

Machine learning-accelerated gradient-based Markov chain Monte Carlo inversion applied to electrical resistivity tomography

Mattia Aleardi¹, Alessandro Vinciguerra^{1,2}, Eusebio Stucchi¹
and Azadeh Hojat^{3,4}

¹University of Pisa, Earth Sciences Department, Pisa, Italy, ²University of Florence, Earth Sciences Department, Florence, Italy, ³Shahid Babonar University of Kerman, Department of Mining Engineering, Kerman, Iran, and ⁴Politecnico di Milano, Department of Civil and Environmental Engineering, Milano, Italy

Received November 2021, revision accepted April 2022

ABSTRACT

Expensive forward model evaluations and the curse of dimensionality usually hinder applications of Markov chain Monte Carlo algorithms to geophysical inverse problems. Another challenge of these methods is related to the definition of an appropriate proposal distribution that simultaneously should be inexpensive to manipulate and a good approximation of the posterior density. Here we present a gradient-based Markov chain Monte Carlo inversion algorithm that is applied to cast the electrical resistivity tomography into a probabilistic framework. The sampling is accelerated by exploiting the Hessian and gradient information of the negative log-posterior to define a proposal that is a local, Gaussian approximation of the target posterior probability. On the one hand, the computing time to run the many forward evaluations needed for both the data likelihood evaluation and the Hessian and gradient computation is decreased by training a residual neural network to predict the forward mapping between the resistivity model and the apparent resistivity value. On the other hand, the curse of dimensionality issue and the computational effort related to the Hessian and gradient manipulation are decreased by compressing data and model spaces through a discrete cosine transform. A non-parametric distribution is assumed as the prior probability density function. The method is first demonstrated on synthetic data and then applied to field measurements. The outcomes provided by the presented approach are also benchmarked against those obtained when a computationally expensive finite-element code is employed for forward modelling, with the results of a gradient-free Markov chain Monte Carlo inversion, and also compared with the predictions of a deterministic inversion. The implemented approach not only guarantees uncertainty assessments and model predictions comparable with those achieved by more standard inversion strategies, but also drastically decreases the computational cost of the probabilistic inversion, making it similar to that of a deterministic inversion.

Key words: Electrical resistivity, Inversion, Tomography.

INTRODUCTION

The main challenge posed by geophysical inverse problems is inherent to their ill-posedness: different combinations of model parameters produce almost the same experimental observations. Such non-uniqueness usually arises from noise

Correspondence

Mattia Aleardi, University of Pisa, Earth Sciences Department, via S. Maria 53, 56126, Pisa, Italy. Email: mattia.aleardi@unipi.it

contamination in the data, insufficient data coverage, and the intrinsic mathematical properties of the forward operator. To properly assess the uncertainties affecting the inverse solution a probabilistic, Bayesian framework is usually adopted (Tarantola, 2005). Differently, a deterministic approach, although guaranteeing rapid convergence towards a best-fitting model, is incapable of accounting for the uncertainties in the predictions. The Bayesian inversion combines prior model uncertainties, data uncertainties (i.e., produced by noise contamination) and modelling errors (i.e., related to the approximated physics used to link the model to the data) into the posterior probability density (PPD) function that is the outcome of the probabilistic inversion. However, an analytical PPD computation is possible only for Gaussian-distributed model parameters and data, and linear forward operators. Otherwise, a numerical assessment of the PPD is needed, and to this end, Markov chain Monte Carlo (MCMC) sampling methods can be adopted (Sambridge and Mosegaard, 2002; Sen and Stoffa, 2013). However, expensive forward model operators and the curse of dimensionality occurring in high-dimensional parameter spaces (Curtis and Lomax, 2001) usually hamper the application of MCMC algorithms to geophysical inversions.

To sample the PPD, MCMC algorithms iteratively sample the parameter space by perturbing the current state of the chain (current model) according to a specified proposal distribution. The generated samples are accepted or rejected according to the Metropolis–Hasting rule. Theoretically, the estimated PPD is independent of the proposal for an infinite number of generated samples. However, from a more practical perspective, the probabilistic sampling is maximally efficient when the proposal is a fair approximation of the target PPD, and thus the choice of such a proposal critically determines the computational efficiency of the MCMC inversion. To solve this issue, some advanced MCMC recipes have been proposed over the last decades (e.g., self-adaptive MCMC algorithms, preconditioned MCMC, and hybrid MCMC approaches; Haario *et al.*, 2001, 2006; Turner and Sederberg, 2012; Sambridge, 2014; Vrugt, 2016; Holmes *et al.*, 2017). As an alternative, gradient-based MCMC (GB-MCMC) sampling (e.g., Hamiltonian Monte Carlo, Langevin Monte Carlo; Sen and Biswaw, 2017; Fichtner and Simutè, 2018; Fichtner and Zunino, 2019; Fichtner *et al.*, 2019; Gebrad *et al.*, 2020; Aleardi and Salusti, 2020; Aleardi, 2020) exploits the gradient information of the negative natural logarithm of the posterior to decrease the number of iterations needed to converge towards a stable posterior (MacKay, 2003; Neal, 2011). It has been demonstrated that the inclusion of this information not only speeds up the probabilistic sam-

pling, but also maximizes the independence of the samples while maintaining high acceptance probabilities. The downside is that derivatives have to be evaluated for each sampled model.

The computational demand of both gradient-free and gradient-based MCMC algorithms can be decreased by running the sampling in reduced model spaces (Lieberman *et al.*, 2010). Several either linear or nonlinear compression techniques based on different basis functions can be employed (Dejtrakulwong *et al.*, 2012; Lochbühler *et al.*, 2014; Aleardi, 2019; Liu and Grana, 2020). Another viable strategy makes use of approximated forward modelling operators. In this context, the regression ability of machine learning algorithms (theoretically able to approximate any nonlinear function) has been extensively exploited (Hansen and Cordua, 2017; Moseley *et al.*, 2020; Song *et al.*, 2021).

The electrical resistivity tomography (ERT) is widely used in a variety of hydrogeological, environmental and engineering problems to infer the subsurface resistivity values (Rucker *et al.*, 2011; Uhlemann *et al.*, 2015; Moradipour *et al.*, 2016; Whiteley *et al.*, 2017; Bièvre *et al.*, 2018; Hojat *et al.*, 2019a; Dahlin, 2020; Hermans and Paepen, 2020; Aleardi *et al.*, 2020; Loke *et al.*, 2020; Norooz *et al.*, 2021). Due to incomplete data coverage and noise contamination, the ERT is an ill-posed problem affected by non-uniqueness and instability (i.e., small variations of the data produce large perturbations in the predictions), and hence, an accurate estimation of the model uncertainty is of primary importance. However, the ERT is routinely solved through deterministic approaches in which optimization algorithms minimize a predefined objective function. Such methods are generally computationally efficient, but provide an estimation of the model (i.e., the most likely solution) without accurately quantifying the associated uncertainty. On the other hand, the computing time needed for multiple forward evaluations (e.g., through a finite element code) hampers the application of standard MCMC approaches to invert ERT data.

In our recent research works on Bayesian ERT inversion, the main comment raised by reviewers and colleagues always concerned the increased computational cost with respect to the local inversion. For specific implementations, the probabilistic ERT is computationally feasible only if dedicated computational resources are used to run the inversion (Aleardi *et al.*, 2020). The computational workload is much higher than that of a deterministic strategy, even if compression methods are used to reduce the number of unknowns (Vinciguerra *et al.*, 2021) or when, instead of MCMC algorithms, less accurate approaches (i.e., ensemble-based algorithms) are used to numerically evaluate the PPD (Aleardi *et al.*, 2021). This

motivated us to exploit the compression capability of the discrete cosine transform (DCT), the regression ability of a machine learning algorithm, and the efficiency of a gradient-based MCMC to make the computational cost of the Bayesian ERT comparable to that of a local inversion even when limited hardware resources are employed. More in detail, we here exploit the geometrical properties of the negative log-posterior to define a proposal density that is a local approximation of the target PPD. In this context, our primary aim was to decrease the computational cost of the probabilistic sampling, while maintaining high acceptance rates, accurate model predictions and uncertainty assessments. The use of a DCT reparameterization of both data and model spaces mitigates the ill-posedness of the problem, the curse of dimensionality issues, and also reduces the computing time for the Hessian and gradient calculation and manipulation. We further decrease the computational cost of the inversion by replacing the standard finite-elements (FE) forward modelling code with the predictions of a properly trained residual neural network (RNN; Glorot and Bengio, 2010; Mo *et al.*, 2020). Note that the modelling error introduced by the imperfect and approximated physics that relates the model to the data is properly accounted for and propagated into the final PPD.

After discussing the theoretical aspects of the proposed algorithm, we first present a synthetic experiment over a simplified subsurface resistivity model, and we finally show the application to field data. The results provided by our method are also compared with the predictions of more standard probabilistic and deterministic approaches. All the inversion tests described here have been run on a notebook equipped with an Intel i7-10750H CPU@2.60GHz with 16 Gb of RAM, and with NVIDIA GeForce RTX 2060. The IP4DI Matlab software (Karaoulis *et al.*, 2013) provided us with the FE forward operator, the code for the Jacobian computation via the adjoint-state method and the code for the deterministic inversion.

METHODOLOGY

Gradient-based Markov chain Monte Carlo

Let us first consider a deterministic inversion framework, in which under the assumption of Gaussian-distributed noise and model parameters, the error function to be minimized can be written as follows (Menke, 2018; Aster *et al.*, 2018):

$$E(\mathbf{m}) = \left\| \mathbf{C}_d^{-\frac{1}{2}} (\mathbf{d} - G(\mathbf{m})) \right\|_2^2 + \left\| \mathbf{C}_m^{-\frac{1}{2}} (\mathbf{m} - \mathbf{m}_{prior}) \right\|_2^2, \quad (1)$$

where \mathbf{m} and \mathbf{d} denote the model parameters and the observed data vectors, respectively; \mathbf{C}_d and \mathbf{C}_m represent the data and prior model covariance matrices; \mathbf{m}_{prior} is the prior model vector, whereas G represents the forward modelling operator. The error function of equation (1) can be iteratively minimized through a local quadratic approximation around the current model \mathbf{m}_k :

$$E(\mathbf{m}_k + \Delta\mathbf{m}) \approx \tilde{E}(\mathbf{m}) = E(\mathbf{m}_k) + \Delta\mathbf{m}^T \nabla_{\mathbf{m}} E(\mathbf{m}_k) + \frac{1}{2} \Delta\mathbf{m}^T \nabla_{\mathbf{m}}^2 E(\mathbf{m}_k) \Delta\mathbf{m} + O(\|\Delta\mathbf{m}\|^3), \quad (2)$$

where $\nabla_{\mathbf{m}} E(\mathbf{m}_k)$ and $\nabla_{\mathbf{m}}^2 E(\mathbf{m}_k)$ denote the first and second derivatives of $E(\mathbf{m})$ around \mathbf{m}_k , respectively, while $\Delta\mathbf{m} = \mathbf{m} - \mathbf{m}_k$. It results that:

$$\nabla_{\mathbf{m}} E(\mathbf{m}_k) = \mathbf{g} = \mathbf{J}^T \mathbf{C}_d^{-1} \Delta\mathbf{d}(\mathbf{m}_k) + \mathbf{C}_m^{-1} (\mathbf{m}_k - \mathbf{m}_{prior}), \quad (3)$$

and

$$\nabla_{\mathbf{m}}^2 E(\mathbf{m}_k) = \mathbf{H} = \left(\mathbf{J}^T \mathbf{C}_d^{-1} \mathbf{J} \right) + \frac{\partial \mathbf{J}^T}{\partial \mathbf{m}^T} \mathbf{C}_d^{-1} (\Delta\mathbf{d}(\mathbf{m}_k) \dots \Delta\mathbf{d}(\mathbf{m}_k)) + \mathbf{C}_m^{-1} = \mathbf{H}_o + \mathbf{B} + \mathbf{C}_m^{-1}, \quad (4)$$

where $\mathbf{B} = \frac{\partial \mathbf{J}^T}{\partial \mathbf{m}^T} \mathbf{C}_d^{-1} (\Delta\mathbf{d}(\mathbf{m}_k) \dots \Delta\mathbf{d}(\mathbf{m}_k))$, and $\Delta\mathbf{d}(\mathbf{m}_k) = G(\mathbf{m}_k) - \mathbf{d}$, with \mathbf{J} representing the Jacobian matrix (i.e. the partial derivative of the data with respect to model parameters). For computational feasibility reasons, an approximated Hessian is usually employed: $\mathbf{H} \approx \mathbf{H}_a = \mathbf{H}_o + \mathbf{C}_m^{-1}$. Therefore, equation (2) can be re-written as:

$$\tilde{E}(\mathbf{m}) = \frac{1}{2} (\mathbf{m} - \mathbf{m}_k + \mathbf{H}_a^{-1} \mathbf{g})^T \mathbf{H}_a (\mathbf{m} - \mathbf{m}_k + \mathbf{H}_a^{-1} \mathbf{g}) + const., \quad (5)$$

Then, the minimizer of $\tilde{E}(\mathbf{m})$ can be computed according to:

$$\mathbf{m}_{k+1} = \mathbf{m}_k - \mathbf{H}_a^{-1} \mathbf{g}. \quad (6)$$

An approximated uncertainty quantification in deterministic inversions can be inferred from the inverse of the Hessian matrix at the convergence point.

Differently, the Bayesian solution of an inverse problem is fully expressed by the posterior probability density (PPD) function in model space:

$$p(\mathbf{m}|\mathbf{d}) = \frac{p(\mathbf{d}|\mathbf{m})p(\mathbf{m})}{p(\mathbf{d})}, \quad (7)$$

where $p(\mathbf{m}|\mathbf{d})$ denotes the PPD, $p(\mathbf{d}|\mathbf{m})$ is the data likelihood function, whereas $p(\mathbf{m})$ and $p(\mathbf{d})$ are the prior distributions of model parameters and data, respectively. In most applications, the data likelihood is derived from the L2 norm difference

between predicted and observed data, under the assumption of Gaussian-distributed noise:

$$p(\mathbf{d}|\mathbf{m}) \propto -0.5 \times (\mathbf{d} - G(\mathbf{m}))^T \mathbf{C}_d^{-1} (\mathbf{d} - G(\mathbf{m})). \quad (8)$$

When the $p(\mathbf{m}|\mathbf{d})$ cannot be expressed in a closed form, Markov chain Monte Carlo (MCMC) sampling algorithms can be used for a numerical assessment of the PPD. These methods iteratively sample the target posterior and define the probability to move from the current model \mathbf{m}_k to the proposed model \mathbf{m}_{k+1} according to the Metropolis–Hasting rule:

$$\alpha = p(\mathbf{m}_{k+1}|\mathbf{m}_k) = \min \left[1, \frac{p(\mathbf{m}_{k+1})}{p(\mathbf{m}_k)} \times \frac{p(\mathbf{d}|\mathbf{m}_{k+1})}{p(\mathbf{d}|\mathbf{m}_k)} \times \frac{q(\mathbf{m}_k|\mathbf{m}_{k+1})}{q(\mathbf{m}_{k+1}|\mathbf{m}_k)} \right], \quad (9)$$

where $q(\cdot)$ is the proposal distribution that draws the new model from the probability distribution $q(\mathbf{m}_{k+1}|\mathbf{m}_k)$. Note that the proposal ratio term vanishes if symmetric proposals are used (i.e., a Gaussian proposal centred on the current state of the chain). If \mathbf{m}_{k+1} is accepted, $\mathbf{m}_k = \mathbf{m}_{k+1}$. Otherwise, \mathbf{m}_k is repeated in the chain and another perturbed model is generated. The ensemble of sampled models after the burn-in period is used to numerically assess the PPD. It is clear that the main computational requirement of an MCMC inversion lies in the many forward modelling evaluations needed to compute the data likelihood for each sampled model.

Now we can formulate the Bayesian inversion in terms of $E(\mathbf{m})$, \mathbf{H} and \mathbf{g} , under Gaussian assumptions for data and model parameters:

$$p(\mathbf{m}) \propto \exp \left(-\frac{1}{2} (\mathbf{m} - \mathbf{m}_{prior})^T \mathbf{C}_m^{-1} (\mathbf{m} - \mathbf{m}_{prior}) \right), \quad (10)$$

$$p(\mathbf{d}|\mathbf{m}) \propto \exp \left(-\frac{1}{2} (\mathbf{d} - G(\mathbf{m}))^T \mathbf{C}_d^{-1} (\mathbf{d} - G(\mathbf{m})) \right), \quad (11)$$

$$p(\mathbf{m}|\mathbf{d}) \propto \exp(-E(\mathbf{m})). \quad (12)$$

A Gaussian approximation of the PPD around \mathbf{m}_k can be found as:

$$p(\mathbf{m}|\mathbf{d}) \approx \tilde{p}(\mathbf{m}|\mathbf{d}) \propto \exp \left(-\frac{1}{2} (\mathbf{m} - (\mathbf{m}_k - \mathbf{H}_a^{-1} \mathbf{g}))^T \mathbf{H}_a (\mathbf{m} - (\mathbf{m}_k - \mathbf{H}_a^{-1} \mathbf{g})) \right). \quad (13)$$

This equation illustrates that the approximated PPD is Gaussian distributed according to $\mathcal{N}(\mathbf{m}_k - \mathbf{H}_a^{-1} \mathbf{g}; \mathbf{H}_a^{-1})$ with mean equal to the minimizer of $\tilde{E}(\mathbf{m})$ and covariance equal to the inverse of the Hessian matrix. An adaptive proposal for each sampled model can be formulated from such local Gaussian approximation as follows:

$$q(\mathbf{m}) \propto \exp \left(-\frac{1}{2} (\mathbf{m} - (\mathbf{m}_k - \lambda \mathbf{H}_a^{-1} \mathbf{g}))^T \frac{\mathbf{H}_a}{\mu^2} (\mathbf{m} - (\mathbf{m}_k - \lambda \mathbf{H}_a^{-1} \mathbf{g})) \right). \quad (14)$$

Then, similarly to standard MCMC sampling, each proposed model is accepted according to the Metropolis–Hasting rule. We must keep in mind that the new proposal is not symmetric (i.e., the proposal distribution is not centred on the current model) and for this reason, the proposal ratio should be evaluated for each sampled model. However, this is not computationally demanding since the proposal is analytically tractable. We also consider the full Hessian and not only its diagonal entries so that posterior correlations between model parameters are taken into consideration. Note that including information about the Hessian matrix means that the proposal depends on the curvature of the negative log-posterior and then compensates for the different parameter illuminations.

After tailoring the proposal density $q(\mathbf{m})$ to the underlying approximation of the PPD, the proposed model can be analytically generated according to:

$$\mathbf{m}_{k+1} = \mathbf{m}_k - \lambda \mathbf{H}_a^{-1} \mathbf{g} + \mu \mathbf{H}_a^{-\frac{1}{2}} \mathbf{n}, \quad (15)$$

with $\mathbf{H}_a^{-1} = \mathbf{H}_a^{-1/2} (\mathbf{H}_a^{-1/2})^T$, where \mathbf{n} is a random vector drawn from the Gaussian distribution $\mathcal{N}(0, \mathbf{I})$, with \mathbf{I} denoting the identity matrix. For a linear problem with Gaussian assumptions and an exact Hessian, the proposed method results in a perfect sampling with an acceptance probability equal to 1 (i.e., all the proposed models are accepted; see Martin *et al.*, 2012). Finally, even though the proposal is derived from a local Gaussian approximation of the PPD, it can be used to sample from whatever type of posterior model and under whatever *a priori* assumption (e.g., non-parametric prior), as it has been done in the following experiments. λ and μ^2 are tunable parameters that determine the step length along the negative gradient direction and the variance of the random perturbation around the minimizer of $\tilde{E}(\mathbf{m})$. These values influence the exploration and exploitation of the algorithm, and then the acceptance rate of the probabilistic sampling. The λ value should be large enough to make the proposal depend on the gradient information, but small enough so that the updating model is not dominated by the deterministic information. The μ^2 value should be large enough to ensure an efficient exploration of the model space, but small enough so that the gradient information is not completely masked by the random update. A proper setting of these parameters can be easily found by the inspection of the acceptance rate and the convergence rate of the sampling towards the steady state. Note that the values of these parameters only influence the efficiency of the sampling and not the final estimated PPD. In particular, in all the following examples, we consider an acceptance rate around

0.8 to be optimal, and in this context all the λ and μ^2 values around 0.2–0.5 and 0.7–0.9, respectively, work well.

The major computational requirement of the implemented approach with respect to gradient-free MCMC algorithms is the need of computing the Jacobian for each sampled model. To this end, when the forward is expressed by a partial differential equation, the adjoint-state method can be used (Plessix, 2006). The Jacobian can also be evaluated using a finite-difference scheme or for weakly nonlinear problems, a linearized approximation of the forward mapping can be employed as well. Large dimensional parameter spaces also need an extra computational workload related to the manipulation of large Hessian matrices and gradient vectors. Therefore, a compression strategy can help to reduce the dimension of the parameter and data spaces (i.e., thus mitigating the ill-conditioning and reducing the dimension of the Hessian matrix and gradient vector), and to also decrease the number of forward evaluations needed for the computation of the Jacobian when a finite-difference strategy is adopted.

Discrete cosine transform

In this section, we briefly introduce the discrete cosine transform (DCT) compression used to reduce the model and data domains. Additional information about this popular compression strategy and its application to solve geophysical inverse problems can be found in Lochbühler *et al.* (2014) and Moghadas and Vrugt (2019).

The basis functions employed by the DCT are cosine functions oscillating with different frequencies. This transformation can be applied to mono or multidimensional signals. For example, for a 2D resistivity model $\rho(x, y)$ with $x = [1, \dots, M_x]$ and $y = [1, \dots, M_y]$ the transformation can be computed as follows:

$$\mathbf{R} = \mathbf{B}_y \boldsymbol{\rho} \mathbf{B}_x^T, \quad (16)$$

where \mathbf{B}_x and \mathbf{B}_y are the matrices with dimensions $M_x \times M_x$ and $M_y \times M_y$, expressing the basis functions, whereas the $M_y \times M_x$ matrix \mathbf{R} contains the DCT coefficients. The DCT concentrates most of the information of the original signal into the low-order coefficients, and hence an approximation of the 2D resistivity model can be computed as follows:

$$\tilde{\rho} = \left(\mathbf{B}_y^q \right)^T \mathbf{R}_{qp} \mathbf{B}_x^p, \quad (17)$$

where $\tilde{\rho}$ is the approximated model, \mathbf{B}_y^q is a $[q \times M_y]$ matrix, with only the first q rows of \mathbf{B}_y ; \mathbf{B}_x^p is a $[p \times M_x]$ matrix with only the first p rows of \mathbf{B}_x ; the matrix \mathbf{R}_{qp} represents the first q

rows and p columns of \mathbf{R} . The q and p values are the retained number of basis functions along the y and x directions. Therefore, the DCT transformation reduces the full $(M_y \times M_x)$ -D model space to a $(q \times p)$ -D DCT-compressed domain (with $p < M_x$ and $q < M_y$). In this context, the $p \times q$ non-zero numerical coefficients of the \mathbf{R}_{qp} matrix become the inverted parameters in the compressed space.

Forward approximation through a trained network

Similar to Aleardi *et al.* (2022), we use an approximated forward operator that in this case significantly speeds up both the computation of the Jacobian matrix (i.e., when a finite-difference scheme is adopted) and the data likelihood evaluation. To this end, we train a residual neural network. The main idea is to use the trained network as a computationally efficient approximation to the forward problem. Multiple resistivity models and associated apparent resistivity data are used to make the residual neural network (RNN) learn the non-linear mapping between the model and the data space. The models forming the training and validation sets are generated according to prior model assumptions, while a 2.5D finite-elements (FE) Matlab modelling code constitutes the forward operator (Karaoulis *et al.*, 2013) that computes the associated apparent resistivity data. We refer the interested reader to Aleardi *et al.* (2022) for more details about this approach. Here, we briefly introduce the reader to the basic RNN principles and we briefly show the adopted network architecture.

Similar to a convolutional neural network, a RNN uses convolutional filters and fully connected layers to extract features from mono- or multidimensional inputs. Such networks can be used to solve both classification or regression problems (Monajemi *et al.*, 2016; Goodfellow *et al.*, 2016). However, in a traditional convolutional neural network, each layer feeds into the next one. Differently, a RNN makes use of shortcuts and skip connections to add the result of a shallow layer directly to the corresponding output of a deeper layer. This strategy helps to prevent the so-called vanishing gradient problem that occurs when training a deep CNN. This results in the degradation problem: the accuracy (i.e., the similarity between desired and computed network responses) gets saturated for a given number of layers and then starts degrading rapidly if additional layers are added.

In our application, the resistivity model is the input of the network, whereas the flattened apparent resistivity section constitutes the output response. The optimization of the network internal parameters is driven by the minimization of the root-mean-square error between desired and computed

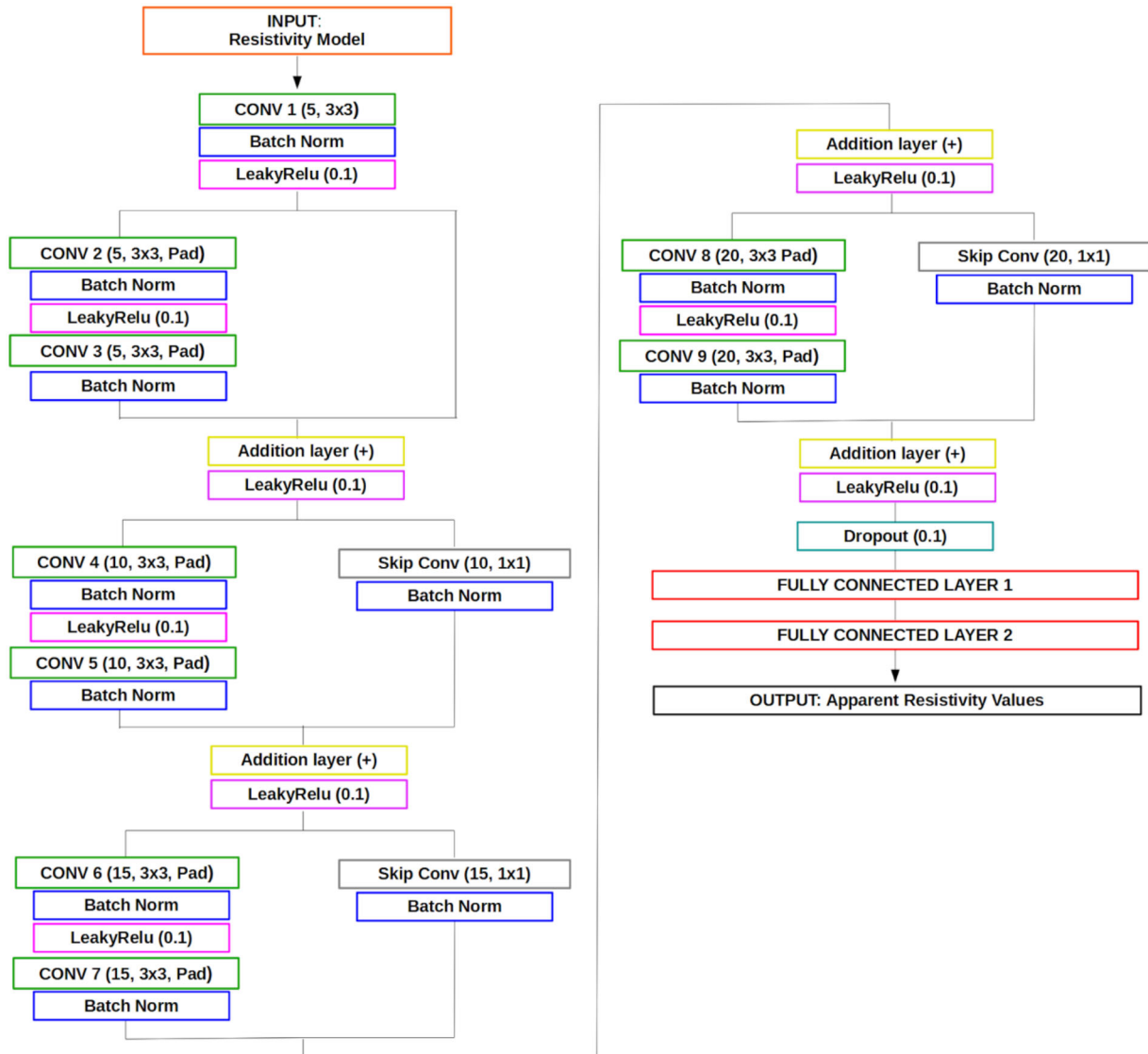


Figure 1 Representation of the employed RNN architecture annotated with key parameters. For example, in the second convolutional layer ‘CONV 2’, the term within bracket (5, 3×3, Pad) indicates that we employ 5 convolutional filters with size 3×3 and that zero-padding is applied. Note that the LeakyRelu with a leakage value of 0.1 is used as the activation function in all the convolutional layers. Skip convolutions are used to adjust features dimensions before additional layers. A dropout of 10% and batch normalization are used to avoid overfitting and as the regularization operator, respectively. The only difference in the synthetic and field data applications concerns the dimension of input and output response. See the text for details. Figure taken from Aleardi *et al.* (2022).

output. The employed network architecture in both the synthetic and field inversions is represented in Figure 1.

The implemented probabilistic inversion

The resistivity values along a 2D profile of dimension ($M_y \times M_x$) and the associated N apparent resistivity values, constitute the model and data vectors, respectively. In common applications, hundreds of model parameters have to be inverted

for and, in this context, the curse of dimensionality problem makes the application of standard probabilistic inversion procedures a formidable computational challenge.

Here we mitigate the curse of dimensionality by compressing the model space with a reduced number of DCT coefficients expressed by a $q \times p$ matrix ($p < M_x$ and $q < M_y$). On the same line, to reduce the dimension of the Hessian matrix, we also employ the DCT to compress the N -dimensional data space to a b -dimensional domain (with $b < N$). Because of its

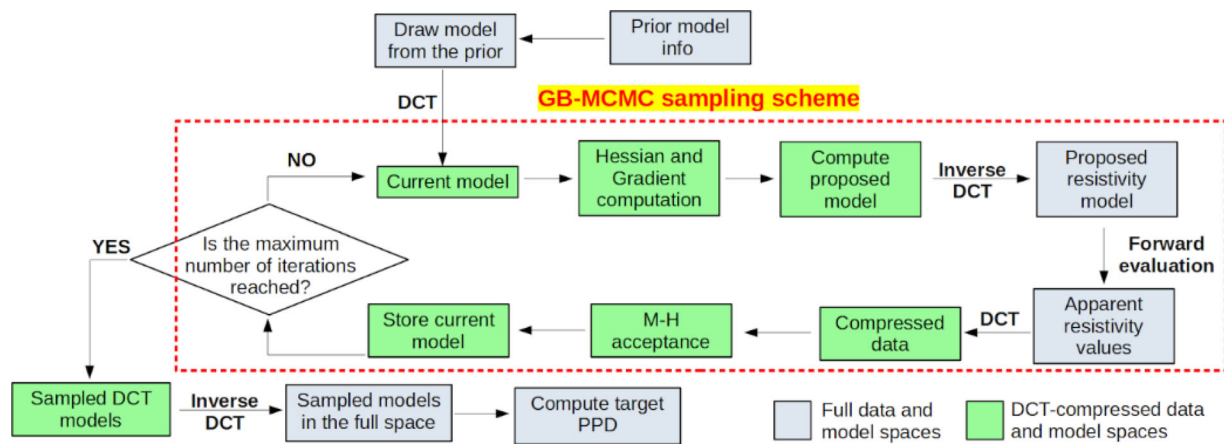


Figure 2 Schematic representation of the GB-MCMC inversion framework. Green and grey rectangles refer to operations performed in the reduced and full spaces, respectively.

trapezoidal shape, the apparent resistivity section cannot be enclosed within a 2D matrix, and thus, it has been flattened to a 1D vector before the DCT projection. More detailed discussions about DCT compression applied to electrical resistivity tomography inversion can be found in Aleardi *et al.* (2021) and Vinciguerra *et al.* (2021). As in these works, prior model realizations and apparent resistivity data have been used to determine the optimal number of basis functions to retain. In particular, we analysed how the explained variability (i.e., the ratio between the variance of the compressed and original signals) changes as the number of considered DCT coefficients increases. Therefore, the GB-MCMC algorithm runs in the compressed ($p \times q$)-D model space and estimates the DCT coefficients expressing the resistivity model from the retained b basis in the data domain. As a consequence, the computation of the likelihood ratio, proposal ratio and prior ratio for each sampled model is performed in the compressed space. A schematic representation of this strategy is given in Figure 2. Note that the multiple forward and inverse DCT projections needed in each iteration can be analytically evaluated with a negligible computational cost. Finally, the sampled models after the burn-in period are projected onto the full, uncompressed, space to numerically compute the statistical properties of the PPD in the original resistivity domain.

A simple Gaussian prior is often employed in probabilistic inversions even if this usually oversimplifies the actual distribution of the model parameters in the investigated area. For example, in the case of multiple litho-fluid facies, the parameter distribution might be better approximated by a multimodal prior, in which each mode is associated with a different facies (Aleardi *et al.*, 2020). Here, in both the synthetic and field applications, we assume multimodal non-parametric priors that

properly model the facies dependency of the resistivity values in the study areas. A stationary spherical variogram model is used to express the lateral and vertical variability pattern. We also assume a Gaussian-distributed noise, so that the data likelihood can be analytically computed from the L2 norm distance between predicted and observed data vectors. The direct sequential simulation algorithm (Soares, 2001) is used to generate random realization from the *a priori* model. The non-parametric prior in the uncompressed space impedes an analytical derivation of the prior in the DCT domain; thus, the prior assumption in the compressed space is numerically computed by applying the kernel density estimation algorithm to prior resistivity realizations projected onto the DCT space. Differently, the assumed Gaussian noise model allows for an analytical derivation of the data covariance matrix in the compressed data domain.

The main limitation of any GB-MCMC approach arises from the need of computing the gradient of the negative log-posterior. For this reason, this strategy is usually applied to problems in which the derivative information can be computed quickly (Neal, 2011). In our application, the Jacobian matrix can be derived by adopting a finite-difference scheme or the adjoint-state method. However, if a common finite-element (FE) forward operator is employed the computation of the Jacobian will make the GB-MCMC inversion computational unfeasible on the limited hardware resource here employed. For example, if we consider a forward finite-difference (FD) scheme and the uncompressed model space, $M_y \times M_x$ forward evaluations are needed to evaluate the Jacobian around each sampled model. The resulting computational workload is impractical, even though the Jacobian computation can be easily parallelized (i.e., each

column of this matrix can be independently computed on a different core). The DCT model compression can partially mitigate this issue because, after compression, only $q \times p$ forward runs are needed to evaluate the Jacobian with the forward FD scheme. However, also in this context the application of the implemented inversion scheme would be very computationally demanding. See the next section for a detailed discussion on the times needed for the Jacobian computation using different approaches and forward operators. Here the computational cost for the Jacobian evaluation is drastically reduced by the approximated RNN forward operator. This not only makes the GB-MCMC inversion feasible on the employed hardware resources but also provides a probabilistic inversion framework much faster than a gradient-free MCMC.

The use of the RNN forward introduces a modelling error related to the approximated physics that maps the model parameters onto the corresponding data. Ignoring this theoretical error might generate overfitting with the observed data and introduce artefacts in the final solution. Therefore, we also properly propagate the error introduced by the network approximation onto the final PPD. To this end, the data covariance matrix C_d is computed as the sum of the noise contaminating the data C_n and the modelling error that takes into account the imperfect physics relating the model to the data C_p (Menke, 2018): $C_d = C_n + C_p$. Both noise and modelling errors are considered to be Gaussian-distributed with a zero mean value. The modelling error matrix is derived from the covariance of the difference between desired and actual network outputs and is computed on the validation set (Hansen and Cordua, 2017).

SYNTHETIC INVERSIONS

We consider a schematic subsurface resistivity model represented by a rectangular block with a resistivity of $50 \Omega\text{m}$ hosted in a homogeneous half-space with resistivity equal to $150 \Omega\text{m}$ (Fig. 3). The high and low resistivity values within the half-space and the rectangular block, respectively, can be thought of as representative of two different litho-fluid classes. The area is discretized with $11 \times 35 = 385$ rectangular cells with vertical and lateral dimensions of 0.5 m and 1.0 m , respectively. The resistivity values within the cells correspond to the model parameters to be estimated. We simulate a Wenner acquisition layout with 36 electrodes with $a = 1.0 \text{ m}$. The maximum a value is 11. This configuration results in 198 data points. In this example, we employ the Wenner layout because it has also been used for the field data acquisition, but the presented inversion framework can be applied to different elec-

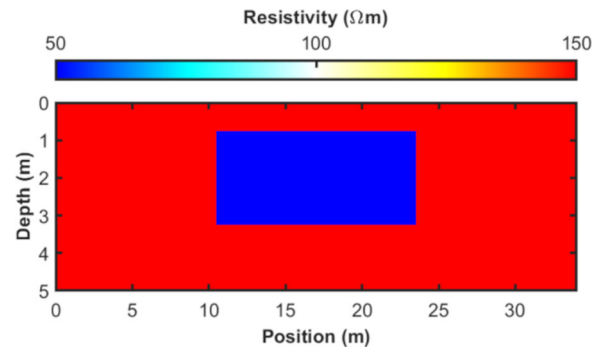


Figure 3 The true model for the synthetic inversion.

trode configurations as well. The finite-element (FE) code was used to compute the noise-free observed dataset that was contaminated with uncorrelated Gaussian noise with a standard deviation equal to 20% of the total standard deviation of the noise-free data.

Figure 4 represents the prior model assumptions used to generate the training and validation sets and also used in the following probabilistic inversions. We employ a non-parametric prior estimated by applying the kernel density estimation algorithm (with an Epanechnikov kernel) to Gaussian perturbations of two resistivity columns extracted from the central part of the true model. This multimodal prior compared to a simple Gaussian assumption guarantees a better representation of the actual distribution of the resistivity values. A spherical variogram is used as the spatial continuity pattern with horizontal and vertical ranges equal to 8 and 3 m, respectively.

We train the residual neural network (RNN) on prior model realizations and associated data to approximate the nonlinear forward modelling operators (Aleardi *et al.*, 2022). As demonstrated in that study, only 2000 models for training and 500 for validations are enough to get accurate forward predictions. The FE code is used to generate the associated apparent resistivity values. Considering a parallel code running on the previously mentioned hardware resources, the generation of the 2000 training examples takes 15 minutes, approximately, while the training running on the GPU is completed in less than 5 minutes. As an example, Figure 5 compares some examples of apparent resistivity pseudosections predicted by the trained RNN and the associated FE datasets taken from the validation set. The close similarity between the actual and desired output confirms that the network can effectively approximate the nonlinear relation linking the model to the data.

The next step is to select the optimal number of discrete cosine transform (DCT) coefficients to compress the data and

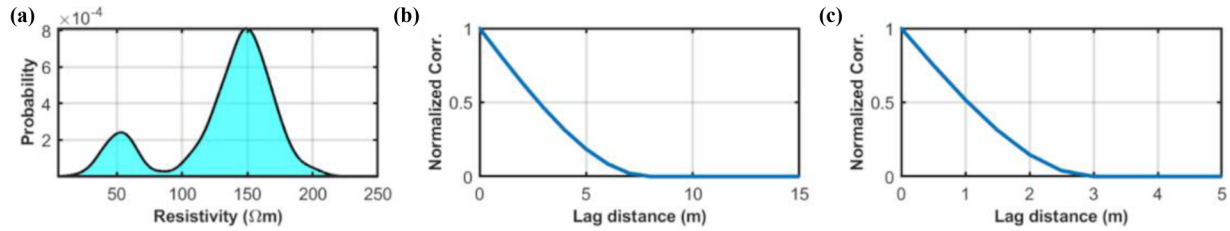


Figure 4 (a) Non-parametric prior distribution for the synthetic example. (b and c) Spatial correlation functions associated with the assumed 2D variogram model for the horizontal and vertical directions, respectively.

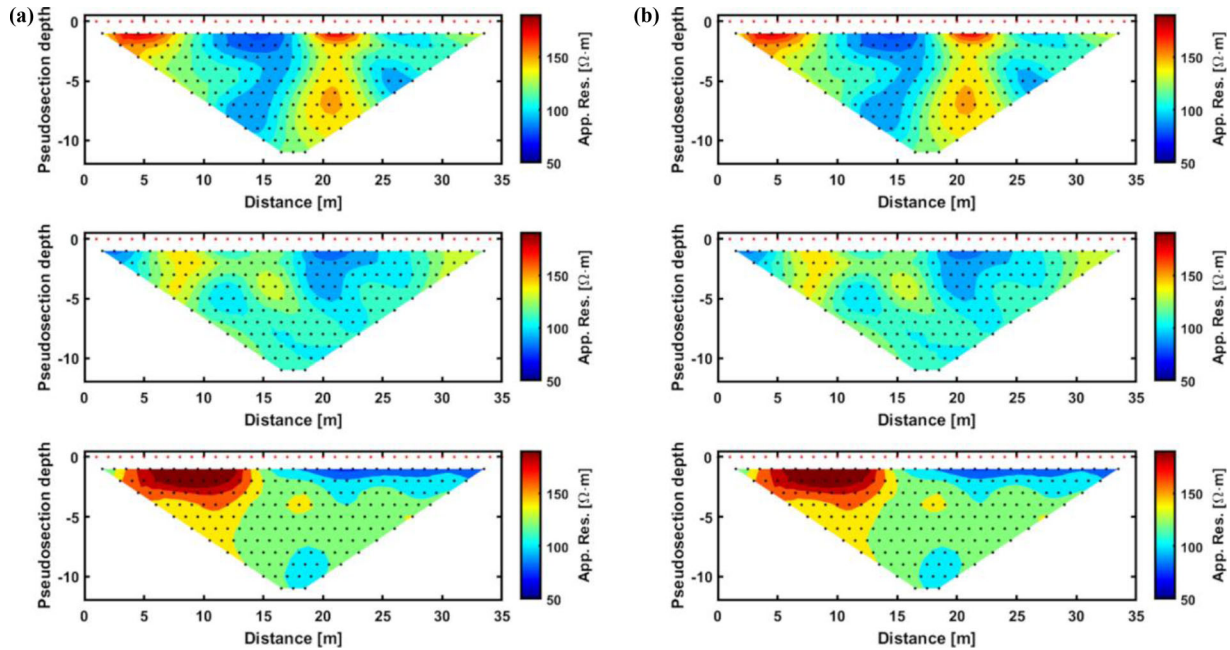


Figure 5 Some comparisons between (a) the network responses and (b) the desired (i.e. FE) output from the validation set.

the model spaces. Here we employ the same strategy as described in Vinciguerra *et al.* (2021) and Aleardi *et al.* (2021) to which we refer the readers for further details. The basic idea is to exploit prior model realizations and associated data to assess how the explained variability (i.e., measured as the ratio between the variance of the compressed on original signal) changes as the number of retained DCT coefficients varies. Figure 6 shows an example of this approach on a model and the associated data vector. In the model space, we observe that just 3, and 5 coefficients along the two DCT dimensions almost completely explain the full variability of the original signal. Similarly, 80 coefficients are needed to successfully approximate the data. This means that the 385-D model space can be sparsely represented by 15 coefficients, while the 198-D data domain is compressed in an 80-D domain: this significantly reduces the dimensions of the Hessian matrices and the time to compute the Jacobian via an FD approach.

After selecting the appropriate number of DCT coefficients, we compare the computing time needed for the Jacobian computation in this synthetic experiment using different approaches and considering the full and compressed data and model spaces. We consider 5 different cases:

1. Jacobian computation through a forward FD approach in the DCT-compressed model and data spaces and using the RNN forward approximation;
2. Jacobian computation through a forward FD approach in the DCT-compressed model and data spaces and using the FE forward code;
3. Jacobian computation through a forward FD approach in the uncompressed model and data spaces and using the RNN forward approximation;
4. Jacobian computation through a forward FD approach in the uncompressed model and data spaces and using the FE routine;

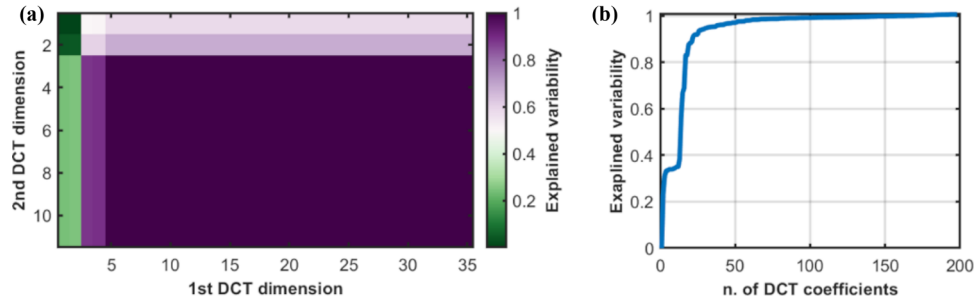


Figure 6 (a) Examples of explained model variability as the number of retained DCT coefficients along the two DCT dimensions increases. The numerical value with coordinate (x, y) indicates the explained variability if the first x , and y DCT coefficients along the first and second DCT dimensions, respectively, are used for compressing the resistivity model. (b) Explained variability as the number of DCT coefficients increases for the data associated with the model shown in (a).

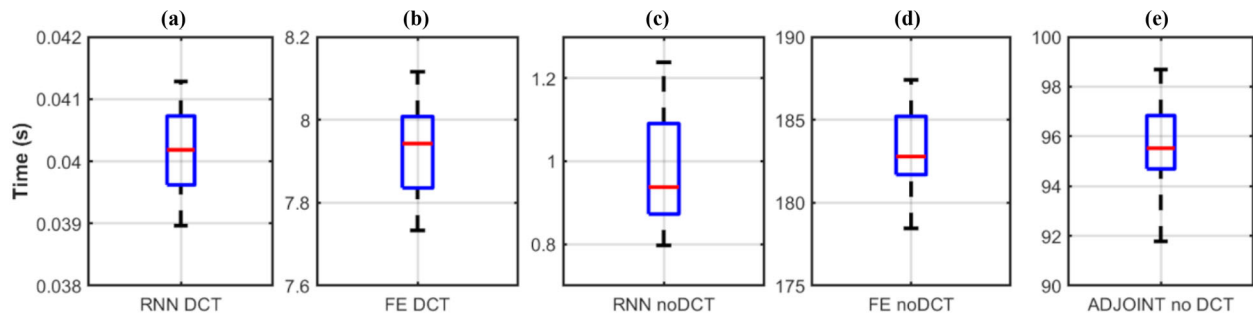


Figure 7 Box plots showing the times needed for the Jacobian computation using different approaches: a) RNN forward + FD scheme in the compressed domain. (b) FE forward + FD scheme in the compressed domain. (c) RNN forward + FD scheme in the uncompressed space. (d) FE forward + FD scheme in the uncompressed space. (e) Adjoint-state method running in the uncompressed space. Note the different scales on the vertical axes.

5. Jacobian computation via the adjoint-state method running in the uncompressed model and data domains.

Note that the code for the adjoint-state method implements this computation only in the full model and data space, and for this reason, this method has not been considered for the Jacobian computation in the DCT space. Also note that when the FD strategy and the FE code are used, the computation of the columns of the Jacobian is distributed across different cores. For each case, we run 50 Jacobian evaluations on a resistivity model extracted from the prior and we evaluate the computing time for each run. The box plots of Figure 7 summarize our results: In both the compressed and full spaces, the RNN forward reduces the computing time for a single Jacobian evaluation of almost three orders of magnitudes with respect to the FE algorithm, while in the full space the adjoint-state method is two times faster than the FD + FE approach. For what concerns our inversion running in the DCT space, Figure 7 demonstrates that when the FE code is used as the forward modelling engine, it is more con-

venient to compute the Jacobian directly on the compressed domain instead of evaluating the Jacobian in the full spaces with the adjoint-state method and then projecting this matrix back onto the DCT domain: This reduces the time for a single Jacobian computation of one order of magnitude (compare Fig. 7b and e).

Figure 8 compares the Jacobians derived in the full model and data domain with the three considered approaches. As expected, it turns out that the FD and adjoint-state methods provide very similar results, but also the approximated RNN forward gives very accurate Jacobian estimations albeit some minor scattering is visible moving along the columns of the matrix. The difference between the Jacobian computed with the RNN forward and the adjoint-state method becomes even smaller if we consider the DCT domain (Fig. 9), this is because the projection onto the compressed space attenuates the lateral scattering visible in Figure 8(a). For this comparison, the Jacobian provided by the adjoint-state in the uncompressed spaces has been analytically projected onto the compressed domain.

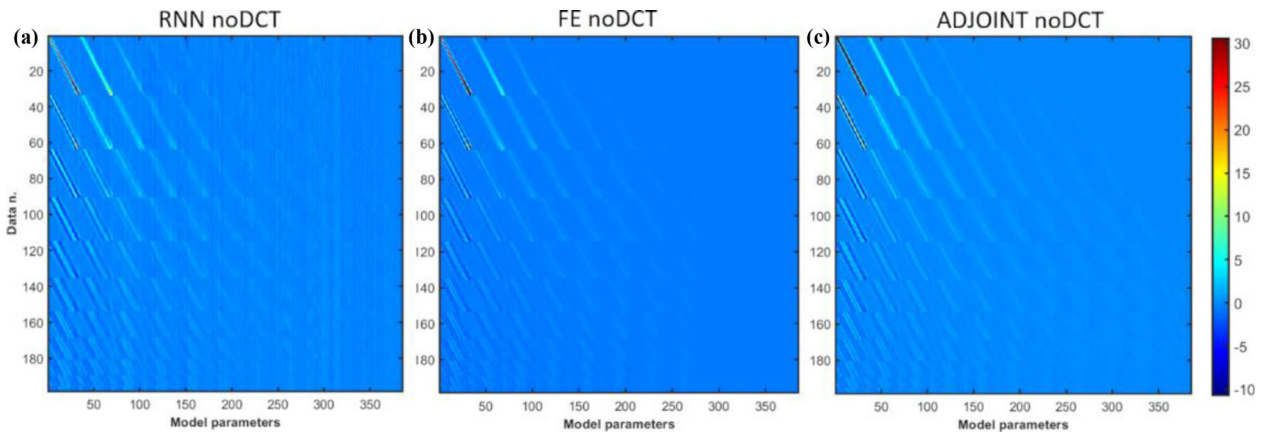


Figure 8 Comparison between the Jacobian matrices computed in the full model and data spaces with three different approaches. (a) RNN forward + FD scheme. (b) FE forward + FD scheme. (c) Adjoint-state method.

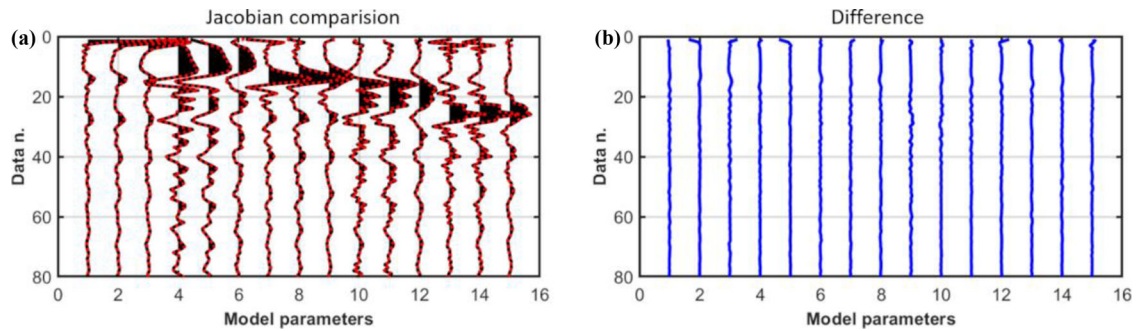


Figure 9 (a) Comparison between the Jacobian matrices computed with the RNN + FD approach in the DCT space (black) and the projection onto the DCT domain of the Jacobian derived with the adjoint-state method in the full space (red). (b) Their sample-by-sample difference.

Our analysis demonstrates that the trained network not only provides quite accurate data predictions but can also be conveniently employed to greatly speed up the Jacobian evaluation. It is worth remembering that the implemented inversion framework does not need an extremely accurate Jacobian because this does not affect the estimated posterior probability density (PPD) but only the computational efficiency of the probabilistic sampling (e.g., the acceptance probability value and the convergence towards the steady state).

We now describe the results obtained in three different inversion experiments all running in the compressed model and data spaces:

- Test 1: Gradient-based Markov chain Monte Carlo (GB-MCMC) inversion in which the RNN forward is used for both the likelihood evaluation and the Jacobian computation via a forward FD scheme.
- Test 2: GB-MCMC inversion in which the FE forward is used for both the likelihood evaluation and the Jacobian computation with a forward FD scheme.

- Test 3: gradient-free MCMC inversion in which the differential evolution Markov chain (DEMC) is used to sample the PPD. The FE code provides the forward operator. This method is the same as the one presented in Vinciguerra *et al.* (2021).

The DEMC used in test 3 is a modification of the standard random walk Metropolis sampling, and it makes use of multiple and interactive chains to improve the efficiency of the probabilistic sampling (see, Vrugt, 2016). For tests 1 and 2, the inversions run for 300 iterations with a burn-in of 10 iterations. Test 3 uses 3000 iterations with a burn-in of 1000. In all cases, 20 chains are used to explore the parameter space, where each chain starts from a different model drawn at random from the prior. The use of multiple chains reduces the risk of entrapment in local maxima of the posterior density and also increases the independence of the samples used to compute the PPD. To assess the quality of the results we first evaluate the most likely solutions and the associated uncertainty. The potential scale reduction factor (PSRF; Brooks and Gelman, 1998) is also used to assess the converge of the

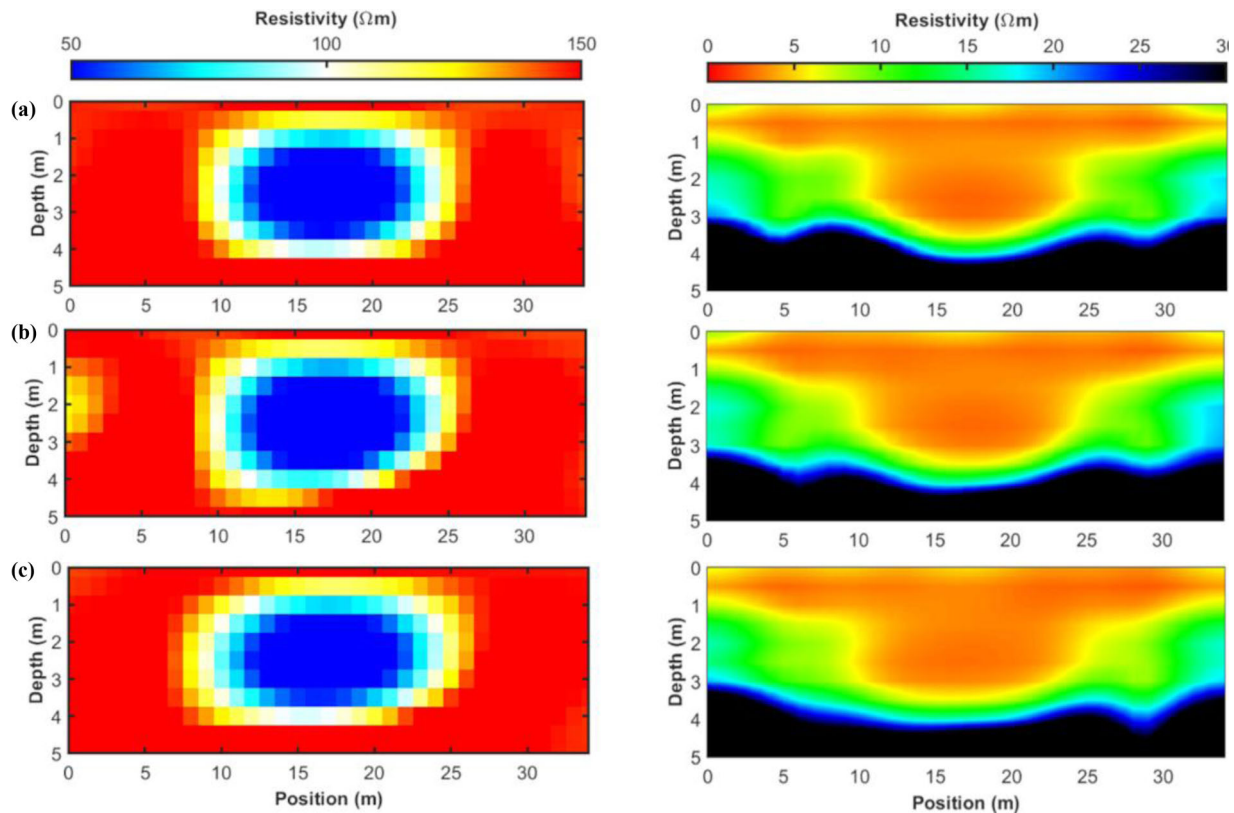


Figure 10 The most likely solutions (left) and the associated posterior standard deviations (right) for the different inversion tests. (a) Test 1. (b) Test 2. (c) Test 3.

sampleings towards a stable posterior model. We also compute the autocorrelations on some model parameters to illustrate that the GB-MCMC sampling decreases the correlations of successively sampled models, thus decreasing the number of samples needed to get accurate uncertainty appraisals. Indeed it is known that the approximation error of the Markov chain is inversely proportional to the number of independent samples. Therefore, for highly correlated samples, the convergence to a stable posterior is usually slower (MacKay, 2003).

Figure 10 shows that all the three tests provide very similar and congruent most likely solutions and uncertainty estimations: the low rectangular resistivity body is successfully located and, as expected, the posterior uncertainties are lower in the central and well-illuminated part of the model and increase towards the bottom and lateral edges of the investigated area due to lower parameter illumination. For a more complete overview of the results, Figures 11 and 12 compare the marginal prior, posterior and true model parameter values in the DCT domain obtained in tests 1 and 2. All the considered 15 DCT coefficients are displayed. The very similar posterior

evaluations estimated in the two tests prove the reliability of the presented approach.

Figure 13(a–c) compares the observed data with the apparent resistivity pseudosections generated on the most likely solution of Figure 10(a) when the RNN and FE codes are employed. On the one hand, the similarity between Figure 13(a) and (b) illustrates that the predictions provided by the implemented inversion successfully reproduce the observed data. Moreover, the good agreement between Figure 13(b) and (c) is a further demonstration of the capability of the trained RNN to predict the forward mapping for a model not seen during the learning stage. Figure 13(d–e) shows instead the predicted data associated with tests 2 and 3, respectively. In all cases, the predicted model successfully reproduces the observations.

In Figure 14 we compare the evolution of the negative log-likelihood, the evolution of the PSRF over iterations, and we also show some examples of autocorrelations estimated from the sampled models for some DCT parameters. From the data likelihood evolution, we conclude that both GB-MCMC inversions reach the steady state within 10 iterations, while 300 iterations are needed by the DEMC algorithm. We can

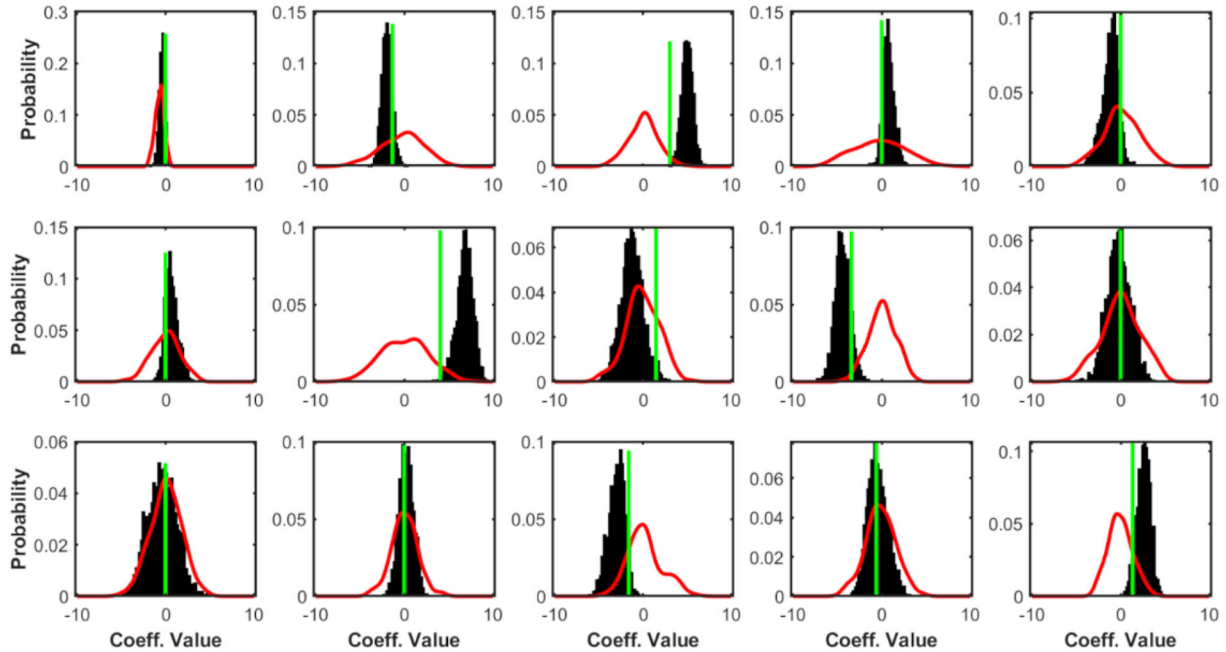


Figure 11 Inversion results in the DCT space for test 1. Each plot refers to one of the fifteen considered DCT coefficients. The red lines represent the marginal priors, the black bars are the marginal posteriors and the green lines indicate the true DCT parameter values.

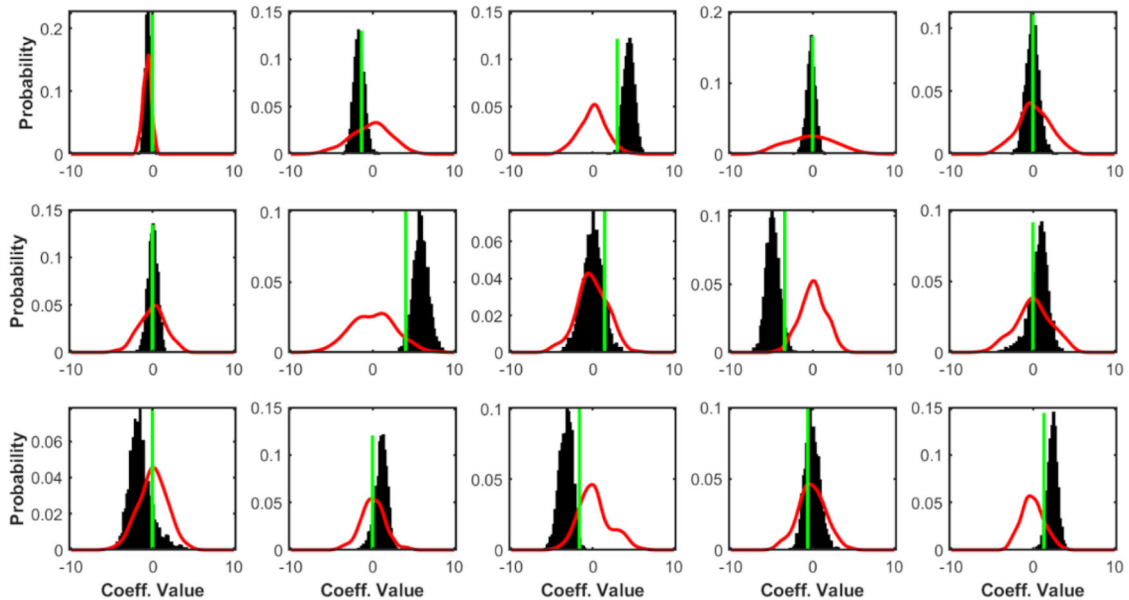


Figure 12 As in Figure 11 but for the second inversion test.

also observe that the two considered GB-MCMC implementations reach the steady state in the same number of iterations, thus demonstrating that the RNN forward provides quite accurate Jacobian matrices comparable to those computed with the FE code. The similar data likelihood values attained in the three tests prove that in all cases the mod-

els sampled after the burn-in period reproduce the observed data with the same accuracy. The PSRF evolutions for all the DCT model parameters illustrate that just 150 iterations are needed by the GB-MCMC inversions to reach a stable PPD estimation, while 1500 iterations are needed by the DEMC. Finally, the comparison of the autocorrelation functions

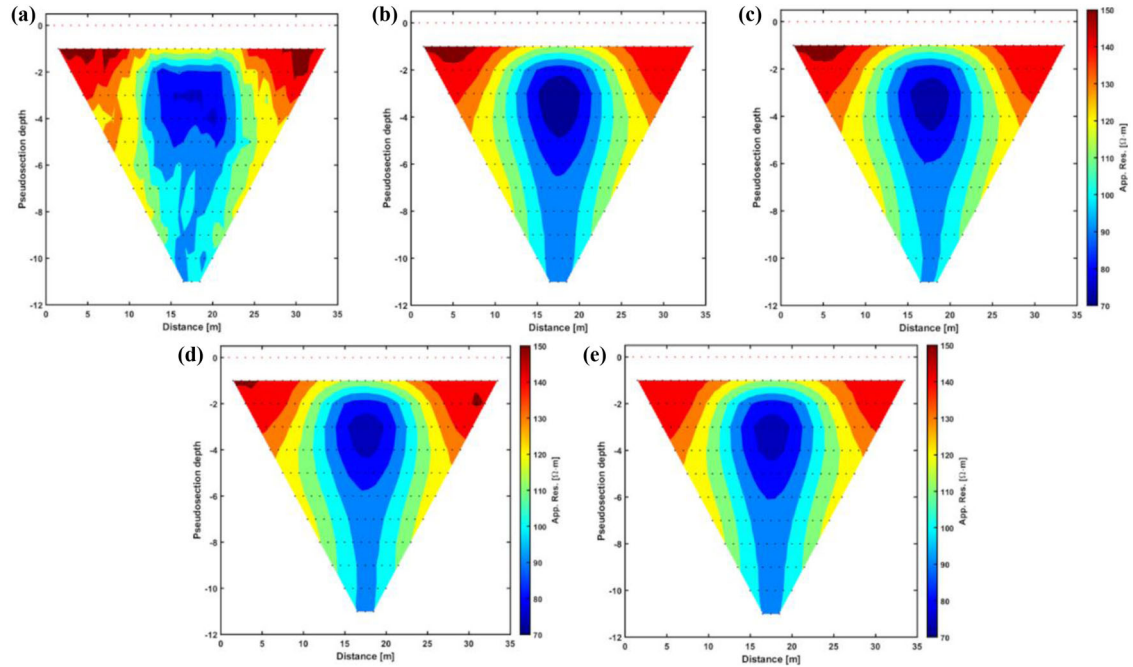


Figure 13 (a) Observed data. (b) Apparent resistivity pseudosection computed on the model shown in Figure 10(a) when the RNN forward operator is used. (c) Apparent resistivity pseudosection computed on the model shown in Figure 10(a) when the FE forward operator is used. (d) Predicted data computed on the most likely model of Figure 10(b). (e) Predicted data computed on the most likely model of Figure 10(c).

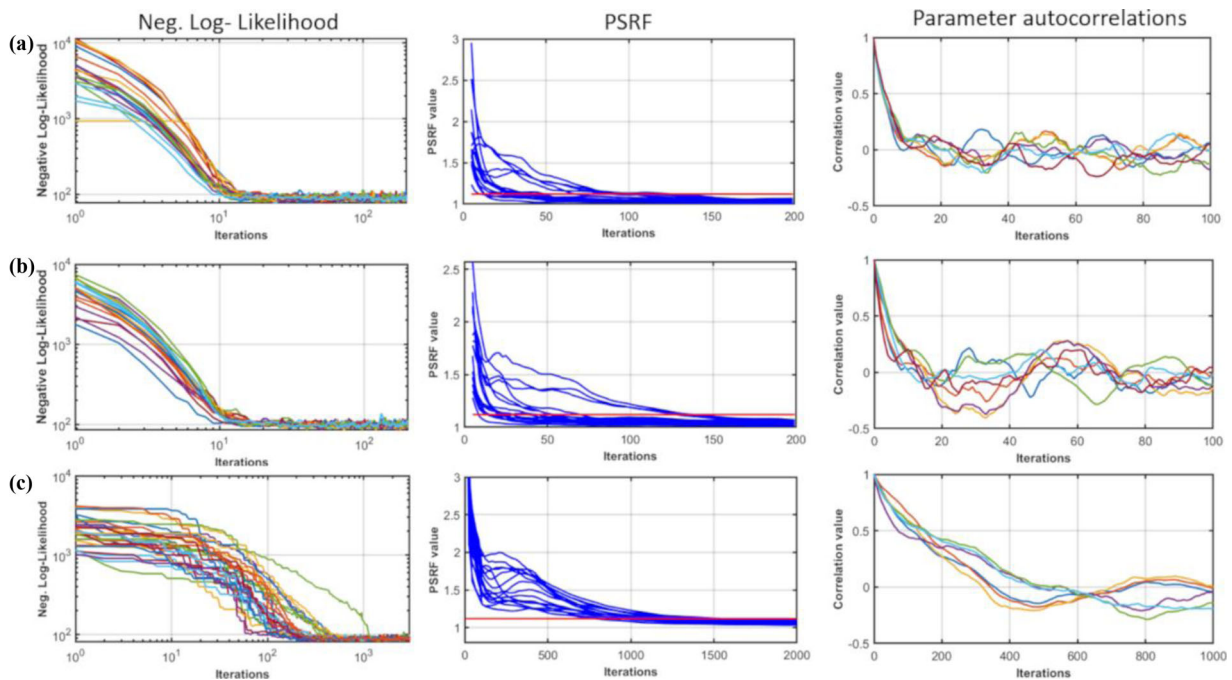


Figure 14 From left to right: Evolution of the negative log-likelihood for the three tests; PSRF value in which the red line indicates the threshold of convergence fixed at 1.1; normalized autocorrelation computed from the sampled models and for some parameters. (a) Test 1. (b) Test 2. (c) Test 3.

Table 1 Summary of the inversion results for the three tests

	Time per iteration considering all the 20 chains (s)	Iterations to converge	Time to converge (minutes)	RMSE data	Coverage ratio (90%)
Test 1	≈ 0.8 s	150	≈ 2	6.25	86.7%
Test 2	≈ 200 s	150	≈ 500	6.17	87.2%
Test 3	≈ 24 s	1500	≈ 600	6.14	87.0%

demonstrates that the correlation drops to zero in 10 iterations in the two GB-MCMC inversions, while hundreds of iterations are needed when the DEMC sampling is adopted. On the one hand, this proves that both the GB-MCMC inversions provide maximally decoupled models, despite the approximated forward employed in test 1. On the other hand, this also illustrates that the GB-MCMC algorithm significantly improves the efficiency of the probabilistic sampling compared to the DEMC. Indeed, the high correlation between successively sampled models is responsible for the slower convergence of the DEMC towards a stable PPD.

As a final and more quantitative assessment of the results, we list in Table 1 and for all the tests, the computing time per iteration (considering all the employed Markov chains), the number of iterations to converge as indicated by the inspection of the PSRF, the time needed to converge toward a stable PPD, the root-mean-square errors (RMSE) between observed and data generated on the most likely solutions previously shown, and finally the 90% coverage ratio. The computing times refer to the hardware resources previously described. We remind that the 90% coverage ratio quantifies the percentage of resistivity values in the true model that fall within the 90% confidence interval as estimated by the probabilistic inversion. All three inversions give very similar data predictions and accuracy in the estimated posterior uncertainties. The major differences concern the times to complete a single iteration and the time needed to converge: In test 1 a single iteration is completed in just 0.8 s. If the RNN forward is replaced by the FE code, this time increases by more than two orders of magnitude, while the DEMC completes a single iteration in 20 s. If we consider the number of iterations needed to converge (i.e., then to achieve a PSRF < 1.1 for all the model parameters) these computing times translate into dramatic differences: in test 1 the probabilistic inversion is completed in 2 minutes, while several hours are needed by the other two approaches. Note that in test 2 the GB-MCMC converges in a much lower number of iterations than the DEMC, but both methods result in similar computational efforts because of the extra time

needed by the gradient-based sampling for the Jacobian evaluation.

These results demonstrate that the proposed approach not only drastically reduces the computational workload of the probabilistic sampling but, more importantly, also provides model estimations and uncertainty assessments comparable to those achieved when an accurate FE modelling is employed or when a more standard, gradient-free MCMC sampling is adopted. As a final remark, we point out that the acceptance probability during the sampling stage for both the GB-MCMC inversions oscillates around 0.8–0.85, while it reduces to 0.25–0.3 during the DEMC inversion. This also indicates that the infusion of the gradient and Hessian information into the sampling framework reduces the time wasted to perform forward evaluations for models that will be rejected by the Metropolis–Hasting rule.

FIELD DATA APPLICATION

We now apply the presented approach to invert a field dataset acquired for levee monitoring along the Parma river in Colorno (Italy). We refer the interested reader to Hojat *et al.* (2019b) for more information about the study area. We invert a single dataset acquired with electrodes buried in a 0.5 m deep trench and employing the Wenner configuration using 48 electrodes with the unit electrode spacing of $a = 2.0$ m. The dataset is corrected for the effect of the soil covering the electrodes (Hojat *et al.*, 2021). The investigated site covers an area that is 94 m wide and 14 m deep, and it is discretized with rectangular cells with vertical and lateral dimensions of 1.0 m and 2.0 m, respectively. This configuration results in $15 \times 47 = 705$ resistivity values to be estimated from 360 data points.

To define the prior assumptions we exploited both the available geological information about the investigated area and the multiple data and associated predicted resistivity sections obtained during the permanent monitoring (see Hojat *et al.*, 2019b). We still employ a non-parametric prior and a

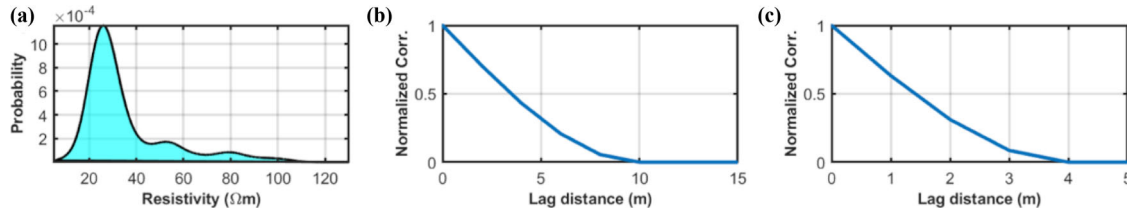


Figure 15 (a) Non-parametric prior distribution for the field data inversion. (b c) The spatial correlation functions associated with the assumed 2D variogram model for the horizontal and vertical directions, respectively.

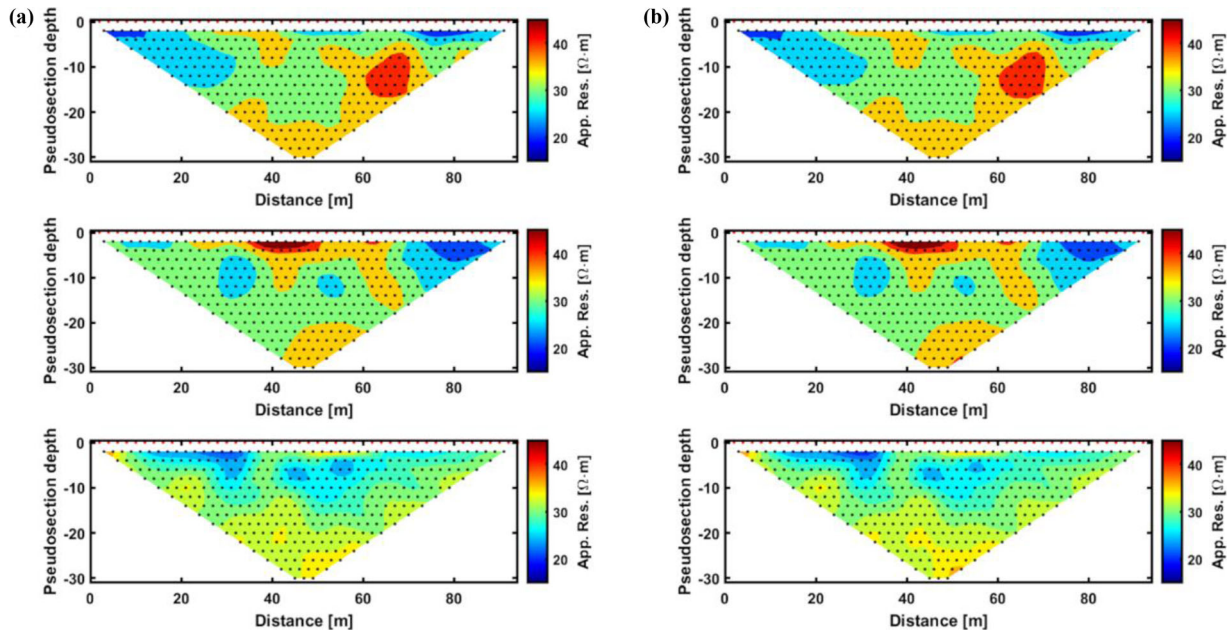


Figure 16 Some comparisons between (a) the network responses and (b) the desired output extracted from the validation set.

spatial variability pattern described by a spherical variogram with lateral and vertical ranges equal to 10 m and 4 m, respectively (Fig. 15). The prior aims to properly model the multimodality of the resistivity distribution in the investigated area where different litho-facies are expected: a low-resistivity clay body that around 2–3 m depth hosts more permeable formations with higher resistivity values associated with sands and gravels.

As in the synthetic case, the residual neural network (RNN) forward has been trained using training and validation sets with 2000, and 500 prior realizations. We maintain the same RNN architecture previously described being the only difference in the dimension of the input image and the output response (in this case a model with 15 rows and 47 columns and 360 apparent resistivity values). In Figure 16 the similarity of the RNN predictions with the corresponding finite-element (FE) responses extracted from the validation set

proves the capability of the trained machine learning model to predict the forward relation.

Similar to Aleardi *et al.* (2021), 150 discrete cosine transform coefficients have been used to compress both model and data space. Figure 17 shows box plots of the time needed for the Jacobian computation with different strategies. Different from the previous synthetic test we limit our attention to the Jacobian computed in the compressed space via a forward FD scheme with the RNN and FE forwards, and the Jacobian computed in the full domains with the adjoint-state method. Note that the considerable time per forward evaluation and the large parameter space make the computation of the Jacobian with the FD + FE strategy impractical in the full model space.

Figure 18 demonstrates that the RNN approximated forward provides accurate Jacobian evaluations with a computational effort three orders of magnitude lower than that

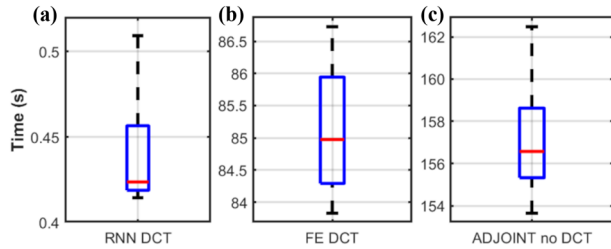


Figure 17 Box plots showing the computing times for the Jacobian computation using different approaches in the field data test: (a) RNN forward + FD scheme in the compressed domain. (b) FE forward + FD scheme in the compressed domain. (c) Adjoint-state method applied in the uncompressed space. Note the different scales along the vertical axes.

associated with the adjoint-state and the FE code. From this analysis, it turns out that the application of the gradient-based Markov chain Monte Carlo (GB-MCMC) inversion will be computationally unfeasible on the employed computational resources without the RNN forward approximation. As discussed by Vinciguerra *et al.* (2021) even the gradient-free DEMC algorithm running in the compressed model space needs several days of computing time to complete this field data inversion. For these reasons, we here compare the results achieved by the GB-MCMC with the RNN forward with the predictions of a standard deterministic (i.e., gradient-based) inversion.

We run the GB-MCMC inversion for 500 iterations, 20 chains, with a burn-in of 10 and starting from prior realizations. Figure 19 compares the predictions yielded by the deterministic method with the most likely solution estimated by the presented approach along with the associated posterior standard deviation. Similar and congruent outcomes are achieved in both cases: the inversion predicts a sand-gravel body at shallow depth hosted in shales. As expected, the uncertainty tends to increase in correspondence of the high resistivity body and at the lateral and bottom edges of the model.

Figure 20 shows the observed data and the data generated by the RNN and FE codes on the model of Figure 19(b), together with the apparent resistivity data computed with the FE code from the deterministic solution (Fig. 19a). Both inversions accurately reproduce the field measurements with very minor differences in the data generated by the RNN and FE codes. This demonstrates that the trained network can properly approximate the forward relation even for realistic resistivity models not seen during the learning procedure.

Finally, Figure 21(a, b) illustrates the evolutions of the potential scale reduction factor for all the model parameters and the negative log-likelihood for the 20 chains. As in the syn-

thetic example, the probabilistic inversion reaches the steady state in 10 iterations, while the full convergence is attained in 400 iterations. Figure 21(c) illustrates that also in this case the GB-MCMC samples maximally decoupled models as the autocorrelation values drop to zero in very few iterations.

As a final remark, we point out that the acceptance probability during the GB-MCMC inversion oscillates around 0.75–0.85, while one single iteration is completed in 2 seconds. This means that the convergence for all the unknown parameters is achieved in 13 minutes, approximately. Since the gradient-based inversion is completed in 6 minutes, we deem that the implemented approach makes the computational cost of the probabilistic sampling comparable with that of a deterministic inversion.

DISCUSSION

Reducing the computational cost of a probabilistic electrical resistivity tomography (ERT) inversion is needed to make this approach more appealing than popular local inversion algorithms. The Bayesian framework provides crucial information regarding the uncertainties affecting the recovered solution. Such estimated model uncertainties can be used to generate different subsurface scenarios in agreement with the experimental data and the prior assumptions. Such models extracted from the posterior add an extra level of information over gradient-based solutions and contribute to a more informed decision-making process in many ERT applications (e.g., monitoring applications, or 3D inversions). For example, the 3D ERT is usually a severely under-determined problem with more unknowns than observations, in which the deterministic solution is highly affected by the initial choice of the starting model (Aguzzoli *et al.*, 2021). In this context, an accurate uncertainty estimation is of fundamental importance to give hints on the ambiguity affecting the predicted resistivity values.

Therefore, this work was aimed at implementing a sampling algorithm for accurate and fast uncertainty assessments in ERT inversion. To this end, we combined the sampling efficiency of a gradient-based Markov chain Monte Carlo (MCMC) algorithm with the fast forward evaluations provided by a trained residual neural network. The network was successfully trained with just 2000 examples, while 500 examples formed the validation set. Note that these values are much lower than the number of forward evaluations needed by the probabilistic sampling. The modelling error introduced by the approximated forward was properly taken into consideration and included in the data covariance

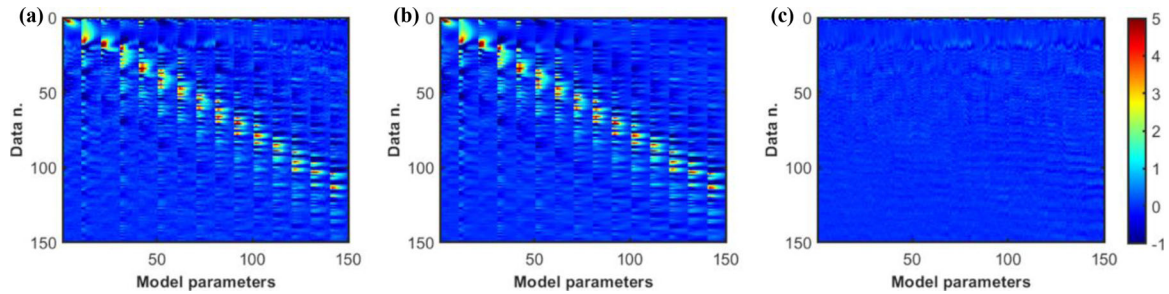


Figure 18 (a) Jacobian computed with the FD scheme and the RNN forward in the compressed model and data spaces. (b) Projection onto the DCT domain of the Jacobian computed with the adjoint method. (c) Difference between (a) and (b). The Jacobian has been computed around a model drawn from the prior.

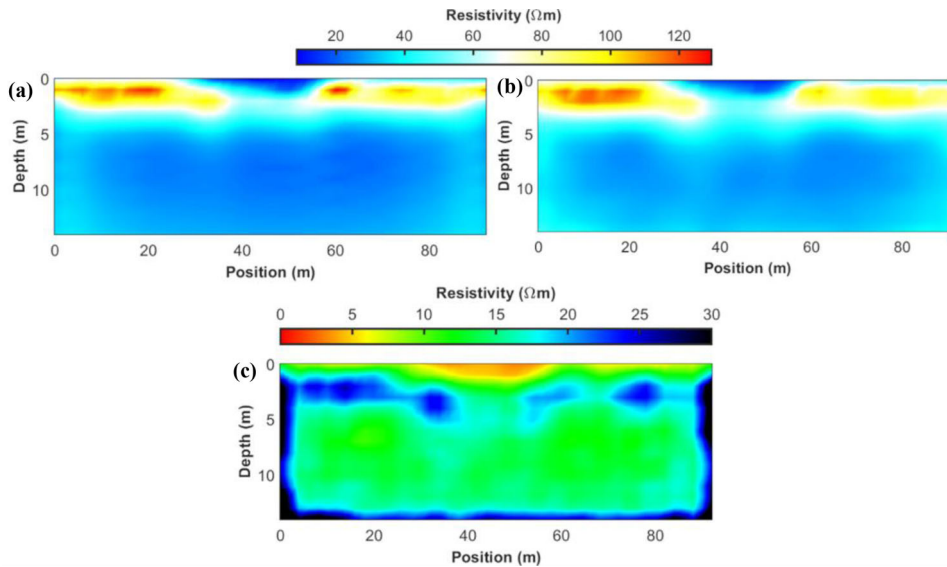


Figure 19 (a) Model predicted by the gradient-based inversion. (b) Most likely solution provided by the GB-MCMC inversion. (c) Posterior standard deviation estimated by the probabilistic inversion.

matrix under a Gaussian assumption. In this regard, the number of examples in the validation sets should not be too small to avoid the underestimation of the modelling error. In the following examples, we have found that 500 models are enough to get an accurate representation of the theoretical error introduced by the network. However, larger validation sets are likely needed for wider investigated areas and in 3D applications. We also point out that a proper network architecture to reliably approximate the forward operator is not that hard to find at least in the 2D scenario we considered here. Indeed, in Aleardi *et al.* (2022) we found that different architectures also provide quite accurate data predictions.

The probabilistic sampling was guided by the gradient and Hessian information of the negative log-posterior. This generates proposal moves that are locally a Gaussian approx-

imation of the posterior probability density (PPD). The proposal is analytically tractable, and this makes the generation of the proposed models and the evaluation of the proposal ratio straightforward. The gradient information guides the sampling towards solutions with higher data likelihood values, while the random perturbation term avoids entrapments in local maxima of the posterior density. In the present work, the number of unknowns in the uncompressed space (few hundreds) makes the projection of the Jacobian from the full space to the compressed domain computationally feasible. However, in the case of larger model spaces (i.e., 3D applications), it would be convenient to compute the Jacobian directly in the compressed domain to avoid the extra workload needed for the projection (Laloy *et al.*, 2019).

A good compromise between exploitation and exploration can be found by setting the two hyperparameters λ and

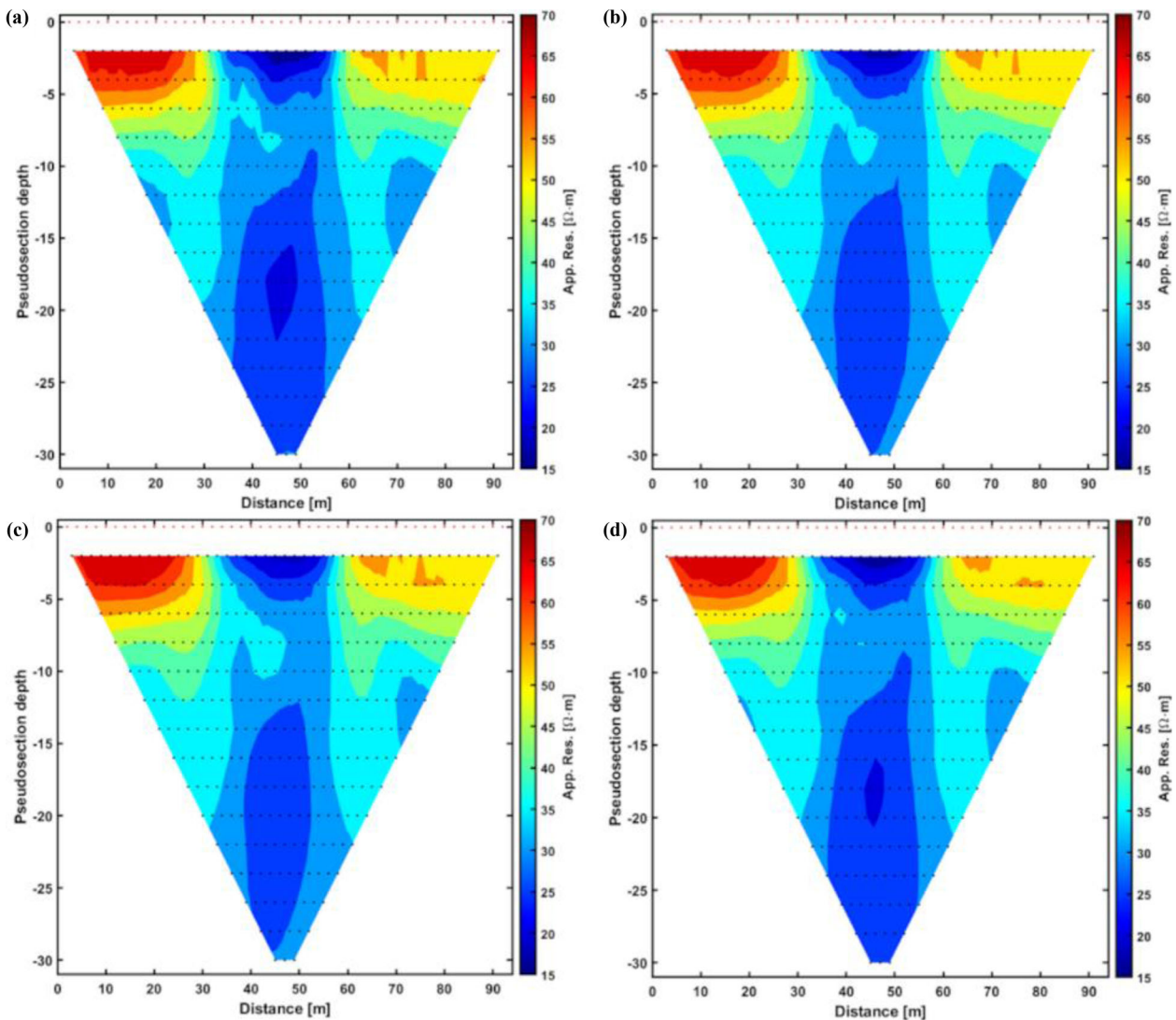


Figure 20 (a) Observed data. (b) Apparent resistivity pseudosection computed on the model shown in Figure 19(b) when the RNN forward operator is used. (c) Apparent resistivity pseudosection computed on the model shown in Figure 19(b) when the FE forward operator is used. (d) Predicted data generated by the deterministic solution with the FE code.

μ . Their proper setting is important for the efficiency of the sampling (i.e., a poorly chosen parameter combination would slow down the convergence toward stable uncertainty assessments) but does not alter the final estimated PPD. We adopted a trial-and-error procedure to set the λ and μ values with the aim to get an acceptance probability of around 0.8 during the sampling stage. Note that the very limited computational cost of the implemented method greatly reduces the human effort for the hyperparameter setting. In the many inversion tests carried out, we also found that the desired acceptance can be achieved by many λ and μ combinations, and hence a proper selection of these parameters is not that hard to find.

The discrete cosine transform (DCT) compression was used to mitigate the ill-posedness of the ERT inverse problem, to reduce the dimension of the model space, and to also decrease the computational efforts for the Hessian and gradient manipulation, and the Jacobian computation via a finite-difference scheme. Another possibility to further reduce the computational workload is to compute the Jacobian matrix only for the very first iterations (i.e. the burn-in period) and hence use the same Jacobian during the sampling stage. This recipe should still guarantee a faster convergence toward the steady state and a more efficient sampling with respect to gradient-free MCMC algorithms. Since the DCT can be

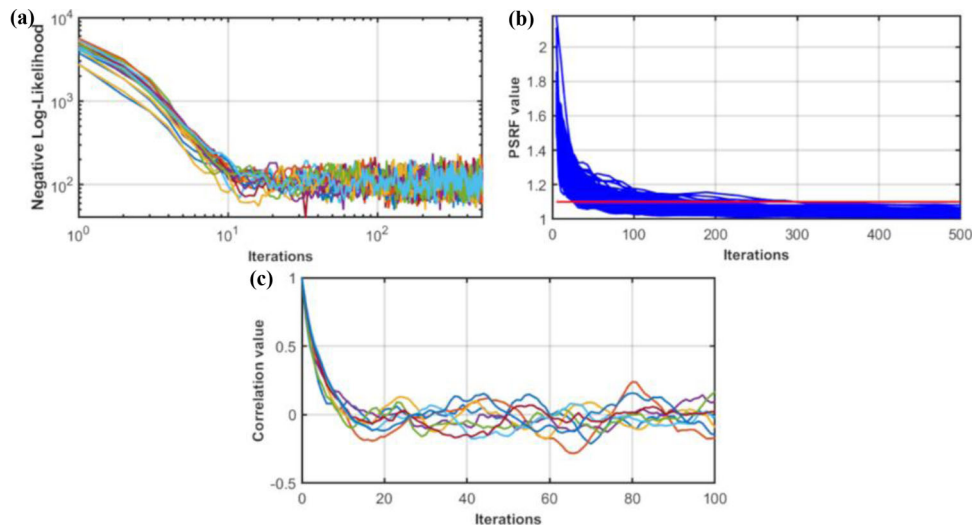


Figure 21 (a) Evolution of the negative log-likelihood for the 20 chains. (b) PSRF value in which the red line indicates the threshold of convergence fixed at 1.1. (c) Normalized autocorrelations computed from the sampled models for some parameters.

applied to any multidimensional signal, the implemented method can be also extended to 3D models. In addition, the linearity of the DCT greatly simplify the proper setting of the number of basis functions to consider (the p and q coefficients) because only some prior realizations are needed to analyse how the recovered model variability is related to the number of retained coefficients (see Aleardi *et al.*, 2021). However, the main limitation of this compression method is related to the large number of coefficients to be retained to accurately represent sharp spatial contrasts in the resistivity values. In these contexts, the DCT can be replaced by other nonlinear compression strategies (i.e., deep generative models) that can better preserve nonlinear features in the solution (Laloy *et al.*, 2017; Jiang and Jafarpour, 2021). However, these strategies can severely complicate the geometry of the posterior distribution (i.e. many local maxima), thus reducing the convergence speed of the Monte Carlo sampling (Lopez-Alvis *et al.*, 2021). Therefore, the optimal choice for the compression method to be used is case dependent as a reasonable compromise must be found between the desired model resolution and the computational cost of the entire inversion procedure.

CONCLUSIONS

We presented a Bayesian approach to the electrical resistivity tomography inversion in which a machine learning algorithm, a compression strategy and a gradient-based Markov chain Monte Carlo (GB-MCMC) were combined to drastically reduce the computational cost of the probabilistic sam-

pling. The regression ability of residual neural network (RNN) was used to approximate the forward operator; the discrete cosine transform was used to compress data and model spaces, whereas the gradient and Hessian information of the posterior density were exploited to increase the efficiency of the MCMC sampling. The synthetic inversions indicated that the presented inversion reaches stable uncertainty estimations with a much lower number of iterations than a gradient-free MCMC algorithm, while the RNN forward reduces of two orders of magnitude the computational cost of the probabilistic inversion compared to the case in which a common FE modelling code is employed. The field example not only showed that the GB-MCMC and the deterministic inversions provide similar final model predictions and data fitting but they are also characterized by comparable computational costs. Therefore, our inversion strategy makes the application of a probabilistic ERT inversion also possible on limited hardware resources.


ACKNOWLEDGEMENTS

Open Access Funding provided by Universita degli Studi di Pisa within the CRUI-CARE Agreement.

DATA AVAILABILITY STATEMENT

Data are available on request from the corresponding author.

ORCID

Mattia Aleardi  <https://orcid.org/0000-0003-1433-0281>

REFERENCES

- Aguzzoli, A., Fumagalli, A., Scotti, A., Zanzi, L. and Arosio, D. (2021) Inversion of synthetic and measured 3D geoelectrical data to study the geomembrane below a landfill. In *4th Asia Pacific Meeting on Near Surface Geoscience & Engineering*, 2021(1), 1–5.
- Aleardi, M. (2019) Using orthogonal Legendre polynomials to parameterize global geophysical optimizations: applications to seismic-petrophysical inversion and 1D elastic full-waveform inversion. *Geophysical Prospecting*, 67(2), 331–348.
- Aleardi, M. (2020) Combining discrete cosine transform and convolutional neural networks to speed up the Hamiltonian Monte Carlo inversion of pre-stack seismic data. *Geophysical Prospecting*, 68(9), 2738–2761.
- Aleardi, M. and Salusti, A. (2020) Hamiltonian Monte Carlo algorithms for target-and interval-oriented amplitude versus angle inversions. *Geophysics*, 85(3), R177–R194.
- Aleardi, M., Vinciguerra, A., Hojat, A. and Stucchi, E. (2022) Probabilistic inversions of electrical resistivity tomography data with a machine learning-based forward operator. *Geophysical Prospecting*. <https://doi.org/10.1111/1365-2478.13189>.
- Aleardi, M., Vinciguerra, A. and Hojat, A. (2020) A geostatistical Markov chain Monte Carlo inversion algorithm for electrical resistivity tomography. *Near Surface Geophysics*, 19(1), 7–26.
- Aleardi, M., Vinciguerra, A. and Hojat, A. (2021) Ensemble-based electrical resistivity tomography with data and model space compression. *Pure and Applied Geophysics*, 1–23.
- Aster, R.C., Borchers, B. and Thurber, C.H. (2018) *Parameter Estimation and Inverse Problems*. Cambridge: Elsevier.
- Bièvre, G., Oxarango, L., Günther, T., Goutaland, D. and Massardi, M. (2018) Improvement of 2D ERT measurements conducted along a small earth-filled dyke using 3D topographic data and 3D computation of geometric factors. *Journal of Applied Geophysics*, 153, 100–112.
- Brooks, S.P. and Gelman, A. (1998) General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7(4), 434–455.
- Curtis, A. and Lomax, A. (2001) Prior information, sampling distributions, and the curse of dimensionality. *Geophysics*, 66(2), 372–378.
- Dahlin, T. (2020) Geoelectrical monitoring of embankment dams for detection of anomalous seepage and internal erosion – experiences and work in progress in Sweden. Fifth International Conference on Engineering Geophysics (ICEG), Al Ain, UAE. <https://doi.org/10.1190/iceg2019-053.1>.
- Dejtrakulwong, P., Mukerji, T. and Mavko, G. (2012) Using kernel principal component analysis to interpret seismic signatures of thin shaly-sand reservoirs. In *SEG Technical Program Expanded Abstracts 2012*. Society of Exploration Geophysicists.
- Fichtner, A. and Simuté, S. (2018) Hamiltonian Monte Carlo inversion of seismic sources in complex media. *Journal of Geophysical Research: Solid Earth*, 123(4), 2984–2999.
- Fichtner, A. and Zunino, A. (2019) Hamiltonian nullspace shuttles. *Geophysical Research Letters*, 46(2), 644–651.
- Fichtner, A., Zunino, A. and Gebraad, L. (2019) Hamiltonian Monte Carlo solution of tomographic inverse problems. *Geophysical Journal International*, 216(2), 1344–1363.
- Gebraad, L., Boehm, C. and Fichtner, A. (2020) Bayesian elastic full-waveform inversion using Hamiltonian Monte Carlo. *Journal of Geophysical Research: Solid Earth*, 125(3), e2019JB018428.
- Glorot, X. and Bengio, Y. (2010) Understanding the difficulty of training deep feedforward neural networks. In Proceedings of the thirteenth international conference on artificial intelligence and statistics, 249–256.
- Goodfellow, I., Bengio, Y. and Courville, A. (2016). *Deep Learning*. Cambridge: MIT Press.
- Haario, H., Saksman, E. and Tamminen, J. (2001) An adaptive Metropolis algorithm. *Bernoulli*, 7(2), 223–242.
- Haario, H., Laine, M., Mira, A. and Saksman, E. (2006) DRAM: efficient adaptive MCMC. *Statistics and computing*, 16(4), 339–354.
- Hansen, T.M. and Cordua, K.S. (2017) Efficient Monte Carlo sampling of inverse problems using a neural network-based forward: applied to GPR crosshole traveltime inversion. *Geophysical Journal International*, 211(3), 1524–1533.
- Hermans, T. and Paepen, M. (2020) Combined inversion of land and marine electrical resistivity tomography for submarine groundwater discharge and saltwater intrusion characterization. *Geophysical Research Letters*, 47(3), e2019GL085877.
- Holmes, C., Krzysztof, L. and Pompe, E. (2017) Adaptive MCMC for multimodal distributions. Technical report. <https://pdfs.semanticscholar.org/c75d/f035c23e3c0425409e70d457cd43b174076f.pdf>.
- Hojat, A., Arosio, D., Di Luch, I., Ferrario, M., Ivanov, V.I., Longoni, L., et al. (2019a) Testing ERT and fiber optic techniques at the laboratory scale to monitor river levees. 25th European Meeting of Environmental and Engineering Geophysics, The Hague, Netherlands, <https://doi.org/10.3997/2214-4609.201902440>.
- Hojat, A., Arosio, D., Longoni, L., Papini, M., Tresoldi, G. and Zanzi, L. (2019b) Installation and validation of a customized resistivity system for permanent monitoring of a river embankment. EAGE-GSM 2nd Asia Pacific Meeting on Near Surface Geoscience and Engineering, Kuala Lumpur, Malaysia. <https://doi.org/10.3997/2214-4609.201900421>.
- Hojat, A., Tresoldi, G. & Zanzi, L. (2021) Correcting the effect of the soil covering buried electrodes for permanent electrical resistivity tomography monitoring systems. 4th Asia Pacific Meeting on Near Surface Geoscience & Engineering, Ho Chi Minh, Vietnam, Online, DOI: 10.3997/2214-4609.202177070.
- Karaoulis, M., Revil, A., Tsourlos, P., Werkema, D.D. and Minsley, B.J. (2013) IP4DI: a software for time-lapse 2D/3D DC-resistivity and induced polarization tomography. *Computers & Geosciences*, 54, 164–170.
- Jiang, A. and Jafarpour, B. (2021) Deep convolutional autoencoders for robust flow model calibration under uncertainty in geologic continuity. *Water Resources Research*, 57(11), e2021WR029754.
- Laloy, E., Héroult, R., Lee, J., Jacques, D. and Linde, N. (2017) Inversion using a new low-dimensional representation of complex binary geological media based on a deep neural network. *Advances in Water Resources*, 110, 387–405.
- Laloy, E., Linde, N., Ruffino, C., Héroult, R., Gasso, G. and Jacques, D. (2019) Gradient-based deterministic inversion of geophysical data with generative adversarial networks: is it feasible?. *Computers & Geosciences*, 133, 104333.

- Lieberman, C., Willcox, K. and Ghattas, O. (2010) Parameter and state model reduction for large-scale statistical inverse problems. *SIAM Journal on Scientific Computing*, 32(5), 2523–2542.
- Liu, M. and Grana, D. (2020) Time-lapse seismic history matching with an iterative ensemble smoother and deep convolutional autoencoder. *Geophysics*, 85(1), M15–M31.
- Lochbühler, T., Breen, S.J., Detwiler, R.L., Vrugt, J.A. and Linde, N. (2014) Probabilistic electrical resistivity tomography of a CO₂ sequestration analog. *Journal of Applied Geophysics*, 107, 80–92.
- Loke, M.H., Papadopoulos, N., Wilkinson, P.B., Oikonomou, D., Simyrdanis, K. and Rucker, D.F. (2020) The inversion of data from very large three-dimensional electrical resistivity tomography mobile surveys. *Geophysical Prospecting*, 68(8), 2579–2597.
- Lopez-Alvis, J., Laloy, E., Nguyen, F. and Hermans, T. (2021) Deep generative models in inversion: the impact of the generator's non-linearity and development of a new approach based on a variational autoencoder. *Computers & Geosciences*, 152, 104762.
- MacKay, D.J. (2003) *Information Theory, Inference and Learning Algorithms*. Cambridge: Cambridge University Press.
- Martin, J., Wilcox, L.C., Burstedde, C. and Ghattas, O. (2012) A stochastic Newton MCMC method for large-scale statistical inverse problems with application to seismic inversion. *SIAM Journal on Scientific Computing*, 34(3), A1460–A1487.
- Menke, W. (2018) *Geophysical Data Analysis: Discrete Inverse Theory*. Orlando, FL: Academic Press.
- Mo, S., Zabarar, N., Shi, X. and Wu, J. (2020) Integration of adversarial autoencoders with residual dense convolutional networks for estimation of non-Gaussian hydraulic conductivities. *Water Resources Research*, 56(2), e2019WR026082.
- Moghadas, D. and Vrugt, J.A. (2019) The influence of geostatistical prior modeling on the solution of DCT-based Bayesian inversion: a case study from chicken creek catchment. *Remote Sensing*, 11(13), 1549.
- Monajemi, H., Donoho, D.L. and Stodden, V. (2016) Making massive computational experiments painless. In 2016 IEEE International Conference on Big Data, 2368–2373.
- Moradipour, M., Ranjbar, H., Hojat, A., Karimi-Nasab, S. and Daneshpajouh, S. (2016) Laboratory and field measurements of electrical resistivity to study heap leaching pad no. 3 at Sarcheshmeh copper mine. 22nd European Meeting of Environmental and Engineering Geophysics, <https://doi.org/10.3997/2214-4609.201602140>.
- Moseley, B., Nissen-Meyer, T. and Markham, A. (2020) Deep learning for fast simulation of seismic waves in complex media. *Solid Earth*, 11(4), 1527–1549.
- Neal, R.M. (2011) MCMC Using Hamiltonian Dynamics. In: Brooks, S., Gelman, A., Jones, G. and Meng, X. (Eds) *Handbook of Markov Chain Monte Carlo*. New York: Chapman and Hall, pp. 113–162.
- Norooz, R., Olsson, P.I., Dahlin, T., Günther, T. and Bernston, C. (2021) A geoelectrical pre-study of Älvkarleby test embankment dam: 3D forward modelling and effects of structural constraints on the 3D inversion model of zoned embankment dams. *Journal of Applied Geophysics*, 191, 104355.
- Plessix, R.E. (2006) A review of the adjoint-state method for computing the gradient of a functional with geophysical applications. *Geophysical Journal International*, 167(2), 495–503.
- Rucker, D.F., Fink, J.B. and Loke, M.H. (2011) Environmental monitoring of leaks using time-lapsed long electrode electrical resistivity. *Journal of Applied Geophysics*, 74(4), 242–254.
- Sambridge, M. (2014) A parallel tempering algorithm for probabilistic sampling and multimodal optimization. *Geophysical Journal International*, 196(1), 357–374.
- Sambridge, M. and Mosegaard, K. (2002) Monte Carlo methods in geophysical inverse problems. *Reviews of Geophysics*, 40(3), 3–1.
- Sen, M.K. and Biswas, R. (2017) Transdimensional seismic inversion using the reversible jump Hamiltonian Monte Carlo algorithm. *Geophysics*, 82(3), R119–R134.
- Sen, M.K. and Stoffa, P.L. (2013) *Global Optimization Methods in Geophysical Inversion*. Cambridge: Cambridge University Press.
- Soares, A. (2001) Direct sequential simulation and cosimulation. *Mathematical Geology*, 33(8), 911–926.
- Song, C., Alkhalifah, T. and Waheed, U.B. (2021) Solving the frequency-domain acoustic VTI wave equation using physics-informed neural networks. *Geophysical Journal International*, 225(2), 846–859.
- Tarantola, A. (2005) Inverse problem theory and methods for model parameter estimation. *Society for Industrial and Applied Mathematics*.
- Turner, B.M. and Sederberg, P.B. (2012) Approximate Bayesian computation with differential evolution. *Journal of Mathematical Psychology*, 56(5), 375–385.
- Uhlemann, S., Wilkinson, P.B., Chambers, J.E., Maurer, H., Merritt, A.J., Gunn, D.A. and Meldrum, P.I. (2015) Interpolation of landslide movements to improve the accuracy of 4D geoelectrical monitoring. *Journal of Applied Geophysics*, 121, 93–105.
- Vinciguerra, A., Aleardi, M., Hojat, A. and Stucchi, E. (2021) Discrete cosine transform for parameter space reduction in bayesian electrical resistivity tomography. *Geophysical Prospecting*, 70(1), 193–209. <https://doi.org/10.1111/1365-2478.13148>.
- 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.
- Whiteley, J., Chambers, J.E. and Uhlemann, S. (2017) Integrated monitoring of an active landslide in lias group mudrocks, north yorkshire, UK. In: Hoyer, S. (Ed.) *GELMON 2017: Fourth International Workshop on GeoElectrical Monitoring*, 27.