



## Research paper

# Unraveling the uncertainty of geological interfaces through data-knowledge-driven trend surface analysis

Lijing Wang <sup>a,d,\*</sup>, Luk Peeters <sup>b</sup>, Emma J. MacKie <sup>c</sup>, Zhen Yin <sup>a</sup>, Jef Caers <sup>a</sup>

<sup>a</sup> Department of Earth and Planetary Sciences, Stanford University, United States of America

<sup>b</sup> Commonwealth Scientific and Industrial Research Organization (CSIRO), Land and Water, Australia

<sup>c</sup> Department of Geological Sciences, University of Florida, United States of America

<sup>d</sup> Earth and Environmental Sciences Area, Lawrence Berkeley National Laboratory, United States of America

## ARTICLE INFO

Dataset link: [https://github.com/lijingwang/da\\_knowledge\\_driven\\_trend\\_surface](https://github.com/lijingwang/da_knowledge_driven_trend_surface)

## Keywords:

Trend surface analysis

Implicit modeling

Level sets

Geological knowledge

Metropolis-Hastings sampling

## ABSTRACT

Modeling complex geological interfaces is a common task in geosciences. Many data sources are available for geological interface modeling, including borehole data and geophysical surveys. Geological knowledge, such as the delineation from geologists, is difficult to quantify but likely adds value to geological interface modeling. To integrate all information, we present a data-knowledge-driven trend surface analysis method to construct stochastic geological interfaces. We design a Metropolis-Hastings sampling framework to sample stochastic trend interfaces and quantify the uncertainty of geological interfaces given all information sources. This method is suitable for both explicit and implicit representations of geological interfaces. We demonstrate our method in three different test cases: modeling stochastic interfaces of Greenland subglacial topography, magmatic intrusion, and buried river valleys in Australia.

## 1. Introduction

Modeling geological interfaces is essential for subsurface characterization and natural resources forecasting. Geologists use various types of datasets to characterize geological interfaces, such as drilling or geophysical surveys. A common geostatistical surface modeling approach is to decompose surface variability into two components: a large-scale trend and a short-scale residual (Norcliffe, 1969; Chiles and Delfiner, 2009; Henrion et al., 2010; Fouedjio et al., 2021). The large-scale trend represents long-range tendencies observed in data, and the short-scale residual is represented by a stationary geostatistical field, often quantified by means of a variogram. Trend surface analysis (Agterberg, 1984), which studies how to model the large-scale trend, is often addressed by a smoothing algorithm, such as a moving window approach (Manchuk and Deutsch, 2011), a linear or polynomial fitting (Kyriakidis and Journel, 1999), or a radial-basis function (RBF) smoothing (Broomhead and Lowe, 1988). From these methods, we know that the goal of trend surface analysis is to model the smooth large-scale tendencies rather than to match or interpolate all data perfectly.

However, these trend surface analysis methods are often deterministic based on a single type of dataset (Pyrcz and Deutsch, 2014). The uncertainty of geological interfaces is left as a small variation within residuals, without modeling uncertainty in the trend surfaces. Additionally, geological knowledge from experts, such as the sequence

of sedimentary layers, the discontinuity from unconformity or faults, the conceptualized shape of orebodies, or the connectivity of buried systems from geologists' sketches, is difficult to make quantitative, hence used in mathematical interpolation methods. Yet, the knowledge likely adds geological realism to geological interface modeling and improves predictions (Rongier et al., 2014; Moradi et al., 2015).

Existing methods to advance geological realism have been proposed but also with some limitations. Object-based or process-based modeling (Paola, 2000; Deutsch and Tran, 2002; Henrion et al., 2010) construct highly realistic geological models but are less flexible to honor multiple datasets and require extra manual efforts to define interface topology. Training image-based geostatistical methods (Lochbühler et al., 2014; Mariethoz and Caers, 2014) address geological realism using the realistic training image, but detailed training images are not always available. Bayesian estimate fusion methods aim to combine estimates with different reliability or confidence in a consistent way (Peeters et al., 2010; Visser and Markov, 2019), but they rely on characterizing the reliability of individual estimates through formal likelihood functions, which can be challenging.

In the context of integrating geological knowledge, particularly regarding stratigraphic layers and faults, the potential field method (Lajaunie et al., 1997) has gained popularity. This method utilizes a 3D scalar field to represent multiple interfaces, where different isosurfaces

\* Corresponding author.

E-mail address: [lijingwang@lbl.gov](mailto:lijingwang@lbl.gov) (L. Wang).

in the scalar field correspond to sequence of layers/interfaces. Toolboxes such as LoopStructural (Grose et al., 2021b), GemPy (de la Varga et al., 2019), and GeoModeller (Calcagno et al., 2008) are available as open-source or commercial software solutions that interpolate potential field interfaces to build 3D geological models directly. LoopStructural, for example, incorporates fault kinetics by identifying fault frames and displacements, modeling pre-fault geometries, and subsequently adding fault offsets and displacements back (Grose et al., 2021a). GemPy offers geophysical inversion capabilities using Bayesian inference, where priors are perturbed initial interpolated potential fields and likelihoods are defined based on geophysical data measurements.

In this paper, we propose a new trend surface analysis method for uncertainty quantification of geological interfaces, given multiple datasets and geological knowledge beyond layers and faults. Rather than interpolating a single surface, this trend surface analysis method constructs multiple stochastic interfaces that meet user-defined criteria provided by data or geological knowledge. To achieve this, we translate both data and knowledge criteria into objective functions and design an optimization framework that stochastically perturbs interfaces by minimizing all objective functions, without relying on an initial interpolation. This optimization framework is applicable to both explicit and implicit modeling of geological interfaces. Implicit methods, which use 3D scalar fields to represent surfaces, are more flexible and become more popular to model complex geological interfaces without heavy representation by many explicit locations on irregular shapes (Fan et al., 2007; Calcagno et al., 2008; Iglesias et al., 2016; Yang et al., 2019; de la Varga et al., 2019; Grose et al., 2021b). Implicit modeling includes potential field modeling for multiple isosurfaces, or if only one interface is of interest, a signed distance function can be used by selecting a specific isosurface (often zero). To quantify the uncertainty of geological interfaces, previous works have used Monte Carlo error propagation (Wellmann et al., 2010; Lindsay et al., 2012; Pakyuz-Charrier et al., 2018, 2019) to estimate the uncertainty on implicit modeling, with an initial interpolated geological model and adding uncertainty in experts interested area. Additionally, Markov chain Monte Carlo methods incorporating additional geophysical measurements have also been proposed to implicit methods for uncertainty quantification, where they perturb a set of input parameters instead of the interface directly (Wellmann et al., 2018; de la Varga et al., 2019; Olieroor et al., 2021) to find the posterior of input parameters. Fouedjio et al. (2021) introduced an interface perturbation approach for optimizing the trend surface. They utilized the probability perturbation optimization (PPO) to identify the optimized correction (residual extension in their paper) for the initial trend estimation with the ultimate goal of matching drillhole measurements. In this paper, we design a Metropolis–Hastings sampling algorithm (Chib and Greenberg, 1995) to solve the optimization problem, where the stochastic realizations represent the uncertainty of geological trend interfaces. Our Metropolis–Hastings sampling algorithm perturbs the entire interface using stochastic velocity field to model the trend directly instead of finding the best correction on top of the initial interpolated field.

We organize this paper as follows. In the methodology section, we first review two different representations of geological interfaces (Section 2.1): explicit and implicit methods. We then propose our trend surface analysis method with data-knowledge integration, integrating different data types and geological knowledge using density functions (Section 2.2). Next, we design a Metropolis–Hastings sampling algorithm in Section 2.3 to sample from density functions on both explicit and implicit interface models, including an examination of the symmetric proposal distribution in Metropolis–Hastings sampling in Section 2.3.2. The result section presents three distinct case studies to show the generalizability of our method, from subglacial topography to magmatic intrusion to palaeovalley (buried valley) modeling. These case studies involve different data sources, geological knowledge, and both explicit and implicit representation of interfaces.

## 2. Methodology

### 2.1. Review: explicit and implicit modeling of interfaces

One way to categorize geological interface modeling methods is by dividing them into explicit methods and implicit methods. The explicit interface requires parameterizing locations of interface  $(x_i, y_i, z_i)$  using meshes or other forms of gridding. This explicit method is often used for elevation or depth models where each horizontal coordinate  $x_i, y_i$  has a unique  $z$  value, thereby the vertical coordinate of interface  $z$  is treated as a function  $z(x_i, y_i)$ . However, the explicit method is hard to express more complex 3D geometries containing irregular shapes.

The implicit method parameterizes an interface by means of an isocontour, or multiple interfaces through multiple isocontours of one three-dimensional continuous scalar field  $\phi(x_i, y_i, z_i)$ . For convenience, here we denote three-dimensional coordinates  $(x, y, z)$  as  $x$ . In implicit modeling,  $\phi(x)$  is still defined on a grid, but we can easily use a regular or adaptive Cartesian grid (Yang et al., 2006; Morgan and Waltz, 2017; Abgrall et al., 2014). If we focus on modeling one interface from one scalar field, locations  $x$  where  $\phi(x) < 0$  are considered to be inside or below the interface, where  $\phi(x) > 0$  are considered to be outside or above the interface. The scalar field  $\phi(x)$  equals 0 when  $x$  is on the interface. A special case of  $\phi(x)$  is called the signed distance function, where  $\phi(x)$  represents the distance to the interface  $\phi(x) = 0$ , and its gradient satisfies  $|\nabla\phi| = 1$  (Osher and Fedkiw, 2003). On the other hand, the potential field method uses multiple iso-contours from one scalar field to model multiple interfaces (Lajaunie et al., 1997), including stratigraphic layers, folds and faults, without any interfaces crossing scenarios. In this paper, we focus on modeling a single interface using the signed distance function.

Implicit methods can represent complex geometries without heavy representation by explicitly knowing all locations on irregular shapes. Another advantage is that we can use the level set equation to easily evolve and perturb the implicit interface and track the change of topology (Fan et al., 2007; Chen and Radke, 2009; Iglesias et al., 2016; Yang et al., 2019), which we will show to be useful in formulating optimization problems. The level set equation advects  $\phi$  with a normal velocity field  $v^*$  in the normal direction of  $\phi$ :

$$\frac{\partial\phi}{\partial t} + v^* |\nabla\phi| = 0 \quad (1)$$

The discretized version of Eq. (1) using the first-order forward Euler method is:

$$\phi^{n+1} = \phi^n - (v^* |\nabla\phi^n|) \Delta t \quad (2)$$

$n$  is the number of perturbations or iterations.  $\Delta t$  is the perturbation step size.

To show how the level set equation works for interface perturbation, we present an example of perturbing an implicit representation of a sphere using Eq. (2) in Fig. 1. We generate a Gaussian velocity field  $v^*$  and extend this velocity on the interface along the normal direction, which is named velocity extension. This new velocity extension field  $F^*$  guarantees that every point along the normal direction of the interface has the same perturbation. Direct perturbation by adding Gaussian fields to the implicit fields, without velocity extension (Yang et al., 2019), can result in unexpected isolated objects. This occurs because the direct addition does not perturb all iso-contours at the same time. Additionally, the velocity extension method avoids the need for recomputing the signed distances using the fast marching method (Adalsteinsson and Sethian, 1999). The velocity extension field  $F^*$  depends on the velocity values on the interface  $v_{\phi^n=0}^*$  and the normal direction of the interface  $\mathbf{n}_{\phi^n=0}$ . In Fig. 1, we perturb the sphere with the velocity extension field  $F^*$  using Eq. (3).

$$\phi^{n+1} = \phi^n - (F^* |\nabla\phi^n|) \Delta t = \phi^n - F^* \Delta t \quad (3)$$

In the following sections, we will perform trend surface analysis with explicit and implicit modeling approaches. Our trend surface

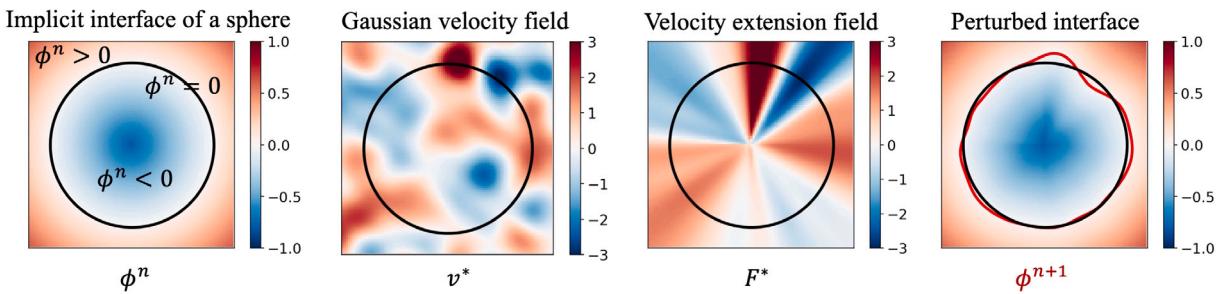


Fig. 1. Example of an implicit interface and its perturbation (2D view) using velocity extension.

analysis aims to construct large-scale trend interfaces that honor various datasets, including borehole data, radar data and electromagnetic survey, geological knowledge and many more. What we do not target to solve is the exact conditioning problem in geostatistical modeling, where we ensure a perfect match or interpolation of all hard data points. Hard data points, such as measurements obtained from boreholes, provide us with direct and precise information about the location of our geological interfaces. Exact conditioning would require an additional step, which can be performed using other approaches such as thresholding (Henrion et al., 2010) or residual/fine-scale analysis (Fouedjio et al., 2021) where the residual is often stationary (Chiles and Delfiner, 2009).

The trend surface analysis has two main steps: (1) formulating objective functions given multiple datasets and geological knowledge (Section 2.2) and (2) sampling trend interfaces stochastically using the Metropolis–Hastings algorithm (Section 2.3).

## 2.2. Formulating data-knowledge-driven objective functions

There are many types of datasets that can inform the trend of geological interfaces, such as drilled boreholes or different geophysical surveys, including electromagnetic, magnetic, radar, sonic, and gravity methods. Geological knowledge, on the other hand, is based on conceptual sketches or interpretations from geologists who leverage their expertise, such as structural geology or petrophysics. This knowledge incorporates insights of the connectivities or geometries of subsurface structures, the geological history, fault kinematics and other relevant delineations. To solve this challenging data-knowledge integration problem for trend surface analysis, we evaluate if an interface model  $z$  or  $\phi$  is a good fit to each dataset or knowledge separately using user-defined objective or loss functions:  $O_1, O_2, O_3, \dots$ . In the result section, we will present three different test cases with various objective functions, including the misfit with borehole data, radar survey, electromagnetic survey, and continuity of subsurface geological features (i.e. palaeovalley river channels) as geological knowledge.

We convert objective functions into density functions  $L_j$ :

$$L_j(\phi) = \exp\left(-\frac{O_j^2(\phi)}{2\omega^2}\right) \quad (4)$$

where  $\omega$  is a hyperparameter. This hyperparameter  $\omega$  impacts the value of density functions and acts as a regularization parameter. A small  $\omega$  is good for matching data exactly but practically makes the sampling difficult. A large  $\omega$  gives more uncertainty but may result in large unrealistic uncertainty. The density function can be compared to a Gaussian distribution:  $f(x) \propto \exp(-\frac{x^2}{2\sigma^2})$ . Therefore we can draw an analogy between  $\omega$  and the standard deviation  $\sigma$ . Depending on the objective or loss functions, we can determine a standard deviation for the loss function (i.e. measurement error) to select the  $\omega$  value. We suggest trying a relatively large  $\omega$  first compared to your loss function, and then tuning it down if there is too large uncertainty. We combine all density functions  $L_j(\phi)$  to construct our final density function  $L(\phi)$ .

$$L(\phi) = \prod_j L_j(\phi) = L_1(\phi)L_2(\phi)L_3(\phi)\dots \quad (5)$$

Density functions are commonly used in Bayesian inversion (Tarantola, 1987; Duijndam, 1988a,b) or optimization algorithms such as simulated annealing (Bertsimas and Tsitsiklis, 1993) or neighborhood algorithm (Sambridge, 1999a,b). This new representation enables us to sample from the combined density function  $L(\phi)$  directly using Markov chain Monte Carlo samplers and thereby quantifying the uncertainty of trend interfaces.

## 2.3. Stochastic sampling of trend interfaces

### 2.3.1. Metropolis–Hastings algorithm

The Metropolis–Hastings algorithm (Chib and Greenberg, 1995) is a Markov chain Monte Carlo that aims to sample iteratively from one target distribution when direct sampling is difficult. The target distribution in this paper is the density function (Eq. (5)). We design a perturbation scheme to update interfaces with a symmetric proposal distribution: zero-mean stationary Gaussian field  $v^*$  with spatial correlation at each iteration step:

$$z^*(x_i, y_i) = z^n(x_i, y_i) - v^*(x_i, y_i)\Delta t \quad (6)$$

$$\begin{aligned} \phi^*(x_i, y_i, z_i) &= \phi^n(x_i, y_i, z_i) - (v^*(x_i, y_i, z_i)|\nabla\phi^n(x_i, y_i, z_i)|)\Delta t \\ &= \phi^n(x_i, y_i, z_i) - F^*(x_i, y_i, z_i)\Delta t \end{aligned} \quad (7)$$

The explicit representation  $z^n(x_i, y_i)$  is directly perturbed by velocity fields  $v^*(x_i, y_i)$ . The implicit representation  $\phi^n(x_i, y_i, z_i)$  is perturbed by velocity fields  $v^*(x_i, y_i, z_i)$  using the level set equation (Eq. (2)) and the velocity extension (Eq. (3)). Instead of perturbing by  $v^*(x_i, y_i, z_i)$  directly,  $\phi^n$  is directly perturbed by the velocity extension field  $F^*(x_i, y_i, z_i)$ .

In each perturbation from  $\phi^n$  to  $\phi^*$ , we accept  $\phi^*$  with the probability  $p$  from the Metropolis–Hastings algorithm:

$$p = \min\left\{1, \frac{L(\phi^*)p(\phi^n|\phi^*)}{L(\phi^n)p(\phi^*|\phi^n)}\right\} \quad (8)$$

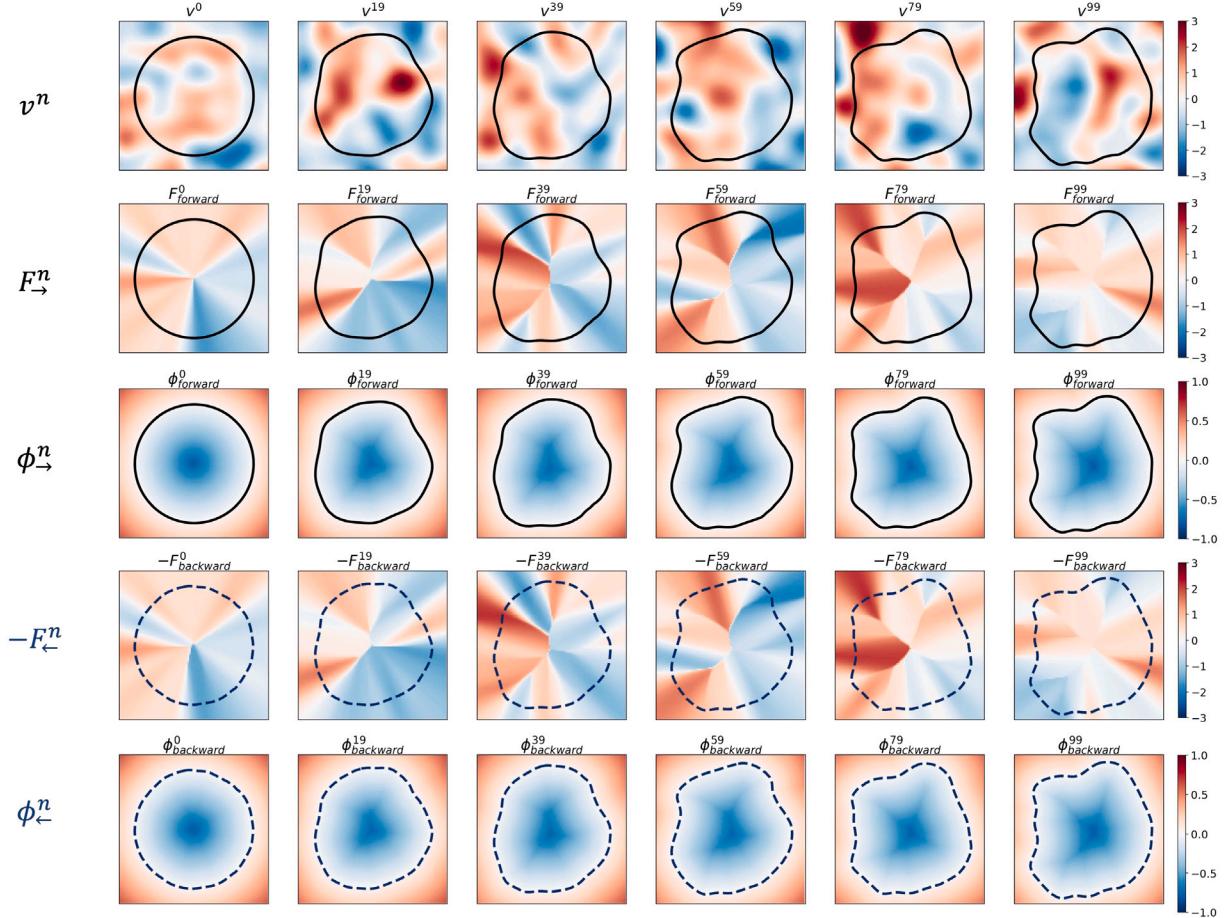
If accepted then  $\phi^{n+1} \leftarrow \phi^*$ , if not accepted then  $\phi^{n+1} \leftarrow \phi^n$ .  $p(\phi^*|\phi^n)$  is the forward proposal distribution from the current state  $\phi^n$  to  $\phi^*$ .  $p(\phi^n|\phi^*)$  is the backward proposal distribution from the perturbed state  $\phi^*$  back to  $\phi^n$ .

The trend interfaces sampled from the Metropolis–Hastings sampler include large-scale non-stationarity, even if the proposal distribution  $p(\phi^*|\phi)$  is based on the stationary Gaussian field  $v^*$ . This is because the target distribution  $L(\phi)$  is a data-knowledge-driven density function that is possible in any non-stationary distribution or complex geometry.

### 2.3.2. Forward and backward proposal distributions of trend interfaces perturbation

The acceptance probability  $p$  (Eq. (8)) depends on the forward and backward proposal distribution  $p(\phi^*|\phi)$  and  $p(\phi|\phi^*)$ . If the proposal distribution is symmetric:

$$p(\phi^*|\phi) = p(\phi|\phi^*) \quad (9)$$



**Fig. 2.** Forward and backward perturbations of a sphere. We start with one sphere and forward perturb this sphere using velocity extension 100 times. The columns show the perturbation from 0 to 100 times. The first three rows show the forward perturbation results. The black contour is  $\phi_{\rightarrow}^n = 0$ . The last two rows show the backward perturbation results. The backward perturbations use the same velocity field but with an opposite sign. The blue dashed contour is  $\phi_{\leftarrow}^n = 0$ .

where the forward proposal equals to the backward proposal, then the acceptance probability becomes fully known to us:

$$p = \min\{1, \frac{L(\phi^*)p(\phi|\phi^*)}{L(\phi)p(\phi^*|\phi)}\} = \min\{1, \frac{L(\phi^*)}{L(\phi)}\} \quad (10)$$

If the forward proposal does not equal to the backward proposal, then we need to evaluate those proposal distributions separately for every perturbation, which makes implementation difficult because it is specific to each case.

In the explicit modeling  $z(x)$ , a symmetric proposal distribution is the distribution of the Gaussian field:

$$z^* = z - v^* \Delta t \quad (11)$$

$$z = z^* + v^* \Delta t = z^* - (-v^*) \Delta t \quad (12)$$

Then the forward proposal  $p(z^*|z) = f(v^*)$ , the backward proposal  $p(z|z^*) = f(-v^*)$ . Our velocity field  $v^*$  is a symmetric Gaussian field. Therefore,  $p(z^*|z) = f(v^*) = f(-v^*) = p(z|z^*)$ .

However, in the implicit modeling, we do not directly perturb  $\phi$  using a stationary velocity field  $v^*$ . We perturb  $\phi$  with the velocity extension field  $F^*$ . Eq. (13) and (14) show that we forward perturb  $\phi$  using  $F_{\rightarrow}$  and backward perturb  $\phi^*$  using  $F_{\leftarrow}$ .

$$\phi^* = \phi - F_{\rightarrow} \Delta t \quad (13)$$

$$\phi = \phi^* - F_{\leftarrow} \Delta t \quad (14)$$

We cancel out  $\phi^*, \phi, \Delta t$  in Eq. (13) and (14), then

$$F_{\rightarrow} = -F_{\leftarrow} \quad (15)$$

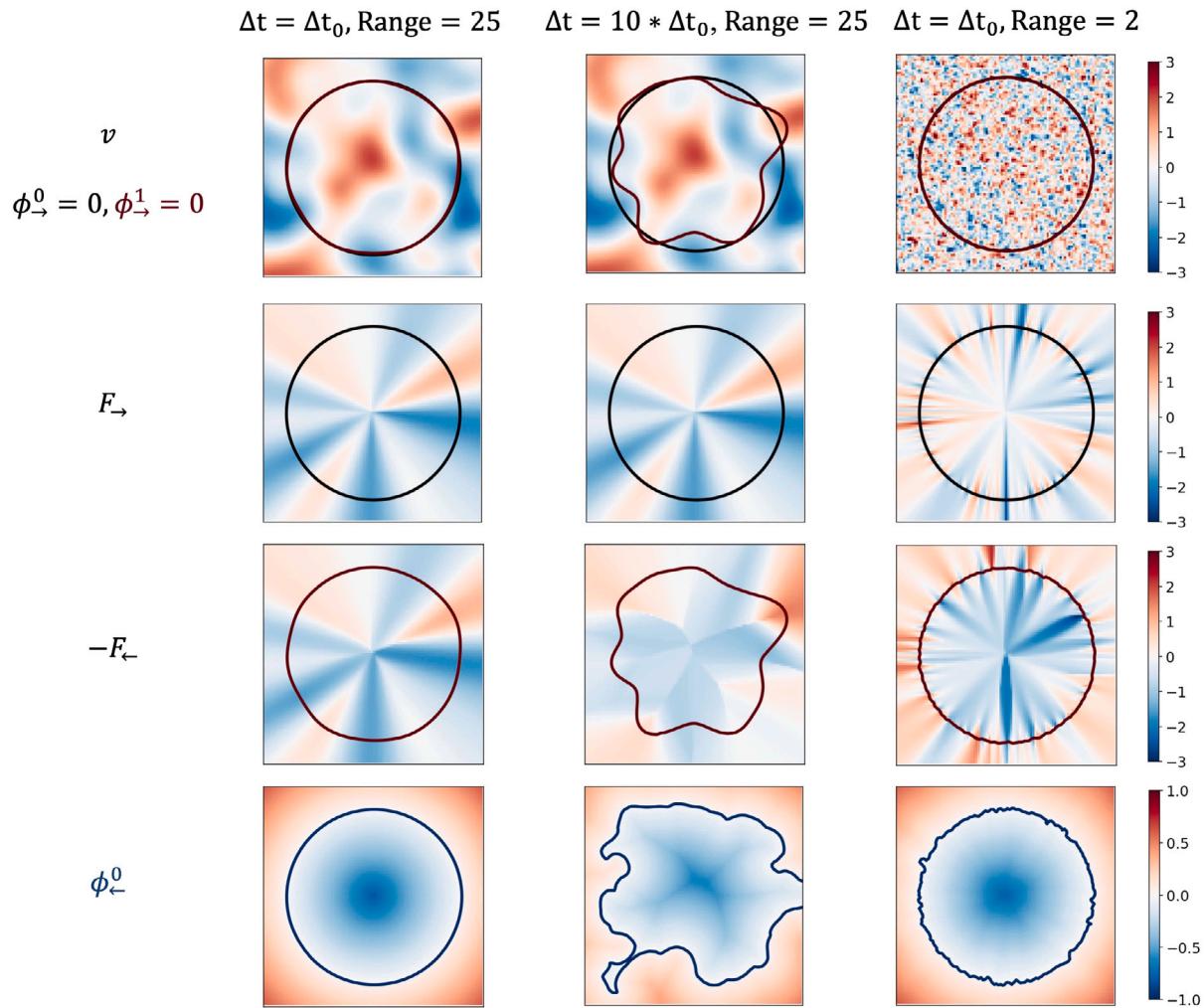
Based on the definition of the velocity extension field (Eq. (3)), if we assume (1) the normal direction of the interface does not change:  $\mathbf{n}_{\phi=0} = \mathbf{n}_{\phi^*=0}$ , (2) the forward velocity field  $v_{\rightarrow}$  has the same value at the interfaces  $\phi^* = 0$  and  $\phi = 0$ , then the backward velocity  $v_{\leftarrow} = -v_{\rightarrow}$  ensures Eq. (15) holds. Then again the forward proposal  $p(\phi^*|\phi) = f(v_{\rightarrow})$ , the backward proposal

$$p(\phi|\phi^*) = f(v_{\leftarrow}) = f(-v_{\rightarrow}) = f(v_{\rightarrow}) = p(\phi^*|\phi) \quad (16)$$

Then the forward and backward probabilities are equal:  $p(\phi^*|\phi) = p(\phi|\phi^*)$  is satisfied under the previous two assumptions and the choice of  $v_{\leftarrow} = -v_{\rightarrow}$ .

The example in Fig. 2 shows that the forward and backward probabilities are approximately equal to each other. We start from an implicit representation of a sphere  $\phi$  and perturb it. We forward perturb the sphere  $n = 100$  times using  $v_{\leftarrow}^n = v^n$  and then backward perturb  $\phi_{\rightarrow}^n$  using the same velocity field with an opposite sign  $v_{\leftarrow}^n = -v_{\rightarrow}^n = -v^n$ . We show  $F_{\rightarrow} = -F_{\leftarrow}$  is approximately satisfied based on the Row 2 and Row 4 of Fig. 2. Each perturbation does not drastically change  $\nabla\phi$ . The backward perturbation  $\phi_{\leftarrow}^0$  (Row 5 of Fig. 2) is back to a sphere.

However, the normal direction of a sampled interface  $\mathbf{n}_{\phi=0}$  might change significantly in stochastic perturbations. For example, if we use a large perturbation step size  $\Delta t$ , then  $\phi^*$  is drastically different from  $\phi$  so that  $\mathbf{n}_{\phi^*=0}$  is different from  $\mathbf{n}_{\phi=0}$ . In this case, our two forward and backward probabilities are not equal. **Another case of this non-symmetry happens when the values of  $v_{\rightarrow}$  are different at two interfaces**, even with a small perturbation. This might be because the Gaussian field  $v_{\rightarrow}$  we used has very small range. In geostatistics, the range refers to a correlation length that represents the spatial



**Fig. 3.** Different scenarios to test if the forward and backward probabilities are equal. The number of grids in this 2D view is  $100 \times 100$ . The last row is  $\phi_0^0$  with 10 times forward perturbation and 10 times backward perturbation.

correlation or continuity of a variable within a spatial domain. It measures how far apart two points need to be before they are not correlated to each other. The range in the velocity field shows the spatial variation of the perturbed velocity. A larger range indicates that the perturbed velocities are similar nearby; a smaller range indicates that the perturbed velocities vary significantly in a close distance.

Fig. 3 shows three different scenarios. Column 1 has a small perturbation step  $\Delta t_0$  and a range equals 25 grids in the velocity field. The total size of the velocity field is  $100 \times 100$  grids. Column 2 has a large ( $\times 10\Delta t_0$ ) perturbation step and the same range. Column 3 has a small perturbation step but a very small range equals 2 grids in the velocity field. Row 2 and Row 3 in Fig. 3 show that  $F_\rightarrow = -F_\leftarrow$  is only approximately satisfied for Column 1. The interface and the normal direction change drastically for only one step in Column 2. Column 3 has a small  $\phi$  change but the velocity field  $v_\rightarrow$  at the two interfaces are very different because of the small range. Row 4 in Fig. 3 shows that we perturb back to a sphere in Column 1, not perturb back to a sphere in Column 2 and Column 3.

To cancel out the forward and backward probabilities in Eq. (10), we need a small enough perturbation and a large enough range for the velocity field. The actual range value often needs trial and error testing. One way of exploring the range value is to sample a varying range within a distribution. The range inference is then a part of acceptance or rejection from the Metropolis–Hastings algorithm. For the step size  $\Delta t$ , one possible choice of the perturbation step size  $\Delta t$  is from the

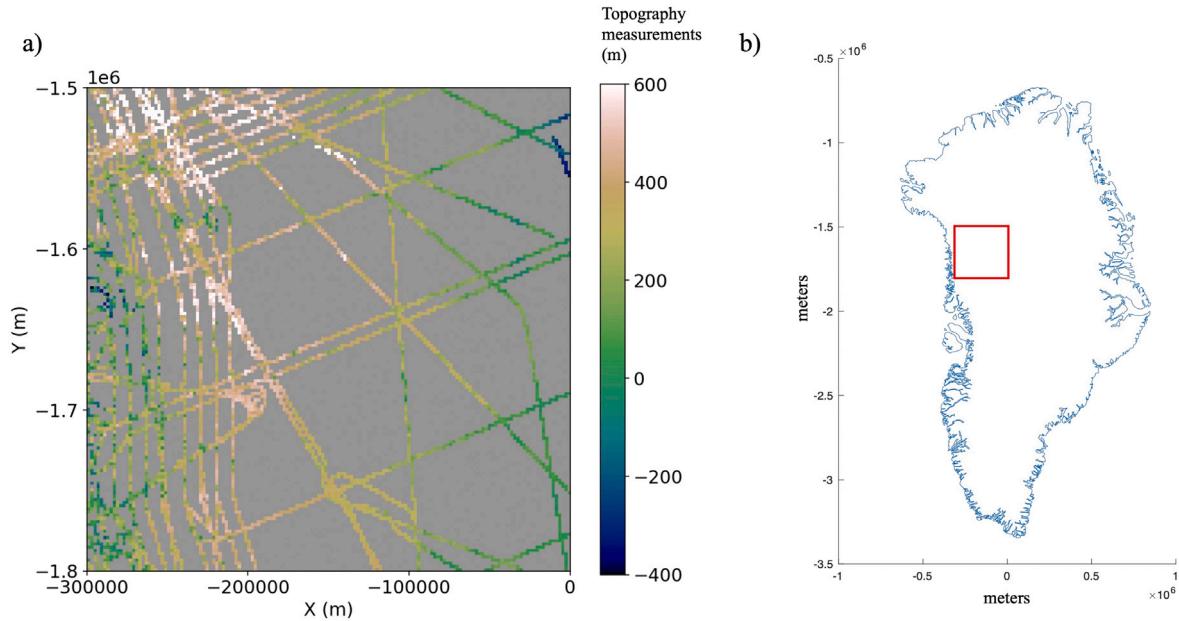
Courant–Friedrichs–Lowy (CFL) condition (Courant et al., 1967):

$$C = \frac{\max(F) * \Delta t}{\Delta x} \leq C_{max} \quad (17)$$

$$\Delta t = \frac{C * \Delta x}{\max(F)} \quad (18)$$

where  $C$  is called the Courant number. To guarantee the numerical stability, the maximum Courant number  $C_{max}$  is typically 1.  $\Delta t_0$  in Fig. 3 fits the CFL condition with  $C_{max} = 1$ . This condition keeps the perturbation step size small enough. Eq. (18) also suggests that  $\Delta t$  is a function of spatial discretization  $\Delta x$ .  $\Delta x$  is the size of each spatial grid.

This velocity field in the perturbation is a regularizer for the stochastic optimization method. We regularize each perturbation by a small change of  $\Delta\phi$  and a velocity field with a large range. In other words, we enforce smoothness of the perturbed interface. A rough velocity field within perturbations cannot satisfy this symmetric proposal distribution as we show in Fig. 3 Column 3. Our stochastic perturbation scheme is aperiodic (no period) and irreducible (always possible to move from one sample to another). And detailed balance is also enforced by the formulation of Metropolis–Hastings algorithm (Chib and Greenberg, 1995). These three properties are the fundamental properties for a Markov chain Monte Carlo (MCMC) algorithm. Therefore, our perturbation scheme is a MCMC method. The samples generated from this MCMC method (after the burn-in stage) quantify trend interface uncertainty given data and knowledge constraints.



**Fig. 4.** Ice-penetrating radar data of subglacial topography in northwest Greenland. (a) Topography measurements (m) with a polar stereographic projection. (b) The locations of our study area in Greenland, represented in polar stereographic coordinates.

### 3. Results

This section has three different interface modeling examples to illustrate our trend surface analysis method with multiple datasets and knowledge constraints. The first case uses the explicit interface modeling, the second and the third case use the implicit modeling approach.

#### 3.1. Test case 1: modeling subglacial topography in Greenland

We apply our trend surface analysis approach first to the interpolation of ice-penetrating radar measurements of subglacial topography in Greenland. Subglacial topography and roughness are critical parameters in ice-sheet models and are a leading source of uncertainty in hydrological models and ice-sheet projections (MacKie et al., 2021; Wernecke et al., 2022; Law et al., 2022). The bed elevation measurements are from the Center for the Remote Sensing of Ice Sheets (CReSIS, 2022) over a  $300 \times 300 \text{ km}^2$  region in northwest Greenland (Fig. 4). This area has irregular radar flight line spacings (Fig. 5a) and a large-scale non-stationary trend (Fig. 5b). Existing trend estimation methods, such as the radial-basis function (RBF) (Broomhead and Lowe, 1988) method, require us to provide a smoothing parameter that is sensitive to flight lines locations, and often use only one deterministic trend to represent the large-scale tendency.

In this case study, we model multiple stochastic trends using our trend surface analysis and compare them with a single deterministic radial-basis function (RBF) (Broomhead and Lowe, 1988) smoothing trend. We use explicit modeling  $z(x)$  for subglacial topography. Instead of doing geophysical inversion, we extract locations of bed topography from the ice-penetrating radar and use these locations in the density function  $L_1$ . The topography at the measured locations from an ice-penetrating radar is  $z_0(x_i, y_i)$ . Then the density function here is

$$L_1(z^n) = \exp\left(-\frac{\frac{1}{K} \sum_{i=1, \dots, K} (z_0^n(x_i, y_i) - z_0(x_i, y_i))^2}{2\omega^2}\right) \quad (19)$$

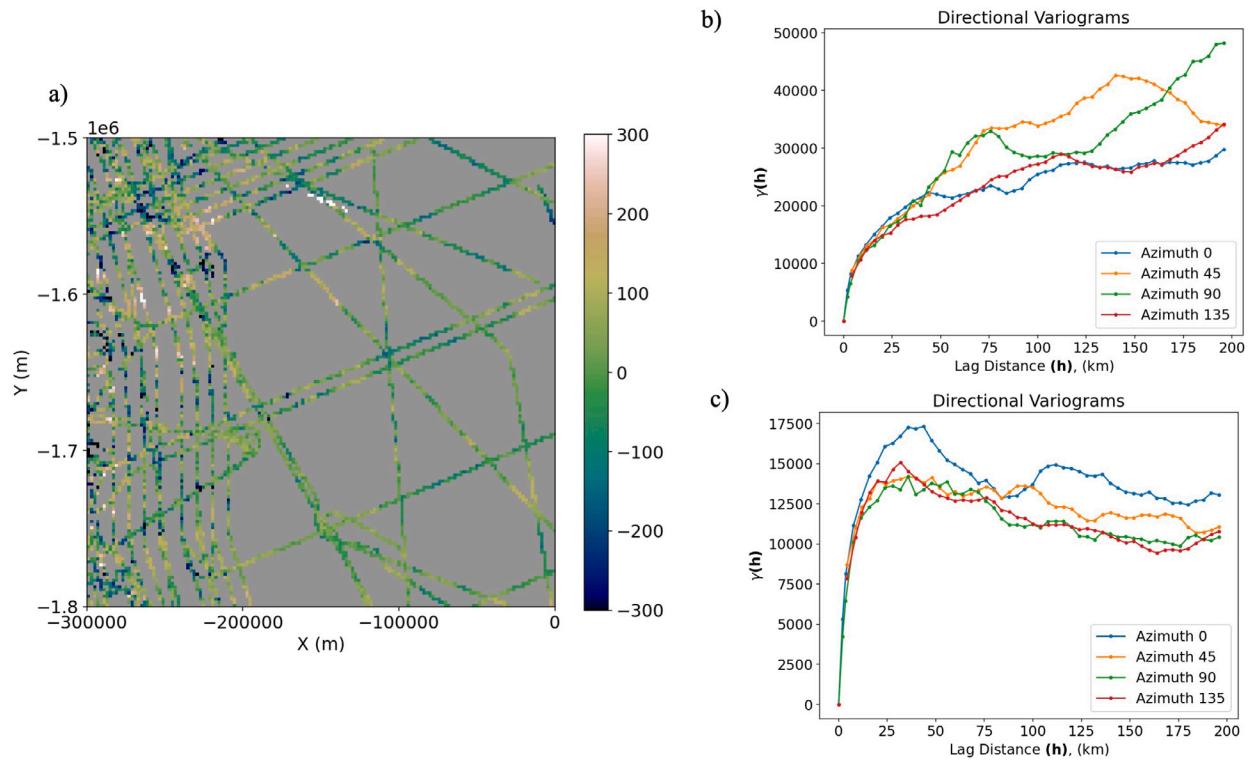
where  $K$  is the total number of measurements. The hyperparameter  $\omega = 500$  in this case.  $z_0^n(x_i, y_i)$  is the trend surface  $z^n$  values at observed locations.

We discretize the domain into cells with a cell size of 2 km and sample stochastic trends. The Metropolis-Hastings sampling result for

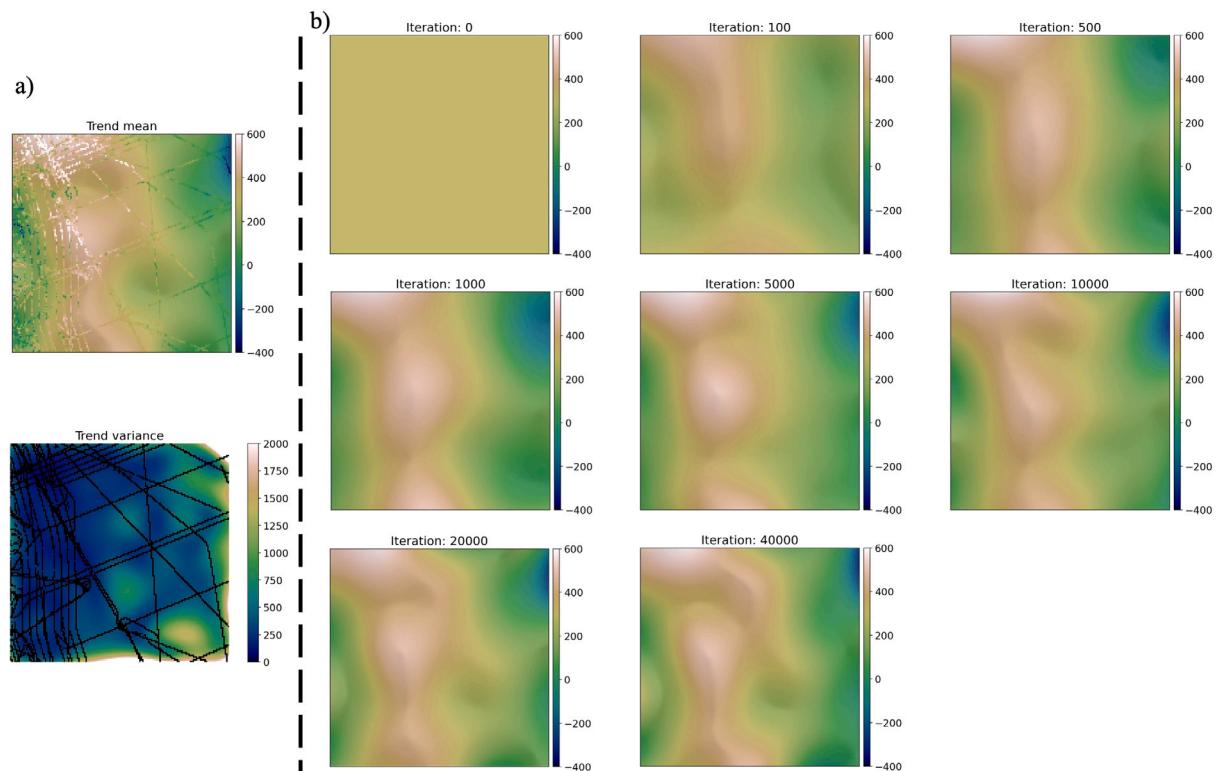
the trend modeling is in Fig. 6. We start from a flat topography and iteratively sample the trend realizations (Fig. 6b). The loss curve (Fig. 7a) decreases and flattens out after 5000 iterations. We take samples after iteration 5000 and calculate the mean and variance of sampled stochastic trends (Fig. 6a). The trend variance is higher in the sparse data area than in the dense flight lines area. The converged acceptance rate (Fig. 7a) is 0.22, which is close to the optimal acceptance rate 0.234 in Gelman et al. (1997). We calculate the autocorrelation for each grid after 5000 iterations and take the average of all autocorrelation. The mean autocorrelation is zero at 6000 lags (Fig. 7c).

There needs to be some trial and error for velocity range estimation in real applications. Here we investigate how different velocity ranges of the proposal stationary Gaussian distribution affect the sampling convergence. We have tested five different ranges: 20 km, 40 km, 60 km, 100 km, 150 km, and sampled with 50,000 iterations. Sampling convergences are monitored by the Gelman–Rubin convergence diagnostic (Gelman and Rubin, 1992). This statistic was originally used to test multiple chains' convergence. It compares the within-chain variance of every parameter of interest to the variance of all chains mixed together. If the variance difference is small, then the statistic will be close to 1. In practice, this statistic  $< 1.2$  means we have achieved convergence in our MCMC sampling. This statistic is also used to test the chain convergence within one chain. We separate one chain with 50,000 samples into two chains: one from iteration 20,000 to 30,000, one from iteration 40,000 to 50,000, and calculate the Gelman–Rubin statistic for different proposal distributions (Fig. 8). Parameters of interest are the mean and variance of the entire sampled area. Fig. 8a shows that the convergence within one chain holds (the statistic  $< 1.2$ ) only with a larger range ( $\geq 60$  km). We also present trace plots (Fig. 8c,d) for a large range (100 km) and a small range (20 km). The variance (orange line, 8d) of the small velocity range increases, and the location 2 topography (red line, 8d) also has drastic changes. A small velocity range adds unnecessary variations, especially in the sparse data area. Therefore, we suggest that the minimum velocity range should be larger than the biggest data gap. Another parameter in the proposal distribution, the perturbation step size  $\Delta t$ , controls the perturbation variance, thereby the acceptance rate. In Case 1, we choose  $\Delta t$  equals 20 m with the acceptance rate 0.22.

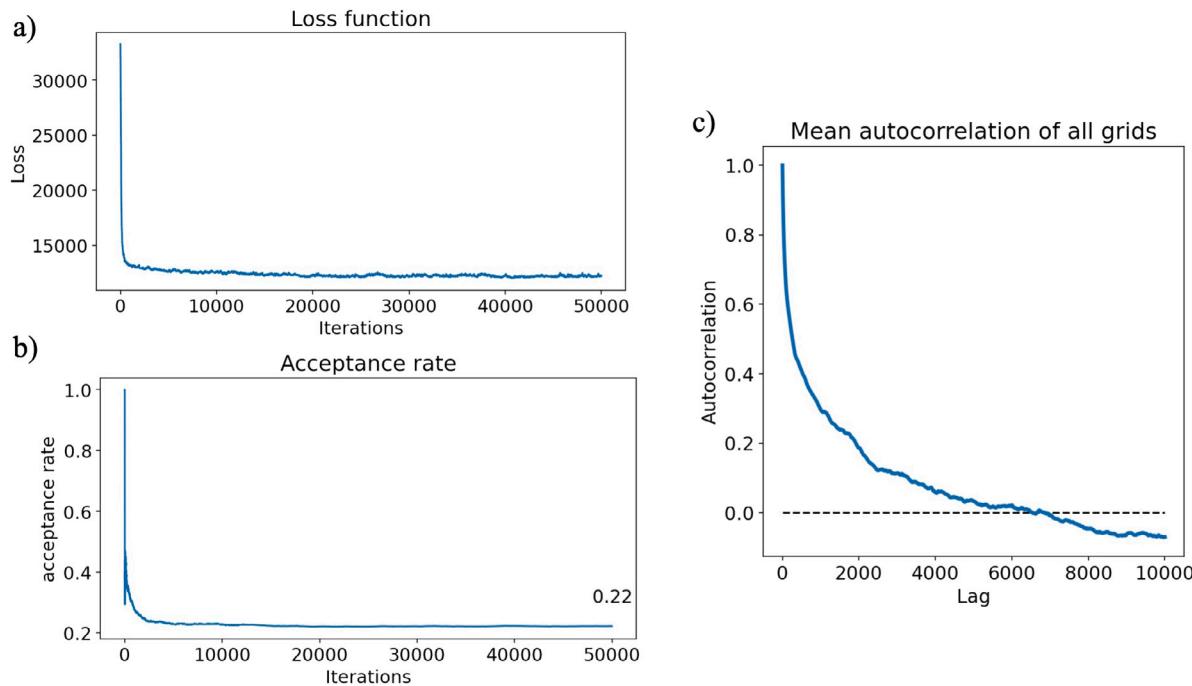
The exact conditioning on the radar flight lines can be performed with additional residual simulations. Given each trend, we compute the



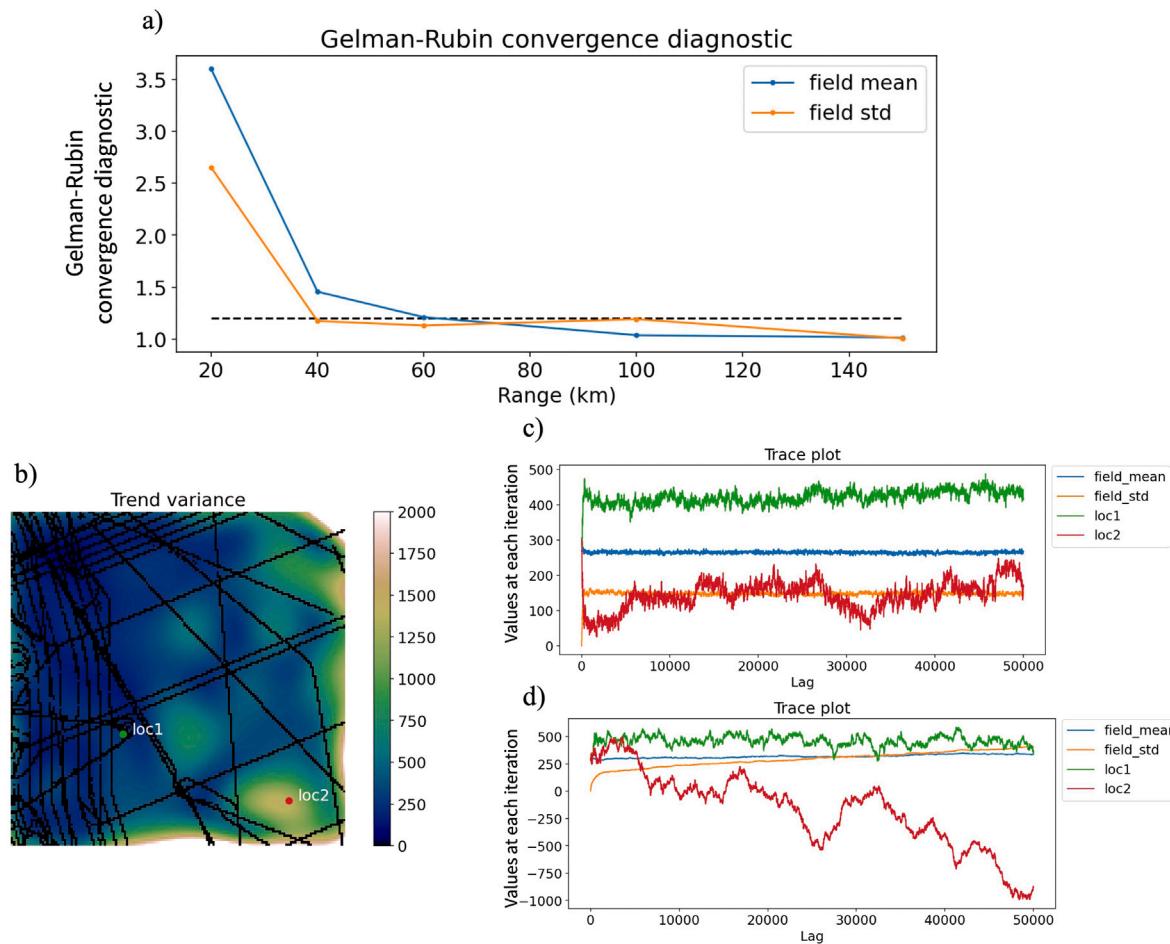
**Fig. 5.** Variogram on original data and residual. (a) Residual data after trend removal. (b) Variogram on the original data. (c) Variogram on the residual data. For (b) and (c), variogram  $\gamma$  units are typically expressed in the square of the units used for the variable being analyzed. Therefore, the unit is the square meter.



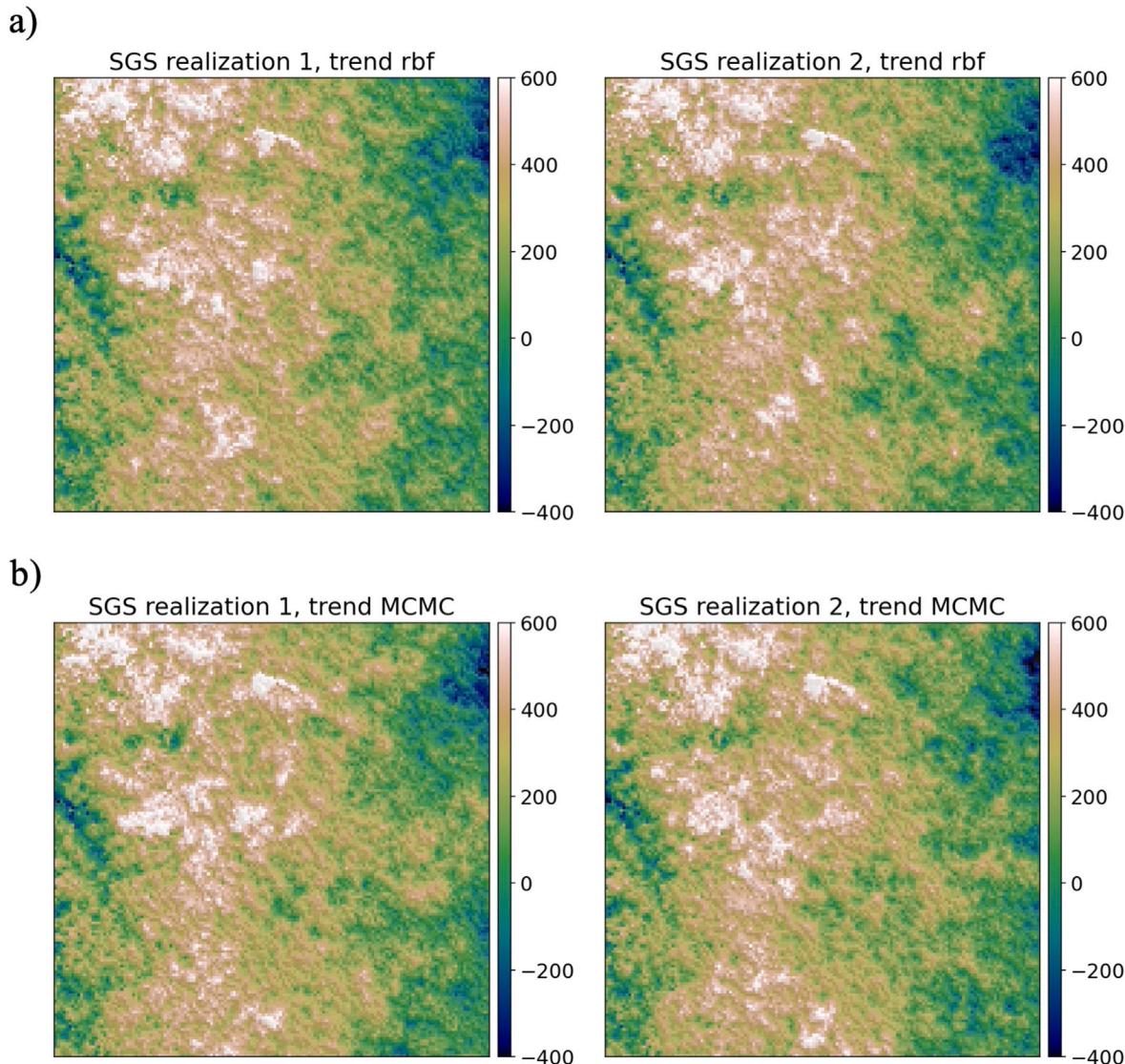
**Fig. 6.** MCMC sampling results: stochastic trend estimations for subglacial topography. (a) Trend mean and variance (b) Iterative sampling results.



**Fig. 7.** Case 1 loss function, acceptance rate and mean autocorrelation: (a) loss function. (b) acceptance rate, with the converged acceptance rate at 0.22. (c) the mean of the autocorrelation for each grid after 5000 iterations, the dashed line represents zero autocorrelation.



**Fig. 8.** Case 1 convergence test with varying velocity field ranges. (a) Gelman-Rubin convergence diagnostic for field mean and field standard deviation with different velocity ranges. The dashed line represents Gelman-Rubin convergence equals 1.2. (b) Two locations for trace plots in (c) and (d). (c) Trace plot with velocity range = 100 km. (d) Trace plot with velocity range = 20 km.



**Fig. 9.** Two sequential Gaussian realizations given (a) one deterministic trend from RBF (b) stochastic trends.

residual at flight lines (Fig. 5a), simulate residual values on all cells (Fig. 9b) using sequential Gaussian simulation (Deutsch and Journel, 1998; Journel and Ying, 2001) and add back each trend. Residual values after the trend removal have no trend in the variogram (Fig. 5c). The sequential simulations in Fig. 9 are full subglacial topography simulations with both trend and residual.

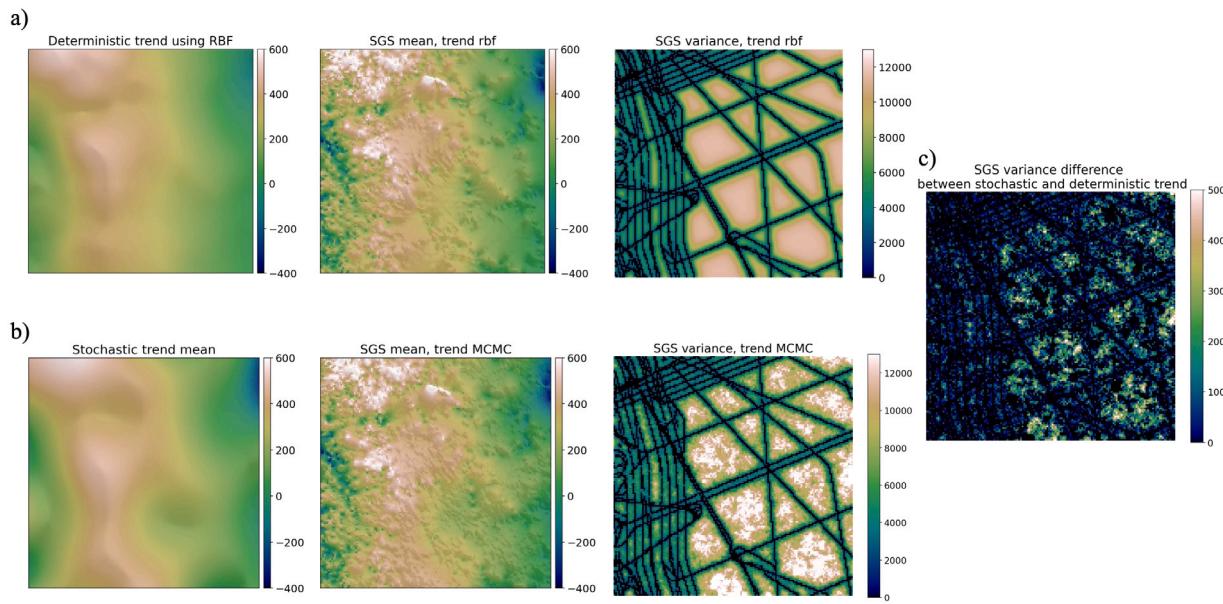
To understand the difference between two trend surface methods: a deterministic RBF trend and our proposed trend surface analysis methods, we compare the mean and variance of their sequential Gaussian simulations in Fig. 10. Both of them have zero variance in measured data locations. Sequential Gaussian simulations with stochastic trends have a larger variance in flight line gaps than with a deterministic RBF trend (Fig. 10c). The variance of sequential Gaussian simulations with a deterministic RBF trend is directly controlled by both the variance of the residual data. The variance of sequential Gaussian simulations with stochastic trends is controlled by the variance of the trends and the residual data. Our comparison shows that our trend surface analysis accounts for the trend uncertainty, which is a part of large-scale uncertainty that has often been ignored in geostatistical surface modeling.

### 3.2. Test case 2: modeling a magmatic intrusion

In the second case, we apply our trend surface analysis method to model a 2D magmatic intrusion, with five synthetic drilled boreholes (Fig. 11a) taken from a magmatic intrusion example (Fig. 11b). The magmatic intrusion can be a source of valuable minerals, such as copper, gold, and iron oxide, and are often the target of mineral exploration efforts. Unlike the first continuous example of subglacial topography, this second intrusion case considers a single categorical body with very coarse boreholes.

We model the intrusion geometry using the implicit model  $\phi(x)$ . The borehole data includes contact points at the interface locations:  $\phi_0(x_i, y_i, z_i) = 0$  and categorical hard data indicating intrusion or non-intrusion (a binary category). The density function  $L_2$  on the contact point measurement is:

$$\begin{aligned} L_2(\phi^n) &= \exp\left(-\frac{\frac{1}{K} \sum_{i=1,\dots,K} (\phi^n(x_i, y_i, z_i) - \phi_0(x_i, y_i, z_i))^2}{2\omega^2}\right) \\ &= \exp\left(-\frac{\frac{1}{K} \sum_{i=1,\dots,K} \phi^n(x_i, y_i, z_i)^2}{2\omega^2}\right) \end{aligned} \quad (20)$$



**Fig. 10.** Deterministic RBF trend and stochastic trends comparison after running sequential Gaussian simulations (SGS) on residual. (a) SGS simulations mean and variance with one deterministic trend (b) SGS simulations mean and variance with stochastic trends (calculated from 100 simulations) (c) Difference of SGS variance between two trend methods.

The density function  $L_3$  on the categorical borehole data is:

$$L_3(\phi^n) = \exp \left( - \sum_{i=1, \dots, K_{\text{intrusion}}} \frac{1}{2 \log(2)\omega^2} \log(1 + \exp(-\phi^n(x_i, y_i, z_i))) \right. \\ \left. - \sum_{i=1, \dots, K_{\text{non-intrusion}}} \frac{1}{2 \log(2)\omega^2} \log(1 + \exp(\phi^n(x_i, y_i, z_i))) \right) \quad (21)$$

Eq. (21) includes the convex loss function for the mismatch of categorical hard data, proposed in Fouedjio et al. (2021). A higher density means the intrusion locations have positive  $\phi^n$  and the non-intrusion locations have negative  $\phi^n$ . The hyperparameter  $\omega = 15$  in these two density functions.

We discretize this modeling area into  $130 \times 230$  cells with 1 m resolution. We start from an ellipse (Fig. 12b) and perturb it with velocity extension (Eq. (13)) using stationary Gaussian velocity fields. Our method does not require knowing exactly the range. The range value can also be uncertain and become a part of inference from the Metropolis–Hastings algorithm. In this case, each perturbation has a velocity range sampled sampling from a uniform distribution of  $U(50, 100)$ , for both depth and  $x$  directions. These parameters such as range of velocity fields are the hyperparameters, structural parameters or global parameters (Kitanidis, 1995; Malinverno and Briggs, 2004; Emerick, 2016; Xiao et al., 2021; Wang et al., 2022), where we can estimate using hierarchical Bayesian methods.

Fig. 12a have the mean and standard deviation of the signed distance function  $\phi^n$  and the discrete representation  $\phi^n > 0$  over all iterations excluding the first 5000 samples. This 5000 cutoff is estimated from the loss function curve (Fig. 13a). The borehole data area has the lowest variance. In areas far from data, especially at the low right boundary of the system, our intrusion models are more uncertain. Similar to Case 1, our sampled interface is a trend interface, which does not perfectly condition on all contact points and borehole data. Additional conditioning steps, such as ensemble smoother (Emerick and Reynolds, 2013; Caers et al., 2022) or randomized quadratic programming (Fouedjio et al., 2021) can be used to further match all the data. Since the trend surface mismatch in Case 2 is insignificant (Fig. 12), we skip this conditioning step.

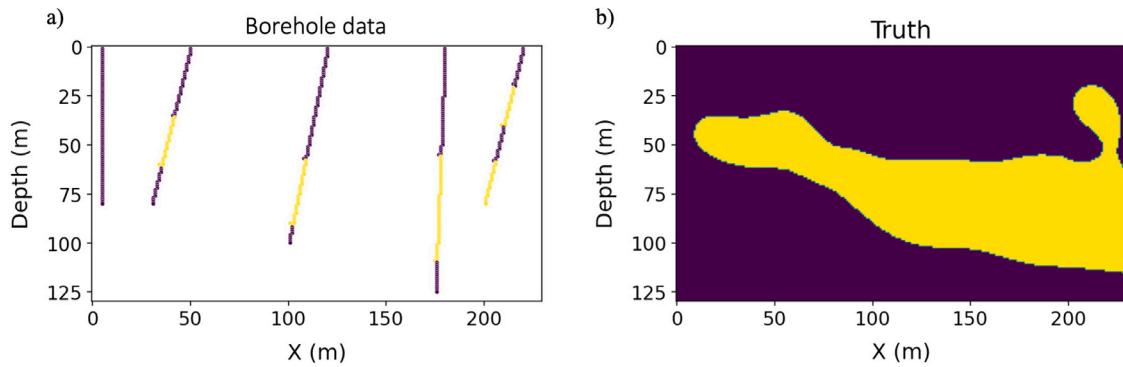
The acceptance rate for this case is very high at 0.68. This is back to the choice of the step size  $\Delta t$ . In implicit model perturbations, we

follow the CFL condition (Eq. (18)) with  $C = 1$ . Therefore, there is an upper limit for the step size  $\Delta t$ , and this capped step size causes a higher acceptance rate in this case study. Bedard (2008) suggests that the optimal acceptance rate might be drastically different than 0.234. Instead of focusing on the acceptance rate, we show the actual samples' uncertainty and trace plots. Fig. 14 shows four trace plots on the field mean of  $\phi^n$ , field standard deviation of  $\phi^n$ , and signed distance values at two locations with small variance and high variance. Within one chain, the value at location 2 has changed drastically and converged to different values. This high variance at location 2 and possibly a bi-modal posterior is not surprising given this sparse borehole data case.

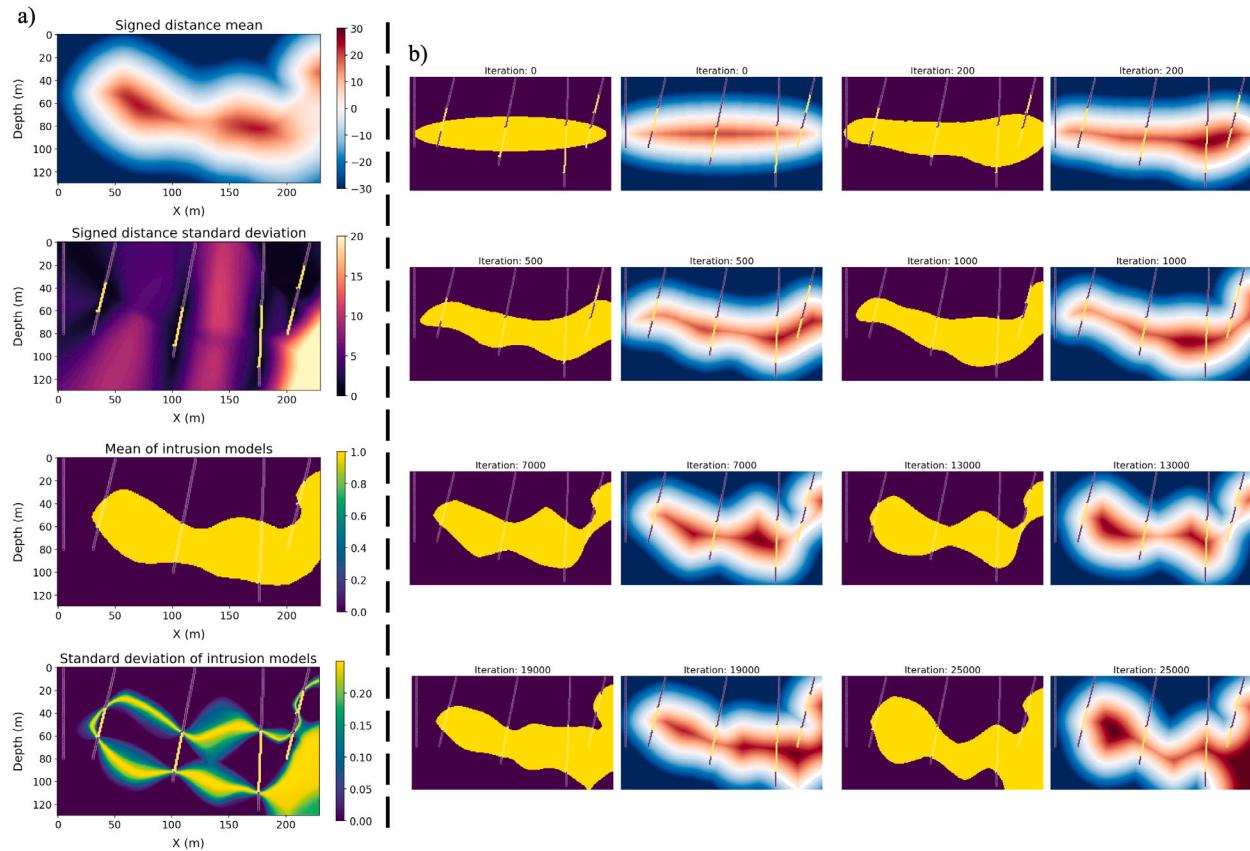
### 3.3. Test case 3: modeling palaeovalley structure in Australia

Palaeovalleys are buried river valleys that are not active surface water systems but good aquifers for groundwater storage. Groundwater in palaeovalleys might become the only available water resource for local communities in remote desert areas in Australia. To map the 3D palaeovalley in the subsurface, the South Australian Government acquired an extensive airborne electromagnetic (AEM) survey with a regional line spacing of 2 km in the Musgrave Province through the Goyder Institute for Water Research's Facilitating Long-term Outback Water Solutions Stage 3 (G-FLOWS S3) project (Munday et al., 2020a,b). More dense AEM surveys are not economically viable. Munday et al. (2020a,b) interpreted the sparse AEM survey and delineated the 2D thalweg (the line connecting the deepest point of river valleys) of ancient buried rivers. This additional geological knowledge is important for modeling complex geological structures. To secure the groundwater supply, more drilling and aquifer testing are in planning. A palaeovalley structure mapping with uncertainty quantification is critical for the future drilling plan.

In this paper, we apply our data-knowledge-driven trend surface analysis in mapping 3D palaeovalley structures, in order to include the information from both the AEM survey and geological knowledge from the 2D thalweg. We combine the sparse AEM survey (Fig. 15b) and the delineated connectivity — thalweg (Fig. 15c) in a pilot area (16 km  $\times$  10 km) that is south of Kaltjiti/Fregon in the Musgrave Province, South Australia (Fig. 15a). We model the palaeovalley structures using the implicit model  $\phi(x)$ .



**Fig. 11.** A synthetic 2D magmatic intrusion. (a) borehole data, (b) a synthetic 2D magmatic intrusion truth.

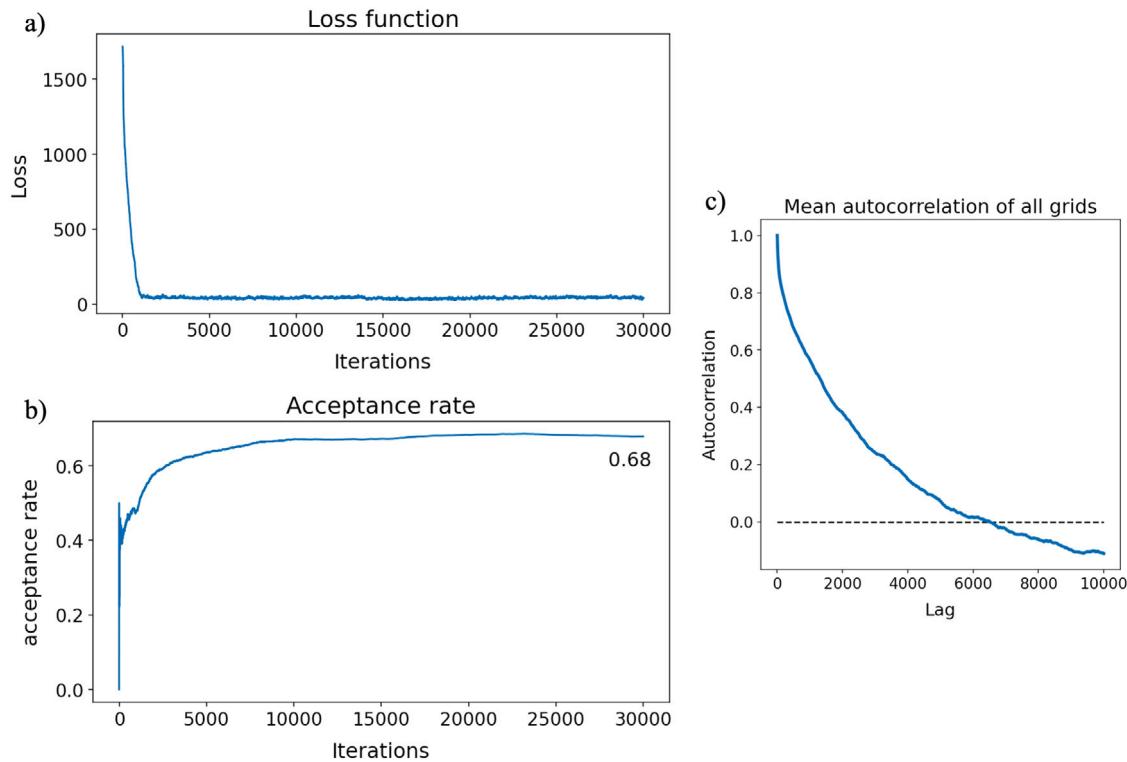


**Fig. 12.** MCMC sampling results: stochastic intrusion simulations. (a) mean and variance for both signed distance and discrete models, (b) Iterative sampling results.

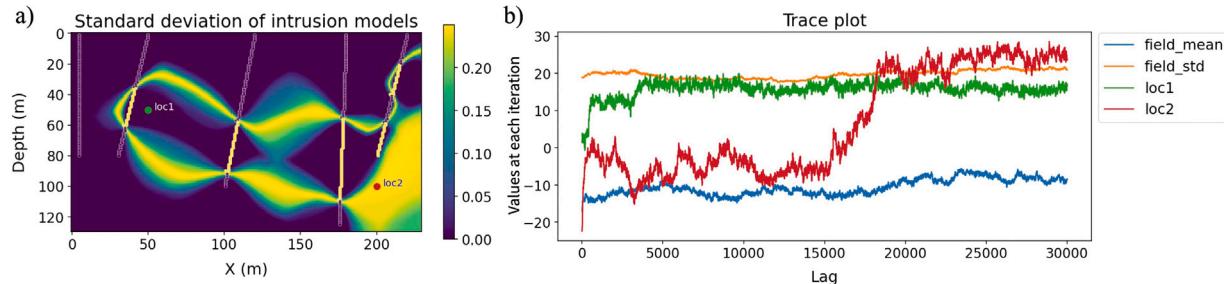
We first define the density function given the sparse AEM survey. After the geophysical inversion for the electromagnetic survey, we have an electrical conductivity model  $\sigma(x_i, y_i, z_i)$  of the subsurface. Different rock types or geological structures have different distributions of electrical conductivity. Interfaces between two structures focus on the contrast of electrical conductivity. Therefore, we construct the density function to maximize the electrical conductivity contrast between different structures. The objective function is to minimize the within-class variance of electrical conductivity  $\sigma$  for different structures in order to maximize the contrast: Eq. (22) is given in Box I where  $\sigma_0$  is the measured electrical conductivity,  $\mu_{\text{palaeovalley}}$  is the average electrical conductivity in palaeovalley,  $\mu_{\text{non-palaeovalley}}$  is the average electrical conductivity in non-palaeovalley. Palaeovalley structures are defined by  $\phi^n > 0$ . This within-class variance density function adopts the idea from the energy minimization in the active contour method (Chan and Vese, 2001) or from the minimizing intra-class intensity variance

in the Otsu's thresholding method (Otsu, 1979) in computer vision segmentation tasks.

We then define the density function for geological knowledge. We construct the distance between the conceptual connectivity of geological structures and the connectivity graph from our interface model  $\phi^n$ . The actual connectivity graphs for each interface model  $\phi^n$  are in Fig. 17. For each sample, we extract the deepest point of the palaeovalley structure (Fig. 17a), then take the area deeper than a threshold of 60 m (Fig. 17c), and finally skeletonize (Zhang and Suen, 1984) the deep area into a connectivity map (Fig. 17c). We use Modified Hausdorff distance to compare two connectivity graphs. Modified Hausdorff distance (MHD) decides the similarity between two objects or two sets of points (Dubuisson and Jain, 1994). It is defined as the average distance between each point in one set to its closest point in the other set. This Modified Hausdorff distance is commonly used to



**Fig. 13.** Case 2 loss function, acceptance rate and mean autocorrelation. (a) loss function. (b) acceptance rate, with the converged acceptance rate at 0.68. (c) the mean of the autocorrelation for each grid after 5000 iterations, the dashed line represents zero autocorrelation.



**Fig. 14.** Case 2 trace plots. (a) Two locations for trace plots in (b). (b) Trace plots for the field mean, standard deviation and two locations.

$$L_4(\phi^n) = \exp\left(-\frac{\sum_{i \in \phi^n > 0} (\sigma_0(x_i, y_i, z_i) - \mu_{palaeovalley})^2 + \sum_{i \in \phi^n < 0} (\sigma_0(x_i, y_i, z_i) - \mu_{non-palaeovalley})^2}{2\omega^2}\right) \quad (22)$$

Box I.

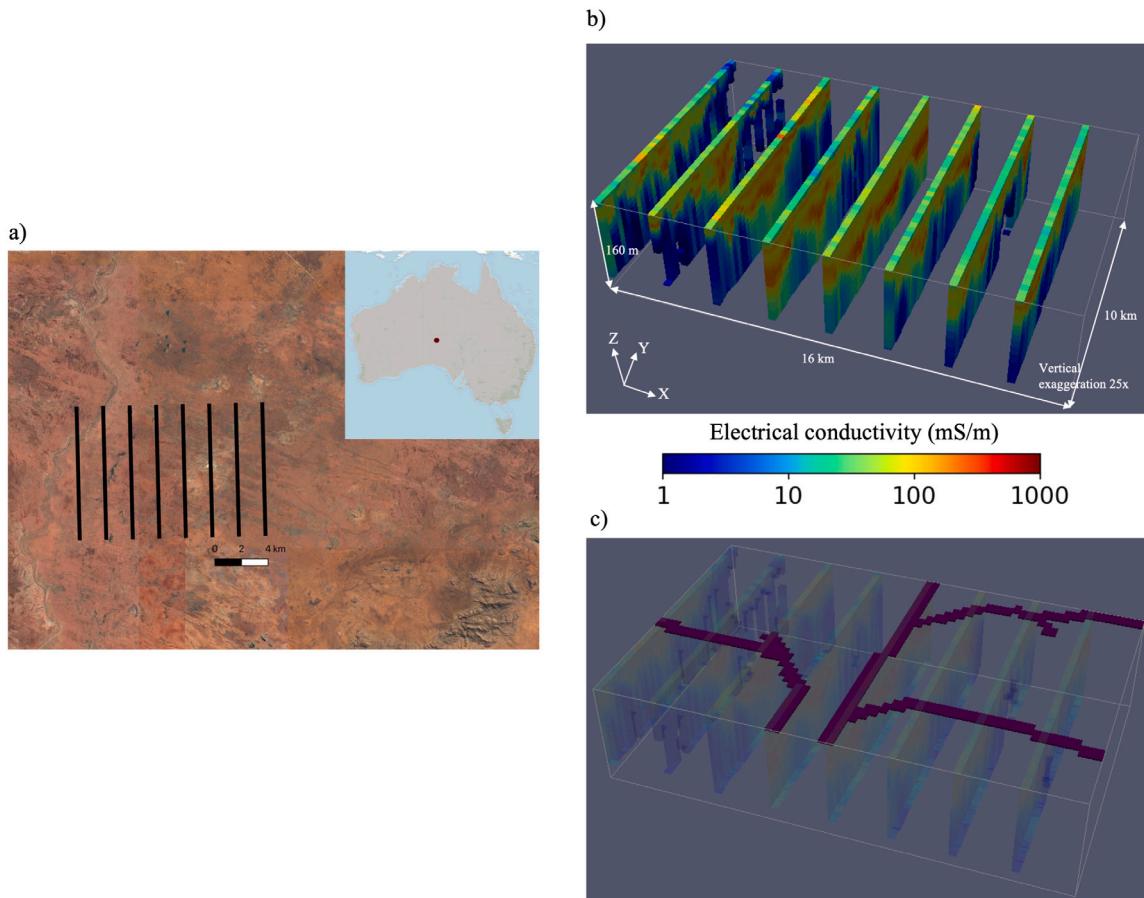
compare shapes in image and object recognition tasks.

$$L_5(\phi^n) = \exp\left(-\frac{MHD^2}{2\omega^2}\right) \quad (23)$$

The hyperparameter  $\omega = 140$  in these two density functions. We discretize this modeling area into  $64 \times 40 \times 80$  grids with  $250 \text{ m} \times 250 \text{ m} \times 2 \text{ m}$  resolution. We start from a simple flat structure (Fig. 16a) and perturb its implicit model with velocity extension in 3D. The simple flat structure consists of two interfaces: the bottom interface separates the in-fill sediment in the palaeovalley from the underlying bedrock, while the top interface separates the palaeovalley sediment from the weathered calcrete layer (Munday et al., 2020b): the weathered crust in semi-arid/arid regions. Note that even we have two interfaces, we still model this structure using one implicit field, where the in-fill

sediment in the palaeovalley has the positive signed distance and the rest has the negative signed distance. The stationary Gaussian velocity field has a uniform distribution for both horizontal and vertical ranges:  $U(2.5 \text{ km}, 5 \text{ km})$  for the horizontal range and  $U(40 \text{ m}, 80 \text{ m})$  for the vertical range. We multiply two density functions in Eqs. (22) and (23) to construct a combined density function and meet criteria from both the AEM survey and the interpreted connectivity.

Fig. 16a have the mean and variance in 3D of the discrete representation  $\phi^n > 0$  over all iterations excluding the first 1500 samples. Fig. 16b presents the iteration of one inline section and one cross-line section. The most uncertain area is at the bottom of the palaeovalley. The palaeovalley boundary separates the high conductive and the context area in low conductivity well. The top interface between the calcrete/surfical sediment and the in-filled sediment has disappeared



**Fig. 15.** Airborne electromagnetic (AEM) data and geological knowledge for the palaeovalley modeling. (a) AEM flight lines in South Australia, (b) electrical conductivity data from the AEM survey, (c) connectivity knowledge/thalweg for the buried valley, provided in Munday et al. (2020b).

after the perturbation because the electrical conductivity contrast is low. Fig. 17 shows that the connectivity has been preserved even with sparse flight lines. Additional geological constraints provide connectivity information between flight lines. Both the AEM loss function and the connectivity loss function converge after 1500 iterations (Fig. 18a, b). The acceptance rate converges to 0.44 (Fig. 18c), and the mean autocorrelation lag of signed distance function values is zero at 500 lags (Fig. 18d). Fig. 19 shows the trace plots for two locations at the boundary of the cross-line section. Their signed distance values  $\phi$  fluctuate similarly to the perturbed boundary up and down.

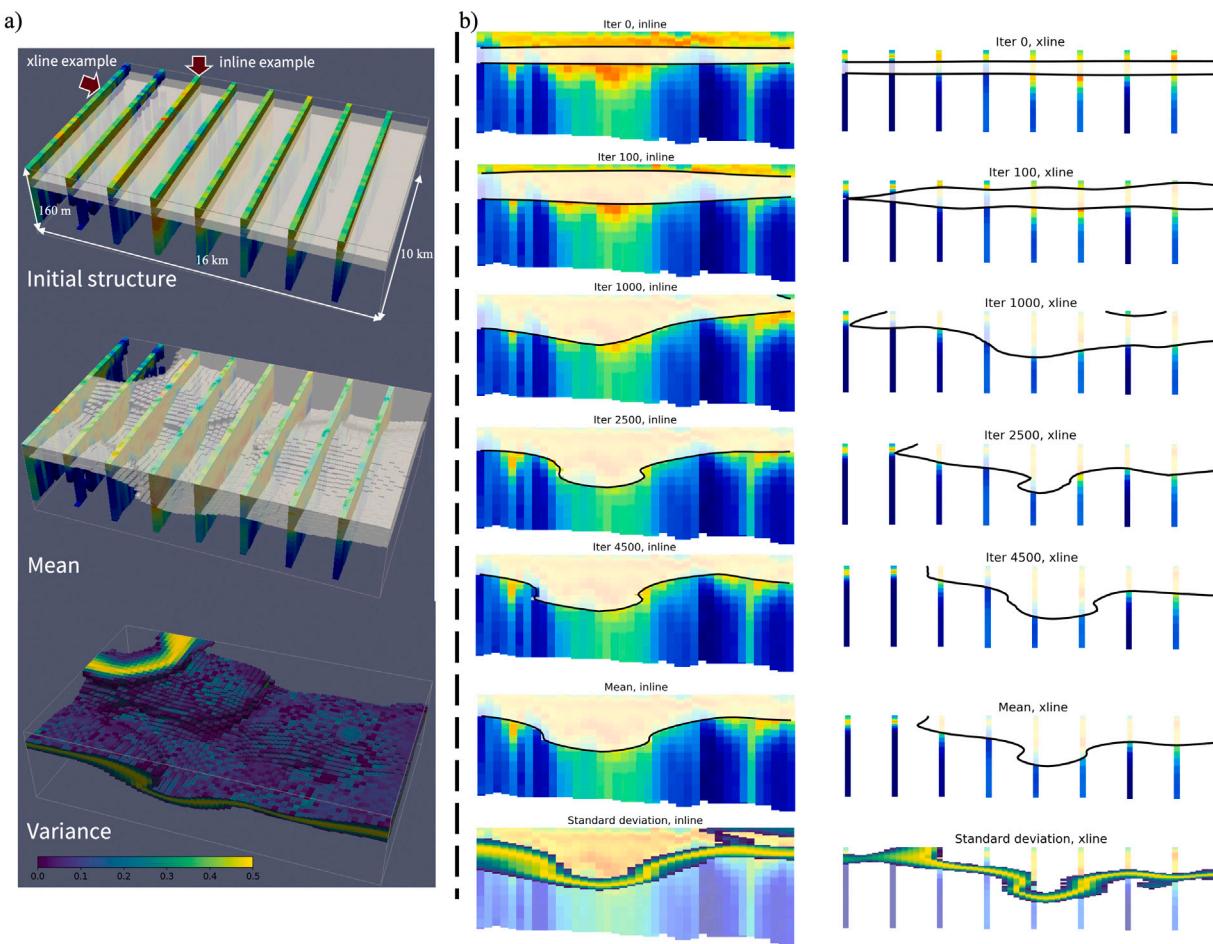
#### 4. Discussion

Our three different case studies demonstrate that the presented data-knowledge-driven trend surface analysis for interface modeling works well in both explicit and implicit modeling and with borehole data, radar data, electromagnetic surveys, and geological knowledge. For more applications in trend interface modeling and uncertainty quantification, we suggest a two-step protocol as we discussed in the methodology section: researchers first define the density function or the target distribution they want to sample from given all available data and knowledge, then use one sampling method to generate stochastic trend interfaces. The integration of data and expert knowledge depends largely on the definition of different density functions. We have presented five different density functions to show how different the density function can be, where many of them are based on the mismatch of data and simulated interfaces. Other inversion approaches such as Approximate Bayesian Computation (Beaumont et al., 2002) have the mismatch defined on summary statistics and then we can transform it into a density function. In this work, we use the inverted geophysical

results instead of conducting the actual geophysical inversion (Test case 1 and 3). However, our method can also be employed for constructing geological interface structures and incorporating the mismatch of geophysical data, such as the mismatch of simulated and observed gravity fields (de la Varga et al., 2019), into the density function. To achieve this, a fast geophysical forward model is required if we use MCMC for quantifying uncertainty. Additional machine learning-based surrogate modeling (Shahriari et al., 2020; Peng et al., 2022; Meles et al., 2022) can be trained to accelerate the forward geophysical model. Therefore, there are many density function options that researchers can explore and combine together. In this paper, we present a Metropolis–Hastings sampling method with stationary Gaussian field perturbations. More advanced MCMC methods, such as Hamilton Monte Carlo (Haario et al., 2006) or adaptive MCMC (Laloy and Vrugt, 2012) can be used in the latter sampling step with different perturbation schemes.

One important advantage of our algorithm is that we can generate a set of stochastic trend interfaces as the prior given limited information. Limited information could be one delineation from geologists or very few drilled boreholes, which is very common in an early exploration stage. Such prior interfaces can directly be used for further inverse problems with new data as the explorations go on. We can also combine the likelihood of the new data with the prior density function to sample the posterior trend interfaces directly. Therefore, our algorithm is also good for Bayesian inverse problems, instead of stochastic optimization problems.

The interfaces we have constructed in this paper are all trend surfaces, which do not exactly match the hard data. Additional inversion or hard conditioning such as geostatistical simulations in Case 1 can be used for final data conditioning. Fouedjio et al. (2021) uses the linear programming method to condition on the hard data. The



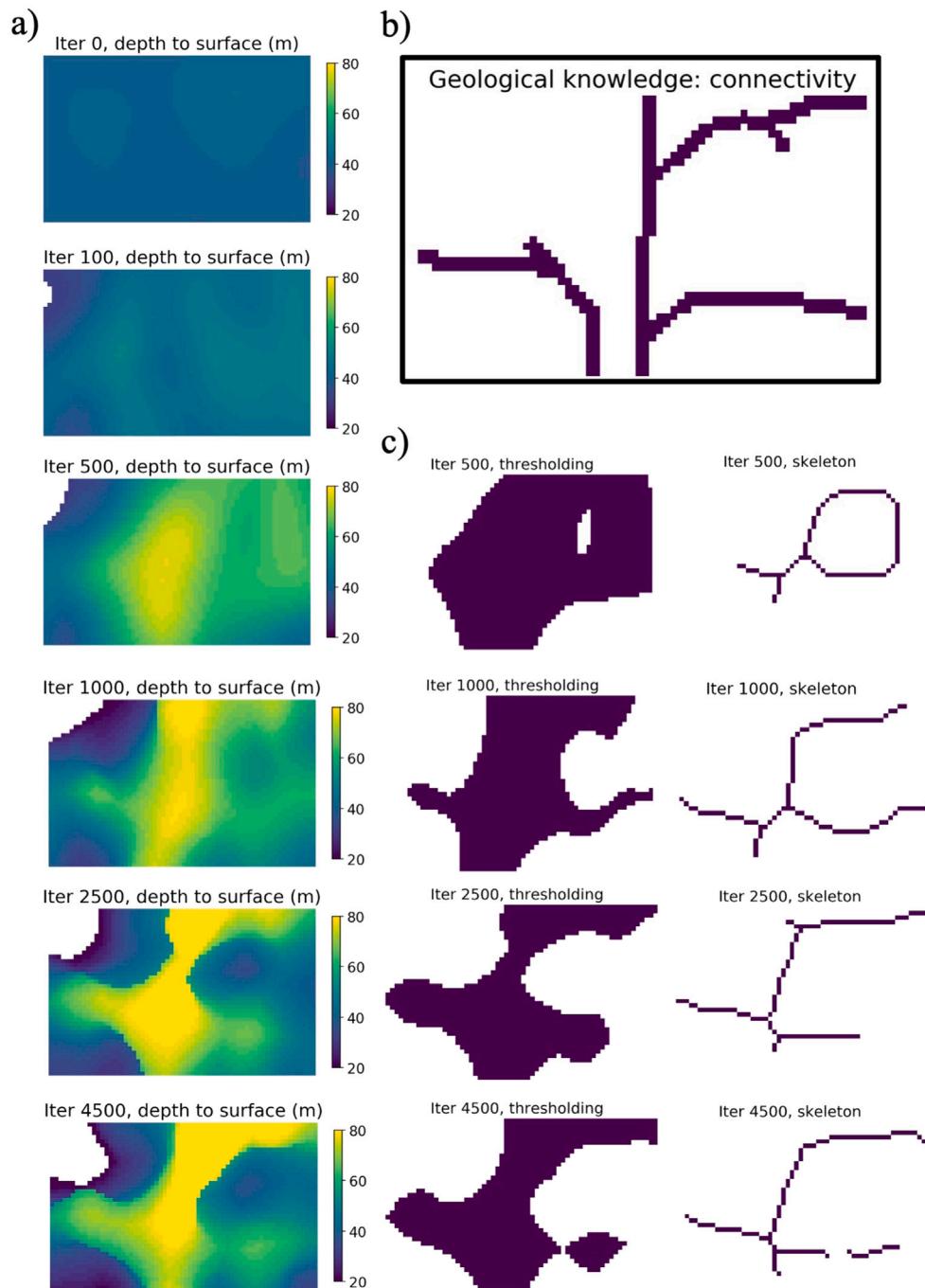
**Fig. 16.** MCMC sampling results: palaeovalley structures. (a) initial structure, mean, and variance for discrete models. (b) iterative sampling results for one inline example and one xline example.

hyperparameter  $\omega$  in density functions acts as the regularizer and controls the target distribution  $L(\phi)$ : we have a better conditioning with a small  $\omega$  but practically the sampling can be very difficult. On the other hand, the range of the velocity field and the step size  $\Delta t$  control the perturbation through the proposal distribution  $p(\phi^*|\phi)$ . In order to have a better convergence, we suggest to use a large range for velocity fields which are larger than the biggest data gap in the system. The step size  $\Delta t$  should satisfy the CFL condition (Eq. (18)) for the implicit method. For the explicit modeling, the step size  $\Delta t$  determines the largest perturbation for each step, thereby influencing the acceptance rate and MCMC convergence. These three parameters  $\omega$ , velocity range and  $\Delta t$  are the hyperparameters or global parameters (Kitanidis, 1995; Malinverno and Briggs, 2004; Emerick, 2016; Xiao et al., 2021; Wang et al., 2022), where we can learn based on the convergence diagnostic using the hierarchical Bayesian method.

Our method can also be applied to an arbitrary number of interfaces. There are two different formulations. The first formulation is based on the potential field method (Lajaunie et al., 1997), where multiple isosurfaces of a scalar field are taken into account. By using the velocity extension, we can perturb the entire potential field and quantify the uncertainty associated with the potential field, and thus the multiple interfaces. This still follows our two main steps: first, defining the density function or target distribution based on all available data and knowledge, and then using our sampling and perturbation method to generate stochastic trend interfaces. This approach allows us to assess the uncertainty of multiple interfaces through multiple iso-surfaces. The second formulation involves working with multiple signed distance functions, each indicating different interfaces. These interfaces can

represent stratigraphic boundaries, fault interfaces, or distinct mineral bodies originating from different sources. In this case, each signed distance function can be perturbed during each iteration using different velocity fields. The density functions can then be evaluated for each iteration. Theoretically, these multiple signed distance functions can still be incorporated into our framework. However, the velocity fields would need to undergo a trial-and-error process to ensure they properly accept the perturbations for multiple fields simultaneously. Previous work has primarily concentrated on modeling two (Li et al., 2017; Zheglova et al., 2018) or arbitrary number (Giraud et al., 2021; Rashidifard et al., 2021) of interfaces using implicit methods based on gradient descent methods or the sensitivity matrix, without incorporating an uncertainty quantification component. By considering these two formulations, we can address the challenges posed by an arbitrary number of interfaces and integrate them into our framework, while accounting for the uncertainties through perturbations and sampling methods.

The discretization of the study domain controls the computational time for our trend surface analysis method. Our suggestion for increasing the efficiency of modeling a very large domain trend is to discretize the field with coarse grids. Hence, the velocity field  $v^*$  has lower dimensions, and calculating the velocity extension field  $F^*$  is less computationally expensive. Additionally, the  $\Delta x$  in Eq. (18) is higher than fine grids as well as the step size  $\Delta t$ . The coarse grids provide a larger step size and thereby increase the perturbation variance. It is important to note that different discretization schemes, particularly with coarser grids, can introduce aliasing issues that may affect the evaluation of density functions. To address this issue, future work can incorporate anti-aliasing techniques, such as Blockworlds proposed by Scalzo et al. (2022).



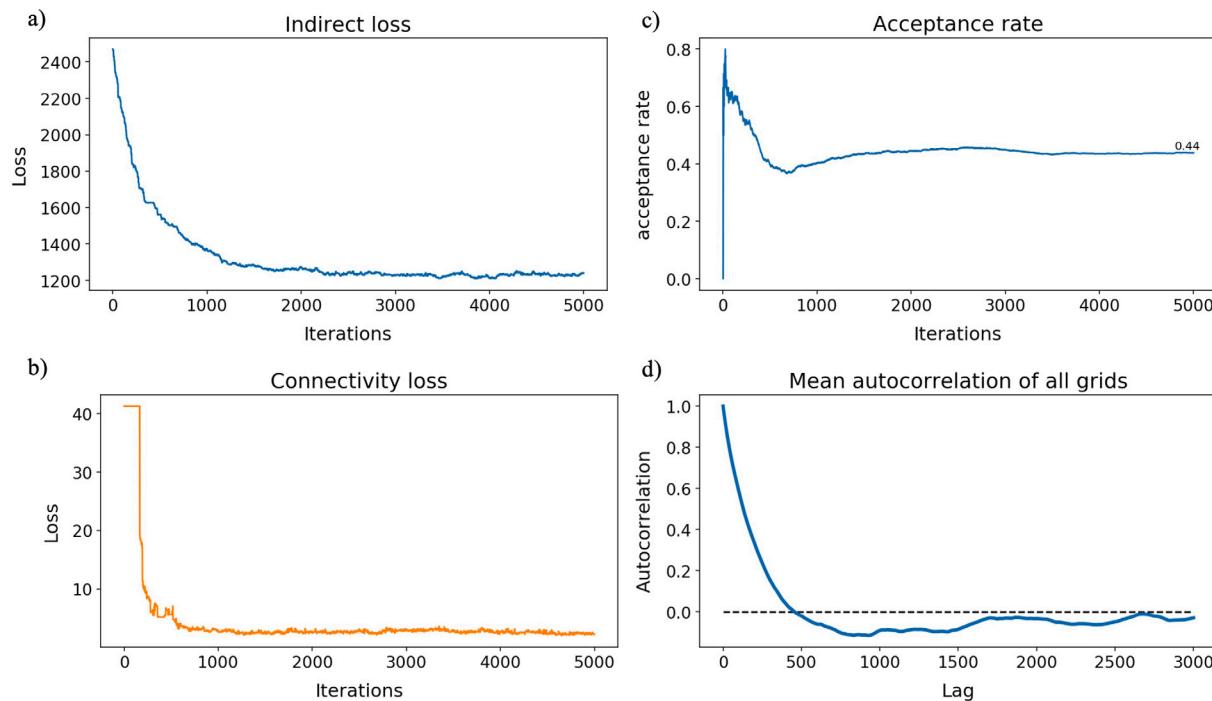
**Fig. 17.** MCMC sampling results: connectivity of the palaeovalley. (a) the palaeovalley depth for MCMC samples. (b) geological knowledge, the connectivity of buried valleys. (c) extracted connectivity for MCMC samples.

Finally, testing the convergence is necessary for MCMC sampling. We have used the Gelman–Rubin convergence diagnostic, the acceptance ratio, and trace plots to understand the MCMC convergence. All test cases have more than 20,000 dimensions (Case 1: 22,500, Case 2: 29,900, Case 3: 204,800). These high-dimensional examples are common in geological modeling. However, testing high-dimensional MCMC convergence is notoriously difficult. We have seen partial parameter convergences where different locations have different convergence behaviors in all three different cases. In the future, we will use more diagnostics, such as multivariate potential scale reduction factor (Brooks and Gelman, 1998) or Kullback Leibler divergence (Dixit and Roy, 2017) with additional dimension reduction methods, to further validate the multivariate convergence. We will also leverage high-performance

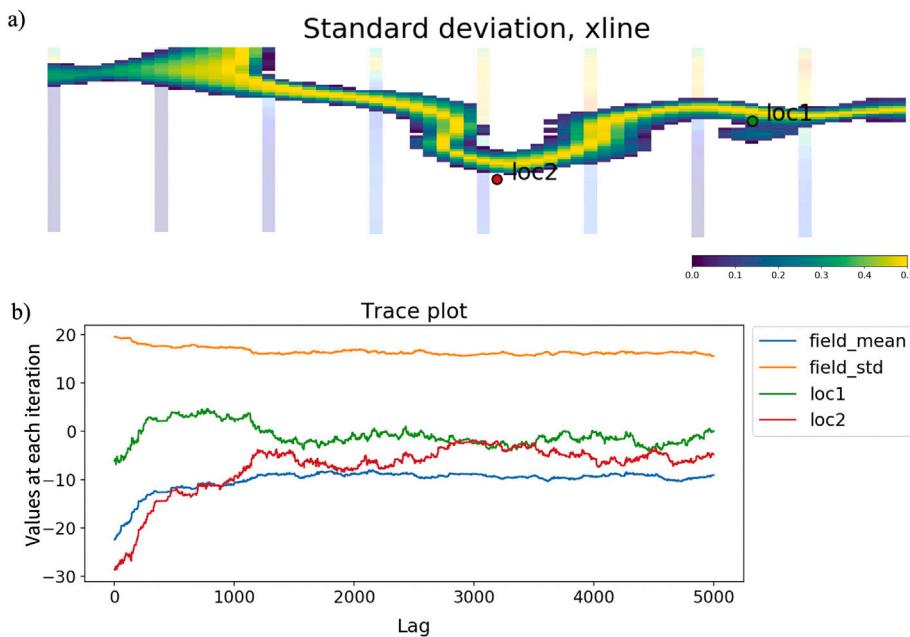
computing to run multiple chains with different initial realizations, and different perturbation schemes, including different stationary Gaussian velocity ranges, or other symmetric distributions which enable easily calculations of the Metropolis–Hastings' acceptance ratio. Having a longer chain or using multiple chains after the convergence is also beneficial for representing the uncertainty of the interface, instead of finding one interface to satisfy datasets and geological knowledge.

## 5. Conclusion

We present a data-knowledge-driven trend surface analysis method for 3D geological interface modeling. Our method integrates multiple datasets with boreholes and geophysical measurements and, most



**Fig. 18.** Case 3 loss function, acceptance rate and mean autocorrelation. (a) indirect data loss. (b) connectivity loss. (c) acceptance rate, with the converged acceptance rate at 0.68. (d) the mean of the autocorrelation for each grid after 1500 iterations, the dashed line represents zero autocorrelation.



**Fig. 19.** Case 3 trace plots. (a) two locations for trace plots in (b). (b) trace plots for the field mean, standard deviation and two locations.

importantly, conceptual geological knowledge. We define different density functions for different information sources, and sample trend interfaces using the Metropolis–Hastings algorithm with stationary Gaussian field perturbations. Our method works for both explicit and implicit interface modeling, where the key advance of the implicit model is to represent complex interfaces and geometries without heavy parameterization. To illustrate our trend surface analysis method, we have tested it on modeling three different interfaces: (1) subglacial topography, (2) magmatic intrusion, (3) palaeovalley structures. Our method has successfully simulated interfaces for all cases with different information sources. In Case 1, we show the sampled subglacial topography served

as stochastic trends in geostatistical simulations and provide more realistic uncertainty quantification than a single deterministic trend. In Case 2, we capture complex intrusion geometries given limited borehole data where the quantified uncertainty can be directly used in future drilling planning such as using efficiency of information (Caers et al., 2022). In Case 3, we built 3D stochastic palaeovalleys for groundwater exploration in South Australia, given sparse geophysical surveys and geological knowledge. We believe our new trend surface analysis tool is beneficial for earth resource explorations such as groundwater and minerals, and for understanding important scientific processes associated with geological interface uncertainty.

## CRediT authorship contribution statement

**Lijing Wang:** Conceptualization, Methodology, Data curation, Formal analysis, Software, Writing – original draft. **Luk Peeters:** Conceptualization, Resources, Writing – review & editing. **Emma J. MacKie:** Data curation for case 1, Writing – review & editing. **Zhen Yin:** Data curation for case 2, Writing – review & editing. **Jef Caers:** Conceptualization, Methodology, Supervision, Writing – review & editing.

## Declaration of competing interest

The authors declare that they have no conflict of interest.

## Data availability

Please access all the data and source code in this GitHub repository: [https://github.com/lijingwang/data\\_knowledge\\_driven\\_trend\\_surface](https://github.com/lijingwang/data_knowledge_driven_trend_surface).

## Acknowledgments

Our computations were made possible by Microsoft Education Azure cloud credits, which are provided through the Stanford Data Science Scholar program to Lijing Wang. Luk Peeters is supported through CSIRO's Deep Earth Imaging Future Science Platform. The authors declare that they have no conflict of interest.

## References

- Abgrall, R., Beaugendre, H., Dobrzynski, C., 2014. An immersed boundary method using unstructured anisotropic mesh adaptation combined with level-sets and penalization techniques. *J. Comput. Phys.* 257, 83–101.
- Adalsteinsson, D., Sethian, J.A., 1999. The fast construction of extension velocities in level set methods. *J. Comput. Phys.* 148 (1), 2–22.
- Agterberg, F.P., 1984. Trend surface analysis. In: *Spatial Statistics and Models*. Springer, pp. 147–171.
- Beaumont, M.A., Zhang, W., Balding, D.J., 2002. Approximate Bayesian computation in population genetics. *Genetics* 162 (4), 2025–2035.
- Bedard, M., 2008. Optimal acceptance rates for Metropolis algorithms: Moving beyond 0.234. *Stochastic Process. Appl.* 118 (12), 2198–2222.
- Bertsimas, D., Tsitsiklis, J., 1993. Simulated annealing. *Statist. Sci.* 8 (1), 10–15.
- Brooks, S.P., Gelman, A., 1998. General methods for monitoring convergence of iterative simulations. *J. Comput. Graph. Statist.* 7 (4), 434–455.
- Broomhead, D.S., Lowe, D., 1988. Radial Basis Functions, Multi-Variable Functional Interpolation and Adaptive Networks. Technical Report, Royal Signals and Radar Establishment, Malvern (United Kingdom).
- Caers, J., Scheidt, C., Yin, Z., Wang, L., Mukerji, T., House, K., 2022. Efficacy of information in mineral exploration drilling. *Nat. Resour. Res.* 1–17.
- Calcagno, P., Chiles, J.-P., Courrioux, G., Guillen, A., 2008. Geological modelling from field data and geological knowledge: Part I. Modelling method coupling 3D potential-field interpolation and geological rules. *Phys. Earth Planet. Inter.* 171 (1–4), 147–157.
- Chan, T.F., Vese, L.A., 2001. Active contours without edges. *IEEE Trans. Image Process.* 10 (2), 266–277.
- Chen, S., Radke, R.J., 2009. Markov chain Monte Carlo shape sampling using level sets. In: 2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops. IEEE, pp. 296–303.
- Chib, S., Greenberg, E., 1995. Understanding the metropolis-hastings algorithm. *Amer. Statist.* 49 (4), 327–335.
- Chiles, J.-P., Delfiner, P., 2009. *Geostatistics: Modeling Spatial Uncertainty*, Vol. 497. John Wiley & Sons.
- Courant, R., Friedrichs, K., Lewy, H., 1967. On the partial difference equations of mathematical physics. *IBM J. Res. Dev.* 11 (2), 215–234. <http://dx.doi.org/10.1147/rd.112.0215>.
- CReSIS, 2022. CReSIS radar depth sounder. Digit. Media URL: <http://data.cresis.ku.edu/>.
- Deutsch, J., 1998. *GSLIB: Geostatistical Software Library and Users Guide*. Oxford University Press.
- Deutsch, C., Tran, T., 2002. FLUVSIM: a program for object-based stochastic modeling of fluvial depositional systems. *Comput. Geosci.* 28 (4), 525–535.
- Dixit, A., Roy, V., 2017. MCMC diagnostics for higher dimensions using Kullback Leibler divergence. *J. Stat. Comput. Simul.* 87 (13), 2622–2638.
- Dubuisson, M.-P., Jain, A.K., 1994. A modified Hausdorff distance for object matching. In: Proceedings of 12th International Conference on Pattern Recognition, Vol. 1. IEEE, pp. 566–568.
- Duijndam, A., 1988a. Bayesian estimation in seismic inversion. Part I: Principles 1. *Geophys. Prospect.* 36 (8), 878–898.
- Duijndam, A., 1988b. Bayesian estimation in seismic inversion. part ii: Uncertainty analysis1. *Geophys. Prospect.* 36 (8), 899–918.
- Emerick, A.A., 2016. Towards a hierarchical parametrization to address prior uncertainty in ensemble-based data assimilation. *Comput. Geosci.* 20 (1), 35–47.
- Emerick, A.A., Reynolds, A.C., 2013. Ensemble smoother with multiple data assimilation. *Comput. Geosci.* 55, 3–15.
- Fan, A.C., Fisher, J.W., Wells, W.M., Levitt, J.J., Willsky, A.S., 2007. MCMC curve sampling for image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 477–485.
- Fouedjio, F., Scheidt, C., Yang, L., Achitzger-Zupančič, P., Caers, J., 2021. A geostatistical implicit modeling framework for uncertainty quantification of 3D geo-domain boundaries: Application to lithological domains from a porphyry copper deposit. *Comput. Geosci.* 157, 104931.
- Gelman, A., Gilks, W.R., Roberts, G.O., 1997. Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Appl. Probab.* 7 (1), 110–120.
- Gelman, A., Rubin, D.B., 1992. Inference from iterative simulation using multiple sequences. *Statist. Sci.* 457–472.
- Giraud, J., Lindsay, M., Jessell, M., 2021. Generalization of level-set inversion to an arbitrary number of geologic units in a regularized least-squares framework. *Geophysics* 86 (4), R623–R637.
- Grose, L., Ailleres, L., Laurent, G., Caumon, G., Jessell, M., Armit, R., 2021a. Modelling of faults in LoopStructural 1.0. *Geosci. Model Dev.* 14 (10), 6197–6213.
- Grose, L., Ailleres, L., Laurent, G., Jessell, M., 2021b. LoopStructural 1.0: time-aware geological modelling. *Geosci. Model Dev.* 14 (6), 3915–3937.
- Haario, H., Laine, M., Mira, A., Saksman, E., 2006. DRAM: efficient adaptive MCMC. *Statist. Comput.* 16 (4), 339–354.
- Henrion, V., Caumon, G., Cherpeau, N., 2010. ODSIM: an object-distance simulation method for conditioning complex natural structures. *Math. Geosci.* 42, 911–924.
- Iglesias, M.A., Lu, Y., Stuart, A.M., 2016. A Bayesian level set method for geometric inverse problems. *Interfaces Free Bound.* 18 (2), 181–217.
- Journel, A.G., Ying, Z., 2001. The theoretical links between sequential Gaussian simulation, Gaussian truncated simulation, and probability field simulation. *Math. Geol.* 33 (1), 31–40.
- Kitanidis, P.K., 1995. Quasi-linear geostatistical theory for inversing. *Water Resour. Res.* 31 (10), 2411–2419.
- Kyriakidis, P.C., Journel, A.G., 1999. Geostatistical space-time models: a review. *Math. Geol.* 31 (6), 651–684.
- Lajaunie, C., Courrioux, G., Manuel, L., 1997. Foliation fields and 3D cartography in geology: principles of a method based on potential interpolation. *Math. Geol.* 29, 571–584.
- Laloy, E., Vrugt, J.A., 2012. High-dimensional posterior exploration of hydrologic models using multiple-try DREAM (ZS) and high-performance computing. *Water Resour. Res.* 48 (1).
- Law, R., Christoffersen, P., MacKie, E., Cook, S., Haseloff, M., Gagliardini, O., 2022. Complex motion of Greenland Ice Sheet outlet glaciers with basal temperate ice.
- Li, W., Lu, W., Qian, J., Li, Y., 2017. A multiple level-set method for 3D inversion of magnetic data. *Geophysics* 82 (5), J61–J81.
- Lindsay, M.D., Ailleres, L., Jessell, M.W., de Kemp, E.A., Betts, P.G., 2012. Locating and quantifying geological uncertainty in three-dimensional models: Analysis of the Gippsland Basin, southeastern Australia. *Tectonophysics* 546, 10–27.
- Lochbühler, T., Pirot, G., Straubhaar, J., Linde, N., 2014. Conditioning of multiple-point statistic facies simulations to tomographic images. *Math. Geosci.* 46, 625–645.
- MacKie, E.J., Schroeder, D.M., Zuo, C., Yin, Z., Caers, J., 2021. Stochastic modeling of subglacial topography exposes uncertainty in water routing at Jakobshavn Glacier. *J. Glaciol.* 67 (261), 75–83.
- Malinverno, A., Briggs, V.A., 2004. Expanded uncertainty quantification in inverse problems: Hierarchical Bayes and empirical Bayes. *Geophysics* 69 (4), 1005–1016.
- Manchuk, J., Deutsch, C.V., 2011. A Short Note on Trend Modeling Using Moving Windows. Centre for Computational Geostatistics, University of Alberta, Edmonton, Canada, CCG Paper (403).
- Mariethoz, G., Caers, J., 2014. *Multiple-Point Geostatistics: Stochastic Modeling with Training Images*. John Wiley & Sons.
- Meles, G.A., Linde, N., Marelli, S., 2022. Bayesian tomography with prior-knowledge-based parametrization and surrogate modelling. *Geophys. J. Int.* 231 (1), 673–691.
- Moradi, M., Basiri, S., Kananian, A., Kabiri, K., 2015. Fuzzy logic modeling for hydrothermal gold mineralization mapping using geochemical, geological, ASTER imagers and other geo-data, a case study in Central Alborz, Iran. *Earth Sci. Inform.* 8 (1), 197–205.
- Morgan, N.R., Waltz, J.I., 2017. 3D level set methods for evolving fronts on tetrahedral meshes with adaptive mesh refinement. *J. Comput. Phys.* 336, 492–512.
- Munday, T., Gilfedder, M., Costar, A., Blaikie, T., Cahill, K., Cui, T., Davis, A., Deng, Z., Flinchum, B., Gao, L., et al., 2020a. Facilitating Long-term Outback Water Solutions (G-FLOWS Stage 3): Final Summary Report. Goyder Institute for Water Research, Adelaide, Australia.
- Munday, T., Taylor, A., Raiber, M., Soerensen, C., Peeters, L., Krapf, C., Cui, T., Cahill, K., Flinchum, B., Smolanko, N., et al., 2020b. Integrated Regional Hydrogeophysical Conceptualisation of the Musgrave Province, South Australia. Goyder Institute for Water Research, Adelaide, SA, Australia.

- Norcliffe, G., 1969. On the use and limitations of trend surface models. *Can. Geogr./Géogr. Can.* 13 (4), 338–348.
- Olieroor, H.K., Scalzo, R., Kohn, D., Chandra, R., Farahbakhsh, E., Clark, C., Reddy, S.M., Müller, R.D., 2021. Bayesian geological and geophysical data fusion for the construction and uncertainty quantification of 3D geological models. *Geosci. Front.* 12 (1), 479–493.
- Osher, S., Fedkiw, R.P., 2003. *Level Set Methods and Dynamic Implicit Surfaces*, Vol. 153. Springer.
- Otsu, N., 1979. A threshold selection method from gray-level histograms. *IEEE Trans. Syst. Man Cybern.* 9 (1), 62–66.
- Pakyuz-Charrier, E., Jessell, M., Giraud, J., Lindsay, M., Ogarko, V., 2019. Topological analysis in Monte Carlo simulation for uncertainty propagation. *Solid Earth* 10 (5), 1663–1684.
- Pakyuz-Charrier, E., Lindsay, M., Ogarko, V., Giraud, J., Jessell, M., 2018. Monte Carlo simulation for uncertainty estimation on structural data in implicit 3-D geological modeling, a guide for disturbance distribution selection and parameterization. *Solid Earth* 9 (2), 385–402.
- Paola, C., 2000. Quantitative models of sedimentary basin filling. *Sedimentology* 47, 121–178.
- Peeters, L., Fasbender, D., Batelaan, O., Dassargues, A., 2010. Bayesian data fusion for water table interpolation: incorporating a hydrogeological conceptual model in kriging. *Water Resour. Res.* 46 (8).
- Peng, Z., Yang, B., Xu, Y., Wang, F., Liu, L., Zhang, Y., 2022. Rapid surrogate modeling of electromagnetic data in frequency domain using neural operator. *IEEE Trans. Geosci. Remote Sens.* 60, 1–12.
- Pyrcz, M.J., Deutsch, C.V., 2014. *Geostatistical Reservoir Modeling*. Oxford University Press.
- Rashidifard, M., Giraud, J., Lindsay, M., Jessell, M., Ogarko, V., 2021. Constraining 3D geometric gravity inversion with a 2D reflection seismic profile using a generalized level set approach: application to the eastern Yilgarn Craton. *Solid Earth* 12 (10), 2387–2406.
- Rongier, G., Collon-Drouaillet, P., Filippioni, M., 2014. Simulation of 3D karst conduits with an object-distance based method integrating geological knowledge. *Geomorphology* 217, 152–164.
- Sambridge, M., 1999a. Geophysical inversion with a neighbourhood algorithm—I. Searching a parameter space. *Geophys. J. Int.* 138 (2), 479–494.
- Sambridge, M., 1999b. Geophysical inversion with a neighbourhood algorithm—II. Appraising the ensemble. *Geophys. J. Int.* 138 (3), 727–746.
- Scalzo, R., Lindsay, M., Jessell, M., Pirot, G., Giraud, J., Cripps, E., Cripps, S., 2022. Blockworlds 0.1. 0: a demonstration of anti-aliased geophysics for probabilistic inversions of implicit and kinematic geological models. *Geosci. Model Dev.* 15 (9), 3641–3662.
- Shahriari, M., Pardo, D., Moser, B., Sobieczky, F., 2020. A deep neural network as surrogate model for forward simulation of borehole resistivity measurements. *Procedia Manuf.* 42, 235–238.
- Tarantola, A., 1987. Inversion of travel times and seismic waveforms. In: *Seismic Tomography*. Springer, pp. 135–157.
- de la Varga, M., Schaaf, A., Wellmann, F., 2019. GemPy 1.0: open-source stochastic geological modeling and inversion. *Geosci. Model Dev.* 12 (1), 1–32.
- Visser, G., Markov, J., 2019. Cover thickness uncertainty mapping using Bayesian estimate fusion: leveraging domain knowledge. *Geophys. J. Int.* 219 (3), 1474–1490.
- Wang, L., Kitanidis, P.K., Caers, J., 2022. Hierarchical Bayesian inversion of global variables and large-scale spatial fields. *Water Resour. Res.* 58 (5), e2021WR031610.
- Wellmann, J.F., De La Varga, M., Murdie, R.E., Gessner, K., Jessell, M., 2018. Uncertainty estimation for a geological model of the Sandstone greenstone belt, Western Australia—insights from integrated geological and geophysical inversion in a Bayesian inference framework. *Geol. Soc. Lond. Spec. Publ.* 453 (1), 41–56.
- Wellmann, J.F., Horowitz, F.G., Schill, E., Regenauer-Lieb, K., 2010. Towards incorporating uncertainty of structural data in 3D geological inversion. *Tectonophysics* 490 (3–4), 141–151.
- Wernecke, A., Edwards, T.L., Holden, P.B., Edwards, N.R., Cornford, S.L., 2022. Quantifying the impact of bedrock topography uncertainty in Pine Island Glacier projections for this century. *Geophys. Res. Lett.* 49 (6), e2021GL096589.
- Xiao, S., Xu, T., Reuschen, S., Nowak, W., Hendricks Fransen, H.-J., 2021. Bayesian inversion of multi-Gaussian log-conductivity fields with uncertain hyperparameters: An extension of preconditioned Crank-Nicolson Markov Chain Monte Carlo with parallel tempering. *Water Resour. Res.* 57 (9), e2021WR030313.
- Yang, L., Hyde, D., Grujic, O., Scheidt, C., Caers, J., 2019. Assessing and visualizing uncertainty of 3D geological surfaces using level sets with stochastic motion. *Comput. Geosci.* 122, 54–67.
- Yang, X., James, A.J., Lowengrub, J., Zheng, X., Cristini, V., 2006. An adaptive coupled level-set/volume-of-fluid interface capturing method for unstructured triangular grids. *J. Comput. Phys.* 217 (2), 364–394.
- Zhang, T.Y., Suen, C.Y., 1984. A fast parallel algorithm for thinning digital patterns. *Commun. ACM* 27 (3), 236–239.
- Zheglova, P., Lelièvre, P.G., Farquharson, C.G., 2018. Multiple level-set joint inversion of traveltime and gravity data with application to ore delineation: A synthetic study. *Geophysics* 83 (1), R13–R30.