Enhancing Estimation Accuracy of Nonstationary Hydrogeological Fields via Geodesic Kernel-based Gaussian Process Regression

Eungyu Park

Department of Geology, Kyungpook National University, Daegu, Republic of Korea

Corresponding author: Eungyu Park (egpark@knu.ac.kr, ORCID: 0000-0002-2293-4686)

Key Points:

- Estimation of hydraulic conductivity in hydrogeological systems using geodesic kernel–Gaussian process regression was proposed.
- Incorporating secondary information can improve estimation accuracy and provide insights into geological structures.
- Importance of accurate hydraulic conductivity estimation for groundwater management and contamination risk assessment was highlighted.

Abstract

In this study, the combined application of geodesic kernel and Gaussian process regression was investigated to estimate nonstationary hydraulic conductivity fields in two-dimensional hydrogeological systems. Particularly, a semi-analytical form of the geodesic distance based on the intrinsic geometry of the manifold was derived and used to define positive definite geodesic covariance matrices that are employed for Gaussian process regression. Furthermore, the proposed approach was applied to a series of synthetic hydraulic conductivity estimation problems and the results show that the incorporation of secondary information, such as geophysical or geological interpretations, can considerably improve the estimation accuracy, especially in nonstationary fields. Moreover, groundwater flow and solute transport simulations based on the estimated hydraulic conductivity fields revealed that the accuracy of the simulations was strongly affected by the inclusion of secondary information. These results suggest that incorporating secondary information into manifold geometry can remarkably improve the estimation accuracy and provide new insights on the underlying structure of geological data. This proposed approach has crucial implications for hydrogeological applications, such as groundwater resource management, safety assessments, and risk management strategies related to groundwater contamination.

Keywords: geodesic kernel, Gaussian process regression, manifold geometry, hydraulic conductivity estimation, nonstationary fields, secondary information.

Plain Language Summary

In this study, a novel method is proposed to estimate the distribution of hydraulic properties in two-dimensional hydrogeological systems. This approach combines geodesic kernel and Gaussian process regression to utilize secondary information, such as geophysical exploration data or field geological insights, and considerably improves the estimation accuracy, particularly for nonstationary fields often observed in real practices. Incorporating secondary information into manifold geometry can yield novel insights into the underlying structure of geological data, allowing for more accurate spatial estimations in nonlinear and nonstationary situations. Moreover, the computational efficiency of the approach is reflected in the simple, semi-analytical form of the derived geodesic kernel equation. These findings have notable implications for subsurface fluid flow and solute transport affected by small scale variability of the media properties for example groundwater resource management, safety assessments, and risk management strategies related to groundwater contamination.

1 Introduction

Geostatistics provides a robust set of methods to estimate subsurface properties based on statistical regularities (e.g., Deutsch & Journel, 1992; Goovaerts, 1997; Chiles & Delfiner, 2009). However, the application of these methods is often limited by the spatial variability of geological processes, which can undermine the assumption of stationarity and make accurate estimation challenging (Cressie, 1986; Cressie, 1993; Yeh & Liu, 2000; Wackernagel, 2003). Nonstationarities in geological processes present a major hindrance to the application of geostatistical methods at large-scale sites, thereby exhibiting substantial heterogeneity in geological structures and processes. This limitation emphasizes the need for advanced and reliable methods to characterize and model the hydraulic conductivity distribution, which plays a vital role in groundwater management and risk assessment (Dagan, 1989; Gorelick & Zheng,

2015). Therefore, it is crucial to improve the existing methods for estimating the hydraulic conductivity distribution in order to enrich our understanding of subsurface properties and efficiently manage groundwater resources and mitigate groundwater contamination risks. Toward this end, both advanced and diversified methods are crucial in capturing the complex spatial variability and nonstationarities in geological processes that are common in large-scale aquifers.

Since the recognition of nonstationarity owing to geological structures and processes is a major challenge in subsurface characterization, considerable research efforts have been made to develop more advanced methods to characterize and model the spatial variability arising from geological processes. Different from the conventional covariance-based approach, multipoint statistics (MPS) is a robust approach to address spatial nonstationarity using pattern recognition from multiple points (Strebelle, 2002). It incorporates additional information from multiple points, thereby improving the accuracy and reliability of subsurface property estimation, particularly in areas exhibiting high heterogeneity in geological structures and processes (Mariethoz et al., 2010; Mariethoz & Caers, 2014). Moreover, spatial generative adversarial network (SGAN) by Laloy et al. (2018) is another robust method that exhibits promising results in capturing and reproducing complex spatial patterns of geological structures and processes. Both MPS and SGAN methods typically rely on training images as a source of nonstationary spatial statistical information. As noted in literature (Mariethoz, 2018; Madsen et al., 2021), these methods are most effective in when there is a severe scarcity or abundance of information present; however, they may not perform optimally in cases where an intermediate level of geological detail from secondary information such as geophysical data is available, which is the issue addressed in this study.

Recently, addressing the nonstationarity problem in covariance-based approaches has received considerable attention. Most popular among these are methods that use geodesic kernel approaches (e.g., Feragen et al., 2014; Jayasumana et al., 2015; Pereira et al., 2022) or kernel convolution approaches (e.g., Higdon et al., 1999; Paciorek, 2003; Fouedjio et al., 2016). These approaches present great potential in capturing complex spatial variations in subsurface properties, which have crucial applications in hydrogeology and other fields where nonstationarity is a common challenge. However, the selection of kernel function and manifold structure requires domain-specific knowledge, which poses a challenge for its applications when the underlying data structure is complex and not well understood. Therefore, further research in hydrogeology and related fields is required to completely address these challenges and realize the true potential of these approaches.

Although the geodesic kernel and kernel convolution approaches exhibit great potential in capturing complex spatial variations in subsurface properties, the former is particularly well-suited for handling nonEuclidean manifolds, which are common in hydrogeology. Considering the irregularity of the subsurface data in hydrogeology, the geodesic kernel approach was selected as the most appropriate choice for this study. This approach combines targeted variable observations (as primary data) with a secondary information-derived manifold to enhance estimation accuracy. The manifold structure utilized herein was guided by geophysical explorations or derived from domain-specific expert knowledge, and can be further optimized with additional data and analysis.

A manifold can provide a more detailed and interpretable representation of the spatial relationships between locations within a given dataset by embedding high-dimensional data in a

low-dimensional space. Differential geometry offers two main approaches for handling manifolds: intrinsic and extrinsic. Intrinsic geometry studies the geometric properties of objects from within, without relying on any external reference system. In contrast, extrinsic geometry studies objects from the outside, by embedding them in a higher-dimensional space. While each approach offers different advantages, the intrinsic approach may be more suitable for providing a more detailed and interpretable representation of the spatial relationships within a given dataset, particularly in the case of the often complex and irregular parameter distributions encountered in geological fields. Nontheless, many of the referable studies have primarily utilized extrinsic geometry, which frequently requires complex numerical methods to compute geodesics on the manifold. To address the existing challenges and to potentially increase its accessibility to hydrogeologists, a computational approach utilizing intrinsic geometry was proposed in this study. By leveraging the underlying manifold structure of the data, this approach has the potential to overcome nonstationarity and other challenges, making it a more accessible solution for hydrogeologists looking to incorporate these techniques into their studies. As an initial study, the main focus is not on providing specific methods for tailoring manifold geometries to particular geological structures, but rather on developing a methodology and demonstrating it through a few synthetic examples. Thus, a detailed description of how manifold geometries are constructed from secondary information is not covered in this manuscript.

In the remaining sections of this manuscript, an approach employing intrinsic geometry is proposed. Particularly, a metric tensor is introduced to define a geodesic distance in the Euclidean space. Additionally, the Gaussian process regression (GPR) method that uses the geodesic kernel is introduced. To implement this approach, a synthetic hydraulic conductivity field is generated to obtain synthetic data that can be utilized as conditioning information. These synthetic data serve as the primary data, and the secondary data with variable amounts of information are used in conjunction for estimation. In the latter part of the implementation, groundwater flow and solute transport simulations are conducted on the estimated results, and they are systematically compared with the true values.

2 Methodology

2.1 Manifold and radial basis functions

Herein, a manifold approach was adopted to address nonstationarity in spatial statistics. A manifold is a topological space that locally resembles Euclidean space. In spatial analysis, manifolds can be used as supplementary information to incorporate prior information, such as the underlying spatial structure, to improve the interpretability of the estimations since the curvature of manifolds can provide crucial information about spatial relationships between data points (Pereira et al., 2022). This study hypothesized that the geometries of the manifolds can be constructed through secondary information, such as geophysical or geological surveys, where the primary data include petrophysical properties such as permeability or porosity.

Owing to their properties of handling irregular or complex geometries and allowing for nonlinear data relationships (Carr et al., 2001), kernel functions, especially radial basis functions (RBFs), have long been used in conventional geostatistics and are known to be well-suited for spatial analysis on a manifold (Feragen et al., 2014). Furthermore, this study hypothesized that the statistical relationships between spatially varying data can be represented via spatial analysis on a manifold using RBFs. The kernel function considered here is a Gaussian RBF, which

satisfies the positive definiteness on a manifold when it is isometric to some Euclidean space (Feragen et al., 2014; Jayasumana et al., 2015; Feragen and Hauberg, 2016; Borovitskiy et al., 2020). Consequently, a Euclidean space equipped with the metric of a manifold (i.e., intrinsic geometry) was used for the spatial analysis to satisfy the isometric condition between the Euclidean space and manifold. In the estimation, GPR will be adopted where they were constrained by the secondary information provided in the form of manifold geometry.

2.2 Intrinsic geometry and metric tensor: spatial relationships on manifolds

The advantage of using intrinsic geometry is that it reduces complexity by measuring the length of an arc between two points within a manifold without referring to a larger space. For example, the arc length on a curved surface in a three-dimensional inner product space of XYZ surface can be examined based on two-dimensional metric space of uv-plane where two spaces (uv and XYZ spaces) are homeomorphic. In this mapping of ψ: $(u, v) \mapsto (X, Y, Z)$ (Figure 1), $u, v \in M$ where (M, r) is a metric space and $X, Y, Z \in \mathcal{V}$, where $\mathcal{V}$ is an inner product space. Furthermore, $X(u, v)$, $Y(u, v)$, and $Z(u, v)$ are uniquely determined using auxiliary variables of u and v by a chart, ψ (a bijective mapping function). From this intrinsic view, a metric tensor is an additional structure on a manifold, allowing arc lengths computations that can be equipped in a Euclidean uv-plane. A metric tensor is defined on the tangent plane at each point p of a manifold and varies smoothly with p (Figure 1). The coordinates for the location of data acquisition lie on the uv plane; however, spatial analysis is based on the location on the XYZ surface corresponding to that of uv plane.

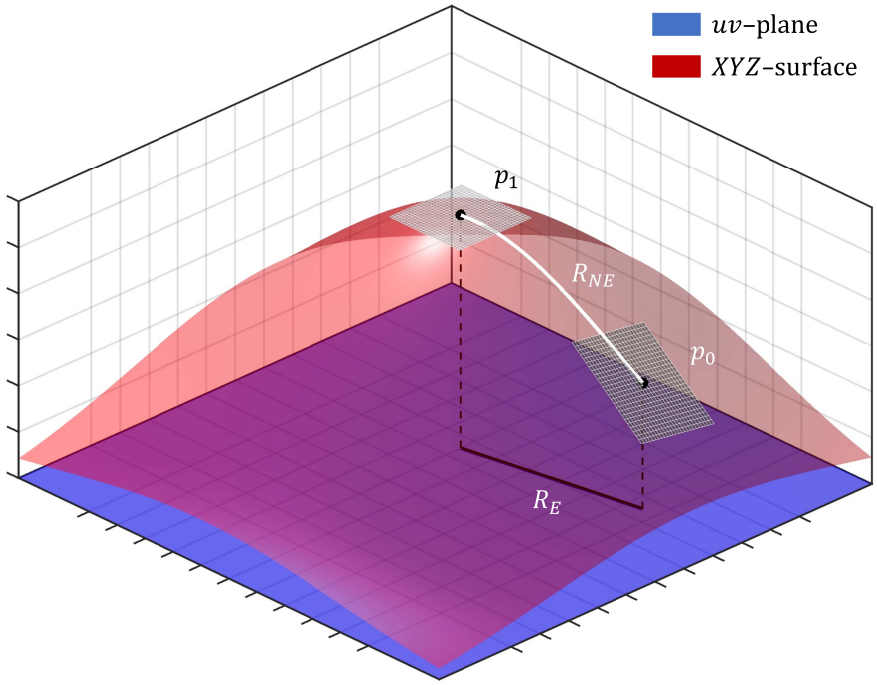

Figure 1. Conceptual schematic of uv plane and XYZ surface. The gray gridded planes are the tangent planes at p_0 and p_1, which are two points on XYZ surface. Arc length of R_{NE} denotes the nonEuclidean distance between p_0 and p_1, whereas R_E denotes the corresponding Euclidean distance. Notably, $R_E \neq R_{NE}$.

Because a manifold defines a detailed spatial relationship between two data locations, these data locations can be analyzed by introducing the metric tensor to the uv plane, allowing the determination of preferential correlation directions (Pereira et al., 2022). The metric tensor can capture the local variations in the spatial relationships between data points, providing a more detailed view of the correlations that exist within the data.

Herein, XYZ surface is an inner product space $\mathcal{V}$ determined using the following mapping function $\psi: M \to \mathcal{V}$:

$$X(u,v) = u,$$
$$Y(u,v) = v, \text{and} \#(1)$$
$$Z(u,v) = f(u,v)$$

where $f: (M \times M) \to \mathbb{R}$. After adopting intrinsic geometry from Eq. (1), the metric tensor for space M can be formulated as follows:

$$g = \begin{bmatrix} \frac{d\mathbf{r}}{du} \cdot \frac{d\mathbf{r}}{du} & \frac{d\mathbf{r}}{du} \cdot \frac{d\mathbf{r}}{dv} \\ \frac{d\mathbf{r}}{du} \cdot \frac{d\mathbf{r}}{dv} & \frac{d\mathbf{r}}{dv} \cdot \frac{d\mathbf{r}}{dv} \end{bmatrix} \#(2)$$

In Eq. (2), $\mathbf{r}$ is a vector valued function parametrically representing a curved surface such that,

$$\mathbf{r}(u,v) = \big(X(u,v), Y(u,v), Z(u,v)\big). \#(3)$$

Moreover, $d\mathbf{r}/du$ and $d\mathbf{r}/dv$ in Eq. (2) are expressed as follows:

$$\frac{d\mathbf{r}}{du} = \frac{dX}{du}\mathbf{e}_X + \frac{dY}{du}\mathbf{e}_Y + \frac{dZ}{du}\mathbf{e}_Z = \mathbf{e}_X + \frac{df(u,v)}{du}\mathbf{e}_Z \text{ and}$$
$$\frac{d\mathbf{r}}{dv} = \frac{dX}{dv}\mathbf{e}_X + \frac{dY}{dv}\mathbf{e}_Y + \frac{dZ}{dv}\mathbf{e}_Z = \mathbf{e}_Y + \frac{df(u,v)}{dv}\mathbf{e}_Z, \#(4)$$

using the parameterizations in Eq. (1), where $\mathbf{e}_X = \partial\mathbf{r}/\partial X$, $\mathbf{e}_Y = \partial\mathbf{r}/\partial Y$, and $\mathbf{e}_Z = \partial\mathbf{r}/\partial Z$ are the tangent space bases. Finally, the metric coefficients are determined by putting Eq. (4) into Eq. (2):

$$g = \begin{bmatrix} 1 + \left(\frac{df(u,v)}{du}\right)^2 & \frac{df(u,v)}{du}\frac{df(u,v)}{dv} \\ \frac{df(u,v)}{du}\frac{df(u,v)}{dv} & 1 + \left(\frac{df(u,v)}{dv}\right)^2 \end{bmatrix}. \#(5)$$

2.3 Geodesic distance

The geodesic distance on a manifold is a notion of distance that measures the shortest path between two points on a curved surface. However, a geodesic distance on a Euclidean space is the shortest path (i.e., a straight line) between two points in the space. In the case of a Euclidean space with a metric tensor that corresponds to the geometry of an embedded manifold, the straight lines in the Euclidean space correspond to the geodesics on the manifold. Using the

metric tensor in Eq. (5), the distances between all points on the uv plane and XYZ surface are equivalent, and uv plane and XYZ surface are isometric to each other.

Generally, to determine a geodesic distance (d_g) between (u_0, v_0) and (u_1, v_1) on a curved space, the following equation can be adopted:

$$d_g = \int \left\| \frac{d\mathbf{r}}{d\lambda} \right\| d\lambda \, \#(6)$$

where the metric tensor in Eq. (5) is used to evaluate the term inside integration as follows:

$$\left\| \frac{d\mathbf{r}}{d\lambda} \right\|^2 = \begin{bmatrix} \frac{du}{d\lambda} & \frac{dv}{d\lambda} \end{bmatrix} \begin{bmatrix} 1 + \left(\frac{df(u,v)}{du} \right)^2 & \frac{df(u,v)}{du}\frac{df(u,v)}{dv} \\ \frac{df(u,v)}{du}\frac{df(u,v)}{dv} & 1 + \left(\frac{df(u,v)}{dv} \right)^2 \end{bmatrix} \begin{bmatrix} \frac{du}{d\lambda} \\ \frac{dv}{d\lambda} \end{bmatrix}. \#(7)$$

As the line spans from (u_0, v_0) to (u_1, v_1), u and v along the straight line can be formulated as $u(\lambda) = u_0 + (u_1 - u_0)\lambda$, $v(\lambda) = v_0 + (v_1 - v_0)\lambda$, and $\lambda \in [0,1]$. From this formulation, the derivatives of u and v with respect to λ are $du/d\lambda = u_1 - u_0$ and $dv/d\lambda = v_1 - v_0$, respectively. Using these formulations, Eq. (7) can be expressed as follows:

$$\left\| \frac{d\mathbf{r}}{d\lambda} \right\|^2 = \left(\frac{du}{d\lambda} \right)^2 g_{11} + 2\frac{du}{d\lambda}\frac{dv}{d\lambda} g_{21} + \left(\frac{dv}{d\lambda} \right)^2 g_{22}. \#(8)$$

Substituting $\|d\mathbf{r}/d\lambda\|$ in Eq. (6) by the square root of Eq. (8), the nonEuclidean arc length representing the correlation between two data locations ((u_0, v_0) and (u_1, v_1)) can be obtained using the following Eq. (9):

$$d_g(u_1, v_1; u_0, v_0) = \int_0^1 \sqrt{(u_1 - u_0)^2 g_{11} + 2(u_1 - u_0)(v_1 - v_0) g_{21} + (v_1 - v_0)^2 g_{22}} d\lambda \, . \#(9)$$

where g_{11}, g_{21} $(= g_{12})$, and g_{22} are the metric tensor elements in Eq. (5). In the equation, $df(u,v)/du$ and $df(u,v)/dv$ can be determined as follows:

$$\frac{df(u,v)}{du} = \frac{df(u,v)}{d\lambda}\frac{1}{u_1 - u_0} \text{ and } \frac{df(u,v)}{dv} = \frac{df(u,v)}{d\lambda}\frac{1}{v_1 - v_0} \, . \#(10)$$

Using Eq. (10), Eq. (9) can be rewritten as follows:

$$d_g(u_1, v_1; u_0, v_0) = \int_0^1 \sqrt{(u_1 - u_0)^2 + (v_1 - v_0)^2 + 4\left(\frac{df(u,v)}{d\lambda} \right)^2} \, d\lambda \, , \#(11)$$

where $d(u_1, v_1; u_0, v_0) = \int_0^1 \sqrt{(u_1 - u_0)^2 + (v_1 - v_0)^2} d\lambda$ represents the Euclidean distance between (u_0, v_0) and (u_1, v_1) for a flat surface. Therefore, the nonEuclidean distance, d_g, is always greater than the Euclidean distance, d, except for the case when the manifold is flat at λ (i.e., $df(u,v)/d\lambda = 0$), as in the problem given by Eq. (1). However, without using Eq. (11), a geodesic distance can be determined from Eq. (9) directly using $df(u,v)/du$ and $df(u,v)/dv$, where the derivatives can be calculated using an analytical or a numerical (i.e., finite difference) method, if $f(u,v)$ in Eq. (1) is given. Additionally, numerical integration in

Eq. (9) may be preferred over analytical integration, when $f(u, v)$ analytical or analytical integrations are not available. Various algorithms can be adopted for the numerical integration. Herein, a simple rectangle rule with 5–10 segmentation was applied for the numerical integration considering gentle manifold shapes. The semi-analytical form of Eq. (9) allows easy computation with only a few lines of code, further simplifying the calculation process. However, for more accurate results, alternative numerical integration methods such as Gaussian quadrature (Abramowitz & Stegun, 1972) could be used instead of the simple rectangle rule, particularly for more complex manifold shapes or when a higher precision is required. Although this study utilized a semi-analytical approach to derive the geodesic distance, numerous completely numerical methods are available for this purpose. Among them, the heat method proposed by Crane et al. (2013) could be used as an alternative approach for numerically determining the geodesic distance.

2.4 Gaussian geodesic kernel and Gaussian process regression

Based on the previous studies (Feragen et al., 2014; Jayasumana et al., 2015; Feragen and Hauberg, 2016; Borovitskiy et al., 2020), a positive definite geodesic kernel on a manifold was defined (Eq. 12), when the manifold is isometric to some Euclidean space.

$$k(u, v; u_0, v_0) = \exp\left(-\frac{{d_g}^2(u, v; u_0, v_0)}{2\rho^2}\right), \#(12)$$

where ρ is the parameter related to the correlation scale of a kernel function.

Reproducing property of a positive definite kernel ensures that the inner product in the associated reproducing kernel Hilbert space can be expressed using evaluations of the kernel at points in the space (Schölkopf et al., 2002). In this case, GPR can be used for the manifold estimation. Here, the Gaussian process is defined over the space of functions on the manifold and can be used to model complex relationships and nonlinear interactions between variables. The use of GPR on manifolds requires the definition of a covariance function that captures the properties of the underlying space. The kernel trick is valid in the case when GPR is applied on this manifold (Patel & Vidal, 2014).

In GPR, the estimation for an uninformed location is modeled as a Gaussian distribution. Moreover, the mean and covariance of the distribution are defined by the observations and a positive definite kernel, which can be seen as a similarity measurement between any given two locations. Although basis functions ($\boldsymbol{\phi}$) are not explicitly used, the estimation ($\mathbf{z}$) in GPR can be expressed in conceptual sense as a linear combination of $\boldsymbol{\phi}$ such that

$$\mathbf{z} = \omega_1\boldsymbol{\phi}_1 + \omega_2\boldsymbol{\phi}_2 + \cdots + \omega_{n_f}\boldsymbol{\phi}_{n_f}, \#(13)$$

where ω represents weights and n_f denotes the number of basis functions used for the estimation. Using Mercer's theorem, a symmetric positive definite kernel can be decomposed into an infinite set of basis functions, and a positive definite kernel can be considered as a generalization of the concept of basis functions. Hence, the covariance matrix from a symmetric positive definite kernel can be written as follows:

$$\boldsymbol{\Sigma} = \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}, \#(14)$$

where the feature matrix $\boldsymbol{\Phi} = \begin{bmatrix}\boldsymbol{\phi}_1 & \cdots & \boldsymbol{\phi}_{n_f}\end{bmatrix}$, and the covariance matrix,

$$\boldsymbol{\Sigma} = \begin{bmatrix} k(\mathbf{x}_1, \mathbf{x}_1) & \cdots & k(\mathbf{x}_1, \mathbf{x}_N) \\ \vdots & \ddots & \vdots \\ k(\mathbf{x}_N, \mathbf{x}_1) & \cdots & k(\mathbf{x}_1, \mathbf{x}_N) \end{bmatrix}, \#(15)$$

is composed of the kernel function in Eq. (12). Furthermore, the weight vector of $\boldsymbol{\omega}$ ($= [\omega_1 \quad \cdots \quad \omega_{n_f}]^{\mathrm{T}}$) can be obtained using a ridge regression as follows:

$$\boldsymbol{\omega} = (\boldsymbol{\Phi}(\mathbf{x}_d)^{\mathrm{T}}\boldsymbol{\Phi}(\mathbf{x}_d) + \sigma^2\mathbf{I})^{-1}\boldsymbol{\Phi}(\mathbf{x}_d)^{\mathrm{T}}\mathbf{d}, \#(16)$$

where $\mathbf{x}_d$ denotes the location vector of data observations and $\boldsymbol{\Phi}(\mathbf{x}_d)$ corresponds to the $\mathbf{x}_d$-th rows of $\boldsymbol{\Phi}$ and $\mathbf{d}$ is a vector with the observed data. Additionally, the estimation can be expressed, from Eqs. (13) and (16), with Woodbury matrix identity (Max, 1950) as follows:

$$\mathbf{z} = \boldsymbol{\Phi}\boldsymbol{\omega} = \boldsymbol{\Sigma}_{0d}(\boldsymbol{\Sigma}_{dd} + \sigma^2\mathbf{I})^{-1}\mathbf{d}\,. \#(17)$$

Moreover, the estimation uncertainty conditioned on observations can be obtained from an updated covariance matrix given by

$$\boldsymbol{\Sigma}^{up} = \boldsymbol{\Sigma} - \boldsymbol{\Sigma}_{0d}(\boldsymbol{\Sigma}_{dd} + \sigma^2\mathbf{I})^{-1}\boldsymbol{\Sigma}_{0d}{}^{\mathrm{T}}, \#(18)$$

where the diagonals of $\boldsymbol{\Sigma}^{up}$ are equivalent to estimation uncertainties.

In Eqs. (17) and (18), σ^2 denotes the observation error, $\boldsymbol{\Sigma}_{0d} = \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}(\mathbf{x}_d)$, and $\boldsymbol{\Sigma}_{dd} = \boldsymbol{\Phi}(\mathbf{x}_d)^{\mathrm{T}}\boldsymbol{\Phi}(\mathbf{x}_d)$, where $\boldsymbol{\Sigma}_{0d}$ and $\boldsymbol{\Sigma}_{dd}$ can be directly evaluated from Eq. (12) without explicit evaluations of $\boldsymbol{\Sigma}$ or $\boldsymbol{\Phi}$ by kernel trick.

3 Results and Discussion

3.1 Case studies of hydraulic conductivity field estimations

Herein, the geodesic kernel approach utilized intrinsic geometry to incorporate a metric tensor (Eq. (5)) into Euclidean space. The metric tensor quantified the distance between points on the manifold surface, as described in Eq. (9), and characterized its curvature. This information was incorporated into the kernel function, as specified in Eq. (12), to calculate the correlation of values between two locations. Accordingly, nonstationary spatial covariance could be easily estimated without any complex numerical techniques to approximate the embedded manifold geometry and the geodesic distances. MATLAB codes and a few illustrative examples developed herein will be published in conjunction with the publication of the manuscript.

To evaluate the estimation characteristics, a few two-dimensional (2D) hypothetical aquifer scenarios were developed. In these scenarios, a 2D domain discretized into $n_x \times n_y$, where $n_x = n_y$= 500, and a grid size of $\Delta x = \Delta y = 1$ m was consistently used throughout this study.

Furthermore, synthetic hydraulic conductivity data using unconditional simulation for a log-transformed hydraulic conductivity field was generated to evaluate the performance of the developed estimator. This simulation targeted the mean and variance of 0 and 1, respectively and assumed a specific nonstationary spatial structure (Figure 2a). A total of 50 locations (red dots) were randomly selected from the simulated field to compare against the actual measurements. This approach allowed the estimation of the accuracy and effectiveness of the developed estimator in predicting hydraulic conductivity in nonstationary conditions. Eq. (1) represents the XYZ surface, and

$$f(u,v) = \alpha_1 \sin\left(\frac{u}{\alpha_2}\right) + \alpha_3 v \#(19)$$

where α_1, α_2, and α_3 represent fitting parameters. In the unconditional simulation, Eqs. (12) and (17) were applied, where $\rho = 100$ m and $\sigma^2 = 10^{-2}$ were used.

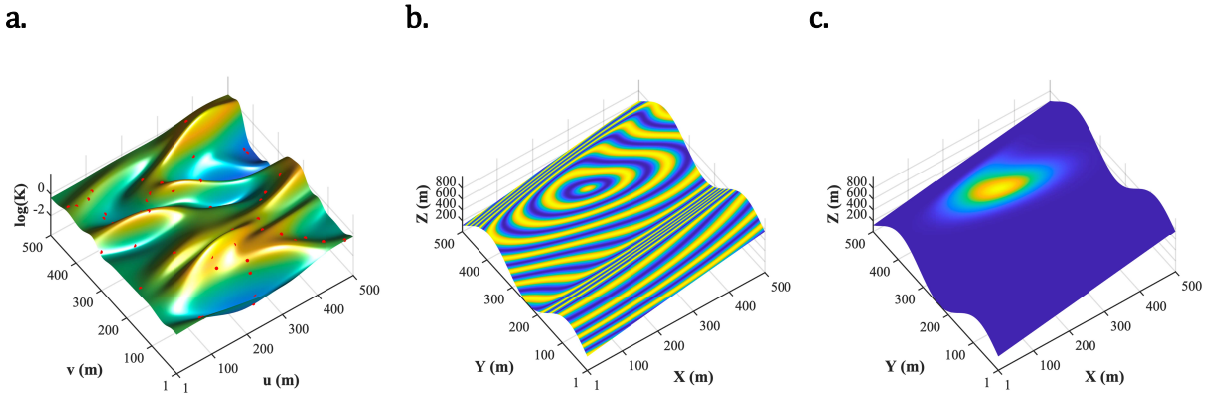

Figure 2. (a) Unconditional simulation result as the reference field for the estimation target with locations of synthetic data acquisition (red dots); (b) XYZ surface used as secondary information in the unconditional simulation with contours (yellow or blue color) of equal geodesic distance (250 m, 350 m); (c) the XYZ surface with RBF centered at (250 m, 350 m), where yellow and blue colors indicate values of 1 and 0, respectively.

The spatial structure constructed herein is similar to the geological fold structures that are frequently observed in sedimentary environments (Figure 2a). In cases of hydraulic conductivity estimation, Figure 2a was used as the reference field for the estimation target against which the performance of different estimations was evaluated. The XYZ surface [Eq. (1) and (19)] and the geodesic distance distribution at an arbitrarily selected location ($u = 250$ and $v = 350$) are represented in Figure 2b, where the arbitrarily chosen fitting parameters of $\alpha_1 = 250$, $\alpha_2 = 50$, and $\alpha_3 = 1$ are used. Figure 2b illustrates the distribution of equal geodesic distances from a given point (250 m, 350 m) on the XYZ surface. The pattern of correlation decay with distance is illustrated in Figure 2c, which depicts the radial basis function (RBF) (Eq. (12)) centered at (250 m, 350 m). The mean of the realized hydraulic conductivity field was observed to be –0.2802, with a standard deviation of 0.8589. Conversely, the mean of the obtained data at 50 locations was found out to be −0.3909, with a standard deviation of 1.001. In the remaining cases, these data were assumed to be the primary data obtained from the target domain, and the estimation target was set to the simulated hydraulic conductivity field in Figure 2a (reference case).

Using these data as conditioning primary information, a few scenarios based on the amount of information on the geological structure of the aquifer, which can be considered as secondary information, were developed. In the first scenario, it was hypothesized that information on the geological structure of the aquifer is lacking, and informative secondary data is not involved in estimation. In terms of the developed method equations, no secondary information scenario could be represented by assigning $f(u,v) = c$ in Eq. (1), which results in the metric tensor (Eq. 5) as

$$g = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \#(20)$$

where c is any constant value and 0 is the value used in this study. Notably, when Eq. (20) was applied to Eq. (9), the geodesic distance, d_g, became the Pythagorean distance, and the estimator (Eq. 17) was essentially identical to simple kriging.

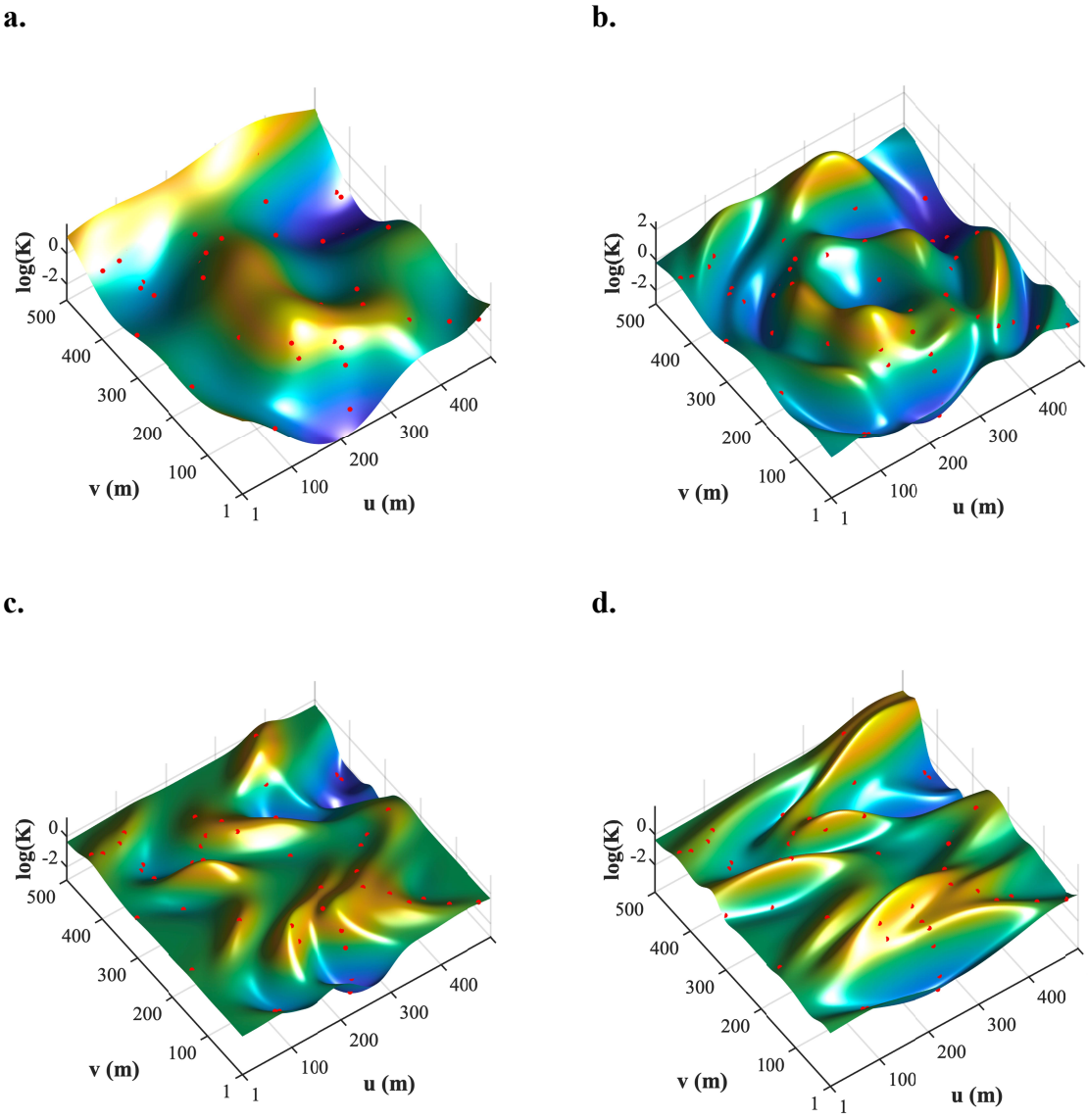

Figure 3. (a) Estimation of the hydraulic conductivity distribution without secondary information; (b) estimation with incorrect secondary information; (c) estimation based on geometry function (Eq. 1) using known geological structures with poorly optimized parameters; (d) estimation based on geometry function using known geological structures with optimized parameters.

Using the same parameter values of $\rho = 100$ m and $\sigma^2 = 10^{-2}$ in Eqs. (12) and (17), the estimated hydraulic conductivity field is represented in Figure 3a. Because it did not improve the estimation of this nonlinear spatial problem, the semi-variogram analysis was not performed to determine the spatial structure of the conditioning data. It is evident that the estimation in Figure

3a exhibits a completely different structure from that in Figure 2a, although the estimate accurately reflects the given conditioning data (red dots). The root mean square error (RMSE) for the conditioning data locations was found to be 0.311, indicating that the conditioning data are not correctly reflected in the estimation. Additionally, the RMSE calculated across the entire domain was 0.7259, and a large RMSE that was close to the standard deviation of the reference case suggests that the estimated field lacks information on the true hydraulic conductivity field across the entire domain. It was found that assuming the absence of geological structure leads to unreliable hydraulic conductivity estimations. To further improve the estimation quality, cross validation can be used to tune the parameter of ρ. However, any additional parameter improvements were not performed, as the improvement in estimation quality is limited for nonstationary fields, such as the reference case, in the absence of secondary information.

In the second scenario, it was hypothesized that the secondary data on the spatial structure were known inaccurately and were used for the estimation. This false information may be attributed to mistakes in geophysical processing or misinterpretation of geology. In terms of Eq. (1), $f(u, v)$ is arbitrarily selected to be

$$f(u,v) = \frac{\alpha_1 - (u-250)^2 + (v-250)^2}{2\alpha_2}\ , \#(21)$$

among which the metric tensor can be analytically derived from Eq. (5). The spatial structure generated using Eq. (21) exhibits a circular shape, which is reminiscent of the shape that some geological domes may assume due to specific geological processes (Figure 3b). Importantly, the conventional semi-variogram analysis is not applicable in this scenario, as it only provides information on linear spatial relationships. The estimated hydraulic conductivity field using $\rho = 50$ m for Eq. (12) and $\sigma^2 = 10^{-2}$ for Eq. (17) is represented in Figure 3b. The conditioning data (red dots) are well reflected in the estimate (red dots in Figure 3b) with an RMSE of 0.0498. However, the RMSE calculated across the entire domain was 0.8195, indicating that the estimation may not accurately capture the variability of the reference field (Figure 2a). Additionally, the spatial structure of the estimate is evidently different from that in the reference field. These observations indicate that the use of misinterpreted secondary information can lead to an erroneous estimation result; moreover, in some instances, it may also result in worse outcomes compared to when no secondary information is used. When using the developed estimator, it is crucial to consider the quality of secondary data that is used to inform the estimation process, as misinformed, biased, or inaccurate data can lead to prejudicial estimation results. Quality assurance and control measures can help mitigate the risk of misinformed secondary data and improve the accuracy and reliability of estimation results; however, the implementation of such measures was beyond the scope of this study.

In the third scenario, it is assumed that the secondary information accurately reflects the structural nonstationarity for the entire domain; therefore, Eq. (19) was applied to Eq. (1). However, it was hypothesized that the fitting parameters of α_1, α_3, and ρ were unknown and had to be estimated. Figure 3c demonstrates the results of GPR estimation using parameters arbitrarily set to $\alpha_1 = 120$, $\alpha_3 = 1.5$, and $\rho = 50$ m, which are used as initial guesses for the parameter estimation. Herein, the periodicity of the folding structure corresponding to α_2 was assumed to be accurately determined through a rigorous secondary data analysis, including geophysical data processing. Based on these parameters, the RMSEs for the conditioning data location and entire domain were found to be 0.0337 and 0.466, respectively. These findings

indicate that although the overall depiction of nonstationarity in the secondary information may be accurate, the estimation outcome can deviate from reality if precision is lacking in the details.

Consequently, the optimization of parameters α_1, α_3, and ρ was performed using the training to testing data ratios heuristically set at 80% and 20%. The interior point method (Nocedal, 2014) encoded in the MATLAB fmincon function was adopted for this parameter optimization. The estimated values of α_1, α_3, and ρ, obtained through parameter estimation, were 297.5336, 1.3732, and 114.5088, respectively. Although these values are close to the target values of 250, 1, and 100, respectively, they differ slightly, indicating that the estimation process may have been influenced by the limited amount of data. Based on the parameter estimations, we found that the conditioning data (red dots) are almost perfectly reflected in the estimate (red dots in Figure 3d) with the RMSE of 0.0166. Furthermore, the RMSE calculated across the entire domain was 0.2266, suggesting that the estimation accurately captures the variability of the reference field. This indicates that, under the given conditions, the developed estimator is effective in modeling the underlying geological processes and can provide reliable estimates in the field of interest when informative secondary data are available.

3.2 Perspectives to flow and transport problems

Two-dimensional steady-state groundwater flow and transient solute transport simulations were conducted to evaluate the impact of an accurate hydraulic conductivity estimation on groundwater flow and solute behavior. MODFLOW-2005 (Harbaugh, 2005) and MT3D-USGS (Bedekar et al., 2016) were applied for the groundwater flow and solute transport simulations, respectively. The groundwater flow simulation included modeling the spatial variations of hydraulic head and flow rates in the aquifer system and applying the estimated hydraulic conductivity distribution as input. Based on the flow simulation results, the solute transport simulation was conducted to model the transport of a contaminant introduced at a specific location (250 m, 100 m) with a constant concentration of 100 mg/L. In the flow simulations, specified heads of 100 m and 90 m were assigned along $y = 1$ and 500 m, respectively, and no flow boundaries were assigned along $x = 1$ and 500 m in all cases. Similarly, in the solute transport simulations, fixed concentration boundaries were assigned along $y = 1$ and 500 m, and zero flux boundaries were assigned along $x = 1$ and 500 m in all cases. A nonreactive solute was selected for the simulations so as to clearly observe the effects of different hydraulic conductivity fields on the solute transport behavior. The dispersivities of the media were set uniformly to 1 and 0.1 m along the longitudinal and transverse directions, respectively, in all simulations. Although allocating uniform dispersivity values to the entire domain may not accurately reflect the real heterogeneity of the aquifer system, this simplification was considered to be sufficient for the purpose of the present study, which focused on the impact of different hydraulic conductivity distributions on solute transport behavior. Using uniform dispersivity values, the effects of hydraulic conductivity on the solute transport behavior were determined, and the performance of the developed method in estimating the hydraulic conductivity distribution under simplified conditions was evaluated. However, it is crucial to note that the simplifications in the simulation parameters may undermine the practicality of the simulation results. Herein, the total simulation period for the solute transport simulation was set to 365 days, which allowed the transport analysis of the deployed solute plume over a realistic timeframe and assessment of the potential impacts on the groundwater quality.

Figure 4 depicts the log-transformed solute plume deployment for a concentration of 10^{-3} mg/L, utilizing the reference hydraulic conductivity field and the three different estimations 365 days after the release of the solute. The application of a logarithmic scale for the concentration values allows better visualization of the spatial distribution and concentration levels of the solute plume. The solute plume for the reference case (Figure 4a) exhibits a complex and nonuniform deployment pattern, which is strongly affected by the heterogeneity of the hydraulic conductivity distribution. This distribution varies spatially and is nonstationary over the domain. However, the solute plumes in the cases without (Figure 4b) or incorrect (Figure 4c) secondary information exhibit noticeable differences in deployment patterns compared to that of the reference case. The solute plume in Figure 4b, which does not account for the geological structure, exhibits a more uniform and homogeneous deployment pattern compared to that observed in the other cases. This observation indicates that neglecting the geological structure in the estimation process may result in oversimplification of the hydraulic conductivity distribution and cause less realistic and inaccurate solute transport modeling.

In cases wherein the interpretation of the geological structure is inaccurate (Figure 4c), the deployment of the solute plume differs significantly from that of the reference case. This result highlights the importance of accurately identifying and incorporating geological information in the estimation process and its effect on the reliability and accuracy of solute transport modeling in groundwater systems. When the geological information is incorporated as a function and the parameters are optimized (Figure 4d), the deployment of the solute plume shows considerable similarity to the reference case. In this scenario, the details of the range and branching pattern of the solute plume is well preserved, suggesting that the developed estimator can accurately capture the important aspects of the hydraulic conductivity distribution. This observation suggests that the developed method is effective in modeling the complex behavior of the hydraulic conductivity distribution and can provide reliable estimates in the field of interest.

Besides the qualitative comparison of plume similarity, the structural similarity index measurement (SSIM, Wang et al., 2004) was employed to provide a more quantitative comparison of the different solute plume deployments. The formulation of the SSIM is given by Eq. (22), as follows:

$$SSIM = \frac{(2\mu_T\mu_e + C_1)(2\sigma_{Te} + C_2)}{(\mu_T^2 + \mu_e^2 + C_1)(\sigma_T^2 + \sigma_e^2 + C_2)}, \#(22)$$

where μ_T and μ_e denote the means of the reference and estimated log-transformed concentrations exceeding –3, respectively; σ_T^2 and σ_e^2 represent the variances of the reference and estimated concentrations, respectively; and σ_{Te} denotes the covariance between the reference and estimated concentrations. For C_1 and C_2, an extremely small number of 2.22×10^{-16} was applied. The SSIM values between the reference case and the simulation using the estimate with no secondary information (Figure 4b) and that using the estimate with misinterpreted secondary information (Figure 4c) were 0.18 and 0.2115, respectively. These low values indicate a lack of similarity between the shapes of the reference and simulated plumes. Conversely, the simulation using the estimated hydraulic conductivity field with the correct geodesic distance function and optimized parameters yielded an SSIM value of 0.7517, suggesting a high degree of similarity between the reference and simulated plumes.

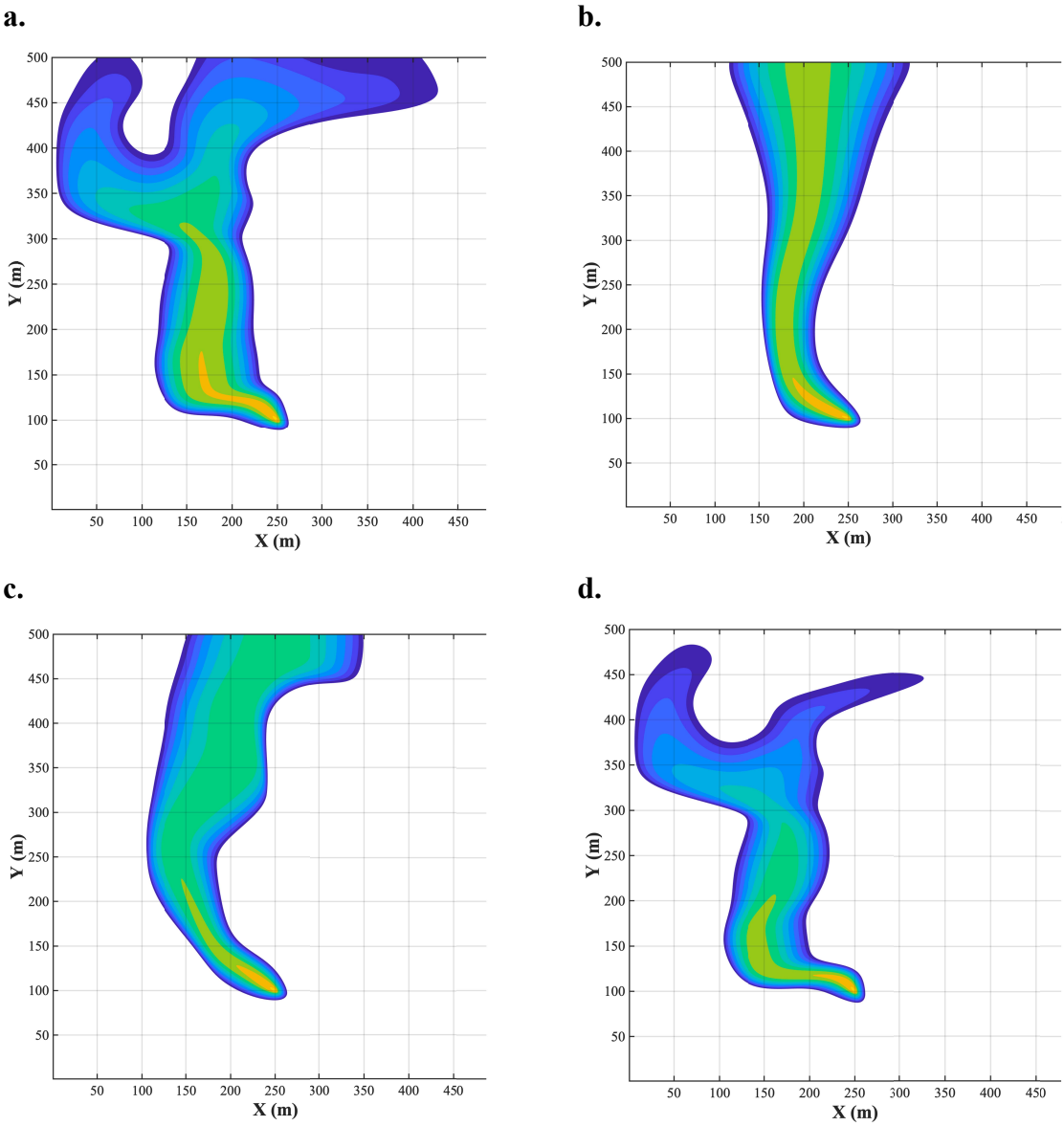

Figure 4. Simulated log-transformed solute plume distribution for a concentration exceeding 10^{-3} mg/L, using (a) the reference hydraulic conductivity field; (b) estimated hydraulic conductivity field without secondary information; (c) estimated hydraulic conductivity field with incorrect secondary information; (d) estimated hydraulic conductivity field based on correct geometry function with optimized parameters.

Overall, the results of the flow and transport simulation confirm that the deployment of the solute plume is highly sensitive to small-scale variability in the hydraulic conductivity distribution (Sudicky et al., 2010). The study findings highlight the need for careful consideration of small-scale variability and nonstationarity in the hydraulic conductivity estimation process; moreover, they underline the importance of developing accurate and reliable methods for estimating the hydraulic conductivity distribution in groundwater systems. More effective and reliable approaches for managing and mitigating the groundwater contamination

risks can be developed by improving our understanding of the effect of hydraulic conductivity variability on solute transport behavior.

4 Conclusions

This study developed and demonstrated the combined application of geodesic kernel and GPR for estimating the distribution of hydraulic properties, which generally show spatially varying and nonstationary characteristics. In this direction, a geodesic distance based on the manifold's intrinsic geometry was derived; moreover, a GPR based on the geodesic kernel was defined. The equations for the geodesic kernel and GPR were of a simple semi-analytical form, which allows their efficient evaluation using only a few lines of computational code.

Furthermore, the proposed approach was applied to several hypothetical scenarios to evaluate its effectiveness and the characteristics of the resulting estimations were analyzed. During the implementation of the developed approach, an unconditional reference hydraulic conductivity field was generated, which served as the foundation for the subsequent conditioning of the primary dataset. To introduce strong spatial nonstationarity in the hydraulic conductivity distribution, secondary information was hypothesized, assuming a specific geological structure that was used in the generation process. The resulting primary dataset was subsequently used in further estimations, with one case using only the primary information and the other two cases incorporating partial amounts of the secondary information to varying degrees. The results of the hydraulic conductivity estimations show the strong dependence of estimation accuracy on the use of secondary information for characterizing the nonstationary field. This sensitivity can be attributed to the high-dimensional nature of the developed estimator. Cross validation can be used to further improve the estimation, allowing optimization of the model parameters of the secondary information. Furthermore, the estimated hydraulic conductivity fields were used to simulate groundwater flow and solute transport, revealing the great dependence of the estimation accuracy of the hydraulic conductivity on the incorporation of secondary information in addition to the primary information, with the case that incorporated optimized secondary information exhibits the highest similarity to the reference case. This finding highlights the significance of utilizing secondary information, especially for characterizing nonstationary aquifers, for accurate hydraulic property estimation to ensure effective groundwater resources management. Notably, the developed approach is susceptible to overfitting similar to that in any high-dimensional method. This leads to poor estimation results in case of misinterpretation of the secondary information. Therefore, future studies should explore techniques to mitigate overfitting and ensure the robustness of the estimation process.

Although further studies are still required in this field, the current study represents a crucial initial step in exploring the method to incorporate secondary information into manifold geometry for geological processes. Results show the potential of the proposed approach to yield remarkable improvements in estimation accuracy and provide novel insights into the underlying structure of geological data. Particularly, the developed estimator can be incorporated into an inversion approach to improve accountability of nonstationary structures in the estimation, leading to more accurate and realistic estimates of hydraulic conductivity distribution. These insights have crucial implications in safety assessments and risk management strategies associated with groundwater contamination, where accurate and reliable estimation of hydraulic conductivity is critical to assess potential risks and develop effective mitigation measures.

Accordingly, further research can significantly enhance our understanding of subsurface geology and associated geological processes.

Acknowledgments

The authors declare that they have no conflict of interest. This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea Government (MSIT) (NRF-2020R1A2C2013606).

Open Research

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

References

Abramowitz, M., & Stegun, I. A. (1972). Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. National Bureau of Standards Applied Mathematics Series 55. Tenth Printing.

Bedekar, V., Morway, E. D., Langevin, C. D., & Tonkin, M. J. (2016). MT3D-USGS version 1: A US Geological Survey release of MT3DMS updated with new and expanded transport capabilities for use with MODFLOW (No. 6-A53). US Geological Survey.

Chiles, J. P., & Delfiner, P. (2009). Geostatistics: modeling spatial uncertainty (Vol. 497). John Wiley & Sons.

Crane, K., Weischedel, C., & Wardetzky, M. (2013). Geodesics in heat: A new approach to computing distance based on heat flow. ACM Transactions on Graphics (TOG), 32(5), 1-11.

Cressie, N. (1986). Kriging nonstationary data. Journal of the American Statistical Association, 81(395), 625-634.

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

Dagan, G., & Nguyen, V. (1989). A comparison of travel time and concentration approaches to modeling transport by groundwater. Journal of contaminant hydrology, 4(1), 79-91.

Deutsch, C. V., & Journel, A. G. (1992). Geostatistical software library and user's guide. Oxford University Press, 8(91), 0-1.

Fouedjio, F., Desassis, N., & Rivoirard, J. (2016). A generalized convolution model and estimation for non-stationary random functions. Spatial Statistics, 16, 35-52.

Goovaerts, P. (1997). Geostatistics for natural resources evaluation. Oxford University Press on Demand.

Gorelick, S. M., & Zheng, C. (2015). Global change and the groundwater management challenge. Water Resources Research, 51(5), 3031-3051.

Harbaugh, A. W. (2005). MODFLOW-2005, the US Geological Survey modular ground-water model: the ground-water flow process (Vol. 6). Reston, VA, USA: US Department of the Interior, US Geological Survey.

Higdon, D. (2002). Space and space-time modeling using process convolutions. In Quantitative methods for current environmental issues (pp. 37-56). Springer, London.

Jayasumana, S., Hartley, R., Salzmann, M., Li, H., & Harandi, M. (2015). Kernel methods on Riemannian manifolds with Gaussian RBF kernels. IEEE transactions on pattern analysis and machine intelligence, 37(12), 2464-2477.

Laloy, E., Hérault, R., Jacques, D., & Linde, N. (2018). Training-image based geostatistical inversion using a spatial generative adversarial neural network. Water Resources Research, 54(1), 381-406.

Madsen, R. B., Møller, I., & Hansen, T. M. (2021). Choosing between Gaussian and MPS simulation: the role of data information content—a case study using uncertain interpretation data points. Stochastic Environmental Research and Risk Assessment, 35(8), 1563-1583.

Mariethoz, G., Renard, P., & Straubhaar, J. (2010). The direct sampling method to perform multiple-point geostatistical simulations. Water Resources Research, 46(11).

Mariethoz, G., & Caers, J. (2014). Multiple-point geostatistics: stochastic modeling with training images. John Wiley & Sons.

Mariethoz, G. (2018). When should we use multiple-point geostatistics?. Handbook of mathematical geosciences: fifty years of IAMG, 645-653.

Max, A. W. (1950). Inverting modified matrices. In Memorandum Rept. 42, Statistical Research Group (p. 4). Princeton Univ.

Nocedal, J., Öztoprak, F., & Waltz, R. A. (2014). An interior point method for nonlinear programming with infeasibility detection capabilities. Optimization Methods and Software, 29(4), 837-854.

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

Patel, V. M., & Vidal, R. (2014, October). Kernel sparse subspace clustering. In 2014 ieee international conference on image processing (icip) (pp. 2849-2853). IEEE.

Pereira, M., Desassis, N., & Allard, D. (2022). Geostatistics for large datasets on Riemannian manifolds: a matrix-free approach. arXiv preprint arXiv:2208.12501.

Schölkopf, B., Smola, A. J., & Bach, F. (2002). Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT press.

Sudicky, E. A., Illman, W. A., Goltz, I. K., Adams, J. J., & McLaren, R. G. (2010). Heterogeneity in hydraulic conductivity and its role on the macroscale transport of a solute plume: From measurements to a practical application of stochastic flow and transport theory. Water Resources Research, 46(1).

Strebelle, S. (2002). Conditional simulation of complex geological structures using multiple-point statistics. Mathematical geology, 34, 1-21.

Wackernagel, H. (2003). Multivariate geostatistics: an introduction with applications. Springer Science & Business Media.

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing, 13(4), 600-612.

Yeh, T. C. J., & Liu, S. (2000). Hydraulic tomography: Development of a new aquifer test method. Water Resources Research, 36(8), 2095-2105.