Hierarchical Bayesian inversion of global variables and large-scale spatial fields

Lijing Wang [1], Peter K. Kitanidis [2], Jef Caers [1]

[1]Department of Geological Sciences, Stanford University
[2]Department of Civil and Environmental Engineering, Stanford University

Key Points:

- Hierarchical Bayesian formulation to invert both uncertain global and spatial variables.
- Machine learning-based inversion method to invert global variables in both linear and non-linear forward models.
- Local principal component analysis (local PCA) approach towards a more efficient local inversion on large-scale spatial fields.

Corresponding author: Lijing Wang, `lijing52@stanford.edu`

Abstract

Bayesian inversion is commonly applied to quantify uncertainty of hydrological variables. However, the focus in Bayesian inversion is more on spatial hydrological properties instead of hyperparameters or global/non-gridded variables. In this paper, we present a hierarchical Bayesian framework to quantify uncertainty of both global and spatial variables. We estimate first the posterior of global variables and then hierarchically estimate the posterior of the spatial field. We propose a machine learning-based inversion method to estimate the joint distribution of data and global variables directly without introducing a statistical likelihood. We also propose a new local dimension reduction method: local principal component analysis (local PCA) to update large-scale spatial fields with local data more efficiently. We illustrate the hierarchical Bayesian formulation with two case studies: one with a linear forward model (volume averaging inversion) and a second with a non-linear forward model (pumping tests). Results show that quantifying global variables uncertainty is critical for assessing uncertainty on predictions. We show how the local PCA approach accelerates the inversion process. Furthermore, we provide an open-source Python package on the hierarchical Bayesian framework including two case studies.

1 Introduction

Estimating hydrological properties and quantifying their uncertainty from observed measurements is vital for groundwater exploration and exploitation decision-making processes. Bayesian inversion is now commonly applied to obtain posterior distributions (Tarantola, 2005). Two popular Bayesian inversion approaches are 1) the Markov chain Monte Carlo (MCMC) method, 2) the Monte Carlo ensemble method.

Many MCMC methods have been well studied to solve Bayesian inversion problems in hydrology and other subsurface reservoirs: Metropolis-Hasting sampling (Oliver et al., 1997; Kuczera & Parent, 1998), DREAM: differential evolution adaptive Metropolis (Vrugt et al., 2008; Vrugt & Ter Braak, 2011; Laloy & Vrugt, 2012), Hamiltonian Monte Carlo (Saley et al., 2016; Fichtner & Simutė, 2018), etc. MCMC performs well if the forward modeling is relatively fast, because MCMC requires many forward model simulations. Therefore this class of methods is relatively computational expensive.

Monte Carlo-based ensemble methods rely on independently sampled ensembles to estimate statistical properties such as covariances: ensemble Kalman filter (Evensen, 2003), ensemble smoother with multiple data assimilation (Emerick & Reynolds, 2013) or the learnable relationship between data and model (or prediction) variables directly: direct forecasting (Satija & Caers, 2015; Hermans et al., 2016; Satija et al., 2017; Athens & Caers, 2019; Yin et al., 2020; J. Park & Caers, 2020), data-space inversion (Sun & Durlofsky, 2017). Other methods directly evaluating the likelihood for each ensemble include rejection sampling (Tarantola, 2005) and Generalized Likelihood Uncertainty Estimation (GLUE) (Beven & Freer, 2001). Monte Carlo methods can be easily parallelized. Monte Carlo methods may however require iterations for non-linear problems (Emerick & Reynolds, 2013)

However, building stochastic spatial fields such as hydraulic conductivity requires considering other uncertainties such as in terms covariances, mean values, boundary conditions, or even multiple types of prior models (H. Park et al., 2013; Scheidt et al., 2015, 2018; Grana, 2020), etc. Those non-spatial variables may also be subject to high uncertainty and are often neglected in Bayesian inversion process. In this paper, we denote those non-spatial variables as global variables, which are also known as hyperparameters or structural parameters (Kitanidis, 1995; Emerick, 2016; Xiao et al., 2021). Even with the simplest linear forward model, the relationship between global variables and data variables are non-linear. Chosen for a single deterministic set may lead to biases (Zhao

& Luo, 2021). Moreover, global and spatial variables are not independent and have a complex relationship: global variables determine the statistical properties for spatial fields, spatial data measurement might inform global variables. Therefore, one objective of this paper is to adapt the hierarchical Bayesian framework to quantify uncertainty of global and spatial variables in a hierarchical fashion, for both linear and non-linear forward models.

Another challenge in hierarchical Bayesian inversion is to invert high-dimensional large spatial fields. We often perform dimension reduction before the actual inversion to make those problems more tractable (Kitanidis & Lee, 2014; Lee & Kitanidis, 2014; Sun & Durlofsky, 2017; Scheidt et al., 2018; Yin et al., 2020; Fouedjio et al., 2021). However, observed data may locate within a small local field and does not inform spatial variables far away. Directly using the lower dimension representation obtained from the full field might cause an ensemble collapse or worse data matching (Chen & Oliver, 2010; Oliver & Chen, 2011). Therefore, the other objective of this paper is to perform efficient local inversions in large-scale spatial fields using local dimension reduction techniques.

We organize this paper as follows. In the methodology section, we first review the hierarchical Bayesian framework for quantifying the uncertainty of global and spatial model variables jointly. We then propose a machine learning-based inversion method within hierarchical Bayes to estimate the posterior distribution of global variables. We limit ourselves within a linear Gaussian case and only focus on the geostatistical part of global variables. Then we move to non-linear forward models and incorporate all possible global variables. We then address the local inversion problem for large-scale spatial fields using a new dimension reduction technique: local principal component analysis (local PCA). The applications section presents hierarchical Bayes inversion on two synthetic studies with linear and non-linear forward models and measurement errors. We compare our machine learning-based inversion method and a statistical sampling method with analytical likelihood functions. We also compare spatial inversions with global PCA and local PCA.

2 Methodology

2.1 Review of hierarchical Bayes: invert both global and spatial variables

The characterization of spatial fields depends on global variables that have no specific grid location, such as variance, range and anisotropy. In the same category, one may include other global variables that are not related to spatial fields, such as hydraulic head or flux at boundaries or specific yield. Global variables, also known as hyperparameters or nuisance parameters, are subject to a high degree of uncertainty and may impact flow prediction.

Therefore, the objective of this paper is to quantify uncertainty of global and spatial variables $\boldsymbol{\theta}$ and $\boldsymbol{m}$ jointly given the observed data $\boldsymbol{d}_{obs}$:

$$p(\boldsymbol{m}, \boldsymbol{\theta} | \boldsymbol{d}_{obs}) \tag{1}$$

This joint posterior distribution can be decomposed as follows:

$$p(\boldsymbol{m}, \boldsymbol{\theta} | \boldsymbol{d}_{obs}) = p(\boldsymbol{m} | \boldsymbol{\theta}, \boldsymbol{d}_{obs}) p(\boldsymbol{\theta} | \boldsymbol{d}_{obs}) \tag{2}$$

One intuitive solution to obtain the joint posterior is to estimate two conditional probabilities hierarchically: first estimating the posterior of global variables $p(\boldsymbol{\theta}|\boldsymbol{d}_{obs})$; then given this global variable posterior, estimating the spatial posterior $p(\boldsymbol{m}|\boldsymbol{\theta}, \boldsymbol{d}_{obs})$. This formulation which considers both uncertainty in global and spatial variables and

solves inversion problem hierarchically is known as hierarchical Bayes in statistics (Gelman et al., 1995; Kitanidis, 1995; Banerjee et al., 2003; Malinverno & Briggs, 2004). In hierarchical Bayes, the posterior distribution of spatial variables $p(\boldsymbol{m}|\boldsymbol{d}_{obs})$ is the marginal posterior distribution by integrating the above joint posterior distribution:

$$p(\boldsymbol{m}|\boldsymbol{d}_{obs}) = \int p(\boldsymbol{m}, \boldsymbol{\theta}|\boldsymbol{d}_{obs})\mathrm{d}\boldsymbol{\theta} = \int p(\boldsymbol{m}|\boldsymbol{\theta}, \boldsymbol{d}_{obs})p(\boldsymbol{\theta}|\boldsymbol{d}_{obs})\mathrm{d}\boldsymbol{\theta} \tag{3}$$

However, integration over all global variables is practically infeasible. We choose to represent this integral using Monte Carlo sampling method presented in Algorithm 1.

Algorithm 1: Hierarchical Bayesian inversion of both global and spatial variables

1) State prior uncertainty of $\boldsymbol{\theta}$: $p(\boldsymbol{\theta})$
2) Estimate posterior distribution of global variables: $p(\boldsymbol{\theta}|\boldsymbol{d}_{obs}) \propto p(\boldsymbol{d}_{obs}|\boldsymbol{\theta})p(\boldsymbol{\theta})$
3) Estimate posterior distribution of spatial variables hierarchically:
for each posterior realization $\boldsymbol{\theta}^*$ sampled from $p(\boldsymbol{\theta}^*) = p(\boldsymbol{\theta}|\boldsymbol{d}_{obs})$ **do**
  3.1) Estimate posterior distribution of spatial variables $\boldsymbol{m}$ given $\boldsymbol{\theta}^*$: $p(\boldsymbol{m}|\boldsymbol{\theta}^*, \boldsymbol{d}_{obs})$
  3.2) Sample $\boldsymbol{m}^*$ from $p(\boldsymbol{m}|\boldsymbol{\theta}^*, \boldsymbol{d}_{obs})$, which gives us one joint posterior realization $(\boldsymbol{m}^*, \boldsymbol{\theta}^*)$
end

An alternative approach to this general formulation is proposed in Kitanidis (1995) by means of a quasi-linear geostatistical theory, which estimates global variables $\boldsymbol{\theta}$ using restricted maximum likelihood.

Note that the relationship between global variables $\boldsymbol{\theta}$ and observed data $\boldsymbol{d}_{obs}$ remains analytically unknown and non-linear in both linear and non-linear forward models. Furthermore, when updating spatial variables, we often have large-scale spatial models and a larger number of unknowns while the observed data may only be collected in a specific local area. This is a local inversion challenge for large-spatial fields.

In the following sections 2.2 and 2.3, we focus on new methods for inverting global variables $\boldsymbol{\theta}$ and subsequently inverting spatial variables $\boldsymbol{m}$ in hierarchical Bayesian formulation. Specifically, we propose a machine learning-based inversion method to invert global variables when the forward model is non-linear (2.2.2). Then, we tackle the local inversion problem for large-spatial fields using a novel local principal component analysis method (2.3).

2.2 Posterior distributions of global variables

In this section, we cover two methods to estimate the posterior distribution of global variables $p(\boldsymbol{\theta}|\boldsymbol{d}_{obs})$: 1) statistical sampling method with analytical likelihood functions and 2) machine learning-based inversion method.

2.2.1 Linear Gaussian inversion: statistical sampling with analytical likelihood functions

We start with the case where the forward model is linear and the prior distributions on spatial variables and error terms are assumed Gaussian:

$$\boldsymbol{m} \sim N(\boldsymbol{\mu}_{\boldsymbol{m}}, \boldsymbol{C}_{\boldsymbol{m}}), \boldsymbol{\varepsilon} \sim N(0, \boldsymbol{\Sigma}),\ \boldsymbol{d} = \mathrm{G}\boldsymbol{m} + \boldsymbol{\varepsilon} \tag{4}$$

Under these conditions, $\boldsymbol{d}$ has a Gaussian distribution given any $\boldsymbol{\theta}$:

$$\boldsymbol{d}|\boldsymbol{\theta} \sim N(\boldsymbol{\mu}_{\boldsymbol{d}|\boldsymbol{\theta}}, \boldsymbol{C}_{\boldsymbol{d}|\boldsymbol{\theta}}) \tag{5}$$

where

$$\boldsymbol{\mu}_{\boldsymbol{d}|\boldsymbol{\theta}} = E(\boldsymbol{d}|\boldsymbol{\theta}) = \mathrm{G}E(\boldsymbol{m}|\boldsymbol{\theta}) + E(\boldsymbol{\varepsilon}) = \mathrm{G}\boldsymbol{\mu}_{\boldsymbol{m}} \tag{6}$$

$$\boldsymbol{C}_{\boldsymbol{d}|\boldsymbol{\theta}} = \mathrm{Cov}(\boldsymbol{d}|\boldsymbol{\theta}) = E(\mathrm{G}(\boldsymbol{m}-\boldsymbol{\mu}_{\boldsymbol{m}})(\boldsymbol{m}-\boldsymbol{\mu}_{\boldsymbol{m}})^T\mathrm{G}^T) + E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T) = \mathrm{G}\boldsymbol{C}_{\boldsymbol{m}}\mathrm{G}^T + \boldsymbol{\Sigma} \tag{7}$$

Therefore, the conditional likelihood $p(\boldsymbol{d}_{obs}|\boldsymbol{\theta})$ is:

$$p(\boldsymbol{d}_{obs}|\boldsymbol{\theta}) = (2\pi)^{-n/2}(\det(\mathrm{G}\boldsymbol{C}_{\boldsymbol{m}}\mathrm{G}^T+\boldsymbol{\Sigma}))^{-1/2}\exp(-\frac{1}{2}(\boldsymbol{d}_{obs}-\mathrm{G}\boldsymbol{\mu}_{\boldsymbol{m}})^T(\mathrm{G}\boldsymbol{C}_{\boldsymbol{m}}\mathrm{G}^T+\boldsymbol{\Sigma}))^{-1}(\boldsymbol{d}_{obs}-\mathrm{G}\boldsymbol{\mu}_{\boldsymbol{m}})) \tag{8}$$

n is the dimension of $\boldsymbol{d}_{obs}$. Global geostatistical variables $\boldsymbol{\theta}$ contribute to $\boldsymbol{\mu}_{\boldsymbol{m}}$ and $\boldsymbol{C}_{\boldsymbol{m}}$ in this likelihood function, such as mean, range, variance, anisotropy.

We multiply this likelihood function with prior uncertainty to derive posterior distributions and sample posterior realizations using sampling method Markov chain Monte Carlo (MCMC):

$$p(\boldsymbol{\theta}|\boldsymbol{d}_{obs}) \propto p(\boldsymbol{d}_{obs}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \tag{9}$$

Any Markov chain Monte Carlo method can be used to sample this posterior in this linear Gaussian setting.

2.2.2 Non-linear inversion: machine learning-based inversion method

In this section, we propose a new machine learning-based inversion method to estimate posterior distribution. Now we extend the setting to any non-linear forward model and without making any Gaussian assumptions on $\boldsymbol{\varepsilon}$ and $\boldsymbol{m}$:

$$\boldsymbol{d} = g(\boldsymbol{m}) + \boldsymbol{\varepsilon} \tag{10}$$

Then the likelihood $p(\boldsymbol{d}_{obs}|\boldsymbol{\theta})$ is:

$$p(\boldsymbol{d}_{obs}|\boldsymbol{\theta}) = \int p(\boldsymbol{d}_{obs}|\boldsymbol{m},\boldsymbol{\theta})p(\boldsymbol{m}|\boldsymbol{\theta})\mathrm{d}\boldsymbol{m} \tag{11}$$

where we need to sample many spatial variables $\boldsymbol{m}$ given one specific global variables $\boldsymbol{\theta}$. Integrating over high-dimensional spatial variables $\boldsymbol{m}$ is practically infeasible. Unlike the derivation in Section 2.2.1, there is no analytical likelihood functions between $\boldsymbol{d}$ and $\boldsymbol{\theta}$ with non-linear forward models. We would also like to include non-geostatistical global variables such as flow properties.

Instead of estimating the likelihood function, we propose to estimate the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{d})$ directly from Monte Carlo samples generated from the prior, which does not need to be Gaussian. This approach is more typical in machine learning-based inversion, where posterior models are learned directly from training samples. Once the posterior distributions are estimated, sampling becomes easier, and there is no need for iterative samplers. Many approaches exist (Tipping & Bishop, 1999; Van Kerm, 2003; Botev et al., 2010; Silverman, 2018) to estimate high-dimensional densities from training data. Our approach will start from the fact that the amount of global variables is typically small. In a 3D spatial field, the spatial covariance has 3 ranges, 3 angles, and a variance (7 parameters), in a 2D field we have 4 parameters.

Our approach directly addresses the non-linear nature of the relationship between data and global variables, even under a linear forward model. For that reason we will train a neural network to predict global variables from the data. Neural networks are essentially regression models, whose output is often interpreted as a conditional expectation (Bishop et al., 1995). The inputs of the neural network are high-dimensional data

variables $\boldsymbol{d}$, the outputs are all global variables $\boldsymbol{\theta}$. Hence, effectively we are performing a dimension reduction, but using the global variables as target. The neural network projection is denoted as $S(.)$. We train $S(.)$ with the mean squared error loss function: $||S(\boldsymbol{d})-\boldsymbol{\theta}||^2$. The training samples are simply generated by forward simulating the data on prior model realizations.

Note that many machine learning methods can be used to learn this lower-dimensional projection $S(.)$: tree-based methods (random forests, gradient boosting, bagging), polynomial regression or neural networks. In this paper, we use neural networks with multiple hidden layers and non-linear activation functions to learn the non-linear relationship between $\boldsymbol{d}$ and $\boldsymbol{\theta}$.

What the neural network effectively does is to turn d into a summary statistic (Vrugt, 2016). Then, the joint distribution $p(\boldsymbol{d}, \boldsymbol{\theta})$ is approximated by the lower dimensional joint distribution $p(S(\boldsymbol{d})), \boldsymbol{\theta})$. We use kernel density estimation to estimate the joint distribution (Van Kerm, 2003; Botev et al., 2010; Scott, 2015; Scheidt et al., 2015), where we directly evaluate the posterior probability $p(\boldsymbol{\theta}|S(\boldsymbol{d}_{obs}))$. Then the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{d}_{obs})$ is approximated by $p(\boldsymbol{\theta}|S(\boldsymbol{d}_{obs}))$. We assume that the machine learning-based non-linear projection $S(\boldsymbol{d})$ preserves all relevant information from $\boldsymbol{d}$ to infer $\boldsymbol{\theta}$. The rest of information $\boldsymbol{d}_{\text{rest}}$ and $\boldsymbol{\theta}$ are assumed to have become independent:

$$p(\boldsymbol{\theta}|\boldsymbol{d}_{\text{rest}}) \approx p(\boldsymbol{\theta}) \tag{12}$$

so that

$$p(\boldsymbol{\theta}|\boldsymbol{d}) = p(\boldsymbol{\theta}|\boldsymbol{d}_{\text{rest}}, S(\boldsymbol{d})) \approx p(\boldsymbol{\theta}|S(\boldsymbol{d})) \tag{13}$$

Algorithm 2: machine learning-based inversion method

1) Sample L global variables realizations from their prior distribution $p(\boldsymbol{\theta})$: $\boldsymbol{\theta}^{(\ell)}, \ell = 1, 2, ..., L$

2) Sample L corresponding spatial variables realizations given each set of global variables $\boldsymbol{\theta}^{(\ell)}$: $\boldsymbol{m}^{(\ell)}$

3) Apply the non-linear forward model $g(.)$ on each $\boldsymbol{m}^{(\ell)}$, sample corresponding data variable $\boldsymbol{d}^{(\ell)} = g(\boldsymbol{m}^{(\ell)}) + \boldsymbol{\varepsilon}^{(\ell)}$ with measurement error $\boldsymbol{\varepsilon}^{(\ell)} \sim N(0, \boldsymbol{\Sigma})$

4) Train a non-linear machine learning projection $S(\boldsymbol{d})$ using the training set $(\boldsymbol{d}^{(\ell)}, \boldsymbol{\theta}^{(\ell)}), \ell = 1, 2, ..., L$

5) Estimate the joint distribution $p(\boldsymbol{\theta}, S(\boldsymbol{d}))$ using kernel density estimation, and estimate the conditional posterior distribution $p(\boldsymbol{\theta}|S(\boldsymbol{d}_{obs}))$

Step 1-3 in Algorithm 2 summarizes how to construct a training dataset from Monte Carlo sampling. Note that we directly incorporate measurement errors $\boldsymbol{\varepsilon}^{(\ell)}$ in sampled data variables $\boldsymbol{d}^{(\ell)}$. If $\boldsymbol{\varepsilon}^{(\ell)}$ are larger and obscure the underlying data, the trained ML model has larger training errors. We use kernel density estimation to directly estimate large training errors. Later in the application section, we will present global variables estimations with different levels of measurement errors. In our setting, the errors can be correlated and don't need to be Gaussian or additive.

2.3 Posterior distribution of large-scale spatial fields: local principal component analysis

The next step in the hierarchical Bayes formulation is to invert high-dimensional spatial variables $p(\boldsymbol{m}|\boldsymbol{\theta}^*, \boldsymbol{d}_{obs})$ given posterior global variables $\boldsymbol{\theta}^* \sim p(\boldsymbol{\theta}|\boldsymbol{d}_{obs})$. In this section, we first focus on a general method on large-spatial field inversions then move to local inversions of large-spatial field with local measurements.

Spatial fields are very high-dimensional with more than thousands or even millions of unknowns, if a gridded parameterization is used. Therefore, dimension reduction techniques are preferred to estimate posterior distribution of spatial variables more efficiently.

One way to reduce this high dimensionality of spatial variables is through principal component analysis (Kitanidis & Lee, 2014; Lee & Kitanidis, 2014; Sun & Durlofsky, 2017; Scheidt et al., 2018; Yin et al., 2020; Fouedjio et al., 2021):

$$\boldsymbol{m} = \boldsymbol{\mu_m} + \boldsymbol{V}\boldsymbol{y} \tag{14}$$

where $\boldsymbol{V}$ is the set of principal components, $\boldsymbol{y}$ is corresponding low-dimensional principal component scores, $\boldsymbol{\mu_m}$ is the mean for $\boldsymbol{m}$.

We perform eigenvalue decomposition (EVD) to decompose covariance matrix of the spatial variables into eigenvalues and eigenvectors. The eigenvectors are termed principal components $\boldsymbol{V}$. Performing EVD on a large $p \times p$ covariance matrix is time-consuming where spatial variables have many p unknowns. We use Monte Carlo sampling and linear algebra properties to overcome this high-dimensional large matrix challenge. First, we sample N spatial variables, remove the mean of spatial fields $\boldsymbol{\mu_m}$ and create the spatial variable matrix X:

$$\boldsymbol{X} = [\boldsymbol{m}^{(1)}, \boldsymbol{m}^{(2)}, ..., \boldsymbol{m}^{(N)}] - \boldsymbol{\mu_m} \tag{15}$$

which is a $p \times N$ matrix. Then instead of performing eigenvalue decomposition on the estimated covariance matrix $\boldsymbol{X}\boldsymbol{X}^T$ ($p \times p$ matrix):

$$\boldsymbol{X}\boldsymbol{X}^T = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T, \boldsymbol{X}\boldsymbol{X}^T v_i = \lambda_i v_i, i = 1, ..., p \tag{16}$$

we perform eigenvalue decomposition on a smaller matrix $\boldsymbol{X}^T\boldsymbol{X}$ ($L \times L$ matrix):

$$\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T, \boldsymbol{X}^T\boldsymbol{X} u_i = \lambda_i u_i, i = 1, ..., N \tag{17}$$

where $\boldsymbol{U}$ is the set of eigenvectors of $\boldsymbol{X}^T\boldsymbol{X}$, with the length of N and $\boldsymbol{\Lambda}$ is the diagonal matrix of eigenvalues. $\boldsymbol{X}\boldsymbol{X}^T$ and $\boldsymbol{X}^T\boldsymbol{X}$ share the same non-zero eigenvalues $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \lambda_2, ..., \lambda_N)$.

The principal components $\boldsymbol{V}$ of $\boldsymbol{X}\boldsymbol{X}^T$ are:

$$\boldsymbol{V} = \boldsymbol{X}\boldsymbol{U} \tag{18}$$

with the length of p. The proof of this derivation is as follows: multiplying $\boldsymbol{X}$ on each side of the equation,

$$\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{X} u_i = \boldsymbol{X}\lambda_i u_i \tag{19}$$

$$\boldsymbol{X}\boldsymbol{X}^T(\boldsymbol{X} u_i) = \lambda_i(\boldsymbol{X} u_i) \tag{20}$$

$$\boldsymbol{X}\boldsymbol{X}^T v_i = \lambda_i v_i \tag{21}$$

The corresponding principal component scores y are:

$$\boldsymbol{y} = \boldsymbol{V}^T\boldsymbol{m} = [v_1 \cdot \boldsymbol{m}, v_2 \cdot \boldsymbol{m}, ..., v_N \cdot \boldsymbol{m}]^T \tag{22}$$

with a maximal dimension N. We can truncate the dimension at the certain amount of variance to reduce dimensions even more. We then estimate the posterior distribution of lower dimensional scores $p(\boldsymbol{y}|\boldsymbol{\theta}^*, \boldsymbol{d}_{obs})$ using Bayes-linear-Gaussian method (Tarantola, 2005) for the linear forward modeling or ensemble smoother with multiple data assimilation (Van Leeuwen & Evensen, 1996; Emerick & Reynolds, 2013) for the non-linear forward modeling. We reconstruct posterior spatial fields with all principal components multiplying posterior lower dimensional scores (Eq. 14).

Principal component analysis focuses on the variance of the full/entire spatial field. However, the observed data may be collected locally and inform that field locally only. Examples are pumping tests or a local high resolution geophysical survey. Then this observed data $\boldsymbol{d}_{obs}$ is unlikely to be informative of principal component scores $\boldsymbol{y}$ on full fields.

Therefore, we propose a dimension reduction method that account for the local information provided by data: local principal component analysis (local PCA). We state a local area (see examples in Section 3.2 on how this is done) and denote the local spatial field as $\boldsymbol{m}_{loc}$, a subvector of $\boldsymbol{m}$. We then perform eigenvalue decomposition on ${\boldsymbol{X}_{loc}}^T \boldsymbol{X}_{loc}$, where $\boldsymbol{X}_{loc} = [\boldsymbol{m}_{loc}^{(1)}, \boldsymbol{m}_{loc}^{(2)}, ..., \boldsymbol{m}_{loc}^{(N)}]$ is a $q \times N$ matrix. q is the number of local spatial variables.

$$ {\boldsymbol{X}_{loc}}^T \boldsymbol{X}_{loc} = \boldsymbol{U}_{loc} \boldsymbol{\Lambda} \boldsymbol{U}_{loc}^T \tag{23} $$

The local principal components $\boldsymbol{V}_{loc}$ for $\boldsymbol{X}_{loc} {\boldsymbol{X}_{loc}}^T$ are:

$$ \boldsymbol{V}_{loc} = \boldsymbol{X}_{loc} \boldsymbol{U}_{loc} \tag{24} $$

The corresponding principal component scores $\boldsymbol{y}'$ are:

$$ \boldsymbol{y}' = {\boldsymbol{V}_{loc}}^T \boldsymbol{m}_{loc} \tag{25} $$

We estimate posterior $\boldsymbol{y}'$ and reconstruct posterior local spatial fields. We also want to have full posterior spatial fields. Therefore, we calculate full principal components $\boldsymbol{V}'$ which are extended from local principal components $\boldsymbol{V}_{loc}$:

$$ \boldsymbol{V}' = \boldsymbol{X} \boldsymbol{U}_{loc} \tag{26} $$

where $\boldsymbol{V}_{loc}$ is a subvector for $\boldsymbol{V}'$ and $\boldsymbol{X}_{loc}$ is a submatrix of $\boldsymbol{X}$. Full principal components $\boldsymbol{V}'$ are continuous at the boundary of local area. The continuity of $\boldsymbol{V}'$ is established through the spatial continuity of Monte Carlo samples $\boldsymbol{X}$.

Algorithm 3: Local principal component analysis

1) Sample N spatial variables realizations given one posterior $\boldsymbol{\theta}^*$: $\boldsymbol{m}^{(\ell)}, \ell = 1, 2, ..., L$
2) State the target local spatial area
3) Construct matrix $\boldsymbol{X} = [\boldsymbol{m}^{(1)}, \boldsymbol{m}^{(2)}, ..., \boldsymbol{m}^{(N)}]$ and submatrix $\boldsymbol{X}_{loc} = [\boldsymbol{m}_{loc}^{(1)}, \boldsymbol{m}_{loc}^{(2)}, ..., \boldsymbol{m}_{loc}^{(N)}]$.
4) Perform eigenvalue decomposition on ${\boldsymbol{X}_{loc}}^T \boldsymbol{X}_{loc}$: ${\boldsymbol{X}_{loc}}^T \boldsymbol{X}_{loc} = \boldsymbol{U}_{loc} \boldsymbol{\Lambda} \boldsymbol{U}_{loc}^T$.
5) Calculate local principal components $\boldsymbol{V}_{loc} = \boldsymbol{X}_{loc} \boldsymbol{U}_{loc}$ and full principal components $\boldsymbol{V}' = \boldsymbol{X} \boldsymbol{U}_{loc}$.
6) Calculate local principal component scores $\boldsymbol{y}' = \boldsymbol{V}_{loc}^T \boldsymbol{m}_{loc}$
7) Denote spatial variables as $\boldsymbol{m} = \boldsymbol{\mu}_{\boldsymbol{m}} + \boldsymbol{V}' \boldsymbol{y}'$ and local spatial variables as $\boldsymbol{m}_{loc} = \boldsymbol{\mu}_{\boldsymbol{m}_{loc}} + \boldsymbol{V}_{loc} \boldsymbol{y}'$

Step 7 in Algorithm 3 denotes spatial variables $\boldsymbol{m}$ and $\boldsymbol{m}_{loc}$ using local PCA. We then reconstruct both posterior local and full spatial fields with all principal components given posterior $\boldsymbol{y}'$.

3 Applications and Discussion

3.1 Linear forward model: volume averaging

We start with a case involving a linear forward model. A simple example is that of linear averaging over some defined volume or area. Such forward models are common, such as measuring averaged moisture content or averaged electrical conductivity from indirect geophysical survey. Our synthetic case has a low resolution averaged property as data. The goal is to estimate spatial fields in high resolution thereby inferring detailed geological features (Fig. 1). The true spatial model $\boldsymbol{m}$ (Fig. 1, left) is 2D spatial Gaussian field, a 3.5 km $\times$ 1.5 km field with the lateral resolution 50m. This spatial field $\boldsymbol{m}$ has corresponding global variables $\boldsymbol{\theta}$: mean = 29, variance = 12, maximum range = 25*(50m), minimum range = 8*(50m), major anisotropic direction = 20 degrees. Prior uncertainty

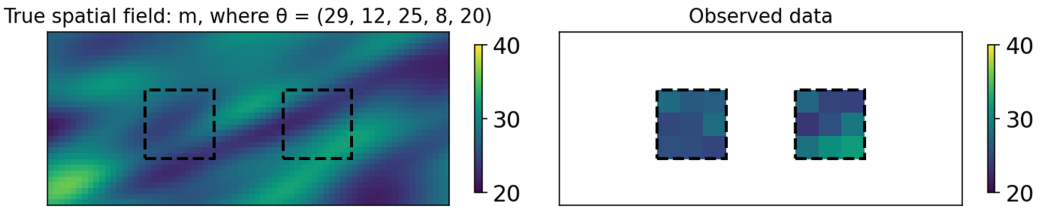

Figure 1. The true spatial field $\boldsymbol{m}$ (left) and the observed low-resolution data $\boldsymbol{d}_{obs}$ (right). The true global variables $\boldsymbol{\theta}$ are: mean = 29, variance = 12, maximum range = 25*(50m), minimum range = 8*(50m), major anisotropic direction = 20 degrees

Global variables θ	Mean	Variance	Maximum range	Minimum range	Anisotropy
Uniform distribution	U(28,12)	U(9,16)	U(15, 30)	U(5,15)	U(0,180)

Table 1. Prior uncertainty of $\boldsymbol{\theta}$ in the linear Gaussian study

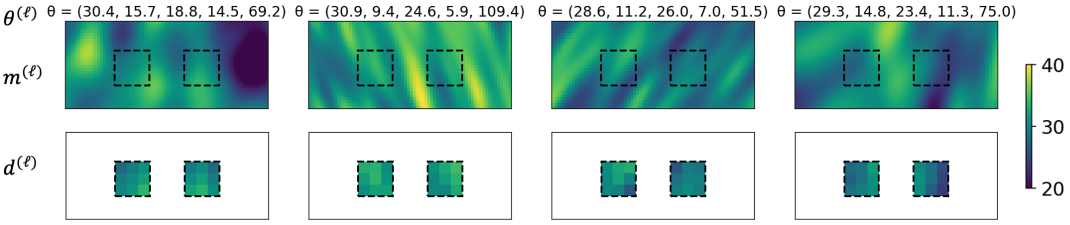

Figure 2. Monte Carlo sampling results (4 realizations): an ensemble of global variables, corresponding spatial variables and data variables

$p(\boldsymbol{\theta})$ is stated in Table 1. We perform linear volume averaging within two rectangular areas. The forward model G is therefore linear averaging 4×4 high resolution spatial fields. The observed data $\boldsymbol{d}_{obs} = \mathrm{G}\boldsymbol{m}$ are averaged values with lateral resolution 200m (Fig. 1, right).

3.1.1 Monte Carlo and sensitivity analysis

Before going further with hierarchical Bayesian inversion, we first explore the relationship between low resolution data variables and global model variables using sensitivity analysis. Sensitivity analysis studies the relationship between the variations of input model variables and output forward modeling responses. In our case, we are interested in whether the variations of global variables change the observed low resolution data, which can be answered by the sensitivity analysis method. We apply a distance-based generalized sensitivity analysis method (DGSA) (Fenwick et al., 2014; J. Park et al., 2016) to an ensemble of model and data variables using Monte Carlo sampling, following the first 3 steps in Algorithm 2. We sample $L = 5,000$ sets of variables as shown in Figure 2. DGSA method first classifies the data variables into multiple different classes using a distance metric, and then see if the global variable distributions in different classes are similar or different. The similarity indicates the data output is insensitive to the global variable, while the difference means the data output is sensitive to the global variable.

Here we perform principal component analysis on data variables first and use the euclidean distance metric on each individual principal component score. In Figure 3, each

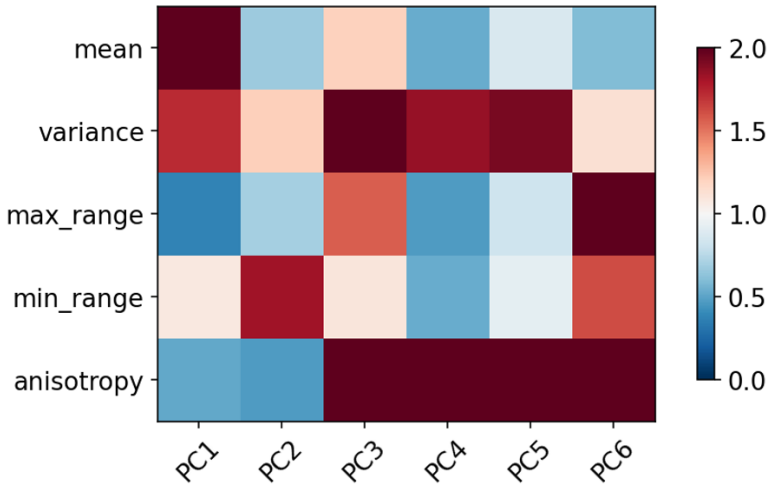

Figure 3. Distance-based Generalized Sensitivity Analysis results with euclidean distance between individual principal component scores: > 1 sensitive, < 1 insensitive

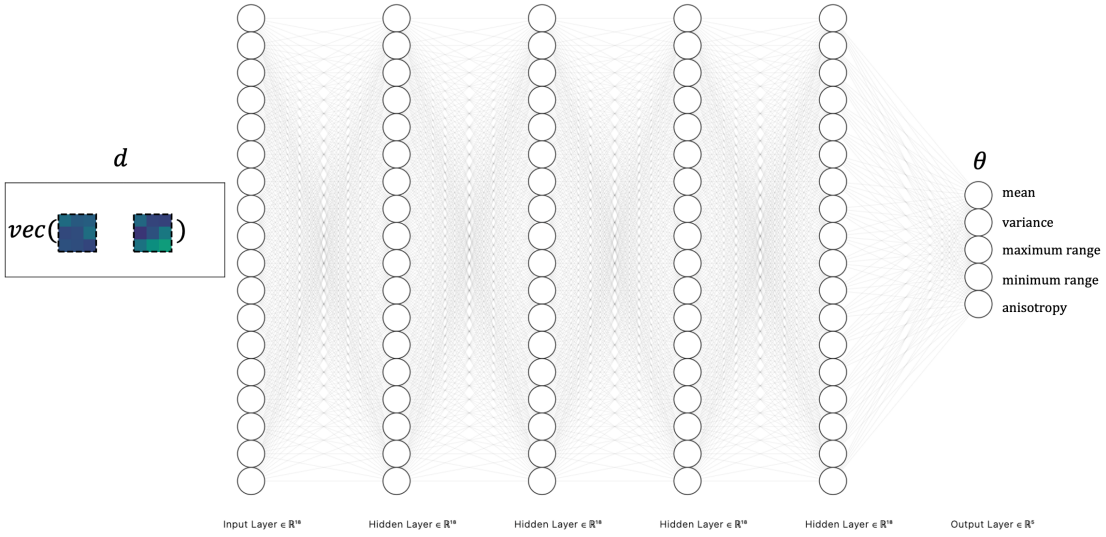

Figure 4. Neural network structure for the machine learning-based dimension reduction

colored block indicates the sensitivity between one global variable and one principal component score from $\boldsymbol{d}$. For example, the bottom right corner red block means the observed 6-th principal component score is sensitive to the major anisotropic direction. We note that the output data variables are sensitive to all global variables. We also learn that the relationship between global model variables and data variables are complex and possibly non-linear.

3.1.2 Non-linear inversion of global variables using machine learning-based inversion method

In the previous Section 3.1.1, we investigate the relationship between observed data and global variables using sensitivity analysis. To perform non-linear inversion on global variables, we use a 5-layer neural network $S(.)$. The inputs are vectorized data variables

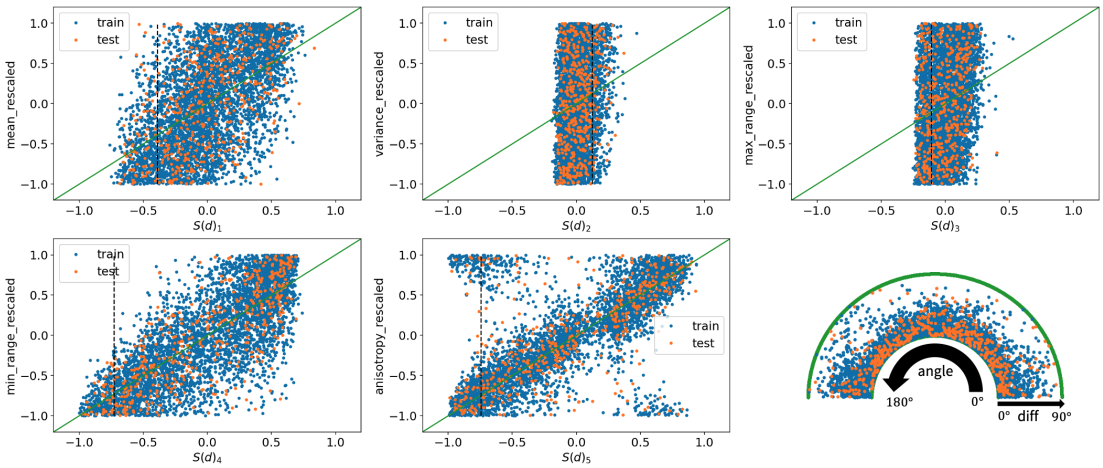

Figure 5. Training results from ML-assisted dimension reduction. First five plots: the x-axis is the lower dimensional $S(d)$, the y-axis is the re-scaled θ, the dashed vertical line is the lower dimensional observed data $S(d_{obs})$. The last plot: anisotropy angle axis is from 0 degree to 180 degree, difference axis between prediction and true angles is from 0 degree to 90 degree.

$\boldsymbol{d}$ and the outputs are re-scaled global variables $\boldsymbol{\theta}$ (Figure. 4), re-scaled into a uniform distribution between -1 and 1. The activation functions for all layers except the last one are the Rectified Linear Unit (ReLU) activation. ReLU activation is a piece-wise linear function. The activation function of the last layer is a hyperbolic tangent function (tanh), to output the value between -1 and 1. We split 5,000 Monte Carlo ensembles into a 4,500 training set and a 500 test set. The loss functions for all variables except the anisotropy is mean squared error. We define a new loss function $\mathcal{L}_{aniso}$ to account for the circular similarity where 0 and 180 degree are same:

$$\mathcal{L}_{aniso} = \min((\theta_{aniso} - \hat{\theta}_{aniso})^2, (180 - |\theta_{aniso} - \hat{\theta}_{aniso}|)^2) \tag{27}$$

which is a circular squared error. The maximum L_1 distance $|\theta_{aniso} - \hat{\theta}_{aniso}|$ between any anisotropy prediction $\hat{\theta}_{aniso}$ and true anisotropy value θ_{aniso} is 90 degree.

Figure. 5 shows the neural network training results. Data variables well predict mean, minimum range and anisotropy. We use kernel density estimation to estimate the joint distribution $p(S(\boldsymbol{d}), \boldsymbol{\theta})$ and then directly evaluate the posterior distribution $p(\boldsymbol{\theta}|S(\boldsymbol{d}_{obs}))$ (Figure. 6). The true values of global variables have high posterior probability.

3.1.3 Non-linear inversion of global variables with different levels of measurement errors

We have not considered any measurement error in the forward model. Now we include different levels of measurement errors with the purpose of quantifying how errors affect posterior estimations. We use eight different levels of measurement errors: $\boldsymbol{\Sigma} = [1, 2, 5, 10, 20, 30, 40, 50\%$ of the data variance var$(\boldsymbol{d})]$ and $\boldsymbol{\varepsilon} \sim N(0, \boldsymbol{\Sigma})$. We add those errors to both data variables $\boldsymbol{d}$ and observed data variables $\boldsymbol{d}_{obs}$ (Figure. 7).

Measurement errors $\boldsymbol{\varepsilon}$ increase machine learning training errors (Figure. 8). We estimate large training errors directly using kernel density estimation. Figure. 9 shows posterior distributions for the mean and the minimum range with different levels of measurement errors $\boldsymbol{\varepsilon}$. The estimated posterior mean stays the same with the increase of $\boldsymbol{\varepsilon}$.

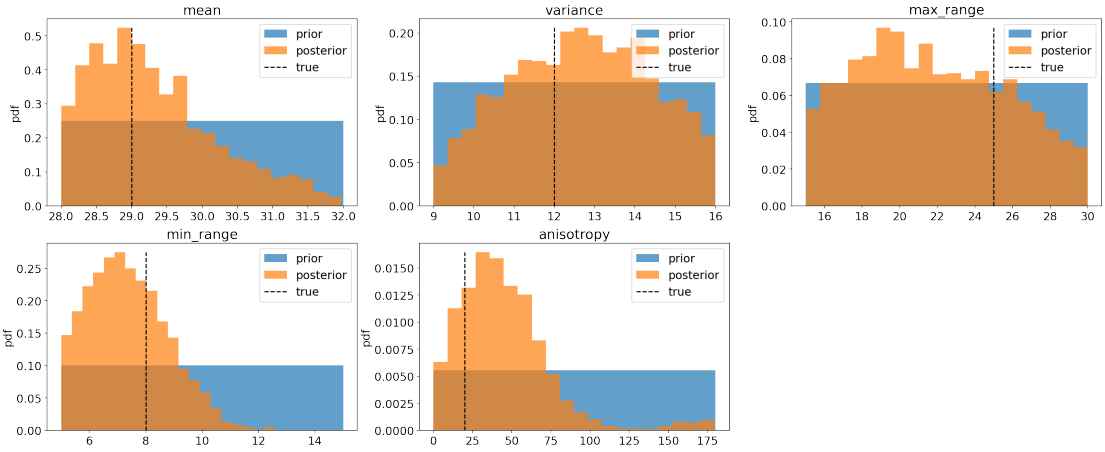

Figure 6. Posterior distribution of global variables using ML-assisted dimension reduction and kernel density estimation

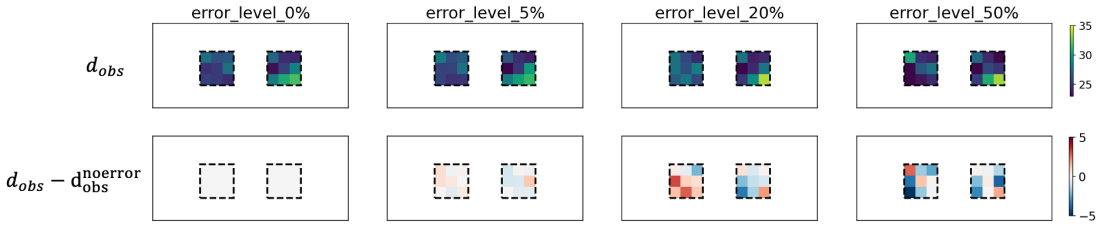

Figure 7. Different levels of measurement errors on observed data

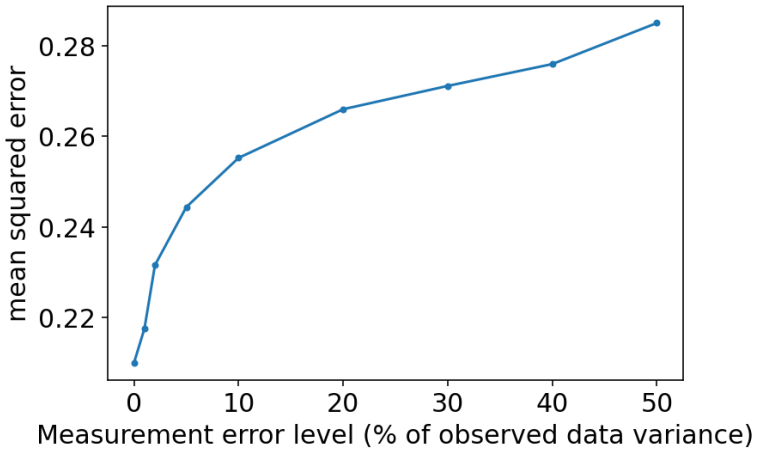

Figure 8. Training errors with different levels of measurement errors

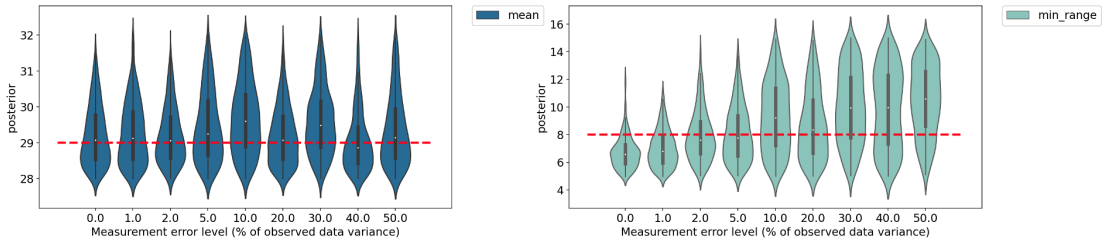

Figure 9. Posterior distribution of mean and minimum range with different levels of measurement error. The red dashed line is the true global variable value.

$\boldsymbol{\varepsilon}$ have different levels of variance but the same zero mean. We have a wider posterior minimum range when $\boldsymbol{\varepsilon}$ increase.

3.1.4 Model comparison: machine learning-based inversion and statistical sampling with analytical likelihood functions

In Section 2.2.1, we formulated a posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{d}_{obs})$ by multiplying a Gaussian likelihood with uniform priors, where $\boldsymbol{m}$ and $\boldsymbol{\varepsilon}$ are Gaussian distributed and the forward model G is linear. Sampling is done using Markov chain Monte Carlo (MCMC). We use the No-U-Turn Sampler (NUTS) MCMC method (Hoffman et al., 2014), which is an extension of Hamilton Monte Carlo (HMC) (Betancourt, 2017). HMC uses the gradient information to avoid the random walk behavior. NUTS method frees users from defining step size and the number of steps in HMC and gives users an automatic and robust MCMC sampler. Posterior distribution of global variables are presented in Figure. 10.

Our goal is to compare two different formulations of the same problem: 1) statistical sampling with analytical likelihood functions and 2) ML-based inversion method. Figure. 10 and 6 show that posterior of anisotropy is narrower in the analytical likelihood formulation for than in ML-based inversion. The reason for this is that each formulation makes different assumptions. In the first formulation, we assume Gaussian $\boldsymbol{m}$ and $\boldsymbol{\varepsilon}$ to obtain an analytical likelihood. In the ML-based inversion no such assumption is made, the posterior $p(\boldsymbol{\theta}|S(\boldsymbol{d}_{obs}))$ is estimated directly from Monte Carlo samples after the ML-based projection $S(.)$. Note how anisotropy > 50 degree (Figure. 11, right)

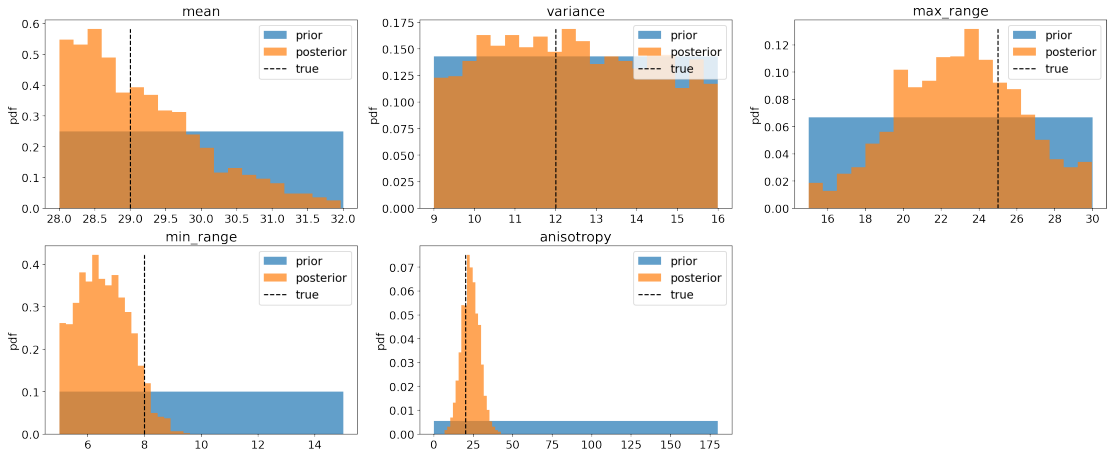

Figure 10. Posterior distribution of global variables using statistical sampling method Markov chain Monte Carlo, with the analytical likelihood function

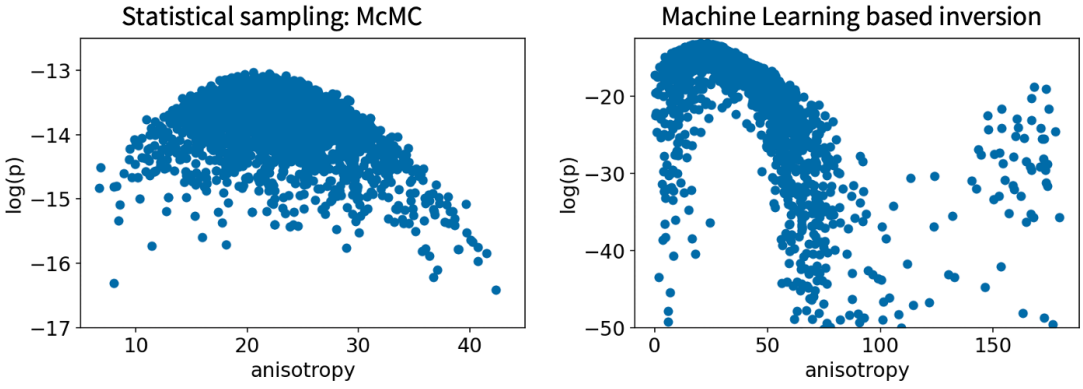

Figure 11. Log(posterior probability) of posterior global variables using 1) statistical sampling: MCMC 2) ML-based inversion

has low posterior probability under the first formulation. As we will show in the next section, this is not due to poor matching of the data (in fact we can match the data perfectly), but due to strict likelihood assumption made under the first formulation.

3.1.5 Why quantifying uncertainty of global variables is important

In this section, we show the importance of quantifying uncertainty of global variables by comparing three scenarios:

- Scenario A: Use a set of deterministic values $\boldsymbol{\theta}^0$ for global variable $\boldsymbol{\theta}$, only invert spatial variable $\boldsymbol{m}$: $p(\boldsymbol{m}|\boldsymbol{d}_{obs}, \boldsymbol{\theta}^0)$
- Scenario B: State prior uncertainty for global variables: $p(\boldsymbol{\theta})$, only invert spatial variable $\boldsymbol{m}$: $p(\boldsymbol{m}|\boldsymbol{d}_{obs}, \boldsymbol{\theta}), \boldsymbol{\theta} \sim p(\boldsymbol{\theta})$
- Scenario C: State prior uncertainty for global variable: $p(\boldsymbol{\theta})$, invert both global and spatial variables using hierarchical Bayes: $p(\boldsymbol{m}, \boldsymbol{\theta}|\boldsymbol{d}_{obs}) = p(\boldsymbol{m}|\boldsymbol{\theta}, \boldsymbol{d}_{obs})p(\boldsymbol{\theta}|\boldsymbol{d}_{obs})$

In all scenarios, we invert spatial variables using Bayes-linear-Gauss inversion (Tarantola, 2005), where given any $\boldsymbol{\theta}$, the posterior of spatial variable $\boldsymbol{m}$ is $N \sim (\boldsymbol{\mu}_{\boldsymbol{m}}^{pos}, \boldsymbol{C}_{\boldsymbol{m}}^{pos})$:

$$\boldsymbol{\mu}_{\boldsymbol{m}}^{pos} = \boldsymbol{\mu}_{\boldsymbol{m}} + \boldsymbol{C}_{\boldsymbol{m}} G^T (G\boldsymbol{C}_{\boldsymbol{m}} G^T + \boldsymbol{\Sigma})^{-1} (\boldsymbol{d}_{obs} - G\boldsymbol{\mu}) \tag{28}$$

$$\boldsymbol{C}_{\boldsymbol{m}}^{pos} = \boldsymbol{C}_{\boldsymbol{m}} - \boldsymbol{C}_{\boldsymbol{m}} G^T (G\boldsymbol{C}_{\boldsymbol{m}} G^T + \boldsymbol{\Sigma})^{-1} G\boldsymbol{C}_{\boldsymbol{m}} \tag{29}$$

Figure. 12 shows inversion results in all scenarios. Scenario A1 has a set $\boldsymbol{\theta}^0 = (29, 12, 25, 8, 45)$, close to the truth. In Scenario A2, we take anisotropy = 110 degree, far from the truth. Spatial uncertainty has been reduced significantly along 45 degree in Scenario A1 and 110 degree in Scenario A2. Scenario A2 has a biased posterior mean outside two observed rectangles. Posterior spatial variables are highly dependent on the chosen values of the global variables.

In Scenario B, we state the prior uncertainty of global variables (Table 1). Spatial uncertainty has been reduced mostly within the observed area. If we do not quantify global variables' uncertainty hierarchically, we end up with too uncertain spatial fields. Scenario B is acceptable only if observed data does not inform any global variables.

Scenario C1 uses ML-based inversion method. Recall in the last section, we have posterior anisotropy values larger than 50 degree. Posterior spatial fields with Anis = 55 and 143 still match the observed data perfectly. Scenario C2 uses statistical sampling with the analytical likelihood function, which has posterior spatial fields less uncertain than the ML-based estimation. Therefore, the wider anisotropy in ML-based formulation is not due to a mismatch of data, it is due to the relaxed linear Gaussian assumption.

3.2 Non-linear forward model: pumping test

In this section, we use a synthetic non-linear forward model: groundwater simulation of a pumping test to illustrate hierarchical Bayesian inversion. We assume one pumping test in the middle of a 1km by 1km spatial field, with a pumping rate 10 m^3/day. This pumping test starts at Day 1, constantly pumps for 2 days, shuts down the pumping and recovers for 4 days. Uncertain global variables include geostatistical parameters for log hydraulic conductivity spatial fields, and specific head boundary condition at the west boundary (Figure 13 left). Specific head boundary condition at east is 10m and the north and south boundary have no flow boundaries. The horizontal resolution for each spatial grid is 10m.

The goal is to invert all global variables including geostatistical parameters and global boundary condition, and log hydraulic conductivity spatial field given the observed draw-

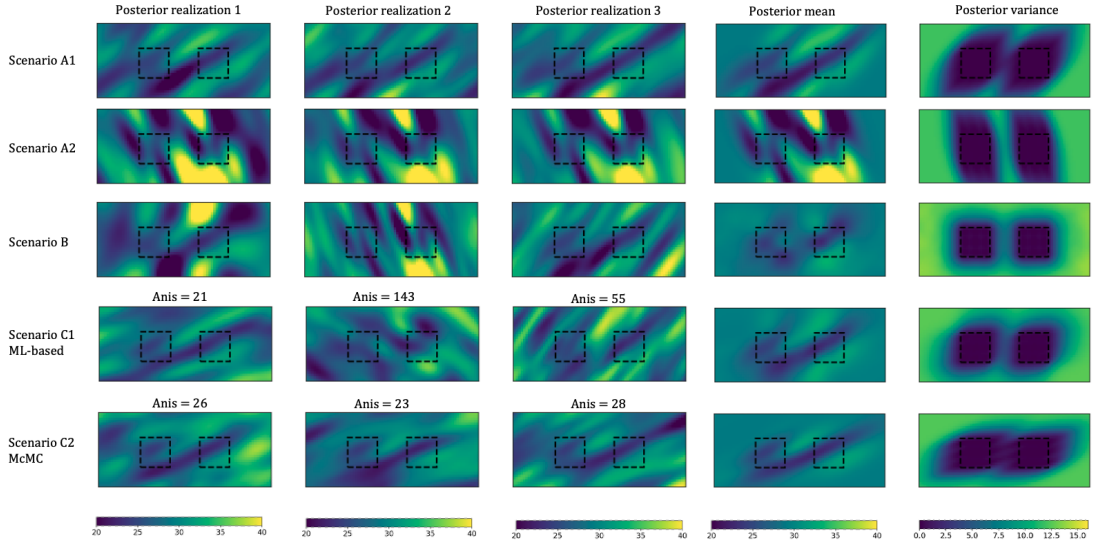

Figure 12. Posterior realizations, posterior mean and variance of spatial variable $\boldsymbol{m}$ in different scenarios

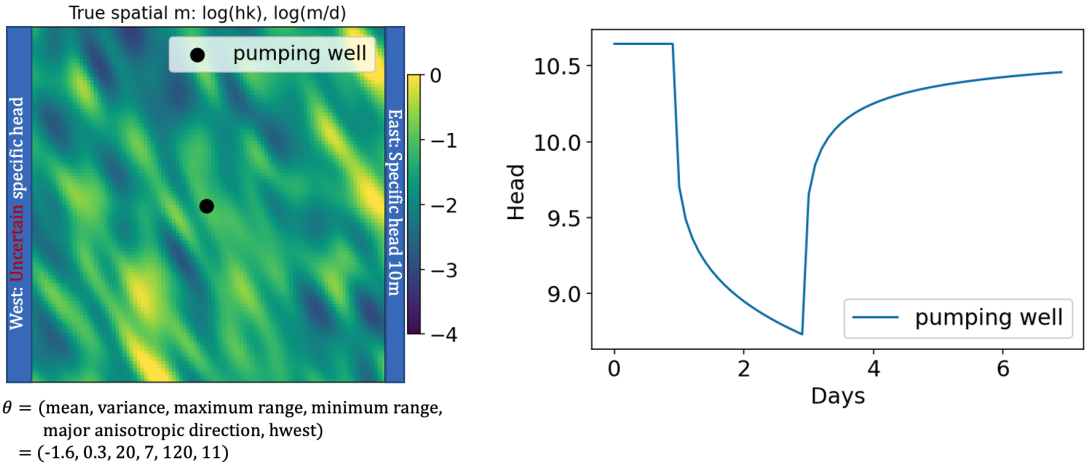

Figure 13. True spatial log hydraulic conductivity and the observed drawdown curve

Global variables θ	Mean	Variance	Max range	Min range	Anisotropy	Specific head at west
Uniform distribution	U(-3, -1)	U(0.1,1)	U(15, 30)	U(5,15)	U(0,180)	U(8,12)

Table 2. Prior uncertainty of $\boldsymbol{\theta}$, non-linear pumping test study

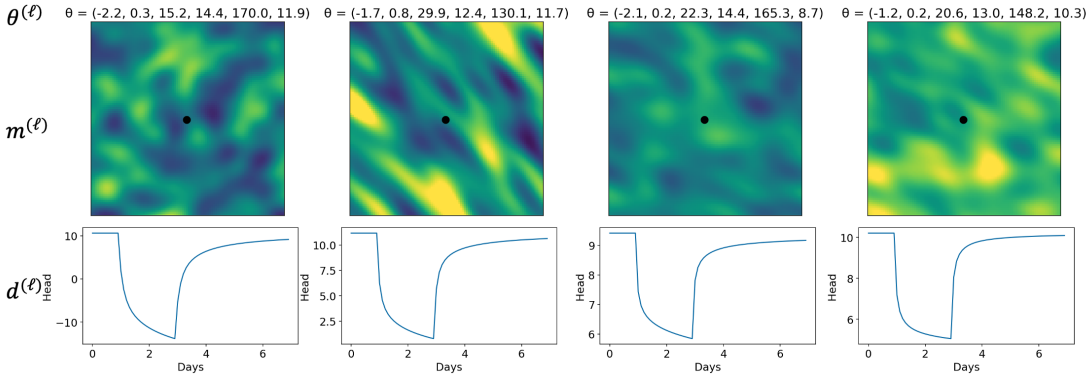

Figure 14. Monte Carlo sampling results for the pumping test case (4 realizations): an ensemble of global variables, corresponding spatial variables and data variables. The colorbar for $\boldsymbol{m}$ is the same with the colorbar in Figure 13

.

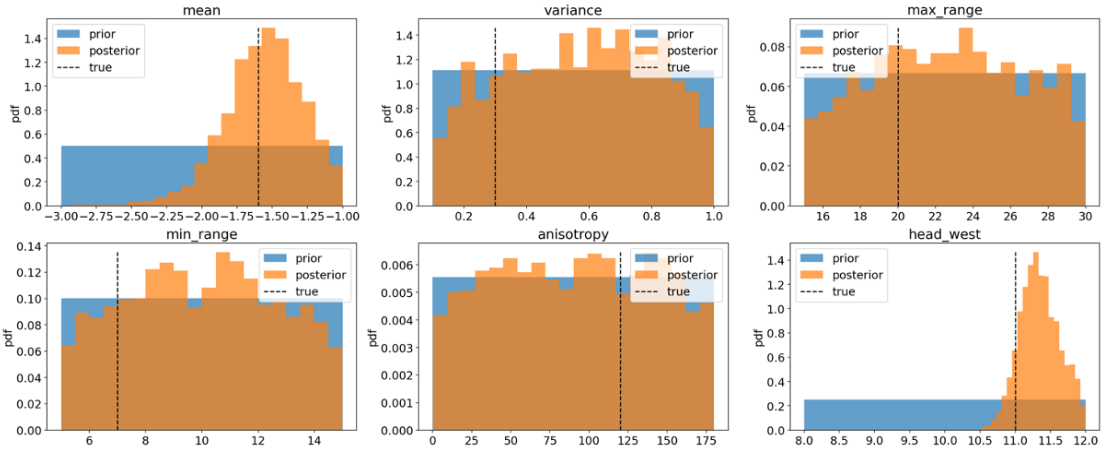

Figure 15. Posterior distribution of global variables in the non-linear case

down curve (Figure 13 right). Prior uncertainty for all global variables are in Table 3.2. The forward groundwater flow model is non-linear: $\boldsymbol{d} = g(\boldsymbol{m}) + \boldsymbol{\varepsilon}$, with no measurement error $\boldsymbol{\varepsilon} = 0$. We obtain a training set by Monte Carlo sampling with 5,000 samples (Figure 14). Then we apply ML-based inversion method to estimate posterior global variables $p(\boldsymbol{\theta}|d_{obs})$ (Figure 15). Posterior log hydraulic conductivity mean and specific head at west have been updated more than other global variables.

3.2.1 Sensitivity analysis with local principal component analysis

In this section, we investigate how the drawdown data informs the spatial field. Specifically, we want to test if the drawdown data is sensitive to the full spatial field or the nearby local spatial field. Using the hierarchical formulation, we sample one $\boldsymbol{\theta}^* = (-1.6, 0.2, 16.5, 8.4, 165, 11.8)$ from $p(\boldsymbol{\theta}|\boldsymbol{d}_{obs})$ and generate $N = 100$ spatial fields. Each spatial field has dimension: 100×100.

We first do sensitivity analysis with PCA on full fields. We refer to PCA on full fields as global PCA. The sensitivity analysis result is shown in Figure 16 (top left). We note that the drawdown is insensitive to any of the first ten principal components of the

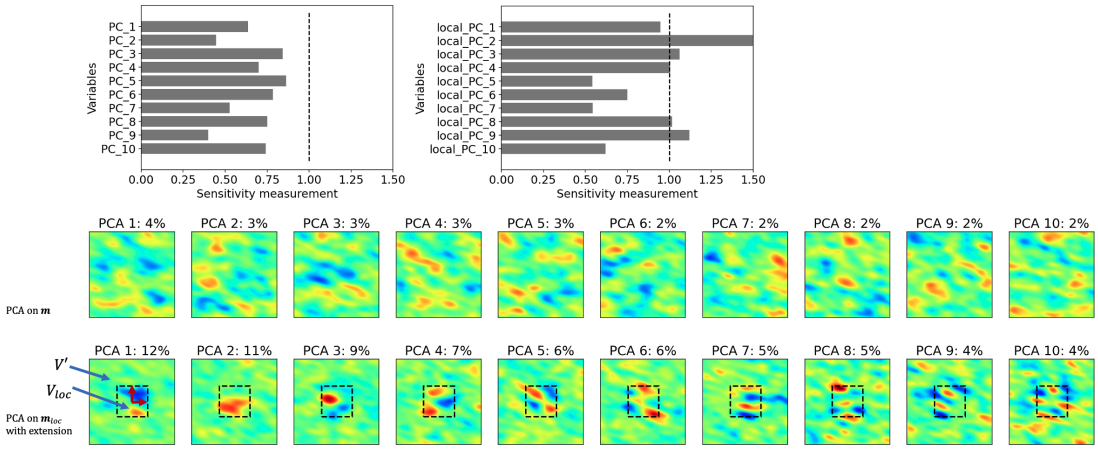

Figure 16. Sensitivity analysis on the principal components from $\boldsymbol{m}$ and $\boldsymbol{m}_{loc}$ and visualizations of those principal components

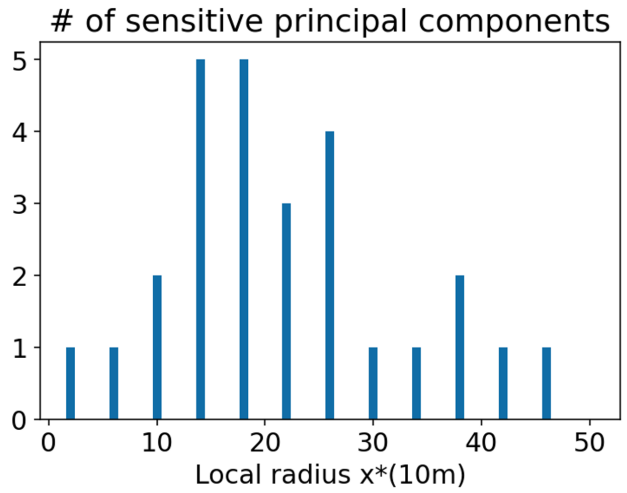

Figure 17. Sensitivity analysis with different local radii. The Y-axis is the number of sensitive principal components in the first 10 PCs

full field $\boldsymbol{m}$. We then perform sensitivity analysis with local PCA. We state the local spatial field $\boldsymbol{m}_{loc}$ with a radius of $18*(10m)$ near the pumping well (red arrows in Figure 16). Local principal components $\boldsymbol{V}_{loc}$ and full principal components $\boldsymbol{V}'$ are shown in Figure 16. The drawdown data is sensitive to five local principal components among first ten. There is no sensitivity to global PC, but there is for local PC. This then means that we need to use the local PC method to invert with the local drawdown data.

We now present a method for determining the size of local area. We perform multiple sensitivity analyses with different radii (Figure 17). The drawdown data is less sensitive to local principal components with too small or too large radii. The most sensitive local area has the local radius around 18*(10m).

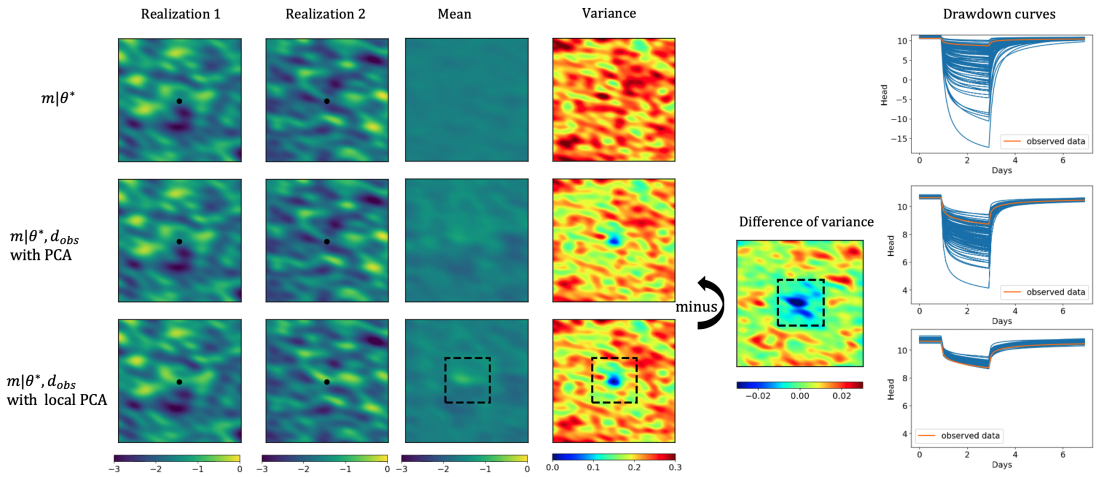

Figure 18. Posterior log hydraulic conductivity fields and corresponding drawdown curves using ensemble smoother with only one iteration. Row 1: spatial fields $\boldsymbol{m}$ given θ^*. Row 2: updated spatial fields using principal component analysis only. Row 3: updated extended spatial fields using local principal component analysis

3.2.2 Local non-linear inversion

To perform spatial inversion, we estimate the posterior distribution of principal component scores $\boldsymbol{y}$ using ensemble smoother with multiple data assimilation (ES-MDA):

$$\boldsymbol{y}_{pos}^{(\ell)} = \boldsymbol{y}^{(\ell)} + \boldsymbol{C}_{yd}(\boldsymbol{C}_{dd} + \alpha * \Sigma)^{-1}(\boldsymbol{d}_{obs} - \boldsymbol{d}^{(\ell)}) = \boldsymbol{y}^{(\ell)} + \boldsymbol{C}_{yd}\boldsymbol{C}_{dd}^{-1}(\boldsymbol{d}_{obs} - \boldsymbol{d}^{(\ell)}) \tag{30}$$

where $\boldsymbol{C}_{yd}$ is the cross-covariance matrix between $\boldsymbol{y}$ and $\boldsymbol{d}$, $\boldsymbol{C}_{dd}$ is the covariance matrix of $\boldsymbol{d}$, α is the inflation coefficient in ES-MDA, and $\Sigma = 0$ means no measurement error. Then ES-MDA is multiple iterations of ensemble smoother method.

We compare two posterior distributions of hydraulic conductivity: invert global PC scores $\boldsymbol{y}$ and invert local PC scores $\boldsymbol{y}'$. In both scenarios, we estimate posterior distributions of first 30 PCs ($> 70\%$ variance) using the ensemble smoother and reconstruct spatial fields with all principal components. Posterior log hydraulic conductivity after the first iteration are in Figure 18. Posterior uncertainties near pumping well (Row 2, Row 3 in Figure 18) have been reduced. We calculate the difference of posterior variances between both cases (Figure 18). Local PCA reduces more variance in the local area than global PCA and it reduces less variance in regions outside the local area than global PCA. Local PCA also provides a better match to the drawdown data.

Figure 19 shows that more iterations improve matching the drawdown data and reducing posterior variance. However, if we solely use global PCA, posterior data variance has not decreased significantly and final posterior drawdown curves match worse than using local PCA once. Using local inversion reduces the number of iterations in ES-MDA.

Back to the hierarchical formulation, we sample different $\boldsymbol{\theta}^*$ from $p(\boldsymbol{\theta}^*|\boldsymbol{d}_{obs})$ and perform local inversions individually with two iterations. Figure 20a shows posterior log hydraulic conductivity realizations with corresponding global variables. The mean of the full spatial field has been updated (Figure 20b). More uncertainty has been reduced near the pumping well. Posterior data variables well match the observed drawdown data (Figure 20c).

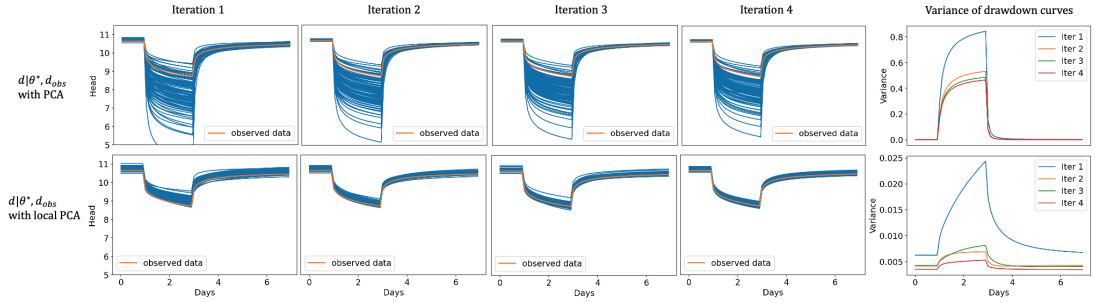

Figure 19. Drawdown curves after multiple iterations

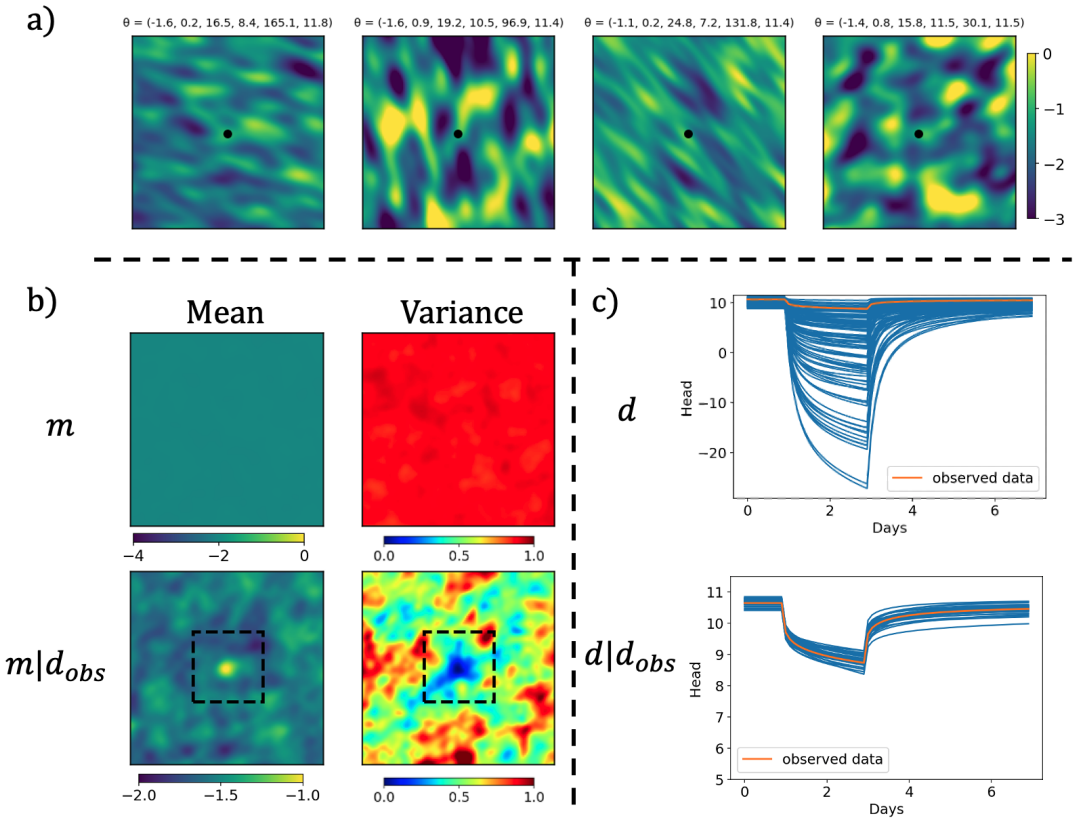

Figure 20. Posterior log hydraulic conductivity: realizations, corresponding global variables, mean, variance and corresponding data variables. a) realizations with corresponding global variables b) Row 1: prior mean and variance of log hydraulic conductivity. Row 2: posterior mean and variance of log hydraulic conductivity c) Prior and posterior data variables.

Note that posterior global variables $\boldsymbol{\theta}^*$ are continuous. We need to sample large amount of $\boldsymbol{\theta}^*$ and estimate posterior spatial variables $\boldsymbol{\theta}^*|\boldsymbol{d}_{obs}$) repeatedly. Therefore, having an efficient inversion method such as the proposed local PCA method on large-scale spatial fields is essential for a hierarchical formulation. We can easily parallelize the spatial inversion stage because posterior global variable samples $\boldsymbol{\theta}^*$ are independent. Another solution to speed up a hierarchical inversion is to select several representative sets of posterior global variables. We can select representatives using any unsupervised clustering methods: K-means (Hamerly & Elkan, 2003), Multi-dimensional Scaling (Mardia, 1978), t-distributed Stochastic Neighbor Embedding (Van der Maaten & Hinton, 2008), etc.

4 Conclusion

In this paper, we present a hierarchical Bayesian formulation to quantify uncertainty of global and spatial variables. In a hierarchical inversion, we estimate posterior global variables first, and then given global posterior, we estimate posterior spatial fields. We propose a machine learning-based inversion method to estimate posterior global variables with 1) both linear and non-linear forward models and 2) measurement errors. We propose a local dimension reduction method: local principal component analysis to efficiently invert large-scale spatial fields with local data. Local principal components are calculated from covariance matrix of local spatial field and save the number of iterations in non-linear inversions.

We present two synthetic case studies: one linear volume average inversion and one non-linear pumping test to illustrate the hierarchical Bayesian formulation. For posterior global variables, we compare our ML-based inversion method and MCMC sampling given the analytical likelihood function. Results from both methods are similar, except MCMC has a narrower estimation on the anisotropy variable because of the Gaussian assumption that leads to a very specific likelihood distribution. This assumption is not made in the machine learning approach. We illustrate why quantifying uncertainty of global variables is important with different scenarios. For posterior large-scale spatial fields, we compare two dimension reduction techniques: global PCA and local PCA. Results show a faster convergence using local PCA given locally collected drawdown data. In the future, we would like to extend our local spatial inversion method to discrete fields such as discrete models (e.g. lithologies, soil types).

References

Athens, N. D., & Caers, J. K. (2019). A monte carlo-based framework for assessing the value of information and development risk in geothermal exploration. *Applied Energy*, *256*, 113932.

Banerjee, S., Carlin, B. P., & Gelfand, A. E. (2003). *Hierarchical modeling and analysis for spatial data*. Chapman and Hall/CRC.

Betancourt, M. (2017). A conceptual introduction to hamiltonian monte carlo. *arXiv preprint arXiv:1701.02434*.

Beven, K., & Freer, J. (2001). Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the glue methodology. *Journal of hydrology*, *249*(1-4), 11–29.

Bishop, C. M., et al. (1995). *Neural networks for pattern recognition*. Oxford university press.

Botev, Z. I., Grotowski, J. F., & Kroese, D. P. (2010). Kernel density estimation via diffusion. *The annals of Statistics*, *38*(5), 2916–2957.

Chen, Y., & Oliver, D. S. (2010). Cross-covariances and localization for enkf in multiphase flow data assimilation. *Computational Geosciences*, *14*(4), 579–601.

Emerick, A. A. (2016). Towards a hierarchical parametrization to address prior

uncertainty in ensemble-based data assimilation. *Computational Geosciences*, *20*(1), 35–47.

Emerick, A. A., & Reynolds, A. C. (2013). Ensemble smoother with multiple data assimilation. *Computers & Geosciences*, *55*, 3–15.

Evensen, G. (2003). The ensemble kalman filter: Theoretical formulation and practical implementation. *Ocean dynamics*, *53*(4), 343–367.

Fenwick, D., Scheidt, C., & Caers, J. (2014). Quantifying asymmetric parameter interactions in sensitivity analysis: application to reservoir modeling. *Mathematical Geosciences*, *46*(4), 493–511.

Fichtner, A., & Simutė, S. (2018). Hamiltonian monte carlo inversion of seismic sources in complex media. *Journal of Geophysical Research: Solid Earth*, *123*(4), 2984–2999.

Fouedjio, F., Scheidt, C., Yang, L., Wang, Y., & Caers, J. (2021). Conditional simulation of categorical spatial variables using gibbs sampling of a truncated multivariate normal distribution subject to linear inequality constraints. *Stochastic Environmental Research and Risk Assessment*, *35*(2), 457–480.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. Chapman and Hall/CRC.

Grana, D. (2020). Bayesian petroelastic inversion with multiple prior models. *Geophysics*, *85*(5), M57–M71.

Hamerly, G., & Elkan, C. (2003). Learning the k in k-means. *Advances in neural information processing systems*, *16*, 281–288.

Hermans, T., Oware, E., & Caers, J. (2016). Direct prediction of spatially and temporally varying physical properties from time-lapse electrical resistance data. *Water Resources Research*, *52*(9), 7262–7283.

Hoffman, M. D., Gelman, A., et al. (2014). The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, *15*(1), 1593–1623.

Kitanidis, P. K. (1995). Quasi-linear geostatistical theory for inversing. *Water resources research*, *31*(10), 2411–2419.

Kitanidis, P. K., & Lee, J. (2014). Principal component geostatistical approach for large-dimensional inverse problems. *Water resources research*, *50*(7), 5428–5443.

Kuczera, G., & Parent, E. (1998). Monte carlo assessment of parameter uncertainty in conceptual catchment models: the metropolis algorithm. *Journal of hydrology*, *211*(1-4), 69–85.

Laloy, E., & Vrugt, J. A. (2012). High-dimensional posterior exploration of hydrologic models using multiple-try dream (zs) and high-performance computing. *Water Resources Research*, *48*(1).

Lee, J., & Kitanidis, P. K. (2014). Large-scale hydraulic tomography and joint inversion of head and tracer data using the principal component geostatistical approach (pcga). *Water Resources Research*, *50*(7), 5410–5427.

Malinverno, A., & Briggs, V. A. (2004). Expanded uncertainty quantification in inverse problems: Hierarchical bayes and empirical bayes. *Geophysics*, *69*(4), 1005–1016.

Mardia, K. V. (1978). Some properties of classical multi-dimensional scaling. *Communications in Statistics-Theory and Methods*, *7*(13), 1233–1241.

Oliver, D. S., & Chen, Y. (2011). Recent progress on reservoir history matching: a review. *Computational Geosciences*, *15*(1), 185–221.

Oliver, D. S., Cunha, L. B., & Reynolds, A. C. (1997). Markov chain monte carlo methods for conditioning a permeability field to pressure data. *Mathematical geology*, *29*(1), 61–91.

Park, H., Scheidt, C., Fenwick, D., Boucher, A., & Caers, J. (2013). History matching and uncertainty quantification of facies models with multiple geological interpretations. *Computational Geosciences*, *17*(4), 609–621.

Park, J., & Caers, J. (2020). Direct forecasting of global and spatial model parameters from dynamic data. *Computers & Geosciences*, *143*, 104567.

Park, J., Yang, G., Satija, A., Scheidt, C., & Caers, J. (2016). Dgsa: A matlab toolbox for distance-based generalized sensitivity analysis of geoscientific computer experiments. *Computers & geosciences*, *97*, 15–29.

Saley, A. D., Jardani, A., Ahmed, A. S., Antoine, R., & Dupont, J.-P. (2016). Hamiltonian monte carlo algorithm for the characterization of hydraulic conductivity from the heat tracing data. *Advances in Water Resources*, *97*, 120–129.

Satija, A., & Caers, J. (2015). Direct forecasting of subsurface flow response from non-linear dynamic data by linear least-squares in canonical functional principal component space. *Advances in Water Resources*, *77*, 69–81.

Satija, A., Scheidt, C., Li, L., & Caers, J. (2017). Direct forecasting of reservoir performance using production data without history matching. *Computational Geosciences*, *21*(2), 315–333.

Scheidt, C., Jeong, C., Mukerji, T., & Caers, J. (2015). Probabilistic falsification of prior geologic uncertainty with seismic amplitude data: Application to a turbidite reservoir case. *Geophysics*, *80*(5), M89–M12.

Scheidt, C., Li, L., & Caers, J. (2018). *Quantifying uncertainty in subsurface systems* (Vol. 236). John Wiley & Sons.

Scott, D. W. (2015). *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons.

Silverman, B. W. (2018). *Density estimation for statistics and data analysis*. Routledge.

Sun, W., & Durlofsky, L. J. (2017). A new data-space inversion procedure for efficient uncertainty quantification in subsurface flow problems. *Mathematical Geosciences*, *49*(6), 679–715.

Tarantola, A. (2005). *Inverse problem theory and methods for model parameter estimation*. SIAM.

Tipping, M. E., & Bishop, C. M. (1999). Mixtures of probabilistic principal component analyzers. *Neural computation*, *11*(2), 443–482.

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, *9*(11).

Van Kerm, P. (2003). Adaptive kernel density estimation. *The Stata Journal*, *3*(2), 148–156.

Van Leeuwen, P. J., & Evensen, G. (1996). Data assimilation and inverse methods in terms of a probabilistic formulation. *Monthly weather review*, *124*(12), 2898–2913.

Vrugt, J. A. (2016). Markov chain monte carlo simulation using the dream software package: Theory, concepts, and matlab implementation. *Environmental Modelling & Software*, *75*, 273–316.

Vrugt, J. A., & Ter Braak, C. J. (2011). Dream (d): an adaptive markov chain monte carlo simulation algorithm to solve discrete, noncontinuous, and combinatorial posterior parameter estimation problems. *Hydrology and Earth System Sciences*, *15*(12), 3701–3713.

Vrugt, J. A., Ter Braak, C. J., Clark, M. P., Hyman, J. M., & Robinson, B. A. (2008). Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with markov chain monte carlo simulation. *Water Resources Research*, *44*(12).

Xiao, S., Xu, T., Reuschen, S., Nowak, W., & Hendricks Franssen, 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 Resources Research*, *57*(9), e2021WR030313.

Yin, Z., Strebelle, S., & Caers, J. (2020). Automated monte carlo-based quantifica-

tion and updating of geological uncertainty with borehole data (autobel v1. 0). *Geoscientific Model Development*, *13*(2), 651–672.

Zhao, Y., & Luo, J. (2021). Bayesian inverse modeling of large-scale spatial fields on iteratively corrected principal components. *Advances in Water Resources*, *151*, 103913. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0309170821000683` doi: https://doi.org/10.1016/j.advwatres.2021.103913