Frequency-domain electromagnetic induction for the prediction of electrical conductivity and magnetic susceptibility using geostatistical inversion and randomized tensor decomposition

Liu, M.\textsuperscript{a,b}, Narciso, J.\textsuperscript{c}\textsuperscript{*}, Grana, D.\textsuperscript{a}, Van De Vijver, E.\textsuperscript{d}, Azevedo, L.\textsuperscript{c}

\textsuperscript{a} former: University of Wyoming, Department of Geology and Geophysics, Laramie, Wyoming, USA

\textsuperscript{b} current: Department of Energy Resources Engineering, Stanford University, Stanford, CA, USA

\textsuperscript{c} CERENA, DER, Instituto Superior Técnico, Universidade de Lisboa, Lisboa, Portugal

\textsuperscript{d} Department of Environment, Ghent University, Gent, Belgium

\textsuperscript{*} corresponding author

Contact Email: joao.narciso@tecnico.ulisboa.pt

(28/07/2023)
ABSTRACT

High-resolution characterization of near-surface systems is crucial for a variety of subsurface applications. Frequency-domain electromagnetic induction (FDEM) has been widely used in near-surface characterization when compared with other geophysical methods due to its flexibility in acquisition and the ability to survey large areas with high-resolution but with relatively low costs. FDEM measurements are sensitive to subsurface electrical conductivity (EC) and magnetic susceptibility (MS). However, the prediction of these properties requires solving a geophysical inverse problem. We combine ensemble smoother with multiple data assimilation (ES-MDA) and model re-parameterization via randomized tensor decomposition (RTD) to simultaneously predict electrical conductivity and magnetic susceptibility from measured FDEM data. ES-MDA is an iterative data assimilation method, which can be applied to nonlinear forward operators and provides multiple posterior realizations conditioned on the geophysical measurements to evaluate the model uncertainty. However, its application is usually computationally prohibitive for large-scale three-dimensional problems. To overcome this limitation, we reduce the model parameters using RTD and then perform the inversion in the low-dimensional model space. The method is applied to synthetic and noisy real data sets. In the synthetic application example, the predicted posterior realizations illustrate the ability of the proposed method to recover the true models of EC and MS accurately. The real case application comprises FDEM data acquired over an arable land characterized by quaternary siliciclastic deposits with geoarchaeological features. We assess the performance of the inversion method at a borehole location not used to constrain the inversion. The inverted models do capture the available log data, illustrating the applicability of the inversion method to noisy real data.
INTRODUCTION

Detailed modelling and characterization of near-surface is key to several applications, such as sustainable development of soil studies, archaeology, and groundwater management (De Smedt et al., 2013, Delefortrie et al., 2014, Simon et al., 2015). This is a challenging task as the near-surface is often characterized by strongly heterogeneous geological properties as the result of complex interacting processes of both natural and anthropogenic origins, which act at different spatiotemporal scales (Morel and Heinrich, 2008).

Due to the complex nature and dynamics of these systems, its characterization using traditional interpolation methods of sparse and discrete direct observations (e.g., borehole data) is not suitable to capture the full spatial variability of the system. Recently, there has been an increased interest in using geophysical data to characterize the near-surface. This is mainly due to the ability to acquire high spatial resolution data over large areas at relatively low-cost, the usability of the existing equipment in different types of terrain, and the ability to image subsurface properties (Everett, 2013) that can be interpreted in terms of geological and physical processes.

Among the most common geophysical methods, frequency-domain electromagnetic induction (FDEM) allows collecting high-resolution data sets timely and efficiently (Hanssens et al., 2020), by providing indirect measurements of two key near-surface properties: electrical conductivity (EC); and magnetic susceptibility (MS). From a simplistic perspective, EC relates mainly to soil salinity, texture, organic matter, moisture content, and bulk density (Doolittle and Brevik, 2014, Everett, 2013, Islam et al., 2014a, Islam et al., 2014b, Reynolds, 2011), while MS tends to be related to the mineralogy of the near-surface rocks, and anthropogenic features (Van De Vijver, 2017). Nevertheless, all these geological properties affect jointly EC and MS.
However, predicting the spatial distribution of EC and MS from the observed FDEM data requires solving a non-linear, ill-conditioned inverse problem with multiple solutions due to measurement errors and uncertainties in the model and observations (Tarantola, 2005), the band-limited nature and resolution of the FDEM data, noise and physical assumptions associated with the forward operators (Qiu et al., 2020). The recorded electromagnetic fields, the in-phase (IP) and quadrature-phase (QP) signal components, are related to EC and MS through a forward operator $F$ that can be mathematically described as

$$
\mathbf{d}_{\text{obs}} = F(\mathbf{m}) + \mathbf{e},
$$

where $\mathbf{d}_{\text{obs}} \in \mathbb{R}^{N_d}$ represents the measured data of dimension $N_d$, $\mathbf{m} \in \mathbb{R}^{N_m}$ represents the EC and MS model of dimension $N_m$, and $\mathbf{e}$ represents the error term associated with the observations errors within the data and assumptions about the physical models used to explain the natural systems under investigation (Tarantola, 2005). The operator $F$ is commonly approximated using 1-D or 2-D numerical models, due to prohibitive computational costs of three-dimensional forward models (Li et al., 2019).

Deterministic algorithms as well as stochastic sampling and optimization methods have been proposed to solve geophysical inverse problems (Tarantola, 2005). Among stochastic approaches the most commonly used methods are Markov chain Monte Carlo (McMC) and ensemble-based methods. FDEM inversion methods for near-surface characterization are generally based on deterministic approaches. These methods have been used successfully applied to model the spatial distribution EC and MS in the near-surface (Deidda et al., 2017, Farquharson et al., 2003, Guillemoteau et al., 2016). However, deterministic inversion methods predict a single best-fit model and have limited capabilities for uncertainty assessment. Due to the non-uniqueness of the
solution of the inversion problem, stochastic inversion methods are generally preferable. In the stochastic approach, the solution can be expressed as an ensemble of models that fit the data within a tolerance and whose variability represents the uncertainty of the solution, which can be used to make informed decisions and quantify risks. Stochastic geophysical inverse methods under a Gaussian assumption comprise the so-called pilot points method that use sequential geostatistical resampling techniques (e.g., Mariethoz et al., 2010; Alcolea et al., 2010; Fu and Gomez-Hernandez, 2009, Hansen et al., 2012, Zahner et al., 2016, Jäggli et al., 2017;), principal component geostatistical approach (Lee and Kitanidis, 2014), methods based on circulant embedding of the covariance matrix (e.g., Hansen et al., 2012, Laloy et al., 2015); and methods that allow for jointly inferring the spatial correlation model (i.e., mean and variogram) together with the two- and three-dimensional spatial distribution of the property field values of interest (Laloy et al., 2015, Hunziker et al., 2017, Wang et al., 2022).

The available literature includes statistical approaches to FDEM inversion, but these are limited to the prediction of EC (Moghadas and Vrugt, 2019) or rely on Gaussian assumptions for the distribution of EC and MS. Trans-dimensional Bayesian inversion of electromagnetic data and Markov chain Monte Carlo methods have been proposed in (Blatter et al., 2018, Minsley, 2011, Ray and Key, 2012). These approaches generally allow an accurate quantification of the posterior distribution; however, the computational cost of the sampling and optimization is generally unfeasible for large geophysical datasets. Ensemble-based methods, such as ensemble smoother and ensemble Kalman filter (Evensen, 2009), provide a reliable alternative to McMC methods, by finding a compromise between model accuracy and computational cost. For example, the Kalman ensemble generator (KEG) method (Bobe et al., 2019) provides such statistical framework for FDEM inversion. A comparison between the KEG and an iterative geostatistical FDEM inversion
method is presented in Narciso et al. (2022). Most publications on the application of machine learning in geophysical inverse problem adopt deep learning algorithms to approximate the forward model and reduce the problem dimension and the computational cost (e.g., Manassero et al., 2020, Puzyrev and Swisinsky, 2021, Qi et al., 2019) or use them directly to approximate the inverse function and replace deterministic inversion methods (Hashemian et al., 2021, Li et al., 2021). For example, (Manassero et al., 2020) propose a reduced-order approach for the inversion of electromagnetic data.

The techniques to reduce the computational time in high-dimensional probabilistic inverse problems, can be generally divided in three categories: (1) approximation of the forward operator (i.e., surrogate modeling), (2) dimensionality reduction of the model and/or data spaces by re-parameterization, and (3) approximating the posterior distribution by making certain assumptions about its probability distribution. In the method proposed herein we explore points (2) and (3) by combining stochastic inversion with dimensionality reduction techniques to perform the inversion in a lower dimensional space. We propose a stochastic nonlinear method based on the ensemble smoother with multiple data assimilation (ES-MDA) (Emerick and Reynolds, 2013) to invert the FDEM data for EC and MS. ES-MDA is a derivative-free optimization method that proves useful when the code of forward simulators is inaccessible, or the sensitivity matrix is challenging to derive. Unlike linear Bayesian methods, ES-MDA does not require a linear approximation of the forward operator, making it advantageous in improving inversion results for non-linear cases. However, ES-MDA poses prohibitive computational costs in large geological models. To address this challenge, we propose using randomized tensor decomposition (RTD) to sparsely re-parameterize the models and update the model parameters in the low-dimensional space. RTD is a high-order linear reduction method that can recover spatial structures between multiple
dimensions of geological models and track uncertainty propagation during model order reduction. Compared with deep-learning-based methods, RTD is easier to integrate into the inversion workflow and is not limited by computing devices.

This method is first applied in a synthetic two-dimensional data set to validate the results obtained and then in a three-dimensional real case application to assess its performance in data contaminated with noise. The next sections describe in detail the modelling steps of the proposed methodology and the results of its application to the synthetic and real data sets.

**METHODOLOGY**

**Forward Response and Sensitivity Modeling**

The FDEM data comprise both the in-phase (IP) and quadrature-phase (QP) components of the electromagnetic field. These data are generally acquired by a loop-loop system characterized by one transmitter coil and one or multiple receiver coils. To link the unknown near-surface properties (i.e., EC and MS) to the measured data, we use a one-dimensional nonlinear approximation of the propagated electromagnetic field (Hanssens et al., 2019). This forward operator calculates the IP and QP responses per transmitter-receiver coil offset located above a model with \( n \) layers, using a superposition of Bessel functions of order 0 and 1, determined by Hankel transform. This transform is numerically calculated by means of a Guptasarma and Singh digital filter (Guptasarma and Singh, 1997). Here, we use the reflection coefficient approximation due to its computational efficiency (Ward and Hohmann, 1987). For low-frequency applications, a quasi-static approximation can be applied, so that dielectric permittivity is negligible, i.e., the acquired signal is mostly dependent on EC and MS. In addition to the IP and QP responses, the
forward operator also calculates the depth of investigation given the characteristics of the acquisition equipment and the model sensitivity (Guptasarma and Singh, 1997).

In this work, geostatistical simulations are used to generate high-resolution subsurface models in 3-D, then a forward geophysical model based on a 1-D approximation is applied to compute the IP and QP predictions, and the simulated models are then updated in 3-D using the ES-MDA. The simulation and update of the models is done in 3-D. To reduce the computational cost, a dimensionality reduction approach is proposed to perform the updating in a lower dimensional space and increase the computational efficiency (Liu et al., 2022a). The use of the 1-D forward model is one of the main limitations of this approach as it is unable to capture the propagation of the EM field in the three directions of space; however, replacing the 1-D approximation would dramatically increase the computational cost of the proposed method. We believe that by applying the ES-MDA with RTD in the real case application and comparing the results at the blind borehole location is enough to assess the robustness of the proposed methodology.

Inverse Method

For the inversion, we adopt numerical approximate methods for the solution of the associated inverse problem. We apply a stochastic method, namely the ES-MDA (Emerick and Reynolds, 2013, Grana et al., 2021), for its computing efficiency and the ability to quantify uncertainty. ES-MDA is derived from Kalman Filter (Evensen, 2009) to overcome the limitation of the operator linearization in non-linear inverse problems and to improve the computational efficiency in large-scale optimization and inverse problems. Like the standard Kalman Filter (KF), the ES-MDA is based on a Bayesian updating approach and the estimation of model parameters from
measurements includes two steps: prediction by the forward model from the prior realizations, and
correction by the measurement according to the likelihood function. In the ES-MDA, the Kalman
gain is empirically estimated from the ensemble of prior models. The ES-MDA updating equation
of model parameters \( m \) of ES-MDA can be written as

\[
\mathbf{m}_u^j = \mathbf{m}_p^j + K (\mathbf{d}_j - \mathbf{d}_p^j),
\]

for \( j = 1, \ldots, N_e \) with \( N_e \) being the ensemble size, where \( \mathbf{m}_p^j \) represents the prior model parameters,
\( \mathbf{m}_u^j \) represents the updated model parameters obtained by assimilating the measurements, \( \mathbf{d}_p^j \) is the
predicted data obtained from \( \mathbf{m}_p^j \) through the forward operator \( F \), \( \hat{\mathbf{d}}_j \) is the observed data with
random perturbation according to the distribution of the noise \( \mathbf{e} \), and \( K \in \mathbb{R}^{N_m \times N_d} \) is the Kalman
gain matrix. In the ES-MDA, the Kalman gain matrix is empirically estimated from the prior
models as

\[
K = C_{md}^p (C_{dd}^p + C_{d})^{-1}
\]

\[
C_{md}^p = \frac{1}{N_e - 1} \sum_{j=1}^{N_e} (\mathbf{m}_p^j - \overline{\mathbf{m}}^p)(\mathbf{d}_p^j - \overline{\mathbf{d}}^p)^T
\]

\[
C_{dd} = \frac{1}{N_e - 1} \sum_{j=1}^{N_e} (\mathbf{d}_p^j - \overline{\mathbf{d}}^p)(\mathbf{d}_p^j - \overline{\mathbf{d}}^p)^T
\]

where \( C_{md}^p \in \mathbb{R}^{N_m \times N_d} \) represents the cross-covariance matrix between the prior model
parameters \( \mathbf{m}^p \) and the corresponding predicted data \( \mathbf{d}^p \), \( C_{dd} \in \mathbb{R}^{N_d \times N_d} \) is the covariance of the
predicted data $\mathbf{d}^p$, $\mathbf{C}_d \in \mathbb{R}^{N_d \times N_d}$ is the covariance matrix of the measurement error $\mathbf{e}$, $\mathbf{\bar{m}}^p$ and $\mathbf{\bar{d}}^p$ are the empirical mean of the ensemble of model variables and predicted data, respectively. In the linear case, where the forward operator $\mathbf{F}$ can be expressed as matrix $\mathbf{G}$, the covariance matrices $\mathbf{C}^p_{\text{md}}$ and $\mathbf{C}^p_{\text{dd}}$ are corresponding to $\mathbf{C}_m \mathbf{G}^T$ and $\mathbf{G} \mathbf{C}_m \mathbf{G}^T$, respectively, where $\mathbf{C}_m$ is the covariance of model parameters. Equations 2 and 3 show that the Kalman Gain matrix controls the trade-off between the prior predictions and updated correction driven by measurements according to their uncertainties. The weights of measurements are large if the measurement errors are small and vice versa. For nonlinear inverse problems, it is necessary to iteratively update the model variables to achieve a satisfactory match between prediction and measurements. One common strategy is to sequentially assimilate observations at each time step (e.g., ensemble Kalman Filter), but this procedure requires to perform forward simulations every time step and thus it is computationally inefficient. Alternatively, in ES, all data available are simultaneously used for model updating. To guarantee the convergence between the model predictions and measurements in nonlinear cases, the simultaneous data assimilation is performed multiple times. This method is referred to as ES-MDA (Emerick and Reynold, 2013).

ES-MDA is an iterative method. An ensemble of prior models is first sampled from a prior distribution and iteratively updated until the models are consistent with the measured data. Each data assimilation step can be interpreted as a Bayesian updating process, where the models updated in the previous iteration are used as the prior at the current step and then corrected by assimilating the observations. The algorithm of ES-MDA can be summarized as follows:

1. Define the ensemble size $N_e$, the number of data assimilations $N$ and the inflation coefficients $\{\alpha_k\}_{k=1,\ldots,N}$ with the constraint $\sum_{k=1}^{N} \alpha_k^{-1} = 1$. 
2. Generate an ensemble of $N_e$ prior realizations $\{m_j\}_{j=1,...,N_e}$ of the EC and MC models conditioned on the available borehole data using, for example, geostatistical simulation algorithms.

3. For $k = 1$ to $N$
   - Compute the geophysical response of each prior realization $\{d^p_j\}_{j=1,...,N_e}$ using the forward modeling.
   - Perturb the observations $\{d_j\}_{j=1,...,N_e}$ for each ensemble member as
     \[
     \tilde{d}_j = d_{obs} + \sqrt{\alpha_i C_{d}^{1/2}} \epsilon_j
     \]
     where $\epsilon_j \sim \mathcal{N}(0, I_{N_d})$.
   - Update model ensemble $\{m_j\}_{j=1,...,N_e}$ using Equations 2-5.

End

The solution of the inverse problem is a linear combination of the updated ensemble models. The ensemble models, at each iteration, are updated according to the residuals between predicted and observed data and the cross-covariance matrix of the residuals and model variables. The initial ensemble must be large enough to represent the prior variability. If the variability of the prior is too small, the uncertainty could be severely underestimated. The number of iterations is established through a trial-and-error approach. Publications on data assimilation in dynamic reservoir modeling show that a number of iterations between 4 and 8 is generally sufficient (Emerick and Reynolds, 2013). In geophysical inverse problems, the large amount of data makes the problem less underdetermined than fluid flow modeling problems, hence 4 iterations are generally sufficient (Grana et al., 2021). However, due to the large number of measurements, a large ensemble is
necessary to avoid uncertainty underestimation or ensemble collapse. The prior model includes the prior distribution of the model variables and the spatial correlation model of the realizations. For datasets with large errors, the prior distribution has a strong impact on the posterior realizations, especially the spatial correlation model. In these cases, alternative methods that predict jointly the model parameters and the spatial correlation model can be used (Laloy et al., 2015, Hunziker et al., 2017, Wang et al., 2022). The vertical correlation can be estimated from well log data, whereas the lateral correlation must be assumed based on prior geological information. For simplicity, the data errors are assumed to be spatially uncorrelated with diagonal covariance matrix; however, if geophysical data are pre-processed for quality control and denoising, the error model could be correlated, and the covariance matrix of the data is banded. The assumption of the banded covariance matrix is generally challenging in practical applications. Large variances of the errors tend to make the prior dominant on the data-driven likelihood function and might lead to a poor data match, whereas small variances of the errors tend to make the likelihood function predominant and might lead to unphysical values of the model variables.

**Model re-parameterization**

Due to the large dimension of the model grids in real applications, the ES-MDA method is often computationally and memory prohibitive. Therefore, we propose to reduce the model parameters using the RTD method and then perform the data assimilation in the reduced model space. After each data assimilation, the reduced model parameters can be back transformed to the full model space using the factor matrices obtained by RTD.

A tensor is a multi-index numerical array, which can be used to represent high-dimensional
data. Conventional multivariate data analysis approaches based on standard flat-view matrix models requires reshaping the data tensor into a matrix or vector and applying classical matrix factorization methods, such as singular value decomposition (SVD) non-negative matrix factorization (NMF), or independent component analysis (ICA) (Cichocki et al., 2015). These methods can be efficiently implemented but they might struggle to capture spatial correlations in multiple dimensions, which limits their performance in high-dimensional data analysis. Tensor decomposition methods are based on multilinear algebra and can exploit the intrinsic multi-dimensional patterns in the model space, as the RTD used in this work. In recent years, many deep-learning-based reduction methods have been proposed to overcome the limitation in geoscience problems (e.g., Laloy et al., 2017, Laloy et al., 2018; Canchumuni et al., 2019; Liu and Grana, 2020; Lopez-Alvis et al., 2021; Mo et al., 2019). However, those methods based on deep neural networks usually require thousands of prior models for training and are relatively difficult to integrate with ES-MDA.

The canonical polyadic (CP) decomposition and the Tucker decomposition are the two most popular tensor decomposition algorithms (Rabanser et al., 2017). The CP decomposition represents a tensor as a linear combination of vectors, whereas the Tucker decomposition decomposes a tensor into a small dense core tensor and a set of factor matrices. We adopt the Tucker decomposition, because it is more suitable for dimensionality reduction in which the core tensor can be regarded as the sparse features extracted from the original tensor data and the factor matrices can be used for back-transformation.

The Tucker decomposition of an \( N \)th order tensor \( \mathbf{X} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_N} \) can be expressed as

\[
\mathbf{X} = \mathbf{G} \times_1 \mathbf{B}^{(1)} \times_2 \mathbf{B}^{(2)} \ldots \times_N \mathbf{B}^{(N)}
\] (7)
where the symbol $\times_n$ represents the tensor-matrix multiplication along mode-$n$, $\mathcal{G}$ is the nondiagonal core tensor that includes the information for the extension of the tensor components, and $\{B^{(n)}\}_{n=1,\ldots,N}$ are factor matrices that represent the principal components in the respective tensor modes. A graphical view of the Tucker decomposition is presented in Figure 1. Tensor decomposition is also non-unique. A decomposition where core tensor and all factor matrices are orthonormal is referred to the Higher-Order Singular Value Decomposition (HOSVD) (De Lathauwer et al., 2000a, De Lathauwer et al., 2000b).

In practice, the 3-dimensional geological models might consist of millions of grid cells. Due to the limited memory and high computational complexity, the conventional tensor decomposition methods are usually not applicable. Randomized algorithms are then used for large-scale tensors. The randomized approach aims to find the low-rank approximation of the unfolding matrices of large-scale tensors via the probabilistic strategy (i.e., random sketching) and then perform matrix factorization on the small matrices. One popular method is the random projection in which a large-scale matrix is reduced using the transformation of a random matrix with given probability distribution. The detailed steps of the randomized HOSVD of an $N$th order tensor data $X \in \mathbb{R}^{l_1 \times l_2 \times \ldots \times l_N}$ are summarized as follows:

1. Define the number of iterations $N_{it}$ and a multilinear rank $(R_1, R_2, \ldots, R_N)$.

2. Initialize the factor matrices $\{B^{(n)} \in \mathbb{R}^{l_n \times R_n}\}_{n=1,\ldots,N}$ as random Gaussian matrices.

3. For $i = 1$ to $N_{it}$:

   For $n = 1$ to $N$:
- $\mathcal{G} = \mathcal{X} \times_{p \neq n} \{ B^{(n)^T} \}$.
- Generate a random matrix $\Omega^{(n)} \in \mathbb{R}^{\prod_{p \neq n} R_p} \times \mathbb{R}^{R_n}$ drawn from Gaussian distribution.
- Compute $W^{(n)} = \mathcal{G}^{(n)} \Omega^{(n)}$ where $\mathcal{G}^{(n)} \in \mathbb{R}^{I_n} \times \mathbb{R}^{\prod_{p \neq n} I_p}$ is the $n$-unfolding matrix of tensor $\mathcal{G}$.
- Compute the orthonormal basis $Q^{(n)} \in \mathbb{R}^{I_n} \times \mathbb{R}^{R_n}$ of $W^{(n)}$ by QR decomposition.

**End**

- Compute the core tensor $\mathcal{G} = \mathcal{G} \times_1 Q^{(1)^T} \times_2 Q^{(2)^T} \cdots \times_N Q^{(N)^T}$

**End**

Herein, we propose the RTD algorithm as a dimensionality reduction method to reduce the dimension of the model and update the variables in a low dimensional model space. This approach allows improving the computational efficiency of the inversion. In this work, we do not differentiate between the uncertainty in the RTD transformation and the uncertainty in the inversion. The larger is the number of ensemble members, the smaller is the underestimation of the uncertainty. Similarly, the larger is the reduction of the model space, the larger is the overestimation of the uncertainty. By adopting a trial-and-error approach, we determine, case by case, the optimal dimension of the model space and of the model ensemble. A detailed analysis of the uncertainty quantification in geophysical inverse problems with model and data reduction is presented in (Grana et al., 2019).
The proposed FDEM inversion method includes the integration of the forward model in section *Forward Response and Sensitivity Modeling*, the inverse method in section *Inverse Method*, and the model reparameterization in section *Model re-parameterization*. The so-obtained method predicts a set of model realizations that represent the posterior distribution of the inverse solution. Figure 2 illustrates the inversion workflow of ES-MDA with RTD. It starts with a set of prior realizations of EC and MS \( \{ m^k_j = 0 \}_{1}^{N_e} \) simulated by geostatistics algorithms. Then, their EM responses \( \{ d^k_j \}_{1}^{N_e} \) are predicted by the forward model and the reduced model parameters \( \{ z^k_j \}_{1}^{N_e} \) are obtained by the RTD. The reduced model parameters are then updated \( \{ z^{k+1}_j \}_{1}^{N_e} \) by assimilating the observations with ES-MDA. It is an iterative procedure in which the prior models in the next iteration \( \{ m_j^{k+1} \}_{1}^{N_e} \) are back transformed from \( \{ z_j^{k+1} \}_{1}^{N_e} \) by the inverse RTD. In the proposed FDEM inversion method both the ES-MDA and the model reduction with RTD affect the uncertainty assessment of the posterior solution. The performance of the ES-MDA depends on the on the number of models in the initial ensemble while the performance on the RTD depends on the dimensionality of the lower-dimensional space. Other stochastic inversion methods with robust uncertainty assessment, such as Markov chain Monte Carlo method (Blatter et al., 2018) could also be adopted.

**SYNTHETIC CASE APPLICATION**

We first apply the proposed inversion method to a two-dimensional synthetic data set (Narciso et al., 2020, 2022) representing a vertical section of the near-surface based on real data collected at a mine tailing in Portugal (Panasqueira). This dataset comprises laboratory measurements of
porosity and particle density obtained from samples collected from two main rock types of the site: fine-shaly sands (which constitutes the predominant rock type), and quartz-schist gravels. First, we generate stochastic realizations of porosity using direct sequential simulation (Soares, 2001) based on a variogram model that represents the expected spatial correlation of porosity. The spatial correlation in the fine shaly sands is larger than in the quartz-schist gravels. Then, we generate stochastic realizations of particle density and water saturation using geostatistical simulations based on the porosity model. The true EC model is then generated from porosity $\phi$ and water saturation $S_w$ using Archie’s equation (Archie, 1942):

$$R_t = a \phi^{-k} R_w S_w^{-n}, \quad (8)$$

where $R_t$ is the electrical resistivity (i.e., the inverse of EC), $a$ is the tortuosity constant, assumed as 0.88, $\phi$ is the porosity, $k$ is the cementation exponent (generally between 1.3 and 2.5 for most sedimentary rocks), $R_w$ the electrical resistivity of the pore fluid (assumed to be constant), $S_w$ is the water saturation and $n$ is the saturation exponent (generally equal to 2) (Archie, 1942). The true MS model is generated using geostatistical simulations (Soares, 2001) based on a variogram model that describes the expected spatial distribution pattern of MS. From the resulting EC and MS models (Figures 3a and 3b), four pseudo-boreholes were extracted equally spaced along the vertical section.

The prior ensembles of EC and MS include 500 geostatistical realizations generated using direct sequential simulation (Soares, 2001). This set of models represents the model parameter space and the histograms of both properties as retrieved from the borehole data. Therefore, this geostatistical simulation algorithm does not assume any parametric distribution for the property of interest. The EC and MS data extracted at the borehole locations are used as conditioning data for
the geostatistical simulation so that all model realization reproduce the borehole data at the borehole locations. Based on the spatial continuity model of the synthetic model, the prior ensemble of realizations is simulated by imposing omnidirectional horizontal exponential variograms for EC and MS with a range of 6 m and 8 m, respectively. The vertical direction is modelled with exponential variograms with a range of 4 m for EC and 6 for MS.

The true FDEM data are generated using the forward model (Hanssens et al., 2019) and considering a loop-loop system setup, characterized by one transmitter coil and multiple receiver coils with two spatial configurations and two offsets per coil configuration. We focus on the horizontal coplanar (HCP) configuration for the offset 1 and 2 m. The data are contaminated by Gaussian noise and the noise level is 10% of the observations.

The model grid includes 400×40 cell in the $i$- and $k$- directions, respectively. The pointwise mean models of the prior EC and MS ensembles (Figures 3c and 3d) reproduce the true EC and MS measurements at the borehole locations. Far from the location of the boreholes, and for distances larger than the variogram range, these models tend to the average value of the distribution. For this 2-D example, both EC and MS model are a third order tensor with a size of 400×40×500 (corresponding to the numbers of model grids in the $i$- and $k$- directions, and ensemble size, respectively). The tensors of EC and MS model are reduced to 40×5×500 by the RTD algorithm with four iterations before data assimilation. We then apply the ES-MDA in the reduced model space. The number of iterations of the ES-MDA is 4 with the inflation coefficients of 9.33, 7.0, 4.0 and 2.0, which are recommended by Emerick and Reynolds (2013). The posterior mean of EC and MS is shown in Figures 4a and 4b, and the posterior standard deviation is shown in Figures 4c and 4d. The absolute residuals between the predicted posterior mean and the true models...
are shown in Figures 4e and 4f. The results capture small- and large-scale features of the true EC and MS models up to the depth of investigation provided directly by the forward operator used in the inversion (Hanssens et al., 2019) which is estimated to be approximately 3 m. The inversion, by construction, reproduces exactly the measurements at the borehole locations. Hence, the posterior standard deviation is zero at the borehole location and it increases with the distance from the borehole locations. Despite the large variability in the values of the QP and IP responses predicted from the prior models for all coil offsets (Figures 5 and 6) the posterior model of QP and IP matches the observed data for most of the observations. At some locations the predicted models show mismatches with the observation. These results might be related simultaneously to the noise component within the data and the uncertainty originated due to the dimensionality reduction technique applied as part of the proposed method (i.e., RTD) (Grana et al., 2019; Liu et al., 2022).

**REAL CASE APPLICATION**

We then apply the proposed ES-MDA with RTD to the FDEM inversion to a real data set from a FDEM survey over an arable land with slight slope and 20 cm of rendzina soil cover, located near Knowlton (Dorset, UK), and containing several archaeological features. The region of interest is characterized by Cretaceous chalk in the shallow subsurface, and calcareous ooze, overlain by Quaternary siliciclastic sand deposits. The Cretaceous formation is characterized by a background susceptibility of zero and a low EC ($\approx 7$ m/Sm), while the sand deposits is strongly magnetic (MS $\approx 1 \times 10^{-3}$) and slightly more conductive that the bedrock. In this area IP anomalies are related to the buried archeology and the background geology provides a large range of QP values (Delefortrie et al., 2018).
The FDEM data are collected using a DUALEM 21HS instrument, with an operating frequency of 9000 Hz in a loop-loop setup, elevated at 0.16 m from the surface pulled by a quadbike. The data acquisition was performed along parallel lines 1 m apart at a speed of ~8 km/h, and a sampling frequency of 8 Hz. For this study, we use the FDEM data collected from one transmitter paired with two coplanar receiver coils, in horizontal mode, at 1 and 2 m from the transmitter (HCP1 and HCP2, respectively). The measured IP and QP data are noisy and with systematic errors; therefore, we apply a calibration before the inversion using the existing EC and MS borehole measurements. We applied a drift correction consisting of tie-line levelling as described in (Delefortrie et al., 2018). However, we did not tackle the striping effect present in the PRP IP signal and the point anomalies observed in HCP QP signal (Delefortrie et al., 2018). These characteristics of the observed signal do affect the quality of the inversion results and the match between predicted and observed data.

EC and MS data are collected in twelve boreholes at intervals of 5-10 cm, reaching a maximum depth of 1.2 m and a minimum of 0.8 m. Eleven boreholes were used to compute horizontal and vertical experimental variograms based on a spherical model (omnidirectional in the horizontal direction) for both EC and MS properties, with horizontal range of 168 m for EC and 282 m for MS, and vertical range of 8 m for EC and 7 for MS.

The model grid includes $531 \times 171 \times 20$ cells in the $i$, $j$- and $k$- directions. A set of 500 geostatistical realizations of EC and MS is then generated conditioned on the borehole data and assuming variogram models fitted to experimental variograms computed from the borehole data. The mean prior models of EC and MS (Figures 7a and 7b) match the borehole measurements and tend to the mean of the direct measurements away from the borehole. For this 3-D case, both EC and MS model is a fourth order tensor with a size of $531 \times 171 \times 20 \times 500$ (corresponding to the
numbers of model grids in the $i$, $j$- and $k$- directions, and ensemble size, respectively). The tensors of EC and MS model are reduced to $20 \times 40 \times 5 \times 500$ by the RTD algorithm with four iterations before data assimilation. We then apply the ES-MDA inversion with 4 iterations and inflation coefficients of 9.33, 7.0, 4.0 and 2.0 (Emerick and Reynolds, 2013). The posterior mean models (Figures 8a and 8b) show a detailed spatial distribution pattern with a layer of continuous high conductivity and susceptibility at around 1 m depth. This value is consistent with the observed depth of the top chalk as interpreted from the existing borehole data (Delefortrie et al., 2018). The posterior standard deviation of EC and MS is shown in Figure 9. As the prior ensemble of EC and MS was constructed with geostatistical simulation, the predicted EC and MS values at the borehole locations are exactly reproduced.

In Figures 10 and 11, we show the comparison between predicted and measured IP and QP data. The predicted data match relatively well the observed data. The mismatch between the predicted and observed data might be due to the noisy nature of the data, as described above, and the one-dimensional approximations of the forward operator that cannot model complex and highly heterogenous lateral distributions of electrical properties. The parametrization of the RTD and the ES-MDA might be partly attributed to the misfit between predicted and observed data.

The computational cost of one updating step for ES-MDA is $O(N_e^2 N_m + N_e^2 N_d)$ where $N_e$, $N_m$ and $N_d$ are the ensemble size, the number of model parameters and observations, respectively (Nino Ruiz et al., 2015). In the real case, the ensemble size is 500; the number of observations is 26,640; the numbers of model parameters with and without reduction are $1,816,020 (171 \times 531 \times 20)$ and $4,000 (20 \times 40 \times 5)$, respectively. The speed-up ratio with model dimension reduction by RTD is roughly 60.14.
DISCUSSION

We proposed a stochastic FDEM inversion method in a reduced space leveraging the benefits of RTD with respect to dimensionality reduction. The method is illustrated in two application examples: one synthetic and one real. First, we consider a 2-D synthetic data set to evaluate the accuracy of the predictions. Then, we apply the proposed method in a real 3-D data set to assess its performance under real noise conditions. In both application examples the data predictions do match the observed FDEM data (Figures 5, 6, 10 and 11). Besides, when observing the residuals between model predictions of EC and MS and the true model (Figure 4e and 4f) they do not exhibit any spatial continuity pattern that is consistent with the true EC and MS spatial continuity pattern. In both application examples the prior ensembles of EC and MS are constructed via geostatistical simulation. While alternative methods can be applied, we suggest this class of methods due to its ability to reproduce direct observations (i.e., borehole data, histograms and spatial continuity patterns as revealed by variogram models). For this reason, when relying on geostatistical simulation to build the prior ensemble, a critical aspect for the success of the proposed inversion method is the availability of borehole data and its spatial distribution within the area of interest. Spatial sampling, including the spatial distributions of conditioning data, has been extensively studied in mining engineering (Journel and Huijbergs, 1978). A limited number of boreholes might affect the accuracy of the inversion and lead to large uncertainties in the predictions. In real applications, the prior distributions and variogram models assumed in the generation of the prior ensembles should account for prior geological information available for the area under
investigation as well as direct measurements from nearby areas.

Figures 5, 6, 10 and 11 show that the P5-P95 interval of the predicted data do not encapsulates entirely the observed data. In other words, there is an underestimation of the predicted uncertainty. This fact might be originated by two complementary reasons: the reparameterization of the model parameter space with the RTD affects the uncertainty assessment (Grana et al., 2019 and Liu et al., 2022b); the ES-MDA has a better performance for non-Gaussian and non-linear inverse problems. Finally, to assess the performance of the inversion locally, we removed one borehole from the conditioning data set. Removing a larger number of conditioning boreholes would decrease the accuracy of the predictions as the estimation of the EC and MS distributions would be poor. The comparison between the predicted properties and the borehole measurements at the location of the borehole not used to constrain the inversion is shown in Figure 12. Despite the limited length of the measured EC and MS, the estimated posterior distribution matches the true EC and MS. Due to the relatively small number of samples the predictions in the deeper part of the model are less reliable.

We use ES-MDA due to its relatively simplicity of implementation and its potential to efficiently assess the posterior distribution in geophysical inversion problems. However, the computational cost of ES-MDA might be prohibitive for large-dimensional inverse problems such as FDEM inversion. For this reason, we combine RTD, a dimensionality reduction technique of the model parameter space, with ES-MDA. The application examples shown herein, show that the coupling of both methodologies is an efficient solution to alleviate the computational burden of ES-MDA without compromising the model predictions and the uncertainty assessment despite assumptions about the prior distributions of the model parameters.
CONCLUSION

We proposed a FDEM inversion method that combines ES-MDA with RTD to predict the spatial distribution of EC and MS. The initial prior ensemble of models is generated using geostatistical simulation, to model the complex and heterogeneous subsurface distributions. Then, RTD coupled with ES-MDA makes the inversion method computationally feasible and applicable to 3-dimensional grids with a large number of cells. This FDEM inversion method was validated on a two-dimensional synthetic data set and then applied to a 3-dimensional real dataset. In both application examples, the predicted models reproduce the measured EC and MS data while allowing assessing the uncertainty of the predictions. The proposed inversion relies on a one-dimensional forward approximation but could be extended to more complex physical models.

CONCLUSION DATA AND MATERIALS AVAILABILITY

The code and synthetic data are freely available on GitHub (https://github.com/theanswer003/ES-RTD-FDEM).

ACKNOWLEDGEMENTS

Leonardo Azevedo and João Narciso acknowledge the support of the CERENA (strategic project FCT-UIDB/04028/2020). João Narciso is funded by the Fundação para a Ciência e a Tecnologia (Portuguese
Foundation for Science and Technology) through the project SFRH/BD/139577/2018. Dario Grana acknowledges the Nielson Energy Fellowship and the School of Energy Resources of the University of Wyoming. We also acknowledge the three anonymous reviewers for their detailed comments and fruitful discussion.
REFERENCES


Van De Vijver, E., 2017, Proximal soil sensing in the context of urban (re)development: an evaluation of multi-receiver electromagnetic induction and stepped-frequency ground penetrating radar at landfills and industrial sites: Ghent, Belgium: Ghent University. Faculty of Bioscience Engineering.


FIGURES CAPTION

**Figure 1** – Tucker decomposition of a third-order tensor.

**Figure 2** – Workflow of ES-MDA with RTD.

**Figure 3** – True and prior mean of EC and MS of the synthetic case: (a) true EC model; (b) true MS model; (c) prior mean of EC; and (d) prior mean of MS. It is clear the influence of the borehole data in the prior mean of EC and MS.

**Figure 4** – Posterior mean, standard deviation (std.) and residual of EC and MS of the synthetic case: (a) posterior mean of EC models; (b) posterior mean of MS models; (c) posterior std. of EC; (d) posterior std. of MS; (e) absolute error between the posterior mean and the true EC; (f) absolute error between the posterior mean and the true MS.

**Figure 5** – Predicted IP data from the prior and posterior EC and MS models of the synthetic case. The red dots are the true measurements with noise; the black lines represent the true data without noise; the intervals in gray and light blue correspond to the region between the percentiles P5 and P95 of the prior and posterior prediction, respectively; the black and blue lines represent the prior mean and posterior mean, respectively.
**Figure 6** – Predicted QP data from the prior and posterior EC and MS models of the synthetic case. The red dots are the true measurements with noise; the black lines represent the true data without noise; the intervals in gray and light blue correspond to the region between the percentiles P5 and P95 of the prior and posterior prediction, respectively; the black and blue lines represent the prior mean and posterior mean, respectively.

**Figure 7** – Prior mean of EC (a) and MS (b) models of the real case. The black dashed lines represent the locations of the X-, Y- and depth slices, and the black dots represent the well locations.

**Figure 8** – Posterior mean of EC (a) and MS (b) models of the real case. The black dashed lines represent the locations of the X-, Y- and depth slices.

**Figure 9** – Posterior standard deviation of EC (a) and MS (b) models of the real case. The black dashed lines represent the locations of the X-, Y- and depth slices.

**Figure 10** – Predicted IP data from the prior and posterior EC and MS models of the real case. The red dots are the true measurements; the intervals in gray and light blue correspond to the region between the percentiles P5 and P95 of the prior and posterior prediction, respectively; the black and blue lines represent the prior mean and posterior mean, respectively.

**Figure 11** – Predicted QP data from the prior and posterior EC and MS models of the real case. The red dots are the true measurements; the intervals in gray and light blue correspond to the region between the percentiles P5 and P95 of the prior and posterior prediction, respectively; the black and blue lines represent the prior mean and posterior mean, respectively.
Figure 12 – Comparison of the predicted and measured EC (a) and MS (b) at the blind well of the real case. The red dots are the true measurements; the intervals in gray and light blue correspond to the region between the percentiles P5 and P95 of the prior and posterior realizations, respectively; the blue lines represent the posterior mean.
Figure 1 – Tucker decomposition of a third-order tensor.

182x85mm (300 x 300 DPI)
Figure 2 – Workflow of ES-MDA with RTD.

175x68mm (300 x 300 DPI)
Figure 3 – True and prior mean of EC and MS of the synthetic case: (a) true EC model; (b) true MS model; (c) prior mean of EC; and (d) prior mean of MS. It is clear the influence of the borehole data in the prior mean of EC and MS.
Figure 4 – Posterior mean, standard deviation (std.) and residual of EC and MS of the synthetic case: (a) posterior mean of EC models; (b) posterior mean of MS models; (c) posterior std. of EC; (d) posterior std. of MS; (e) absolute error between the posterior mean and the true EC; (f) absolute error between the posterior mean and the true MS.

650x274mm (300 x 300 DPI)
Figure 5 – Predicted IP data from the prior and posterior EC and MS models of the synthetic case. The red dots are the true measurements with noise; the black lines represent the true data without noise; the intervals in gray and light blue correspond to the region between the percentiles P5 and P95 of the prior and posterior prediction, respectively; the black and blue lines represent the prior mean and posterior mean, respectively.
Figure 6 – Predicted QP data from the prior and posterior EC and MS models of the synthetic case. The red dots are the true measurements with noise; the black lines represent the true data without noise; the intervals in gray and light blue correspond to the region between the percentiles P5 and P95 of the prior and posterior prediction, respectively; the black and blue lines represent the prior mean and posterior mean, respectively.
Figure 7 – Prior mean of EC (a) and MS (b) models of the real case. The black dashed lines represent the locations of the X-, Y- and depth slices, and the black dots represent the well locations.
Figure 8 – Posterior mean of EC (a) and MS (b) models of the real case. The black dashed lines represent the locations of the X-, Y- and depth slices.

243x308mm (300 x 300 DPI)
Figure 9 – Posterior standard deviation of EC (a) and MS (b) models of the real case. The black dashed lines represent the locations of the X-, Y- and depth slices.
Figure 10 – Predicted IP data from the prior and posterior EC and MS models of the real case. The red dots are the true measurements; the intervals in gray and light blue correspond to the region between the percentiles P5 and P95 of the prior and posterior prediction, respectively; the black and blue lines represent the prior mean and posterior mean, respectively.
Figure 11 – Predicted QP data from the prior and posterior EC and MS models of the real case. The red dots are the true measurements; the intervals in gray and light blue correspond to the region between the percentiles P5 and P95 of the prior and posterior prediction, respectively; the black and blue lines represent the prior mean and posterior mean, respectively.
Figure 12 – Comparison of the predicted and measured EC (a) and MS (b) at the blind well of the real case. The red dots are the true measurements; the intervals in gray and light blue correspond to the region between the percentiles P5 and P95 of the prior and posterior realizations, respectively; the blue lines represent the posterior mean.

182x205mm (300 x 300 DPI)