Manifold Embedding Based on Geodesic Distance for Non-stationary Subsurface Characterization Using Secondary Information

Eungyu Park[1*], Jize Piao[2], Hyunggu Jun[1], Yong-Sung Kim[3], Heejun Suk[2], and Weon Shik Han[4]

[1]Department of Geology, Kyungpook National University, 80 Daehak-ro, Daegu 41566, Republic of Korea (ORCID: 0000-0002-2293-4686)
[2]Korea Institute of Geoscience and Mineral Resources, 124 Gwahak-ro, Yuseong-gu, Daejeon 34132, Republic of Korea
[3]INSUNG D&M Inc., 73 Dorim-ro, Guro-gu, Seoul 08312, Republic of Korea
[4]Department of Geosystem Sciences, Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul 03722, Republic of Korea

Correspondence to: Eungyu Park (egpark@knu.ac.kr)

Abstract. In geological characterization, the traditional methods that rely on the covariance matrix for continuous variable estimation often either neglect or oversimplify the challenge posed by subsurface non-stationarity. This study presents an innovative methodology using ancillary data such as geological insights and geophysical exploration to address this challenge directly, with the goal of accurately delineating the spatial distribution of subsurface petrophysical properties, especially, in large geological fields where non-stationarity is prevalent. This methodology is based on the geodesic distance on an embedded manifold and is complemented by the level-set curve as a key tool for relating the observed geological structures to intrinsic geological non-stationarity. During validation, parameters ρ and β were revealed to be the critical parameters that influenced the strength and dependence of the estimated spatial variables on secondary data, respectively. Comparative evaluations showed that our approach performed better than a traditional method (i.e., kriging), particularly, in accurately representing the complex and realistic subsurface structures. The proposed method offers improved accuracy, which is essential for high-stakes applications such as contaminant remediation and underground repository design. This study focused primarily on two-dimensional models. There is a need for three-dimensional advancements and evaluations across diverse geological structures. Overall, this research presents novel strategies for estimating non-stationary geologic media, setting the stage for improved exploration of subsurface characterization in the future.

1 Introduction

The challenge in subsurface investigations, particularly those involving hydrogeology, is to define petrophysical attributes such as hydraulic properties accurately. This challenge is exacerbated in large spatial domains characterised by a high level of heterogeneity (Hewett, 1986; Adams & Gelhar, 1992; Boggs et al., 1992; Hu & Chugunova, 2008; Park et al., 2021). Direct measurements of primary variables, such as permeability, porosity, storage coefficient and dispersivity, often show considerable spatial variability. These variations can be attributed to inherent geological processes, and they make the interpretation of data from single, localised samples an exceedingly complex task.

In addition to direct measurements, secondary data, which are often derived by methods such as geophysical techniques, play a key role in subsurface investigations. Such data provide invaluable insights into the spatial distribution of primary variables, thereby enriching our understanding of the less-explored regions (e.g. Yaramanci et al., 1999; Soupios et al., 2007; Doetsch et al., 2010; Mao et al., 2015). Although the profound importance of secondary data in revealing the spatial variability of subsurface properties has been recognised, the search for an unambiguous methodology to fully exploit its potential is still ongoing.

Traditionally, secondary data have been used in subsurface analysis to provide focused insights. For example, through amplitude interpretation in seismic exploration, correlations between seismic velocities and rock properties were established (Cooper et al., 1965; Rubin et al., 1992; Hyndman et al., 1994; Lumley, 2001; Pride et al., 2003). Similarly, electrical resistivity tomography sheds light on subsurface resistivity; since resistivity is related to pore fluid properties, the subsurface resistivity acts as an indicator of properties such as porosity and water saturation (Kemna et al., 2002; Dietrich et al., 2014). In addition, geological surveys provide a comprehensive picture of the subsurface lithological distributions, which influence variations in hydrogeological parameters (e.g. D'Affonseca et al., 2020). Nevertheless, the precise nature of these correlations remains a subject of active investigation.

Besides these traditional analyses, site investigations, when combined with secondary data analysis, provide ample detailed insights (Batu, 1998; Kerrou et al., 2013). However, there is an ongoing debate suggesting that these findings may be heavily influenced by the individual expertise and heuristic interpretations of the practitioner, rather than a universally accepted methodology. Considering the inherent limitations and uncertainties of secondary information, it is clear that the existing methods cannot fully capture the complex heterogeneities of subsurface stratigraphy, thus highlighting the urgent need for innovative methods to utilise secondary data sets holistically.

In subsurface studies, the intrinsic non-stationarity, which is an unavoidable feature of the real subsurface in practical-scale problems, is a predominant challenge. This problem is especially pronounced when considering the robust spatial correlations of hydraulic properties (Cressie, 1986; Cressie, 1993; Yeh and Liu, 2000; Higdon et al., 2022; Piao and Park, 2023). In conventional geostatistics, we often assume that these spatial correlations are consistent across an entire study area; however, actual observations often contradict this assumption. In large areas subjected to different geological processes, the directions that indicate strong spatial correlations can vary considerably (Piao and Park, 2023). For example, while site surveys mainly provide information on strike and dip directions at shallow depths, advanced techniques such as seismic exploration provide information from greater depths and reveal the more complex spatial interactions of the subsurface. By combining these data sources, we can discover the detailed subsurface heterogeneities shaped by a range of geological processes. However, some of the predominantly used methods, such as cokriging, occasionally cannot these subtleties, particularly when mapping spatial variations. Often these shortcomings are due to fundamental errors such as the assumption of stationarity (Strebelle, 2002).

In the field of hydrogeological characterization, the distinct directionality of both conductive and non-conductive layers plays a crucial role in determining groundwater flow trajectories and the intricacies of solute migration. During analytical considerations, any oversights, data gaps or limitations in estimation techniques can lead to profound misconceptions regarding

flow dynamics and solute dispersion in aquifer systems. These discrepancies are accentuated in multiphase flow situations. These include situations with unsaturated flow (e.g., Suk & Park, 2019), interactions between groundwater and hydrocarbons (e.g., Qin et al., 2007), or situations in which supercritical CO_2 interfaces with brine (e.g., Han et al., 2010). In these complex arenas, small errors in measuring directionality can lead to large consequences, underscoring the acute sensitivity of flow mechanics to nuanced shifts in subsurface properties. Therefore, the careful and accurate delineation of these characteristics is critical for hydrogeological exploration.

In the following discussion, we examine closely the complexities of non-stationarity in covariance-based methods. Recent research focused on this issue, with most techniques relying on the geodesic kernel—as done by Feragen et al. (2015), Jayasumana et al. (2015) and Pereira et al. (2022)—and kernel convolution methods, as reported by Higdon (1998), Higdon et al. (2022), Paciorek (2003) and Fouedjio et al. (2016). These cutting-edge techniques represent a notable departure from classical geostatistical methods, such as kriging. The unique strength of these new techniques lies in providing an adaptive framework that seamlessly integrates spatially variable statistics. This adaptability stems from the integration of Riemannian manifolds or tailor-made non-stationary covariance functions. Hence, these methods excel at capturing and modelling the nuances of directional variation in spatial associations—a task that traditional kriging struggles to accomplish with the same finesse.

Recently, innovative methods designed to capture complex spatial variations have been applied in hydrogeological studies. For example, Piao and Park (2023) used the intrinsic geometry in manifold embedding customised for non-stationary field characterization in hydrogeology. Their study highlighted the importance of understanding of the nuances in the distribution of hydraulic properties. In particular, the use of non-stationarity enhanced spatial field estimates led to drastic improvements in estimation accuracy.

Notably, the existing literature lacks explicit guidance on the construction of manifolds in a geological context. The processes of creating manifolds and extracting associated geometric information from secondary data are of paramount importance. Such processes provide a crucial link between secondary geological data and practical estimation techniques. Without this crucial link, the full potential of these methods cannot be realised, resulting in a significant gap in non-stationary field characterization. In this study, a basic framework for using secondary data is developed via a contextual approach. By extracting spatial insights from these data, we provide a perspective for estimating primary variables, especially in areas that are characterised by pronounced variations in directional dependence accross spatial locations. Here, we emphasise that this study was not aimed at using secondary information at specific locations to improve estimates based on correlations between primary and secondary data, as is done traditionally. Instead, this study proposes a more judicious and strategic use of secondary data in subsurface characterization within a geological context.

2 Targeted secondary data for analysis

Section 2 reviews the specific types of secondary data that extremely important for the proposed methodology. The emphasis is on the data that provide directional information about spatial correlations that are inherently non-stationary. Such data types

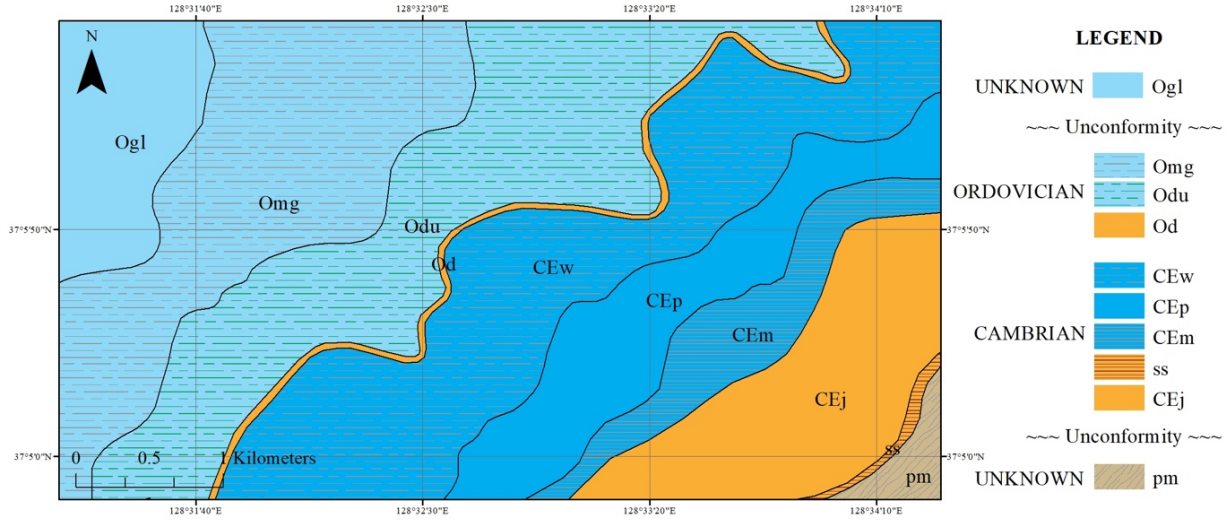
provide insight into the intricacies of geological formations and their variable orientations, thereby improving our understanding and challenging the traditional geostatistical paradigms.

Geological observations often tend to have non-stationary correlations. This nature of geological observations is incompatible with the traditionally accepted assumption of stationarity in the conventional geostatistical methods such as kriging. To explain this, see the geologic map in Fig. 1(a) showing surface sedimentary rock formations, commonly known as folds, with distinct spatial variations in orientation. This map shows the Joseon Supergroup within the Taebaeksan Basin in Gangwon-do, South Korea. The dominant features seen in the map are the lower Palaeozoic (Cambrian–Ordovician) sedimentary and metasedimentary rocks such as sandstone, shale, limestone, dolomite, quartzite and slate (see Son & Lee, 1966; Choi et al., 2016).

From Fig. 1a, we see that the sedimentary formations have a clear stratigraphic sequence, indicating their sequential deposition. Over geologic time scales, these formations experienced various tectonic forces and diagenetic processes, resulting in their current surface expressions. The boundaries of these formations, which have uneven thickness and intricate patterns, provide important insights into the dynamics of their depositional environments.

Such boundary patterns are important for inferring subsurface petrophysical properties. The continuity along these boundaries suggests that depositional environments remained relatively consistent in a direction parallel to these boundaries. This implies a greater degree of similarity in lithological and petrophysical properties along these lines. Conversely, as deposition progressed in a direction perpendicular to these boundaries, variations in the depositional conditions become more pronounced. Thus, we can hypothesise that the spatial correlation of petrophysical properties is more extensive along the depositional strike (parallel to the formation boundary). In contrast, this correlation decreases more rapidly when examined in the normal direction to the boundary.

(a)



(b)

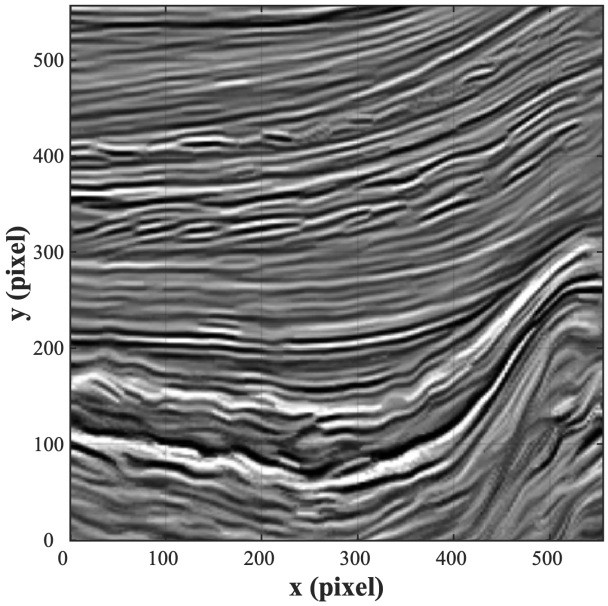


Figure 1. **Geologic and geophysical representations of non-stationary spatial correlations: (a) A part of the geologic map of the Joseon Supergroup within the Taebaeksan Basin in Gangwon-do, South Korea, showing sedimentary rock formations with varying orientations. (b) The seismic profile of the F3 block in the southern North Sea, showing subsurface horizons with non-stationary orientations over a span of 5 km in the x-direction and 0.88 s (two-way travel time) in the y-direction. Reflectors show varying dips, indicating a geologic structure influenced by the underlying Permian salt dome intrusion.**

Similarly, Fig. 1(b) shows a seismic profile where the horizons represent non-stationary orientations. This seismic profile represents the 691st inline of a three-dimensional (3D) seismic profile acquired in the F3 block in the southern North Sea. The extracted profile spans 5 km in the x-direction and 0.88 s (two-way travel time) in the y-direction. The pixel spacing is approximately 9 m in the x-direction and approximately 0.0015 s in the y-direction. The data-acquisition site is characterised by an anticlinal dip of the upper sedimentary layers because of the underlying Permian salt dome intrusion, resulting in a complex geologic structure. In the profile, the right part shows steeply dipping reflectors, while the dip decreases towards the upper layers (Schroot & Schüttenhelm, 2003). In contrast, the left part has relatively flat reflectors. In general, the petrophysical properties are similar within a given reflector; however, there are discrepancies among different layers, implying that physical proximity does not necessarily correspond to similarity in petrophysical properties. When predicting subsurface properties from limited geophysical data such as well logs, it is essential to consider the subsurface reflection structure. Modern image processing techniques, such as those explored (in part) in this study, can effectively delineate these features.

In an undisturbed geologic environment in the absence of major deformation such as faults or plutonic intrusions, hydraulic properties often follow the direction of sedimentary deposition. This orientation results from the consistent geologic processes in similar depositional and diagenetic environments over time. The key indicators of this orientation include strike directions determined from field surveys, lithologic boundaries on geologic maps and seismic reflectors in exploration data. These markers provide the secondary information that is essential for geostatistical estimators when inferring petrophysical properties.

The main hypothesis of this study is that petrophysical properties are predominantly correlated in the sedimentary depositional direction, especially in aquifers with sedimentary matrices, regardless of their consolidation state. The objective of this study was to identify those linear features from secondary data that help elucidate the geologic layering and estimate primary petrophysical properties such as permeability. It is imperative to recognise that these boundary orientations can vary because of deformation, indicating directional non-stationarity influenced by specific geologic conditions during and after deposition. Therefore, the study of these linear features can enhance our understanding of geological formations and improves the accuracy of hydrogeological evaluations. This paper describes a method to obtain this information and incorporate it into the estimation of the primary variable.

3 Theory

3.1 Geostatistical implications of the manifold

In geostatistics, the manifold is not merely a geometric construct; it is a fundamental tool for understanding spatial variation because it provides a framework for capturing spatial non-stationarity in the directions of correlation. Conceptually, the manifold provides a blueprint—a structured representation—that delineates the intricate spatial shifts in correlation within a given domain.

Consider a manifold as a surface residing in the XYZ-space, parameterised by the uv-plane. This relationship can be represented mathematically by as

$$\begin{aligned} X &= u \\ Y &= v \\ Z &= f(u,v) \end{aligned}, \tag{1}$$

where the function $f(u,v)$ captures the intricate shape and undulations of the manifold, encapsulating spatial heterogeneities in subsurface parameters. The uv-plane provides a base reference—a standard metric space—upon which the nuances of a manifold are projected or resolved. X and Y correspond directly to u and v, respectively, and the expression $Z = f(u,v)$ captures the depth of a manifold, highlighting the spatial nonstationarities in correlation orientations.

Figure 2 shows two representations of the manifold concept. Figure 2(a) shows a 3D visualisation of the manifold. The grey surface delineates the manifold. Superimposed on this manifold is a blue surface, which represents a specific level. The intersection of this level with the manifold defines a curve that is referred to as the level-set curve (in Sect. 3.2), analogous to geological strikes. The magenta line indicates the direction of the level-set curve (interchangeably, strike in the geological context) on the manifold, and the green line indicates the dip direction, which is aligned with the steepest gradient (i.e., dip), regardless of ascending or descending dip.

Figure 2(b) shows the two-dimensional (2D) projections of linear segments, indicating strike and dip directions at designated measurement locations corresponding to Figure 2a. The red line in this 2D projection corresponds to the magenta line in the 3D visualisation, representing the strike direction. In contrast, the blue line in the 2D projection mirrors the green line in the

3D visualisation, indicating the direction of the gradient at a given manifold point. Note that the length of the blue line in the projection appears shortened because of the angle between the manifold and the actual 2D plane. In the most extreme case, when the manifold is oriented vertically, this blue line indicating the dip direction may appear to have virtually zero length in the projection.

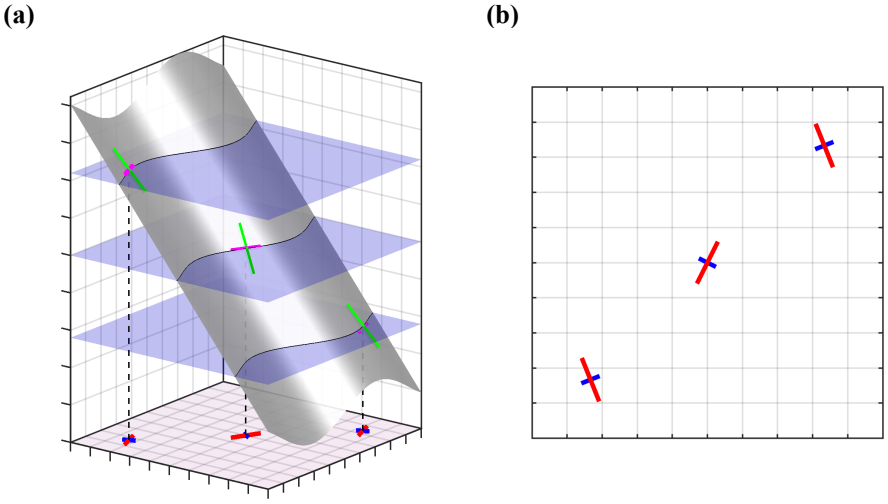


Figure 2. Conceptual illustrations of the manifold in geostatistics: (a) A three-dimensional (3D) representation of the manifold (grey surface) with a given level (blue surface). The intersection of the manifold and the level forms a level-set curve. The magenta line indicates the direction of the level-set curve (or geological strike) on the manifold, while the green line indicates the dip direction, aligned with the steepest gradient; (b) two-dimensional (2D) projections of the linear segments corresponding to the strike (red line) and dip (blue line) directions at specific measurement locations related to (a). The change in length of the blue line in the 2D projection reflects the angle between the manifold and the actual 2D plane.

The level-set curve, suggesting the locations of geological formation boundaries, offers significant insight into the geometry of the manifold and its implications for spatial estimation. In the visual representation, manifold measurements in 3D are projected onto a 2D plane. However, in practical applications, the process operates in reverse: information is gathered from a 2D plane (i.e., secondary data) and utilised to construct the 3D manifold. As discussed in the following sections, the orthogonality theorem suggests a perpendicular relationship between the direction of this curve and the gradient of the manifold. This relationship suggests that by understanding the direction of the level-set curves, one can understand the geometric complexity of a manifold. The central goal of this study was to extract information about manifold geometry from secondary data sources (e.g., strikes) in the 2D plane and use these data to reconstruct the 3D manifold for spatial estimation.

3.2 Level-set curve of a manifold

To further appreciate the importance of the manifold in geostatistics, it is imperative to analyze the concept of the level-set curve. Consider a scalar function $f: \mathcal{M} \to \mathbb{R}$. The level-set curve, denoted as L_c, for a given value c is defined as

$$L_c = \{\mathbf{p} \in \mathcal{M} : f(\mathbf{p}) = c\}, \quad (2)$$

where $\mathbf{p}$ represents a point on the manifold, $\mathcal{M}$. The level-set curve consists of those points on the manifold for which the function f attains a constant value of c. Given the spatial representations of a manifold, this curve serves as an essential reference that captures constant-value contours on the manifold and thus has substantial relevance in geostatistical estimates.

3.3 Spatial correlation in the context of the level-set curve

In spaces defined by a smooth manifold, $\mathcal{M}$, the concept of geodesic distance becomes paramount. This distance represents the shortest path between two points on a curved surface and differs from the traditional Euclidean distance used on flat surfaces (conventional geostatistics).

Mathematically, given two points $\mathbf{p}$ and $\mathbf{q}$ that are members of $\mathcal{M}$, their geodesic distance, denoted as $d_g(\mathbf{p}, \mathbf{q})$, can be written as

$$d_g(\mathbf{p}, \mathbf{q}) = \inf\left\{\int_0^1 \|\gamma'(\lambda)\| d\lambda : \gamma: [0,1] \to \mathcal{M}, \gamma(0) = \mathbf{p}, \gamma(1) = \mathbf{q}\right\}, \quad (3)$$

where the term 'inf' stands for infimum, or the greatest lower bound.

The key insight here is that when the conventional Euclidean distance is consistent in every direction, the geodesic distance is minimum in the direction of the level-set curve. Specifically, this occurs when both points, $\mathbf{p}$ and $\mathbf{q}$, lie on this curve. This phenomenon suggests that the strongest spatial correlation is found in the direction of the level-set curve.

This understanding emphasises the central role of the level-set curve in geostatistical evaluations. Recognising that points on this curve share a heightened spatial affinity allows for the more accurate modelling and analysis of spatial patterns.

3.4 Utilising secondary data for insights

In the absence of significant deformation such as joints, faults or plutonic intrusions, hydraulic properties are typically assumed to extend primarily in the horizontal direction of the formation. This tendency is attributed to the fact that these properties develop simultaneously under similar geologic conditions. Because of this inherent characteristic, secondary data become increasingly important in geostatistical analysis. This is especially true when identifying linear structures that represent the most spatially correlated direction at a given location. These structures often reflect the underlying geologic processes and their resulting spatial patterns. As mentioned in Sect. 3.3., this direction coincides with that of the level-set curve. A closer examination shows that secondary data can reveal these important spatial orientations, making in-depth geostatistical studies possible.

In Sect. 4, we examine in greater detail the process of determining the most spatially correlated direction using secondary data and discuss real-world examples based on geologic maps and seismic profiles.

3.5 Deriving manifold gradient from secondary data

Secondary data holds the potential to provide vectors indicating the maximum correlation directions. Using the notation ∇f to represent the gradient of f on the manifold $\mathcal{M}$, the direction corresponding to the highest spatial correlation at a given point, $\mathbf{p}$, is discerned to be orthonormal to $\nabla f(\mathbf{p})$. This assertion is based on a cardinal mathematical theorem that states that if a function f is differentiable, its gradient at a point will either be zero or perpendicular to its level set at that point.

Thus, the direction of the level-set curve is perpendicular to the direction of the gradient on a manifold. Specifically, the gradient of f at point $\mathbf{p}$ is orthogonal to the level-set curve, L_c, at the same point. This relationship can be expressed mathematically for any vector, $\mathbf{w}$, tangent to L_c at point $\mathbf{p}$ as

$$\nabla f(\mathbf{p}) \cdot \mathbf{w} = 0. \tag{4}$$

These orthogonal relations are further encapsulated by the equations:

$$\frac{\partial L_c}{\partial v} = -\frac{\partial f}{\partial u} \text{ and } \frac{\partial L_c}{\partial u} = \frac{\partial f}{\partial v}. \tag{5}$$

The orthogonal relationship in Eq. (5) indicates that one component of the gradient is positive, while the other is negative—indicating that they are perpendicular.

By recognising and internalising this mathematical interplay, a robust framework can be developed to decipher the spatial nuances of a manifold. Thus, geostatistical evaluations become more thorough and insightful.

3.6 Challenges in manifold reconstruction

Reconstructing a manifold involves more than determining the direction of the gradient, ∇f. It also requires understanding the magnitude of the variation. While secondary data provide profound insights, they occasionally lack the granularity or precision required for meticulous manifold reconstruction. To overcome these challenges, often iterative strategies or complementary methods are required.

Although the orthogonality between the level-set curve and the gradient of a manifold at a given point is mathematically defined in Eq. (5) when f is given, for manifold reconstruction from the level-set curve, especially when information is constrained owing to the lower dimensionality of the level-set curve compared to the manifold, the following modification is required:

$$\left(\frac{\partial f}{\partial u}\right)_{est} = -\beta \frac{\partial L_c}{\partial v} \text{ and } \left(\frac{\partial f}{\partial v}\right)_{est} = \beta \frac{\partial L_c}{\partial u}, \tag{6}$$

where β is introduced as a scalar factor whose exact value is undetermined in the theoretical context, and $(\partial f/\partial u)_{est}$ and $(\partial f/\partial v)_{est}$ are the estimated manifold gradients in the directions of u and v. The inclusion of β provides flexibility, allowing the model to capture manifold intricacies by adjusting the magnitude of the gradient based on empirical evidence. Thus, β becomes an essential parameter, and it provides the necessary degree of freedom to ensure accurate manifold reconstructions from limited data.

For an impact of this parameter to the estimated field and the specifics of our empirical approach, which involves determining the reasonable β values for different cases, readers are directed to Sect. 5.

3.7 Gaussian process regression based on geodesic kernel

In the domain of spatial structure derivation and geodesic distance computation, statistical methodologies are indispensable to make use of extracted spatial nuances. Although several estimation techniques can be used considering the spatial affinity embedded within manifolds, Gaussian Process Regression (GPR) is a suitable choice, especially when coupled with a kernel tailored to geodesic distances. The GPR can be used with the traditional kriging methods, ensuring consistency with the existing practices.

Departing from the realms of traditional kriging and GPR, our proposed methodology offers a unique space to accommodate the non-stationary spatial distributions inherent in variable estimation.

Our methodology is based is the role of the kernel in assessing the similarity of data points. If geological constructs are regarded as manifolds, the geodesic kernel is a quintessential measure of such similarities. Given two manifold points, $\mathbf{p}_i$ and $\mathbf{p}_j$, the geodesic kernel can be expressed as

$$k(\mathbf{p}_i, \mathbf{p}_j) = \exp\left(-\frac{d_g(\mathbf{p}_i, \mathbf{p}_j)^2}{2\rho^2}\right), \tag{7}$$

where $d_g(\mathbf{p}_i, \mathbf{p}_j)$ represents the geodesic distance between the two points, and ρ (similar to β) is a critical, yet undetermined parameter that defines the correlation scale. Methods to determine β and ρ will be discussed in the Results and Discussion section (Sect. 5).

The geodesic kernel is adopted considering the nuanced spatial relationship inherent in geological structures. By anchoring the kernel to geodesic distances, the similarity measure is inherently aligned with the manifold geometry, thereby enhancing the predictive power of GPR.

The introduction of the geodesic kernel ensures that the similarity measure closely follows the innate geometry of the manifold. This alignment optimises the predictive power of GPR. Once the geodesic kernel is computed, it can be seamlessly integrated into the GPR framework. Thus, we obtain a model that accurately captures spatial dependencies in geological structures, addressing both broad patterns and fine-grained variations.

The equation used for internal estimation in this model is

$$\mathbf{z}^* = \mathbf{\Omega}\mathbf{z}_{obs} = \mathbf{\Sigma}_{ab}(\mathbf{\Sigma}_{bb} + \sigma^2\mathbf{I})^{-1}\mathbf{z}_{obs}, \tag{8}$$

where $\mathbf{z}^*$ represents estimates of the primary variable at all sampled locations $(Z_{o1}, \cdots, Z_{oK})$; $\mathbf{z}_{obs}$ is a vector of N observed values of the primary variables; $\mathbf{\Omega}$ denotes a $K \times N$ weight matrix; $\mathbf{\Sigma}_{ab}$ and $\mathbf{\Sigma}_{bb}$ represent $K \times N$ and $N \times N$ covariance matrices, respectively, representing relationships of unknown–known points and known–known points, respectively. In the equation, $\sigma^2\mathbf{I}$ with dimension $N \times N$ introduces regularisation into the estimation, and the constant σ^2 refers to the reliability of the primary data observations, accounting for possible measurement errors.

In the future, we will focus on optimising ρ and exploring multi-kernel methods to improve the model performance. Combining the geodesic kernel with GPR itself represents a notable step in geostatistical analysis, paving the way for advanced research and real-world applications (Feragen et al., 2015; Jayasumana et al., 2015; Pereira et al., 2022; Piao & Park, 2023).

3.8 Calculation of geodesic distance

In geostatistical studies, especially in manifold learning, it is essential to compute the geodesic distance between two points on a manifold. This distance represents the shortest path between two given points on a curved surface and can be regarded as the equivalent of a 'straight line' distance in Euclidean spaces. In a geological context, this distance is critical because it characterises the spatial relationships between different geological processes or structures.

To compute the geodesic distance between two points $\mathbf{p}_i$ and $\mathbf{p}_j$ on the manifold, described by the coordinates (u_i, v_i) and (u_j, v_j), respectively, the following formula is used (Piao & Park, 2023):

$$d_g(\mathbf{p}_i, \mathbf{p}_j) = \int_0^1 \sqrt{(u_j - u_i)^2 g_{11} + 2(u_j - u_i)(v_j - v_i) g_{21} + (v_j - v_i)^2 g_{22}}\, d\lambda\,. \tag{10}$$

Here, g represents the metric tensor of the manifold, and it is a function that captures how distances vary across the manifold as a function of direction. More specifically, the metric tensor is defined as follows:

$$g = \begin{bmatrix} 1 + \left(\frac{\partial f(u,v)}{\partial u}\right)_{est}^2 & \left(\frac{\partial f(u,v)}{\partial u}\right)_{est} \left(\frac{\partial f(u,v)}{\partial v}\right)_{est} \\ \left(\frac{\partial f(u,v)}{\partial u}\right)_{est} \left(\frac{\partial f(u,v)}{\partial v}\right)_{est} & 1 + \left(\frac{\partial f(u,v)}{\partial v}\right)_{est}^2 \end{bmatrix}. \tag{11}$$

The components g_{11}, g_{21} and g_{22} of the metric tensor are crucial to understanding the intrinsic geometry of the manifold, and they describe how distances change as a function of direction on the manifold surface. The geodesic distance is determined using Eq. (10) and Eq. (11) by numerical integration. In this study, Legendre–Gauss quadrature with 20 abscissa and weights were adopted to improve the accuracy and numerical efficiency. For more details of the equations and the derivation, see Piao and Park (2023).

By systematically applying the formulas given in Sect. 3.6–3.8, an approximate measure of geodesic distance on geological manifolds can be obtained. Such calculations pave the way for richer insights into the spatial intricacies of manifolds, thereby enhancing our understanding of geological structures and processes.

4 Method

4.1 Deriving spatial structures from supplementary data

This subsection describes the methods used to obtain the key ancillary data essential to this study. Secondary data are presented mainly in the form of imagery, particularly, geologic maps and seismic profiles. In addition, information derived from

geological surveys, particularly, the strike and dip directions of formations, is invaluable and is incorporated as ancillary data. From these sources, we derive an in-depth understanding of the spatial dependencies present in complex geological settings. The emphasis is on linear features as they inherently capture the directional correlation of the spatial distribution, which is critical to understanding the spatial coherence of petrophysical attributes.

From the provided data, linear segments along geological structures that indicate spatial coherence (namely, the level-set curves and directions), are extracted. For this extraction, sampling points within an image are identified, and windows centred on each of these points are created for information extraction. Although several sampling methods, such as pure random, grid, and Latin hypercube, are available, the choice is often a matter of preference. In this study, a quasi-random sampling method, namely, the Sobol sequence, was used. The window size, which is essential for extracting information precisely, is determined empirically. It is essential to ensure that the number of sampling points does not inadvertently introduce redundancy and that the spatial variability in the image is considered. Hence, the window size (s_{win}) was determined to be the ceiling value of $\gamma \times \min(n_u, n_v)$, where γ was set as 0.025 in this study. Here, n_u and n_v are the number of pixels in the horizontal and vertical directions, respectively. The total number of sampling points, symbolised by n_{sam}, where the window was applied and the linear features were detected, is expressed as

$$n_{sam} = \frac{2n_u n_v}{{S_{win}}^2}. \tag{12}$$

Within each specified sampling window, a systematic approach consisting of four key tasks was adopted to identify any salient linear feature, if present:

(1) **Enhancement of edge structure in images**: Given the heterogeneity in the quality of images derived from secondary data, techniques that enhance the inherent linear features are essential. The Canny edge detection method is particularly effective for this purpose. It can be implemented using the edge function in MATLAB's Image Processing Toolbox.

(2) **Digitisation of linear features**: After the edge enhancement phase, the Hough transform is adopted to convert the delineated edges into distinct linear segments that are confined within the specified sampling windows. The segment selection procedure is based on MATLAB's Hough transform capabilities (the 'houghlines' function) to locate the start and end points of these linear trajectories.

(3) **Calculation of slopes for directional representation**: Individual linear segments in isolation may have no intrinsic value for this study, as the primary interest is in the representative direction for each window. Therefore, only the slopes of the identified linear segments within a window were retained for further processing. For delineating linear features, only lines exceeding a length of $s_{win}/2$ were considered. The coordinates of the start and end points of each lineament are denoted as (u_k^s, v_k^s) and (u_k^e, v_k^e), respectively, where k ranges from 1 to K (where K is the cumulative number of linear segments in the sampling window). Then, the slope of the detected linear segment, represented as α_k, can be calculated using the following formula:

$$\alpha_k = \frac{v_k^s - v_k^e}{u_k^s - u_k^e}. \tag{13}$$

(4) **Determination of the representative slope**: Given the inherent variability in image quality and clarity, multiple slopes indicative of different orientations may be detected within a single window. Such diversity can lead to inconsistencies. Hence, it is necessary to have a single, representative slope that adequately captures the underlying directionality. To address this concern, median slopes were selected from the set of calculated slopes, considering the inherent robustness of median slope and its ability to remain unaffected by outliers. Consequently, a median is extracted from all α_k values to denote a representative slope, α, for that particular sampling window. Knowing α, the angle of structural inclination, θ, can be derived as $\theta = \arctan(\alpha)$. Next, the partial derivatives of the level-set curve direction with respect to u and v, symbolised as $\partial L_c/\partial u$ and $\partial L_c/\partial v$, respectively, are defined as follows:

$$\frac{\partial L_c}{\partial u} = \cos(\theta) \text{ and } \frac{\partial L_c}{\partial v} = \sin(\theta)\,. \tag{14}$$

4.2 Conversion from localised level-set curve direction to manifold gradient fields

First, at the discrete sampled locations, the gradients $(\partial f/\partial u)_{est}$ and $(\partial f/\partial v)_{est}$ can be obtained directly using Eq. (6). This equation provides a straightforward means of conversion based on the orthonormal relationship between the detected level-set curve direction and its corresponding manifold gradient at these specific points.

However, there are challenges in obtaining a comprehensive representation over the entire domain, especially when the gradients are insufficient. Equation (6) facilitates the computation for the sampled locations; the geodesic distances, from Eq. (10) and Eq. (11), require that these gradients should be uniformly distributed over the entire domain. To overcome this problem, the gradient fields must be interpolated over the entire spatial extent. This process, which is based on the locally detected level-set curve directions, aims to provide a seamless and continuous gradient representation that reproduces all local intricacies.

Hence, GPR (Sect. 3.7) is used in this study. GPR can handle uneven data sets and hence is a suitable tool for this interpolation task. GPR appreciates the spatial interplay between points, focusing on the Euclidean distance while omitting complicated manifold subtleties. In particular, parameter ρ from Eq. (7) is set empirically by considering the structural variations of an image.

4.3 Measure of structural similarity

Given two images, **A** and **B**, with matching dimensions $n_u \times n_v$, their similarity can traditionally be quantified by comparing the basic statistical parameters such as mean and variance. Typically, the correlation coefficient between **A** and **B** is used for this comparison.

However, methods that focus exclusively on direct pixel-by-pixel comparisons may not fully capture the true similarity between the estimated and the actual image. To overcome this shortcoming, a more sophisticated metric named the structural similarity index measure (SSIM) was developed (Wang et al., 2004).

The SSIM was designed to evaluate the perceptual quality of images, but it goes beyond mere numerical disparities. the SSIM emphasises the local patterns of pixel intensities within images, more closely aligning with the perception of the human visual system. The SSIM is expressed mathematically as

$$\mathrm{SSIM}(\mathbf{A},\mathbf{B}) = \frac{(2\mu_{\mathbf{A}}\mu_{\mathbf{B}} + C_1)(2\sigma_{\mathbf{AB}} + C_2)}{(\mu_{\mathbf{A}}{}^2 + \mu_{\mathbf{B}}{}^2 + C_1)(\sigma_{\mathbf{A}}{}^2 + \sigma_{\mathbf{B}}{}^2 + C_2)}, \tag{15}$$

where $\mu_{\mathbf{A}}$ and $\mu_{\mathbf{B}}$ are the means of **A** and **B**, respectively; $\sigma_{\mathbf{A}}{}^2$ and $\sigma_{\mathbf{B}}{}^2$ are the variances of **A** and **B**, respectively; $\sigma_{\mathbf{AB}}$ is the covariance of **A** and **B**; and C_1 and C_2 are constants to avoid instability (both are unity in this study).

The SSIM defined in Eq. (15) ranges from −1 to 1. An SSIM value of 1 indicates that the two images being compared are perceptually identical. Conversely, a value of −1 indicates complete structural dissimilarity. In most real-world scenarios, the SSIM fluctuates between 0 and 1: values closer to 0 indicate less structural similarity, while those closer to 1 indicate higher similarity. Therefore, higher SSIM values indicate better perceptual quality when two images are juxtaposed.

5 Results and discussion

5.1 Geological map as secondary data

The geologic map shown in Fig. 1(a) delineates the formation boundaries of the constituent geologies (Sect. 2). These boundaries provide a deeper understanding of the stratigraphic order, which is fundamental to recognising the hydrogeologic properties of sedimentary formations. In such terrains, correlation scales tend to be longer along bedding planes than in perpendicular directions because of the uniform depositional conditions that prevailed. Hence, nearby points within a single lithostratigraphic unit formed during an identical geochronological span tend to exhibit analogous petrophysical properties. Notably, angular unconformities are not considered in this context. Instead, all boundaries are assumed to have been parallel to the overlying and underlying formations to some degree during deposition.

Building on this understanding, we can extract directional information by applying the method described in Sect. 4.1. Figure 3 shows the results of boundary identification combined with the derived tangential slopes at selected locations, effectively highlighting the detected geologic boundaries from the geologic map shown in Fig. 1(a). In addition, the tangential lines associated with these boundary curves are shown as line segments. During delineation, the window covered an area of 6 × 6 pixels, and the total number of sampling points was 3730, considering n_u and n_v as 308 and 218 pixels, respectively. These identified linear features act as indicators of the directions of the level-set curve at their respective sampling points. These features can be interpreted as the gradient of a manifold describing the directional shifts of spatial correlations throughout the modelling domain.

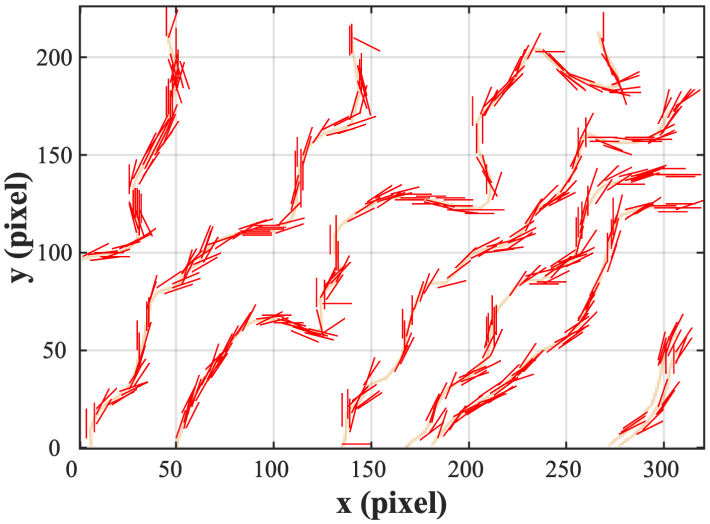
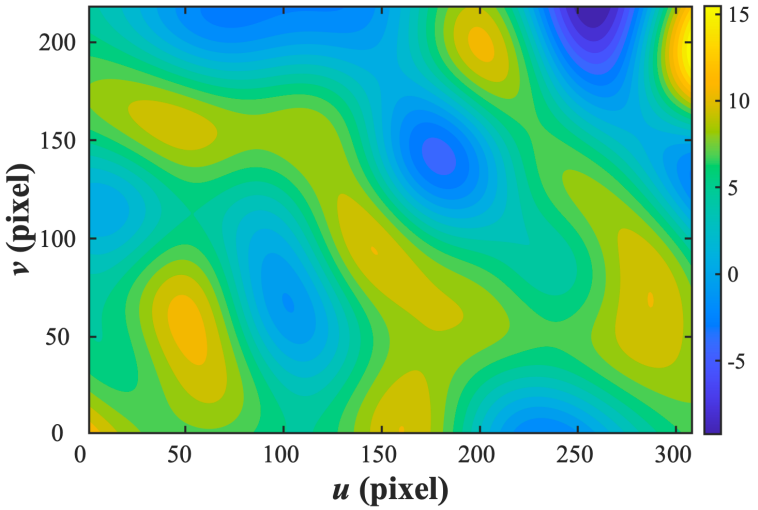


Figure 3. Boundary identification results paired with the tangential gradients derived at selected locations showing the detected geologic boundaries and associated tangential lines. The linear features indicate the level-set curve directions at their corresponding sampling points, representing the gradient of a manifold representing directional shifts of spatial correlations.

Figure 4 shows the interpolated gradients, labelled as $(\partial f/\partial u)_{est}$ and $(\partial f/\partial v)_{est}$, with ρ as 40 pixels. This value, $\rho = 40$ (Eq. 7), was empirically determined considering the general trend of lithological boundaries (Sect. 4.2). In general, a smaller ρ value in GPR can capture more detailed structures at the cost of overfitting. To compensate for the potential inaccuracies in linear feature delineation, we adopted a σ^2 value of 1×10^{-1} (Eq. 8) during the interpolation process. Observations from the figure indicate pronounced gradient variations throughout the domain, suggesting the potential non-stationarity and directional oscillations in spatial correlations. In contrast, the conventional methods such as kriging, which are based on Euclidean principles, theoretically produce zero values for both $\partial f/\partial u$ and $\partial f/\partial v$. The geodesic distance calculated using Eq. (10) and Eq. (11) were based on these interpolated gradients, $(\partial f/\partial u)_{est}$ and $(\partial f/\partial v)_{est}$.

(a)



(b)

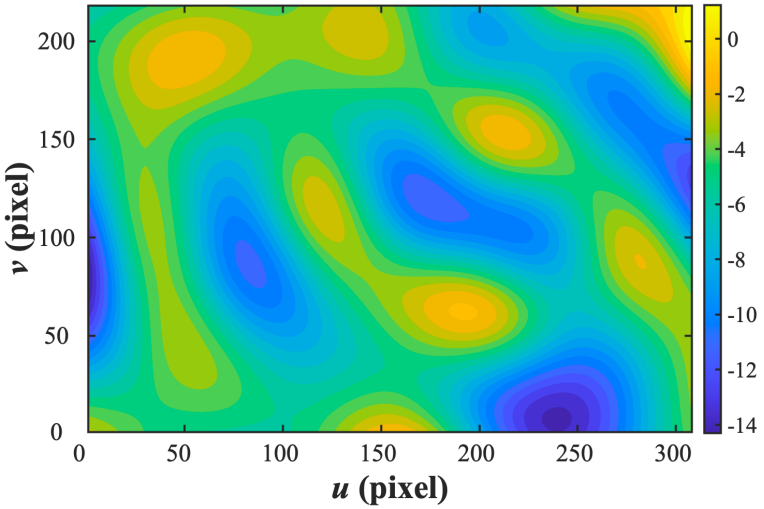


Figure 4. Interpolated gradients labelled $(\partial f/\partial u)_{est}$ and $(\partial f/\partial v)_{est}$ for a ρ value of 40 pixels. (a) and (b) show the estimated gradient fields in the u and v directions, respectively.

In light of the calculations discussed, the unconditionally simulated results shown in Fig. 5 provide a credible representation of the hydraulic conductivity distribution over the region shown, informed by its geologic background. This plot is based on 50 data points that are randomly distributed and follow a normal distribution at random locations (indicated by white dots with black borders). These markers represent the log-transformed hydraulic conductivities with the log-transformed mean and variance both equal to 1. The geodesic kernel GPR methods explained in Sect. 3.7, based on manifold embedding, were used to process these figures. The resulting hypothetical predictions are consistent with the discernible non-stationary directional patterns seen in the ancillary geologic map data (Figure 1a) and reflect a sedimentary basin setting. Meanwhile, the derived results agree with the factual data sets, such as hydraulic conductivities derived from aquifer evaluations, at selected monitoring

sites. During this computational exercise, ρ was set to 100 pixels, and β was fixed at 10—consistent with Case 1 (refer to baseline scenario), to facilitate subsequent juxtapositions with alternative ρ and β configurations.

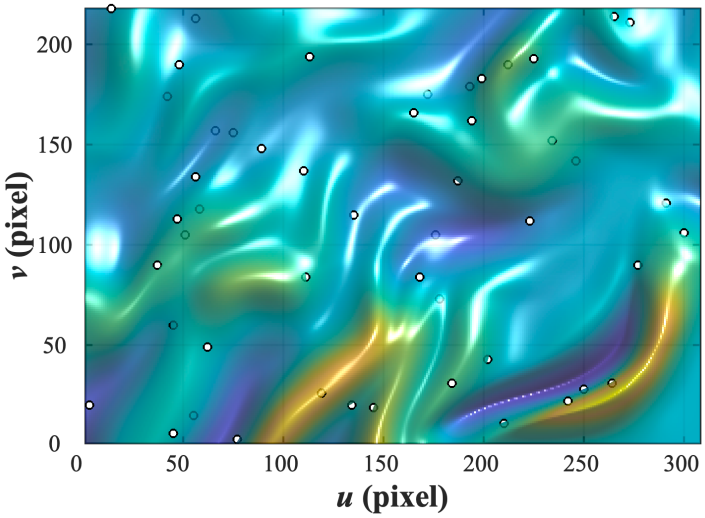


Figure 5. Unconditionally simulated representation of the petrophysical property (e.g., hydraulic conductivity) distribution over the studied region (Fig. 1a), based on geologic secondary information. The estimated distribution is based on 50 data points, represented by white dots with black outlines.

From the analytical exploration, results of two additional test scenarios were evaluated: $\rho = 150$ and $\beta = 10$ (Case 2) and $\rho = 100$ and $\beta = 5$ (Case 3), both shown in Fig. 6. The figure shows that a higher ρ value (Case 2), which corresponds to the correlation scale used in traditional geostatistical techniques such as kriging, results in reduced resolution of the structural variations compared to Case 1. This observation is consistent with the understanding that longer correlation-scale structures may lack the finer scale variations inherent in the embedded manifold. In Case 3, a lower β value is correlated with reduced magnitudes of $(\partial f/\partial u)_{est}$ and $(\partial f/\partial v)_{est}$. This reduction suggests that geometrical subtleties of the manifold are not adequately reflected in the generated estimates. Thus, β inherently measures the degree to which the variability of the manifold affects the estimation. At its limit, as β approaches zero (i.e., a flat manifold obtained by nullifying Z in Eq. (1)), the result will reproduce the isotropic correlation scale results that are typical of the conventional kriging. From Case 3, it is evident that β plays a critical role in determining the alignment of the estimates with the embedded manifold.

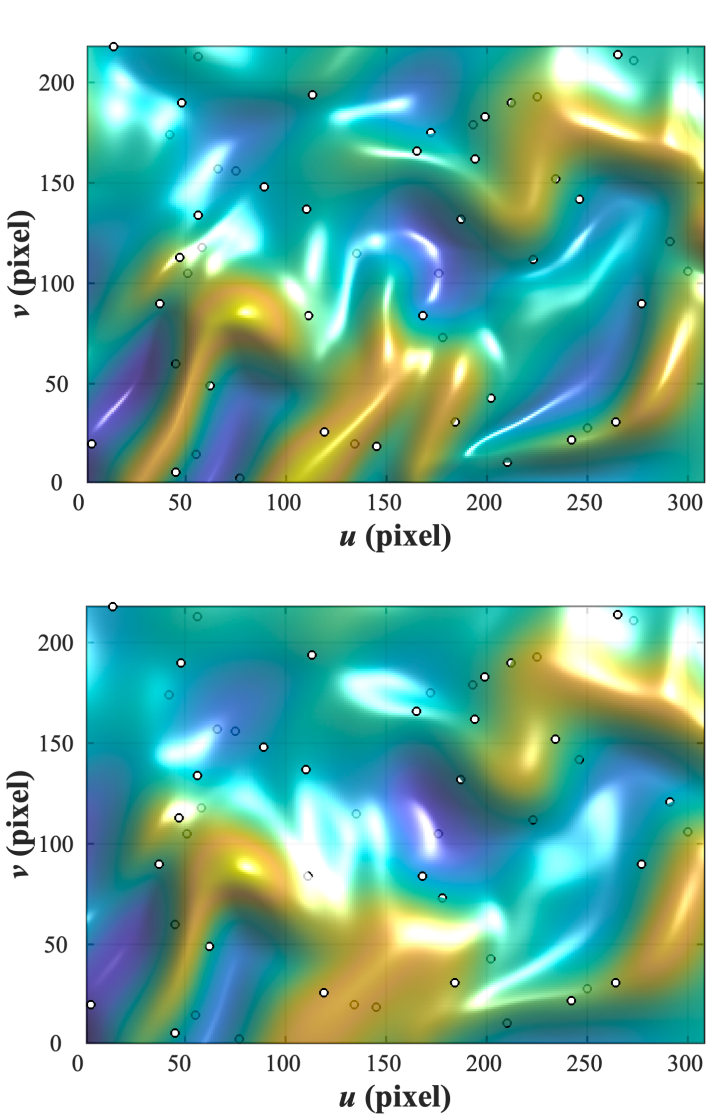


Figure 6. Comparatively simulated results for hydraulic conductivity distribution informed by the geologic background. Two test scenarios are presented: (a) Case 2 with parameters $\rho = 150$ and $\beta = 10$, and (b) Case 3 with parameters $\rho = 100$ and $\beta = 5$. The white dots with black borders indicate the 50 random data used in Case 1.

The parameters ρ and β are elucidated as the integral determinants of the morphological configuration of the projected field. It is necessary to calibrate them carefully to implement our proposed algorithmic approach. The adoption of cross-validation methods, as outlined by Piao and Park (2023), was proposed as a feasible strategy to delineate these empirical coefficients.

In the field of continuous variable estimation, non-stationary estimation has been elusive historically; most such estimations were based on zoning methods. However, the conventional zoning approach, which assumes null correlation between different zones, may be ineffective when variables across zones exhibit correlations. The proposed method is as an effective alternative under such circumstances. In particular, while geological processes leading to non-stationary correlation orientations were

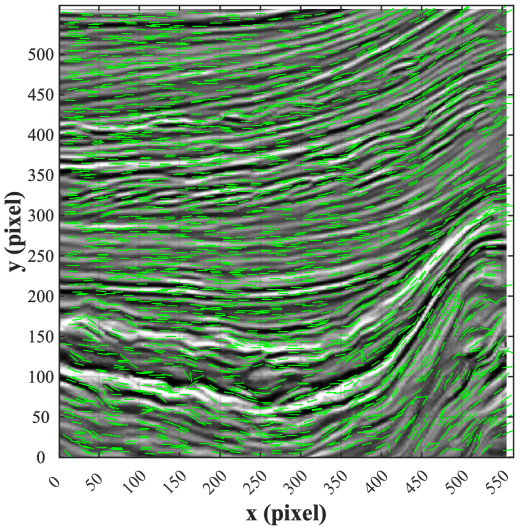
recognised, they were mostly ignored or adjusted to reflect a dominant direction. These modifications can lead to inferior results, especially in groundwater flow and solute transport simulations where the directionality of conductive layers can be critical. Furthermore, in cases where geologic maps are not available, field-derived formation orientation data, such as strike and dip directions, can be invaluable for reconstructing the manifold geometry. In this context, the strike direction can indicate the orientation of the level-set curve, and the dip, the manifold geometry. The integration of such field data into the methodological framework of this study remains a work in progress, and detailed results are expected in future publications.

5.2 Seismic profile as secondary data

Seismic data form an important complementary tool to reveal the underlying subsurface structures. As shown in Fig. 1(b), seismic reflectors, shown on a profile, play a key role in identifying the arrangement of different lithological layers. These profiles not only provide insight into stratigraphy, which is essential for defining petrophysical properties in sedimentary basins, but also are important for reservoir characterization, particularly in the petroleum industry. Although seismic data are often associated with hydrocarbon exploration, these data are versatile enough to be extended to other hydrogeological applications, including groundwater resource management and the identification of ideal CO_2 storage sites. This subsection illustrates how the techniques reported herein effectively leverage seismic profiles, similar to the geologic maps, for subsurface analysis.

Figure 7 shows the results of linear feature identification for two window sizes: 14 pixels, as shown in Fig. 7(a), and 28 pixels, as shown in Fig. 7(b), with number of sample points determined to be 3042 and 722, respectively. For the seismic profile, the pixel counts along the x and y axes are 555 (n_u) and 557 (n_v) for window sizes of 14 and 28 pixels, respectively, reflecting a resolution that is approximately 4.6 times that of the geological map example (See Sect. 5.1, where n_u and n_v are 308 and 218, respectively). The delineated tangential linear features serve as proxies for the level-set curve directions (i.e., $\partial L_c/\partial u$ and $\partial L_c/\partial v$). From Fig. 7, we see that a higher level of structural detail is captured when a smaller window size is used. Thus, we conclude that a smaller window size should be chosen in cases where a detailed manifold gradient pattern can improve the estimation accuracy.

(a)



(b)

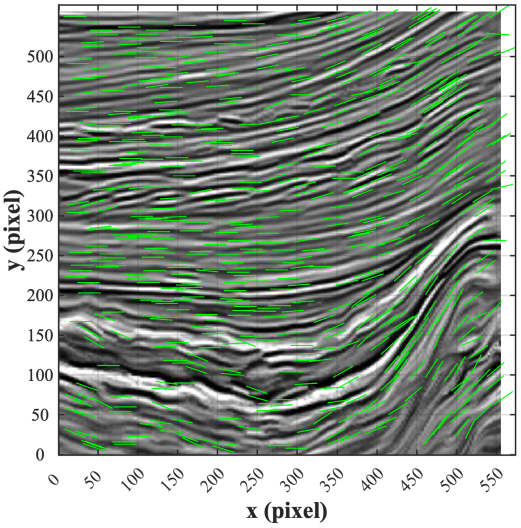
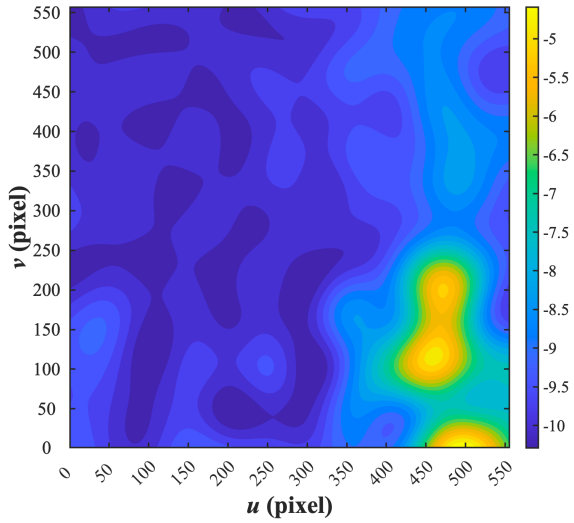


Figure 7. Linear feature identification in a seismic profile for different window sizes: (a) A window size of 14 pixels with 3042 sample points; (b) a window size of 28 pixels with 722 sample points.

Figure 8 shows the interpolated gradients of $(\partial f/\partial u)_{est}$ and $(\partial f/\partial v)_{est}$, which were derived from the proxies for the level-set curve directions at the sampling locations. Only the 28-pixel window size was considered for this particular demonstration for the computational efficiency. As in the previous section, GPR was used for regression. In the estimation, ρ was empirically determined as 50 pixels and σ^2 was taken as 1×10^{-1}. Consistent with the findings reported in Sect. 5.1, noticeable variations were observed in the interpolated gradients throughout the domain. This pattern highlights a non-stationary spatial relationship, resulting in a non-unitary matrix of g in Eq. (11). The gradients $(\partial f/\partial u)_{est}$ and $(\partial f/\partial v)_{est}$ were then used to calculate the geodesic distance.

(a)



(b)

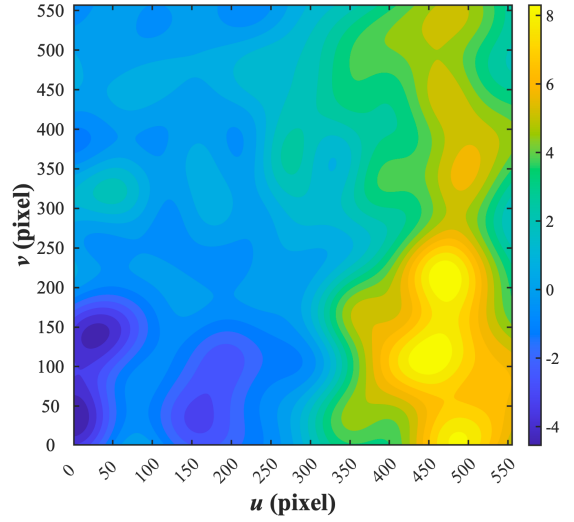


Figure 8. Interpolated gradients, labelled $(\partial f/\partial u)_{est}$ and $(\partial f/\partial v)_{est}$, derived from proxies for the level-set curve directions, using a window size of 28 pixels, where (a) shows the estimated gradient field in the u direction, and (b) shows the estimated gradient field in the v direction.

Given that the seismic profile is vertical to the surface, dispersed data cannot be easily obtained by the conventional methods. Hence, a two-step approach was adopted. In the first step, hypothetical data were generated by unconditional simulation. In the second step, the unconditional simulation from the first step was treated as an actual field. Within this simulated field, three artificial boreholes were introduced, and full petrophysical properties (e.g., permeability) were assumed at these borehole locations. This framework allowed a comparison between the conditional simulation based on the second phase and the unconditional simulation from the first phase to assess the reproducibility of the proposed method. For the unconditional generation, 100 randomly distributed data points, following a normal distribution, were assigned to random locations

distributed uniformly in both the u and v directions. Thus, following the approach described in the previous section (Sect. 4), these data can be interpreted as log-transformed hydraulic conductivity or permeability values. In the next phase, the three boreholes were set at u = 100, 250 and 400 pixels along the x-axis. The data from these boreholes and the secondary data were considered available for estimation in this second phase. The GPR was used as a regressor in both the unconditional simulation and the conditional estimation.

The results of the unconditional simulation, based on 100 random data points, are shown in Fig. 9. These results show patterns similar to the sequential distribution of high and low permeability values, which is a typical distribution seen in sedimentary basins. For this simulation, ρ = 150 pixels, β = 10, and $\sigma^2 = 1 \times 10^{-1}$, consistent with the previous scenario (Sect. 5.1.). From these simulated results, three hypothetical boreholes were selected at x = 100, 250 and 400 pixels. The log-transformed permeability at these boreholes was assumed to be known, and their patterns, as shown in Fig. 10, indicate sedimentological sequences influenced by the characteristic transgression and regression of the sea level.

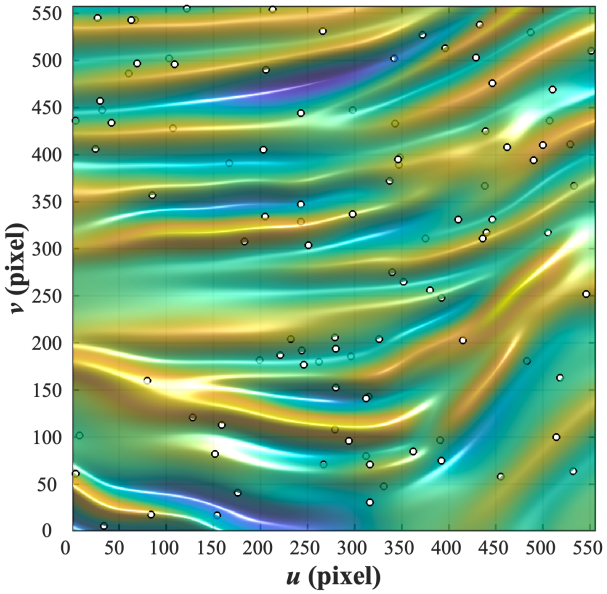


Figure 9. Unconditional simulation results showing patterns reminiscent of the sequential high and low permeability distributions typical of sedimentary basins. This simulation was based on 100 randomly distributed data points (white dots with black outlines) with parameters set to ρ = 150 pixels and β = 10.

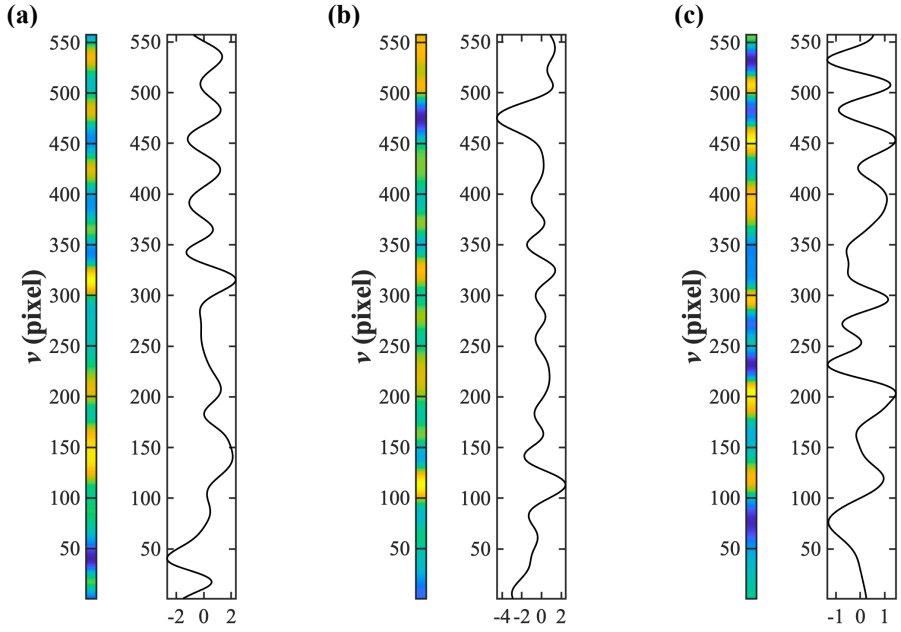


Figure 10. Patterns of log-transformed petrophysical property (e.g. hydraulic conductivity) at three hypothetical borehole locations at (a) u = 100, (b) u = 250 and (c) u = 400 pixels.

Statistical evaluation of the selected hypothetical boreholes and the entire domain provides remarkable insights. The minimum and maximum log-transformed permeabilities were−4.54 and 2.44, respectively, for boreholes and −4.55 and 2.73, respectively, for the entire domain. The mean permeability for the boreholes was −0.001, while that for the entire domain was 0.093. Furthermore, the recorded variances were 1.1125 and 0.9979 for the boreholes and the entire domain, respectively. A comparison of the histograms of the two data sets revealed a pronounced negative skew. The value of the Kullback–Leibler divergence, as a measure of histogram similarity, was 0.03, suggesting that the two distributions were essentially identical. This result underscores the suitability of the selected boreholes for conditional estimation.

A simulation was then conditioned on the three hypothetical boreholes using the same seismic profile from the unconditional simulation as that for the secondary data. The results obtained using the simple GPR, kriging and proposed method were compared. Kriging yielded remarkably poorer results mainly because it could not handle the non-stationarity of the field. For kriging, the correlation scales in the x and y directions were derived from the correlogram, which was derived from the unconditional simulation that accurately represented the spatial statistics of the actual field. The correlation scales in the x and y directions were 143 (denoted as ρ_x) and 27 (denoted as ρ_y) pixels, respectively. Additionally, the geodesic kernel in Eq. (7) was modified to an anisotropic Euclidean distance expressed as follows:

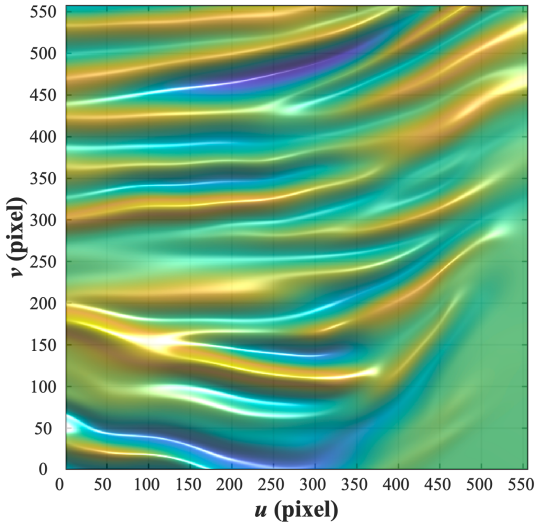
$$k_{kriging}(\mathbf{p}_i, \mathbf{p}_j) = \exp\left[-\frac{(x_i - x_j)^2}{2{\rho_x}^2} - \frac{(y_i - y_j)^2}{2{\rho_y}^2}\right].$$

From the above equation, kriging clearly did not consider the rotations of ρ_x and ρ_y.

Figure 11 shows the results obtained using the proposed method and kriging. The differences are clear: kriging struggles with the non-stationarity of the field, yielding results that are incongruent with the real scenario. In contrast, a visual inspection confirms that the results of the proposed method closely match the true field represented by the unconditional simulation. With regard to the correlation between the true field and estimates, the correlation coefficients of the proposed method is 0.9, while that of kriging is only 0.51. This stark difference in correlation coefficients highlights the inherent superiority of the proposed method, even when kriging is informed with precise spatial statistics. Although the basic statistics such as means and variances show some variation across the true field, results of the proposed method, and kriging results, they are relatively consistent overall.

The SSIM values were compared; the proposed method had an SSIM value of 0.915, while that of kriging was 0.706. This considerable difference further emphasises that compared to the kriging results, the estimates obtained by the proposed method are much closer to the actual data.

(a)



(b)

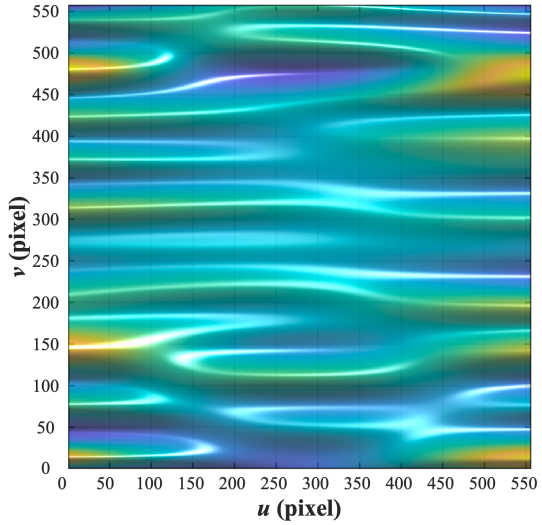
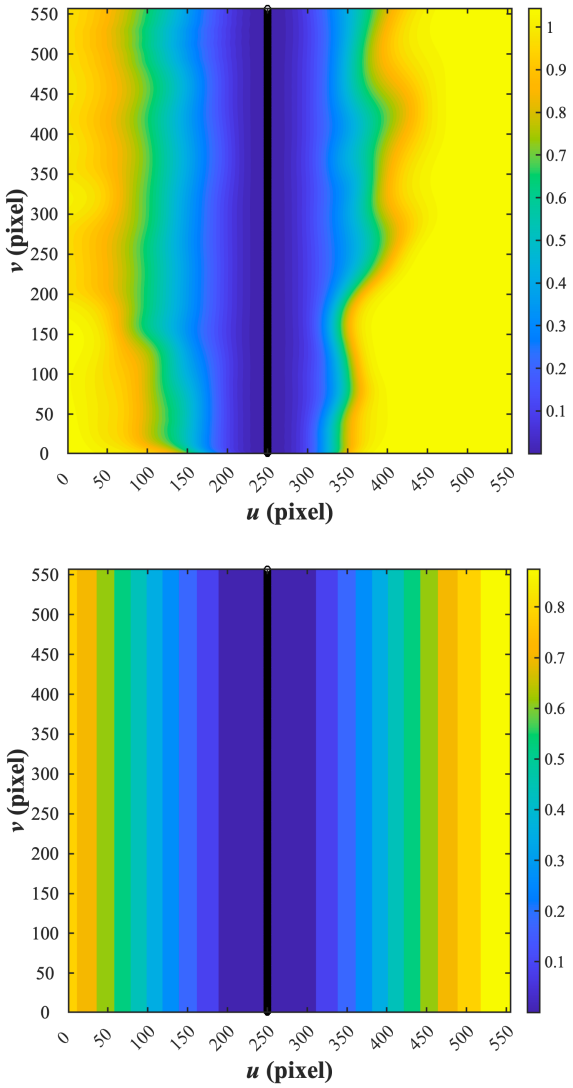


Figure 11. Comparative visualisation of the estimated fields using conditioning data from Fig. 10 and two methods: (a) the estimated field produced by the proposed method, and (b) the estimated field derived from kriging.

In addition, the relative uncertainty inherent in both the developed and kriging methods was evaluated. The results are shown in Fig. 12. For a clearer visualisation of the uncertainty spread across the domain, only the borehole positioned at $u = 250$ pixels was considered. It is critical to note that the proposed method is based on geodesic distance, a metric that is supported by secondary data. Consequently, the relative uncertainty is not simply distributed based on the linear distance from the borehole. Instead, it exhibits a nuanced distribution that reflects the integration of secondary information. This nuance is inherently logical: different lithologies have different correlation scales. Thus, while some lithological units may exhibit extended correlation scales, others may exhibit a more localised nature. In stark contrast, the uncertainty propagation of the kriging method is predominantly influenced by the simple Euclidean distance from the borehole. Therefore, its uncertainty distribution shows a gradual and homogenous increase with the distance from the borehole. This difference between the two methods emphasises the robustness of the proposed method in capturing the complexity of spatial relationships influenced by multiple data sources, as opposed to the simpler and distance-dependent approach of kriging.

(a)



(b)

Figure 12. Comparative visualisation of the relative uncertainty distributions for the proposed method and kriging considering a borehole located at u = 250 pixels: the uncertainty distribution for (a) the proposed method and (b) kriging.

Analysis of the aforementioned cases clearly indicate the advantages of the proposed method over conventional kriging. Our method, which is based on the assimilation of secondary data, can effectively delineate the complex petrophysical parameter distribution of the target area. Consequently, it can yield more accurate predictions in the simulation of subsurface flows than the predictions based on traditional kriging. The traditional method, by design, often oversimplifies spatial relationships using a limited set of spatial parameters.

In more complex scenarios, such as solute transport simulations and multiphase flows, the difference between the proposed method and conventional techniques becomes particularly sharp. In these situations, even minute granular-scale variabilities, especially when intertwined with structural connectivity, can drastically affect simulation results. Consider, for example, the

critical tasks of evaluating the potential impact of contaminant sources for strategic mitigation, assessing the risk of CO_2 leakage because of an imperfect cap rock, or investigating the suitability of sites for the geological disposal of high-level radioactive waste. In these critical contexts, compared to the traditional methods, the proposed method has much higher accuracy because of which this method is capable of providing insights that are both more reliable and more actionable.

6 Summary and conclusion

Subsurface non-stationarity has always been a formidable challenge in geological characterization. The traditional methods based on the covariance matrix to estimate continuous variables often either ignore or oversimplify this complex problem. In contrast, this study adopted a direct approach. By incorporating ancillary data, we formulated a rigorous theory and methodology that gives us accurate estimates of the spatial distribution of subsurface petrophysical properties. This methodology is particularly relevant given the widespread non-stationarity inherent in large-scale geological fields. The geodesic distance on the embedded manifold, which is the foundation of the method proposed herein, was introduced as a fundamental tool. Field observations were linked to the intrinsic geological non-stationarity by using the level-set curve. This curve serves as a key indicator for interpreting the manifold information from observed geological structures and effectively addresses the spatial variations in the correlation direction characteristic of inhomogeneous geological processes.

During the implementation phase, using the geological map and seismic profile as secondary data, we found that parameters ρ and β play a crucial role. Specifically, ρ indicates correlation strength and is analogous to the correlation length in the conventional geostatistics, and β governs the dependence of the estimates on the secondary data. Therefore, a careful calibration of these parameters is essential. Comparative analyses showed that the proposed method significantly outperforms conventional methods, such as kriging, especially in terms of reproducing subsurface structures with subtle shifts in correlation direction. In particular, the uncertainty in our method encompasses both data proximity and the complex correlation structures inherent in the secondary data. That is, the uncertainty exhibits non-stationarity, mirroring the estimates. This representation agrees well with geological contexts and provides a more rational and intuitive representation of subsurface uncertainties.

The accuracy of estimating subsurface media distribution is extremely important, especially in critical applications such as contaminant remediation design and underground repository siting. In these contexts, even small inaccuracies can have profound negative consequences. The theories and methods presented herein offer promising solutions to these complex subsurface challenges. By skilfully leveraging the subtleties of secondary data, our approach facilitates the accurate characterization of petrophysical properties that exhibit non-stationarity.

However, the study has some limitations. The study mainly focused on 2D methods. There is an urgent need to explore 3D frameworks to further enhance the practicality of the proposed method. The present application focuses on the distribution of layered petrophysical properties. However, evaluations over a wider range of geological structures, especially those associated with remarkable subsurface geological deformations (such as faults and plutonic intrusions) are yet to be performed. In future

study, efforts should be made to validate the adaptability of the methodology to various secondary data sets and assessing its effectiveness over a broader range of geologic processes.

In conclusion, this paper reports innovative theoretical foundations and practical method for estimating spatial distributions characterised by non-stationarity in geological media. These contributions are significant advances towards bridging the prevailing knowledge gaps. The profound implications of these outcomes for addressing the existing challenges in subsurface characterization are evident, and a solid foundation has been established to facilitate diverse research efforts in the future.

Code availability. All software programs were written in MATLAB. All the executable software used in this study are available through a public data repository once the manuscript is accepted for publication.

Data availability. All the data used in this study are available through a public data repository once the manuscript is accepted for publication.

Author contributions. EP and JP developed the theory, method and the code. EP, JP, HJ, WSH and HS discussed results and validate the results. EP and YSK wrote the paper with contributions from all authors.

Competing interests. The authors declare that they have no conflict of interest.

Financial support. This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea Government (MSIT) (RS-2023-002772264).

Acknowledgements. The authors acknowledge the use of AI technologies, including OpenAI's ChatGPT and Google's Bard, for their assistance in the conceptual development of this work. These tools were instrumental in improving English expression, a critical tool for non-native English-speaking authors to achieve scientific accuracy. DeepL's language refinement services also significantly augmented the linguistic quality of the manuscript.

References

Adams, E. E., and Gelhar, L. W.: Field study of dispersion in a heterogeneous aquifer: 2. Spatial moments analysis, Water Resour. Res., 28(12), 3293-3307, https://doi.org/10.1029/92WR01757, 1992.

Batu, V.: Aquifer hydraulics: a comprehensive guide to hydrogeologic data analysis, John Wiley & Sons, 1998.

Boggs, J. M., Young, S. C., Beard, L. M., Gelhar, L. W., Rehfeldt, K. R., and Adams, E. E.: Field study of dispersion in a heterogeneous aquifer: 1. Overview and site description, Water Resour. Res., 28(12), 3281-3291, https://doi.org/10.1029/92WR01756, 1992.

Choi, D. K., Lee, J. G., Lee, S. B., Park, T. Y. S., and Hong, P. S.: Trilobite biostratigraphy of the lower Paleozoic (Cambrian–Ordovician) Joseon Supergroup, Taebaeksan Basin, Korea, Acta Geologica Sinica-English Edition, 90(6), 1976-1999, https://doi.org/10.1111/1755-6724.13016, 2016.

Cooper Jr, H. H., Bredehoeft, J. D., Papadopulos, I. S., and Bennett, R. R.: The response of well-aquifer systems to seismic waves, J. Geophys. Res., 70(16), 3915-3926, https://doi.org/10.1029/JZ070i016p03915, 1965.

Cressie, N.: Kriging nonstationary data, J. Am. Stat. Assoc., 81(395), 625-634, https://doi.org/10.2307/2288990, 1986.

Cressie, N.: Aggregation in geostatistical problems, Springer Netherlands, pp. 25-36, 1993.

D'Affonseca, F. M., Finkel, M., and Cirpka, O. A.: Combining implicit geological modeling, field surveys, and hydrogeological modeling to describe groundwater flow in a karst aquifer, Hydrogeol. J., 28(8), 2779-2802, https://doi.org/10.1007/s10040-020-02220-z, 2020.

Dietrich, S., Weinzettel, P. A., and Varni, M.: Infiltration and drainage analysis in a heterogeneous soil by electrical resistivity tomography, Soil Sci. Soc. Am. J., 78(4), 1153-1167, https://doi.org/10.2136/sssaj2014.02.0062, 2014.

Doetsch, J., Linde, N., Coscia, I., Greenhalgh, S. A., and Green, A. G.: Zonation for 3D aquifer characterization based on joint inversions of multimethod crosshole geophysical data, Geophysics, 75(6), G53-G64, https://doi.org/10.1190/1.3496476, 2010.

Feragen, A., Lauze, F., and Hauberg, S.: Geodesic exponential kernels: When curvature and linearity conflict, Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 3032-3042, https://doi.org/10.48550/arXiv.1411.0296, 2015.

Fouedjio, F., Desassis, N., and Rivoirard, J.: A generalized convolution model and estimation for non-stationary random functions, Spatial Stat., 16, 35-52, https://doi.org/10.1016/j.spasta.2016.01.002, 2016.

Han, W. S., McPherson, B. J., Lichtner, P. C., and Wang, F. P.: Evaluation of trapping mechanisms in geologic CO_2 sequestration: Case study of SACROC northern platform, a 35-year CO_2 injection site, Am. J. Sci., 310(4), 282-324, https://doi.org/10.2475/04.2010.03, 2010.

Hewett, T. A.: Fractal distributions of reservoir heterogeneity and their influence on fluid transport, SPE Annual Technical Conference and Exhibition, SPE-15386, https://doi.org/10.2118/15386-MS, 1986.

Higdon, D.: A process-convolution approach to modeling temperatures in the North Atlantic Ocean, Environ. Ecol. Stat., 5, 173-190, https://doi.org/10.1023/A:1009666805688, 1998.

Hu, L. Y., and Chugunova, T.: Multiple-point geostatistics for modeling subsurface heterogeneity: A comprehensive review, Water Resour. Res., 44(11), https://doi.org/10.1029/2008WR006993, 2008.

Hyndman, D. W., Harris, J. M., & Gorelick, S. M.: Coupled seismic and tracer test inversion for aquifer property characterization, Water Resour. Res., 30(7), 1965-1977, https://doi.org/10.1029/94WR00950, 1994.

Jayasumana, S., Hartley, R., Salzmann, M., Li, H., and Harandi, M.: Kernel methods on Riemannian manifolds with Gaussian RBF kernels, IEEE Trans. Pattern Anal. Mach. Intell., 37(12), 2464-2477, https://doi.org/10.1109/TPAMI.2015.2414422, 2015.

Kemna, A., Vanderborght, J., Kulessa, B., and Vereecken, H.: Imaging and characterisation of subsurface solute transport using electrical resistivity tomography (ERT) and equivalent transport models, J. Hydrol., 267(3-4), 125-146, https://doi.org/10.1016/S0022-1694(02)00145-2, 2002.

Kerrou, J., Renard, P., Cornaton, F., and Perrochet, P.: Stochastic forecasts of seawater intrusion towards sustainable groundwater management: application to the Korba aquifer (Tunisia), Hydrogeol. J., 21(2), 425-440, https://doi.org/10.1007/s10040-012-0911-x, 2013.

Lumley, D. E.: Time-lapse seismic reservoir monitoring, Geophysics, 66(1), 50-53, https://doi.org/10.1190/1.1444921, 2001.

Mao, D., Revil, A., Hort, R. D., Munakata-Marr, J., Atekwana, E. A., and Kulessa, B.: Resistivity and self-potential tomography applied to groundwater remediation and contaminant plumes: Sandbox and field experiments, J. Hydrol., 530, 1-14, https://doi.org/10.1016/j.jhydrol.2015.09.031, 2015.

Paciorek, C. J.: Nonstationary Gaussian processes for regression and spatial modelling, Carnegie Mellon University, Doctoral dissertation, 2003.

Park, E., Kim, K. Y., and Suk, H.: A basin-scale aquifer characterization using an inverse analysis based on groundwater level fluctuation in response to precipitation: Practical application to a watershed in Jeju Island, South Korea, J. Hydrol.: Regional Studies, 37, 100933, https://doi.org/10.1016/j.ejrh.2021.100933, 2021.

Pereira, M., Desassis, N., and Allard, D.: Geostatistics for large datasets on Riemannian manifolds: a matrix-free approach, arXiv preprint, arXiv:2208.12501, https://doi.org/10.48550/arXiv.2208.12501, 2022.

Piao, J., and Park, E.: Enhancing Estimation Accuracy of Nonstationary Hydrogeological Fields via Geodesic Kernel-Based Gaussian Process Regression, J. Hydrol., 130150, https://doi.org/10.1016/j.jhydrol.2023.130150, 2023.

Pride, S. R., Harris, J. M., Johnson, D. L., Mateeva, A., Nihel, K. T., Nowack, R. L., Rector, J. W., Spetzler, H., Wu, R., Yamomoto, T., Berryman, J. G., and Fehler, M.: Permeability dependence of seismic amplitudes, The Leading Edge, 22(6), 518-525, https://doi.org/10.1190/1.1587671, 2003.

Qin, X. S., Huang, G. H., Chakma, A., Chen, B., and Zeng, G. M.: Simulation-based process optimization for surfactant-enhanced aquifer remediation at heterogeneous DNAPL-contaminated sites, Sci. Total Environ., 381(1-3), 17-37, https://doi.org/10.1016/j.scitotenv.2007.04.011, 2007.

Rubin, Y., Mavko, G., and Harris, J.: Mapping permeability in heterogeneous aquifers using hydrologic and seismic data, Water Resour. Res., 28(7), 1809-1816, https://doi.org/10.1029/92WR00154, 1992.

Schroot, B. M., & Schüttenhelm, R. T.: Shallow gas and gas seepage: expressions on seismic and other acoustic data from the Netherlands North Sea, J. Geochem. Explor., 78, 305-309, https://doi.org/10.1016/S0375-6742(03)00112-2, 2003.

Son, C. M., and Lee, D. S.: Explanatory text of the Geological Map of Ogdong Sheet, Geological Survey of Korea, Sheet-6925-III, scale 1: 50,000, 30p, 1966.

Soupios, P. M., Kouli, M., Vallianatos, F., Vafidis, A., and Stavroulakis, G.: Estimation of aquifer hydraulic parameters from surficial geophysical methods: A case study of Keritis Basin in Chania (Crete–Greece), J. Hydrol., 338(1-2), 122-131, https://doi.org/10.1016/j.jhydrol.2007.02.028, 2007.

Strebelle, S.: Conditional simulation of complex geological structures using multiple-point statistics, Math. Geol., 34, 1-21, https://doi.org/10.1023/A:1014009426274, 2002.

Suk, H., and Park, E.: Numerical solution of the Kirchhoff-transformed Richards equation for simulating variably saturated flow in heterogeneous layered porous media, J. Hydrol., 579, 124213, https://doi.org/10.1016/j.jhydrol.2019.124213, 2019.

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P.: Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process., 13(4), 600-612, https://doi.org/10.1109/TIP.2003.819861, 2004.

Yaramanci, U., Lange, G., and Knödel, K.: Surface NMR within a geophysical study of an aquifer at Haldensleben (Germany), Geophys. Prospect., 47(6), 923-943, https://doi.org/10.1046/j.1365-2478.1999.00161.x, 1999.

Yeh, T. C. J., and Liu, S.: Hydraulic tomography: Development of a new aquifer test method, Water Resour. Res., 36(8), 2095-2105, https://doi.org/10.1029/2000WR900114, 2000.