_{\boldsymbol{v},l}}/f_{V_{\boldsymbol{v},l}}$. The benefits of this strategy are twofold. First, it simplifies the multivariate dependence specification by focusing on the bivariate random vectors $(U_{\boldsymbol{v},l}, V_{\boldsymbol{v},l})$. The multivariate dependence will be induced by bivariate distributions through the model's mixture formulation. Second, the strategy allows for the construction of models with a pre-specified family of marginal distributions, facilitating the study of local variability. For example, it is common in discrete geostatistical data modeling to include covariates through the (transformed) mean of the marginal distribution.

The second feature of this strategy relies on an extension of the first-order strict stationarity result from Zheng et al. (2023). Based on that result, an NNMP has stationary marginal pmf f_Y if $f_{U_{\boldsymbol{v},l}} = f_{V_{\boldsymbol{v},l}} = f_Y$, for all $\boldsymbol{v}$ and all l. Here, we generalize the result such that discrete NNMPs can be built from pre-specified spatially varying marginal pmfs $g_{\boldsymbol{v}}$, where $g_{\boldsymbol{v}}$ is the marginal pmf of $Y(\boldsymbol{v})$. The generalization of the stationarity proposition applies to all NNMPs. For the interest of this article, we summarize the result in the following proposition for discrete NNMPs.

> **Proposition 1.** *Consider a discrete NNMP model for spatial process* $\{Y(\boldsymbol{v}) : \boldsymbol{v} \in \mathcal{D}\}$, *and a collection of spatially varying pmfs* $\{g_{\boldsymbol{v}} : \boldsymbol{v} \in \mathcal{D}\}$. *If, for each* $\boldsymbol{v}$, *the marginal pmfs of the mixture component bivariate distributions are such that* $f_{U_{\boldsymbol{v},l}} = g_{\boldsymbol{v}}$ *and* $f_{V_{\boldsymbol{v},l}} = g_{\boldsymbol{v}_{(l)}}$, *the discrete NNMP has marginal pmf* $g_{\boldsymbol{v}}$ *for* $Y(\boldsymbol{v})$, *for every* $\boldsymbol{v} \in \mathcal{D}$.

A natural example for $\{g_{\boldsymbol{v}} : \boldsymbol{v} \in \mathcal{D}\}$ is a family of distributions with (at least) one of its parameters indexed in space, that is, $g_{\boldsymbol{v}}(\cdot) \equiv g(\cdot \mid \theta(\boldsymbol{v}), \boldsymbol{\xi})$, in particular, through spatially varying covariates. Using a link function for $\theta(\boldsymbol{v})$, we can include such covariates that provide additional spatially referenced information. A more general example involves partitioning the domain into several regions, where in each region, $g_{\boldsymbol{v}}$ is associated with a different family of marginal distributions. A relevant application is estimation of the abundance of a species that shows overdispersion in most areas, but underdispersion in areas where the species is less prevalent (Wu et al., 2015). Overall, Proposition 1 provides flexibility for construction of discrete-valued spatial models with specific marginal pmfs.

We develop next a key component of the methodology, that is, discrete copula NNMP model construction and inference. Given a family of marginal pmfs $g_{\boldsymbol{v}}$, we create spatial copulas for random vectors $(U_{\boldsymbol{v},l}, V_{\boldsymbol{v},l})$. We begin with copulas for a set of base random vectors (U_l, V_l), and extend them to be spatially dependent by modeling the copula parameter that controls the dependence structure as spatially varying. Together with Proposition 1, this strategy allows for construction of discrete NNMPs with marginal pmfs in general families.

3 | DISCRETE COPULA NNMP MODELS

3.1 | Copula functions

A bivariate copula function $C : [0,1]^2 \to [0,1]$ is a distribution function whose marginals are uniform distributions on $[0,1]$. Following Sklar (1959), given a random vector (Z_1, Z_2) with joint probability distribution F and marginals F_1 and F_2, there exists a copula function C such that $F(z_1, z_2) = C(F_1(z_1), F_2(z_2))$. If F_1 and F_2 are continuous, C is unique. In this case, the copula density is $c(z_1, z_2) = \partial C(F_1(z_1), F_2(z_2))/(\partial F_1 \partial F_2)$, and the joint density is $f(z_1, z_2) = c(z_1, z_2) f_1(z_1) f_2(z_2)$, where f_1 and f_2 are the densities of F_1 and F_2, respectively.

If both marginals are discrete, the copula C is only unique on the set $\mathrm{Ran}(F_1) \times \mathrm{Ran}(F_2)$, where $\mathrm{Ran}(F_j)$ consists of all possible values of F_j, $j = 1, 2$ (Joe, 2014). Nevertheless, if C is a copula and F_1 and F_2 are discrete distribution functions, then $F(z_1, z_2) = C(F_1(z_1), F_2(z_2))$ is a valid joint distribution; in practice, we select a parametric family for C (Panagiotelis et al., 2012; Smith and Khaled, 2012; Song et al., 2009). Note that, in contrast with the continuous case, when the marginals

are discrete, some popular dependence measures, such as Kendall's τ, will depend on the marginals (Denuit and Lambert, 2005; Genest and Nešlehová, 2007). Consequently, the Kendall's τ of the random vector (Z_1, Z_2) will not be equivalent to the Kendall's τ of the copula. Without loss of generality, hereafter, we assume the bivariate copula carries a single parameter.

3.2 | Copula NNMPs for discrete geostatistical data

Here, we introduce copula NNMPs with discrete marginals, with focus on using copulas to specify the bivariate distributions of the mixture components. Dropping the dependence on l for clarity, consider a random vector (U, V) with discrete marginal distributions F_U, F_V, and marginal pmfs f_U, f_V. Let $a_u = F_U(u-1)$ and $b_u = F_U(u)$. Analogous definitions of a_v and b_v apply for V. The joint pmf $f_{U,V}$ of (U, V) is obtained by finite differences,

$$f_{U,V}(u,v) = C(b_u, b_v) - C(b_u, a_v) - C(a_u, b_v) + C(a_u, a_v). \tag{7}$$

Let $c(u,v) = f_{U,V}(u,v)/(f_U(u)f_V(v))$, such that $f_{U,V}(u,v) = c(u,v)f_U(u)f_V(v)$, using a notation that is analogous to that of the joint density when (U, V) is continuous. Therefore, the conditional pmf, $f_{U|V}(u \mid v) = c(u,v)f_U(u)$.

To specify the distribution of base random vector (U_l, V_l), we use copula C_l with parameter η_l. For a parsimonious location-dependent model, we create spatially varying copulas $C_{\boldsymbol{v},l}$ on $(U_{\boldsymbol{v},l}, V_{\boldsymbol{v},l})$ by extending η_l to $\eta_l(\boldsymbol{v})$. In practice, we associate $\eta_l(\boldsymbol{v})$ to a spatial kernel that depends on $\boldsymbol{v} \in \mathcal{D}$ through a link function. Using Proposition 1 with a family of marginal pmfs $g_{\boldsymbol{v}}$, the joint pmf on $(U_{\boldsymbol{v},l}, V_{\boldsymbol{v},l})$ is $f_{U_{\boldsymbol{v},l},V_{\boldsymbol{v},l}}(u,v) = c_{\boldsymbol{v},l}(u,v)f_{U_{\boldsymbol{v},l}}(u)f_{V_{\boldsymbol{v},l}}(v)$, where $f_{U_{\boldsymbol{v},l}} = g_{\boldsymbol{v}}$ and $f_{V_{\boldsymbol{v},l}} = g_{\boldsymbol{v}_{(l)}}$, and the conditional pmf is $f_{\boldsymbol{v},l}(u \mid v) = c_{\boldsymbol{v},l}(u,v)g_{\boldsymbol{v}}(u)$. Finally, the conditional pmf of the discrete copula NNMP model is given by

$$p(y(\boldsymbol{v}) \mid \boldsymbol{y}_{\mathrm{Ne}(\boldsymbol{v})}) = \sum_{l=1}^{L} w_l(\boldsymbol{v})\; c_{\boldsymbol{v},l}(y(\boldsymbol{v}), y(\boldsymbol{v}_{(l)}))\; g_{\boldsymbol{v}}(y(\boldsymbol{v})), \tag{8}$$

where the marginal pmf for $Y(\boldsymbol{v})$ is $g_{\boldsymbol{v}}$.

Recall that an NNMP model involves two sets of locations, the reference and nonreference sets. As done in practice, we take the reference set $\mathcal{S}$ to correspond to the observed locations, and consider a generic finite set $\mathcal{U}$ such that $\mathcal{S} \cap \mathcal{U} = \emptyset$. Then, the joint pmf $p(\boldsymbol{y}_{\mathcal{V}})$ over set $\mathcal{V} = \mathcal{S} \cup \mathcal{U}$ describes the NNMP distribution over any finite set of locations that includes the observed locations. In general, for a discrete NNMP, an explicit expression for $p(\boldsymbol{y}_{\mathcal{V}})$ is not available, since working with a bivariate discrete distribution and its conditional pmf is difficult. However, using copulas to specify the bivariate mixture component yields a structured conditional pmf and allows for the study of the joint pmf. The following proposition provides an explicit expression for $p(\boldsymbol{y}_{\mathcal{V}})$ under a discrete copula NNMP. The proof of the proposition can be found in the Supporting Information.

Proposition 2. *Consider a discrete copula NNMP model for spatial process* $\{Y(\boldsymbol{v}) : \boldsymbol{v} \in \mathcal{D}\}$, *with* $\mathcal{S} = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\}$ *and* $\mathcal{U} = \{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m\}$, *where* $n \geq 3$, $m \geq 1$, *and* $\mathcal{S} \cap \mathcal{U} = \emptyset$. *Take* $\mathcal{V} = \mathcal{S} \cup \mathcal{U}$, *and let* $\boldsymbol{y}_{\mathcal{V}} = (y(\boldsymbol{s}_1), \ldots, y(\boldsymbol{s}_n), y(\boldsymbol{u}_1), \ldots, y(\boldsymbol{u}_m))^\top$. *Then the joint pmf of* $\boldsymbol{y}_{\mathcal{V}}$ *is* $p(\boldsymbol{y}_{\mathcal{V}}) = p(\boldsymbol{y}_{\mathcal{U}} \mid \boldsymbol{y}_{\mathcal{S}})\, p(\boldsymbol{y}_{\mathcal{S}})$, *where*

$$\begin{aligned} p(\boldsymbol{y}_{\mathcal{S}}) &= \prod_{i=1}^{n} g_{\boldsymbol{s}_i}(y(\boldsymbol{s}_i)) \sum_{l_n=1}^{n_L} \cdots \sum_{l_2=1}^{2_L} w_{\boldsymbol{s}_n,l_n} \cdots w_{\boldsymbol{s}_2,l_2} c_{\boldsymbol{s}_n,l_n} \cdots c_{\boldsymbol{s}_2,l_2}, \\ p(\boldsymbol{y}_{\mathcal{U}} \mid \boldsymbol{y}_{\mathcal{S}}) &= \prod_{i=1}^{m} g_{\boldsymbol{u}_i}(y(\boldsymbol{u}_i)) \sum_{\tilde{l}_m=1}^{L} \cdots \sum_{\tilde{l}_1=1}^{L} w_{\boldsymbol{u}_m,\tilde{l}_m} \cdots w_{\boldsymbol{u}_1,\tilde{l}_1} c_{\boldsymbol{u}_m,\tilde{l}_m} \cdots c_{\boldsymbol{u}_1,\tilde{l}_1}. \end{aligned} \tag{9}$$

where $w_{\boldsymbol{s}_i,l_i} \equiv w_{l_i}(\boldsymbol{s}_i)$ *and* $c_{\boldsymbol{s}_i,l_i} \equiv c_{\boldsymbol{s}_i,l_i}(y(\boldsymbol{s}_i), y(\boldsymbol{s}_{(i,l_i)}))$, *for* $l_i = 1, \ldots, i_L$, $i = 3, \ldots, n$, *and* $w_{\boldsymbol{u}_i,\tilde{l}_i} \equiv w_{\tilde{l}_i}(\boldsymbol{u}_i)$ *and* $c_{\boldsymbol{u}_i,\tilde{l}_i} \equiv c_{\boldsymbol{u}_i,\tilde{l}_i}(y(\boldsymbol{u}_i), y(\boldsymbol{u}_{(i,\tilde{l}_i)}))$, *for* $\tilde{l}_i = 1, \ldots, L$, $i = 1, \ldots, m$.

We note that Proposition 2 also applies when $\boldsymbol{y}_{\mathcal{V}}$ is continuous. It indicates that, given the sequence of pmfs $g_{\boldsymbol{v}}$, the joint pmf of $\boldsymbol{y}_{\mathcal{V}}$ is determined by the collection of bivariate copulas, motivating the use of different copula families to construct discrete NNMPs. To balance flexibility and scalability, our strategy is to take all copulas C_l in one family with the same link function for the copula parameters. Table 1 presents three examples with copula parameters modeled via

TABLE 1 Examples of spatial copulas $C_{\boldsymbol{\nu},l}$ and corresponding link functions, $k : \mathcal{D} \times \mathcal{D} \to [0,1]$.

	$C_{\boldsymbol{\nu},l}(z_1, z_2)$	Link function
Gaussian	$\Phi_2(\Phi^{-1}(z_1),\ \Phi^{-1}(z_2))$	$\rho_l(\boldsymbol{\nu}) = k(\boldsymbol{\nu}, \boldsymbol{\nu}_{(l)})$
Gumbel	$\exp(-\{(-\log z_1)^{\eta_l(\boldsymbol{\nu})} + (-\log z_2)^{\eta_l(\boldsymbol{\nu})}\}^{1/\eta_l(\boldsymbol{\nu})})$	$\eta_l(\boldsymbol{\nu}) = (1 - k(\boldsymbol{\nu}, \boldsymbol{\nu}_{(l)}))^{-1}$
Clayton	$(z_1^{-\delta_l(\boldsymbol{\nu})} + z_2^{-\delta_l(\boldsymbol{\nu})} - 1)^{-1/\delta_l(\boldsymbol{\nu})}$	$\delta_l(\boldsymbol{\nu}) = 2k(\boldsymbol{\nu}, \boldsymbol{\nu}_{(l)})/(1 - k(\boldsymbol{\nu}, \boldsymbol{\nu}_{(l)}))$

Note: The bivariate cdf Φ_2 corresponds to the standard bivariate Gaussian distribution with correlation $\rho \in (0,1)$, and the cdf Φ corresponds to the standard univariate Gaussian distribution.

a link function $k : \mathcal{D} \times \mathcal{D} \to [0,1]$. In particular, the Gumbel and Clayton copulas are asymmetric. They exhibit greater dependence in the positive and negative tails, respectively. In the first simulation example, we demonstrate that when the underlying spatial dependence is non-Gaussian, it may be appropriate to choose asymmetric copulas. We present next an example of a discrete copula NNMP construction.

Example 1. *Gaussian copula NNMP with negative binomial marginals*. For the family of marginal pmfs $g_{\boldsymbol{\nu}}$, consider the negative binomial distribution with mean $\alpha(\boldsymbol{\nu})$ and dispersion parameter r, denoted as $\text{NB}(\alpha(\boldsymbol{\nu}), r)$. Therefore, $g_{\boldsymbol{\nu}}(y) = \binom{y+r-1}{y} (p(\boldsymbol{\nu}))^r (1 - p(\boldsymbol{\nu}))^y$, with $p(\boldsymbol{\nu}) = r/(\alpha(\boldsymbol{\nu}) + r)$. To include a vector of covariates $\boldsymbol{x}(\boldsymbol{\nu})$, we take a log-link function for $\alpha(\boldsymbol{\nu})$ such that $\log(\alpha(\boldsymbol{\nu})) = \boldsymbol{x}(\boldsymbol{\nu})^\top \boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is a vector of regression parameters. We first specify Gaussian copulas C_l with correlation parameters ρ_l for the base random vectors (U_l, V_l). We then modify the correlation parameters ρ_l using a correlation function k for all l such that $\rho_l(\boldsymbol{\nu}) := k(\boldsymbol{\nu}, \boldsymbol{\nu}_{(l)})$, creating a sequence of spatially varying copulas $C_{\boldsymbol{\nu},l}$. The resulting model is given by (8) with $g_{\boldsymbol{\nu}} = \text{NB}(\alpha(\boldsymbol{\nu}), r)$.

In summary, as shown in Example 1, the construction of a discrete copula NNMP requires specification of a family of spatial copulas and that of marginal pmfs. More examples with different families of copulas and marginals are illustrated in Section 5.

3.3 | Inference for discrete copula NNMPs

We develop a framework for discrete copula NNMP inference, based on transforming the discrete random variables to continuous ones by adding auxiliary variables, using the continuous extension (CE) approach in Denuit and Lambert (2005). Working with continuous marginals improves computational efficiency and stability: the likelihood requires only nL bivariate copula density evaluations; and, computing the conditional pmf using the finite differences in (7) is bypassed, thus avoiding numerical instability especially for copulas that are not analytically available, such as the Gaussian copula. Moreover, this framework makes more efficient the key task of spatial prediction over unobserved sites by avoiding computation that involves inverting the conditional cdf based on (7).

We associate each $Y(\boldsymbol{\nu})$ with a continuous random variable $Y^*(\boldsymbol{\nu})$, such that $Y^*(\boldsymbol{\nu}) = Y(\boldsymbol{\nu}) - O(\boldsymbol{\nu})$, where $O(\boldsymbol{\nu})$ is a continuous uniform random variable on $(0,1)$, independent of $Y(\boldsymbol{\nu})$ and of $O(\boldsymbol{\nu}')$, for $\boldsymbol{\nu}' \neq \boldsymbol{\nu}$. We refer to $Y^*(\boldsymbol{\nu})$ as the continued $Y(\boldsymbol{\nu})$ by $O(\boldsymbol{\nu})$. Let $Q_{\boldsymbol{\nu}}$ and $g_{\boldsymbol{\nu}}$ be the marginal cdf and pmf of $Y(\boldsymbol{\nu})$, respectively. Then, the marginal cdf and density of $Y^*(\boldsymbol{\nu})$ are $Q^*_{\boldsymbol{\nu}}(y^*(\boldsymbol{\nu})) = Q_{\boldsymbol{\nu}}([y^*(\boldsymbol{\nu})]) + (y^*(\boldsymbol{\nu}) - [y^*(\boldsymbol{\nu})])g_{\boldsymbol{\nu}}([y^*(\boldsymbol{\nu})+1])$, and $g^*_{\boldsymbol{\nu}}(y^*(\boldsymbol{\nu})) = g_{\boldsymbol{\nu}}([y^*(\boldsymbol{\nu})+1])$, respectively, where $[x]$ denotes the integer part of x.

Based on marginal densities $g^*_{\boldsymbol{\nu}}$, we take spatial copulas $C^*_{\boldsymbol{\nu},l} = C_{\boldsymbol{\nu},l}$ for continuous random vectors $(U^*_{\boldsymbol{\nu},l}, V^*_{\boldsymbol{\nu},l})$, with marginals $f_{U^*_{\boldsymbol{\nu},l}} = g^*_{\boldsymbol{\nu}}$ and $f_{V^*_{\boldsymbol{\nu},l}} = g^*_{\boldsymbol{\nu}_{(l)}}$, using copulas $C_{\boldsymbol{\nu},l}$ from the original NNMP model. The joint density on $(U^*_{\boldsymbol{\nu},l}, V^*_{\boldsymbol{\nu},l})$ is $f_{U^*_{\boldsymbol{\nu},l}, V^*_{\boldsymbol{\nu},l}}(u,v) = c^*_{\boldsymbol{\nu},l}(u,v) g^*_{\boldsymbol{\nu}}(u) g^*_{\boldsymbol{\nu}_{(l)}}(v)$, and the conditional density is $f^*_{\boldsymbol{\nu},l}(u \mid v) = c^*_{\boldsymbol{\nu},l}(u,v) g^*_{\boldsymbol{\nu}}(u)$, where $c^*_{\boldsymbol{\nu},l}$ is the copula density. Denote by $\boldsymbol{y}^*_{\text{Ne}(\boldsymbol{\nu})}$ the vector that contains the continued elements of $\boldsymbol{y}_{\text{Ne}(\boldsymbol{\nu})}$, and $\boldsymbol{o}_{\text{Ne}(\boldsymbol{\nu})}$ the vector of auxiliary variables for elements of $\boldsymbol{y}_{\text{Ne}(\boldsymbol{\nu})}$. Then, the implied model on $y^*(\boldsymbol{\nu})$ is

$$p(y^*(\boldsymbol{\nu}) \mid D^*(\boldsymbol{\nu})) = \sum_{l=1}^{L} w_l(\boldsymbol{\nu})\ c^*_{\boldsymbol{\nu},l}(y^*(\boldsymbol{\nu}), y^*(\boldsymbol{\nu}_{(l)}))\ g^*_{\boldsymbol{\nu}}(y^*(\boldsymbol{\nu})), \tag{10}$$

where $y^*(\boldsymbol{\nu}) = y(\boldsymbol{\nu}) - o(\boldsymbol{\nu})$, and $D^*(\boldsymbol{\nu}) = \{\boldsymbol{y}^*_{\text{Ne}(\boldsymbol{\nu})}, o(\boldsymbol{\nu}), \boldsymbol{o}_{\text{Ne}(\boldsymbol{\nu})}\}$. Based on Proposition 1, model (10) has marginal density $g^*_{\boldsymbol{\nu}}$ for $Y^*(\boldsymbol{\nu})$. To recover $y(\boldsymbol{\nu})$, we first generate $y^*(\boldsymbol{\nu})$ from the extended model, and then set $y(\boldsymbol{\nu}) = [y^*(\boldsymbol{\nu}) + 1]$.

We note that our approach to continuous extension may appear to be similar to that for SGCMs (e.g., Hughes, 2015; Madsen, 2009) in which the CE is used for the data marginals with a single multivariate copula. However, unlike the SGCM, the discrete NNMP consists of many bivariate copulas. Each bivariate copula is associated with a mixture component vector whose marginals are location-dependent. Thus, the result of Proposition 1 is key both for modeling purposes as well as for inference for discrete copula NNMPs, using the CE approach.

Regarding the existing literature, statistical inference for spatial copula models based on the CE approach is typically conducted by maximizing the expected likelihood with respect to the auxiliary variables. As discussed in Hughes (2015), the expectation involves a potentially high-dimensional integral which requires Monte Carlo approximation; for particular models, the Monte Carlo sample size that determines the approximation error can also depend on the spatial dependence among the data. We develop inferential methods under the Bayesian framework. The Bayesian approach builds from the joint distribution of the data and auxiliary variables. This avoids the challenging task of specifying the Monte Carlo sample size to obtain a good approximation to the expectation, required under the maximum likelihood estimation approach. Moreover, posterior simulation based on (10) takes advantage of copula properties for continuous random variables, thus providing efficient computation for both model estimation and prediction.

4 | BAYESIAN IMPLEMENTATION

4.1 | Hierarchical model formulation

Assume that $\boldsymbol{y}_S = (y(\boldsymbol{s}_1), \ldots, y(\boldsymbol{s}_n))^\top$ is a realization of a discrete copula NNMP with spatially varying marginal pmfs through spatially dependent covariates, $g_{\boldsymbol{s}_i}(y(\boldsymbol{s}_i)) \equiv g(y(\boldsymbol{s}_i) \mid \boldsymbol{\beta}, \boldsymbol{\xi})$. Here, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$, where β_0 is an intercept and $(\beta_1, \ldots, \beta_p)^\top$ is the regression parameter vector for covariates $\boldsymbol{x}(\boldsymbol{s}_i)$, and $\boldsymbol{\xi}$ collects all other parameters of $g_{\boldsymbol{s}_i}$. The copula parameter is modeled through a link function k with parameter(s) $\boldsymbol{\phi}$.

We use the CE approach associating each $y(\boldsymbol{s}_i)$ with $y^*(\boldsymbol{s}_i)$, such that $y^*(\boldsymbol{s}_i) = y(\boldsymbol{s}_i) - o_i$, where $o_i \equiv o(\boldsymbol{s}_i)$ is uniformly distributed on $(0, 1)$, independent of $y(\boldsymbol{s}_i)$ and of o_j, for $j \neq i$. Moreover, denote by $\boldsymbol{\zeta}$ the parameter of the cutoff point kernel for the mixture weights, defined in Section 2.1.

The formulation of the mixture weights allows us to augment the model with a sequence of auxiliary variables, $\{t_i : i = 3, \ldots, n\}$, where t_i is a Gaussian random variable with mean $\mu(\boldsymbol{s}_i)$ and variance κ^2. The augmented model for the data can be expressed as

$$
\begin{aligned}
& y(\boldsymbol{s}_i) = y^*(\boldsymbol{s}_i) + o_i, \quad o_i \overset{i.i.d.}{\sim} \text{Unif}(0,1), \ i = 1, \ldots, n, \\
& y^*(\boldsymbol{s}_1) \mid \boldsymbol{\beta}, \boldsymbol{\xi} \ \sim g^*_{\boldsymbol{s}_1}(y^*(\boldsymbol{s}_1)), \quad y^*(\boldsymbol{s}_2) \mid y^*(\boldsymbol{s}_1), \boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{\phi} \sim f^*_{\boldsymbol{s}_2,1}(y^*(\boldsymbol{s}_2) \mid y^*(\boldsymbol{s}_1)), \\
& y^*(\boldsymbol{s}_i) \mid \{y^*(\boldsymbol{s}_{(il)})\}_{l=1}^{i_L}, t_i, \boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{\phi}, \boldsymbol{\zeta} \overset{ind.}{\sim} \sum_{l=1}^{i_L} f^*_{\boldsymbol{s}_i,l}(y^*(\boldsymbol{s}_i) \mid y^*(\boldsymbol{s}_{(il)}))\, \mathbb{1}_{(r^*_{\boldsymbol{s}_i,l-1}, r^*_{\boldsymbol{s}_i,l})}(t_i), \ i = 3, \ldots, n, \\
& t_i \mid \boldsymbol{\gamma}, \kappa^2 \overset{ind.}{\sim} N(t_i \mid \gamma_0 + \gamma_1 s_{i1} + \gamma_2 s_{i2}, \kappa^2), \ i = 3, \ldots, n,
\end{aligned}
$$

where $f^*_{\boldsymbol{s}_i,l}(y^*(\boldsymbol{s}_i) \mid y^*(\boldsymbol{s}_{(il)})) = c^*_{\boldsymbol{s}_i,l}(y^*(\boldsymbol{s}_i), y^*(\boldsymbol{s}_{(il)})) g^*_{\boldsymbol{s}_i}(y^*(\boldsymbol{s}_i))$, and $r^*_{\boldsymbol{s}_i,l} = \log\{r_{\boldsymbol{s}_i,l}/(1 - r_{\boldsymbol{s}_i,l})\}$, for $l = 1, \ldots, i_L$, and $i = 3, \ldots, n$.

The full Bayesian model is completed with prior specification for parameters $\boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{\phi}, \boldsymbol{\zeta}, \boldsymbol{\gamma} = (\gamma_0, \gamma_1, \gamma_2)^\top$ and κ^2. The priors for $\boldsymbol{\xi}$, $\boldsymbol{\phi}$, and $\boldsymbol{\zeta}$ depend on the choices of the pmf $g_{\boldsymbol{s}_i}$, the copula $C^*_{\boldsymbol{s}_i,l}$, and the kernel k', respectively. For parameters $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, and κ^2, we consider $N(\boldsymbol{\beta} \mid \mu_\beta, \boldsymbol{V}_\beta)$, $N(\boldsymbol{\gamma} \mid \mu_\gamma, \boldsymbol{V}_\gamma)$, and $\text{IG}(\kappa^2 \mid u_{\kappa^2}, v_{\kappa^2})$ priors, where IG denotes the inverse gamma distribution. We obtain the joint posterior distribution given by

$$
\begin{aligned}
p(\boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{\phi}, \boldsymbol{\zeta}, \boldsymbol{\gamma}, \kappa^2, \{t_i\}_{i=3}^n, \{o_i\}_{i=1}^n \mid \boldsymbol{y}_S) \propto\ & N(\boldsymbol{\beta} \mid \mu_\beta, \boldsymbol{V}_\beta) \times p(\boldsymbol{\xi}) \times p(\boldsymbol{\phi}) \times p(\boldsymbol{\zeta}) \times N(\boldsymbol{\gamma} \mid \mu_\gamma, \boldsymbol{V}_\gamma) \\
& \times \text{IG}(\kappa^2 \mid u_{\kappa^2}, v_{\kappa^2}) \times N(\boldsymbol{t} \mid \boldsymbol{D}\boldsymbol{\gamma}, \ \kappa^2 \mathbf{I}_{n-2}) \times \prod_{i=1}^{n} \text{Unif}(o_i \mid 0, 1) \times g^*_{\boldsymbol{s}_1}(y(\boldsymbol{s}_1) - o_1 \mid \boldsymbol{\beta}, \boldsymbol{\xi}) \\
& \times f^*_{\boldsymbol{s}_2,1}(y(\boldsymbol{s}_2) - o_2 \mid y(\boldsymbol{s}_1) - o_1, \boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{\phi}) \times \prod_{i=3}^{n} \sum_{l=1}^{i_L} f^*_{\boldsymbol{s}_i,l}(y(\boldsymbol{s}_i) - o_i \mid y(\boldsymbol{s}_{(il)}) - o_{(il)}, \boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{\phi}) \mathbb{1}_{(r^*_{\boldsymbol{s}_i,l-1}, r^*_{\boldsymbol{s}_i,l})}(t_i),
\end{aligned}
$$

where $p(\boldsymbol{\zeta})$, $p(\boldsymbol{\phi})$, and $p(\boldsymbol{\zeta})$ are priors on $\boldsymbol{\zeta}, \boldsymbol{\phi}, \boldsymbol{\zeta}$, respectively, $o_{(il)} \equiv o(\boldsymbol{s}_{(il)})$, the vector $\boldsymbol{t} = (t_3, \ldots, t_n)^\top$, and the matrix $\boldsymbol{D}$ is $(n-2) \times 3$ such that the ith row is $(1, s_{2+i,1}, s_{2+i,2})$.

4.2 | Model estimation, validation, and prediction

We outline the MCMC sampler for parameters $(\boldsymbol{\beta}, \boldsymbol{\xi}, \boldsymbol{\phi}, \boldsymbol{\zeta}, \boldsymbol{\gamma}, \kappa^2)$, and latent variables $\{t_i\}_{i=3}^n$ and $\{o_i\}_{i=1}^n$. We note that there is a set of configuration variables $\{\ell_i\}_{i=3}^n$ in one-to-one correspondence with t_i, that is, $\ell_i = l$ if and only if $t_i \in (r^*_{\boldsymbol{s}_i,l-1}, r^*_{\boldsymbol{s}_i,l})$, for $l = 1, \ldots, i_L$.

The updates for parameters $\boldsymbol{\beta}$, $\boldsymbol{\xi}$, and $\boldsymbol{\phi}$ require Metropolis steps, since they enter in copula densities $c^*_{\boldsymbol{s}_i,l}$. We use a Metropolis step also for kernel k' parameter $\boldsymbol{\zeta}$, which is involved in the definition of the mixture weights. The posterior full conditional distribution of $\boldsymbol{\gamma}$ is $N(\boldsymbol{\gamma} \mid \boldsymbol{\mu}^*_{\boldsymbol{\gamma}}, \boldsymbol{V}^*_{\boldsymbol{\gamma}})$, where $\boldsymbol{V}^*_{\boldsymbol{\gamma}} = (\boldsymbol{V}^{-1}_{\boldsymbol{\gamma}} + \kappa^{-2}\boldsymbol{D}^\top\boldsymbol{D})^{-1}$ and $\boldsymbol{\mu}^*_{\boldsymbol{\gamma}} = \boldsymbol{V}^*_{\boldsymbol{\gamma}}(\boldsymbol{V}^{-1}_{\boldsymbol{\gamma}}\boldsymbol{\mu}_{\boldsymbol{\gamma}} + \kappa^{-2}\boldsymbol{D}^\top\boldsymbol{t})$. An inverse gamma prior for κ^2 yields conjugate posterior full conditional distribution $\mathrm{IG}(\kappa^2 \mid u_{\kappa^2} + (n-2)/2, v_{\kappa^2} + \sum_{i=3}^n (t_i - \mu(\boldsymbol{s}_i))^2/2)$.

Regarding the updates for latent variables, the posterior full conditional distribution for each t_i can be expressed as $\sum_{l=1}^{i_L} q_l(\boldsymbol{s}_i)\ \mathrm{TN}(t_i \mid \mu(\boldsymbol{s}_i), \kappa^2; r^*_{\boldsymbol{s}_i,l-1} < t_i \le r^*_{\boldsymbol{s}_i,l})$, for $i = 3, \ldots, n$, where TN denotes the truncated normal distribution over the indicated interval, and $q_l(\boldsymbol{s}_i) \propto w_l(\boldsymbol{s}_i)\ c^*_{\boldsymbol{s}_i,l}(y^*(\boldsymbol{s}_i), y^*(\boldsymbol{s}_{(il)}))$, for $l = 1, \ldots, i_L$. Hence, each t_i can be readily updated by sampling from the lth truncated normal with probability proportional to $q_l(\boldsymbol{s}_i)$. For variables o_i, the posterior full conditional distribution of o_1 is proportional to $\prod_{\{j:\boldsymbol{s}_{(j,\ell_j)}=\boldsymbol{s}_1\}} c^*_{\boldsymbol{s}_j,\ell_j}(y(\boldsymbol{s}_j) - o_j, y(\boldsymbol{s}_1) - o_1)$, and that of o_i, $i \ge 2$, is proportional to $c^*_{\boldsymbol{s}_i,\ell_i}(y(\boldsymbol{s}_i) - o_i, y(\boldsymbol{s}_{(i,\ell_i)}) - o_{(i,\ell_i)}) \prod_{\{j:\boldsymbol{s}_{(j,\ell_j)}=\boldsymbol{s}_i\}} c^*_{\boldsymbol{s}_j,\ell_j}(y(\boldsymbol{s}_j) - o_j, y(\boldsymbol{s}_i) - o_i)$, where $\ell_2 = 1$ and $o_{(i,\ell_i)} \equiv o(\boldsymbol{s}_{(i,\ell_i)})$. We update each o_i with an independent Metropolis step with a $\mathrm{Unif}(0,1)$ proposal distribution.

The likelihood of the continued model admits the form $g_{\boldsymbol{s}_1}(y^*(\boldsymbol{s}_1)) \prod_{i=2}^n p(y^*(\boldsymbol{s}_i) \mid D^*(\boldsymbol{s}_i))$. The product formulation allows for model validation, using a generalization of the randomized quantile residuals proposed by Dunn and Smyth (1996) for independent data. Specifically, we define the marginal quantile residual, $r_1 = \Phi^{-1}(Q^*_{\boldsymbol{s}_1}(y^*(\boldsymbol{s}_1)))$, and the ith conditional quantile residual, $r_i = \Phi^{-1}(F(y^*(\boldsymbol{s}_i) \mid D^*(\boldsymbol{s}_i)))$, $i = 2, \ldots, n$, where F is the conditional cdf of $y^*(\boldsymbol{s}_i)$. If the model is correctly specified, the residuals r_i, $i = 1, \ldots, n$, would be independent and identically distributed as a standard Gaussian distribution.

Finally, we turn to posterior predictive inference at a new location $\boldsymbol{v}_0$. If $\boldsymbol{v}_0 \notin S$, for each posterior sample, we first compute the cutoff points $r_{\boldsymbol{v}_0,l}$, such that $r_{\boldsymbol{v}_0,l} - r_{\boldsymbol{v}_0,l-1} = k'(\boldsymbol{v}_0, \boldsymbol{v}_{(0l)}) / \sum_{l=1}^L k'(\boldsymbol{v}_0, \boldsymbol{v}_{(0l)})$, and the weights $w_l(\boldsymbol{v}_0) = G_{\boldsymbol{v}_0}(r_{\boldsymbol{v}_0,l}) - G_{\boldsymbol{v}_0}(r_{\boldsymbol{v}_0,l-1})$, for $l = 1, \ldots, L$. We then generate $y^*(\boldsymbol{v}_0)$ based on (10), and set $y(\boldsymbol{v}_0) = [y^*(\boldsymbol{v}_0) + 1]$. If $\boldsymbol{v}_0 \equiv \boldsymbol{s}_i \in S$, we generate $y(\boldsymbol{v}_0)$ similarly, the difference being that we now use the posterior samples for the mixture weights obtained from the MCMC algorithm.

5 | DATA ILLUSTRATIONS

To illustrate the proposed methodology, we present two synthetic data examples and a real data analysis. The goal of the first simulation experiment is to investigate the flexibility of discrete copula NNMPs, using different copula functions to define the NNMP mixture components. In the second experiment, we demonstrate the inferential and computational advantages of our approach for count data modeling, compared to SGLMMs. Implementation details for the models are provided in the Supporting Information. Since our purpose is primarily demonstrative, we took $L = 10$ for the simulation experiments. A comprehensive sensitivity analysis for L was conducted for the real data application of Section 5.3, with details provided in the Supporting Information.

In both simulated data examples, we ran the MCMC algorithm for each copula NNMP model for 20,000 iterations, discarding the first 4000 iterations, and collecting posterior samples every four iterations. The SGLMM models were implemented using the spBayes package in R (Finley et al., 2007); we ran the algorithm for 40,000 iterations and collected posterior samples every five iterations, with the first 20,000 as burn-in.

We compare models based on parameter estimates, root mean squared prediction error (RMSPE), 95% credible interval width (95% CI width), 95% credible interval coverage rate (95% CI cover), continuous ranked probability score (CRPS; Gneiting and Raftery, 2007), energy score (ES; Gneiting and Raftery, 2007), and variogram score of order one (VS; Scheuerer and Hamill, 2015). The energy score is a multivariate extension of the CRPS, while the variogram score

examines pairwise differences of the components of the multivariate quantity. Both the ES and VS allow for comparison of model predictive performance with respect to dependence structure.

5.1 | First simulation experiment

We first generated sites over a regular grid of 120×120 resolution on a unit square domain, and then simulated data from $y(\boldsymbol{\nu}) = F_Y^{-1}(F_Z(z(\boldsymbol{\nu})))$, where F_Y corresponds to the Poisson distribution with rate parameter $\lambda_0 = 5$, and $z(\boldsymbol{\nu})$ is the skew-Gaussian random field from Zhang and El-Shaarawi (2010) with stationary marginal distribution F_Z. More specifically, $z(\boldsymbol{\nu}) = \sigma_1 \, |\omega_1(\boldsymbol{\nu})| + \sigma_2 \, \omega_2(\boldsymbol{\nu})$, where both $\omega_1(\boldsymbol{\nu})$ and $\omega_2(\boldsymbol{\nu})$ are standard Gaussian processes with exponential correlation function based on range parameter 0.1. The density of F_Z is $f_Z(z) = 2\, N(z \mid 0, \sigma_1^2 + \sigma_2^2)\, \Phi(\sigma_1 z/(\sigma_2\sqrt{\sigma_1^2 + \sigma_2^2}))$, where $\sigma_1 \in \mathbb{R}$ controls the skewness, and $\sigma_2 > 0$ is a scale parameter. We took $\sigma_2 = 1$, and set $\sigma_1 = 1, 3, 10$ which corresponds to different levels of skewness.

We considered three discrete copula NNMPs with stationary Poisson marginals, that is, $g_{\boldsymbol{\nu}} = f_Y$, for all $\boldsymbol{\nu}$, where f_Y is the Poisson pmf with rate λ. The three models correspond to the copulas in Table 1, with the link function k given by an exponential correlation function with range parameter denoted by ϕ_1, ϕ_2, and ϕ_3 for the Gaussian, Gumbel, and Clayton copula models, respectively. We specified the cutoff point kernel through an exponential correlation function with range parameter ζ_1, ζ_2, and ζ_3 for the Gaussian, Gumbel, and Clayton copula models, respectively. The Bayesian models are fully specified with an IG(3, 1) prior for the ϕ and ζ parameters, and with $N(\boldsymbol{\gamma} \mid (-1.5, 0, 0)^\top,\ 2\mathbf{I}_3)$ and $\mathrm{IG}(\kappa^2 \mid 3, 1)$ priors. Finally, the prior for the rate parameter λ was taken as $\mathrm{Ga}(1, 1)$, where $\mathrm{Ga}(a, b)$ denotes the gamma distribution with mean a/b. We simulated 1000 responses and used 800 of them to fit the three NNMP models. The remaining 200 observations were used for model comparison.

Table 2 provides estimates for the rate parameter λ of the Poisson marginal distribution, and predictive performance metrics. For all three cases for $\sigma_1 = 1, 3, 10$, the Gumbel model yields the more accurate estimates for λ. In particular, the Gumbel model's 95% credible intervals include the true parameter value, whereas those of the Gaussian and Clayton models failed to cover it when $\sigma_1 = 1$ and $\sigma_1 = 10$, respectively. Regarding predictive performance, the Gumbel model outperforms to a smaller or larger extent the other two models across different scenarios. Predictive random fields under the three models are provided in the Supporting Information. We found that prediction by the Clayton model was not able to recover large values. Compared to the Gaussian model, the Gumbel model recovered large values slightly better. Overall, this example demonstrates that, when the underlying spatial dependence is driven by non-Gaussian processes, it is practically useful to consider copulas from asymmetric families, including use of appropriate model comparison tools.

5.2 | Second simulation experiment

We generated data over a grid of sites with 120×120 resolution, uniformly on the square $[0, 1] \times [0, 1]$, using a Poisson SGLMM with $y(\boldsymbol{\nu}) \mid \eta(\boldsymbol{\nu}) \sim \mathrm{Pois}(\eta(\boldsymbol{\nu}))$, and $\log(\eta(\boldsymbol{\nu})) = \beta_0 + \nu_1\beta_1 + \nu_2\beta_2 + z(\boldsymbol{\nu})$, where $\boldsymbol{\nu} = (\nu_1, \nu_2)$, and $z(\boldsymbol{\nu})$ is a

TABLE 2 First simulation example: Posterior mean and 95% credible interval estimates for the rate parameter λ of the Poisson NNMP marginal distribution, and scores for comparison of Gaussian, Gumbel, and Clayton copula NNMP models, under each of the three simulation scenarios for σ_1.

	$\sigma_1 = 1$			$\sigma_1 = 3$			$\sigma_1 = 10$		
	λ			λ			λ		
Gaussian	4.55 (4.16, 4.94)			4.71 (4.37, 5.07)			4.88 (4.55, 5.22)		
Gumbel	4.78 (4.39, 5.21)			4.88 (4.56, 5.24)			4.94 (4.66, 5.23)		
Clayton	5.33 (4.99, 5.68)			5.25 (4.96, 5.56)			5.36 (5.08, 5.65)		
	CRPS	**ES**	**VS**	**CRPS**	**ES**	**VS**	**CRPS**	**ES**	**VS**
Gaussian	0.69	12.77	94,855	0.85	15.54	124,893	0.93	16.98	138,592
Gumbel	0.69	12.58	92,278	0.85	15.32	120,932	0.92	16.71	134,774
Clayton	0.75	14.34	125,800	0.90	17.36	164,148	1.00	18.70	174,123

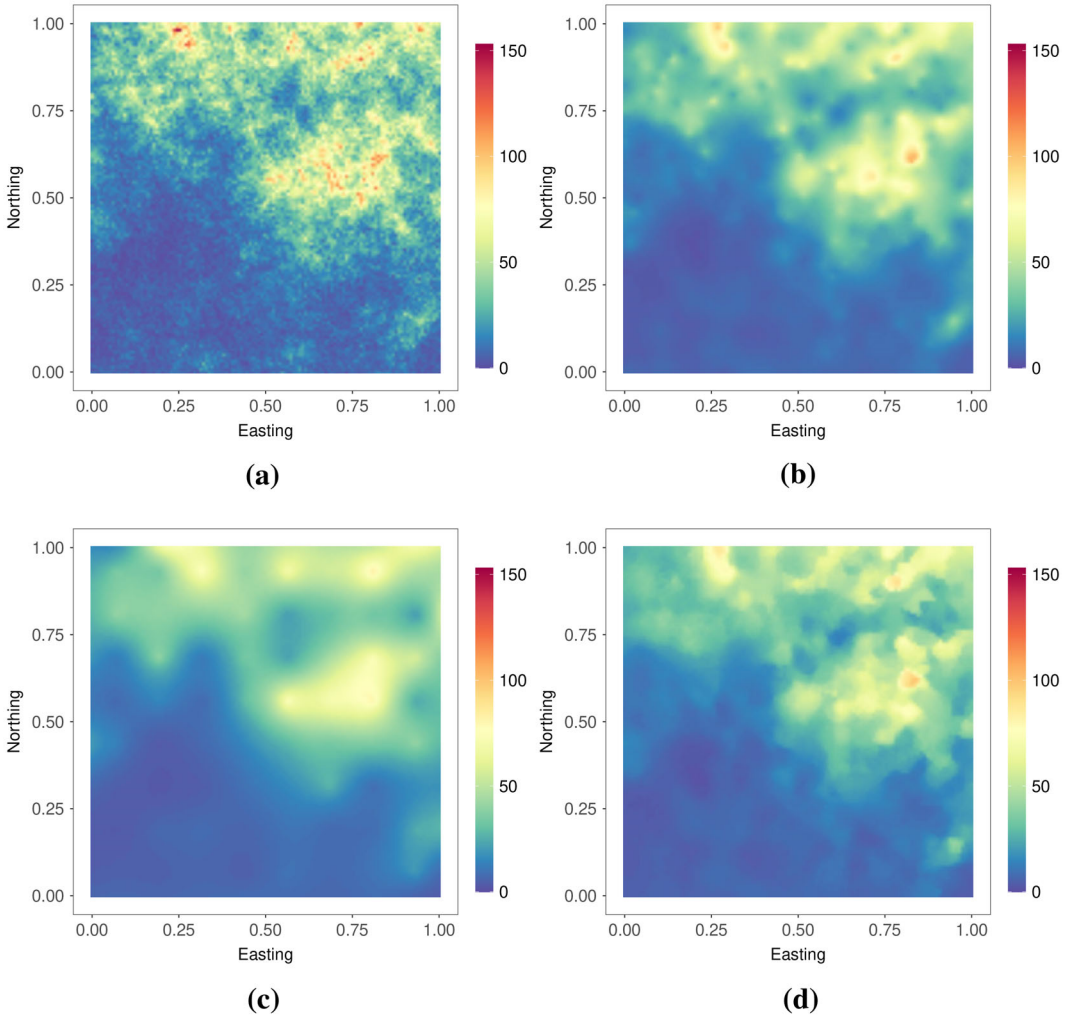

FIGURE 1 Second simulation example. Interpolated surfaces of the true model and posterior median estimates of the SGLMM-GP, SGLMM-GPP, and NBNNMP models. (a) True $y(\boldsymbol{\nu})$. (b) SGLMM-GP. (c) SGLMM-GPP. (d) NBNNMP.

zero-centered Gaussian process (GP) with variance parameter $\sigma^2 = 0.2$ and an exponential correlation function with range parameter $\phi_0 = 1/12$. We set the regression coefficients $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^\top = (1.5, 1, 2)^\top$, resulting in a random field with a trend, as shown in Figure 1a.

We considered three models. The first is the negative binomial NNMP model (NBNNMP) with a Gaussian copula, as discussed in Example 1. The second model (SGLMM-GP) is a Poisson SGLMM with a GP prior assigned to $z(\boldsymbol{\nu})$. For the last model (SGLMM-GPP), we considered a Poisson SGLMM with spatial random effects $z(\boldsymbol{\nu})$ corresponding to a Gaussian predictive process (GPP; Banerjee et al., 2008), with 10×10 knots placed on a grid over the domain. We chose the number of knots such that the computing times for the SGLMM-GPP and NBNNMP models are similar. As in the first simulation example, all models were fit to 800 observations and compared on the basis of 200 additional observations.

The regression coefficients for all models were assigned mean-zero, dispersed normal priors. We worked with an exponential correlation function for all models, used for $\rho_l(\boldsymbol{\nu})$ of the Gaussian copula in the NBNNMP model, and as the correlation function for the GP and GPP in the SGLMMs. The range parameter was assigned an inverse gamma prior IG(3, 1) for the NBNNMP model, and a uniform prior Unif(1/30, 1/3) for the other two models. The cutoff point kernel of the NBNNMP was also specified an exponential correlation function, with an IG(3, 1) prior for the range parameter. The variance parameter for the SGLMM models was assigned an inverse gamma prior IG(2, 1). For the logit Gaussian distribution parameters $\boldsymbol{\gamma}$ and κ^2 of the NBNNMP, we used $N((-1.5, 0, 0)^\top,\ 2\mathbf{I}_3)$ and IG(3, 1) priors, respectively. Finally, we placed a Ga(1, 1) prior on the NBNNMP dispersion parameter r.

TABLE 3 Second simulation example: Posterior mean and 95% credible interval estimates for the regression parameters, performance metrics, and computing time, under the NBNNMP model and the two SGLMM models.

	True	NBNNMP	SGLMM-GP	SGLMM-GPP
β_0	1.5	1.61 (1.29, 1.97)	1.53 (1.22, 1.81)	1.41 (1.02, 1.73)
β_1	1	0.90 (0.51, 1.31)	0.70 (0.25, 1.15)	0.91 (0.43, 1.34)
β_2	2	1.94 (1.51, 2.32)	2.18 (1.91, 2.53)	2.25 (1.81, 2.84)
RMSPE	-	9.06	8.88	10.00
95% CI cover	-	0.98	0.97	0.78
95% CI width	-	37.02	32.24	19.02
CRPS	-	4.58	4.52	5.37
ES	-	92.07	91.41	107.46
VS	-	5,175,591	5,199,629	6,378,263
Time (min)	-	11.18	935.02	11.68

Estimates of the regression parameters and performance metrics for out-of-sample prediction are provided in Table 3. We observe that, overall, the NBNNMP model provided the more accurate estimation for $\boldsymbol{\beta}$. Regarding predictive performance, the NBNNMP model outperformed the SGLMM-GPP model by a large margin, and was comparable to the SGLMM-GP model, which corresponds to the data generating process for this simulation experiment. Moreover, the last row of the table highlights the NBNNMP model's huge gains in computing time compared to the SGLMM-GP model.

Figure 1b–d plots the posterior median estimates of the random field for the three models. The SGLMM-GPP yields an overly smooth estimate, whereas the SGLMM-GP and NBNNMP models provide similar estimates that approximate well the true surface. Overall, this example illustrates the inferential and computational advantages of discrete copula NNMPs for modeling count data.

5.3 | North American Breeding Bird Survey data analysis

The primary source of information on population evolution for birds is count data surveys. Since 1966, the North American Breeding Bird Survey (BBS) has been conducted to monitor bird population change. There are over 4000 sampling units in the survey, each with a 24.5-mile roadside route. Along each route, volunteer observers count the number of birds by sight or sound, in a 3-min period at each of 50 stops (Pardieck et al., 2020). The BBS data are often used to determine temporal or geographical patterns of relative abundance. Spatial maps of relative abundance are crucial for ecological studies.

We are interested in the Northern Cardinal, a bird species that is prevalent in Eastern United States. Figure 2a shows the number of birds observed in 2019, with the sizes of the circle radii proportional to the number of birds at each sampling location. The dataset was extracted with the help of the R package *bbsAssistant* (Burnett et al., 2019); it contains 1515 irregular sampling locations. From Figure 2a we observe that the counts tend to increase as latitude decreases, and we thus take latitude as a covariate to account for the long range variability in the population.

We considered the Gaussian copula NBNNMP model defined in Example 1, with spatially varying marginal $\text{NB}(\exp(\boldsymbol{x}(\boldsymbol{v})^\top\boldsymbol{\beta}), r)$, where $\boldsymbol{\beta} = (\beta_0, \beta_1)^\top$. We used the same link functions and prior specifications as in Section 5.2. We first examined model performance under different values of L. Overall, parameter estimates were quite robust. The estimates of mixture weights suggested that the effective number of neighbors for each location was quite consistent for L between 10 and 20. Also, there was no discernible differences for out-of-sample predictive performance. Therefore, we took $L = 20$ as a reasonable upper bound. We also compared NBNNMP models with the three copulas listed in Table 1, using the same link functions for copulas as in Section 5.1. The three models were evaluated based on their predictive performance. Overall, the Gaussian copula outperformed the other two. Details of these analyses are provided in the Supporting Information.

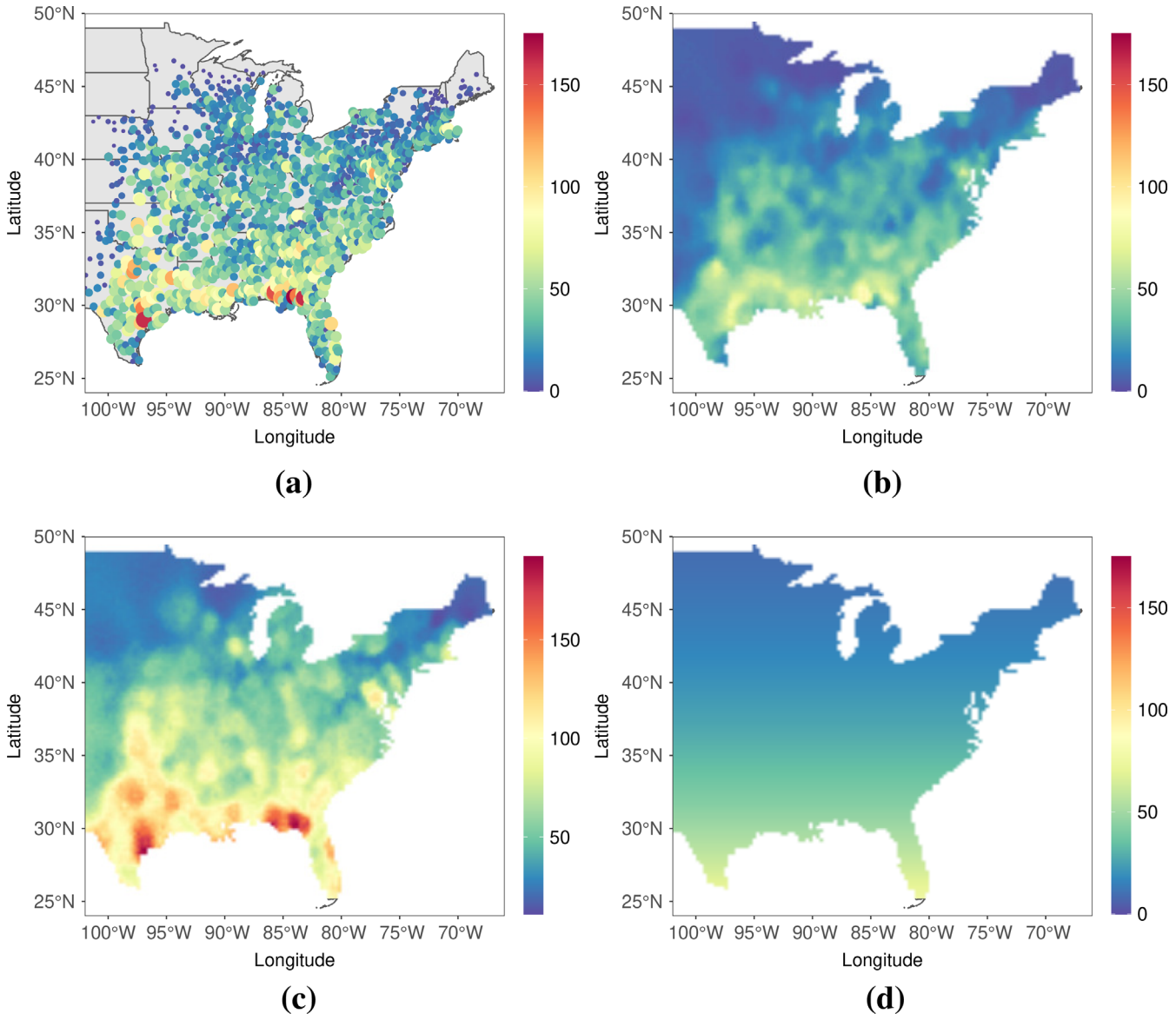

FIGURE 2 North American Breeding Bird Survey data analysis: (a) observed counts for 2019 BBS of Northern Cardinal, with circle radius proportional to the observed counts; (b) median of the posterior predictive distribution for Northern Cardinal count; (c) 95% CI widths of the posterior predictive distribution for Northern Cardinal count; (d) posterior mean of $\exp(\boldsymbol{x}(\boldsymbol{v})^\top \boldsymbol{\beta})$.

We proceeded to analyze the BBS data with the Gaussian copula NBNNMP model with $L = 20$. The posterior mean and 95% credible interval estimates of the regression parameters β_0 and β_1 are 6.53 (5.61, 7.38) and −0.09 (−0.11, −0.06), respectively, suggesting an increasing trend in the Northern Cardinal counts as the latitude decreases. The corresponding estimates of the dispersion parameter r are 1.88 (1.55, 2.21), indicating that there is overdispersion over the domain. Figure 2b,c show the posterior predictive median of the counts and the 95% posterior predictive credible interval width, respectively. Figure 2b displays the domain's spatial variability. The estimated uncertainty, as shown in Figure 2c, is meaningful, as areas with high uncertainty correspond to those where the observed counts are quite heterogeneous. Figure 2d provides a spatial map of the mean of the negative binomial marginals, which depicts a North–South trend. Model checking results are shown in Figure 3, including a posterior summary of the Gaussian quantile-quantile plot, and the histogram and spatial plot of the posterior means of the residuals. The results suggest good model fit.

Finally, we compared the NBNNMP with the SGLMM-GP model (details are given in the Supporting Information). The parameter estimates of $\boldsymbol{\beta}$ were quite close under the two models. On the other hand, the NBNNMP model resulted

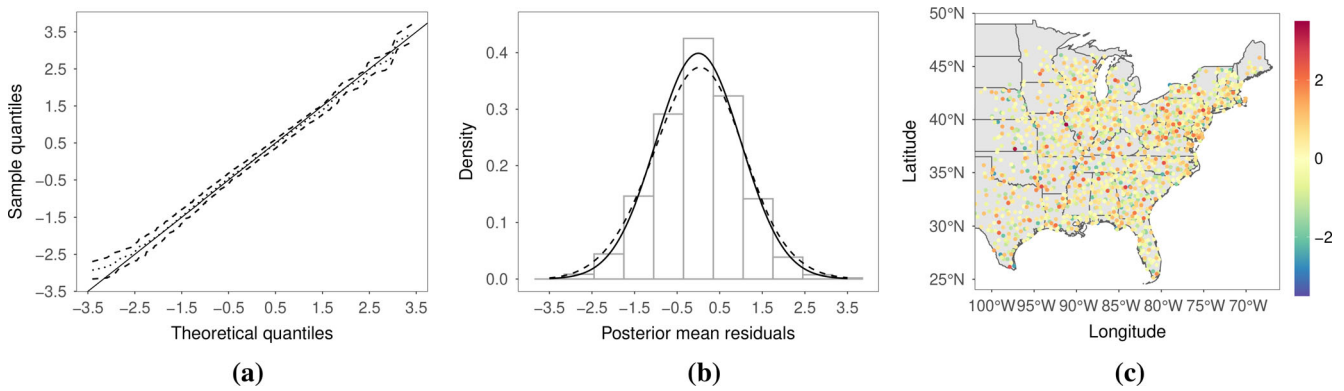

FIGURE 3 North American Breeding Bird Survey data analysis. Randomized quantile residual analysis: (a) Dotted and dashed lines correspond to the posterior mean and 95% interval bands, respectively; (b) solid and dashed lines are the standard Gaussian density and the kernel density estimate of the posterior means of the residuals, respectively; (c) spatial plot of the posterior means of the residuals. (a) Quantile-quantile plot. (b) Histogram.

in better out-of-sample predictive performance, and, notably, it was substantially more efficient to implement, with computing time 110 times faster than that for the SGLMM-GP model.

6 | DISCUSSION

We have introduced a new class of models for discrete geostatistical data, with particular focus on using different families of bivariate copulas to build modeling and inference. Compared to traditional SGLMM methods, the proposed class of models is scalable, and is able to accommodate complex dependence structures.

It is worth mentioning that the discrete copula NNMPs allow for the use of any family of bivariate copulas under a simple model structure. This compares favorably to discrete vine copula models (Panagiotelis et al., 2012) where a multivariate pmf is decomposed into bivariate copulas under trees, which results in an approach that, in the spatial context, would be much more complicated than the one proposed in this article. Furthermore, the computational burden for their model likelihood evaluations grows quadratically in n.

This work considers the strategy of using a single copula family for all bivariate distributions. Exploring the alternative which builds from different copula families for the bivariate distributions remains an interesting question to investigate. We can cast this as a model selection problem and develop algorithms to select models; see examples in Panagiotelis et al. (2017) and Gruber and Czado (2018) in the context of regular vine copula models. Different copula families for bivariate distributions yield more flexibility for the model to capture complex dependence, albeit at the cost of computational scalability. If the main purpose of the application is prediction, rather than model selection, one could explore calibrating the prediction using all candidate copula families. This could be done, for example, with the pseudo Bayesian model averaging approach, where the weight for each model is estimated based on stacking (Yao et al., 2018).

Regarding implementation, inference for discrete copula NNMPs is conducted based on the CE approach. This approach may allow discrete copula NNMPs to make use of alternative algorithms for faster computation, which are currently being developed for continuous NNMP models. Moreover, with the CE approach, it is possible to develop a class of NNMPs for a multivariate response that consists of both continuous and discrete components, while at the same time retaining computational efficiency.

Finally, we remark on some general research directions. Discrete NNMPs provide direct spatial modeling of discrete data. In fact, they can also be used for latent process modeling, for example, by introducing a latent process for the probability field for binary data or the intensity field for count data. This is similar to the common use of SGLMMs, the difference being that discrete NNMPs assume conditionally dependent observations. For spatial classification problems (Berrett & Calder, 2016), a comparison between the discrete NNMP and the SGLMM is of interest, in particular for recently proposed SGLMMs that consider nearest-neighbor approaches for the latent process; see, for example, Saha et al. (2022). In

this context, an interesting research direction is to explore different bivariate Bernoulli distributions (e.g., Dai et al., 2013) to construct NNMPs, as an alternative to using copulas.

ACKNOWLEDGMENTS

This work is part of the Ph.D. dissertation of X. Zheng, completed at University of California, Santa Cruz. The research was supported in part by the National Science Foundation under awards MMS 2050012, DMS 2153277, and SES 1950902. The authors wish to thank two reviewers and an associate editor for useful comments. Open access publishing facilitated by University of Wollongong, as part of the Wiley - University of Wollongong agreement via the Council of Australian University Librarians.

REFERENCES

Bai, Y., Kang, J., & Song, P. X.-K. (2014). Efficient pairwise composite likelihood estimation for spatial-clustered data. *Biometrics*, *70*(3), 661–670.

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.

Berrett, C., & Calder, C. A. (2016). Bayesian spatial binary classification. *Spatial Statistics*, *16*, 72–102.

Burnett, J. L., Wszola, L., & Palomo-Muñoz, G. (2019). BbsAssistant: An R package for downloading and handling data and information from the North American Breeding Bird Survey. *Journal of Open Source Software*, *4*(44), 1768.

Christensen, O. F., Roberts, G. O., & Sköld, M. (2006). Robust Markov chain Monte Carlo methods for spatial generalized linear mixed models. *Journal of Computational and Graphical Statistics*, *15*(1), 1–17.

Christensen, O. F., & Waagepetersen, R. (2002). Bayesian prediction of spatial count data using generalized linear mixed models. *Biometrics*, *58*(2), 280–286.

Dai, B., Ding, S., & Wahba, G. (2013). Multivariate Bernoulli distribution. *Bernoulli*, *19*(4), 1465–1483.

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.

Denuit, M., & Lambert, P. (2005). Constraints on concordance measures in bivariate discrete data. *Journal of Multivariate Analysis*, *93*(1), 40–57.

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.

Dunn, P. K., & Smyth, G. K. (1996). Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, *5*(3), 236–244.

Finley, A. O., Banerjee, S., & Carlin, B. P. (2007). spBayes: An R package for univariate and multivariate hierarchical point-referenced spatial models. *Journal of Statistical Software*, *19*(4), 1–24.

Genest, C., & Nešlehová, J. (2007). Aprimer on copulas for count data. *ASTIN Bulletin: The Journal of the IAA*, *37*(2), 475–515.

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*(477), 359–378.

Gruber, L. F., & Czado, C. (2018). Bayesian model selection of regular vine copulas. *Bayesian Analysis*, *13*(4), 1111–1135.

Guan, Y., & Haran, M. (2018). A computationally efficient projection-based approach for spatial generalized linear mixed models. *Journal of Computational and Graphical Statistics*, *27*(4), 701–714.

Han, Z., & De Oliveira, V. (2016). On the correlation structure of Gaussian copula models for geostatistical count data. *Australian & New Zealand Journal of Statistics*, *58*(1), 47–69.

Han, Z., & De Oliveira, V. (2019). Maximum likelihood estimation of Gaussian copula models for geostatistical count data. *Communications in Statistics-Simulation and Computation*, *49*, 1957–1981.

Hughes, J. (2015). copCAR: A flexible regression model for areal data. *Journal of Computational and Graphical Statistics*, *24*(3), 733–755.

Jin, B., Peruzzi, M., & Dunson, D. B. (2023). Bag of DAGs: Inferring directional dependence in spatiotemporal processes. https://doi.org/10.48550/arXiv.2112.11870

Joe, H. (2014). *Dependence modeling with copulas*. Boca Raton, FL: CRC Press.

Katzfuss, M., & Guinness, J. (2021). A general framework for Vecchia approximations of Gaussian processes. *Statistical Science*, *36*(1), 124–141.

Kazianka, H. (2013). Approximate copula-based estimation and prediction of discrete spatial data. *Stochastic Environmental Research and Risk Assessment*, *27*(8), 2015–2026.

Kazianka, H., & Pilz, J. (2010). Copula-based geostatistical modeling of continuous and discrete data including covariates. *Stochastic Environmental Research and Risk Assessment*, *24*(5), 661–673.

Madsen, L. (2009). Maximum likelihood estimation of regression parameters with spatially dependent discrete data. *Journal of Agricultural, Biological, and Environmental Statistics*, *14*(4), 375–391.

Panagiotelis, A., Czado, C., & Joe, H. (2012). Pair copula constructions for multivariate discrete data. *Journal of the American Statistical Association*, *107*(499), 1063–1072.

Panagiotelis, A., Czado, C., Joe, H., & Stöber, J. (2017). Model selection for discrete regular vine copulas. *Computational Statistics & Data Analysis*, *106*, 138–152.

Pardieck, K., Ziolkowski, D., Jr., Lutmerding, M., Aponte, V., & Hudson, M. (2020). *North American Breeding Bird Survey dataset 1966–2019: US Geological Survey data release*.

Peruzzi, M., Banerjee, S., & Finley, A. O. (2022). Highly scalable Bayesian geostatistical modeling via meshed Gaussian processes on partitioned domains. *Journal of the American Statistical Association*, *117*(538), 969–982.

Recta, V., Haran, M., & Rosenberger, J. L. (2012). A two-stage model for incidence and prevalence in point-level spatial count data. *Environmetrics*, *23*(2), 162–174.

Saha, A., Datta, A., & Banerjee, S. (2022). Scalable predictions for spatial probit linear mixed models using nearest neighbor Gaussian processes. *Journal of Data Science*, *20*(4), 533–544.

Scheuerer, M., & Hamill, T. M. (2015). Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities. *Monthly Weather Review*, *143*(4), 1321–1334.

Sengupta, A., & Cressie, N. (2013). Hierarchical statistical modeling of big spatial datasets using the exponential family of distributions. *Spatial Statistics*, *4*, 14–44.

Sengupta, A., Cressie, N., Kahn, B. H., & Frey, R. (2016). Predictive inference for big, spatial, non-Gaussian data: Modis cloud data and its change-of-support. *Australian & New Zealand Journal of Statistics*, *58*(1), 15–45.

Sklar, M. (1959). Fonctions de repartition an dimensions et leurs marges. *Publications de l'Institut de Statistique de L'Université de Paris*, *8*, 229–231.

Smith, M. S., & Khaled, M. A. (2012). Estimation of copula models with discrete margins via Bayesian data augmentation. *Journal of the American Statistical Association*, *107*(497), 290–303.

Song, P. X.-K., Li, M., & Yuan, Y. (2009). Joint regression analysis of correlated data using Gaussian copulas. *Biometrics*, *65*(1), 60–68.

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.

Wikle, C. K. (2002). *Spatial modeling of count data: A case study in modelling breeding bird survey data on large spatial domains*. In A. Lawson & D. Denison (Eds.), *Spatial cluster modelling* (pp. 199–209). Chapman & Hall.

Wu, G., Holan, S. H., Nilon, C. H., & Wikle, C. K. (2015). Bayesian binomial mixture models for estimating abundance in ecological monitoring studies. *The Annals of Applied Statistics*, *9*(1), 1–26.

Yang, L., Frees, E. W., & Zhang, Z. (2020). Nonparametric estimation of copula regression models with discrete outcomes. *Journal of the American Statistical Association*, *115*(530), 707–720.

Yao, Y., Vehtari, A., Simpson, D., & Gelman, A. (2018). Using stacking to average Bayesian predictive distributions (with discussion). *Bayesian Analysis*, *13*(3), 917–1007.

Zhang, B., & Cressie, N. (2020). Bayesian inference of spatio-temporal changes of Arctic sea ice. *Bayesian Analysis*, *15*(2), 605–631.

Zhang, H. (2002). On estimation and prediction for spatial generalized linear mixed models. *Biometrics*, *58*(1), 129–136.

Zhang, H., & El-Shaarawi, A. (2010). On spatial skew-Gaussian processes and applications. *Environmetrics*, *21*(1), 33–47.

Zheng, X., Kottas, A., & Sansó, B. (2023). Nearest-neighbor mixture models for non-Gaussian spatial processes. https://doi.org/10.48550/arXiv.2107.07736

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article. The data and code are available at https://github.com/xzheng42/nnmp-examples-env-2023.

How to cite this article: Zheng, X., Kottas, A., & Sansó, B. (2023). Bayesian geostatistical modeling for discrete-valued processes. *Environmetrics*, e2805. https://doi.org/10.1002/env.2805