Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/uasa20

# A Nonstationary Soft Partitioned Gaussian Process Model via Random Spanning Trees

Zhao Tang Luo, Huiyan Sang & Bani Mallick

**To cite this article:** Zhao Tang Luo, Huiyan Sang & Bani Mallick (23 Aug 2023): A Nonstationary Soft Partitioned Gaussian Process Model via Random Spanning Trees, Journal of the American Statistical Association, DOI: 10.1080/01621459.2023.2249642

**To link to this article:** https://doi.org/10.1080/01621459.2023.2249642

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2022, VOL. 00, NO. 0, 1–12: Theory and Methods
https://doi.org/10.1080/01621459.2023.2249642



# A Nonstationary Soft Partitioned Gaussian Process Model via Random Spanning Trees

Zhao Tang Luo, Huiyan Sang, and Bani Mallick

Department of Statistics, Texas A&M University, College Station, TX

**ABSTRACT**

There has been a long-standing challenge in developing locally stationary Gaussian process models concerning how to obtain flexible partitions and make predictions near boundaries. In this work, we develop a new class of locally stationary stochastic processes, where local partitions are modeled by a soft partition process via predictive random spanning trees that leads to highly flexible spatially contiguous subregion shapes. This valid nonstationary process model knits together local models such that both parameter estimation and prediction can be performed under a unified and coherent framework, and it captures both discontinuities/abrupt changes and local smoothness in a spatial random field. We propose a theoretical framework to study the Bayesian posterior concentration concerning the behavior of this Bayesian nonstationary process model. The performance of the proposed model is illustrated with simulation studies and real data analysis of precipitation rates over the contiguous United States. Supplementary materials for this article are available online.

**ARTICLE HISTORY**
Received May 2021
Accepted July 2023

**KEYWORDS**
Bayesian posterior concentration; Locally stationary models; Nonstationary Gaussian process; Random spanning trees

## 1. Introduction

Gaussian processes (GPs) have been a widely used modeling tool in spatial statistics, machine learning, and computer model uncertainty quantification. In the past decades, nonstationary GPs have attracted much attention due to their flexibility in modeling varying spatial dependence structures over the spatial domain. Despite significant progress on nonstationary spatial process models in the literature, developing flexible, computationally efficient, and theoretically justified nonstationary models remains an important but challenging open problem. Locally stationary processes for nonstationary spatial data have gained great popularity in the literature, due to their advantages in adapting to local and nonstationary data features and naturally allowing for reduced computations using local model results. Such models are also useful for detecting discontinuities in spatial fields that are commonly encountered in subsurface geology, public health, real estate, and social-demographic and economic studies, to name a few. However, there are several vital questions surrounding such methods: (i) How many subregions to use? (ii) How to identify locally stationary partitions (subregions)? (iii) How to achieve consensus predictions from local models, especially at boundaries?

Some existing work on locally stationary GPs relies on predetermined subregions. Park, Huang, and Ding (2011) used uniform grids for roughly evenly distributed data points. For unevenly distributed data points, k-d tree partitions with rectangular shapes (Shen, Ng, and Seeger 2006) and spatial hierarchical clustering algorithms (Heaton, Christensen, and Terres 2017) were used in the literature. Gerber and Nychka (2021) considered an overlapping domain partitioning method and used a parallel cross-validation algorithm to estimate local covariance parameters and perform spatial predictions. Model-based partitioning methods provide an alternative to these approaches. Risser et al. (2019) and Zhang and Williamson (2019) considered GP models based on Gaussian mixture clustering of spatial locations, where the number of clusters is pre-specified instead of being learned from data. Bolin, Wallin, and Lindgren (2019) developed a mixture of GP models for data on a uniform grid, where the clusters are modeled by a Markov random field and hence may not be spatially contiguous. The binary-treed GP models proposed in Gramacy and Lee (2008) partition the input space into nonoverlapping regions by making binary splits recursively, hence only producing rectangular-shaped clusters with boundaries always parallel to the input-space axes. Kim, Mallick, and Holmes (2005) assumed the partition is defined by a number of centering locations such that points within a cluster are closer to its center than any other cluster centers, which leads to convex-polygon-shaped clusters (a.k.a. Voronoi cells). The Voronoi tessellation-based method was extended in Pope et al. (2021) by allowing a subregion to be formed by multiple convex polygons but without guarantee of spatial contiguity of subregions, and in Gosoniu and Vounatsou (2011) by assuming a mixture of cell-specified models with distance-based weights. Despite the benefits of locally stationary models, a common criticism is that they lack a coherent global process for inference and prediction. Moreover, the constraints imposed on the shape of clusters in the current literature considerably limit the applications and interpretability of local stationary models in real problems, where it is of interest for practitioners to detect and locate spatial nonstationarities that may have highly irregular structures. Most recently, the spanning-treed partitioning model has been demonstrated as an efficient modeling tool for highly flexible spatially contiguous cluster shapes (Li and Sang 2019;

**CONTACT** Huiyan Sang *huiyan@stat.tamu.edu* Department of Statistics, Texas A&M University, College Station, TX.
Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JASA*.

Teixeira, Assunção, and Loschi 2019; Luo, Sang, and Mallick 2021b). Nonetheless, these works have been restricted to the partition of a finite set of observed locations in regression settings. Besides the aforementioned locally stationary GP models, a variety of nonstationary covariance functions of GPs have been proposed to model the heterogeneity of spatial dependence based on the ideas of kernel convolution, dimension expansion, spatial deformation, basis representations, and stochastic partial differential equations, to name a few. We refer interested readers to Risser (2016) and Fouedjio (2017) for a comprehensive review.

In light of these challenges and limitations, our contribution is to develop a new class of nonstationary GP models with flexible and easily interpretable dependence structures. The proposed nonstationary model is constructed from locally stationary stochastic processes on a partitioned domain. We propose a general framework to extend a spatially contiguous partition model on a finite set of reference knots to the whole spatial domain, by introducing a soft space partition process that uses neighborhood information. To address the key and challenging issue of learning space partitions with flexible shapes and sizes, we assign a spanning-treed partition prior on the finite reference set. Built upon the latent space partition, a valid global spatial process model called the soft partitioned GP (SPGP), is further defined to knit together local models, such that the predictive distributions admit Gaussian mixture forms that can lead to better performance in prediction and uncertainty quantification near partition boundaries. The idea of building spatial processes from finite-dimensional models has shown great promise in recent literature (see, e.g., Lindgren, Rue, and Lindström 2011; Datta et al. 2016). Our formulation adds to this line of work, but the motivation and model specifications are vastly different from the existing literature. The framework of constructing SPGP is general and can adopt any other space partition priors such as binary trees, Voronoi tessellations, and product partition models.

We also make a theoretical contribution to studying the Bayesian posterior concentration concerning the infill asymptotic behavior of this Bayesian nonstationary process model. To the best of our knowledge, Bayesian theoretical properties of locally partitioned GP models have not been investigated in the literature. We design an efficient Bayesian estimation and prediction algorithm that automatically learns local partitions and other model parameters from data to detect discontinuities (abrupt changes), understand spatial dependence structures, and perform spatial prediction by the Bayesian model averaging. Moreover, the modeling framework allows flexible choices of reference knots, which, if selected to be smaller sets, naturally deliver a speed-up computation algorithm. We offer several other computational strategies to take advantage of tree structures, linear algebra tricks, and recently developed fast algorithms for GP models with massive spatial data.

The rest of this article is organized as follows. Section 2 describes a general framework to construct the SPGP model and develops its theoretical properties. In Section 3, we discuss some computational strategies. We then demonstrate the model performance with synthetic data in Section 4 and with real precipitation data in Section 5. Finally, Section 6 concludes the article with some discussions. Technical proofs, details of posterior inference, and supplementary results on the synthetic and real data are provided in supplementary materials. Our code is publicly available at *https://github.com/ztluostat/SPGP*.

## 2. Soft Partitioned Gaussian Process Models

In many environmental applications, spatial data often exhibit a dependence structure that is not homogeneous in space. Oftentimes data within a subregion are relatively homogeneous while there could be substantial differences across subregions. To introduce a valid global process model for spatial estimation and prediction while characterizing spatially heterogeneous dependence, we begin by introducing a soft partition process that probabilistically models the cluster membership of any given location in Section 2.2. The soft partition process over the whole study domain $\mathcal{D} \subseteq \mathbb{R}^d$ is extended from a so-called predictive spanning tree partition prior model over a fixed finite set of locations in $\mathcal{D}$. Conditional on this soft partition process, we then define a valid soft partitioned Gaussian process in Section 2.3.

### 2.1. Notations

Let $\mathcal{S} = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\} \subseteq \mathcal{D}$ denote a set of observed locations, $\mathcal{S}^* = \{\mathbf{s}_1^*, \ldots, \mathbf{s}_m^*\} \subseteq \mathcal{D}$ denote a set of pre-specified distinct *reference knots* that may or may not coincide with $\mathcal{S}$, and $\mathcal{V}$ denote *any* arbitrary finite set in $\mathcal{D}$. In this article, we assume both $\mathcal{S}$ and $\mathcal{S}^*$ are fixed sets. We use $\pi_k(\mathcal{V}) = \{\mathcal{V}_1, \ldots, \mathcal{V}_k\}$ to denote a partition of $\mathcal{V}$ into $k$ disjoint subsets (also called clusters). Given $\pi_k(\mathcal{V})$, let $\mathbf{z}(\mathcal{V}) = \{z(\mathbf{s})\}_{\mathbf{s}\in\mathcal{V}}$ be the vector of cluster memberships of locations in $\mathcal{V}$ such that $z(\mathbf{s}) = j \in \{1, \ldots, k\}$ with probability one if $\mathbf{s} \in \mathcal{V}_j$. Finally, let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ denote an arbitrary undirected spatial graph with vertices $\mathcal{V}$ and edges $\mathcal{E}$, where two nearby locations (with respect to the Euclidean distance metric $d$) are connected by an edge.

In spatial settings, it is desired to impose spatial contiguity constraints on partitions such that each cluster can be interpreted as a spatially connected subregion. We say $\pi_k(\mathcal{V})$ is a *spatially contiguous partition* (or contiguous partition for short) with respect to $\mathcal{G}$ if each induced subgraph of $\mathcal{G}$ on the vertex subset $\mathcal{V}_j$ is *connected*.

### 2.2. A Spatial Soft Partition Process

The backbone of our soft partition process (SPP) is a finite partition model $\pi_k(\mathcal{S})$ on the observed locations. While it is natural to restrict $\pi_k(\mathcal{S})$ to be spatially contiguous with respect to a spatial graph on $\mathcal{S}$, it is computationally challenging to build a probabilistic contiguous partition model directly on $\mathcal{S}$ when the number of observations is large, due to the large number of vertices and edges in the spatial graph for observations. Instead, we construct a contiguous partition model on the reduced reference knot set $\mathcal{S}^*$, and then define $\pi_k(\mathcal{S})$ on the observed locations via a predictive soft assignment equation.

In a nutshell, we will construct the SPP model in three steps to be described in the following three sections. In the first step, we will introduce a random spanning tree (RST) partition prior for $\pi_k(\mathcal{S}^*)$ on the reference knots. In the second step, we

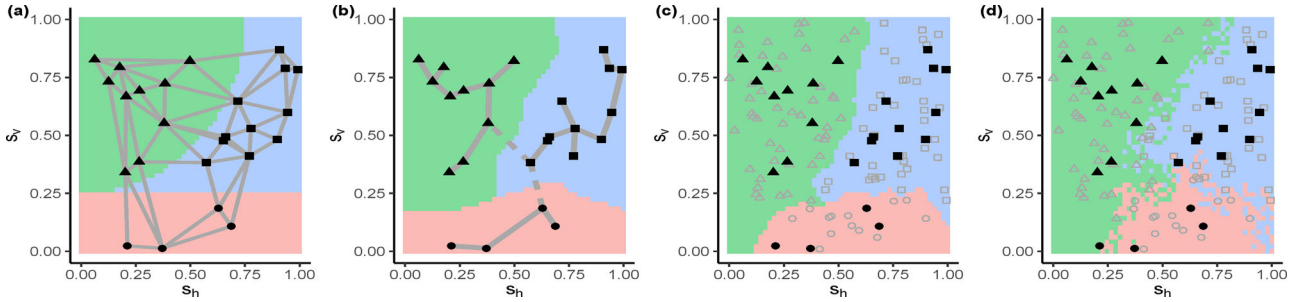


**Figure 1.** (a) A Delaunay triangulation graph on $\mathcal{S}^*$ and the true space partition. (b) $\pi_3(\mathcal{S}^*)$ obtained by removing the two dashed edges from a spanning tree graph on $\mathcal{S}^*$, and a realization of 1-SPP with $\mathcal{S} = S^*$ given $\pi_3(\mathcal{S}^*)$. (c) A realization of 1-SPP with $\mathcal{S} \neq S^*$ given the same $\pi_3(\mathcal{S}^*)$ as in (b). Locations in $\mathcal{S}$ are marked by gray color. (d) A realization of 3-SPP given the same $\mathcal{S} \neq S^*$ and $\pi_3(\mathcal{S}^*)$ as in (c).

define a predictive soft assignment equation to model $\pi_k(\mathcal{S})$ given $\pi_k(\mathcal{S}^*)$. Finally, we extend the finite partition model on $(\mathcal{S}^*, \mathcal{S})$ to a partition process on $\mathcal{D}$ again through a predictive soft assignment equation given $\pi_k(\mathcal{S}^*)$ and $\pi_k(\mathcal{S})$. Despite we focus on an RST prior for $\pi_k(\mathcal{S}^*)$ in this article, the proposed framework for constructing SPPs is generic in the sense that SPPs can be built upon any prior model for $\pi_k(\mathcal{S}^*)$.

### 2.2.1. Random Spanning Tree Prior for $\pi_k(\mathcal{S}^*)$

We first describe the prior model for spatially contiguous $\pi_k(\mathcal{S}^*)$ (or equivalently, $\mathbf{z}(\mathcal{S}^*)$). Motivated by the success of spanning tree models for capturing contiguous partitions in linear regression settings (see, e.g., Luo, Sang, and Mallick 2021b), we assign an RST partition prior for $\pi_k(\mathcal{S}^*)$. This prior simplifies a graph partition problem into a compact representation based on spanning tree cuts, without sacrificing the richness of its support (see also Proposition 1).

Let $\mathcal{G} = (\mathcal{S}^*, \mathcal{E})$ be a spatial graph with vertex set $\mathcal{S}^*$ and edge set $\mathcal{E}$. Guided by our theoretical results (see Assumption SD in Section 2.5), $\mathcal{G}$ can be specified as a radius-based nearest neighbor (R-NN) graph that connects a node with its neighboring nodes within a certain distance, or a Delaunay triangulation graph (Chew 1987) with edges longer than a large threshold removed to avoid those artifact edges connecting two remote points near domain boundaries. Figure 1(a) shows an example of Delaunay triangulation graphs.

A spanning tree of $\mathcal{G}$ is defined as a subgraph $\mathcal{T} = (\mathcal{S}^*, \mathcal{E}_\mathcal{T}), \mathcal{E}_\mathcal{T} \subseteq \mathcal{E}$ that connects all vertices without any cycle. Let $\omega_{i,i'}$ be the weight of the edge $(\mathbf{s}_i^*, \mathbf{s}_{i'}^*)$ in $\mathcal{G}$ and $\boldsymbol{\omega} = \{\omega_{i,i'} : (\mathbf{s}_i^*, \mathbf{s}_{i'}^*) \in \mathcal{E}\}$. A minimum spanning tree (MST) is a spanning tree that has the minimal sum of edge weights $\sum_{(\mathbf{s}_i^*, \mathbf{s}_{i'}^*) \in \mathcal{E}_\mathcal{T}} \omega_{i,i'}$. A well-known property of spanning trees is that after a set of $k-1$ edges is removed from $\mathcal{T}$, we obtain a graph with $k$ connected components. By treating the $j$th connected component as cluster $\mathcal{S}_j^*$, we obtain a contiguous partition $\pi_k(\mathcal{S}^*)$. We say $\pi_k(\mathcal{S}^*)$ is *induced* by $\mathcal{T}$ in this case. See Figure 1(b) for an example of $\pi_3(\mathcal{S}^*)$ induced from a spanning tree. The estimation of $\pi_k(\mathcal{S}^*)$ amounts to learning the spanning tree (which may not be unique) and its removed edges that induce the true partition. A prior on $\pi_k(\mathcal{S}^*)$ can therefore be assigned hierarchically, by first placing priors on the number of clusters $k$ and the spanning trees in $\mathcal{G}$, and then the positions of the $k-1$ removed edges.

Formally, conditional on $\mathcal{T}$ and $k$ we assume a uniform prior on all possible partitions induced by $\mathcal{T}$ (or equivalently, a uniform prior on which $k-1$ edges in $\mathcal{T}$ are removed):

$$p\left\{\pi_k(\mathcal{S}^*) \mid k, \mathcal{T}\right\} \propto \mathbb{1}\left\{\pi_k(\mathcal{S}^*) \text{ is induced by } \mathcal{T} \text{ and has } k \text{ clusters}\right\}, \quad (1)$$

where $\mathbb{1}(\cdot)$ is an indicator function. Regarding the prior on $\mathcal{T}$, a seemingly natural choice is to assume a discrete uniform distribution on all possible spanning trees of $\mathcal{G}$. However, it is challenging to sample from this uniform distribution. We opt to place an iid uniform prior on edge weights $\boldsymbol{\omega}$ instead, which induces a prior model on the spanning tree space via

$$\mathcal{T} = \text{MST}(\boldsymbol{\omega}), \quad \omega_{i,i'} \overset{\text{iid}}{\sim} \text{Unif}(0,1), \quad (2)$$

where $\text{MST}(\boldsymbol{\omega})$ means an MST of the graph $\mathcal{G}$ based on edge weights $\boldsymbol{\omega}$. This MST space constructed from random edge weights consists of all possible spanning trees of $\mathcal{G}$. We will show in Section 3 that this prior also leads to an exact and fast spanning tree sampler, taking advantage of Prim's algorithm for MST constructions.

Finally, we assume a truncated geometric distribution on $k$ such that

$$\mathbb{P}(k = j) \propto (1-c)^j, \quad \text{for } j = 1, \ldots, \bar{k}_m, \ 0 \leq c < 1, \quad (3)$$

where $\bar{k}_m$ is the pre-specified maximum number of clusters and $c$ is a hyperparameter controlling the decaying rate of the prior probability so that models with a large number of clusters can be penalized. This allows the number of clusters to be learned from the data. When $\mathcal{S} = \mathcal{S}^*$, guided by our theoretical results in Section 2.5, we recommend specifying $\bar{k}_m$ such that it scales with $\sqrt{n \log n}$ (see Assumption P1).

The next proposition states that the support of the RST partition prior contains any spatially contiguous partitions on $\mathcal{S}^*$ (and hence $\mathcal{S}$ when $\mathcal{S} = \mathcal{S}^*$) with no more than $\bar{k}_m$ clusters.

*Proposition 1.* Let $\pi_k(\mathcal{S}^*) = \{\mathcal{S}_1^*, \ldots, \mathcal{S}_k^*\}$ be an arbitrary spatially contiguous partition. Then $\pi_k(\mathcal{S}^*)$ is within the support of the prior defined by (1), (2), and (3) if $k \leq \bar{k}_m$.

### 2.2.2. A Soft Partition Model for $\pi_k(\mathcal{S})$

We now introduce the so-called *predictive RST partition prior* for $\pi_k(\mathcal{S})$, which is constructed from the RST partition prior, denoted as $\text{RST}(\mathcal{G})$, on a reduced reference knots set $\mathcal{S}^*$ as described in Section 2.2.1. More precisely, conditional on the

cluster memberships $\mathbf{z}(\mathcal{S}^*)$ associated with $\pi_k(\mathcal{S}^*)$, we model each cluster membership $z(\mathbf{s})$, $\mathbf{s} \in \mathcal{S}$ independently as follows:

$$z\big(\mathbf{s}|\mathbf{z}(\mathcal{S}^*)\big) = \boldsymbol{\Psi}(\mathbf{s}, \mathcal{S}^*)\mathbf{z}(\mathcal{S}^*), \quad \mathbf{s} \in \mathcal{S} \tag{4a}$$

$$\boldsymbol{\Psi}(\mathbf{s}, \mathcal{S}^*) \sim \text{Multinomial}(1, \boldsymbol{\alpha}(\mathbf{s}, \mathcal{S}^*)) \tag{4b}$$

$$\mathbf{z}(\mathcal{S}^*) \sim \text{RST}(\mathcal{G}) \tag{4c}$$

where $\boldsymbol{\alpha}(\mathbf{s}, \mathcal{S}^*) = \big(\alpha(\mathbf{s}, \mathbf{s}_1^*), \ldots, \alpha(\mathbf{s}, \mathbf{s}_m^*)\big) \in \mathbb{R}^{1\times m}$ is a row vector of membership assignment probabilities that sum to 1, and $\Psi(\mathbf{s}, \mathcal{S}^*) = \big(\psi(\mathbf{s}, \mathbf{s}_1^*), \ldots, \psi(\mathbf{s}, \mathbf{s}_m^*)\big) \in \mathbb{R}^{1\times m}$ is a random binary row vector with only one entry being 1 that follows a multinomial distribution with event probabilities $\boldsymbol{\alpha}(\mathbf{s}, \mathcal{S}^*)$. The model in (4) defines a predictive soft assignment scheme at the observed locations which assumes $\mathbf{s}$ shares the same cluster membership as $\mathbf{s}_\ell^*$ with probability $\alpha(\mathbf{s}, \mathbf{s}_\ell^*)$. For spatial problems, it is reasonable to assume a location is more likely to share the same cluster membership as one of its nearby neighbors. We thus, assume that when $\mathbf{s} \notin \mathcal{S}^*$, $\boldsymbol{\alpha}(\mathbf{s}, \mathcal{S}^*)$ has nonzero probabilities only at the $L$ nearest reference knots, and for each of these $L$ neighbors $\alpha(\mathbf{s}, \mathbf{s}_\ell^*) = 1/L$ (uniform weighting) or $\alpha(\mathbf{s}, \mathbf{s}_\ell^*) \propto 1/d(\mathbf{s}, \mathbf{s}_\ell^*)$ (inverse distance weighting). When $\mathbf{s} \in \mathcal{S}^*$, $d(\mathbf{s}, \mathbf{s}_\ell^*) = 0$ for some $\ell$, and $\boldsymbol{\alpha}(\mathbf{s}, \mathcal{S}^*)$ only has one nonzero entry with value 1 at the $\ell$th column. Alternatively, one may use other probability weighting functions such as the hat basis function commonly adopted in the finite element type of methods such as INLA (Lindgren, Rue, and Lindström 2011).

#### 2.2.3. Predictive Soft Partition Process

Conditional on the finite partition prior model for $\mathcal{S}$ and $\mathcal{S}^*$ in (4), we now turn attention to the construction of the SPP model for any arbitrary finite set in $\mathcal{D}$. Let $\mathcal{U}$ be any finite subset of $\mathcal{D}$ such that $\mathcal{U} \cap (\mathcal{S} \cup \mathcal{S}^*) = \emptyset$. We follow a similar predictive soft assignment scheme as in (4) and assume that

$$z\big(\mathbf{u}|\mathbf{z}(\mathcal{S}), \mathbf{z}(\mathcal{S}^*)\big) = \boldsymbol{\Psi}(\mathbf{u}, \mathcal{S})\mathbf{z}(\mathcal{S}), \quad \mathbf{u} \in \mathcal{U} \tag{5a}$$

$$\boldsymbol{\Psi}(\mathbf{u}, \mathcal{S}) \sim \text{Multinomial}(1, \boldsymbol{\alpha}(\mathbf{u}, \mathcal{S})) \tag{5b}$$

where $\Psi(\mathbf{u}, \mathcal{S})$ and $\boldsymbol{\alpha}(\mathbf{u}, \mathcal{S})$ are defined similarly as in (4) to reflect the partitioning uncertainty at new locations. When $\psi(\mathbf{u}, \mathbf{s}_i) = 1$ for some $i$, a new location $\mathbf{u} \in \mathcal{U}$ is assigned to the same cluster as the $i$th observation.

It is easy to see that the constructions in (4) and (5) define a stochastic process $\{z(\mathbf{v}) : \mathbf{v} \in \mathcal{D}\}$ given $\pi_k(\mathcal{S}^*)$ and $\pi_k(\mathcal{S})$ that takes value in $\{1, \ldots, k\}$, such that the joint distribution for any finite set $\mathcal{V} \subseteq \mathcal{D}$ satisfies $p\big(\mathbf{z}(\mathcal{V})\big) = \prod_{\mathbf{v}\in\mathcal{V}} p\big(z(\mathbf{v})\big)$, where $p\big(z(\mathbf{v})\big)$ is a degenerated distribution on $j$ if $\mathbf{v} \in \mathcal{S}_j^* \cup \mathcal{S}_j$, or a categorical distribution with event probabilities depending on $\boldsymbol{\alpha}(\mathbf{v}, \mathcal{S})$ and $\pi_k(\mathcal{S})$ if $\mathbf{v} \notin \mathcal{S}^* \cup \mathcal{S}$. We refer to this process as an $L$ nearest neighbor *soft partition process* ($L$-SPP) conditional on $\pi_k(\mathcal{S})$ and $\pi_k(\mathcal{S}^*)$. We treat $L$ as a hyperparameter, and its selection will be discussed in Section 3.1.

Figure 1 shows three examples of realized $\pi_k(\mathcal{D})$ from the $L$-SPP model, when (i) $\mathcal{S} = \mathcal{S}^*$ and $L = 1$, (ii) $\mathcal{S}^* \neq \mathcal{S}$ with $m < n$ and $L = 1$, and (iii) $\mathcal{S}^* \neq \mathcal{S}$ and $L = 3$. Let $V_{\mathcal{S}^*}(\mathbf{s}^*) = \{\mathbf{v} \in \mathcal{D} : d(\mathbf{v}, \mathbf{s}^*) < d(\mathbf{v}, \mathbf{s}^{*\prime}) \text{ for any } \mathbf{s}^{*\prime} \in \mathcal{S}^* \text{ and } \mathbf{s}^{*\prime} \neq \mathbf{s}^*\}$ be the Voronoi cell with nucleus $\mathbf{s}^*$ in the Voronoi tessellation of $\mathcal{D}$ based on $\mathcal{S}^*$. When $L = 1$ and $\mathcal{S}^* = \mathcal{S}$, $L$-SPP reduces to a hard space partition model for $\pi_k(\mathcal{D})$, whose $j$th subregion $\mathcal{D}_j$ becomes the unioned Voronoi cells, $\cup_{\mathbf{s}^*\in\mathcal{S}_j^*} V_{\mathcal{S}^*}(\mathbf{s}^*)$. When $\mathcal{S} \neq \mathcal{S}^*$, $\pi_k(\mathcal{D})$ is jointly determined by both $\pi_k(\mathcal{S}^*)$ and $\pi_k(\mathcal{S})$. Note that when $L = 3$, the partition boundary in $\pi_k(\mathcal{D})$ is soft, reflecting the uncertainty near the boundary. When denser $\mathcal{S}^*$ and $\mathcal{S}$ are used, the determined $\pi_k(\mathcal{D})$ better approximates the true partition.

### 2.3. A Soft Partitioned Gaussian Process

Given a realization of $\pi_k(\mathcal{S})$ (conditional on $\pi_k(\mathcal{S}^*)$), we allow each $\mathbf{w}(\mathcal{S}_j) = \{w(\mathbf{s}) : \mathbf{s} \in \mathcal{S}_j\}$ to be a realization of different zero-mean Gaussian processes characterized by a covariance function $C(\cdot, \cdot|\boldsymbol{\theta}_j)$, that is,

$$\mathbf{w}(\mathcal{S}_j)|\pi_k(\mathcal{S}), \pi_k(\mathcal{S}^*) \sim \text{N}_{n_j}\big\{\mathbf{0},\ \mathbf{C}(\mathcal{S}_j, \mathcal{S}_j|\boldsymbol{\theta}_j)\big\} \tag{6}$$

independently for all $j = 1, \ldots, k$, where $n_j = |\mathcal{S}_j|$ is the number of observed locations in cluster $\mathcal{S}_j$. The joint distribution of $\mathbf{w}(\mathcal{S}) = \{\mathbf{w}(\mathcal{S}_1), \ldots, \mathbf{w}(\mathcal{S}_k)\}$ conditional on $\pi_k(\mathcal{S})$ is therefore Gaussian with a block-diagonal covariance matrix whose $j$th block is $\mathbf{C}(\mathcal{S}_j, \mathcal{S}_j|\boldsymbol{\theta}_j)$.

To extend (6) to a legitimate spatial process on $\mathcal{D}$, we first define the distribution of $\mathbf{w}(\mathcal{U})$ given $\mathbf{w}(\mathcal{S})$ for any finite set $\mathcal{U}$ that is disjoint from $\mathcal{S}$. Given a realization of the $L$-SPP $\mathbf{z}(\mathcal{U}) = \big(z(\mathbf{u}_1), \ldots, z(\mathbf{u}_r)\big)$ conditional on $\pi_k(\mathcal{S})$ and $\pi_k(\mathcal{S}^*)$, $\mathcal{U}$ can be partitioned into clusters $\mathcal{U}_j = \{\mathbf{u} \in \mathcal{U} : z(\mathbf{u}) = j\}$. The conditional distribution of $\mathbf{w}(\mathcal{U})$ given $\mathbf{w}(\mathcal{S}_j)$, $\mathbf{z}(\mathcal{U}_j)$, $\pi_k(\mathcal{S})$, and $\pi_k(\mathcal{S}^*)$ is assumed to be

$$\begin{aligned}&\mathbf{w}(\mathcal{U}_j)|\mathbf{w}(\mathcal{S}_j), \mathbf{z}(\mathcal{U}_j), \pi_k(\mathcal{S}), \pi_k(\mathcal{S}^*)\\ &\quad\sim \text{N}_{r_j}\big\{\boldsymbol{\mu}(\mathcal{U}_j|\mathcal{S}_j, \boldsymbol{\theta}_j),\ \mathcal{S}(\mathcal{U}_j|\mathcal{S}_j, \boldsymbol{\theta}_j)\big\},\end{aligned} \tag{7}$$

independently for $j = 1, \ldots, k$, where $r_j = |\mathcal{U}_j|$, and

$$\boldsymbol{\mu}(\mathcal{U}_j|\mathcal{S}_j, \boldsymbol{\theta}_j) = \mathbf{C}(\mathcal{U}_j, \mathcal{S}_j|\boldsymbol{\theta}_j)\mathbf{C}^{-1}(\mathcal{S}_j, \mathcal{S}_j|\boldsymbol{\theta}_j)\mathbf{w}(\mathcal{S}_j), \tag{8}$$

$$\begin{aligned}\mathcal{S}(\mathcal{U}_j|\mathcal{S}_j, \boldsymbol{\theta}_j) &= \mathbf{C}(\mathcal{U}_j, \mathcal{U}_j|\boldsymbol{\theta}_j)\\ &\quad - \mathbf{C}(\mathcal{U}_j, \mathcal{S}_j|\boldsymbol{\theta}_j)\mathbf{C}^{-1}(\mathcal{S}_j, \mathcal{S}_j|\boldsymbol{\theta}_j)\mathbf{C}(\mathcal{S}_j, \mathcal{U}_j|\boldsymbol{\theta}_j).\end{aligned} \tag{9}$$

Combining (6) and (7), for *any* finite subset $\mathcal{V}$ of $\mathcal{D}$ with the associated cluster memberships $\mathbf{z}(\mathcal{V})$, the density of $\mathbf{w}(\mathcal{V})$ given $\mathbf{z}(\mathcal{V})$, $\pi_k(\mathcal{S})$ and $\pi_k(\mathcal{S}^*)$ is given by

$$\begin{aligned}p\big(\mathbf{w}(\mathcal{V})|\mathbf{z}(\mathcal{V})\big) = \int p\big(\mathbf{w}(\mathcal{U})|\mathbf{w}(\mathcal{S}), \mathbf{z}(\mathcal{U})\big)\, p(\mathbf{w}(\mathcal{S}))\\ \prod_{\{\mathbf{s}\in\mathcal{S}\setminus\mathcal{V}\}} \text{d}(\mathbf{w}(\mathbf{s}))\quad \text{where } \mathcal{U} = \mathcal{V}\setminus\mathcal{S}.\end{aligned} \tag{10}$$

The dependence on $\pi_k(\mathcal{S})$, $\pi_k(\mathcal{S}^*)$ and parameters $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k)$ is made implicit in (10) for conciseness. Note that if $\mathcal{V} \subseteq \mathcal{S}$ then $p(\mathbf{w}(\mathcal{U})|\mathbf{w}(\mathcal{S}), \mathbf{z}(\mathcal{U})) = 1$ and if $\mathcal{S} \setminus \mathcal{V} = \emptyset$ then the integration in (10) is not needed. A mean-zero GP on $\mathcal{D}$ is therefore defined by (10) conditional on an $L$-SPP on $\mathcal{D}$, with a covariance function

$$\begin{aligned}&C^{\ddagger}(\mathbf{v}, \mathbf{v}'|z(\mathbf{v}), z(\mathbf{v}'), \boldsymbol{\Theta})\\ &\quad= \begin{cases} C\big(\mathbf{v}, \mathbf{v}'|\boldsymbol{\theta}_j\big), & \text{if } \mathbf{v}, \mathbf{v}' \in \mathcal{D}_j \text{ for some } j \in \{1, \ldots, k\},\\ 0, & \text{otherwise,}\end{cases}\end{aligned}$$

where $\mathcal{D}_j = \{\mathbf{s} \in \mathcal{D} : z(\mathbf{s}) = j\}$ is the collection of all locations in $\mathcal{D}$ that are assigned to the $j$th cluster.

Marginalizing out the $L$-SPP realization on $\mathcal{V} \setminus (\mathcal{S} \cup \mathcal{S}^*)$ conditional on $\pi_k(\mathcal{S})$ and $\pi_k(\mathcal{S}^*)$, the density of $p(\mathbf{w}(\mathcal{V}))$ for *any* finite subset $\mathcal{V} \subseteq \mathcal{D}$ is therefore given by a Gaussian mixture

$$p\big(\mathbf{w}(\mathcal{V})|\pi_k(\mathcal{S}),\pi_k(\mathcal{S}^*)\big) = \sum_{\mathbf{z}(\mathcal{V}\setminus(\mathcal{S}\cup\mathcal{S}^*))} p\big(\mathbf{w}(\mathcal{V})|\mathbf{z}(\mathcal{V})\big)p\big(\mathbf{z}(\mathcal{V})|\pi_k(\mathcal{S}),\pi_k(\mathcal{S}^*)\big), \quad (11)$$

where the summation is over all possible combinations of cluster memberships $\mathbf{z}(\mathcal{V} \setminus (\mathcal{S} \cup \mathcal{S}^*))$. As shown in Supplementary Section S1, the density (11) satisfies Kolmogorov's consistency criteria and thus implies a valid spatial process on $\mathcal{D}$ conditional on $\pi_k(\mathcal{S})$ and $\pi_k(\mathcal{S}^*)$, which we call an $L$ nearest neighbor *soft partitioned Gaussian process* ($L$-SPGP) conditional on $\pi_k(\mathcal{S})$ and $\pi_k(\mathcal{S}^*)$. The covariance function of this process is given by

$$C^{\ddagger}(\mathbf{v},\mathbf{v}'|\pi_k(\mathcal{S}),\pi_k(\mathcal{S}^*),\boldsymbol{\Theta}) = \sum_{j=1}^{k} \kappa_j^{\mathbf{v},\mathbf{v}'} C(\mathbf{v},\mathbf{v}'|\boldsymbol{\theta}_j), \quad (12)$$

where the weights $\kappa_j^{\mathbf{v},\mathbf{v}'} = \kappa_j^{\mathbf{v}}\kappa_j^{\mathbf{v}'}$ and $\kappa_j^{\mathbf{v}}$ is the conditional probability that $\mathbf{v}$ belongs to the $j$th cluster given $\pi_k(\mathcal{S})$ and $\pi_k(\mathcal{S}^*)$. Note that for any location $\mathbf{v} \in \mathcal{D}$, $\kappa_j^{\mathbf{v}} = \sum_{\ell=1}^{L} \alpha_j \mathbb{1}(N_{\mathbf{v},\ell} \in \mathcal{S}_j)$ if $\mathbf{v} \notin \mathcal{S} \cup \mathcal{S}^*$, and $\kappa_j^{\mathbf{v}} = 1$ if $\mathbf{v} \in \mathcal{S}_j \cup \mathcal{S}_j^*$, where $N_{\mathbf{v},\ell}$ is the $\ell$ nearest neighbor of $\mathbf{v}$ in $\mathcal{S}$. Therefore, the conditional covariance function is completely determined by the neighborhood structure and the base covariance function $C$. In particular, $C^{\ddagger}(\mathbf{v},\mathbf{v}'|\pi_k(\mathcal{S}),\pi_k(\mathcal{S}^*),\boldsymbol{\Theta})$ reduces to $C(\mathbf{v},\mathbf{v}'|\boldsymbol{\theta}_j)$ if $\mathbf{v},\mathbf{v}' \in \mathcal{I}_j$, where $\mathcal{I}_j$ is the interior space of the $j$th cluster defined as $\mathcal{I}_j := \mathcal{S}_j^* \cup \mathcal{S}_j \cup \{\mathbf{u} \in \mathcal{D}\setminus(\mathcal{S}\cup\mathcal{S}^*) : N_{\mathbf{u},\ell} \in \mathcal{S}_j \text{ for all } \ell = 1,\dots,L\}$. If $C$ is taken to be a stationary and mean square continuous covariance function, then $L$-SPGP is locally stationary and mean square continuous within $\mathcal{I}_j$ for any choice of $L \geq 1$. When $L > 1$ and the inverse distance weighting is used for $\boldsymbol{\alpha}$, the covariance function of SPGP is continuous everywhere following (12), and hence SPGP is also mean square continuous (Adler and Taylor 2007) everywhere given $\pi_k(\mathcal{S}^*)$.

Note that this process can also be viewed as a finite *mixture* of GPs defined on $\mathcal{D}$, where each mixture component is $\mathrm{GP}(0, C(\cdot,\cdot|\boldsymbol{\theta}_j))$ and the spatially varying mixture weights are determined by the soft partition process model.

For further illustration of $L$-SPGP, let us consider two examples.

*Example 1.* Let $\mathcal{S}$ be the locations where the realization of the process $\{w(\mathbf{v})\}$ is observed and $\mathbf{u} \notin \mathcal{S}$ be a location on which we want to do prediction. The conditional (also called predictive or kriging) distribution is given by a Gaussian mixture

$$w(\mathbf{u})|\mathbf{w}(\mathcal{S}),\pi_k(\mathcal{S}),\pi_k(\mathcal{S}^*) \sim \sum_{\ell=1}^{L} \alpha_\ell \,\mathrm{N}_1\left(\boldsymbol{\mu}(\mathbf{u}|\mathcal{S}_{j(\ell)},\boldsymbol{\theta}_{j(\ell)}), \mathcal{S}(\mathbf{u}|\mathcal{S}_{j(\ell)},\boldsymbol{\theta}_{j(\ell)})\right), \quad (13)$$

where $j(\ell) = z(N_{\mathbf{u},\ell})$. The uncertainty of cluster memberships of $\mathbf{u}$ for prediction is captured by the Gaussian mixture structure. We refer to the mean and variance of the Gaussian mixture in (13) as the kriging mean and kriging variance at $\mathbf{u}$, respectively. Note that each mixture component in (13) may not be distinct; the $\ell$th and $\ell'$th components are identical whenever $j(\ell) = j(\ell')$. When $j(1) = \cdots = j(L) = J$ (i.e., when $\mathbf{u} \in \mathcal{I}_J$), (13) reduces to the same predictive distribution given by (7) using only the observations in $\mathcal{S}_j$ and the local covariance function with $\boldsymbol{\theta}_j$.

In general, the number of neighbors $L$ controls the smoothness of the kriging mean at $\mathbf{u}$ near the boundary set $\mathcal{D} \setminus \cup_{j=1}^{k} \mathcal{I}_j$. As $L$ increases, we have a larger boundary set, allowing for capturing partitioning uncertainty in a larger area, and the smoothing effects are stronger within the boundary set. See Figure S2 in Supplementary Section S3.1 for an illustration of the kriging means and standard deviations (SDs) across $\mathcal{D} = [0,1]^2$ with various values of $L$ and inverse distance weighted soft assignment probabilities.

*Example 2.* When $L = 1$, SPGP becomes a piecewise GP in the sense that it takes the form $\mathrm{GP}(0, C(\cdot,\cdot|\boldsymbol{\theta}_j))$ in $\mathcal{D}_j = \cup_{\mathbf{s}^* \in \mathcal{S}_j^*} V_{\mathcal{S}^*}(\mathbf{s}^*)$, the unioned Voronoi cells corresponding to $\mathcal{S}_j^*$. Our method allows Voronoi cells to be merged together to form space partitions with highly flexible shapes. In contrast, the piecewise GP model in Kim, Mallick, and Holmes (2005) treats each Voronoi cell as a cluster that can only have convex polygon shapes.

### 2.4. Bayesian Soft Partitioned Gaussian Process Regressions

We embed the proposed $L$-SPGP into a spatial regression setting. Consider a point-referenced response variable $y(\mathbf{s}) \in \mathbb{R}$ at a generic location $\mathbf{s} \in \mathcal{D}$ along with a vector of covariates $\mathbf{x}(\mathbf{s}) \in \mathbb{R}^p$. We denote the collection of responses and the design matrix corresponding to a generic finite subset $\mathcal{V}$ of $\mathcal{D}$ by $\mathbf{y}(\mathcal{V})$ and $\mathbf{X}(\mathcal{V})$, respectively.

We consider a spatial regression model specified as

$$y(\mathbf{s}) = \boldsymbol{\beta}^{\mathsf{T}}\mathbf{x}(\mathbf{s}) + w(\mathbf{s}), \quad \mathbf{s} \in \mathcal{D},$$

where the residual process $w(\mathbf{s})$ is modeled as a zero-mean $L$-SPGP conditional on $\pi_k(\mathcal{S})$ and $\pi_k(\mathcal{S}^*)$ as discussed in Section 2.3. Finally, we assign a random spanning tree prior to $\pi_k(\mathcal{S}^*)$ following Section 2.2.1, and the soft partition prior model to $\pi_k(\mathcal{S})$ following (4). The hierarchical model for observations can be written as

$$\mathbf{y}(\mathcal{S}_j)|\boldsymbol{\beta},\boldsymbol{\Theta},\pi_k(\mathcal{S}),\pi_k(\mathcal{S}^*) \overset{\text{ind.}}{\sim} \mathrm{N}_{n_j}\left\{\mathbf{X}(\mathcal{S}_j)\boldsymbol{\beta},\ \mathbf{C}(\mathcal{S}_j,\mathcal{S}_j|\boldsymbol{\theta}_j)\right\}, \quad (14a)$$

$$\boldsymbol{\beta}|\lambda \sim \mathrm{N}_p\left(\boldsymbol{\mu}_\beta, \lambda\mathbf{I}_p\right), \quad \lambda \sim \mathrm{IG}(a_\lambda, b_\lambda), \quad (14b)$$

$$z\big(\mathbf{s}|\pi_k(\mathcal{S}^*)\big) = \boldsymbol{\Psi}(\mathbf{s},\mathcal{S}^*)\mathbf{z}(\mathcal{S}^*),\ \boldsymbol{\Psi}(\mathbf{s},\mathcal{S}^*) \overset{\text{ind.}}{\sim} \mathrm{Multinomial}(1,\boldsymbol{\alpha}(\mathbf{s},\mathcal{S}^*)),\ \text{for } \mathbf{s} \in \mathcal{S}, \quad (14c)$$

$$\big(\pi_k(\mathcal{S}^*),k,\mathcal{T}\big) \sim p\big(\pi_k(\mathcal{S}^*)|k,\mathcal{T}\big)p(k)p(\mathcal{T}), \quad (14d)$$

where $p(\pi_k|k,\mathcal{T})$, $p(\mathcal{T})$, and $p(k)$ are specified in (1), (2), and (3), respectively. Note that we assume all clusters share the same coefficients. One may instead assume cluster-specified coefficients; however, we argue that this may cause identifiability issues between spatially varying regression means and spatial random effects and hence a poor parameter estimation performance, though we may still obtain reasonable prediction accuracy of the responses.

We complete the hierarchical model by specifying the local covariance function. One popular choice is the *stationary*

Matérn family (Banerjee et al. 2004) including both isotropic models $C(\mathbf{s},\mathbf{s}'|\boldsymbol{\theta}) = \sigma^2\rho(\mathbf{s},\mathbf{s}'|\phi,\nu) + \tau^2\mathbb{1}(\mathbf{s}=\mathbf{s}')$, where $\sigma^2$, $\phi$, $\nu$ and $\tau^2$ are the variance, range, smoothness and nugget effect variance parameters, respectively, and geometric *anisotropic* models. Priors for local covariance parameters are assigned following standard GP models.

Finally, consider a new location $\mathbf{u} \notin \mathcal{S}$ where we intend to predict the response $y(\mathbf{u})$ given $\mathbf{x}(\mathbf{u})$ and $\mathbf{y}(\mathcal{S})$. Following (13), the posterior predictive distribution of $y(\mathbf{u})$ is

$$y(\mathbf{u})|\mathbf{y}(\mathcal{S}),\boldsymbol{\beta},\boldsymbol{\Theta},\pi_k(\mathcal{S}),\pi_k(\mathcal{S}^*) \sim \sum_{\ell=1}^{L}\alpha_\ell \,\mathrm{N}_1\left(\widetilde{\boldsymbol{\mu}}(\mathbf{u}|\mathcal{S}_{j(\ell)},\boldsymbol{\theta}_{j(\ell)}),\mathcal{S}(\mathbf{u}|\mathcal{S}_{j(\ell)},\boldsymbol{\theta}_{j(\ell)})\right), \quad (15)$$

with $\widetilde{\boldsymbol{\mu}}(\mathbf{u}|\mathcal{S}_j,\boldsymbol{\theta}_j) = \mathbf{x}(\mathbf{u})^\mathsf{T}\boldsymbol{\beta} + \mathbf{C}(\mathbf{u},\mathcal{S}_j|\boldsymbol{\theta}_j)\mathbf{C}^{-1}(\mathcal{S}_j,\mathcal{S}_j|\boldsymbol{\theta}_j)\{\mathbf{y}(\mathcal{S}_j) - \mathbf{X}(\mathcal{S}_j)\boldsymbol{\beta}\}$, and $\mathcal{S}(\mathbf{u}|\mathcal{S}_{j(\ell)},\boldsymbol{\theta}_{j(\ell)})$ taking the same form as in (9).

### 2.5. Theoretical Properties

In this section, we establish posterior concentration results for the SPGP regression model under the assumption that $\mathcal{D} = [0,1]^2$ and the true spatial field is a piecewise smooth function. Our theoretical results can be easily extended to a more general domain that is homeomorphic to the unit square with the Euclidean metric and a bi-Lipschitz homeomorphism. Throughout this section, we focus on the case where $\mathcal{S}^* = \mathcal{S}$. Assuming $y(\mathbf{s})$ has zero mean for simplicity, our model can be written as

$$y(\mathbf{s}) = \widetilde{w}(\mathbf{s}) + \varepsilon(\mathbf{s}), \quad \varepsilon(\mathbf{s}) \sim \mathrm{N}_1\left\{0,\tau^2(\mathbf{s})\right\} \quad (16)$$

where $\widetilde{w}(\mathbf{s})$ is assigned an $L$-SPGP prior with local isotropic Matérn parameters $\{\sigma_j^2,\phi_j\}$ and a common smoothness parameter $\nu$, and $\varepsilon(\mathbf{s})$ is the nugget effect with a piecewise constant variance $\{\tau_j^2\}$.

We adopt the following notations. Given two positive sequences $\{a_n\}$ and $\{b_n\}$, $a_n = o(b_n)$ means $\lim_{n\to\infty}(a_n/b_n) = 0$ and $a_n \asymp b_n$ means $0 < \liminf_{n\to\infty}(a_n/b_n) \le \limsup_{n\to\infty}(a_n/b_n) < \infty$. The posterior given data $(\mathbf{y}(\mathcal{S}),\mathcal{S})$ is denoted by $\Pi_n(\cdot|\mathbf{y}(\mathcal{S}),\mathcal{S})$.

We first state the assumptions on the true data generating process. We assume the responses are generated according to (16) with a piecewise smooth true mean function $\widetilde{w}^*(\mathbf{s})$ and a piecewise constant true nugget variance $\tau^{*2}(\mathbf{s})$. More precisely, we let $\pi^*_{k^*}(\mathcal{D}) = \{\mathcal{D}_1^*,\ldots,\mathcal{D}_{k^*}^*\}$ be the true contiguous partition of $[0,1]^2$ with some fixed $k^*$ and a fixed boundary set $\mathcal{B}^* \subset [0,1]^2$ (see Supplementary Section S1.2 for the definition). We assume the following smoothness conditions on the true spatial field in each $\mathcal{D}_j^*$.

*Assumption T.* We assume $\widetilde{w}^*(\mathbf{s})$ and $\tau^{*2}(\mathbf{s})$ satisfy

$$\widetilde{w}^*(\mathbf{s}) = \sum_{j=1}^{k^*}\widetilde{w}_j^*(\mathbf{s})\mathbb{1}(\mathbf{s}\in\mathcal{D}_j^*), \quad \tau^*(\mathbf{s}) = \sum_{j=1}^{k^*}\tau_j^*\mathbb{1}(\mathbf{s}\in\mathcal{D}_j^*),$$

for some functions $\widetilde{w}_j^* \in C^\beta[0,1]^2 \cap H^\beta[0,1]^2$ and constants $\tau_j^* > 0$ that are fixed as $n$ grows, where $C^\beta[0,1]^2$ and $H^\beta[0,1]^2$ are the Hölder space and the Sobolev space of regularity $\beta$, respectively. Further, we assume that $\widetilde{w}_j^*(\cdot)$ is within the support of a GP prior with an isotropic Matérn covariance $\sigma_j^{2*}\rho(\cdot,\cdot|\phi_j^*,\nu)$ for some constants $\sigma_j^{*2},\phi_j^*$, and a known $\nu \ge \beta$.

We adopt a random design framework where the number of sampling locations within a fixed domain diverges to infinity. We assume the following on the spatial design and spatial graph of $n$ points $\mathbf{s}_1,\ldots,\mathbf{s}_n$ in $\mathcal{D}$.

*Assumption SD.* Given $n \in \mathbb{N}$, we assume $\mathcal{S}$ is a sequence of $n$ independent points where each point is distributed on $[0,1]^2$ with a probability density function $p_s$ such that $0 < p_s^{\min} \le p_s(\mathbf{s}) \le p_s^{\max} < \infty$. We assume the spatial graph on $\mathcal{S}$ is constructed by (i) the R-NN graph with a radius $\gamma_1 \asymp \sqrt{\log n/n}$ and $\gamma_1 > \gamma_0$, where $\gamma_0$ is the maximum edge length of the MST on $\mathcal{S}$; or (ii) the Delaunay triangulation graph where the edges are removed if they are longer than $\gamma_2$, where $\gamma_2 \asymp \sqrt{\log n/n}$ and $\gamma_2 > \gamma_0$.

Given the true space partition and a spatial graph $\mathcal{G} = (\mathcal{S},\mathcal{E})$, we say an edge $(\mathbf{s}_i,\mathbf{s}_{i'}) \in \mathcal{E}$ is across the true boundary $\mathcal{B}^*$ if $\mathbf{s}_i \in \mathcal{D}_j^*$ and $\mathbf{s}_{i'} \in \mathcal{D}_{j'}^*$ for some $j \ne j'$. If the set of all edges within a spanning tree $\mathcal{T}$ that are across $\mathcal{B}^*$ is removed, one obtains a partition of $\mathcal{S}$, denoted by $\pi^*_{k^*_\mathcal{T}}(\mathcal{S})$, that is nested in the true partition $\pi^*_{k^*}(\mathcal{S}) = (\mathcal{S}\cap\mathcal{D}_1^*,\ldots,\mathcal{S}\cap\mathcal{D}_{k^*}^*)$ of $\mathcal{S}$, and with the number of clusters $k^*_\mathcal{T} \ge k^*$. Assumption SD guarantees that the maximum number of edges across $\mathcal{B}^*$ in any spanning tree scales with $\sqrt{n\log n}$ with probability tending to 1 (Luo, Sang, and Mallick 2021b). This implies $k^*_\mathcal{T} \le c_1\sqrt{n\log n}$ for some constant $c_1 > 0$ and any $\mathcal{T}$, which plays a crucial role in establishing the prior concentration around the true model.

We further assume the prior satisfies the following condition, which guarantees the partition $\pi^*_{k^*_\mathcal{T}}(\mathcal{S})$ is within the support of the prior given an arbitrary spanning tree $\mathcal{T}$. It also regularizes the partition model so that the number of obtained clusters is not too large.

*Assumption P1.* We assume $\bar{k}_n$ satisfies $c_1\sqrt{n\log n} \le \bar{k}_n \le c_1'\sqrt{n\log n}$ for some positive constants $c_1' > c_1$.

We are now ready to state our first posterior concentration result. We denote by $p(y|\mathbf{s})$ the conditional density of the response given the sampled location, whereas the true one is denoted by $p^*(y|\mathbf{s})$. Note that $p(y|\mathbf{s})$ depends on the partition and covariance parameters. The following theorem shows that $p(y|\mathbf{s})$ concentrates in a weak neighborhood of $p^*(y|\mathbf{s})$ asymptotically under a random spatial design for $\mathcal{S}$. Its proof is deferred to Supplementary Section S1.

*Theorem 1 (Weak consistency).* Let $g$ be any bounded continuous function. Define the weak $\epsilon$-neighborhood of true density $p^*(y|\mathbf{s})$ for any $\epsilon > 0$ as

$$W_{g,\epsilon} = \left\{p : \left|\int g(y|\mathbf{s})p(y|\mathbf{s})p_s(\mathbf{s})dyd\mathbf{s} - \int g(y|\mathbf{s})p^*(y|\mathbf{s})p_s(\mathbf{s})dyd\mathbf{s}\right| < \epsilon\right\}, \text{ for any } g.$$

Under Assumptions T, SD, and P1, the posterior distribution satisfies $\Pi_n\left(W^c_{g,\epsilon} \mid \mathbf{y}(\mathcal{S}), \mathcal{S}\right) \to 0$ almost surely under $p^*(y|\mathbf{s})p_s(\mathbf{s})$ for any $g$.

To establish posterior contraction rate results, we need additional assumptions on the priors and the spatial graph. Let $\epsilon_n$ be a sequence going to zero such that $\epsilon_n \asymp (\log n/n)^\delta$ with some constant $0 < \delta < \min\{\beta/(8\nu + 8 - 4\beta),\ 1/4 - 1/(2\alpha)\}$, where $\alpha = \lfloor \nu \rfloor$.

*Assumption P2.* Let $\Pi_\phi$, $\Pi_\sigma$, and $\Pi_\tau$ denote the priors for $\phi, \sigma^2$, and $\tau$, respectively.

(P2-1) Assume that $\nu \geq \max(3, \beta)$.

(P2-2) There exist sequences $\widetilde{\phi}_n$, $\widetilde{\sigma}_n$ and $M_n$ satisfying that, as $n \to \infty$,

$$
\begin{aligned}
&-\log \Pi_\phi(\phi < \widetilde{\phi}_n^{-1})/(n\epsilon_n^2) \to +\infty,\\
&-\log \Pi_\sigma(\sigma^2 > \widetilde{\sigma}_n^{-2})/(n\epsilon_n^2) \to +\infty,\\
&\bar{k}_n(M_n/\epsilon_n)^{2/\alpha} = o(n\epsilon_n^2),\\
&M_n^2\widetilde{\sigma}_n^2\widetilde{\phi}_n^{-2\alpha}/(n\epsilon_n^2) \to +\infty.
\end{aligned}
$$

(P2-3) $\Pi_\tau$ is supported on $[a, b] \subset \mathbb{R}$ with $0 < a \leq \tau_j^{*2} \leq b < +\infty$ for all $j = 1, \ldots, k^*$.

*Assumption SG.* Let $\xi_n(k)$ be the number of unique spatially contiguous partitions with $k$ clusters of the graph $\mathcal{G}$ on $\mathcal{S}$. We assume $\mathcal{G}$ is constructed such that $\log\left(\max_{1\leq k\leq \bar{k}_n} \xi_n(k)\right) = O(n\epsilon_n^2)$.

Assumptions (P2-1) and (P2-2) on the priors of covariance functions allow us to construct a sieve on $\widetilde{w}$ that has a desired tail probability and metric entropy; similar assumptions can be found in Ghosal and Roy (2006) and Payne et al. (2020). Assumption (P2-3) is a standard assumption in the literature for nonparametric regressions with GP priors (see van der Vaart and van Zanten 2008; Bhattacharya, Pati, and Dunson 2014, among others), which is used to construct a sieve on $\tau^2$. Assumption SG excludes some graphs that are too dense and constrains the complexity of the space of all possible partitions so that the test functions with a desired probability of Type-I errors exist.

The next theorem suggests that the posterior contracts with rate $\epsilon_n$ at $p^*(y|\mathbf{s})$ with respect to the expected total variation distance. Note that this rate is slower than the minimax rate for customary GP regressions with Matérn kernels (van der Vaart and van Zanten 2011) as we pay a price for estimating the unknown partition structure using the flexible RST prior. Detailed proof is provided in Supplementary Section S1.

*Theorem 2 (Posterior contraction).* Under the same assumptions in Theorem 1 as well as Assumptions P2 and SG, the posterior distribution satisfies

$$
\Pi_n\left(\int \left|p(y|\mathbf{s}) - p^*(y|\mathbf{s})\right| p_s(\mathbf{s})dyd\mathbf{s} \geq M\epsilon_n \mid \mathbf{y}(\mathcal{S}), \mathcal{S}\right) \longrightarrow 0
$$

almost surely under $p^*(y|\mathbf{s})p_s(\mathbf{s})$ for some constant $M > 0$.

## 3. Computational Strategies

### 3.1. Estimation

The unknown parameters of the proposed SPGP regression model involve the soft RST partition parameters $\left(\boldsymbol{\Psi}(\mathcal{S}, \mathcal{S}^*), \pi_k(\mathcal{S}^*), k, \mathcal{T}\right)$ with $\boldsymbol{\Psi}(\mathcal{S}, \mathcal{S}^*) = \{\boldsymbol{\Psi}(\mathbf{s}, \mathcal{S}^*)\}_{\mathbf{s}\in\mathcal{S}}$, the cluster-specified covariance parameters $\boldsymbol{\Theta} = \left\{\tau_j^2, \sigma_j^2, \widetilde{\boldsymbol{\theta}}_j\right\}_{j=1:k}$ with local correlation parameters $\widetilde{\boldsymbol{\theta}}_j$, and the global parameters $(\boldsymbol{\beta}, \lambda)$. Conditional on $\left(\boldsymbol{\Psi}(\mathcal{S}, \mathcal{S}^*), \pi_k(\mathcal{S}^*), k, \mathcal{T}\right)$ and $\boldsymbol{\Theta}$, the global parameters can be updated via standard Bayesian inference methods. Specifically, we sample $\boldsymbol{\beta}$ and $\lambda$ from their posterior conditional distributions, which follow a multivariate normal and an inverse gamma distribution, respectively. The detailed forms are included in Supplementary Section S2.1.

Below, we focus on the adaptive estimation of the soft partition parameters and the local covariance parameters conditional on $(\boldsymbol{\beta}, \lambda)$. As the number of clusters is assumed unknown, this trans-dimensional inference is done via a tailored reversible jump Markov chain Monte Carlo (RJ-MCMC) sampler (Green 1995; Luo, Sang, and Mallick 2021b). Taking advantage of the tree structure, each trans-dimensional move can be achieved by simply adding and/or deleting an edge in the tree. The acceptance ratio of the proposed RJ-MCMC move involves the calculation of likelihood ratios, a major computation bottleneck in Bayesian GP models. We will show that each move under SPGP only changes the cluster memberships of a smaller subset of observations, and hence only the likelihood ratios involving this subset of data need to be calculated. An additional advantage of the locally stationary model is that it allows estimating cluster-specific parameters $\boldsymbol{\Theta}$ using only the data in each subregion. By doing so, SPGP naturally leads to a reduced computation from fitting a global GP model to a number of local GP models. In this article, when a cluster contains a large number of observations, we employ the nearest neighbor GP (NNGP) methods (Datta et al. 2016) to speed up the local likelihood calculation.

Specifically, we reparameterize the covariance function by setting $\sigma_j^2 = \tau_j^2\bar{\sigma}_j^2$ and place a conjugate inverse gamma prior for $\boldsymbol{\tau}^2 = \{\tau_j^2\}_{j=1:k}$ that allows us to integrate $\tau_j^2$ out analytically when we update the partitions and other cluster-specific parameters, which improves mixing and convergence of our sampler.

To collect samples from $\left(\{\bar{\sigma}_j^2, \widetilde{\boldsymbol{\theta}}_j\}_{j=1:k}, \boldsymbol{\Psi}(\mathcal{S}, \mathcal{S}^*), \pi_k(\mathcal{S}^*), k, \mathcal{T}\right)|(\boldsymbol{\beta}, \lambda)$, one of the four moves—*birth*, *death*, *change*, and *hyper*—is performed with probabilities $r_b(k)$, $r_d(k)$, $r_c(k)$, and $r_h(k)$, respectively. The first three moves modify the partition $\pi_k(\mathcal{S}^*)$ given $\mathcal{T}$ and $\boldsymbol{\Psi}(\mathcal{S}, \mathcal{S}^*)$. The use of $\mathcal{S}^*$ with $m \ll n$ allows us to work on a smaller graph and therefore a much smaller space of $\pi_k(\mathcal{S}^*)$, which speeds up convergence and encourages better mixing.

In a *birth* move, one of the $k$ clusters in $\pi_k(\mathcal{S}^*)$ is randomly chosen with equal probabilities, and then the chosen cluster is split into two by randomly removing an edge in $\mathcal{T}$ that connects vertices in the cluster. Suppose that $\mathcal{S}^*_{j_0}$ is chosen to be split into $\mathcal{S}^*_{j_1}$ and $\mathcal{S}^*_{j_2}$. In the case where $\mathcal{S}^* \neq \mathcal{S}$, $\mathcal{S}_{j_0}$ is also split into two clusters $\mathcal{S}_{j_1}$ and $\mathcal{S}_{j_2}$ according to (4a) given $\boldsymbol{\Psi}(\mathcal{S}, \mathcal{S}^*)$. One of $\mathcal{S}_{j_1}$ and $\mathcal{S}_{j_2}$ is uniformly chosen to inherit the parameters $(\bar{\sigma}^2, \widetilde{\boldsymbol{\theta}})$ from the original cluster. As there is no conjugate prior for $\bar{\sigma}^2$ or $\widetilde{\boldsymbol{\theta}}$, standard Metropolis-Hastings (M-H) updates can lead to

low efficiency. To address this, following Payne et al. (2020), the $(\bar{\sigma}^2, \widetilde{\boldsymbol{\theta}})$ for the other new cluster, say $\mathcal{S}_{j_2}$, are chosen to maximize $p\left\{\mathbf{y}(\mathcal{S}_{j_2})|\bar{\sigma}^2, \widetilde{\boldsymbol{\theta}}, -\right\} p(\bar{\sigma}^2)p(\widetilde{\boldsymbol{\theta}})$, where $p(\bar{\sigma}^2)$ and $p(\widetilde{\boldsymbol{\theta}})$ are the prior densities for $\bar{\sigma}^2$ and $\widetilde{\boldsymbol{\theta}}$, respectively, and $p\left\{\mathbf{y}(\mathcal{S}_j)|\bar{\sigma}_j^2, \widetilde{\boldsymbol{\theta}}_j, -\right\}$ is the likelihood function of $\mathbf{y}(\mathcal{S}_j)$ with $\tau^2$ integrated out. The M-H ratio is therefore

$$(1-c)\times\frac{r_d(k+1)}{r_b(k)}\times\frac{p\left\{\mathbf{y}(\mathcal{S}_{j_1})|\bar{\sigma}_{j_1}^2, \widetilde{\boldsymbol{\theta}}_{j_1}, -\right\} p\left\{\mathbf{y}(\mathcal{S}_{j_2})|\bar{\sigma}_{j_2}^2, \widetilde{\boldsymbol{\theta}}_{j_2}, -\right\}}{p\left\{\mathbf{y}(\mathcal{S}_{j_0})|\bar{\sigma}_{j_0}^2, \widetilde{\boldsymbol{\theta}}_{j_0}, -\right\}}, \tag{17}$$

which only involves likelihood functions on subsets of $\mathcal{S}$. Opposite to the *birth* move, a *death* move randomly merges two adjacent clusters in $\pi_k(\mathcal{S}^*)$. Specifically, an edge in $\mathcal{T}$ that connects two distinct clusters in $\pi_k(\mathcal{S}^*)$ is uniformly selected and then the two clusters are merged. The corresponding two clusters in $\pi_k(\mathcal{S})$ are also merged accordingly. The parameters $(\bar{\sigma}^2, \widetilde{\boldsymbol{\theta}})$ of the merged clusters are chosen using a similar maximum a posteriori (MAP) approach as in the birth move. The M-H ratio is analogous to (17). In a *change* move, a *death* move is performed followed by a *birth* move, so that the number of clusters is unchanged. This move is designed to encourage better mixing of the Markov chain.

A *hyper* move updates $\mathcal{T}$ and $\boldsymbol{\Psi}(\mathcal{S}, \mathcal{S}^*)$ given other parameters. Specifically, $\mathcal{T}$ is updated using the exact sampler similar as in Luo, Sang, and Mallick (2021b), which adaptively learns a desired spanning tree spatial order to better recover the true partition. We sample the edge weight $\omega_{i,i'}$ of $\mathcal{G}$ from iid Unif$(1/2, 1)$ if the vertices $\mathbf{s}_i^*$ and $\mathbf{s}_{i'}^*$ are in different clusters under $\pi_k(\mathcal{S}^*)$, and otherwise from iid Unif$(0, 1/2)$. A new spanning tree is the MST generated by Prim's algorithm using the new edge weights. Given $\pi_k(\mathcal{S}^*)$, we use a Gibbs sampler to iteratively sample $\boldsymbol{\Psi}(\mathbf{s}, \mathcal{S}^*)$ from its full conditional distribution, which admits a closed form Multinomial$(1, \widetilde{\boldsymbol{\alpha}}(\mathbf{s}, \mathcal{S}^*))$ distribution. The $\ell$th element $(\ell = 1, \ldots, m)$ of $\widetilde{\boldsymbol{\alpha}}(\mathbf{s}, \mathcal{S}^*)$ is given by $\widetilde{\alpha}(\mathbf{s}, \mathbf{s}_\ell^*) \propto p\{\mathbf{y}(\mathcal{S})|z(\mathbf{s}) = z(\mathbf{s}_\ell^*), -\}\alpha(\mathbf{s}, \mathbf{s}_\ell^*)$ if $\mathbf{s}_\ell^*$ is among the $L$ nearest reference knots of $\mathbf{s}$, and $\widetilde{\alpha}(\mathbf{s}, \mathbf{s}_\ell^*) = 0$ otherwise, where $p\{\mathbf{y}(\mathcal{S})|z(\mathbf{s}) = z(\mathbf{s}_\ell^*), -\}$ is the likelihood when $\mathbf{s}$ is assigned to the same cluster as $\mathbf{s}_\ell^*$ with other cluster memberships $\mathbf{z}(\mathcal{S} \setminus \{\mathbf{s}\})$ fixed. Note that for any $\mathbf{s} \in \mathcal{S}$ whose $L$ nearest neighbors in $\mathcal{S}^*$ all belong to the same cluster, $\widetilde{\alpha}(\mathbf{s}, \mathbf{s}_\ell^*) = \alpha(\mathbf{s}, \mathbf{s}_\ell^*)$, suggesting we can simply update $\boldsymbol{\Psi}(\mathbf{s}, \mathcal{S}^*)$ from its prior. Thanks to the $L$-SPP and local GP formulation, in practice, we can employ several computation tricks to further reduce the computation of the likelihood, including rank-one Cholesky update/downdate (Golub and Van Loan 2013, Section 6.5.4). Details can be found in Supplementary Section S2.1.

Finally, we update the parameters $\{\tau_j^2\}_{j=1:k}$ by sampling from their inverse gamma full conditionals, whose closed forms are given in Supplementary Section S2.1. Given posterior samples of all the parameters, one can choose $L$ by standard model selection techniques such as deviance information criterion (DIC; Spiegelhalter et al. 2002).

### 3.2. Prediction

Posterior predictive inference in the SPGP model can be achieved via (15). Let $\mathcal{U} = \{\mathbf{u}_1, \ldots, \mathbf{u}_r\}$ be a collection of locations where the responses are unobserved. Conditional on a posterior draw of the parameters, we can sample $\mathbf{z}(\mathcal{U})$ according to (5) and then $\mathbf{y}(\mathbf{u})$ from (15). The detailed algorithm is provided in Supplementary Section S2. Note that the prediction algorithm is parallelizable, as predictions at each $\mathcal{U}_j$ are independent and only depend on data in $\mathcal{S}_j$ at a time given a sample of cluster memberships of $\mathcal{U}$ and $\mathcal{S}$.

Thanks to the Gaussian mixture structure in the predictive distributions, the prediction uncertainty at points near boundaries will be reflected by their oftentimes multi-modal predictive distributions. As discussed in Section 2.3, the kriging mean predictive surface around the estimated boundary becomes smoother as $L$ grows. The surface can be further smoothed by using Bayesian model averaging to account for model estimation uncertainties (Gramacy and Lee 2008). We remark that the usual kriging mean and SD estimates may not be the ideal choice to summarize the possibly multi-modal prediction results of the SPGP model near boundaries. Instead, we recommend using the highest posterior density (HPD) region to capture multimodality by potentially disjoint HPD intervals.

## 4. Simulation Studies

In this section, we assess the performance of the SPGP regression model by some simulated data. We consider a squared spatial domain $\mathcal{D} = [0, 1]^2$ that is partitioned into two subregions $\mathcal{D}_1^*$ and $\mathcal{D}_2^*$ with the true partition boundary $\mathcal{B}^*$ given by a circle of radius 0.3 centered at $(0.5, 0.5)$. We generate $n = 2000$ uniform locations $\mathcal{S}$ for training and $r = 400$ hold-out locations $\mathcal{U}$ for prediction. 75% of $\mathcal{U}$ are sampled near $\mathcal{B}^*$, allowing us to assess the prediction performance primarily near $\mathcal{B}^*$ where the abrupt changes happen. See Figure 2(a) for the sampled locations. The responses are generated from (16) using isotropic Matérn covariance functions, where the true parameters of the processes in $\mathcal{D}_1^*$ and $\mathcal{D}_2^*$ have well-separated microergodic parameters $\vartheta = \sigma^2/\phi^{2\nu}$, and $\nu$ is treated as known. It is shown in Zhang (2004) that $\vartheta$ matters more in prediction and can be consistently estimated under the infill asymptotic framework, while $\sigma^2$ or $\phi$ cannot. We choose $m = 616$ reference knots $\mathcal{S}^*$ using a data-driven approach detailed in Supplementary Section S3.2 such that more knots are placed near the initially estimated boundary. We utilize NNGP approximation with 15 neighbors to speed up computation and select the optimal value of $L$ from $\{1, \ldots, 5\}$ by DIC. The detailed data generation process, prior specifications, and other model choices can be found in Supplementary Section S3.2.

We compare the SPGP model with axis-parallel binary decision treed GP (TGP) models (Gramacy and Lee 2008), nonstationary GP (NSGP) models developed in Paciorek and Schervish (2006) and Risser and Turek (2020), and stationary GP (SGP) spatial regressions with isotropic Matérn covariance functions (see, e.g., Banerjee, Carlin, and Gelfand 2014).

We first examine the partition recovery performance of the SPGP model with $L = 3$ chosen by DIC and compare it with that of TGP. Figure 2 shows the MAP estimates of the partition and the cluster-specified log microergodic parameter estimates. The partition estimated by SPGP is fairly consistent with the truth considering that the estimation results are based on just one realization of the random field. On the other hand, due

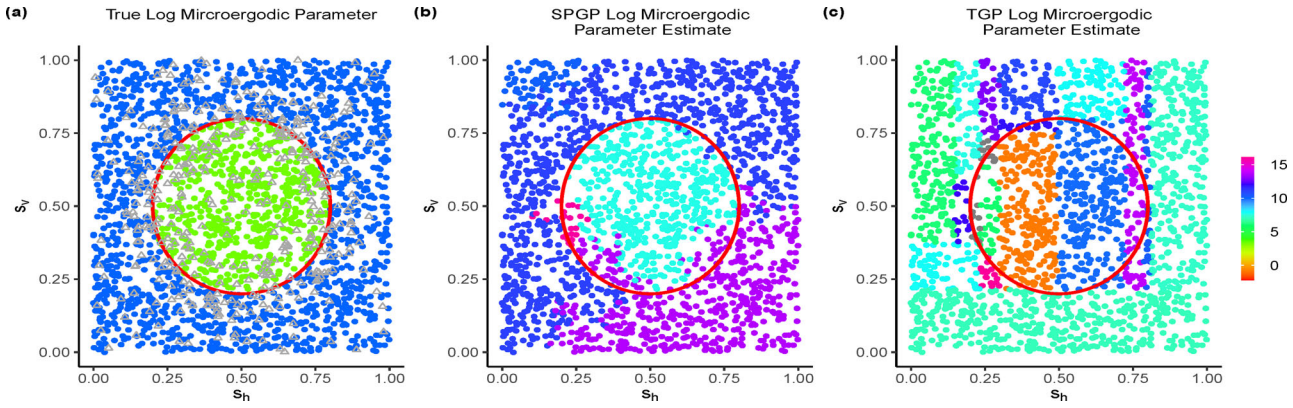


**Figure 2.** (a) True log microergodic parameters at training locations. Hold-out locations are marked by gray triangles. (b), (c) MAP estimates of log microergodic parameters given by SPGP and TGP. The true boundary is marked by the red circle.

**Table 1.** Performance metrics of SPGP and its competing methods.

| | SPGP | TGP | NSGP | SGP |
|---|---|---|---|---|
| In-sample ARI | **0.429** | 0.153 | — | — |
| Hold-out ARI | **0.422** | 0.119 | — | — |
| $\text{MSE}_\vartheta$ | **13.316** | 16.974 | — | 16.610 |
| $\text{MSPE}_y$ | **0.041** | **0.041** | 0.046 | **0.041** |
| Mean $\text{CRPS}_y$ | **0.083** | 0.086 | 0.106 | 0.098 |
| Mean $\text{LogS}_y$ | **−0.655** | −0.562 | −0.227 | −0.113 |

Bold values highlight the result of the best performing model.

to the use of axis-parallel treed partitions, TGP approximates the true partition with a much larger number of rectangular clusters. The superior partition recovery performance of SPGP is also evidenced by the higher in-sample and hold-out MAP adjusted Rand indices (ARIs; Hubert and Arabie 1985) in the first two rows of Table 1. In particular, the higher ARI for hold-out locations suggests that the membership prediction from SPGP agrees more with the ground truth. We then consider the estimation accuracy of covariance parameters measured by the mean square error of the MAP estimates of the log microergodic parameters $\log(\vartheta(\mathbf{s}))$, denoted by $\text{MSE}_\vartheta$. For SGP, $\widehat{\vartheta}(\mathbf{s})$ reduces to a constant $\widehat{\sigma}^2/\widehat{\phi}^{2\nu}$. The resulting $\text{MSE}_\vartheta$'s in the third row of Table 1 indicates SPGP has the smallest estimation error, suggesting that it can estimate the spatial covariance parameters most accurately among the three methods. In particular, as shown in Figure 2, SPGP provides more accurate microergodic parameter estimation in the upper part of the outer true cluster. This is possibly because TGP has to approximate this region by many small rectangles and learn $\vartheta$ within each of them using fewer data points, while the estimated partition from SPGP has more data points in each cluster for covariance parameter estimation.

Next, we analyze the performance of out-of-sample prediction. As shown in the fourth row of Table 1, the SPGP and TGP models have the lowest mean squared prediction error (MSPE). We argue that, nonetheless, MSPE is not the most ideal metric to evaluate the performance of probabilistic prediction as it may not fully take into account the potential multimodality in the posterior predictive distributions. As a result, it is more sensible to compare scoring rules (Gneiting and Raftery 2007) such as average continuous ranked probability score (CRPS) and logarithmic score (LogS), which are presented in the last two rows in Table 1. The SPGP model has the best scores overall among all models. The superior performance of SPGP over TGP is partly due to its ability to produce a more accurate predictive distribution, which reflects partition uncertainties around the true boundary. This makes SPGP more robust to misclassification of cluster memberships, thanks to its soft partition process model with $L = 3$.

Finally, Figure 3 displays the posterior mean predictive surfaces and posterior prediction SDs from SPGP, TGP, and NSGP. We also include the kriging results from the model where the true partition and other parameters are known as a benchmark. As expected, both SPGP and TGP generate similar, accurate predictions in the interior of the true clusters, while one can observe some differences around the true boundary. Due to the Gaussian mixture predictive distributions and Bayesian model averaging, we obtain a relatively smooth kriging mean surface from SPGP near the true boundary rather than a sharp jump, but overall it still well captures the jump in the true surface. As desired, the prediction SDs are higher around the true boundary, capturing the uncertainty from the unknown partition. Note that the prediction uncertainty, characterized by lower posterior prediction SD values, is smaller near regions with a smaller jump in the true field. On the surface from TGP, some discontinuities can be observed near the estimated boundaries, and the shape of the estimated boundaries is less flexible, with some noticeable artifact axis-parallel patterns. In addition, some high prediction uncertainty regions from TGP do not align with the true boundary. Despite capturing the pattern of the true field, the predictive surface of NSGP is less smooth compared with the truth and the SD plot from NSGP is less informative on where the true boundary is. This is possibly because the nonstationarity modeling components in NSGP are misspecified and/or NSGP is better suited for the case where the change of covariance is relatively smooth. In contrast, the advantage of our method is more prominent when the true covariance function has irregular abrupt changes or clustering patterns.

Since it is insufficient to visualize a possibly multimodal posterior predictive distribution via its mean and SD, we further examine the plots of predictive densities for selected locations (see Figure S4 in Supplementary Section S3.2. Our results confirm that SPGP can quantify prediction uncertainty in a desirable way, where the higher mode appears near the true value and the corresponding 95% HPD interval also covers the true value. In contrast, the posterior predictive densities from TGP

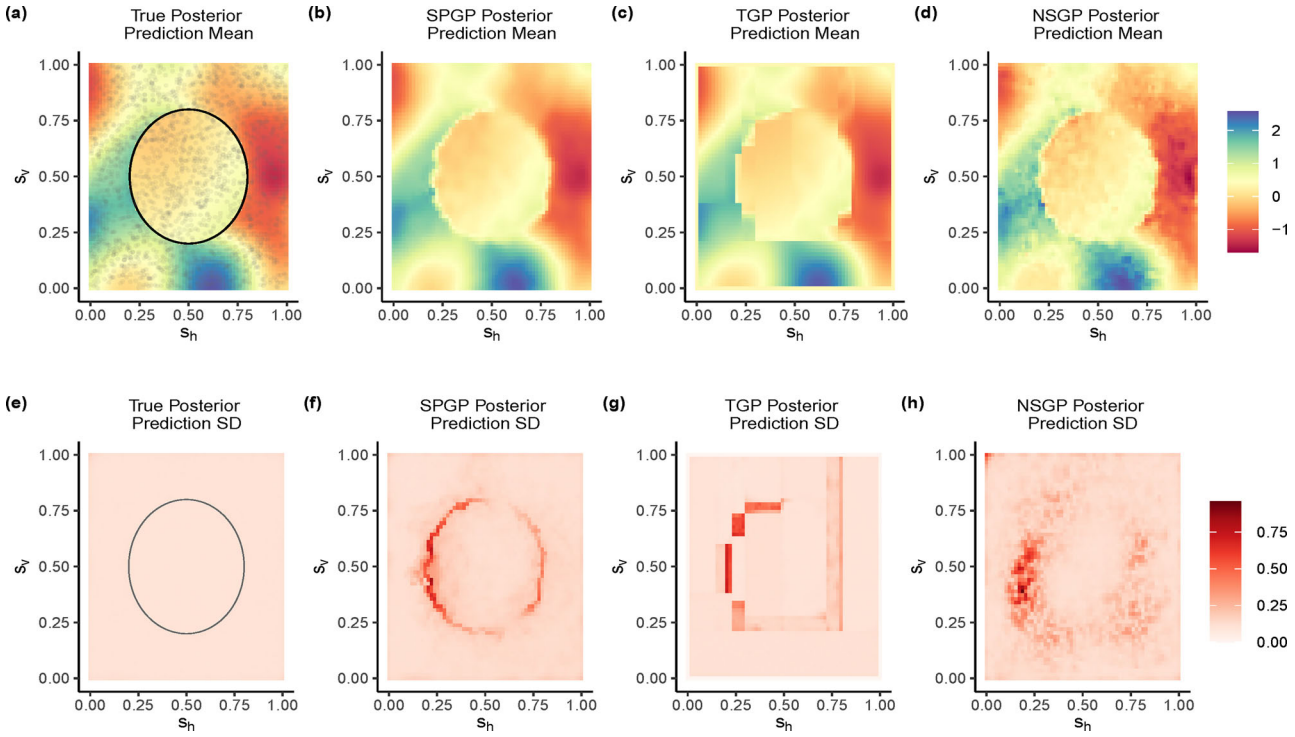


**Figure 3.** Posterior mean predictive surfaces (a)–(d) and SD surfaces (e)–(h) form the true data generating model, SPGP, TGP, and NSGP. The black circle represents the true boundary and the grey points represent observed locations.

and NSGP fail to capture the multimodality when prediction locations are near the true boundaries.

We have also investigated the scenarios where the data are generated from a piecewise anisotropic process or a globally nonstationary process in Paciorek and Schervish (2006). Our results again confirm the competitive performance of SPGP. See Supplementary Sections S3.3 and S3.4 for details.

## 5. Real Data Analysis

We apply the SPGP regression model to analyze the precipitation data over the contiguous United States (CONUS). The dataset consists of daily average precipitation over the 2018 water year (October 1, 2017 to September 30, 2018) obtained from the Global Historical Climatology Network-Daily database (GHCN-D), and was analyzed in Risser and Turek (2020). As noted in Risser and Turek (2020), the precipitation data in the western half of the CONUS is highly nonstationary due to the heterogeneous topography and the diverse physical phenomena related to precipitation. As a result, we focus on the precipitation data measured at GHCN-D stations located to the west of 90°W and use $n = 1689$ uniformly selected locations out of 1939 stations for model training and hold out the rest for testing. We perform a logarithmic transform of the precipitation rates following Risser and Turek (2020) so that the GP assumption is more applicable. The observed locations and the associated log precipitation rates are shown in Figure 4(a). The goal of this analysis is to demonstrate how well the SPGP model recovers the local stationarity structure in the precipitation data and predicts the precipitation at unobserved locations, especially around partition boundaries.

To model the log precipitation rates, we consider the SPGP regression with $\mathcal{S}^* = \mathcal{S}$, a spatially constant intercept, and geometric anisotropic Matérn local covariance functions. We compare the SPGP model with TGP and NSGP. The detailed specifications of all models can be found in Supplementary Section S4. We also perform predictive analysis of the log precipitation rates at hold-out locations in the same manner as in Section 4.

Figure 4(b)–(c) shows the MAP estimates for partitions from SPGP and TGP. The partition given by SPGP can be largely explained by the topography in the CONUS: Cluster 1 covers the Interior Plains and the Interior Highlands to the east of the Rocky Mountains, while Cluster 3 corresponds to the mountainous regions including the Rocky Mountain System, the Intermontane Plateaus, and most parts of the Pacific Mountains. The small Cluster 2 mainly consists of the desert region in southern California with low precipitation rates. This suggests that SPGP can capture the geographic heterogeneity in the precipitation data. The TGP model identifies more clusters, some of which partly overlap with the clusters from SPGP but others are quite different. For example, in the partition from SPGP, the northern Montana region shares the same cluster membership with the regions to its east, while this is not the case in TGP. One possible reason is that the binary trees used in TGP may partition an irregularly shaped region into several subregions with horizontal or vertical boundaries. Another possible reason is that the TGP model uses less flexible separable exponential covariance functions compared with the geometric anisotropic ones in SPGP.

As in the simulation studies, we use MSPE, average CRPS, and average LogS to quantify the performance of predicting out-of-sample log precipitation rates. Table 2 summarizes the results

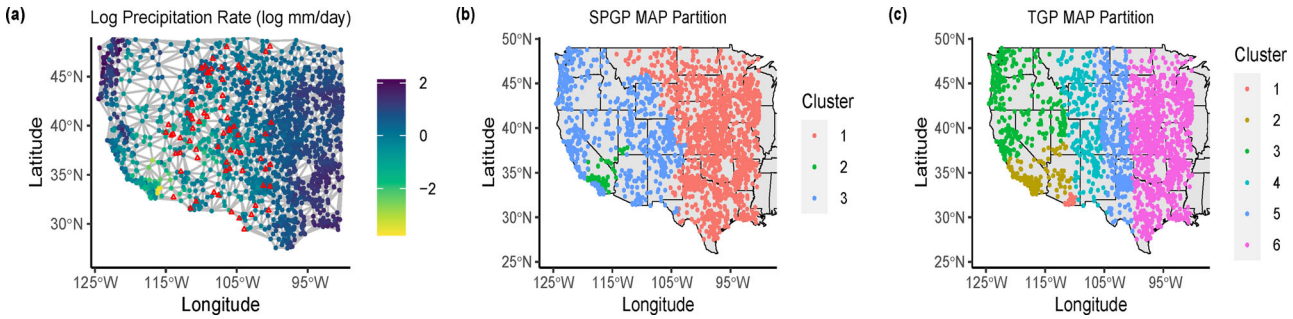


**Figure 4.** (a) Log precipitation rate at $n = 1689$ stations and the Delaunay triangulation graph used for model fitting. $r = 75$ hold-out locations near the Rocky Mountains are marked as red triangles. (b)–(c) MAP partition estimates of the training locations given by SPGP and TGP.

**Table 2.** Prediction performance for the precipitation data on $r_1 = 75$ hold-out locations.

| | SPGP ($L = 1$) | SPGP ($L = 3$) | SPGP ($L = 5$) | TGP | NSGP |
|---|---|---|---|---|---|
| MSPE | **0.073** | **0.073** | 0.075 | 0.093 | 0.081 |
| Mean CRPS | 0.145 | **0.143** | **0.143** | 0.159 | 0.152 |
| Mean LogS | −0.001 | −0.017 | **−0.019** | 0.076 | 0.083 |

Bold values highlight the result of the best performing model.

based on $r_1 = 75$ hold-out locations between 100°W and 115°W near the Rocky Mountains that contain many boundary points identified by SPGP and TGP. The SPGP models achieve the best predictive performance in all three metrics. We have also investigated the prediction results based on $r_2 = 175$ hold-out locations that are not near the Rocky Mountains area, which suggest the comparable performance of all models under this prediction scenario. The details are provided in supplementary Section S4. In summary, our results indicate that the gain in the prediction performance when using SPGP over other methods is more prominent for boundary locations. We have also examined the predictive surfaces and SDs at equally spaced points. The results from all three models look similar and are provided in supplementary Section S4.

## 6. Conclusions and Discussion

In this article, we have developed a novel soft partitioned Gaussian process to capture locally stationary spatial structures. Our process is constructed conditional on a predictive random spanning tree soft partition process in the spatial domain. We embed it into a Bayesian hierarchical spatial modeling framework, leading to a soft partitioned GP regression model. The prediction of SPGP uses a mixture of $L$ Gaussian distributions, where $L$ is the number of nearest neighbors used for determining cluster memberships.

The proposed SPGP model can be extended in several directions. SPGP can be applied in a spatial GLM framework for the analysis of non-Gaussian spatial responses. Another future research direction is to extend the univariate process into multivariate cases, possibly with multiple spanning-treed partitions and tree-based graphical models (Gao, Datta, and Banerjee 2021). Extension to soft partitioned versions of other types of stochastic processes is also possible if their conditional distributions are available. It is known that nearest neighbor graphs and Delaunay triangular meshes are capable of capturing more complex geometries. Therefore, a promising direction of future research is to extend our graph-based SPGP to build locally stationary processes on complex domains (Luo, Sang, and Mallick 2021a). To model smoother functions, it is possible to extend SPGP by considering an additive form of multiple SPGP models with different partitions (Luo, Sang, and Mallick 2022; Maia, Murphy, and Parnell 2022). On the computational side, we have demonstrated that scalability can be straightforwardly achieved by specifying a small-sized set of reference knots and using likelihood approximation methods such as NNGP (Datta et al. 2016). It is possible to treat the choice of reference knots as unknown parameters so that the optimal choice of knots can be learned from data. Moreover, we will investigate the use of other block-based scalable GP approximations (see, e.g., Konomi, Sang, and Mallick 2014; Zhang, Sang, and Huang 2019; Peruzzi, Banerjee, and Finley 2020) for local likelihood computations.

Our theoretical results on the SPGP models suggest that the posterior distribution of the conditional density concentrates in a weak or total variation neighborhood, and we establish a contraction rate for the latter case. We remark that the rate can be potentially improved if the complexity of the RST partition space can be better bounded. For linear prediction (or kriging) problems, posterior asymptotic efficiency can possibly be established following a similar spirit of Li (2022). Besides the precipitation data in Section 5, SPGP has many other potential applications beyond spatial statistics such as the photometric redshift data in cosmology (Fadikar, Wild, and Chaves-Montero 2021). We leave these as future works.

## Supplementary Materials

**Supplementary File:** The supplementary consists of (a) The proofs of Kolmogorov Consistency of SPGP, Propositions 1, and Theorems 1 and 2; (b) Detailed algorithms of Bayesian posterior inference and prediction; (c) Additional simulation studies to investigate the performance of SPGP; (d) Additional real data analysis results.

## Acknowledgements

The authors thank the referees and the editor for their valuable comments. The authors also thank Dr. Mark Risser for providing the CONUS precipitation data.

## Funding

The research of Zhao Tang Luo and Huiyan Sang was partially supported by NSF grant no. NSF DMS-2210456 and 2220231. The research of Bani Mallick was partially supported by NSF grant no. NSF CCF-1934904 and

National Cancer Institute of the National Institutes of Health grant under award number R01CA194391.

## References


Adler, R. J., and Taylor, J. E. (2007), *Random Fields and Geometry* (Vol. 80), New York: Springer. [5]

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2014), *Hierarchical Modeling and Analysis for Spatial Data*, Boca Raton, FL: CRC Press. [8]

Banerjee, S., Gelfand, A., Knight, J. R., and Sirmans, C. (2004), "Spatial Modeling of House Prices Using Normalized Distance-Weighted Sums of Stationary Processes," *Journal of Business & Economic Statistics*, 22, 206–213. [6]

Bhattacharya, A., Pati, D., and Dunson, D. (2014), "Anisotropic Function Estimation Using Multi-Bandwidth Gaussian Processes," *Annals of Statistics*, 42, 352–381. [7]

Bolin, D., Wallin, J., and Lindgren, F. (2019), "Latent Gaussian Random Field Mixture Models," *Computational Statistics & Data Analysis*, 130, 80–93. [1]

Chew, L. P. (1987), "Constrained Delaunay Triangulations," in *Proceedings of the Third Annual Symposium on Computational Geometry*, pp. 215–222. [3]

Datta, A., Banerjee, S., Finley, A. O., and Gelfand, A. E. (2016), "Hierarchical Nearest-Neighbor Gaussian Process Models for Large Geostatistical Datasets," *Journal of the American Statistical Association*, 111, 800–812. [2,7,11]

Fadikar, A., Wild, S. M., and Chaves-Montero, J. (2021), "Scalable Statistical Inference of Photometric Redshift via Data Subsampling," in *International Conference on Computational Science*, pp. 245–258, Springer. [11]

Fouedjio, F. (2017), "Second-Order Non-stationary Modeling Approaches for Univariate Geostatistical Data," *Stochastic Environmental Research and Risk Assessment*, 31, 1887–1906. [2]

Gao, L., Datta, A., and Banerjee, S. (2021), "Hierarchical Multivariate Directed Acyclic Graph Auto-Regressive (mdagar) Models for Spatial Diseases Mapping," arXiv preprint arXiv:2102.02911. [11]

Gerber, F., and Nychka, D. W. (2021), "Parallel Cross-Validation: A Scalable Fitting Method for Gaussian Process Models," *Computational Statistics & Data Analysis*, 155, 107113. [1]

Ghosal, S., and Roy, A. (2006), "Posterior Consistency of Gaussian Process Prior for Nonparametric Binary Regression," *The Annals of Statistics*, 34, 2413–2429. [7]

Gneiting, T., and Raftery, A. E. (2007), "Strictly Proper Scoring Rules, Prediction, and Estimation," *Journal of the American Statistical Association*, 102, 359–378. [9]

Golub, G. H., and Van Loan, C. F. (2013), *Matrix Computations*, Baltimore MD: JHU Press. [8]

Gosoniu, L., and Vounatsou, P. (2011), "Non-stationary Partition Modeling of Geostatistical Data for Malaria Risk Mapping," *Journal of Applied Statistics*, 38, 3–13. [1]

Gramacy, R. B., and Lee, H. K. H. (2008), "Bayesian Treed Gaussian Process Models with an Application to Computer Modeling," *Journal of the American Statistical Association*, 103, 1119–1130. [1,8]

Green, P. J. (1995), "Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination," *Biometrika*, 82, 711–732. [7]

Heaton, M. J., Christensen, W. F., and Terres, M. A. (2017), "Nonstationary Gaussian Process Models Using Spatial Hierarchical Clustering from Finite Differences," *Technometrics*, 59, 93–101. [1]

Hubert, L., and Arabie, P. (1985), "Comparing Partitions," *Journal of Classification*, 2, 193–218. [9]

Kim, H.-M., Mallick, B. K., and Holmes, C. (2005), "Analyzing Nonstationary Spatial Data Using Piecewise Gaussian Processes," *Journal of the American Statistical Association*, 100, 653–668. [1,5]

Konomi, B. A., Sang, H., and Mallick, B. K. (2014), "Adaptive Bayesian Nonstationary Modeling for Large Spatial Datasets Using Covariance Approximations," *Journal of Computational and Graphical Statistics*, 23, 802–829. [11]

Li, C. (2022), "Bayesian Fixed-Domain Asymptotics for Covariance Parameters in a Gaussian Process Model," *The Annals of Statistics*, 50, 3334–3363. [11]

Li, F., and Sang, H. (2019), "Spatial Homogeneity Pursuit of Regression Coefficients for Large Datasets," *Journal of the American Statistical Association*, 114, 1050–1062. [1]

Lindgren, F., Rue, H., and 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, 73, 423–498. [2,4]

Luo, Z. T., Sang, H., and Mallick, B. (2021a), "BAST: Bayesian Additive Regression Spanning Trees for Complex Constrained Domain," in *Advances in Neural Information Processing Systems* (Vol. 34), pp. 90–102. [11]

Luo, Z. T., Sang, H., and Mallick, B. (2022), "BAMDT: Bayesian Additive Semi-multivariate Decision Trees for Nonparametric Regression," in *International Conference on Machine Learning*, pp. 14509–14526, PMLR. [11]

Luo, Z. T., Sang, H., and Mallick, B. K. (2021b), "A Bayesian Contiguous Partitioning Method for Learning Clustered Latent Variables," *Journal of Machine Learning Research*, 22, 1–51. [2,3,6,7,8]

Maia, M., Murphy, K., and Parnell, A. C. (2022), "GP-BART: A Novel Bayesian Additive Regression Trees Approach Using Gaussian Processes," arXiv preprint arXiv:2204.02112. [11]

Paciorek, C. J., and Schervish, M. J. (2006), "Spatial Modelling Using a New Class of Nonstationary Covariance Functions," *Environmetrics: The Official Journal of the International Environmetrics Society*, 17, 483–506. [8,10]

Park, C., Huang, J., and Ding, Y. (2011), "Domain Decomposition Approach for Fast Gaussian Process Regression of Large Spatial Data Sets," *Journal of Machine Learning Research*, 12, 1697–1728. [1]

Payne, R. D., Guha, N., Ding, Y., and Mallick, B. K. (2020), "A Conditional Density Estimation Partition Model Using Logistic Gaussian Processes," *Biometrika*, 107, 173–190. [7,8]

Peruzzi, M., Banerjee, S., and Finley, A. O. (2020), "Highly Scalable Bayesian Geostatistical Modeling via Meshed Gaussian Processes on Partitioned Domains," *Journal of the American Statistical Association*, 117, 969–982. [11]

Pope, C. A., Gosling, J. P., Barber, S., Johnson, J. S., Yamaguchi, T., Feingold, G., and Blackwell, P. G. (2021), "Gaussian Process Modeling of Heterogeneity and Discontinuities Using Voronoi Tessellations," *Technometrics*, 63, 53–63. [1]

Risser, M. D. (2016), "Nonstationary Spatial Modeling, with Emphasis on Process Convolution and Covariate-Driven Approaches," arXiv preprint arXiv:1610.02447. [2]

Risser, M. D., Calder, C. A., Berrocal, V. J., and Berrett, C. (2019), "Nonstationary Spatial Prediction of Soil Organic Carbon: Implications for Stock Assessment Decision Making," *The Annals of Applied Statistics*, 13, 165–188. [1]

Risser, M. D., and Turek, D. (2020), "Bayesian Inference for High-Dimensional Nonstationary Gaussian Processes," *Journal of Statistical Computation and Simulation*, 90, 2902–2928. [8,10]

Shen, Y., Ng, A., and Seeger, M. (2006), "Fast Gaussian Process Regression Using kd-trees," in *Proceedings of the 19th Annual Conference on Neural Information Processing Systems*, number CONF. [1]

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002), "Bayesian Measures of Model Complexity and Fit," *Journal of the Royal Statistical Society*, Series B, 64, 583–639. [8]

Teixeira, L. V., Assunção, R. M., and Loschi, R. H. (2019), "Bayesian Space-Time Partitioning by Sampling and Pruning Spanning Trees," *Journal of Machine Learning Research*, 20, 1–35. [2]

van der Vaart, A., and van Zanten, H. (2011), "Information Rates of Nonparametric Gaussian Process Methods," *Journal of Machine Learning Research*, 12, 2095–2119. [7]

van der Vaart, A. W., and van Zanten, J. H. (2008), "Rates of Contraction of Posterior Distributions based on Gaussian Process Priors," *The Annals of Statistics*, 36, 1435–1463. [7]

Zhang, B., Sang, H., and Huang, J. Z. (2019), "Smoothed Full-Scale Approximation of Gaussian Process Models for Computation of Large Spatial Data Sets," *Statistica Sinica*, 29, 1711–1737. [11]

Zhang, H. (2004), "Inconsistent Estimation and Asymptotically Equal Interpolations in Model-based Geostatistics," *Journal of the American Statistical Association*, 99, 250–261. [8]

Zhang, M. M., and Williamson, S. A. (2019), "Embarrassingly Parallel Inference for Gaussian Processes," *Journal of Machine Learning Research*, 20, 1–26. [1]