AIMS Mathematics, 10(11): 26351–26388.
DOI: 10.3934/math.20251159
Received: 10 August 2025
Revised: 29 September 2025
Accepted: 17 October 2025
Published: 14 November 2025

https://www.aimspress.com/journal/Math

*Research article*

# Spatial distributional estimation via ensemble spatial analysis

**Alvaro F. Egaña[1,*], Gonzalo Díaz[1], Felipe Navarro[1], Mohammad Maleki[2] and Juan F. Sánchez-Pérez[3]**

[1] Advanced Laboratory of Geostatistical Supercomputing (ALGES), Advanced Mining Technology Center (AMTC), Universidad de Chile, Santiago, Chile

[2] Department of Metallurgical and Mining Engineering, Universidad Católica del Norte, Antofagasta, Chile

[3] Department of Applied Physics and Naval Technology, Universidad Politécnica de Cartagena, Murcia, Spain

* **Correspondence:** Email: alvaro.egana@amtc.uchile.cl; Tel: +56229780724.

**Abstract:** This paper introduces a novel data-driven methodology for estimating the full random field (random function) of a regionalized variable, building upon and extending the concepts of ensemble spatial interpolation and adaptive ensemble spatial analysis. The core contribution is the development of a rigorous theoretical framework that proves how the method implicitly learns the underlying spatial dependence structure. By inferring local predictive distributions through adaptively defined, overlapping spatial partitions, the approach ensures that spatial coherence is governed by a learned statistical copula. Formal proof is provided demonstrating that this dependency structure converges to the actual structure of the regionalized variable. Therefore, the resulting spatial distributional estimator is consistent, captures all the uncertainty of the random field, and naturally allows for the generation of multiple equally probable realizations. This non-parametric strategy offers a flexible alternative, inherently preserving spatial patterns and capturing fine-scale variability without requiring prior model specification. Experiments conducted in a detailed case study applied to geostatistical simulation, using both synthetic and real datasets, confirm the effectiveness and computational efficiency of the method, demonstrating its ability to recover local statistics and spatial structure with greater robustness compared to conventional techniques.

**Keywords:** geostatistics; computational geostatistics; generative geostatistics; non-linear geostatistics; distributional geostatistics; geostatistical simulation; empirical copula; data-driven methods

**Mathematics Subject Classification:** 60G60, 62M30, 62M40, 86A32

## 1. Introduction

Accurate spatial prediction and robust uncertainty quantification are paramount when modelling regionalized variables, which inherently exhibit spatial dependence and variation. This uncertainty arises from factors such as limited data sampling, measurement inaccuracies, and the intrinsic variability of the underlying phenomena. In critical fields like mining, environmental sciences, and oilfield characterization, where operational and strategic decisions often hinge on spatial estimates, a comprehensive understanding of this uncertainty is indispensable.

To meet this critical need, a diverse array of geostatistical estimation and simulation algorithms has been developed over the years. Widely adopted methods, including turning-bands simulation [1, 2] and sequential Gaussian simulation [3–5], are designed to generate multiple, equally probable realizations of a spatial variable. Each realization represents a plausible configuration of reality, meticulously respecting the statistical and spatial characteristics inferred from observed data. The analysis of variability across these multiple realizations provides a rigorous assessment of the inherent uncertainty within spatial models.

Despite their established utility, conventional geostatistical estimation and simulation methods often present several practical challenges. A significant hurdle is the demand for exhaustive parameter estimation, such as variogram modelling, which frequently necessitates specialized knowledge and can be both subjective and time-consuming [6, 7]. Furthermore, regionalized variables in real-world scenarios often display complex spatial structures and heterogeneity that are difficult to capture with precision, potentially compromising the reliability of estimation outcomes [8, 9]. An additional complication arises from the widespread assumption of stationarity in many traditional geostatistical approaches, where statistical properties are presumed constant throughout the domain. In practice, this assumption is often violated, as natural spatial variables frequently exhibit non-stationary behavior. Adapting classical algorithms to effectively manage non-stationarity remains a non-trivial endeavor [10–12]. Other techniques have been designed for complex situations with local changes in the regionalized variable [13–15], where it is necessary to specify parameters, trends or fields, which increases the complexity of modelling and the expertise required of the user. Moreover, some regionalized variables—such as metallurgical recovery rates—are intrinsically non-additive. In these instances, values observed in small-scale samples cannot be meaningfully averaged to infer block-scale values, thereby limiting the direct applicability of conventional geostatistical frameworks [16–18].

Given these inherent challenges, data-driven methodologies have emerged as compelling alternatives to traditional geostatistical modelling. Recent advancements in machine learning have demonstrated substantial promise for modelling spatial variables, especially in complex, high-dimensional [19, 20], and non-linear environments [21–23]. These techniques, while not yet data-driven, tend to require fewer restrictive modelling assumptions, such as normality or stationarity of observations, and have the flexibility necessary to capture complex and non-linear interactions between variables. [24, 25]. Furthermore, they often streamline the modelling process by reducing the number of parameters necessitating manual adjustment.

In this study, we propose a novel data-driven methodology for the estimation of the full spatial random field. This method has evolved from the foundational concepts of *ensemble spatial interpolation* (ESI), as introduced by Egaña et al. [26] and Menafoglio et al. [27], and its subsequent

development into *adaptive ensemble spatial analysis* (ESA) proposed by Egaña et al. [28]. Our approach achieves this by inferring local predictive distributions through adaptively defined and overlapping spatial partitions. The spatial coherence of the estimated field is implicitly preserved through the underlying statistical relationships, which are governed by a learned statistical copula that emerges from shared local information. This offers a powerful, non-parametric, and highly flexible alternative to conventional geostatistical estimation techniques, with minimal modelling effort that is intended for a much broader geoscience audience. Although this type of modelling may be less intuitive from a geological or geoscientific point of view, it allows for greater expressiveness at a statistical level, with the possibility of representing complex and non-linear forms of spatial dependence. As a direct application of this estimation, the method readily allows for the generation of multiple, equally probable realizations, enabling a robust quantification of spatial uncertainty and demonstrating a particular ability to capture fine-scale variability, including the nugget effect, without explicit parameterization.

The paper is structured as follows: Section 2 contains the central message of this work and describes in detail the theoretical results that support the proposed model. Section 3 presents a detailed case study of the model applied to the formulation of an *adhoc* geostatistical simulation method. Section 4 presents the results of the case study. Section 5 analyses these results in detail, exploring their theoretical, methodological, and practical implications, as well as outlining directions for future research. Section 6 concludes the article by highlighting the main theoretical contribution. For completeness, the supplementary material contains the statements of the relevant theorems used in Section 2.

## 2. Spatial distributional estimation

### 2.1. Overview of the proposed model

Before introducing the formal details, we summarize the structure of the proposed spatial distributional estimation (SDE) model. Namely, at each discretized location $\mathbf{x}_i \in \mathcal{X}$, we estimate a local empirical distribution $\widehat{F}_i^{(T)}$ from a neighborhood-informed ensemble of conditional realizations. These neighborhoods are adaptively defined through a data-driven partitioning mechanism that preserves local spatial information while ensuring global coverage. Crucially, instead of fitting a single joint model across all locations, the method constructs a local marginal distribution at each site and implicitly infers the joint structure via the dependencies induced by overlapping neighborhoods. This leads to a collection of measurable random elements $\{\widehat{F}_i^{(T)}\}_{i=1}^N$, whose dependence structure is governed by an estimated copula $\widehat{C}_T$. The resulting model is entirely non-parametric, decentralized, and spatially coherent by design. The following section provides the formal foundation of this construction, with particular attention to the measurability, dependence, and convergence properties of the estimated distributions.

### 2.2. Ensemble spatial interpolation

The ESI method provides a flexible, data-driven framework for spatial prediction. It is designed to estimate the posterior predictive distribution of a spatial variable at unsampled locations, without requiring explicit assumptions of stationarity or manual modelling of spatial continuity [26]. ESI is

particularly suitable for complex and heterogeneous spatial domains where classical geostatistical techniques may have difficulties due to the need for variographic modelling or domain-specific parameter adjustment. For an overview of ESI, let $\mathcal{S} \subset \mathbb{R}^d$ be the spatial domain, and let $Z : \mathcal{S} \to \mathbb{R}$ be the spatial variable of interest. Given a finite set of observations $\mathcal{O} = \{(\mathbf{x}_i, z_i)\}_{i=1}^n$, with $\mathbf{x}_i \in \mathcal{S}$ and $z_i = Z(\mathbf{x}_i)$, the objective is to infer the predictive distribution $p(Z(\mathbf{x}) \mid \mathcal{O})$ for all $\mathbf{x} \in \mathcal{S}$. Thus, the ESI method is based on a hierarchical generative model:

$$p(Z, \mathcal{P}) = p(Z \mid \mathcal{P}) \cdot p(\mathcal{P}), \tag{2.1}$$

where $\mathcal{P} \in \Pi(\mathcal{S})$ denotes a random partition (or tessellation) of the spatial domain. The partition $\mathcal{P} = \{\mathcal{P}_k\}_{k=1}^K$ divides $\mathcal{S}$ into disjoint regions, and within each region $\mathcal{P}_k$, a local interpolation model is used to estimate the value of $Z(\mathbf{x})$ at locations $\mathbf{x} \in \mathcal{P}_k$. The predictive distribution at a query point $\mathbf{x}$ is obtained by marginalizing over the ensemble of partitions in such a way that:

$$p(Z(\mathbf{x})) = \mathbb{E}_{\mathcal{P}} \left[ p(Z(\mathbf{x}) \mid \mathbf{x} \in \mathcal{P}_k) \right]. \tag{2.2}$$

The term *ensemble* comes from the fact that the expression (2.2) encodes a bagging-type process in ensemble learning schemes [29]. Figure 1 shows the basic components of the ESI method, an example of tessellations and the distribution sampling applied after the ESI framework (see Section 2.3.1).

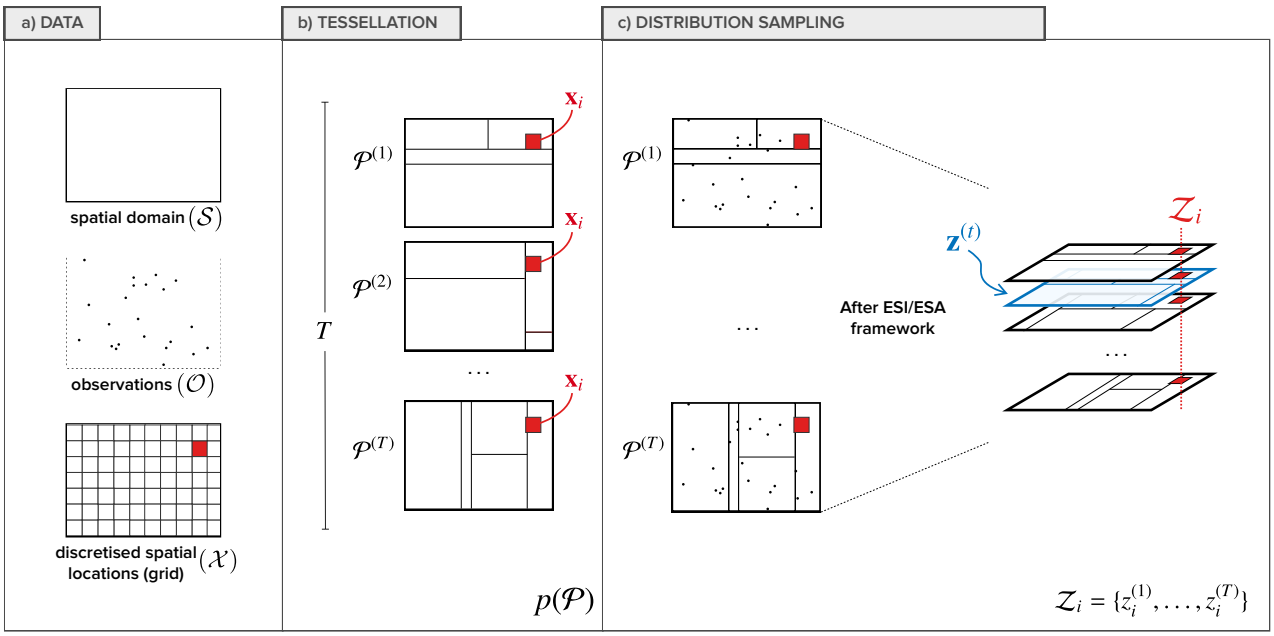


**Figure 1.** SDE model flow: Tessellation generation and distribution sampling. a) Basic components to apply the ESI/ESA framework. b) An example of a random partition over the spatial domain. c) The distribution sampling after the ESI/ESA framework.

### 2.2.1. Adaptive ensemble spatial analysis

In general, within each cell $\mathcal{P}_k$, prediction can be performed using any interpolation function. The basic idea of this function is that it should not be too complex because it is intensively used. For this,

the first versions proposed were a simple *inverse distance weighting* (IDW) or a kind of radial basis function version of kriging (where the covariance model is fixed for all cells, so as to avoid variogram analysis), both with quite good results [26]. Later, Egaña et al. [28] proposed a variation of ESI, which they called *adaptive ensemble spatial analysis* (ESA), using a type of local interpolator that is a locally adaptive version of the IDW method, which analyses (optimizes) the IDW parameters according to the data present in each cell of each partition. Namely, let

$$\mathcal{N}_{\mathcal{P}_k}(\mathbf{x}) = \{(\mathbf{x}_j, z_j) \mid \mathbf{x}_j \in \mathcal{P}_k\} \subset \mathcal{O}, \tag{2.3}$$

be the set of neighboring observations of (i.e., in the same partition cell $\mathcal{P}_k$ as) $\mathbf{x}$. Then, for a fixed exponent $p$, in the original IDW version, the interpolated value for all elements in $\mathcal{P}_k$ is given by:

$$\hat{Z}(\mathbf{x}) = \frac{\sum_j w_j z_j}{\sum_j w_j}, \quad \text{where} \quad w_j = \frac{1}{d(\mathbf{x}_j, \mathbf{x})^p} \quad \text{and} \quad (\mathbf{x}_j, z_j) \in \mathcal{N}_{\mathcal{P}_k}(\mathbf{x}). \tag{2.4}$$

In its adaptive version, the distance metric $d(\cdot, \cdot)$ incorporates local anisotropy and orientation through a linear transformation:

$$d(\mathbf{x}_j, \mathbf{x}) = \left\| \mathbf{A}_\theta (\mathbf{x}_j - \mathbf{x}) \right\|_2, \tag{2.5}$$

where $\mathbf{A}_\theta$ is a scaling-rotation matrix parameterized by $\theta = (a_f, \phi)$, representing an anisotropy factor and an azimuthal orientation, respectively. The parameters $p, a_f, \phi$ are optimized locally in each cell by minimizing a cross-validated error, such as the mean absolute error (MAE) under leave-one-out (LOO) validation:

$$(p_k, a_{f,k}, \phi_k)^* = \arg\min_{p, a_f, \phi} \sum_i \left| z_i - \hat{Z}_i^{\text{LOO}} \right|. \tag{2.6}$$

This new adaptive version allows for better capture of the local properties of the regionalized variable, requiring fewer partitions to find convergence in the results [28]. It is also interesting to note that the overlap that occurs between cells for each position analyzed allows spatial dependencies to be maintained. This makes it possible to implicitly capture anisotropy (even in very complex situations) and other spatial characteristics, without the need for explicit parameterization.

### 2.2.2. ESA posterior samples

In practice, in order to implement the model for the predictive distribution of the expression (2.2), multiple partitions $\{\mathcal{P}^{(1)}, \ldots, \mathcal{P}^{(T)}\}$ are sampled from the partition distribution $p(\mathcal{P})$, typically generated using stochastic tessellation processes such as the Mondrian process [26] or the Voronoi process [27]. Each partition induces a complete interpolation map over $\mathcal{S}$. The final output is a sample from the predictive distribution at each location:

$$\{z^{(1)}(\mathbf{x}), \ldots, z^{(T)}(\mathbf{x})\}, \quad \forall\, \mathbf{x} \in \mathcal{S}. \tag{2.7}$$

These posterior samples are called *ESA samples* and are the final stage of the ESA methodology flow, whose estimate is given by

$$z^*(\mathbf{x}) = \mathcal{T}(z^{(1)}(\mathbf{x}), \ldots, z^{(T)}(\mathbf{x})), \quad \forall\, \mathbf{x} \in \mathcal{S}, \tag{2.8}$$

where $\mathcal{T}$ is a statistic that minimizes some loss function (usually $\mathcal{T}$ is the mean and the loss function is based on the mean squared error or MSE). However, what matters to us here is that the ESA samples provide the basis for the random field estimation described in the next section.

### *2.3. Random field estimation*

As mentioned above, the SDE model leverages local predictive distribution samples obtained from ESA. Therefore, instead of aggregating ESA samples to form a single interpolated estimate, the SDE model uses them to estimate the entire random field.

#### 2.3.1. Local distributional estimation

For a set $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \subset \mathcal{S}$ of discretized spatial locations (e.g., a regular grid), the ESA method provides, for each location $\mathbf{x}_i$, a set of predictive samples based on neighborhood information gathered from multiple ensemble partitions.

Let us define the local neighborhood of $\mathbf{x}_i$ as:

$$\mathcal{N}(\mathbf{x}_i) = \bigcup_{t=1}^{T} \bigcup_{k=1}^{K(t)} \mathcal{N}_{\mathcal{P}_k^{(t)}}(\mathbf{x}_i) \subset \mathcal{O}, \tag{2.9}$$

where $\mathcal{N}_{\mathcal{P}_k^{(t)}}(\mathbf{x}_i)$ denotes the set of observed locations contributing to the prediction at $\mathbf{x}_i$ under the $k$-th cell of the $t$-th tessellation (partition). This union accounts for the variability introduced by the ensemble strategy and captures a rich and diverse local context for the estimation process.

Using the aggregated information from these neighborhoods, we define:

$$\mathcal{Z}_i = \{z_i^{(1)}, \ldots, z_i^{(T)}\} \subset \mathbb{R}, \tag{2.10}$$

as the set of ESA samples at location $\mathbf{x}_i$ –Figure 1 c) shows the relation between $\mathcal{Z}_i$ and $\mathbf{z}^{(t)}$ for distribution sampling after the application of the ESI/ESA framework. With these samples, the most basic model we can build at $\mathbf{x}_i$ is the empirical distribution:

$$\widehat{F}_i^{(T)}(z) := \frac{1}{T} \sum_{t=1}^{T} \mathbb{I}\{z_i^{(t)} \leq z\}, \tag{2.11}$$

where $\mathbb{I}\{\cdot\}$ is the indicator function. In this way, the estimated random field, conditionally defined with respect to the neighborhood system $\{\mathcal{N}(\mathbf{x}_i)\}_{i=1}^N$, is constructed completely locally (position by position) as follows:

$$\tilde{\mathbf{Z}}^{(T)} \mid \{\mathcal{N}(\mathbf{x}_i)\}_{i=1}^N = (\tilde{Z}_1^{(T)} \mid \mathcal{N}(\mathbf{x}_1), \ldots, \tilde{Z}_N^{(T)} \mid \mathcal{N}(\mathbf{x}_N)), \quad \text{where} \tag{2.12}$$

$$\tilde{Z}_i^{(T)} \mid \mathcal{N}(\mathbf{x}_i) \sim \widehat{F}_i^{(T)}(z) \quad \forall\, i \in \{1, \cdots, N\}. \tag{2.13}$$

The notation $\tilde{\mathbf{Z}}^{(T)} \mid \{\mathcal{N}(\mathbf{x}_i)\}_{i=1}^N$ emphasizes that each local distribution $\widehat{F}_i^{(T)}$ is constructed from the predictive ESA samples $\mathcal{Z}_i$ derived from the specific neighborhood structure $\mathcal{N}(\mathbf{x}_i)$ around each location.

2.3.2. Copula-based joint distribution and spatial structure

In this section, we aim to formalize how the spatial dependency structure inherent in the estimated random field can be recovered when estimating local empirical distributions pointwise using the SDE model. Although the field is defined through local distributions, the overlapping structure of the neighborhoods ensures that dependencies between points are preserved and summarized by a unique joint copula, yielding a coherent spatial model.

To this end, let us consider the following proposition. Its proof is based on the fact that each local empirical distribution is measurable with respect to the information contained in its neighborhood. Overlapping neighborhoods induce a direct dependency between the corresponding distributions. In the case of non-overlapping neighborhoods, the connectivity of adaptive partition cells guarantees a chain of overlapping neighborhoods connecting any two points. This spatial connectivity ensures mutual dependence between all local distributions. According to Sklar's theorem [30], this implies the existence of a unique copula that describes their joint distribution, capturing the underlying spatial dependency structure.

**Proposition 1.** *Under the SDE framework, the joint distribution of the estimated random field* $(\tilde{\mathbf{Z}}^{(T)} \mid \{\mathcal{N}(\boldsymbol{x}_i)\}_{i=1}^N)$ *is governed by an implicitly learned copula* $\widehat{C}_T$*, which emerges from the overlapping structure of local neighborhoods* $\mathcal{N}(\mathbf{x}_i)$ *used to learn* $\widehat{F}_i^{(T)}$ $\forall\, \mathbf{x}_i \in \mathcal{X}$.

*Proof.* Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $Z : \mathcal{S} \to \mathbb{R}$ be a real-valued spatial random field defined over a domain $\mathcal{S} \subset \mathbb{R}^d$. Consider a finite set of discretized spatial locations $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \subset \mathcal{S}$. For each location $\mathbf{x}_i \in \mathcal{X}$, the SDE method produces a local sample set $\mathcal{Z}_i = \{z_i^{(1)}, \ldots, z_i^{(T)}\}$, from which the empirical distribution function is defined as

$$\widehat{F}_i^{(T)}(z) := \frac{1}{T} \sum_{t=1}^{T} \mathbb{I}\{z_i^{(t)} \le z\}. \tag{2.14}$$

Each empirical distribution $\widehat{F}_i^{(T)}$ is constructed based on conditioning data supported on the local neighborhood $\mathcal{N}(\mathbf{x}_i)$. Let $\mathcal{F}_{\mathcal{N}(\mathbf{x}_i)} \subseteq \mathcal{F}$ denote the sub-$\sigma$-algebra generated by the random variables $\{Z(\mathbf{s}) : \mathbf{s} \in \mathcal{N}(\mathbf{x}_i)\}$. By construction, each empirical distribution

$$\widehat{F}_i^{(T)} : (\Omega, \mathcal{F}_{\mathcal{N}(\mathbf{x}_i)}) \to (\mathcal{D}, \mathcal{B}), \tag{2.15}$$

is measurable with respect to its local sub-$\sigma$-algebra, where $\mathcal{D}$ is the space of cumulative distribution functions on $\mathbb{R}$ and $\mathcal{B}$ its Borel $\sigma$-algebra.

Statistical dependence between $\widehat{F}_i^{(T)}$ and $\widehat{F}_j^{(T)}$ arises naturally from the spatial configuration of their respective neighborhoods. If the neighborhoods overlap,

$$\mathcal{N}(\mathbf{x}_i) \cap \mathcal{N}(\mathbf{x}_j) \neq \emptyset, \tag{2.16}$$

then the corresponding sub-$\sigma$-algebras $\mathcal{F}_{\mathcal{N}(\mathbf{x}_i)}$ and $\mathcal{F}_{\mathcal{N}(\mathbf{x}_j)}$ share common information, which induces dependence between $\widehat{F}_i^{(T)}$ and $\widehat{F}_j^{(T)}$ – Figure 2. If the neighborhoods do not overlap,

$$\mathcal{N}(\mathbf{x}_i) \cap \mathcal{N}(\mathbf{x}_j) = \emptyset, \tag{2.17}$$

dependence may still be present due to the spatial connectivity of the ensemble partitions. Since each $\mathcal{N}(\mathbf{x}_i)$ is contained by a union of overlapping partition cells, the collection of partition cells containing $\{\mathcal{N}(\mathbf{x}_i)\}_{i=1}^N$ constitutes a connected covering of the domain $\mathcal{S}$. More precisely, as shown in Figure 3, for any pair $\mathbf{x}_i, \mathbf{x}_j \in \mathcal{X}$, there exists a finite sequence

$$\{\mathbf{x}_{k_0}, \mathbf{x}_{k_1}, \ldots, \mathbf{x}_{k_p}\} \subset \mathcal{X}, \tag{2.18}$$

with $\mathbf{x}_{k_0} = \mathbf{x}_i$ and $\mathbf{x}_{k_p} = \mathbf{x}_j$, such that

$$\mathcal{N}(\mathbf{x}_{k_l}) \cap \mathcal{N}(\mathbf{x}_{k_{l+1}}) \neq \emptyset, \quad \forall l = 0, \ldots, p-1. \tag{2.19}$$

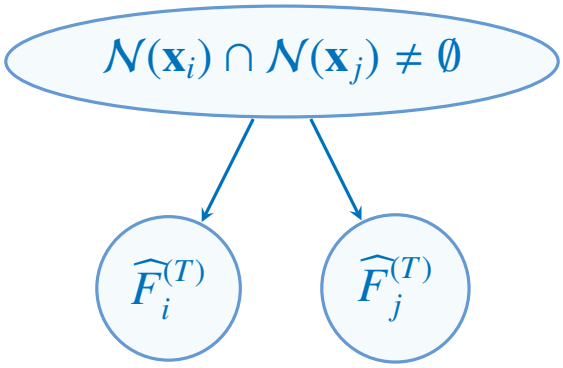


**Figure 2.** Dependence between $\widehat{F}_i^{(T)}$ and $\widehat{F}_j^{(T)}$ through having a common set of observations that generated them.

To verify the previous statement, suppose for the sake of contradiction that there exist two neighborhoods $\mathcal{N}(\mathbf{x}_i)$ and $\mathcal{N}(\mathbf{x}_j)$ such that:

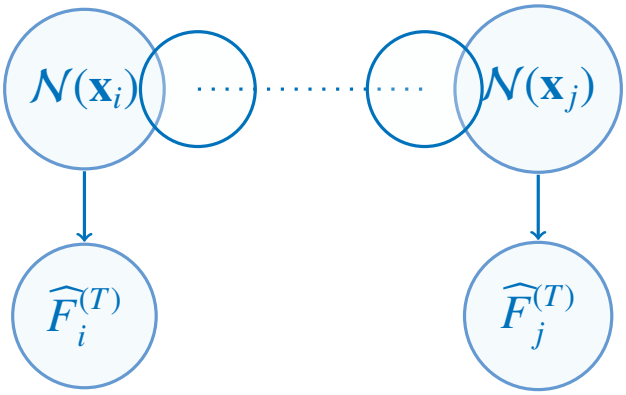


**Figure 3.** There is a path of dependencies (via intersections of the extended sets of observations) that make $\widehat{F}_i^{(T)}$ and $\widehat{F}_j^{(T)}$ dependent on each other.

(1) Their neighborhoods are disjoint: $\mathcal{N}(\mathbf{x}_i) \cap \mathcal{N}(\mathbf{x}_j) = \emptyset$.

(2) There exists no finite chain of overlapping neighborhoods connecting $\mathcal{N}(\mathbf{x}_i)$ to $\mathcal{N}(\mathbf{x}_j)$.

Since each neighborhood is defined as the union of the observations contained in partition cells of the ensemble tessellations, this assumption implies that the sets of observations associated with $\mathcal{N}(\mathbf{x}_i)$

and $\mathcal{N}(\mathbf{x}_j)$ belong to two disconnected families of partition cells. By construction, the partition cells cover the entire spatial domain in a connected manner. Specifically, the collection of partition cells cannot be split into two disjoint subsets with no adjacency or overlap. This ensures that any two sets of observations contained in the cells must be connected through a chain of overlapping or adjacent cells. Therefore, the assumption that $\mathcal{N}(\mathbf{x}_i)$ and $\mathcal{N}(\mathbf{x}_j)$ correspond to disconnected families of partition cells and their observations contradict the fundamental property of any partitioning scheme.

In order to formalize this connectivity scheme, we construct a discrete neighborhood graph $G = (V, E)$, whose vertex set $V$ and corresponds to the locations $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$. An edge $(\mathbf{x}_i, \mathbf{x}_j) \in E$ exists if and only if their respective neighborhoods overlap, i.e., $\mathcal{N}(\mathbf{x}_i) \cap \mathcal{N}(\mathbf{x}_j) \neq \emptyset$. The chain of overlapping neighborhoods that connects any two points, $\mathbf{x}_i$ and $\mathbf{x}_j$, corresponds exactly to the existence of a path in the graph $G$ between vertices $\mathbf{x}_i$ and $\mathbf{x}_j$. The spatial connectivity of the domain, guaranteed by the partitioning scheme, implies that this graph $G$ is fully connected (directly or indirectly via a path) by construction. To verify how this affects mutual dependencies in our random field, we must refer to some well-known concepts and results from *probabilistic graphical models* theory [31]. In this regard, we consider the following lemma.

**Lemma:** Let the graph $G = (V, E)$ be defined with $V = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ and $(\mathbf{x}_i, \mathbf{x}_j) \in E$ if $\mathcal{N}(\mathbf{x}_i) \cap \mathcal{N}(\mathbf{x}_j) \neq \emptyset$. Let $U_i := \widehat{F}_i^{(T)}(z_i)$ for $i = 1, \ldots, N$ be the vector of transformed marginal variables. Assuming that the joint distribution of $U = (U_1, \ldots, U_N)$ satisfies the *global Markov property* with respect to $G$ and is also *faithful* to $G$, then for any $i \neq j$, there is a path in $G$ that connects $\mathbf{x}_i$ and $\mathbf{x}_j$ if and only if $U_i$ and $U_j$ are not conditionally independent.

*Proof.* The proof general idea is that a path in $G$ between $\mathbf{x}_i$ and $\mathbf{x}_j$ means that there is no set $S \subset V \setminus \{\mathbf{x}_i, \mathbf{x}_j\}$ that graphically separates $\mathbf{x}_i$ from $\mathbf{x}_j$. By the global Markov property, any graphical separation corresponds to a statistical independence in the distribution. The faithfulness assumption provides the converse argument, i.e., any statistical independence must correspond to a graphical separation. Since there is no graphical separation in this case, there cannot be statistical independence between $U_i$ and $U_j$. Formally:

( $\Longrightarrow$ ): To prove that if there is a path in $G$ that connects $\mathbf{x}_i$ and $\mathbf{x}_j$, then $U_i$ and $U_j$ are not conditionally independent, we will reason by contradiction. The premise of our method implies that the graph $G$ is fully connected, thus we can always assume that there is a path between $\mathbf{x}_i$ and $\mathbf{x}_j$ in $G$. This means that there is no set $S \subset V \setminus \{\mathbf{x}_i, \mathbf{x}_j\}$ that graphically separates them. Suppose, for the sake of contradiction, that $U_i$ and $U_j$ are conditionally independent given some set of variables $U_{S'}$, where $S' \subset V \setminus \{\mathbf{x}_i, \mathbf{x}_j\}$. By the faithfulness assumption, this statistical independence $(U_i \perp U_j \mid U_{S'})$ must imply a corresponding graphical separation in $G$. Thus, the set of vertices $S'$ must graphically separate $\mathbf{x}_i$ and $\mathbf{x}_j$ ($\Rightarrow|\Leftarrow$).

( $\Longleftarrow$ ): On the other hand, to prove that if $U_i$ and $U_j$ are not conditionally independent, then there is a path in $G$ that connects $\mathbf{x}_i$ and $\mathbf{x}_j$, we will reason by contraposition. The contrapositive of this statement is that if there is no path in $G$ that connects $\mathbf{x}_i$ and $\mathbf{x}_j$, then $U_i$ and $U_j$ are conditionally independent. Thus, if there is no path between $\mathbf{x}_i$ and $\mathbf{x}_j$ in $G$, they are neither directly nor indirectly connected. This implies that there is a set $S \subset V \setminus \{\mathbf{x}_i, \mathbf{x}_j\}$ that graphically separates them, i.e., $\mathbf{x}_i \perp_G \mathbf{x}_j \mid S$. By the global Markov property, a graphical separation implies a statistical conditional independence. Thus, $\mathbf{x}_i \perp_G \mathbf{x}_j \mid S \implies U_i \perp U_j \mid U_S$. This means that if there is no path in the graph $G$, the corresponding variables are conditionally independent. ■

It should be noted that both assumptions (Markov global property and faithfulness) are valid by

construction in the SDE scheme.

*Remark on global Markov property:* The SDE construction is intrinsically local, i.e., each empirical distribution $\widehat{F}_i^{(T)}$ is generated solely from observations within $\mathcal{N}(\mathbf{x}_i)$. This means that the random element $U_i := \widehat{F}_i^{(T)}(Z_i)$ is measurable with respect to the sub-$\sigma$-algebra $\mathcal{F}_{\mathcal{N}(\mathbf{x}_i)}$. This locality implies conditional independence in the sense that if two neighborhoods are disjoint and separated by others, the information in one cannot influence the other except through the overlapping chain. This is exactly what formalizes the global Markov property, which states that conditional independence corresponds to graphical separation in $G$. It is worth to mention that it is the same hypothesis as in classical Markov random fields (MRFs) on a network, where the random field satisfies Hammersley-Clifford's theorem [32] which states that the joint probability factors according to the neighborhood graph if it is strictly positive [32, 33]. Thus, in our construction $U$ is naturally Markovian with respect to $G$ (for further details, see the supplementary material).

*Remark on faithfulness:* The phenomenon of *unfaithfulness* corresponds to "accidental cancellations" that could produce independence along paths and is well known in probabilistic graphical model theory. This generally occurs in parameter configurations that produce unfaithful topologies in dependency graphs, such as $G$. The independence constraints that would produce an unfaithful probability correspond to non-trivial algebraic equalities between the model parameters (e.g., linear constraints such as $c + ab = 0$ in simple illustrative cases). Unless these equalities are maintained identically (a degenerate structural constraint), their solution sets have zero Lebesgue measure [31, 34, 35]. Within the framework of the SDE, this may be associated with three phenomena: (i) the set of overlapping neighborhoods, (ii) the local parameters associated with each cell in the local interpolator, and (iii) certain rare pathological configurations of the observations. Cases (i) and (ii) do not pose a problem, since it is assumed, by construction, that the SDE partition generator produces configurations with a non-degenerate probability (i.e., without atoms, or at least absolutely continuous) and that the local parameters depend continuously on the geometry of the partition. This makes the probability of sampling an exactly unfaithful configuration zero. Case (iii) is more of a practical nature and could occur if some observations, with highly punctual support, represent a pure *nugget effect*; such cases, given their punctual support, are also zero Lebesgue measure sets and therefore have zero probability of occurrence.

Thus, this fully connected graph $G$, far from being a mere combinatorial abstraction, accurately and faithfully reflects the information connectivity induced by the SDE overlapping neighborhood system. The global Markov property and faithfulness ensure that the fully statistical dependence structure of the local distributions $\{\widehat{F}_i^{(T)}\}_{i=1}^N$ is captured by the topology of $G$. Consequently, the guaranteed connectivity of the SDE method implies that the family $\{\widehat{F}_i^{(T)}\}_{i=1}^N$ forms a collection of mutually dependent distributional elements, as there is always a path of information connecting them.

This intrinsic dependence, which is not assumed but rather emerges from the overlapping neighborhood structure, and assuming each $\widehat{F}_i^{(T)}$ is continuous, which holds almost surely for sufficiently large $T$ as a consequence of the Glivenko–Cantelli theorem [36, 37], are the necessary conditions for applying Sklar's theorem [30]. The theorem guarantees the existence of a unique copula

$$\widehat{C}_T : [0,1]^N \to [0,1], \tag{2.20}$$

such that, for all $(u_1, \ldots, u_N) \in [0,1]^N$,

$$\mathbb{P}(\widehat{F}_1^{(T)}(Z_1) \le u_1, \ldots, \widehat{F}_N^{(T)}(Z_N) \le u_N) = \widehat{C}_T(u_1, \ldots, u_N). \tag{2.21}$$

This copula fully captures the joint dependence structure induced by the overlapping neighborhood system. □

2.3.3. Convergence of the estimated copula

Let $C$ denote the true copula associated with the joint distribution $H$ of $Z(\mathbf{x}_1), \ldots, Z(\mathbf{x}_N)$, and let $\widehat{C}_T$ denote the copula implicitly defined by the SDE framework using $T$ ensemble samples.

We are interested in the convergence

$$\widehat{C}_T \xrightarrow[T\to\infty]{\mathcal{L}} C, \tag{2.22}$$

understood as weak convergence of distribution functions or uniform convergence on compact subsets of $[0,1]^N$.

We verify this convergence by establishing the following conditions.

**(C1) Consistent marginal estimation.** For each $\mathbf{x}_i \in \mathcal{X}$, the empirical distribution

$$\widehat{F}_i^{(T)}(z) := \frac{1}{T}\sum_{t=1}^{T} \mathbb{I}\{z_i^{(t)} \le z\} \tag{2.23}$$

converges almost surely to the conditional marginal distribution

$$F_i(z) := \mathbb{P}(Z(\mathbf{x}_i) \le z \mid \mathcal{F}_{\mathcal{N}(\mathbf{x}_i)}). \tag{2.24}$$

*Proof.* Each sample $z_i^{(t)}$ is drawn via ESA using information from an adaptive neighborhood $\mathcal{N}_{\mathcal{P}_k^{(t)}}(\mathbf{x}_i)$. Under the assumption that the underlying tessellation process is stationary and ergodic, the strong law of large numbers (SLLN) for conditionally i.i.d. sequences ensures that [38]

$$\widehat{F}_i^{(T)}(z) \xrightarrow[T\to\infty]{a.s.} \mathbb{E}\left[\mathbb{I}\{Z(\mathbf{x}_i) \le z\} \mid \mathcal{F}_{\mathcal{N}(\mathbf{x}_i)}\right] = F_i(z). \tag{2.25}$$

■

*Remark 1.* The assumption of stationarity in the tessellation process ensures that the statistical properties of the partitioning mechanism do not change across ensemble iterations. In particular, it guarantees that the neighborhoods $\mathcal{N}_{\mathcal{P}_k^{(t)}}(\mathbf{x}_i)$ used to generate each sample $z_i^{(t)}$ are drawn from a distribution that is invariant with respect to $t$. Without this invariance, the empirical average would reflect a mixture of non-homogeneous sampling schemes, thereby violating the conditions required for almost sure convergence to a well-defined conditional distribution.

*Remark 2.* In this context, ergodicity implies that averaging over $T$ ensemble realizations yields convergence to conditional expectations with respect to the neighborhood $\mathcal{N}(\mathbf{x}_i)$. That is, for any measurable function $g$,

$$\frac{1}{T}\sum_{t=1}^{T} g(z_i^{(t)}) \xrightarrow[T\to\infty]{a.s.} \mathbb{E}[g(Z(\mathbf{x}_i)) \mid \mathcal{F}_{\mathcal{N}(\mathbf{x}_i)}]. \tag{2.26}$$

**(C2) Stability of the dependence structure.** For any pair $i \neq j \in \{1, \cdots, N\}$, the empirical copula

$$\widehat{C}_{ij}^{(T)}(u, v) := \frac{1}{T} \sum_{t=1}^{T} \mathbb{I}\left\{\widehat{F}_i^{(T)}(z_i^{(t)}) \leq u,\ \widehat{F}_j^{(T)}(z_j^{(t)}) \leq v\right\}. \tag{2.27}$$

converges in probability to the true pairwise copula $C_{ij}(u, v)$ of the pair $(Z(\mathbf{x}_i), Z(\mathbf{x}_j))$.

*Proof.* Assuming that $\mathbf{z}^{(t)} = (z_1^{(t)}, \ldots, z_N^{(t)})$ denote the $t$-th ensemble sample drawn from the SDE framework (Figure 1), the pairs $(z_i^{(t)}, z_j^{(t)})$ are spatially dependent due to overlapping neighborhoods and data reuse as proved in Proposition 1. Although the Glivenko–Cantelli theorem is classically stated for i.i.d. sequences, convergence results also hold under weak dependence assumptions [36, 37]. Specifically, suppose that the sequence of ensemble samples satisfies an $\alpha$-mixing condition, that is, there exists a sequence of $\sigma$-algebras $(\mathcal{A}_t)$, adapted to the ensemble sampling process, such that the strong mixing coefficients

$$\alpha(n) := \sup_t \{ \sup_{A \in \mathcal{A}_t,\, B \in \mathcal{A}_{t+n}} \{|\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)|\}\} \tag{2.28}$$

satisfy

$$\sum_{n=1}^{\infty} \alpha(n)^{\delta/(2+\delta)} < \infty \quad \text{for some } \delta > 0. \tag{2.29}$$

Under this assumption, the empirical process associated with the transformed pairs $(\widehat{F}_i^{(T)}(z_i^{(t)}), \widehat{F}_j^{(T)}(z_j^{(t)}))$ satisfies a uniform law of large numbers. Therefore,

$$\widehat{C}_{ij}^{(T)}(u, v) \xrightarrow[T \to \infty]{\mathbb{P}} C_{ij}(u, v) \tag{2.30}$$

holds uniformly on $[0, 1]^2$. This convergence implies that the estimated pairwise copula recovers the true spatial dependence structure as $T \to \infty$. ■

*Remark 3.* In practice, in our context the $\alpha$-mixing condition reflects the idea that ensemble samples $z_i^{(t)}$ and $z_i^{(t+n)}$ generated for the location $\mathbf{x}_i$ at distant indices $t$ and $t + n$ become increasingly decorrelated as $n \to \infty$. For the SDE framework, this means that the dependence between distinct ensemble realizations diminishes when the tessellation process introduces sufficient variability across ensemble iterations. Such variability ensures that the aggregated statistics do not remain locked into a single spatial configuration, enabling convergence results akin to those of independent sampling.

**(C3) Compactness and equicontinuity of the copula process.** The sequence of empirical copulas $\{\widehat{C}_T\}_{T \in \mathbb{N}}$ is relatively compact and equicontinuous in $C([0, 1]^N)$, the space of continuous functions on the unit hypercube with the uniform norm.

*Proof.* Let $\mathbf{z}^{(t)} = (z_1^{(t)}, \ldots, z_N^{(t)})$ denote again the $t$-th ensemble sample drawn from the SDE framework (Figure 1). Each marginal observation $z_i^{(t)}$ is transformed via its corresponding empirical distribution into a pseudo-observation $u_i^{(t)} := \widehat{F}_i^{(T)}(z_i^{(t)}) \in [0, 1]$. The empirical copula is then defined by:

$$\widehat{C}_T(\mathbf{u}) := \frac{1}{T} \sum_{t=1}^{T} \mathbb{I}\left\{u_1^{(t)} \leq u_1, \ldots, u_N^{(t)} \leq u_N\right\} = \frac{1}{T} \sum_{t=1}^{T} \prod_{i=1}^{N} \mathbb{I}\left\{u_i^{(t)} \leq u_i\right\}. \tag{2.31}$$

which corresponds to the empirical joint distribution function of the transformed data.

This construction satisfies several regularity properties:

(1) The map $z_i^{(t)} \mapsto u_i^{(t)}$ is non-decreasing and Lipschitz continuous (with constant 1) if $\widehat{F}_i^{(T)}$ is continuous, which holds almost surely when $T$ is large.

(2) The indicator functions $\mathbb{I}\{u_i^{(t)} \leq u_i\}$ are bounded between 0 and 1, so $\widehat{C}_T(\mathbf{u}) \in [0,1]$ for all $\mathbf{u} \in [0,1]^N$, and thus the sequence $\{\widehat{C}_T\}_{T\in\mathbb{N}}$ is uniformly bounded.

(3) To show equicontinuity, note that small perturbations in the evaluation point $\mathbf{u}$ modify only a few terms in the empirical sum. More precisely, for any $\epsilon > 0$, there exists $\delta > 0$ such that

$$\|\mathbf{u}_1 - \mathbf{u}_2\|_\infty < \delta \quad \Rightarrow \quad |\widehat{C}_T(\mathbf{u}_1) - \widehat{C}_T(\mathbf{u}_2)| < \epsilon, \tag{2.32}$$

for all $T$. This follows from the fact that the transformed pseudo-observations $\mathbf{u}^{(t)}$ are smoothly spread over the unit hypercube due to the continuity of the marginal distributions $\widehat{F}_i^{(T)}$, and that the indicator function is almost everywhere continuous in $\mathbf{u}$.

These properties ensure that the sequence $\{\widehat{C}_T\}_{T\in\mathbb{N}}$ is both uniformly bounded and equicontinuous. Therefore, by the Arzelà–Ascoli theorem [37], it is relatively compact in $C([0,1]^N)$. This compactness result is essential, as it guarantees the existence of uniformly convergent subsequences, and, combined with the convergence of the marginals and stability of the pairwise dependence structure, leads to the convergence of the empirical copula $\widehat{C}_T$ to the true copula $C$.

■

*Remark 4.* The existence of such $\delta$ follows from the fact that the pseudo-observations $\{\mathbf{u}^{(t)}\}_{t=1}^T \subset [0,1]^N$, being obtained through continuous transformations of the marginal data, are asymptotically dense in the unit hypercube. Consequently, for any $\epsilon > 0$, one can always find a sufficiently small $\delta > 0$ such that the number of pseudo-observations affected by a perturbation of size $\delta$ in the evaluation point $\mathbf{u}$ remains uniformly small relative to $T$.

2.3.4. Consequence and generative interpretation

Under conditions (C1)–(C3), the convergence

$$\sup_{\mathbf{u}\in[0,1]^N} \left|\widehat{C}_T(\mathbf{u}) - C(\mathbf{u})\right| \xrightarrow[T\to\infty]{\mathbb{P}} 0 \tag{2.33}$$

follows from classical results in empirical process theory, including the uniform consistency of empirical copulas [36] and weak convergence theorems for copula processes [37]. In statistical terms, this means that $\widehat{C}_T$ is a consistent estimator of $C$.

Although the above already provides the certainty we were looking for, this can also be reinforced from another perspective. Since $[0,1]^N$ is a compact metric space and each $\widehat{C}_T$ defines a valid probability measure over it, the sequence $\{\widehat{C}_T\}_{T\in\mathbb{N}}$ is tight. By Prokhorov's theorem [38], this implies that the sequence is relatively compact in the topology of weak convergence. Combined with the uniqueness of the limit, guaranteed by Sklar's theorem under continuous marginals [30], we conclude that the sequence $\widehat{C}_T$ converges in law to the true copula $C$.

*Remark 5.* In compact metric spaces, such as $[0,1]^N$, tightness and relative compactness are equivalent. Therefore, verifying tightness (e.g., via uniform boundedness and equicontinuity of the associated stochastic process) implies that the sequence $\{\widehat{C}_T\}_{T\in\mathbb{N}}$ of empirical copulas is relatively compact. This is a key step in establishing convergence in distribution to the true copula $C$.

**Corollary (Consistency of the joint distribution).** Under conditions (C1)–(C3), the joint distribution of estimated random field $(\tilde{\mathbf{Z}}^{(T)} \mid \{\mathcal{N}(\mathbf{x}_i)\}_{i=1}^N)$ converges in law to the true joint distribution $H$, i.e.,

$$(\tilde{\mathbf{Z}}^{(T)} \mid \{\mathcal{N}(\mathbf{x}_i)\}_{i=1}^N) \xrightarrow[T\to\infty]{\mathcal{L}} \mathbf{Z} \sim H(\mathbf{z}) = C(F_1(z_1), \ldots, F_N(z_N)). \tag{2.34}$$

This confirms that the SDE model asymptotically recovers the full joint distribution of the spatial random field, including both marginals and dependence structure. In this sense, spatial dependence is conveyed smoothly through neighborhood geometry and data reuse, rather than through the direct parameterization of a global model or a family of known parametric copulas.

Finally, from a generative modelling perspective, this mechanism constitutes a remarkable non-parametric approximation of what in classical geostatistics is known as a *random function*, or what in this framework corresponds to the true joint spatial distribution $p(Z(\mathbf{x}_1), \ldots, Z(\mathbf{x}_N) \mid \mathcal{O})$, based on:

$$p(Z(\mathbf{x}_1), \ldots, Z(\mathbf{x}_N) \mid \mathcal{O}) \approx \prod_{i=1}^{N} p(Z(\mathbf{x}_i) \mid \mathcal{N}(\mathbf{x}_i)). \tag{2.35}$$

### *2.4. Practical considerations and applications*

In practice, several considerations affect the implementation of the SDE framework.

(a) **Quality and quantity of observations.** Condition (C1), which ensures consistent marginal estimation, relies on the convergence of empirical distributions computed at each location $\mathbf{x}_i$. This convergence improves with larger ensemble size $T$ and higher data quality in the local neighborhoods. Sparse or noisy data in $\mathcal{N}(\mathbf{x}_i)$ can degrade the accuracy of $\widehat{F}_i^{(T)}$, making the estimation less reliable. Moreover, unbalanced sampling across space may induce heterogeneous estimation quality. Thus, while theoretical convergence holds under general assumptions, practical performance depends critically on both the number and informativeness of the observations that contribute to each neighborhood.

(a) **Empirical distributions vs continuous estimates.** Throughout this analysis, we have considered empirical distributions $\widehat{F}_i^{(T)}$ due to their theoretical tractability and convergence properties. However, in applied settings, continuous approximations such as kernel density estimators (KDEs) or Gaussian mixture models (GMMs) with strictly positive variances may be preferred. These estimators preserve the essential assumptions behind Conditions (C1)–(C3):

→ They produce smooth transformations $z \mapsto \widehat{F}_i(z)$ that improve numerical inversion and sampling.

→ They maintain consistency under standard regularity assumptions.

→ They lead to well-behaved pseudo-observations $u_i^{(t)} := \widehat{F}_i(z_i^{(t)})$ over $[0,1]$, supporting the construction of equicontinuous empirical copulas.

(a) **Point estimation and scenario simulation.** The distributional estimation model proposed in this work supports both point estimation and the generation of spatially consistent scenarios. Point estimation adheres to the same core philosophy as the ESA framework, in which local information is adaptively aggregated to inform estimates at each location. This approach not only preserves spatial consistency with the underlying data but also enables data augmentation through the ensemble structure. Empirical evidence has shown that the ESA approach achieves results comparable to, and in challenging cases superior to, those of classical geostatistical methods [28].

In addition, the model introduces a new class of geostatistical simulation methods, which we refer to as *ensemble spatial simulation* (ESS). This technique, in accordance with the theoretical basis already presented, has two key properties: (1) it generates results that are highly consistent with the spatial structure expressed in the data, due to the structure imposed by the latent copula $\widehat{C}_T$, and (2) it focuses on local statistics due to the pointwise fitting of local distributions ($\widehat{F}_i(z)$). The performance of the ESS is examined in detail through empirical studies presented in the following sections.

These tools enable spatial inference without requiring parametric assumptions on the global spatial dependence structure, aligning with the data-driven philosophy of the SDE framework.

## 3. Case study

As mentioned above, solely as a case study, we use the SDE model to introduce a new class of geostatistical simulation methods, which we refer to as ensemble spatial simulation (ESS), based on independent sampling at each location, while preserving the spatial structure through the learned copula.

### *3.1. Simulation of scenarios*

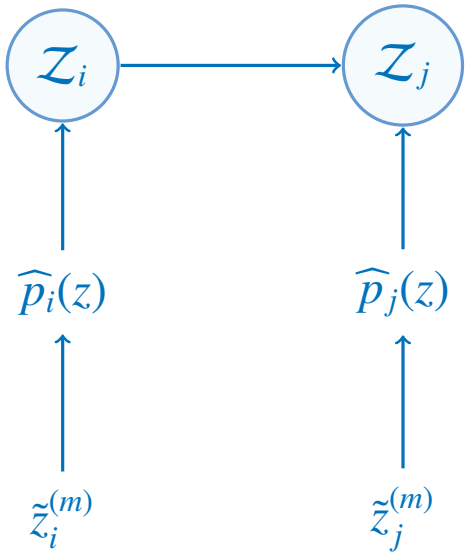


**Figure 4.** The samples $\tilde{z}_i^{(m)}$ and $\tilde{z}_j^{(m)}$ are dependent on each other because of the dependence between $\mathcal{Z}_i$ and $\mathcal{Z}_j$.

In this configuration, we assume that a univariate probabilistic model $\widehat{p}_i(z)$ can be fitted to $\mathcal{Z}_i$, yielding a smoothed estimate of the local predictive density for $Z(\mathbf{x}_i)$ that generalizes the empirical distribution $\widehat{F}_i^{(T)}(z)$. Once local densities are obtained $\widehat{p}_i(z)$ for all $i = 1, \ldots, N$, $M$ spatial realizations are generated by independent sampling, while preserving the spatial structure through the mutual statistical dependencies demonstrated in Proposition 1 – Figure 4:

$$\tilde{z}_i^{(m)} \sim \widehat{p}_i(z), \quad m = 1, \ldots, M. \tag{3.1}$$

Then, each realization $\tilde{Z}^{(m)} \in \mathbb{R}^N$ is given by:

$$\tilde{Z}^{(m)} = (\tilde{z}_1^{(m)}, \ldots, \tilde{z}_N^{(m)})^\top. \tag{3.2}$$

So the complete output is a matrix:

$$\mathbf{S} = [\tilde{Z}^{(1)}, \ldots, \tilde{Z}^{(M)}] \in \mathbb{R}^{N \times M}, \tag{3.3}$$

which represents a set of spatial scenarios consistent with the inferred posterior uncertainty of ESA.

### *3.2. Local predictive models and configurations*

To assess the sensitivity and robustness of ESS, two local predictive models for estimating $\widehat{p}_i(z)$ at each spatial location $\mathbf{x}_i$ are selected:

#### 3.2.1. Kernel density estimation

Kernel density estimation (KDE) is a non-parametric method that estimates a probability density function by summing kernel functions placed at each data point. The general form is:

$$\widehat{p}_i(z) = \frac{1}{T \cdot h_i} \sum_{j=1}^{T} K\left(\frac{z - z_i^{(j)}}{h_i}\right), \tag{3.4}$$

where $T$ is the size of the ESA sample at $\mathbf{x}_i$ ($\mathcal{Z}_i$), $K(\cdot)$ is the kernel function, and $h_i$ is the associated bandwidth.

In our study, we consider two kernel types:

- Gaussian kernel: $K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}$. It provides a smooth, bell-shaped contribution.
- Top-hat (rectangular) kernel: $K(u) = \frac{1}{2}$ for $|u| \leq 1$ and 0 otherwise. It is useful for capturing sudden local changes in density. This is especially necessary in geostatistical data to account for the *nugget effect* that the data may present.

As for bandwidth selection, the Silverman method is used to select the bandwidth automatically. Therefore, our bandwidth is given by:

$$h_i = 0.9 \min\left(\sigma_i, \frac{\mathrm{IQR}_i}{1.34}\right) T^{-1/5}, \tag{3.5}$$

where $\sigma_i$ is the standard deviation and $\mathrm{IQR}_i$ is the interquartile range of the point ESA sample $\mathcal{Z}_i$. It is worth mentioning that the method of Silverman is preferred over the rule of Scott due to its slightly

increased robustness to outliers and skewed distributions, common in geostatistical data, providing a smoother estimate.

#### 3.2.2. Gaussian mixture models

Gaussian mixture models (GMM) represent a probability distribution as a sum of multiple Gaussian distributions, effective for modelling complex, multimodal or skewed distributions:

$$\widehat{p_i}(z) = \sum_{k=1}^{K} \pi_{k,i} \mathcal{N}(z|\mu_{k,i}, \sigma_{k,i}^2), \tag{3.6}$$

where $\pi_{k,i}$ is the weight of the $k$-th component at location $\mathbf{x}_i$, and $\mathcal{N}(z|\mu_{k,i}, \sigma_{k,i}^2)$ is the Gaussian probability density function with mean $\mu_{k,i}$ and variance $\sigma_{k,i}^2$ for the $k$-th component at the same location.

For this, we consider two fitting methods:

- Expectation-maximization (EM): An iterative algorithm for finding maximum likelihood estimates. In general, it is computationally efficient and provides good local optima for unimodal/bimodal distributions. However, it is prone to overfitting with too many components or sparse data, which can occur in many cases of the samples $\mathcal{Z}_i$.

- Variational inference (VI): Another type of inference method that approximates the true posterior distribution by minimizing KL divergence. It is less prone to overfitting due to its inherent regularization structure and provides more robust estimates for smaller sample sizes or sparse data. However, it generally requires greater computational effort.

Both EM and VI are tested with one component, assuming local Gaussianity, and two components to capture bimodality or asymmetry.

### *3.3. Datasets*

Two datasets are used: (i) a simulated dataset for controlled testing, and (ii) a real-world dataset for practical assessment. In both cases, it is essential to define a base scenario, which serves as the "ground truth" and from which a number of samples are extracted to generate various simulated versions.

#### 3.3.1. Simulated data

For controlled testing, 100 Gaussian simulations were generated using the *turning bands algorithm* (TBSIM) [1], which generates stationary and spatially correlated Gaussian fields by integrating one-dimensional random processes. This provides a benchmark to assess the ability of ESS to implicitly reproduce known spatial structures.

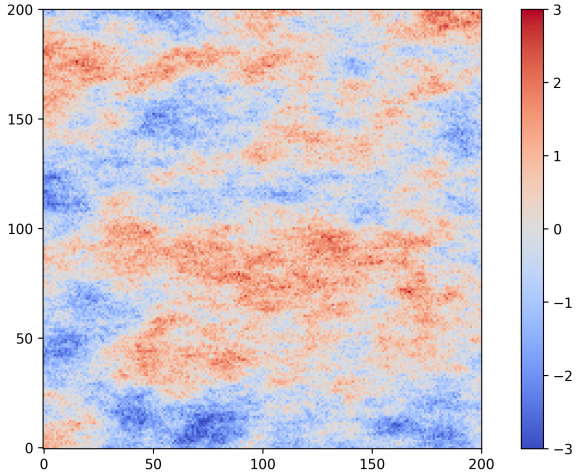


**Figure 5.** Base scenario used in the simulated data case.

One simulation is chosen as the base scenario (Figure 5). From this, three conditioning data sets are extracted to simulate varying data availability (Figure 6):

(a) 5% Random sample: Moderately dense sampling.

(b) 1% Random sample: Sparse sampling scenario.

(c) Reduced sample (50 samples): Fixed small sample size, regardless of domain size.

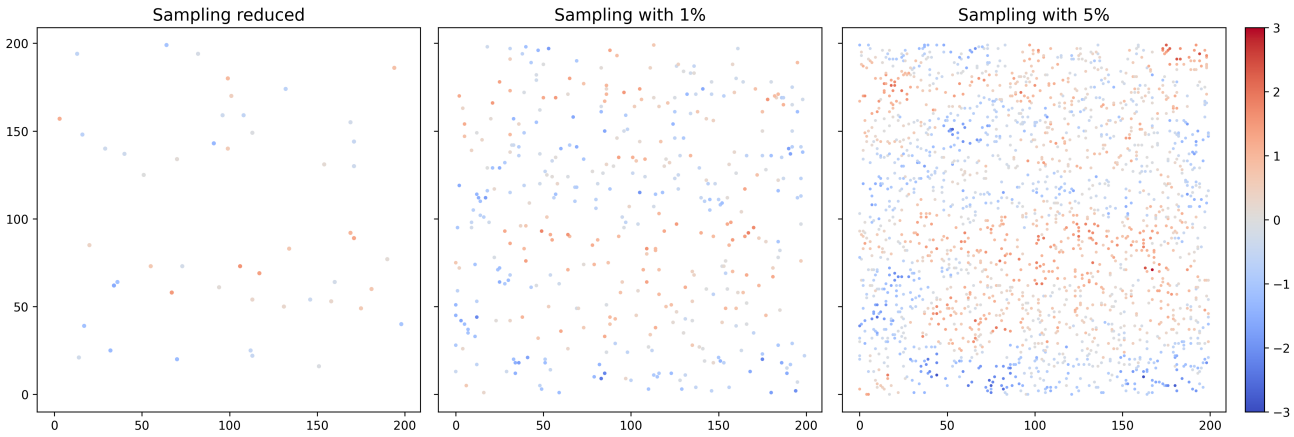


**Figure 6.** Conditioning locations for each sampling case: Reduced sample (left), 1% Sampling (middle), and 5% Sampling (right).

### 3.3.2. Real data

To demonstrate practical applicability, a set of real conditioning data is utilized. It is built upon a real data set taken from the Río Blanco–Los Bronces porphyry copper deposit, which is a breccia complex located in the Chilean central Andes. These data, obtained from [26], correspond to a 2D cross-section (north-east horizontal plane) of a three-dimensional drill-hole information. The conditioning data locations (400 data samples) are presented in Figure 7. The highest values of copper grade are

concentrated in the center (mainly in south and north places) of the study area, whereas the lowest ones can be observed in the east and west zones.

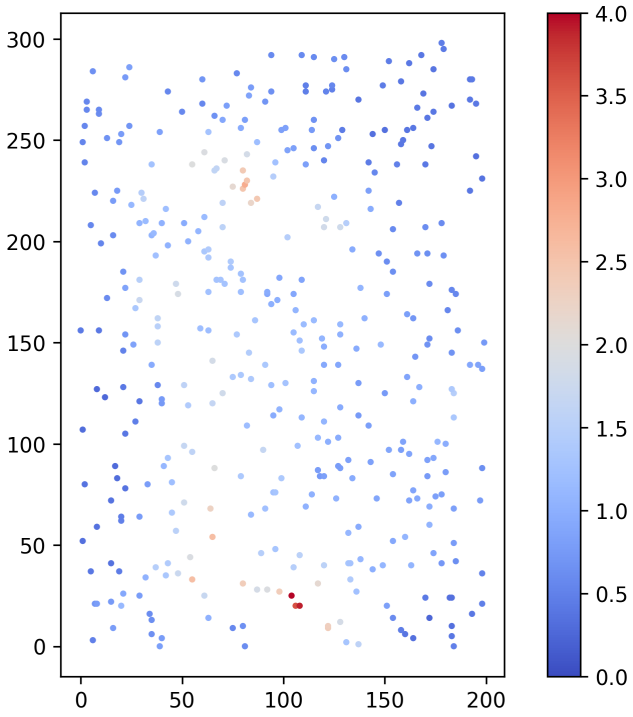


**Figure 7.** Conditioning locations for the real data case.

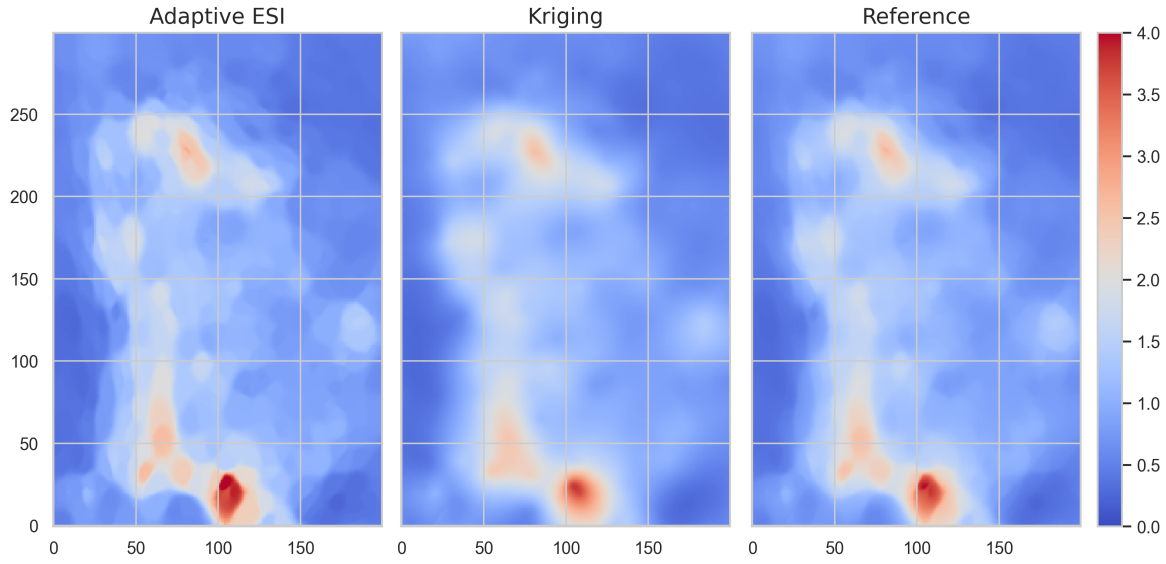


**Figure 8.** Estimations used in the real data case: ESA estimation (left), kriging estimation (middle), and Average of both (right) used as the reference in this case.

Since there is no "true" baseline scenario for the real data, the average of two estimates is considered as the baseline scenario, one estimate using ordinary kriging (which tends to favor TBSIM) and one

estimate using ESA (which tends to favor our proposed ESS model) – see Figure 8.

## 4. Results

This section presents and discusses the results obtained from applying the ESS algorithm with different local predictive model configurations on both simulated and real-world datasets.

### *4.1. Visual analysis of ESS realizations*

Figures 9 and 10 present visual examples of ESS realizations for simulated (1% sampling) and real conditioning datasets, respectively. A remarkable observation from both of them is how the ESS model proves its capability to preserve the spatial structure. Despite the fact that the ESS model is simulated by independent spatial sampling at each location, the global spatial structure is effectively preserved through the implicit learning of the underlying latent copula $\widehat{C}$ that governs the spatial relationships within the data. The resulting ESS simulations are highly consistent with the ground truth (for the simulated data) and with each other, evidencing high precision in capturing essential spatial patterns. This inherent data-driven approach addresses a well-known limitation of traditional Gaussian and TBSIM simulations (the rightmost column in both Figures 9 and 10), which can sometimes lead to practically unfeasible scenarios if their explicit variogram models do not accurately reflect actual spatial variability or if the simulation model is not modified to adapt to harder constraints [12].

Upon closer inspection, the KDE simulations generally exhibit a more textured appearance than the GMM simulations. This is to be expected, since KDE is an approach that can more directly reflect the more marked local variability present in the ensemble. On the other hand, a somewhat contradictory phenomenon is observed in KDE configurations, which is that the Gaussian kernel tends to show more texture than the Top-hat kernel. This could be attributed to the local bandwidth $h_i$ chosen in the Top-hat kernel, which, if relatively wide, could produce an unwanted smoothing effect, thereby reducing the visible texture. In contrast, the Gaussian kernel, although often associated with smoothing, seems to preserve more fine-scale details in these specific contexts, possibly due to a more nuanced interaction with the local data density.

This observation about bandwidth and smoothing seems to be corroborated by GMM simulations. The 2-component GMM configurations tend to appear smoother than their 1-component counterparts. This suggests the local presence of more than one significant (local) population, which the 2-component model attempts to fit, potentially resulting in a more generalized and smooth local distribution. An interesting consequence of this is that, in general, GMM realizations tend to be smooth with highly localized areas of texture. This pattern suggests that ESS, through its construction of local models, favors the reproduction of local effects and characteristics over strictly global ones, allowing for adaptive behavior in heterogeneous spatial domains. Both EM and VI fitting methods for GMMs show similar overall visual characteristics, with VI potentially offering slightly more robust estimates in regions of sparse data due to its inherent regularization.

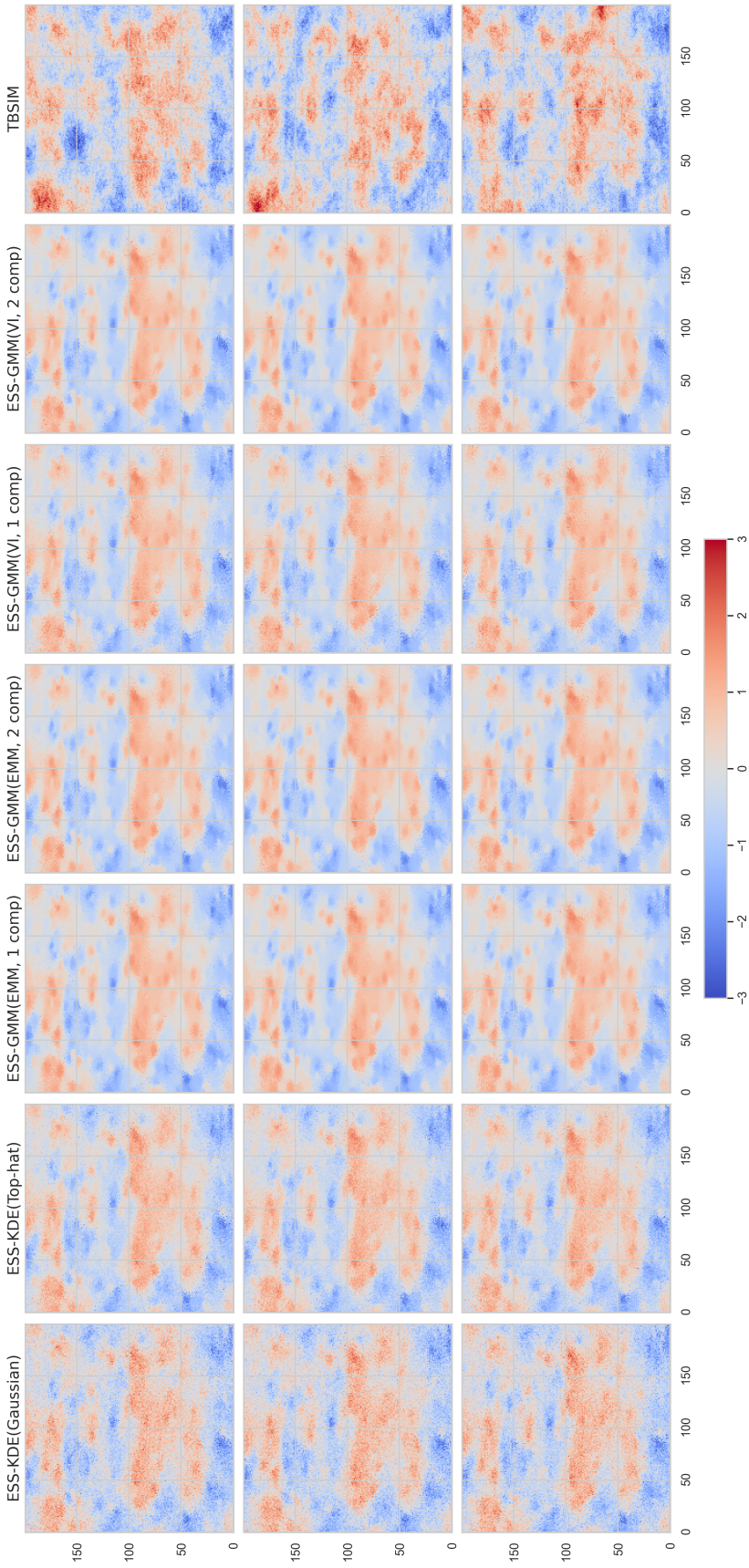


**Figure 9.** Example realizations from ESS for simulated data (1% sampling). Each row represents a different realization, and columns show results for various local predictive model configurations compared to example realizations generated using the TBSIM algorithm (final column, rightmost). From left to right, configurations are: KDE Gaussian, KDE Top-hat, GMM (EM, 1 comp), GMM (EM, 2 comp), GMM (VI, 1 comp), GMM (VI, 2 comp), and TBSIM.

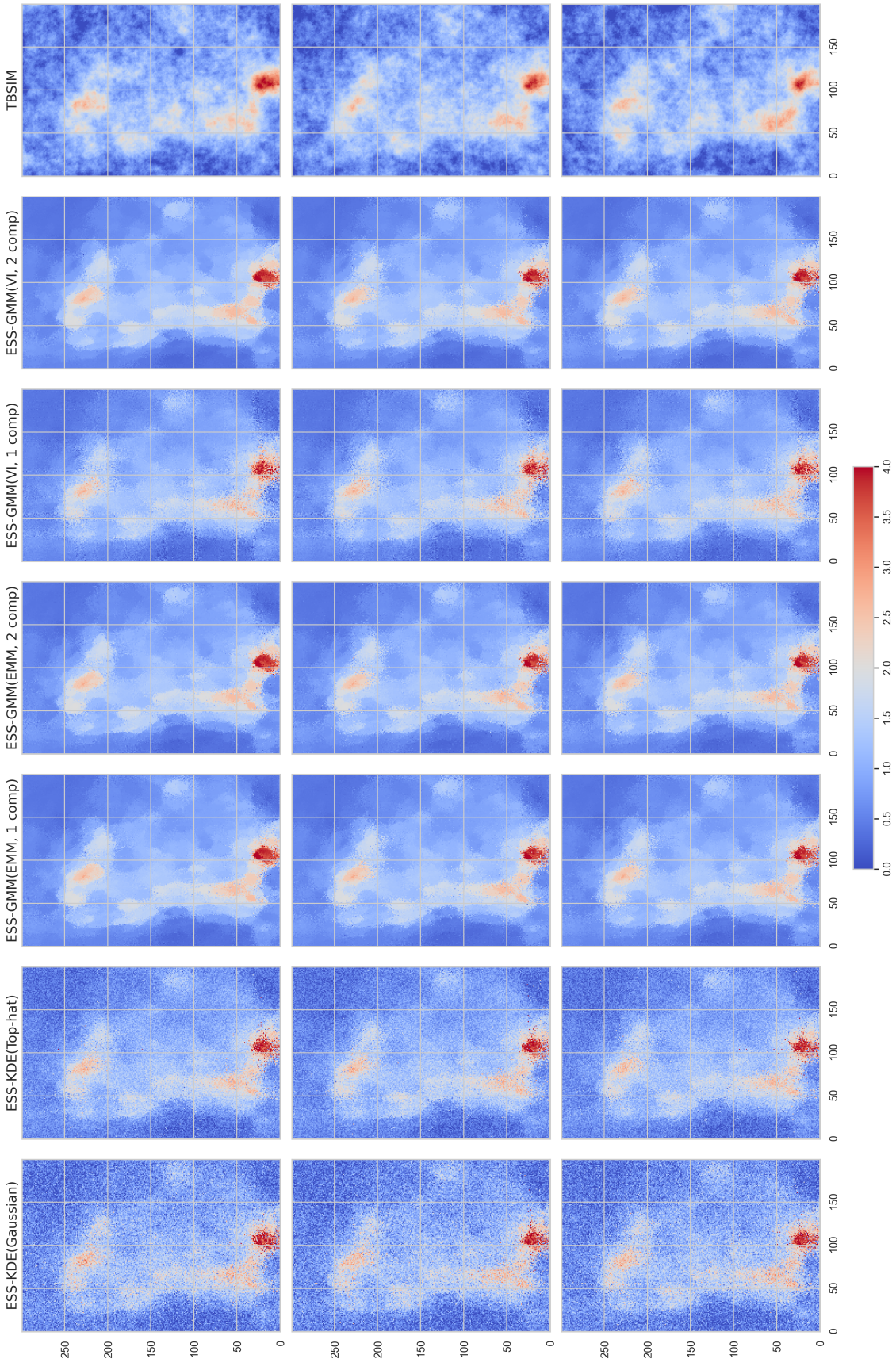


**Figure 10.** Example realizations from ESS for real conditioning data. Each row represents a different realization, and columns show results for various local predictive model configurations. From left to right, configurations are: KDE Gaussian, KDE Top-hat, GMM (EM, 1 comp), GMM (EM, 2 comp), GMM (VI, 1 comp), GMM (VI, 2 comp).

## 4.2. *Global statistics check*

### 4.2.1. First and second order

This section presents a straightforward inspection of the fundamental first-order and second-order statistics of the ensemble of scenarios: mean, standard deviation, and median.

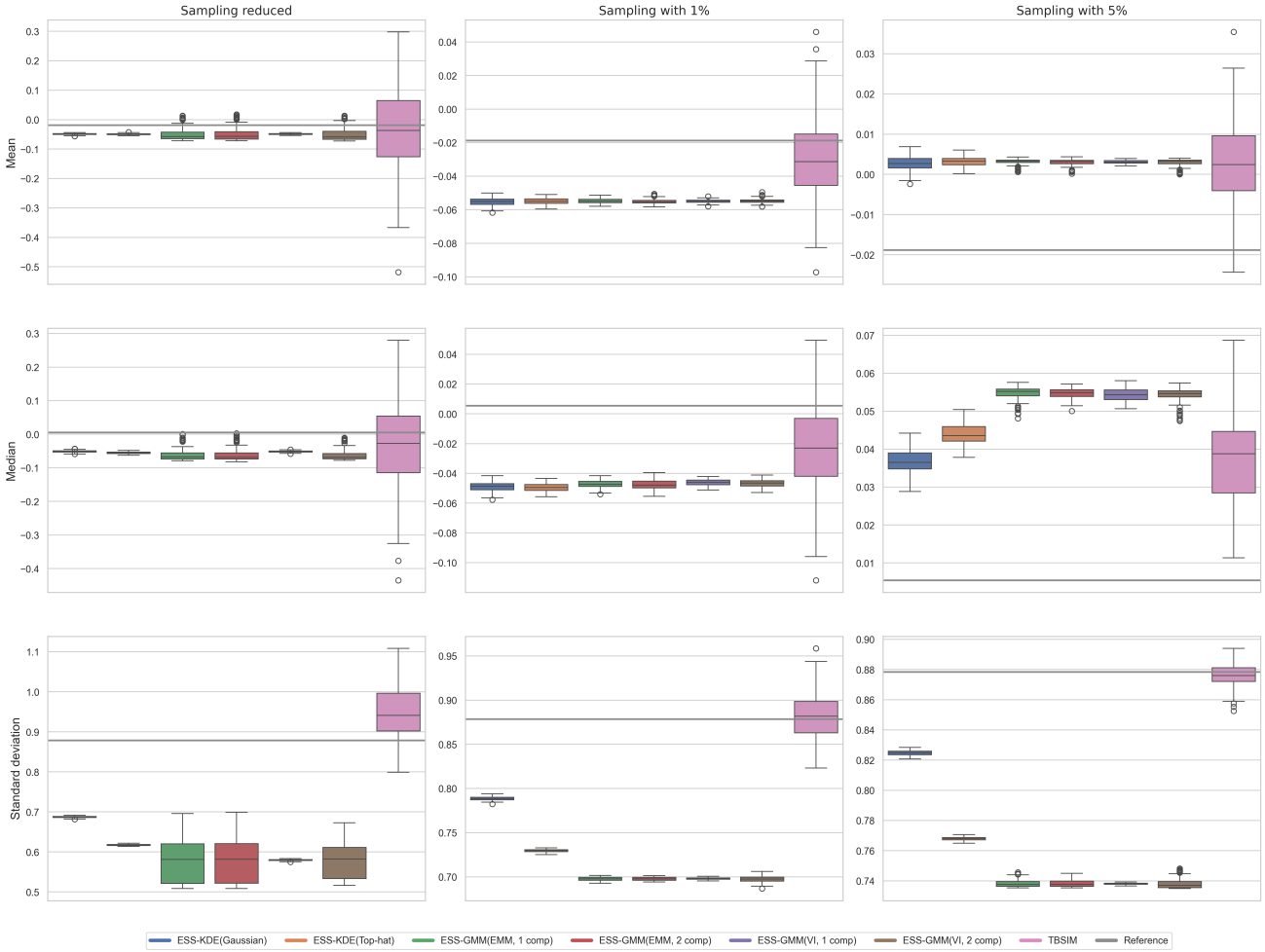


**Figure 11.** Box plots of first-order and second-order statistics for the synthetic case. The solid horizontal line in gray represents the corresponding statistic in the base scenario.

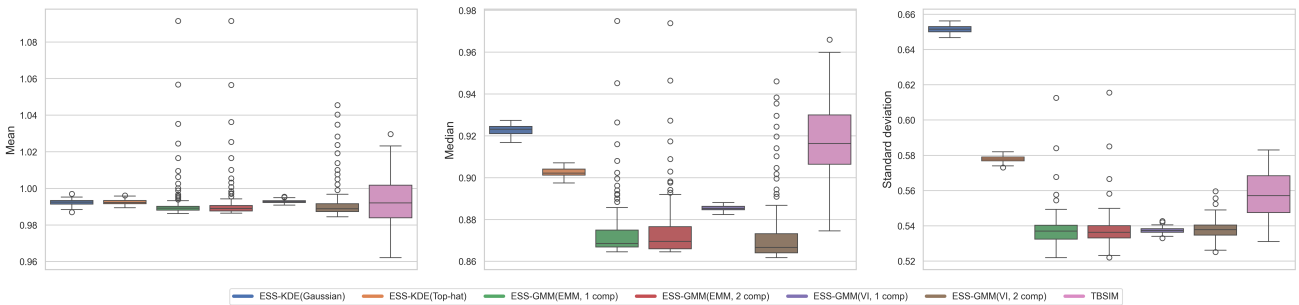


**Figure 12.** Box plots of first-order and second-order statistics for the real data case.

Figure 11 displays box plots of these statistics for the synthetic case, where the solid horizontal line represents the corresponding statistic in the base scenario. Similarly, Figure 12 shows the same for the real data case.

The means appear quite stable across all cases (synthetic and real), although slightly biased in the 1% and 5% sampling scenarios for the synthetic data. It is important to note, however, that considering the scale of the plots, this bias is not significant.

Regarding the standard deviation, it is observed that in the synthetic cases, it tends to be lower in all scenarios for ESS, whereas TBSIM remains stable around the ground-truth. This can be attributed to the construction of the TBSIM model, which is specifically designed to replicate global second-order statistics. While this might initially appear as a limitation of ESS, it can be explained by the fact that the ESS model inherently allows less "freedom" in choosing scenarios, as it anchors itself to the data, thereby producing a smoothing effect in the results. In the real data case, a more or less expected outcome is observed, namely that ESS-KDE demonstrates a better ability to capture the local behavior of the spatial structure, owing to the inherent properties of its fitting method.

The case of the median is particularly interesting, as it appears very stable in the reduced scenario, likely due to the low expressiveness of the limited amount of data. It gradually becomes more biased as the data volume increases. In the 5% case, it is once again corroborated that the choice of the fitting method plays a crucial role in capturing local spatial properties. This is particularly evident in the case of real data.

#### 4.2.2. Structure of spatial continuity

This section presents the variograms calculated along the principal directions ($X$ and $Y$ axes). Variograms are statistical tools used in geostatistics to quantify the spatial autocorrelation of a regionalized variable. They describe the degree of spatial dependence between two data points as a function of their separation distance and direction, thereby summarizing the second-order statistics of the spatial field.

Figure 13 shows the average variograms of the scenarios for different sampling cases. The first row displays variograms along the $X$-axis, and the second row along the $Y$-axis. The noisy gray line represents the reference variogram, calculated from the ground truth. It can be observed that TBSIM most closely matches the reference, which is expected given its design specifically for reproducing second-order statistics. However, for the 1% and 5% sampling cases, ESS produces variograms that, although exhibiting a slightly lower sill (explained by the reduction in variance mentioned previously), successfully reproduce the shape of the reference variogram. This indicates that ESS effectively captures spatial continuity. Under reduced sampling, ESS loses more of the variogram structure compared to TBSIM, likely due to its focus on capturing local statistics exclusively from the data. This is further corroborated by the remarkable ability of ESS to capture the nugget effect directly from the data across all sampling cases. For algorithms like TBSIM (or traditional Gaussian simulation), this is impossible unless explicitly incorporated into their hyperparameters.

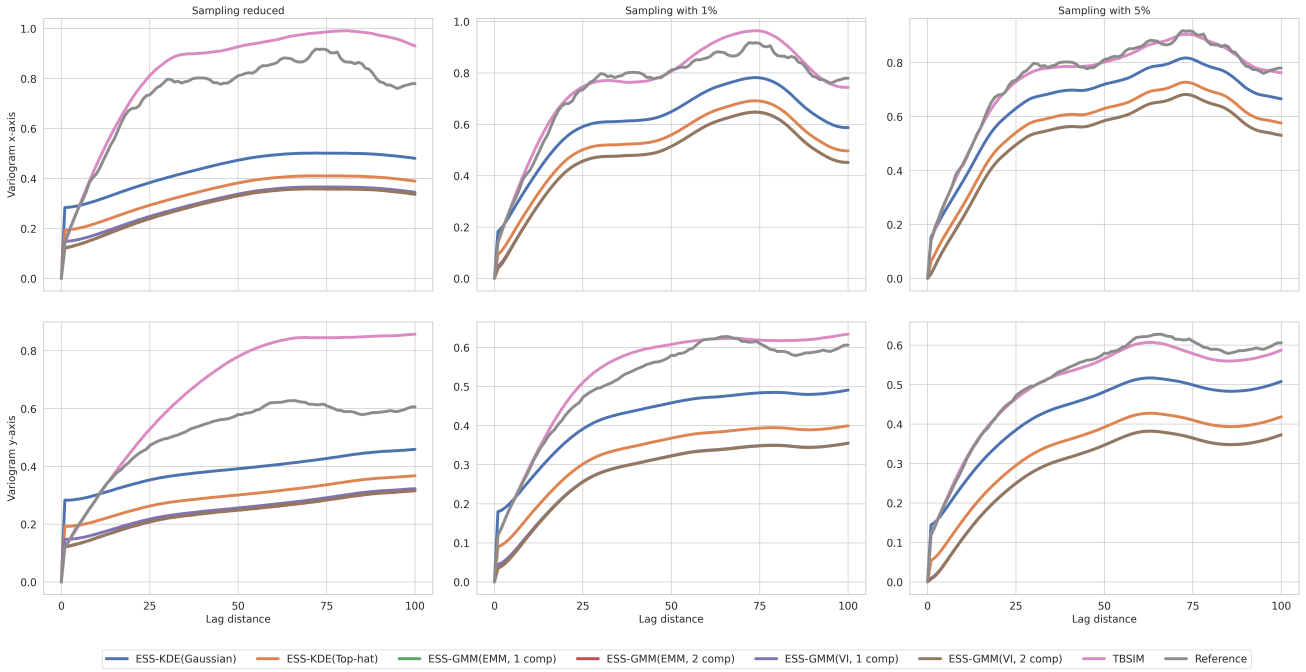


**Figure 13.** Average variograms for the synthetic data scenarios. The first row shows variograms along the $X$-axis, the second row along the $Y$-axis. The noisy gray line represents the reference variogram calculated from the ground truth.

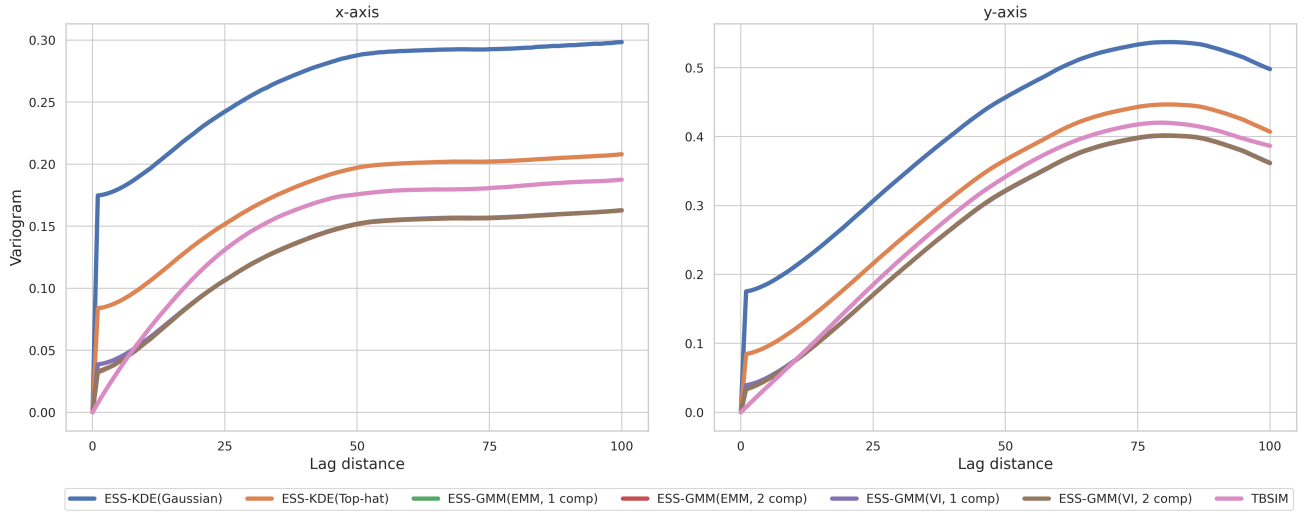


**Figure 14.** Average variograms for the real data simulations. The first row shows variograms along the $X$-axis, the second row along the $Y$-axis.

Conversely, Figure 14 displays the average variograms for simulations using real data. Here, both ESS and TBSIM demonstrate a similar capacity to capture the structure of spatial continuity, as evidenced by the consistent shape of their variograms. However, a clear distinction emerges when it is noted that TBSIM fails to capture the nugget effect (because it was not explicitly included in its hyperparameters), whereas ESS successfully captures it in all real data cases. It should be noted in

this regard that the proposed ESS methodology infers spatial dependence through a latent (non-parametric) copula structure learned directly from the data, while the TBISM method requires the explicit characterization of the modelled variogram as input (and, therefore, it is easier for it to establish a correspondence between this spatial continuity reference and the resulting variograms).

### *4.3. Local statistics check*

To assess the reproduction of local statistics, we employed the following methodology. A sample of random positions within the base scenario was selected. For each position, the empirical distribution (histogram) of values within a $w \times w$ window centered on that position was computed, both in the base scenario and in each generated simulation. The Jensen-Shannon distance (JSD), $D_{JS}(P||Q)$, between the empirical distributions of the windows in the base scenario ($P$) and those in each simulation ($Q$) was then calculated. The JSD is a symmetric and finite measure of similarity between two probability distributions. It is based on the Kullback-Leibler (KL) divergence and defined as:

$$D_{JS}(P||Q) = \frac{1}{2}D_{KL}(P||M) + \frac{1}{2}D_{KL}(Q||M), \tag{4.1}$$

where $M = \frac{1}{2}(P + Q)$. The JSD ranges from 0 to 1, with 0 indicating identical distributions and 1 indicating maximum dissimilarity. Its properties, including its relationship with KL divergence, allow it to compare distributions across all their moments, making it a robust metric for assessing the similarity of empirical distributions. The average JSD across all selected positions was then computed; a lower average JSD indicates better reproduction of local distributions. We tested this for window sizes $w = 7, 13, 19$, where smaller $w$ values are assumed to capture more localized statistics.

Figure 15 presents the results for the synthetic data. It can be seen that for all sampling cases and for the smallest window size ($w = 7$), ESS consistently shows better performance in capturing local behavior. The ESS-KDE configuration yields the best results, which is consistent with its observed ability to capture the nugget effect. As the window size increases, the performance of TBSIM improves, which is consistent with its design. Notably, even for the largest window size ($w = 19$), ESS remains competitive with TBSIM in the 5% sampling case.

However, with real data (Figure 16), where TBSIM is no longer operating under its ideal conditions (i.e., when explicit variogram modelling for real data might be challenging or less accurate), ESS demonstrates better performance across all cases.

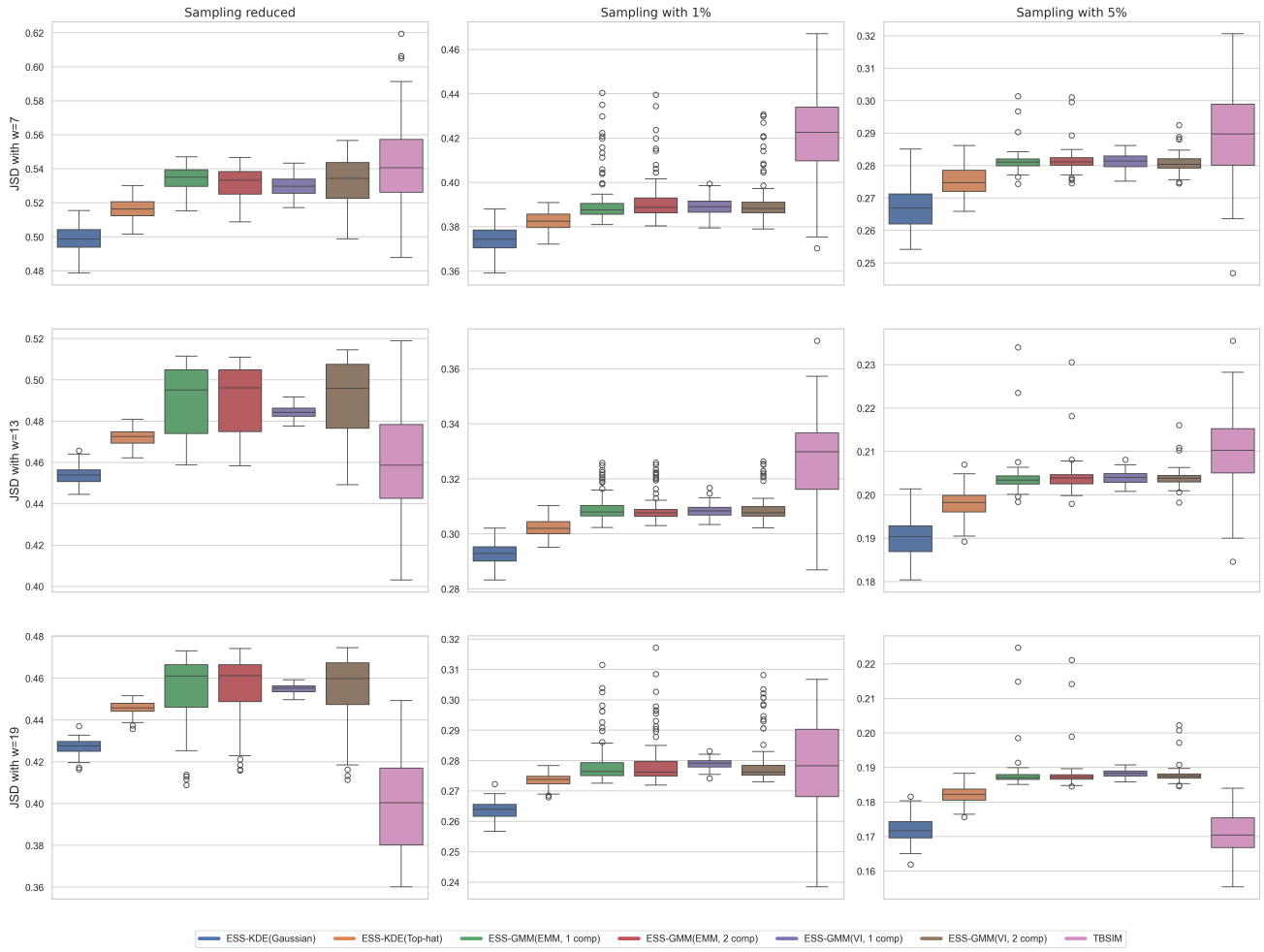


**Figure 15.** Jensen-Shannon distance comparison for synthetic data. Box plots show the distribution of JSD values for different sampling cases and window sizes ($w$=7, 13, 19). Lower values indicate better local statistics reproduction.

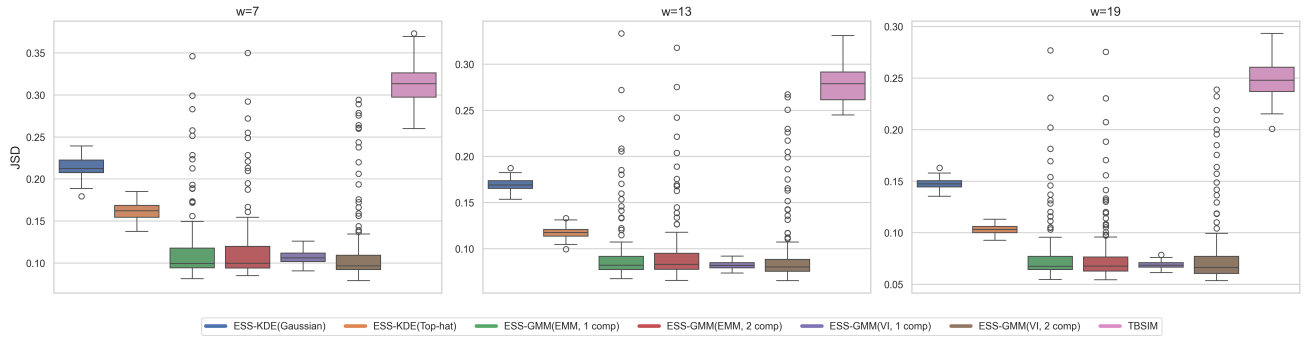


**Figure 16.** Jensen-Shannon distance comparison for real data. Box plots show the distribution of JSD values for different window sizes ($w$=7, 13, 19). Lower values indicate better local statistics reproduction.

### 4.4. Similarity measurement

It is crucial to quantify the observation from the visual inspection section that ESS, as a data-driven model, does not generate infeasible simulations. While this is inherently non-trivial to measure, given its dependence on the specific characteristics of the $Z(\mathbf{x}_i)$ distribution, measuring the 'similarity' between the base scenario and the generated simulations seems a pertinent approach, assuming that infeasible scenarios tend to be less similar to the base scenario.

To quantify this similarity, we employ the structural similarity index measure (SSIM), a widely recognized metric in image analysis literature [39]. Formally, SSIM between two images (or spatial fields) $x$ and $y$ is defined as:

$$\text{SSIM}(x, y) = [l(x, y)]^{\alpha} \cdot [c(x, y)]^{\beta} \cdot [s(x, y)]^{\gamma}, \tag{4.2}$$

where $l(x, y)$ is the luminance comparison function, $c(x, y)$ is the contrast comparison function, and $s(x, y)$ is the structure comparison function. These are typically defined as:

$$\begin{aligned} l(x, y) &= \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \\ c(x, y) &= \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \\ s(x, y) &= \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}. \end{aligned} \tag{4.3}$$

Here, $\mu_x, \mu_y$ are the means of $x, y$; $\sigma_x, \sigma_y$ are the standard deviations of $x, y$; and $\sigma_{xy}$ is the covariance of $x, y$. $C_1, C_2, C_3$ are small constants to prevent division by zero, and $\alpha, \beta, \gamma$ are weights, typically set to 1. The SSIM metric yields values in the range $[0, 1]$, where a value of 1 signifies perfect similarity between the compared images. Similar to the local statistics check, this method also employs a moving window over the images to compute the statistics. We also tested for window sizes $w = 7, 13, 19$.

Figure 17 shows the results for the synthetic data. It can be seen that ESS generally demonstrates better performance across all cases. Interestingly, ESS-KDE (particularly with the Gaussian kernel) shows a decrease in performance in some cases, becoming comparable to TBSIM. This does not imply the generation of infeasible scenarios; rather, this is explained by the known tendency of SSIM to heavily penalize *salt and pepper* noise [39], which is highly compatible with the nugget effect effectively captured by these models. This phenomenon is observed more explicitly in the real data case shown in Figure 18.

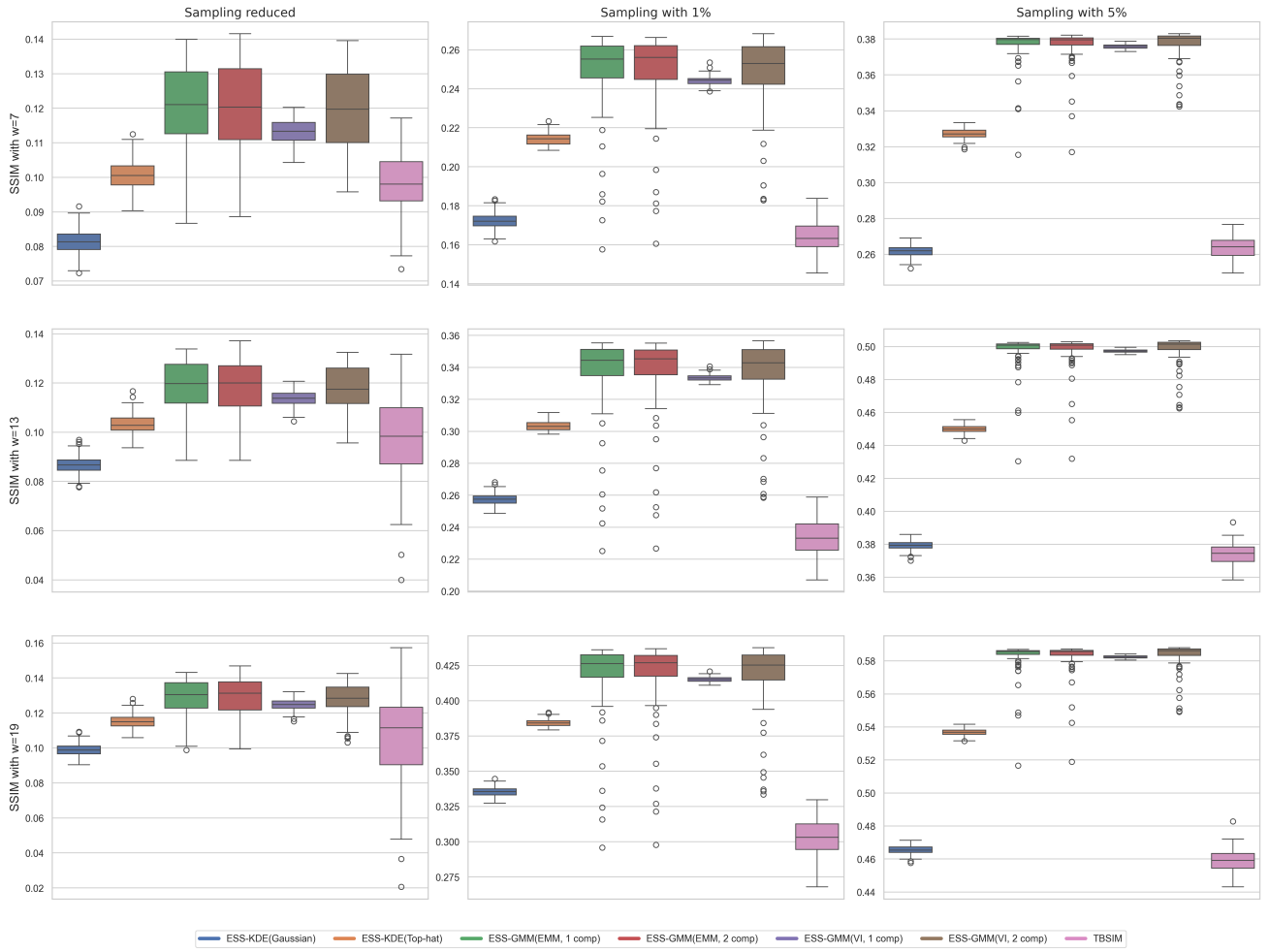


**Figure 17.** SSIM comparison for synthetic data. Box plots show the distribution of SSIM values for different sampling cases and window sizes ($w$=7, 13, 19). Higher values indicate greater similarity.

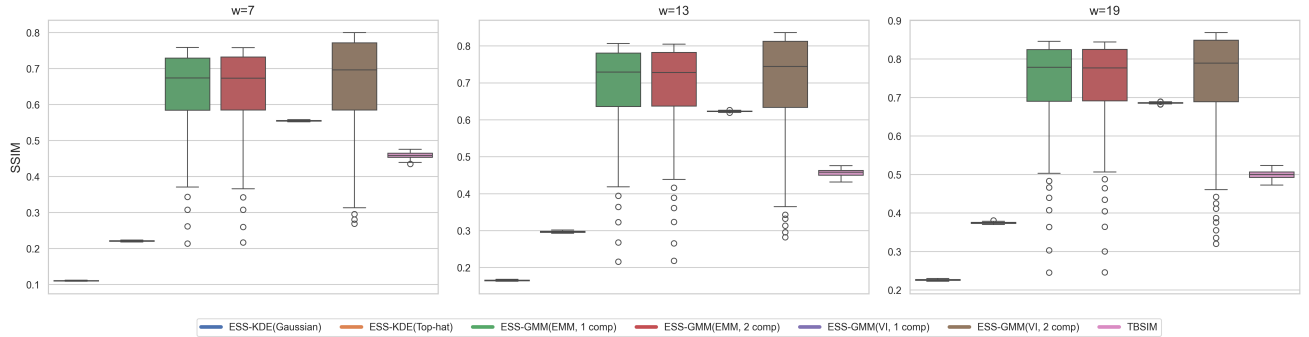


**Figure 18.** SSIM comparison for real data. Box plots show the distribution of SSIM values for different window sizes ($w$=7, 13, 19). Higher values indicate greater similarity.

### 4.5. Computational performance

A relevant aspect of the proposed methodology is its speed in performing a simulation of a spatial variable. In order to contrast this aspect between the different ESS variants with the way of estimating the local distribution, as well as with respect to the TBSIM simulation methodology, an experiment was carried out in which 1000 realizations of the base scenario used in the simulated data case (Figure 5). Table 1 shows the results of the experiments performed for the different sampling cases considered (reduced, 1% and 5%), shown in Figure 6. In this experiment, we considered the separation of the total time into: (i) modelling time $t_M$ by an expert person before the application of the simulation methodology, (ii) base time $t_B$ which includes the different pre-computations necessary to optimize the referred process to generate each realization, and (iii) the simulation time $t_S$ considered to generate the 1000 realizations of the spatial variable.

**Table 1.** Calculation times for simulation methods (in seconds) to generate 1000 realizations.

| Methods | Sampling reduced | | | Sampling with 1% | | | Sampling with 5% | | |
|---|---|---|---|---|---|---|---|---|---|
| | $t_M$ | $t_B$ | $t_S$ | $t_M$ | $t_B$ | $t_S$ | $t_M$ | $t_B$ | $t_S$ |
| ESS-KDE(Gaussian) | 0 | 11 | 4 | 0 | 285 | 4 | 0 | 5919 | 4 |
| ESS-KDE(Top-hat) | 0 | 11 | 8 | 0 | 285 | 8 | 0 | 5919 | 7 |
| ESS-GMM(EMM, 1 comp) | 0 | 11 | 88 | 0 | 285 | 80 | 0 | 5919 | 71 |
| ESS-GMM(EMM, 2 comp) | 0 | 11 | 85 | 0 | 285 | 81 | 0 | 5919 | 71 |
| ESS-GMM(VI, 1 comp) | 0 | 11 | 60 | 0 | 285 | 58 | 0 | 5919 | 55 |
| ESS-GMM(VI, 2 comp) | 0 | 11 | 207 | 0 | 285 | 196 | 0 | 5919 | 163 |
| TBSIM | * | 7 | 354 | * | 210 | 353 | * | 5509 | 356 |

It is verified that for the different ESS variants, we have the same times $t_B$ (as well as the null time $t_M = 0$ in each case). This is due to the fact that this time is considered in the stages prior to the generation of the realizations: (i) partitioning, (ii) parameter adjustment in each cell, and (iii) creation of the ESA samples. On the other hand, the generation of the realizations (considered in time $t_S$) includes the adjustment of the local distribution and the sampling of the 1000 values for each position, information with which each of the 1000 spatial realizations is constructed. Naturally, the impact of each local distribution adjustment methodology is verified at each of the reported times $t_S$.

[*] It should be noted that obtaining results with the TBSIM methodology depends on parameters that are adjusted by expert judgement. This naturally entails an associated modelling time $t_M$, which depends to a large extent on the expertise of the person carrying out the analysis and the means used to do so (e.g. using semi-automatic tools will probably take less time than just manually adjusting the parameters). In contrast, the automatic methodology proposed in ESS has no time associated with modelling.

On the other hand, the choice of parameters in TBSIM has an impact on the computational time associated with the computation prior to a realization, i.e., the base time $t_B$. In this sense, the time reported in Table 1 with the TBSIM method was obtained considering a kriging conditioning that uses a sample search in a neighborhood of 50 units, both in the $x$-axis and in the $y$-axis. Increasing the search neighborhood, while ideal in theoretical terms, is impractical in terms of how it increases the associated computational time for this methodology. Thus, unlike the ESS methodology that considers

all available samples for the estimation of each position in space, the TBSIM method only considers one neighborhood to obtain computation times of the same order as ESS.

Comparing the different sampling cases studied, it is verified that the time base $t_B$ increases as the amount of conditioning data increases.

## 5. Discussion

The proposed framework extends ensemble-based approaches such as ensemble spatial interpolation (ESI) and adaptive ensemble spatial analysis (ESA). These methods generate multiple conditional predictions for each location, which form the basis for estimating local distributions. A central theoretical result established in this work is that, although these distributions are computed independently, the overlapping neighborhood structure induces a unique statistical dependency across the spatial domain. This dependency is rigorously characterized by a statistical latent copula that emerges from the data, showing that the spatial structure can be captured without the need to specify a global model explicitly. Furthermore, it is proven that, as the number of ensemble predictions increases, the estimated marginals converge to the true conditional distributions, and the copula describing spatial dependencies converges to the true structure. These results ensure that the full joint distribution of the spatial field is consistently approximated in a non-parametric and data-driven manner.

An additional theoretical aspect concerns the tessellation process. For the copula convergence to hold, the process must be stationary, ergodic, and generate weakly independent samples. While these requirements are difficult to guarantee in observational data, they can be transferred to a latent sampling process designed to satisfy them, thereby providing a theoretical basis for the approach.

Therefore, the uncertainty associated with the proposed model is completely encoded intrinsically in the estimated joint distribution of the data. Thus, any probabilistic statement about the field (marginal, conditional, or joint) can be derived directly from this distribution. This represents a new approach compared to traditional geostatistical frameworks, where the quantification of uncertainty is often external to the model and requires measures such as coverage rates, prediction intervals, or calibration metrics calculated over realizations to assess the variability of predictions.

Just to show a relevant case study, the framework was evaluated in one of its most prominent and promising applications, namely the generation of simulated scenarios through joint spatial simulation (ESS). Visual inspection of the generated simulations suggested that the model is capable of reproducing the overall spatial structure and texture of the data. This observation is consistent with the theoretical assertion that spatial coherence can arise solely from local information. Different local estimation methods appeared to influence the visual results. In this regard, KDEs tended to capture small-scale variability and features such as the nugget effect, while GMMs often produced smoother and more uniform results, reflecting their adaptation to multimodal local populations.

In terms of statistical properties, ESS produced overall averages and variability that generally matched the original data, although in some cases the simulations showed lower dispersion. This behavior can be interpreted, on the one hand, as greater adherence to the observed data and, on the other, as SDE seeking to approximate the complete distribution, with all its moments, compromising adherence to specific moments (first and second order, as in the case of TBSIM). The choice of local estimator appeared to affect the ability to capture small-scale features, and kernel-based methods

provided richer variability.

Spatial continuity was also reasonably reproduced, suggesting that the framework can capture how spatial values vary with distance and direction. It should be noted that the nugget effect (representing microscale variability or measurement error) appeared naturally in the simulations, without the need to specify it manually. However, it should be remembered that ESS is only a case study in this work and that there is still some way to go before it can compete with established methods such as TBSIM in this regard.

Performance at the local scale was encouraging. Quantitative assessments indicated that the method is effective in capturing fine-scale variability, particularly when kernel-based estimators were used. A standard image similarity index was employed to evaluate spatial realism, and in most cases the proposed approach compared favorably to conventional methods. Slight reductions in similarity were occasionally observed, possibly reflecting the preservation of small-scale noise, which the similarity metric may penalize despite its fidelity to real structure.

From a practical standpoint, the method showed computational efficiency. It reduces the need for manual setup and expert intervention, and enables fast simulation generation – especially when kernel-based estimators are employed. These properties suggest that the framework could be well suited for applications requiring large numbers of spatial realizations.

Looking ahead, several avenues of research appear promising. One direction involves developing more adaptive, data-driven techniques for selecting and tuning local distribution estimators, which could improve performance in complex and heterogeneous environments. Another concerns the systematic evaluation of robustness and sensitivity under challenging conditions such as sparse data and high noise levels, in order to better understand the practical limitations of the method. Although the theoretical results provide guarantees of asymptotic convergence, further work is needed to investigate finite-sample effects, approximation errors associated with different estimator choices, and the impact of variable data quality. Scalability also remains an important challenge; advances in algorithmic design, including parallel computing and efficient approximation methods, may allow the approach to be applied to very large datasets and higher-dimensional problems. Extensions to incorporate auxiliary information such as covariates, temporal dynamics, or multivariate spatial fields could further broaden the applicability of the framework. Finally, integrating recent advances in machine learning may enable more accurate and flexible local distribution estimation, better capturing intricate spatial dependencies that conventional methods might overlook. The development of advanced tools for uncertainty quantification and visualization also represents an important step, as these would support decision-making by allowing users to interpret the confidence and variability of spatial predictions more effectively.

## 6. Conclusions

We have introduced a novel, non-parametric, and decentralized geostatistical methodology for estimating spatial random fields. The central theoretical contribution lies in the identification of a latent statistical copula that naturally emerges when local probability distributions are combined through overlapping adaptive neighborhoods. This emergent copula captures spatial dependence without requiring a predefined global model and, under mild regularity conditions, converges to the true dependency structure. In this way, the framework provides a *consistent non-parametric estimator*

of the joint distribution of a spatial random field.

This theoretical insight may have important implications for spatial statistics. It indicates that global spatial coherence could emerge from local information alone, with latent copulas serving as a possible link between local estimation and global structure. Such a perspective could open new avenues for understanding and modelling spatial processes in a decentralized, data-driven manner.

Future theoretical work should further characterize the convergence properties of the emergent latent copula, study finite-sample effects, and extend the framework to more complex settings such as multivariate or spatio-temporal fields. Further exploration of these foundations could consolidate the role of local–to–global latent copula emergence as a cornerstone in modern non-parametric geostatistics.

## Author contributions

Alvaro F. Egaña: Conceptualization, Methodology, Software, Formal Analysis, Writing - Original Draft, Writing - Review & Editing, Project Administration. Gonzalo Díaz: Methodology, Software, Formal Analysis, Visualization, Writing - Review & Editing. Felipe Navarro: Methodology, Software, Formal Analysis, Visualization, Writing - Review & Editing. Mohammad Maleki: Writing - Review & Editing. Juan F. Sánchez-Pérez: Writing - Review & Editing. All authors have read and agreed to the published version of the manuscript.

## Use of Generative-AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

## Acknowledgments

This material is based on work supported by grants from the Chilean National Agency for Research and Development (ANID), specifically through the Advanced Mining Technology Center (AMTC) of the University of Chile, Basal Project AFB230001. We are especially grateful for the support of the entire team of the Advanced Lab for Geostatistical Super Computing (ALGES) of the AMTC and the Department of Mining Engineering of the University of Chile.

## Conflict of interest

Juan F. Sánchez-Pérez is the Guest Editor of special issue "Modeling and applied mathematics to engineering and applied physics" for AIMS Mathematics. Juan F. Sánchez-Pérez was not involved in the editorial review and the decision to publish this article. The authors declare no conflict of interest.

## References

1. X. Emery, C. Lantuéjoul, Tbsim: A computer program for conditional simulation of three-dimensional Gaussian random fields via the turning bands method, *Comput. Geosci.*, **32** (2006), 1615–1628. https://doi.org/10.1016/j.cageo.2006.03.001

2. X. Emery, A turning bands program for conditional co-simulation of cross-correlated Gaussian random fields, *Comput. Geosci.*, **34** (2008), 1850–1862. https://doi.org/10.1016/j.cageo.2007.10.007
3. M. Nowak, G. Verly, The practice of sequential Gaussian simulation, In: *Geostatistics Banff 2004*, Dordrecht: Springer, 2005, 387–398. https://doi.org/10.1007/978-1-4020-3610-1
4. C. V. Deutsch, A sequential indicator simulation program for categorical variables with point and block data: BlockSIS, *Comput. Geosci.*, **32** (2006), 1669–1681. https://doi.org/10.1016/j.cageo.2006.03.005
5. J. G. Manchuk, C. V. Deutsch, A flexible sequential Gaussian simulation program: USGSIM, *Comput. Geosci.*, **41** (2012), 208–216. https://doi.org/10.1016/j.cageo.2011.08.013
6. P. D. Sampson, P. Guttorp, Nonparametric estimation of nonstationary spatial covariance structure, *J. Amer. Stat. Assoc.*, **87** (1992), 417, 108–119. https://doi.org/10.1080/01621459.1992.10475181
7. E. Pardo-Igúzquiza, P. A. Dowd, Comparison of inference methods for estimating semivariogram model parameters and their uncertainty: The case of small data sets, *Comput. Geosci.*, **50** (2013), 154–164. https://doi.org/10.1016/j.cageo.2012.06.002
8. G. De Marsily, F. Delay, J. Gonçalvès, Ph. Renard, V. Teles, S. Violette, Dealing with spatial heterogeneity, *Hydrogeol. J.*, **13** (2005), 161–183. https://doi.org/10.1007/s10040-004-0432-3
9. L. Keeney, The development of a novel method for integrating geometallurgical mapping and orebody modelling, PhD thesis, University of Queensland, 2010.
10. J.-P. Chiles, P. Delfiner, *Geostatistics: Modeling Spatial Uncertainty*, John Wiley & Sons, 2012. https://doi.org/10.1007/s11004-012-9429-y
11. M. Maleki, X. Emery, Joint simulation of stationary grade and non-stationary rock type for quantifying geological uncertainty in a copper deposit, *Comput. Geosci.*, **109** (2017), 258–267. https://doi.org/10.1016/j.cageo.2017.08.015
12. R. Ferrer, X. Emery, M. Maleki, F. Navarro, Modeling the uncertainty in the layout of geological units by implicit boundary simulation accounting for a preexisting interpretive geological model, *Natural Resour. Res.*, **30** (2021), 4123–4145. https://doi.org/10.1007/s11053-021-09964-9
13. J. B. Boisvert, J. G. Manchuk, C. V. Deutsch, Kriging in the presence of locally varying anisotropy using non-Euclidean distances, *Math. Geosci.*, **41** (2009), 585–601. https://doi.org/10.1007/s11004-009-9229-1
14. P. A. Bostan, G. B. Heuvelink, S. Z. Akyurek, Comparison of regression and kriging techniques for mapping the average annual precipitation of Turkey, *Int. J. Appl. Earth Obser. Geoinf.*, **19** (2012), 115–126. https://doi.org/10.1016/j.jag.2012.04.010
15. D. Arroyo, X. Emery, Simulation of intrinsic random fields of order k with a continuous spectral algorithm, *Stochastic Environ. Res. Risk Assess.*, **32** (2018), 3245–3255. https://doi.org/10.1007/s00477-018-1516-2
16. S. Dunham, J. Vann, Geometallurgy, Geostatistics and project value—does your block model tell you what you need to know, *Proceedings of the Project Evaluation Conference*, 2007, 19–20.
17. P. Carrasco, J.-P. Chilès, S. A. Séguret, Additivity, metallurgical recovery, and grade, *Proceedings of the 8th International Geostatistics Congress*, 2008.

18. P. H. A. Campos, J. F. C. Leite Costa, V. C. Koppe, M. A. Arcari Bassani, Geometallurgy-oriented mine scheduling considering volume support and non-additivity, *Min. Tech.*, **131** (2022), 1–11. https://doi.org/10.1080/25726668.2021.1963607

19. J. L. Deutsch, K. Palmer, C. V. Deutsch, J. Szymanski, T. H. Etsell, Spatial modeling of geometallurgical properties: Techniques and a case study, *Nat. Resour. Res.*, **25** (2016), 161–181. https://doi.org/10.1007/s11053-015-9276-x

20. U. E. Kaplan, Y. Dagasan, E. Topal, Mineral grade estimation using gradient boosting regression trees, *Int. J. Min., Reclam. Environ.*, **35** (2021), 728–742. https://doi.org/10.1080/17480930.2021.1949863

21. N. K. Dumakor-Dupey, S. Arya, Machine learning—a review of applications in mineral resource estimation, *Energies*, **14** (2021), 4079. https://doi.org/10.3390/en1414407

22. S. I. Cevik, J. M. Ortiz, Machine learning in the mineral resource sector: An overview, Technical report, Queen's University, 2020. Available from: https://qspace.library.queensu.ca/server/api/core/bitstreams/afc8a395-c84e-45ff-9dc8-7970c131d2aa/content.

23. K. Dachri, M. Bouabidi, K. Naji, K. Nouar, I. Benzakour, A. Oummouch, et al., Predictive insights for copper recovery: A synergistic approach integrating variability data and machine learning in the geometallurgical study of the Tizert deposit, Morocco, *J. Afr. Earth Sci.*, **212** (2024), 105208. https://doi.org/10.1016/j.jafrearsci.2024.105208

24. M. Kanevski, V. Timonin, A. Pozdnukhov, *Machine Learning for Spatial Environmental Data: Theory, Applications, and Software*, New York: EPFL Press, 2009.

25. A. Burkov, *The Hundred-Page Machine Learning Book*, Andriy Burkov, Quebec City, QC, Canada, 2019. Available from: http://ema.cri-info.cm/wp-content/uploads/2019/07/2019BurkovTheHundred-pageMachineLearning.pdf.

26. A. Egaña, F. Navarro, M. Maleki, F. Grandon, F. Carter, F. Soto, Ensemble spatial interpolation: A new approach to natural or anthropogenic variable assessment, *Nat. Resour. Res.*, **30** (2021), 3777–3793. https://doi.org/10.1007/s11053-021-09860-2

27. A. Menafoglio, G. Gaetani, P. Secchi, Random domain decompositions for object-oriented kriging over complex domains, *Stochastic Environ. Res. Risk Assess.*, **32** (2018), 3421–3437. https://doi.org/10.1007/s00477-018-1596-z

28. A. F. Egaña, M. J. Valenzuela, M. Maleki, J. F. Sánchez-Pérez, G. Díaz, Adaptive ensemble spatial analysis, *Sci. Rep.*, **15** (2025), 26599. https://doi.org/10.1038/s41598-025-08844-z

29. T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2 Eds., New York: Springer, 2009. https://doi.org/10.1007/978-0-387-84858-7

30. R. B. Nelsen, *An Introduction to Copulas*, 2 Eds., New York: Springer, 2006. https://doi.org/10.1007/0-387-28678-0

31. D. Koller, N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*, MIT Press, 2009.

32. J. M. Hammersley, P. Clifford, Markov fields on finite graphs and lattices, Oxford University, 1971. Available from: https://ora.ox.ac.uk/objects/uuid:4ea849da-1511-4578-bb88-6a8d02f457a6.

33. S. L. Lauritzen, D. J. Spiegelhalter, Local computations with probabilities on graphical structures and their application to expert systems, *J. R. Stat. Soc.: Ser. B*, **50** (1988), 157–224.

34. P. Spirtes, C. Glymour, R. Scheines, *Causation, Prediction, and Search*, MIT Press, 2 Eds., 2000.

35. J. Pearl, *Causality: Models, Reasoning, and Inference*, Cambridge University Press, 2 Eds., 2009.

36. P. Deheuvels, La fonction de dépendance empirique et ses propriétés. Un test non paramétrique d'indépendance, *Bulletins de l'Académie Royale de Belgique*, **65** (1979), 274–292.

37. A. W. van der Vaart, J. A. Wellner, *Weak Convergence and Empirical Processes: With Applications to Statistics*, New York: Springer, 1996. https://doi.org/10.1007/978-1-4757-2545-2

38. P. Billingsley, *Convergence of Probability Measures*, 2 Eds., New York: Wiley, 1999. https://doi.org/10.1002/9780470316962

39. Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, Image quality assessment: From error visibility to structural similarity, *IEEE Trans. Image Process.*, **13** (2004), 600–612. https://doi.org/10.1109/TIP.2003.819861

## Supplementary

*Code availability*

The Spatialize library, which implements the proposed methodology (SDE with the case study of ESS), is available on GitHub (https://github.com/alges/spatialize).

*Statements of relevant theorems*

**[Sklar theorem]** Let $H : \mathbb{R}^d \to [0, 1]$ be a joint cumulative distribution function (CDF) with marginals $F_1, \ldots, F_d$. Then there exists a copula $C : [0, 1]^d \to [0, 1]$ such that for all $\mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$,

$$H(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d)). \tag{6.1}$$

If each $F_i$ is continuous, then the copula $C$ is unique. If some of the $F_i$ are not continuous, then $C$ is uniquely determined on $\mathrm{Ran}(F_1) \times \cdots \times \mathrm{Ran}(F_d)$.

Conversely, if $C$ is a copula and $F_1, \ldots, F_d$ are univariate CDFs, then the function

$$H(x_1, \ldots, x_d) := C(F_1(x_1), \ldots, F_d(x_d)) \tag{6.2}$$

is a valid joint CDF with marginals $F_1, \ldots, F_d$.

**[Glivenko–Cantelli theorem]** Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed (i.i.d.) real-valued random variables with cumulative distribution function $F$. Let $F_n$ be the empirical distribution function defined by:

$$F_n(x) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{X_i \leq x\}. \tag{6.3}$$

Then,

$$\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \xrightarrow[n \to \infty]{\text{a.s.}} 0. \tag{6.4}$$

In other words, $F_n$ converges uniformly almost surely to the true distribution function $F$.

**[Arzelà–Ascoli theorem]** Let $\mathcal{F} \subset C(K)$, the space of continuous real-valued functions on a compact metric space $K$, equipped with the uniform norm. If $\mathcal{F}$ is uniformly bounded and equicontinuous, then $\mathcal{F}$ is relatively compact in $C(K)$. That is, every sequence $\{f_n\} \subset \mathcal{F}$ has a uniformly convergent subsequence.

**[Uniform consistency of empirical copulas theorem]** Let $\{(X_t^{(1)}, \ldots, X_t^{(N)})\}_{t=1}^T$ be an i.i.d. sample from a continuous multivariate distribution $H$ with continuous marginals $F_1, \ldots, F_N$ and copula $C$. Let $\widehat{F}_i$ denote the empirical distribution of the $i$-th margin, and let $\widehat{C}_T$ be the empirical copula defined by

$$\widehat{C}_T(\mathbf{u}) := \frac{1}{T} \sum_{t=1}^{T} \prod_{i=1}^{N} \mathbb{I}\left\{\widehat{F}_i(X_t^{(i)}) \le u_i\right\}, \quad \mathbf{u} \in [0,1]^N. \tag{6.5}$$

Then,

$$\sup_{\mathbf{u}\in[0,1]^N} \left|\widehat{C}_T(\mathbf{u}) - C(\mathbf{u})\right| \xrightarrow[T\to\infty]{a.s.} 0. \tag{6.6}$$

**[Weak convergence of the empirical copula process theorem]** Under suitable regularity conditions, the empirical copula process

$$\sqrt{T}\left(\widehat{C}_T(\mathbf{u}) - C(\mathbf{u})\right) \tag{6.7}$$

converges in distribution in $\ell^\infty([0,1]^N)$ to a tight Gaussian process with mean zero and covariance structure determined by $C$.

**[Prokhorov theorem]** Let $\{\mu_T\}_{T\ge1}$ be a family of probability measures on a *separable metric space*. If the family is *tight*, that is, for every $\epsilon > 0$, there exists a compact set $K_\epsilon \subseteq \mathbb{R}^d$ such that

$$\mu_T(K_\epsilon) > 1 - \epsilon \quad \text{for all } T, \tag{6.8}$$

then the family $\{\mu_T\}$ is *relatively compact* in the topology of weak convergence. That is, every sequence $\{\mu_{T_k}\}$ has a subsequence that converges in distribution to a probability measure.

**[Hammersley–Clifford theorem]** This theorem establishes a direct equivalence between graphical separation properties and the factorization of the joint distribution into local components. Before stating the theorem, it is important to consider the following two definitions.

***[Global Markov property]:*** Let $G = (V, E)$ be an undirected graph, and let $\{X_v : v \in V\}$ denote a collection of random variables indexed by $V$. A strictly positive probability distribution $P$ on these variables is said to satisfy the *global Markov property* with respect to $G$ if, for any disjoint subsets $A, B, S \subseteq V$,

$$S \text{ separates } A \text{ and } B \text{ in } G \quad \Longrightarrow \quad X_A \perp\!\!\!\perp X_B \mid X_S \text{ under } P.$$

***[Clique]:*** A clique $C \subseteq V$ is a subset of vertices such that every pair of distinct vertices in $C$ is connected by an edge in $G$. A clique is *maximal* if it is not properly contained in any other clique.

***[Theorem statement [32]]:*** Let $P$ be a strictly positive probability distribution over $\{X_v : v \in V\}$. Then the following are equivalent:

(1) $P$ satisfies the global Markov property with respect to $G$.

(2) $P$ admits a factorization of the form

$$P(x_1, \ldots, x_{|V|}) = \frac{1}{Z} \prod_{C\in\mathcal{C}} \psi_C(x_C),$$

where $C$ is the set of maximal cliques of $G$, $\psi_C$ are strictly positive potential functions, and $Z$ is a normalizing constant.

*Remark.* This result implies that once a distribution is strictly positive, the conditional independence relations encoded by the graph $G$ are exactly those reflected in its factorization. In our setting, the neighborhood graph $G$ constructed from overlapping partitions plays the role of such an undirected graph. By ensuring that $G$ is fully connected, we guarantee that dependencies can be transmitted along any path of overlapping neighborhoods. Consequently, the latent copula learned from local marginals can be understood as an estimator of the joint distribution that is consistent with the factorization induced by $G$.


AIMS Press